ELSEVIER Stochastic Processes and their Applications 71 (1997) 165 185

stochastic processes and their applications

# Mixing times for uniformly ergodic Markov chains

David Aldous[a,]*, László Lovász[b], Peter Winkler[c]

[a]*Department of Statistics, University of California at Berkeley, CA 94 720, USA*
[b]*Department of Computer Science, Yale University, New Haven, CT 06510, USA*
[c]*Bell Laboratories 2C-379, Murray Hill, NJ 07960, USA*

Received 16 January 1996; received in revised form 5 March 1997

**Abstract**

Consider the class of discrete time, general state space Markov chains which satisfy a "uniform ergodicity under sampling" condition. There are many ways to quantify the notion of "mixing time", i.e., time to approach stationarity from a worst initial state. We prove results asserting equivalence (up to universal constants) of different quantifications of mixing time. This work combines three areas of Markov theory which are rarely connected: the potential-theoretical characterization of optimal stopping times, the theory of stability and convergence to stationarity for general-state chains, and the theory surrounding mixing times for finite-state chains. © 1997 Elsevier Science B.V.

*Keywords:* Markov chain; Minorization; Mixing time; Randomized algorithm; Stopping time

## 1. Introduction

Our topic lies near the intersection of three different areas of the theory of (discrete time, general state space) Markov chains.

(a) Potential theory, as treated in e.g. Revuz (1984) or Dellacherie and Meyer (1983). This theory classically focused on transient chains, but does include results on *recurrent potential* and its relation to hitting times for recurrent chains, which are our concern (see also Syski, 1992).

(b) The theory of convergence to stationarity for general state space chains, treated in Orey (1971) and Nummelin (1984) and in particular given a recent very clear exposition by Meyn and Tweedie (1993).

(c) The theory surrounding *mixing times*, i.e. quantitative measures of times to approach stationarity, for finite-state chains. This is treated (in the reversible setting) in the forthcoming book Aldous and Fill (1997). See also Diaconis (1988) for the case of random walks on groups, and Sinclair (1993) and Motwani and Raghavan (1995) for uses in the theory of algorithms.

* Corresponding author. E-mail: aldous@stat.berkeley.edu.

*PII* S0304-4149(97)00037-9

These areas have developed rather independently, and the connections are not easy to find in the monographs above. The purpose of this paper is to record two related results which explicitly use aspects of all three areas. These results (Theorems 1 and 3) assert the equivalence (up to universal constants) of different formalizations of "mixing time" in (essentially) the context of uniformly ergodic general state-space chains.

In Section 1.1 we recall the underlying algorithmic motivation for studying mixing times. In Section 1.2 we describe, as mathematical background, known results from each of the three areas (a)–(c) above. Section 1.3 states our new results, and Section 1.4 interprets the conceptual significance of the new results.

### *1.1. Mixing times and randomized algorithms*

One motivation for the study of mixing times comes from computer science, more exactly from the analysis of sampling algorithms, which has been an active area over the last ten years. In randomized algorithms solving a variety of computational tasks (approximate enumeration, volume computation, integration, simulated annealing, generation of contingency tables etc.) the key element is to sample from a given distribution $\pi$ over a known but large and complicated set. The basic method is to construct an ergodic Markov chain with stationary distribution $\pi$, and then run the chain for an appropriately large number of steps. The details vary according to the goal of the algorithm, which might be to estimate an average $\int f \mathrm{d}\pi$, or to bound the $\pi$-probability of some set of unlikely states, or to generate typical realizations from $\pi$ for illustrative purposes. The number of steps required by a particular algorithm (as a function of the Markov chain) will depend on some algorithm-specific notion of "mixing time", i.e. the number of steps until the distribution approaches stationarity. Three such notions are mentioned below. Even for the more restricted issue of quantifying the distance between the time-$t$ distribution and $\pi$ there are several answers: total variation distance (i.e. $l_1$ distance for densities), the analogous $l_2$ or $l_\infty$ distances, Kullback–Leibler distance, etc.

In a sampling algorithm, we may want to generate a single state from the stationary distribution, starting from some fixed state (determined by the rest of the algorithm). The minimum mean time to do so is a definition of a mixing time from a given state. If we do not have more information about the starting state, we have to use the maximum over all starting states, which we will call the *mixing time* $\mathscr{T}_{\mathrm{mix}}$ (precise definitions will be given later). But it may be the case that we need to generate several 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* $\mathscr{T}_{\mathrm{reset}}$. Alternatively, we may use the Markov chain to find an element from a specified, 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* $\mathscr{T}_{\mathrm{set}}$.

This paper is motivated by a foundational question.

> Does it make sense to undertake a mathematical analysis of a given chain being used in some sampling algorithm, without paying attention to the algorithmic use of the samples?

If the mixing times for the given chain associated with different algorithms were incomparable, then it would not make sense. Fortunately this is not the case. For reversible chains, it has been known for a long time (Aldous, 1982) that various mixing times (including the three mixing times above) are within absolute constant factors of each other, assuming that they are finite at all. The case of non-reversible chains is a bit more complicated, but results of this paper show that many mixing measures fall into three groups only, where measures in the same group are within absolute constant factors of each other, one group is always "above" the other two, and these two are related in an interesting way through time-reversal.

To discuss a celebrated example, consider a convex body $K$ in $\mathbb{R}^n$ and suppose that we want to generate a uniformly distributed point in it. We assume that the body is in isotropic position (i.e. a uniform random point $(X_1, \ldots, X_n)$ of $K$ has $EX_i = 0$ and $EX_iX_j = 1_{(i=j)}$). Choose an appropriately small $\delta > 0$, say $\delta = 1/\sqrt{n}$, and start a random walk from a point $s$ by stepping distance $\delta$ in a uniformly chosen random direction. (If this step takes us outside the body, we choose another direction, until we finally are able to make a step.) The stationary distribution of this walk is close to uniform. Mixing properties of this walk were analyzed in several papers (Lovász and Simonovits, 1993; Kannan et al., 1997). It turns out that the mixing time of the walk is $O(n^3)$ independently of the body (it may even be $O(n^2)$; this is an open question). Our results than say that a number of other mixing measures have the same order of magnitude.

Other algorithmic contexts where mixing times have been studied include sampling from log-concave distributions (Frieze, 1994), matchings in graphs (Jerrum and Sinclair, 1989; Motwani and Raghavan, 1995), and Metropolis-type algorithms (Diaconis and Saloff-Coste, 1996).

### *1.2. Background mathematical results*

We set the stage by first quoting one standard theorem from each of the three areas mentioned initially. None of these theorems is recent. Theorem A, in explicit form, is due to Baxter and Chacon (1976), though seems implicit in the earlier works of Dinges (1974) and Rost (1971) (see also Pitman, 1977): extensions can be found in Revuz (1978) and finite-state applications in Lovász and Winkler (1995). Theorem B is part of Theorem 16.0.2 of Meyn and Tweedie (1993), who describe its history, tracing the various parts of the cycle of equivalences to dates between 1941 and 1980. Theorem C is from Aldous (1982).

Write $(X(t); t = 0, 1, 2, \ldots)$ for a Markov chain with transition kernel $P(x, A)$ on a measurable state space $\mathscr{X}$. Suppose an invariant probability measure $\pi$ exists. Write $\|\nu\|$ for the total variation norm on signed measures on $\mathscr{X}$, so that for probability measures $\mu_1, \mu_2$ we have

$$\|\mu_1 - \mu_2\| = 2 \sup_A |\mu_1(A) - \mu_2(A)|.$$

Here we have used the "functional analysis" normalization, even though in the "mixing time" literature it is common to divide the right sides by 2. Write $H_A \geqslant 0$ for the first hitting time on $A$. Write $E_\mu(\cdot)$ and $\mathscr{L}_\mu(\cdot)$ for expectation and distribution w.r.t. the initial distribution $\mu$. For probability measures $\mu, \theta$ define

$$h(\mu, \theta) = \inf\{E_\mu T : T \text{ is a randomized stopping time, } \mathscr{L}_\mu(X(T)) = \theta\}.$$

Abusing notation slightly, write $h(x, \theta)$ instead of $h(\delta_x, \theta)$ for the case of an initial distribution $\delta_x$ concentrated at $x$.

