TOTAL VARIATION CUTOFF IN BIRTH-AND-DEATH CHAINS

JIAN DING, EYAL LUBETZKY AND YUVAL PERES

ABSTRACT. The *cutoff phenomenon* describes a case where a Markov chain exhibits a sharp transition in its convergence to stationarity. In 1996, Diaconis surveyed this phenomenon, and asked how one could recognize its occurrence in families of finite ergodic Markov chains. In 2004, the third author noted that a necessary condition for cutoff in a family of reversible chains is that the product of the mixing-time and spectral-gap tends to infinity, and conjectured that in many settings, this condition should also be sufficient. Diaconis and Saloff-Coste (2006) verified this conjecture for continuous-time birth-and-death chains, started at an endpoint, with convergence measured in *separation*. It is natural to ask whether the conjecture holds for these chains in the more widely used *total-variation* distance.

In this work, we confirm the above conjecture for all continuous-time or lazy discrete-time birth-and-death chains, with convergence measured via total-variation distance. Namely, if the product of the mixing-time and spectral-gap tends to infinity, the chains exhibit cutoff at the maximal hitting time of the stationary distribution median, with a window of at most the geometric mean between the relaxation-time and mixing-time.

In addition, we show that for any lazy (or continuous-time) birth-and-death chain with stationary distribution π, the separation $1 - p^t(x,y)/\pi(y)$ is maximized when x, y are the endpoints. Together with the above results, this implies that total-variation cutoff is equivalent to separation cutoff in any family of such chains.

1. INTRODUCTION

The *cutoff phenomenon* arises when a finite Markov chain converges abruptly to equilibrium. Roughly, this is the case where, over a negligible period of time known as the *cutoff window*, the distance of the chain from the stationary measure drops from near its maximum to near 0.

Let (X_t) denote an aperiodic irreducible Markov chain on a finite state space Ω with transition kernel $P(x,y)$, and let π denote its stationary distribution. For any two distributions μ, ν on Ω, their *total-variation distance* is

Research of J. Ding and Y. Peres was supported in part by NSF grant DMS-0605166.

defined to be

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A\subset\Omega} |\mu(A) - \nu(A)| = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)| \,.$$

Consider the worst-case total-variation distance to stationarity at time t,

$$d(t) := \max_{x\in\Omega} \|\mathbf{P}_x(X_t \in \cdot) - \pi\|_{\mathrm{TV}} \,,$$

where $\mathbf{P}_x$ denotes the probability given $X_0 = x$. The total-variation *mixing-time* of (X_t), denoted by $t_{\mathrm{MIX}}(\varepsilon)$ for $0 < \varepsilon < 1$, is defined to be

$$t_{\mathrm{MIX}}(\varepsilon) := \min\{t : d(t) \le \varepsilon\} \,.$$

Next, consider a family of such chains, $(X_t^{(n)})$, each with its corresponding worst-distance from stationarity $d_n(t)$, its mixing-times $t_{\mathrm{MIX}}^{(n)}$, etc. We say that this family of chains exhibits *cutoff* iff the following sharp transition in its convergence to stationarity occurs:

$$\lim_{n\to\infty} \frac{t_{\mathrm{MIX}}^{(n)}(\varepsilon)}{t_{\mathrm{MIX}}^{(n)}(1-\varepsilon)} = 1 \quad \text{for any } 0 < \varepsilon < 1 \,. \tag{1.1}$$

Our main result is an essentially tight bound on the difference between $t_{\mathrm{MIX}}(\varepsilon)$ and $t_{\mathrm{MIX}}(1-\varepsilon)$ for general *birth-and-death* chains; a birth-and-death chain has the state space $\{0,\ldots,n\}$ for some integer n, and always moves from one state to a state adjacent to it (or stays in place).

We first state a quantitative bound for a single chain, then deduce a cutoff criterion. Let $\mathtt{gap}$ be the spectral-gap of the chain (that is, $\mathtt{gap} := 1 - \lambda$ where λ is the largest absolute-value of all nontrivial eigenvalues of the transition kernel P), and let $t_{\mathrm{REL}} := \mathtt{gap}^{-1}$ denote the relaxation-time of the chain. A chain is called *lazy* if $P(x,x) \ge \frac{1}{2}$ for all $x \in \Omega$.

Theorem 1. *For any $0 < \varepsilon < \frac{1}{2}$ there exists an explicit $c_\varepsilon > 0$ such that every lazy irreducible birth-and-death chain (X_t) satisfies*

$$t_{\mathrm{MIX}}(\varepsilon) - t_{\mathrm{MIX}}(1-\varepsilon) \le c_\varepsilon \sqrt{t_{\mathrm{REL}} \cdot t_{\mathrm{MIX}}(\tfrac{1}{4})} \,. \tag{1.2}$$

As we later show, the above theorem extends to continuous-time chains, as well as to δ-lazy chains, which satisfy $P(x,x) \ge \delta$ for all $x \in \Omega$.