Consider the hypothesis

$$G(x, \cdot) \equiv \lim_{t \to \infty} \sum_{s=0}^{t} (P^s(x, \cdot) - \pi(\cdot)) \quad \text{exists } \forall x, \tag{1}$$

where the limit is w.r.t. total variation. If (1) holds, then $G(x, \cdot)$ is a signed measure with $G(x, \mathscr{X}) = 0$. It may not be true that $G(x, \cdot) \ll \pi$, but it is easy to see that $G(x, \cdot)$ decomposes as the sum of a positive measure singular w.r.t. $\pi$, and a signed measure with some density $g(x, \cdot)$ w.r.t. $\pi$. We call $g = g(x, y)$ the *recurrent potential density*.

**Theorem A.** *Suppose* (1). *If the negative part of* $(\mu - \sigma)G$ *has a density* $-\phi(y)$ *w.r.t.* $\pi$ *then*

$$h(\mu, \sigma) = \text{ess sup } \phi.$$

*Otherwise,* $h(\mu, \sigma) = \infty$.

To state the second theorem, define

$$d(t) = \sup_x \| P^t(x, \cdot) - \pi \|. \tag{2}$$

If $d(t) \to 0$ the chain is called *uniformly ergodic*. It is well known that $d(t)$ is submultiplicative, so if $d(t) \to 0$ then the convergence is geometrically fast. Next, a *petite structure* is a collection $\{C, \mu, m, \delta\}$ where $C$ is a subset of $\mathscr{X}$, $\mu$ is a probability distribution on $\mathscr{X}$, $m \geqslant 1$, $\delta > 0$ and

$$K_m(x, \cdot) \equiv (m+1)^{-1} \sum_{t=0}^{m} P^t(x, \cdot) \geqslant \delta\mu(\cdot) \quad \forall x \in C. \tag{3}$$

Call $C$ a *petite set* if it is part of some petite structure $\{C, \mu, m, \delta\}$.

**Theorem B.** *The following are equivalent.*

(i) *The chain is uniformly ergodic.*

(ii) *There exist* $m < \infty$, $\delta > 0$ *and a probability measure* $\mu$ *such that*

$$P^m(x, \cdot) \geqslant \delta\mu(\cdot) \quad \forall x.$$

(iii) *The chain is aperiodic and there exists a petite set* $C$ *such that*

$$\sup_x E_x H_C < \infty.$$

(iv) *The chain is aperiodic and there exist a petite set $C$, a bounded function $V(x) \geqslant 0$ and constants $\beta > 0$, $b < \infty$ such that*

$$E_x V(X_1) - V(x) \leqslant -\beta, x \notin C$$
$$\leqslant b, x \in C.$$

To state the third theorem, define

$$\mathscr{T}_{\text{mix}} = \sup_x h(x, \pi), \qquad \mathscr{T}_{\text{set}} = \sup_{x, A : \pi(A) > 0} \pi(A) E_x H_A,$$

$$\mathscr{T}_{\text{continuize}} = \inf\{t : \tilde{P}^t(x, \cdot) - \pi \| \leqslant e^{-1} \forall x\}$$

where

$$\tilde{P}^t(x, \cdot) = \sum_{i=0}^{\infty} \frac{e^{-t} t^i}{i!} P^i(x, \cdot) \tag{4}$$

is the transition kernel for the associated continuous-time chain. Finally, in the finite-state case, where $h(x, y)$ and $h(\pi, y)$ are finite for all $x, y$, we may define

$$\mathscr{T}_G = \sup_x \sum_y |h(x, y) - h(\pi, y)| \pi(y).$$

So $\mathscr{T}_{\text{G}}$ measures variability of mean hitting times as a function of starting state. (See (9) for a more generally applicable redefinition of $\mathscr{T}_{\text{G}}$).

**Theorem C.** *For each pair* $(\mathscr{T}_i, \mathscr{T}_j)$ *from* $\{\mathscr{T}_{\text{mix}}, \mathscr{T}_{\text{set}}, \mathscr{T}_{\text{G}}, \mathscr{T}_{\text{continuize}}\}$ *there is a constant* $K_{i,j} < \infty$ *such that for every irreducible* reversible *chain on every* finite *state space* $\mathscr{X}$,

$$\mathscr{T}_i \leqslant K_{i,j} \mathscr{T}_j.$$

Though the hypotheses and conclusions of Theorems A–C are somewhat different, it seems intuitively clear that they refer in part to the same idea: the relation between means of stopping times and convergence to stationarity. Means of stopping times are explicit in Theorem A, in Theorem B(iii) and the definitions of $\mathscr{T}_{\text{mix}}$ and $\mathscr{T}_{\text{set}}$ in Theorem C. And as regards convergence, the parameters $\mathscr{T}_{\text{mix}}$ and $\mathscr{T}_{\text{continuize}}$ in Theorem C provide quantifications of the uniform ergodicity assertion in Theorem B (i), while in Theorem A one expects the size of the measures $G(x, \cdot)$ to be related to the speed of convergence of the sum in (1).

### *1.3. Statement of new results*

The goal of our paper, in brief, is to establish quantitative bounds like those in Theorem C in the continuous-space setting of Theorem B.

The setting we shall adopt is best described as "uniform ergodicity, but without assuming aperiodicity". More precisely, define $\bar{d}(t)$ as "$d(t)$ for the uniformly-sampled chain", i.e.,

$$\bar{d}(t) = \sup_x \| P_x(X(U_t) \in \cdot) - \pi(\cdot) \| = \sup_x \| K_t(x, \cdot) - \pi(\cdot) \|$$

where $K_t$ was defined at (3) and where $U_t$ denotes a random variable distributed uniformly on $\{0, 1, 2, \dots, t\}$, independent of the chain. If $\bar{d}(t) \to 0$, call the chain *uniformly ergodic under sampling* (UES). Minor modifications to the proof of Theorem B would establish the parallel result

**Theorem B*.** *The following are equivalent.*

(i) *The chain is UES.*

(ii) *There exist $m \geq 1$, $\delta > 0$ and a probability measure $\mu$ such that*

$$K_m(x, \cdot) \geq \delta \mu(\cdot) \forall x.$$

(iii, iv) *The corresponding statements in Theorem B, without the "aperiodic" assertion.*

Our goal is to give a "quantitative" version of Theorem $B^*$. That is, we replace assertions of the form

there exist objects $\{a, b, \dots\}$ satisfying requirements $\{R, S, T \dots\}$

by parameters $\mathscr{T}$ defined via

$\mathscr{T}$ is the minimum, over all choices of objects $\{a, b, \dots\}$ satisfying requirements $\{R, S, T \dots\}$, of a certain numerical function of $\{a, b, \dots\}$.

Applying this procedure to the four parts of Theorem $B^*$ leads to the following four definitions.

$$\mathscr{T}_{\text{uniform}}(c) = \min\{t : \bar{d}(t) \leq c\}, \quad 0 < c < 1. \tag{5}$$

$\mathscr{T}_{\text{minorize}}$ is the infimum of $\delta^{-1} m$ over all $\{m, \delta, \mu\}$ in Theorem $B^*$ (ii). (6)

$\mathscr{T}_{\text{petite}}$ is the infimum of $\delta^{-1}(m + \sup_x E_x H_C)$ over all petite structures $\{C, \mu, m, \delta\}$. (7)

$\mathscr{T}_{\text{drift}}$ is the infimum of $\delta^{-1}(m + \max(b, \beta^{-1} \sup_x V(x)))$ over all petite structures $\{C, \mu, m, \delta\}$ and all $\{V, \beta, b\}$ satisfying the inequality in Theorem B (iv). (8)

But it is almost obvious (see Section 6.2) that in fact $\mathscr{T}_{\text{drift}} = \mathscr{T}_{\text{petite}}$, so we need not consider $\mathscr{T}_{\text{drift}}$ separately. We shall also consider parameters equal or similar to those in Theorem C. Redefine $\mathscr{T}_{\text{G}}$ as

$$\mathscr{T}_{\text{G}} = \sup_x \| G(x, \cdot) \|. \tag{9}$$

This is consistent with the previous definition in the finite-state case, where it is classical (see the discussion of the fundamental matrix in [17]) that

$h(i,j) = (G(j,j) - G(i,j))/\pi(j)$, $h(\pi,j) = G(j,j)/\pi(j)$ and so $\pi(j)(h(i,j) - h(\pi,j)) = -G(i,j)$. Next, we give two weaker variants of $\mathscr{T}_{\text{mix}}$. The first requires only approximately attaining the target distribution $\pi$:

$$\mathscr{T}_{\text{stop}}(c) = \sup_x \inf\{E_x T : \|\mathscr{L}_x X(T) - \pi\| \leqslant c\}, \quad 0 < c < 1. \tag{10}$$

The second replaces $\pi$ by some target distribution $\sigma$ of our choice.

$$\mathscr{T}_{\text{forget}} = \inf_\sigma \sup_\mu h(\mu, \sigma) = \inf_\sigma \sup_x h(x, \sigma). \tag{11}$$

**Theorem 1.** *A chain is UES if and only if one of the parameters* $\{\mathscr{T}_G, \mathscr{T}_{\text{set}}, \mathscr{T}_{\text{forget}}, \mathscr{T}_{\text{minorize}}, \mathscr{T}_{\text{petite}}, \mathscr{T}_{\text{uniform}}(c), 0 < c < 1, \mathscr{T}_{\text{stop}}(c), 0 < c < 1\}$ *is finite, in which case all of these parameters are finite. For each pair* $(\mathscr{T}_i, \mathscr{T}_j)$ *of parameters in that set, there is a constant* $K_{i,j} < \infty$ *such that* $\mathscr{T}_i \leqslant K_{i,j}\mathscr{T}_j$ *for every UES chain.*

More concisely, call these parameters *equivalent*. In addition to quantifying Theorem $B^*$, Theorem 1 shows that part of Theorem C remains true in the non-reversible setting. One might hope that $\mathscr{T}_{\text{mix}}$ remained equivalent to these parameters in the non-reversible setting, but this hope is dashed by

**Example 2.** *The winning streak chain.* Take $\mathscr{X} = \{0, 1, 2, \ldots\}$ and $P(x, x+1) = p$, $P(x, 0) = 1 - p$ for fixed $0 < p < 1$. So $\pi(x) = (1-p)p^x$. By considering $\sigma = \delta_0$ we have $\mathscr{T}_{\text{forget}} = 1/(1-p)$. But an elementary calculation gives $E_0 H_x = (1/\pi(x)) - (1/(1-p))$ and so $\mathscr{T}_{\text{mix}} \geqslant h(0, \pi) = \sum_x \pi(x) E_0 H_x = \infty$.

It turns out that $\mathscr{T}_{\text{mix}}$ is related instead to yet another parameter. Define, for $0 < c < 1$,

$$\mathscr{T}_{\text{separate}}(c) = \min\{t : P_x(X(U_t) \in \cdot) \geqslant (1 - c)\pi(\cdot) \forall x\}. \tag{12}$$

For a UES chain the parameters $\mathscr{T}_{\text{mix}}$ and $\mathscr{T}_{\text{separate}}(c)$ may be infinite, but they are equivalent.

**Theorem 3.** $\mathscr{T}_{\text{mix}} \leqslant [1/2(1-c)]\,\mathscr{T}_{\text{separate}}(c)$; $0 < c < 1$. *Conversely, if* $1/c$ *is an integer then* $\mathscr{T}_{\text{separate}}(c) \leqslant (4/c^2)\,\mathscr{T}_{\text{mix}}$.

To connect this with recurrent potential, note that Theorem A gives

$$h(x, \pi) = \text{ess sup}_y(-g(x, y)). \tag{13}$$

Thus $\mathscr{T}_{\text{mix}}$ can be defined directly in terms of the recurrent potential density $g$ as

$$\mathscr{T}_{\text{mix}} = \sup_x \text{ess sup}_y(-g(x, y)).$$

We should emphasize that Theorems 1 and 3 are not really difficult or deep. Our proofs use the same mix of ingredients as the proof of Theorem C, with occasional modifications which use Theorem A in place of considering mean hitting times on single states. Textbooks sometimes leave the impression that general-state chains require different techniques than finite-state chains, but from the quantitative viewpoint this is not so: our proofs were originally written for finite-state chains but then extended to the UES setting with only minor rephrasing.

Some further results dealing with time-reversals (and requiring some measure-theoretic technicalities) will be given in Section 5.

### *1.4. Interpretation of results*

In the setting of Section 1.1, there is a specific Markov chain which we use to obtain samples for some ultimate algorithmic use. For an analysis of the number of steps needed, the ultimate use affects the notion of "mixing time" needed. The significance of our results is that one can to some extent "decouple" mathematical analysis of the chain from the ultimate algorithmic use of the samples, because many different mixing times are equivalent up to constants. In other words, for a sequence of Markov chains with size-parameter $n$, Theorem 1 says there is a well-defined "order of magnitude of mixing times" $t(n)$ such that each parameter in Theorem 1 is $\Theta(t(n))$. In contrast, Theorems B and $B^*$ are typically uninformative in this context.

Of course, to actually bound mixing times for specific chains is a more interesting and important problem. Our results do not directly help, beyond providing flexibility in what one needs to prove to obtain an order-of-magnitude bound. (For instance, in obtaining upper bounds the freedom of choice of $\sigma$ in $\mathscr{T}_{\text{forget}}$ may be helpful; in obtaining lower bounds the freedom of choice of $x$ and $A$ in $\mathscr{T}_{\text{set}}$ may be helpful.)

We remark that most of the algorithmic problems of Section 1.1 are so hard that one cannot get the correct order of magnitude bound for mixing times. On the other hand, in the more highly-structured setting of card-shuffling and random walks on groups, one can often do rather precise calculations of mixing times: see for instance the analysis (Beyer and Diaconis, 1992) of the riffle shuffle. Our work is perhaps most relevant to examples whose complexity is such that one can get only the correct order of magnitude. Here are two recent examples. Chung and Graham (1996) analyze the chain on states $\{0, 1\}^n$ in which two coordinates $i, j$ are chosen at random, and the parity of $x_i$ is changed if $x_j = 1$. They show the mixing time is $\Theta(n \log n)$. Diaconis and Saloff-Coste (1996) study simple symmetric random walk on a convex subset of the two-dimensional lattice, and show that the mixing time is $\Theta(\text{diameter}^2)$.

In the setting of random walks on groups, the main focus of study has been

$$\mathscr{T}(c) = \min\{t : d(t) \leqslant c\} \tag{14}$$

and the *cut-off phenomenon* [8], in place of the time averaged analog $\mathscr{T}_{\text{uniform}}(c)$. While this is natural in examples, there seems no elegant "equivalence theory" analogous to Theorem 1 for $\mathscr{T}(c)$, and indeed Corollary 9 later indicates how $\mathscr{T}(c)$ may behave undesirably. The underlying difficulty is to quantify aperiodicity. Since periodicity is irrelevant for algorithmic sampling purposes, the Theorem 1 mixing times are more natural in that context.

## 2. Some technical tools

*The minorization construction.* Let $(Y(t))$ have kernel $Q$ satisfying the minorization condition $Q(x,\cdot) \geqslant \delta\mu(\cdot)\,\forall x$, for some $\delta > 0$ and some probability measure $\mu$. Obviously we can construct a randomized stopping time $T$ with geometric($\delta$) distribution such that $Y(T)$ has distribution $\mu$ and is independent of both the starting state and the value of $T$: in particular

$$\mathscr{L}Y(T) = \mu, \qquad ET = 1/\delta. \tag{15}$$

*Elaborations of Theorem A.* We need to use some ingredients of the proof of Theorem A, so we shall outline parts of the proof. See Baxter and Chacon (1976), Lovasz and Winkler (1995) and Aldous and Fill (1977) for more details.

Fix $\mu, \sigma$ and consider a stopping time $T$ with $E_\mu T < \infty$ and $\mathscr{L}_\mu X(T) = \sigma$. Write $\psi(\cdot) = E_\mu \sum_{t=0}^{T-1} 1_{(X(t)\in\cdot)}$. Then $\psi$ is one solution of the identity $\psi - \psi P = \mu - \sigma$. Assuming (1), a particular solution of this identity is $\psi_0 = (\mu - \sigma)G$ and then the general solution is $\psi = (\mu - \sigma)G + c\pi$ for some constant $c$. Since $\psi(\mathscr{X}) = E_\mu T$ we have $c = E_\mu T$. To summarize:

$$\psi(\cdot) \equiv E_\mu \sum_{t=0}^{T-1} 1_{(X(t)\in\cdot)} = (\mu - \sigma)G + (E_\mu T)\pi. \tag{16}$$

Since $\psi \geqslant 0$, (16) implies

$$E_\mu T \geqslant \text{ess sup} \frac{\mathrm{d}(\sigma - \mu)G}{\mathrm{d}\pi}. \tag{17}$$

The proof of Theorem A is completed via a *filling scheme* construction, which defines inductively a certain decreasing sequence $A_t$ of random subsets such that

$$T = \min\{t : X(t) \in A_t\} \tag{18}$$

achieves equality in (17).

## 3. Proof of Theorem 1

The proof is structured as three cycles of inequalities, in which $0 < c < 1$ is arbitrary. The first cycle is

$$\mathscr{T}_{\text{stop}}(c) \leqslant 4\mathscr{T}_{\text{set}}/c^2, \tag{19}$$

$$\mathscr{T}_{\text{set}} \leqslant \mathscr{T}_{\text{G}} \leqslant \frac{2}{1-c}\mathscr{T}_{\text{stop}}(c). \tag{20}$$

These imply that $\mathscr{T}_{\text{stop}}(c) \leqslant [8/c^2(1-c')]\,\mathscr{T}_{\text{stop}}(c')$. So the parameters $\{\mathscr{T}_{\text{set}}, \mathscr{T}_{\text{G}}, \mathscr{T}_{\text{stop}}(c), 0 < c < 1\}$ are all *equivalent*, i.e. ratios are bounded by constants.

The second cycle is

$$\mathscr{T}_{\mathrm{stop}}(c) \leqslant \tfrac{1}{2}\,\mathscr{T}_{\mathrm{uniform}}(c), \tag{21}$$

$$\mathscr{T}_{\mathrm{uniform}}(c') \leqslant \frac{2\mathscr{T}_{\mathrm{stop}}(c)}{c'-c}, \quad c' > c. \tag{22}$$

These show that the parameters $\{\mathscr{T}_{\mathrm{uniform}}(c),\ 0 < c < 1\}$ are equivalent to the parameters above. The third cycle is

$$\mathscr{T}_{\mathrm{forget}} \leqslant \mathscr{T}_{\mathrm{petite}} \leqslant \mathscr{T}_{\mathrm{minorise}} \leqslant 43\mathscr{T}_{\mathrm{uniform}}(\tfrac{1}{4}), \tag{23}$$

$$\mathscr{T}_{\mathrm{uniform}}(c) \leqslant \frac{4}{c}\,\mathscr{T}_{\mathrm{forget}}. \tag{24}$$

These imply equivalence of the remaining parameters $\{\mathscr{T}_{\mathrm{forget}}, \mathscr{T}_{\mathrm{petite}}, \mathscr{T}_{\mathrm{minorize}}\}$.

*The first cycle.* Fix some initial distribution. The fact that a minimal-mean stopping time $T$ with $\mathscr{L}X(T) = \pi$ can be constructed via a filling scheme (18) implies

$$P(T \geqslant t) \leqslant P(H_{A_t} \geqslant t), \quad P(T \geqslant t) \leqslant \pi(A_t).$$

Using the definition of $\mathscr{T}_{\mathrm{set}}$ and the inequalities above,

$$P(T \geqslant t) \leqslant P(H_{A_t} \geqslant t) \leqslant t^{-1}EH_{A_t} \leqslant \frac{\mathscr{T}_{\mathrm{set}}}{t\pi(A_t)} \leqslant \frac{\mathscr{T}_{\mathrm{set}}}{tP(T \geqslant t)}$$

and so $P(T \geqslant t) \leqslant \sqrt{\mathscr{T}_{\mathrm{set}}/t}$, in particular

$$P(T > \lfloor 4\mathscr{T}_{\mathrm{set}}/c^2 \rfloor) \leqslant c/2.$$

But $\| \mathscr{L}X(\min(T,t)) - \pi \| \leqslant 2P(T > t)$ and $\min(T,t)$ is a stopping time with mean at most $t$, so by definition of $\mathscr{T}_{\mathrm{stop}}(c)$ we have $\mathscr{T}_{\mathrm{stop}}(c) \leqslant 4\mathscr{T}_{\mathrm{set}}/c^2$, which is (19).

Fix $\mu$ and $A$ and write $\sigma = \mathscr{L}_{\mu}X(H_A)$. Consider (16) with $T = H_A$: since $\psi = 0$ on $A$ we have $\mathrm{d}(\sigma - \mu)G/\mathrm{d}\pi = E_x H_A$ on $A$. Then

$$\pi(A)E_{\mu}H_A = (\sigma - \mu)G(A) \leqslant \tfrac{1}{2}\|(\sigma - \mu)G\| \leqslant \mathscr{T}_{\mathrm{G}}.$$

So by definition of $\mathscr{T}_{\mathrm{set}}$ we have $\mathscr{T}_{\mathrm{set}} \leqslant \mathscr{T}_{\mathrm{G}}$, which is the first inequality of (20).

Fix $x$. By Theorem A, for any distribution $\sigma$ and any set $A$,

$$h(x,\sigma) \geqslant \frac{(\sigma - \delta_x)G(A)}{\pi(A)}.$$

Rearranging, and using the fact $\pi G(\cdot) = 0$ (a simple consequence of (1)),

$$\begin{aligned} -G(x,A) &\leqslant h(x,\sigma) - \sigma G(A) = h(x,\sigma) + (\pi - \sigma)G(A) \\ &\leqslant h(x,\sigma) + \tfrac{1}{2}\|\pi - \sigma\|\mathscr{T}_{\mathrm{G}}. \end{aligned}$$

By definition of $\mathscr{T}_{\mathrm{stop}}(c)$, minimizing over $\{\sigma : \|\sigma - \pi\| \leqslant c\}$ gives

$$-G(x,A) \leqslant \mathscr{T}_{\mathrm{stop}}(c) + \frac{c}{2}\mathscr{T}_{\mathrm{G}}.$$

Applying this to $A = \{y : g(x, y) < 0\}$ and maximizing over $x$ gives, by definition of $\mathscr{T}_{\mathrm{G}}$,

$$\tfrac{1}{2}\mathscr{T}_{\mathrm{G}} \leqslant \mathscr{T}_{\mathrm{stop}}(c) + \frac{c}{2}\mathscr{T}_{\mathrm{G}}.$$

In other words, $\mathscr{T}_{\mathrm{G}} \leqslant [2/(1-c)]\,\mathscr{T}_{\mathrm{stop}}(c)$, which is the second inequality of (20).

*The second cycle.* Inequality (21) follows from the definitions of $\mathscr{T}_{\mathrm{stop}}(c)$ and $\mathscr{T}_{\mathrm{uniform}}(c)$, because $EU_t = t/2$. Now fix $x$, consider $T$ as in the definition of $\mathscr{T}_{\mathrm{stop}}(c)$, so that $E_x T \leqslant \mathscr{T}_{\mathrm{stop}}(c)$ and $\|\mathscr{L}_x X(T) - \pi\| \leqslant c$. The latter implies $\|\mathscr{L}_x X(T + U_t) - \pi\| \leqslant c$. And

$$\begin{aligned}
&\|\mathscr{L}_x X(U_t) - \mathscr{L}_x X(T + U_t)\| \\
&\quad\leqslant \|\mathscr{L}_x U_t - \mathscr{L}_x(T + U_t)\| \\
&\quad\leqslant 2t^{-1}E_x T \text{ because } \|\mathscr{L}U_t - \mathscr{L}(a + U_t)\| \leqslant 2t^{-1}a \\
&\quad\leqslant 2t^{-1}\mathscr{T}_{\mathrm{stop}}(c).
\end{aligned}$$

By the triangle inequality

$$\|\mathscr{L}_x X(U_t) - \pi\| \leqslant c + 2t^{-1}\mathscr{T}_{\mathrm{stop}}(c). \tag{25}$$

In other words, $\mathscr{T}_{\mathrm{uniform}}(c') \leqslant t$ whenever $c + 2t^{-1}\mathscr{T}_{\mathrm{stop}}(c) \leqslant c'$. Rearranging gives $\mathscr{T}_{\mathrm{uniform}}(c') \leqslant 2\mathscr{T}_{\mathrm{stop}}(c)/(c' - c)$, $c' > c$, which is (22).

*The third cycle.* We start by proving (24), the proof being similar to the proof of (22). Let $\sigma$ attain the *inf* in the definition of $\mathscr{T}_{\mathrm{forget}}$. So given an initial state $x$, we can choose $S_x$ and $S_\pi$ such that

$$\mathscr{L}_x X(S_x) = \sigma,\ E_x S_x \leqslant \mathscr{T}_{\mathrm{forget}},\ \mathscr{L}_\pi X(S_\pi) = \sigma,\ E_\pi S_\pi \leqslant \mathscr{T}_{\mathrm{forget}}.$$

Then $\mathscr{L}_x X(S_x + U_t) = \mathscr{L}_\pi X(S_\pi + U_t) = \mathscr{L}_\sigma X(U_t)$, and so

$$\begin{aligned}
\|\mathscr{L}_x X(U_t) - \mathscr{L}_\pi X(U_t)\| &\leqslant \|\mathscr{L}_x X(U_t) - \mathscr{L}_x X(S_x + U_t)\| \\
&\quad + \|\mathscr{L}_\pi X(U_t) - \mathscr{L}_\pi X(S_\pi + U_t)\| \\
&\leqslant \|\mathscr{L}_x U_t - \mathscr{L}_x(S_x + U_t)\| + \|\mathscr{L}_\pi(U_t) - \mathscr{L}_\pi(S_\pi + U_t)\| \\
&\leqslant 2t^{-1}(E_x S_x + E_\pi S_\pi) \\
&\leqslant 4t^{-1}\mathscr{T}_{\mathrm{forget}}.
\end{aligned}$$

Since $\mathscr{L}_\pi X(U_t) = \pi$, we have established (24).

Given a petite structure $\{C, \mu, m, \delta\}$, we have

$$Q(x, \cdot) \equiv P_x(X(H_C + U_m) \in \cdot) \geqslant \delta\mu(\cdot) \quad \forall x \in \mathscr{X}.$$

Use (15) to construct a stopping time attaining distribution $\mu$ with mean $\leqslant \delta^{-1}\sup_x E_x(H_C + U_m) = \delta^{-1}(m/2 + \sup_x E_x H_C)$. This implies $\mathscr{T}_{\mathrm{forget}} \leqslant \mathscr{T}_{\mathrm{petite}}$, the first inequality in (23). The second inequality, $\mathscr{T}_{\mathrm{petite}} \leqslant \mathscr{T}_{\mathrm{minorize}}$, is immediate by taking $C = \mathscr{X}$ in the definition of $\mathscr{T}_{\mathrm{petite}}$. The third inequality requires a preliminary lemma.

**Lemma 4.** *Let $Q$ be a transition kernel for which $\pi$ is invariant, and suppose*

$$\sup_x \|Q(x,\cdot) - \pi(\cdot)\| \leqslant \tfrac{1}{4}. \tag{26}$$

*Then there exists a set $A$ with $\pi(A) \geqslant 1/2$ such that, for all probability distributions $\mu$,*

$$\frac{\mathrm{d}\mu Q}{\mathrm{d}\pi} \geqslant \frac{1}{4}(1 - \|\mu - \pi\|) \quad \textit{on } A.$$

**Proof.** Let $B_x$ be the set where $\mathrm{d}Q(x,\cdot)/\mathrm{d}\pi \leqslant \frac{1}{2}$, or more precisely where the measure $Q(x,\cdot) - \frac{1}{2}\pi(\cdot)$ is negative. Write $B = \{(x,y): y \in B_x\} \subset \mathscr{X} \times \mathscr{X}$. For each $x$ we have $Q(x, B_x) \leqslant \frac{1}{2}\pi(B_x)$, and so

$$\tfrac{1}{2}\pi(B_x) \leqslant \pi(B_x) - Q(x, B_x) \leqslant \tfrac{1}{2}\|Q(x,\cdot) - \pi(\cdot)\|.$$

Using (26), $\pi(B_x) \leqslant \frac{1}{4}$ and hence $\pi \times \pi(B) \leqslant \frac{1}{4}$. Write $B^y = \{x: (x,y) \in B\}$ and define $A = \{y: \pi(B^y) \leqslant \frac{1}{2}\}$. Then $\pi(A^c) \leqslant \pi \times \pi(B)/1/2 \leqslant 1/2$. By definition we have $\mathrm{d}Q(x,\cdot)/\mathrm{d}\pi(y) \geqslant \frac{1}{2}$ for $x \in \mathscr{X} \backslash B^y$, and so

$$\frac{\mathrm{d}\mu Q}{\mathrm{d}\pi}(y) \geqslant \tfrac{1}{2}\,\mu(\mathscr{X}\backslash B^y) \geqslant \frac{1}{2}(\pi(\mathscr{X}\backslash B^y) - \frac{1}{2}\,\|\mu - \pi\|) \geqslant \frac{1}{2}\left(\frac{1}{2} - \frac{1}{2}\|\mu - \pi\|\right) \text{ on } A,$$

the final inequality by definition of $A$. □

Now set $t = \mathscr{T}_{\text{uniform}}(\frac{1}{4})$ and let $U$ and $U'$ be independent, uniform on $\{0, 1, \ldots, t\}$. Let $Q$ be the kernel associated with $X(U')$, so that (26) holds by definition. Let $A$ be the set guaranteed by the lemma. For fixed $x$, write $\mu = \mathscr{L}_x X(U)$, and then

$$\frac{\mathrm{d}\mathscr{L}_x X(U + U')}{\mathrm{d}\pi} = \frac{\mathrm{d}\mu Q}{\mathrm{d}\pi} \geqslant \frac{1}{4}\left(1 - \frac{1}{4}\right) = \frac{3}{16} \text{ on } A$$

by the lemma. In other words, if $R(x,\cdot)$ is the kernel associated with $X(U + U')$ and if we set $\mu = \pi(\cdot \mid A)$ then $R(x,\cdot) \geqslant \frac{3}{32}\mu(\cdot)\forall x$, because $\pi(A) \geqslant 1/2$ by the lemma. It is elementary that $P(U_{2t} = i) \geqslant \frac{1}{2}P(U + U' = i)\forall i$, and so $\mathscr{L}_x X(U_{2t}) \geqslant \frac{3}{64}\mu\,\forall x$. So by definition of $\mathscr{T}_{\text{minorize}}$ we have $\mathscr{T}_{\text{minorize}} \leqslant \frac{64}{3}2t \leqslant 43\mathscr{T}_{\text{uniform}}(1/4)$.

## 4. Proof of Theorem 3

Fix $t$, write

$$s(t) = \inf\{c: P_x(X(U_t) \in \cdot) \geqslant (1 - c)\pi(\cdot)\forall x\}$$

and consider the chain $Y$ with transition kernel $K_t(x,\cdot) = P_x(X(U_t) \in \cdot)$. Construction (15) gives a stopping time $S$ for $Y$ satisfying $\mathscr{L}Y(S) = \pi$ and $ES = 1/(1 - s(t))$. This in turn specifies a stopping time $T = U_t^{(1)} + \cdots + U_t^{(S)}$ for $X$ satisfying $\mathscr{L}X(T) = \pi$ and $ET = (ES)(EU_t) = t/[2(1 - s(t))]$. Putting $t = \mathscr{T}_{\text{separate}}(c)$ gives

$$\mathscr{T}_{\text{mix}} \leqslant \frac{\mathscr{T}_{\text{separate}}(c)}{2(1 - c)}.$$

For a reverse inequality, the central idea is contained in the following lemma, analogous to Lemma 4. Write $N(t,\cdot) = \sum_{n=0}^{t-1} 1_{(X(n)\in\cdot)}$.

**Lemma 5.** $\frac{\mathrm{d}}{\mathrm{d}\pi} E_\sigma N(t,\cdot) \geqslant (1 - \frac{\|\sigma-\pi\|}{2})t - \mathscr{T}_{\mathrm{mix}}$.

**Proof.** Write $\delta = \|\sigma - \pi\|/2$. Decompose the initial distribution $\sigma$ as $\sigma = \pi + \delta(\alpha - \beta)$, where $\alpha = (\sigma - \pi)^+/\delta$ and $\beta = (\pi - \sigma)^+/\delta$. Note that

$$\text{ess sup}\,\frac{\mathrm{d}\beta}{\mathrm{d}\pi} \leqslant 1/\delta \tag{27}$$

and that

$$E_\sigma N(t,\cdot) = E_\pi N(t,\cdot) + \delta(E_\alpha N(t,\cdot) - E_\beta N(t,\cdot)) \geqslant t\pi(\cdot) - \delta E_\beta N(t,\cdot). \tag{28}$$

Next observe that for any stopping time $S$ with $\mathscr{L}_\pi X(S) = \pi$ (note this refers to the *stationary* chain) we have by (16)

$$E_\pi N(S,\cdot) = (E_\pi S)\pi(\cdot). \tag{29}$$

By definition of $\mathscr{T}_{\mathrm{mix}}$ there exists a stopping time $T$ with $E_x T \leqslant \mathscr{T}_{\mathrm{mix}}$ and $\mathscr{L}_x X(T) = \pi$ for all $x$. So

$$\begin{aligned}\int \beta(\mathrm{d}x)E_x N(T,\cdot) &\leqslant \frac{1}{\delta}\int \pi(\mathrm{d}x)E_x N(T,\cdot) \quad \text{by (27)}\\ &= \frac{1}{\delta}\left(\int \pi(\mathrm{d}x)E_x T\right)\pi(\cdot) \quad \text{by (29)}\\ &\leqslant \frac{1}{\delta}\mathscr{T}_{\mathrm{mix}}\,\pi(\cdot)\end{aligned}$$

and then

$$\begin{aligned}E_\beta N(t,\cdot) &\leqslant \int \beta(\mathrm{d}x)E_x N(T+t,\cdot)\\ &= \int \beta(\mathrm{d}x)E_x N(T,\cdot) + E_\pi N(t,\cdot)\\ &\leqslant \frac{1}{\delta}\mathscr{T}_{\mathrm{mix}}\,\pi(\cdot) + t\pi(\cdot).\end{aligned}$$

The lemma now follows, using (28). □

To deduce the reverse inequality in Theorem 3, fix an initial state $x$ and a time $s$. Then

$$E_x N(s+t,\cdot) \geqslant E_x N(U_s + t,\cdot) = E_\sigma N(t,\cdot)$$

for $\sigma = \mathscr{L}_x X(U_s)$. By (25) with $c = 0$ we have $\|\sigma - \pi\| \leqslant 2\mathscr{T}_{\text{mix}}/s$ and so using Lemma 5

$$\frac{\mathrm{d}}{\mathrm{d}\pi} E_x N(s+t, \cdot) \geqslant t - \mathscr{T}_{\text{mix}} t/s - \mathscr{T}_{\text{mix}}.$$

In other words,

$$\frac{\mathrm{d}}{\mathrm{d}\pi} K_{s+t-1}(x, \cdot) \geqslant \frac{t - \mathscr{T}_{\text{mix}} t/s - \mathscr{T}_{\text{mix}}}{s+t-1} \geqslant \frac{t - (1 - \frac{t}{s})\mathscr{T}_{\text{mix}}}{s+t}.$$

Provided $1/c$ is an integer we may set $s = \frac{2}{c}\mathscr{T}_{\text{mix}}$ and $t = (\frac{4}{c^2} - \frac{2}{c})\mathscr{T}_{\text{mix}}$. The right side becomes $1 - c$, so we have proved

$$\mathscr{T}_{\text{separate}}(c) \leqslant s + t = \frac{4\mathscr{T}_{\text{mix}}}{c^2}. \tag{30}$$

## 5. Time-reversals

A kernel $P$ with stationary distribution $\pi$ and the time-reversed kernel $P^*$ are related by the following identity for measures on $\mathscr{X} \times \mathscr{X}$.

$$\pi(\mathrm{d}x)P^*(x, \mathrm{d}y) = \pi(\mathrm{d}y)P(y, \mathrm{d}x). \tag{31}$$

It is perhaps surprising that for a UES chain the time-reversed (or *dual*) chain need not be UES. For instance, the time-reversal of Example 2 is the chain with $p^*(x, x-1) = 1$, $x \geqslant 1$ and $p(0, x) = (1-p)p^x$, $x \geqslant 0$, which is plainly not UES. This lack of symmetry suggests study of the class of processes whose time-reversals are UES. For each parameter $\mathscr{T}$ we may define a parameter $\mathscr{T}^*$ as "$\mathscr{T}$ for $P^*$". For instance, $\mathscr{T}^*_{\text{G}} = \sup_x \|G^*(x, \cdot)\|$ where $G^*$ is defined in terms of $P^*$ as in (1). Theorem 1 implies equivalence of the "starred" parameters therein, but what does this mean for the $P$-chain? The parameters involving stopping times for the $P^*$-chain have no very clear interpretation as parameters for the $P$-chain, but it turns out (Lemma 6) that the remaining parameters $\{\mathscr{T}^*_{\text{G}}, \mathscr{T}^*_{\text{minorize}}, \mathscr{T}^*_{\text{uniform}}(c)\}$ can be expressed directly in terms of the $P$-chain. But first we must deal with a technical issue. If we use (31) as a *definition* of the time-reversed kernel $P^*(x, \cdot)$ of a given kernel $P$, then $P^*(x, \cdot)$ is defined uniquely only up to $\pi$-null sets of $x$-values. This matters because the parameters $\mathscr{T}$ were defined as *sups*, rather than *ess sups*, over $x$. Issues like this are frequently resolved by imposing topological assumptions, but for our purposes we may just adopt the following simple though inelegant assumption.

**Assumption.** $P$ and $P^*$ are related by (31). Furthermore, in the definition of each parameter $\mathscr{T}$ for both $P$ and $P^*$, using $\sup_x$ and $\operatorname{ess\,sup}_x$ give the same value.

It is straightforward to check that, given $P$ and $P^*$ related by (31), we can delete a single $\pi$-null set from $\mathscr{X}$ so that the second requirement holds on the remaining space. In this sense, the assumption involves "no loss of generality". To see the need for some such assumption, consider a finite state space chain containing both transient states and a single recurrent class $\mathscr{R}$. Then there exists a (unique) stationary distribution supported on $\mathscr{R}$, the chain is UES, so Theorems 1 and 3 are meaningful (and true), for parameters defined as maxima over the whole state space. But there is no natural way to define $P^*$ outside $\mathscr{R}$, and hence no natural way of making results like Theorem D true. Our technical assumption has the effect of pruning the state space down to $\mathscr{R}$.

Relation (31) easily implies that the density components $k_t(x, y)\pi(dy)$ and $g(x, y)\pi(dy)$ of $K_t(x, \cdot)$ and $G(x, \cdot)$ are related to their duals by symmetry:

$$k_t^*(x, y) = k_t(y, x), g^*(x, y) = g(y, x), \text{ a.e. } (\pi \times \pi). \tag{32}$$

Now using this symmetry and our technical assumption, it is easy to relate certain "starred" parameters to their unstarred versions. The parameters in Theorem 3 can be written as

$$\mathscr{T}_{\text{mix}} = \text{ess sup}_{(x,y)}(-g(x, y))$$

$$\mathscr{T}_{\text{separate}}(c) = \min\{t : \text{ess inf}_{(x,y)} t^{-1} k_t(x, y) \geq (1 - c)\}.$$

So by symmetry they are unchanged by time-reversal. The next lemma (proved in Section 5.1) expresses the "starred" versions of certain parameters in Theorem 1 in terms of the $P$-chain.

**Lemma 6.** (a) *$\mathscr{T}^*_{\text{minorize}}$ is the infimum of $\delta^{-1}m$ over all triples $\{m, \delta, V\}$ such that*

$$V \geq 0; \qquad \int V \, d\pi = 1; \qquad K_m(x, \cdot) \geq \delta V(x)\pi(\cdot) \quad \forall x.$$

(b)

$$\mathscr{T}^*_{\text{uniform}}(c) = \min\{t : d^*(t) \leq c\}, \quad 0 < c < 1,$$

*where*

$$d^*(t) = 2 \text{ ess sup}_y \int (k_t(x, y) - 1)^- \pi(dx),$$

*where $k_t(x, y)\pi(dy)$ is the density component of $K_t(x, \cdot)$*

(c) *Write $G = G^+ - G^-$ for the Hahn–Jordan decomposition of $G$ as a difference of positive kernels, and write $|G| = G^+ + G^-$. Then*

$$\mathscr{T}^*_G = \text{ess sup} \frac{d\pi|G|}{d\pi}.$$

Whereas the original parameters in Theorem 1 all explicitly involved *sups* over initial states, the starred parameters in Lemma 6 have a different flavor: roughly, they involve approximate minorization at a terminal time. A more natural parameter with that flavor is

$$\mathscr{T}_{\text{reset}} = \int h(x, \pi)\pi(\mathrm{d}x).$$

In Section 5.1 we prove the simple bounds

**Lemma 7.** $\frac{1}{2}\mathscr{T}^*_{\mathrm{G}} \leqslant \mathscr{T}_{\text{reset}} \leqslant \frac{1}{2}\mathscr{T}^*_{\text{minorize}}$.

Combining with Theorem 1 (applied to time-reversed chains) gives

**Corollary 8.** *The parameters* $\{\mathscr{T}_{\text{reset}}, \mathscr{T}^*_{\mathrm{G}}, \mathscr{T}^*_{\text{minorize}}, \mathscr{T}^*_{\text{uniform}}(c),\ 0 < c < 1\}$ *are equivalent. The time-reversed chain is UES iff one (all) of these parameters are finite.*

We observed above that the parameters in Theorem 3 are unchanged by time-reversal. So (e.g. because $\mathscr{T}_{\text{stop}}(c) \leqslant \mathscr{T}_{\text{mix}}$) if $\mathscr{T}_{\text{mix}}$ is finite then both the chain and its time-reversal are UES. The converse also holds: if a chain and its time-reversal are both UES then $\mathscr{T}_{\text{mix}}$ is finite. (So our results imply the equivalence of $\{\mathscr{T}_{\text{mix}}, \mathscr{T}_{\text{set}}, \mathscr{T}_{\mathrm{G}}\}$ for reversible chains, which was part of Theorem C). In fact,

$$\mathscr{T}_{\text{mix}} \leqslant 2(\mathscr{T}_{\text{set}} + \mathscr{T}_{\text{reset}}).$$

Because $A = \{x : h(x, \pi) \leqslant 2\mathscr{T}_{\text{reset}}\}$ has $\pi(A) \geqslant \frac{1}{2}$, and so $\sup_x E_x H_A \leqslant 2\mathscr{T}_{\text{set}}$. Thus for any initial distribution, we run the chain until hitting $A$, then until an optimal stopping time attaining $\pi$, and this stopping time has mean $\leqslant \sup_x E_x H_A + 2\mathscr{T}_{\text{reset}}$.

Theorems 1 and 3 reflect the spirit of Theorems B and C in dealing with inequalities. Theorem A is in the spirit of standard results on maximal coupling and minimal strong stationary times (see Section 6.3) giving “optimal constructions” or “min–max characterizations”. Lovász and Winkler (1997) proved another remarkable identity in the same spirit.

**Theorem D.** *For a finite-state irreducible chain,* $\mathscr{T}^*_{\text{reset}} = \mathscr{T}_{\text{forget}}$.

Assuming this extends to our general-space setting, one could use Theorem D in place of Lemma 7 in establishing Corollary 8.

### *5.1. Proofs*

**Proof of Lemma 6.** In the definition of $\mathscr{T}_{\text{minorize}}$, the probability measure $\mu$ must satisfy $\mu \ll \pi$. Setting $V = \mathrm{d}\mu/\mathrm{d}\pi$ and using the symmetry relation (32) leads to the stated expression for $\mathscr{T}^*_{\text{minorize}}$. Next, we may rewrite the definition of $\bar{d}(t)$ as

$$\bar{d}(t) = 2 \text{ ess sup}_x \int (k_t(x, y) - 1)^- \pi(\mathrm{d}y)$$

and then the expression for $\mathscr{T}^*_{\mathrm{uniform}}(c)$ follows by symmetry. Similarly, writing $\mathscr{T}_{\mathrm{G}} = 2$ ess $\sup_x \int g^-(x, y)\pi(\mathrm{d}y)$ gives by symmetry

$$\mathscr{T}^*_{\mathrm{G}} = 2 \text{ ess } \sup_y \int g^-(x, y)\pi(\mathrm{d}x) \tag{33}$$

and the stated expression for $\mathscr{T}^*_{\mathrm{G}}$ follows because $\pi G = 0$.

**Proof of Lemma 7.**

$$\begin{aligned}\tfrac{1}{2}\mathscr{T}^*_{\mathrm{G}} &= \text{ess } \sup_y \int g^-(x, y)\pi(\mathrm{d}x) \quad \text{by (33)}\\ &\leqslant \int \text{ess } \sup_y(-g(x, y))\pi(\mathrm{d}x)\\ &= \int h(x, \pi)\pi(\mathrm{d}x) \quad \text{by (13)}\\ &= \mathscr{T}_{\mathrm{reset}}.\end{aligned}$$

For the second inequality, consider $\{m, \delta, V\}$ as in the definition of $\mathscr{T}^*_{\mathrm{minorize}}$, so that the chain $Y$ with kernel $Q = K_m$ satisfies

$$Q(x, \cdot) \geqslant \delta V(x)\pi(\cdot) \quad \forall x.$$

Let $Y(0)$ have distribution $\pi$ and define a randomized stopping time $T \geqslant 1$ for $Y$ by

$$P(T = t + 1 \mid T > t,\ Y(t) = x,\ Y(t + 1) = y) = \delta V(x)\frac{1}{\dfrac{\mathrm{d}Q(x, \cdot)}{\mathrm{d}\pi}(y)}.$$

One can verify inductively that

$$\begin{aligned}&P(Y(t)\in\cdot \mid T > t) = \pi(\cdot),\\ &P(Y(t + 1)\in\cdot \mid T = t + 1,\ Y(t) = x) = \pi(\cdot),\\ &P(T = t + 1 \mid T > t) = \delta.\end{aligned}$$

So $ET = 1/\delta$ and $Y(T)$ has distribution $\pi$, independent of $Y(0)$. Since $Y$ is the uniformly-sampled $X$-chain, this construction gives a stopping time $S$ for $X$ such that $X(S)$ has distribution $\pi$, independent of $X(0)$. So $\mathscr{T}_{\mathrm{reset}} \leqslant ES = m/2\delta$ and thus $\mathscr{T}_{\mathrm{reset}} \leqslant \frac{1}{2}\mathscr{T}^*_{\mathrm{minorize}}$.

## 6. Discussion

### 6.1. Transforming a chain

Let $X(t)$ be a chain with forget time $\mathscr{T}_{\mathrm{forget}}$. Corollary 9 gives three constructions of chains of the form $\tilde{X}(t) = X(N(t))$ for different definitions of $N(t)$, and relates their forget times $\tilde{\mathscr{T}}_{\mathrm{forget}}$ to $\mathscr{T}_{\mathrm{forget}}$.

**Corollary 9.** (a) [jump chain]. *Define*

$$N(t+1)=\min\{s>N(t)\colon X(s)\neq X(N(t))\}.$$

*Then* $\tilde{\mathscr{T}}_{\text{forget}}\leqslant\mathscr{T}_{\text{forget}}$.

(b) [slowed-down chain]. *Take* $N(t+1)=N(t)$ *or* $N(t)+1$, *with*

$$P(N(t+1)=N(t)+1\,|\,X(N(t))=x)=a(x).$$

*Then* $\tilde{\mathscr{T}}_{\text{forget}}\leqslant a^{-1}\mathscr{T}_{\text{forget}}$, *where* $a=\inf_x a(x)$.

(c) [chain watched only on $A$]. *Take*

$$N(t+1)=\min\{s>N(t)\colon X(s)\in A\}.$$

*Then* $\tilde{\mathscr{T}}_{\text{forget}}\leqslant\mathscr{T}_{\text{forget}}$.

The proofs are immediate, by considering the minimizing $\sigma$ in the definition of $\mathscr{T}_{\text{forget}}$, and using $\tilde{\sigma}=\sigma$ (cases (a, b)) or $\tilde{\sigma}=P_\sigma(X(H_A)\in\cdot)$ (case (c)) as target distributions in the definition of $\tilde{\mathscr{T}}_{\text{forget}}$. Intuitively, any reasonable definition of "mixing time" should satisfy similar inequalities. But note that with the traditional definition using total variation at fixed times (14), inequality (a) fails (the jump chain may be periodic) and the other inequalities do not seem simple to establish.

### *6.2. Technical remarks*

(a) We have used "uniform smoothing" rather than "geometric smoothing" throughout, though there is no essential difference. Our statement of Theorem B skips some further, similar-style, assertions in [23, Theorem 16.0.2]. Our statement of the drift condition (iv) is superficially different from theirs, but is clearly equivalent. Our statement was chosen to highlight the quantitative equality $\mathscr{T}_{\text{petite}}=\mathscr{T}_{\text{drift}}$, which is a consequence of the following observations. Given a petite set $C$, the function $V(x)=E_xH_C$ satisfies the inequality in Theorem B (iv) with $\beta=1$ and $b=\sup_xE_xH_C$. Conversely, if $V$ satisfies the inequality in Theorem B (iv) then $E_xH_C\leqslant V(x)/\beta$ by the obvious supermartingale argument.

(b) In the deterministic chain $X(t)=t$ modulo $n$, the parameters in Theorems 1 and 3 are $\theta(n)$. This example shows that in Theorem 1 we cannot replace $\mathscr{T}_{\text{uniform}}(c)$ by $\mathscr{T}_{\text{continuize}}$ or by any smoothing essentially weaker than uniform.

(c) We glossed over two related technical points. For a periodic chain the limit (1) defining $G$ may not exist; and in the setting of Theorem 1 we do not know a priori that $G$ exists. What is important about $G$ is that it satisfies

$$(I-P)G=I-\Pi,$$

where $I$ is the identity kernel and $\Pi(x,\cdot)=\pi(\cdot)$. In the period-$d$ setting where $P^d$ is uniformly ergodic on each cyclic component, we can modify (1) by taking averages over $\{t,t+1,\ldots,t+d-1\}$, and then the $t\to\infty$ average of these limits exists. Using the general-space decomposition of a periodic chain into cyclic components (Meyn and Tweedie, 1993, Section 5.4.1; Revuz, 1994, Section 6.3), one can show that a UES chain is of this form, and so $G$ exists.

(d) In our proofs we assumed that infima in the definitions of parameters $\mathscr{T}$ are attained. Pedantically, we should have considered attaining $\mathscr{T} + \varepsilon$, and then let $\varepsilon \to 0$.

(e) The parameters $\mathscr{T}_G$ and $\mathscr{T}_{\text{mix}}$ give probabilistic interpretations of certain $L^1$ and $L^\infty$ norms of $G$, so it is natural to ask whether the analogous $L^2$ norm $\sup_x \int_{y:g(x,y)<0} g^2(x,y)\pi(\mathrm{d}y)$ has a probabilistic interpretation as a mixing time.

(f) Our proofs used the same general set of techniques as in Aldous (1982), though at the level of detailed proofs the overlap with Aldous (1982), Baxter and Chacon (1976) and Lovasz and Winker (1995) is quite small.

(g) A more complicated example of a UES chain whose time-reversal is not UES can be found in Revuz (1994, Exercise 8.3.11).

### 6.3. *Conceptual remarks*

(a) The asymptotic geometric rate of convergence of a chain is controlled by its *spectral gap*. That parameter is rather different from our mixing time parameters. See Aldous and Fill (1997) for an extensive discussion in the reversible setting.

(b) Fix an initial distribution $\mu$ and define $s(t) = \min\{c : \mathscr{L}_\mu X(t) \geqslant (1-c)\pi\}$. A *strong stationary time* is a stopping time $T$ such that $X(T)$ has distribution $\pi$ and is independent of $T$ (e.g. the minorization construction (15) gave a strong stationary time). Such a $T$ must satisfy the inequalities $P(T > t) \geqslant s(t) \forall t$; and it is easy to construct an optimal strong stationary time $T$ satisfying $P(T > t) = s(t) \forall t$. See Diaconis and Fill (1990) for developments of such theory. This construction, and the conceptually similar notion (Goldstein, 1979; Lindvall, 1992) of *maximal coupling*, are in the same spirit as Theorem A.

(c) One of the themes of Meyn and Tweedie (1993) is a sequence of theorems, in the general format of Theorem B, which treat successively stronger notions of convergence (ergodicity, geometric ergodicity, $V$-uniform ergodicity, uniform ergodicity), and relate each to drift and "return time to petite sets" conditions. Their presentation thus emphasizes "general" results such as the existence of minorizing measures as a consequence of irreducibility, and the "split chain" construction. But results at that level of generality are inherently non-quantitative. We are deliberately approaching these results from the opposite direction in order to get quantitative results. Whether analogs of Theorem 1 hold for these more general notions of convergence is an interesting question.

(d) Informally, our parameters are defined to "scale as time". This is easier to formalize in continuous time: if $X(t)$ has parameter value $\mathscr{T}$ then $X^*(t) = X(ct)$ should have parameter value $c^{-1}\mathscr{T}$. For two such parameters in continuous time, the existence of some universal inequality $\mathscr{T}_2 \leqslant \psi(\mathscr{T}_1)$ clearly implies a linear inequality $\mathscr{T}_2 \leqslant K_{1,2}\mathscr{T}_1$. Thus the existence of linear inequalities in Theorems 1 and 3 is not surprising.

(e) A quite different setting where mixing times might be studied is "randomly-perturbed chaos". Consider a topological space $\mathscr{X}$ and a continuous function $f : \mathscr{X} \to \mathscr{X}$ for which $\pi$ is invariant, and suppose we define kernels $P^{(n)}(x,\cdot)$ such

$P^{(n)}(x,\cdot) \to \delta_{f(x)}$ weakly and whose stationary distributions $\pi^{(n)} \to \pi$ in total variation. Theorem 1 implies there is a well-defined order of magnitude $\Theta(t(n))$ for the parameters therein, providing an indirect formalization of the time until the underlying deterministic process $(f^i(x);\ i \geqslant 0)$ becomes chaotic. Getting explicit results in this setting seems challenging, even in simple-looking examples like the following. Fix $0 \leqslant a < 1$ and consider the random walk on the reals modulo 1 whose step-distribution is Normal$(a, \sigma^2)$ modulo 1. If $a = 0$, clearly the mixing time parameters in Theorems 1 and 3 are $\Theta(\sigma^{-2})$ as $\sigma \to 0$. For general $a$, the behavior of our parameters as $\sigma \to 0$ is related to equidistribution of $\{ia \bmod 1\colon i = 1, 2, \ldots\}$. Heuristically it appears that for typical $a$ the mixing times are $\Theta(\sigma^{-2/3})$, but this appears non-elementary to prove.

## Acknowledgements

We thank J. Baxter, D. Revuz and R. Syski for helpful comments on the literature.

## References

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

Aldous D. J., Fill, J. A., 1997. Reversible Markov Chains and Random Walks on Graphs. Book in preparation.

Baxter J. R., Chacon, R. V., 1976. Stopping times for recurrent Markov processes. Illinois J. Math. 20, 467–475.

Bayer, D., Diaconis, P., 1992. Trailing the dovetail shuffle to its lair. Ann. Appl. Probab 2, 294–313.

Chung, F. R. K., Graham, R. L., 1996. Stratified random walks on the $n$-cube. Technical report, Univ. Pennsylvania.

Dellacherie, C., Meyer, P.-A., 1983. Probabilités et Potentiel: Théorie Discrète du Potentiel. Hermann, Paris.

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

Diaconis, P., 1996. The cut-off phenomenon in finite Markov chains. Proc. Nat. Acad. Sci. USA 93, 1659–1664.

Diaconis, P., Fill, J. A., 1990. Strong stationary times via a new form of duality. Ann. Probab. 18, 1483–1522.

Diaconis P., Saloff-Coste, L., 1996. Nash inequalities for finite Markov chains. J. Theoret. Probab. 9, 459–510.

Diaconis, P., Saloff-Coste, L., 1996. Walks on generating sets of groups. Probab. Theory Related Fields 105, 393–421.

Dinges, H., 1974. Stopping sequences. In: Séminaire de Probabilités VIII, Lecture Notes in Math., vol. 381, Springer, Berlin, pp. 27–36.

Frieze, A., Kannan, R., Polson, N., 1994. Sampling from log-concave distributions. Ann. Appl. Probab. 4, 812–837.

Goldstein, S., 1979. Maximal coupling. Z. Wahrsch. Verw. Gebiete 46, 193–204.

Jerrum M., Sinclair, A., 1989. Approximating the permanent. SIAM J. Comput. 18, 1149–1178.

Kannan, R., Lovász, L., Simonovits, M., 1997. Random walks and an $O(n^5)$ volume algorithm. Random Struct. Alg. 8, to appear.

Kemeny, J.G., Snell, J.L., 1960. Finite Markov Chains. Van Nostrand, Princeton, NJ.

Lindvall, T., 1992. Lectures on the Coupling Method. Wiley, New York.

Lovász, L., Simonovits, M., 1993. Random walks in a convex body and an improved volume algorithm. Random Struct. Alg. 4, 359–412.

Lovász, L., Winkler, P., 1995. Efficient stopping rules for Markov chains. In: Proc. 27th ACM Symp. Theory of Computing, pp. 76–82.

Lovász, L., Winkler, P., 1997. Fast mixing in a Markov chain. In preparation.

Lovász, L., Winkler, P., 1997. Reversal of Markov chains and the forget time. In preparation.

Meyn, S.P., Tweedie, R.L., 1993. Markov Chains and Stochastic Stability. Springer, Berlin.

Motwani, R., Raghavan, P., 1995. Randomized Algorithms. Cambridge University Press. Cambridge.

Nummelin, E., 1984. General Irreducible Markov Chains and Non-Negative Operators. Cambridge University Press, Cambridge.

Orey, S., 1971. Limit Theorems for Markov Chain Transition Probabilities. Van Nostrand, Princeton, NJ

Pitman, J.W., 1977. Occupation measures for Markov chains. Adv. Appl. Probab. 9, 69–86.

Revuz, D., Remarks on the filling scheme for recurrent Markov chains. Duke Math. J. 45, 681–689.

Revuz, D., 1984. Markov Chains, 2nd ed. North-Holland, Amsterdam.

Rost, H., 1971. The stopping distributions of a Markov process. Invent. Math. 14, 1–16.

Sinclair, A.J., 1993. Algorithms for Random Generation and Counting. Birkhauser, Basel.

Syski. R., 1992. Passage Times for Markov Chains. IOS Press, Amsterdam.