The notion of a cutoff-window relates Theorem 1 to the cutoff phenomenon. A sequence w_n is called a *cutoff window* for a family of chains $(X_t^{(n)})$ if the following holds: $w_n = o(t_{\mathrm{MIX}}^{(n)}(\frac{1}{4}))$, and for any $\varepsilon > 0$ there exists some $c_\varepsilon > 0$ such that, for all n,

$$t_{\mathrm{MIX}}^{(n)}(\varepsilon) - t_{\mathrm{MIX}}^{(n)}(1-\varepsilon) \le c_\varepsilon w_n \,. \tag{1.3}$$

Equivalently, if t_n and w_n are two sequences such that $w_n = o(t_n)$, one may define that a sequence of chains exhibits cutoff at t_n with window w_n iff

$$\begin{cases} \lim_{\lambda\to\infty} \liminf_{n\to\infty} d_n(t_n - \lambda w_n) = 1 \ , \\ \lim_{\lambda\to\infty} \limsup_{n\to\infty} d_n(t_n + \lambda w_n) = 0 \ . \end{cases}$$

To go from the first definition to the second, take $t_n = t_{\text{MIX}}^{(n)}(\frac{1}{4})$.

Once we compare the forms of (1.2) and (1.3), it becomes clear that Theorem 1 implies a bound on the cutoff window for any general family of birth-and-death chains, provided that $t_{\text{REL}}^{(n)} = o(t_{\text{MIX}}^{(n)}(\frac{1}{4}))$.

Theorem 1 will be the key to establishing the criterion for total-variation cutoff in a general family of birth-and-death chains.

1.1. **Background.** The cutoff phenomenon was first identified for the case of random transpositions on the symmetric group in [11], and for the case of random walks on the hypercube in [1]. It was given its name by Aldous and Diaconis in their famous paper [3] from 1985, where they showed that the top-in-at-random card shuffling process (repeatedly removing the top card and reinserting it to the deck at a random position) has such a behavior. Saloff-Coste [25] surveys the cutoff phenomenon for random walks on finite groups.

Though many families of chains are believed to exhibit cutoff, proving the occurrence of this phenomenon is often an extremely challenging task, hence there are relatively few examples for which cutoff has been rigorously shown. In 1996, Diaconis [7] surveyed the cutoff phenomenon, and asked if one could determine whether or not it occurs in a given family of aperiodic and irreducible finite Markov chains.

In 2004, the third author [24] observed that a necessary condition for cutoff in a family of reversible chains is that the product $t_{\text{MIX}}^{(n)}(\frac{1}{4}) \cdot \mathsf{gap}(n)$ tends to infinity with n, or equivalently, $t_{\text{REL}}^{(n)} = o(t_{\text{MIX}}^{(n)}(\frac{1}{4}))$; see Lemma 2.1. The third author also conjectured that, in many natural classes of chains,

$$\textit{Cutoff occurs if and only if } t_{\text{REL}}^{(n)} = o(t_{\text{MIX}}^{(n)}(\tfrac{1}{4})) \ . \tag{1.4}$$

In the general case, this condition does not always imply cutoff : Aldous [2] and Pak (private communication via P. Diaconis) have constructed relevant examples (see also [5],[6] and [21]). This left open the question of characterizing the classes of chains for which (1.4) holds.

One important class is the family of birth-and-death chains; see [10] for many natural examples of such chains. They also occur as the magnetization chain of the mean-field Ising Model (see [12],[20]).

In 2006, Diaconis and Saloff-Coste [10] verified a variant of the conjecture (1.4) for birth-and-death chains, when the convergence to stationarity is measured in *separation*, that is, according to the decay of $\text{sep}(\mathbf{P}_0(X_t \in \cdot), \pi)$,

where $\mathrm{sep}(\mu,\nu) = \sup_{x\in\Omega}(1-\frac{\mu(x)}{\nu(x)})$. Note that, although $\mathrm{sep}(\mu,\nu)$ assumes values in $[0,1]$, it is in fact not a metric (it is not even symmetric). See, e.g., [4, Chapter 4] for the connections between mixing-times in total-variation and in separation.

More precisely, it was shown in [10] that any family of continuous-time birth-and-death chains, started at 0, exhibits cutoff in separation if and only if $t_{\mathrm{REL}}^{(n)} = o(t_{\mathrm{sep}}^{(n)}(\frac{1}{4};0))$, where $t_{\mathrm{sep}}(\varepsilon;s) = \min\{t : \mathrm{sep}(\mathbf{P}_s(X_t\in\cdot),\pi) < \varepsilon\}$. The proof used a spectral representation of passage times [16, 17] and duality of strong stationary times. Whether (1.4) holds with respect to the important and widely used total-variation distance, remained unsettled.

1.2. **Total-variation cutoff.** In this work, we verify the conjecture (1.4) for arbitrary birth-and-death chains, with the convergence to stationarity measured in total-variation distance. Our first result, which is a direct corollary of Theorem 1, establishes this for lazy discrete-time irreducible birth-and-death chains. We then derive versions of this result for continuous-time irreducible birth-and-death chains, as well as for δ-lazy discrete chains (where $P(x,x)\geq\delta$ for all $x\in\Omega$). In what follows, we omit the dependence on n wherever it is clear from the context.

Corollary 2. *Let $(X_t^{(n)})$ be a sequence of lazy irreducible birth-and-death chains. Then it exhibits cutoff in total-variation distance iff $t_{\mathrm{MIX}}^{(n)}\cdot\mathsf{gap}(n)$ tends to infinity with n. Furthermore, the cutoff window size is at most the geometric mean between the mixing-time and relaxation time.*

As we will later explain, the given bound $\sqrt{t_{\mathrm{MIX}}\cdot t_{\mathrm{REL}}}$ for the cutoff window is essentially tight, in the following sense. Suppose that the functions $t_M(n)$ and $t_R(n)\geq 2$ denote the mixing-time and relaxation-time of $(X_t^{(n)})$, a family of irreducible lazy birth-and-death chains. Then there exists a family $(Y_t^{(n)})$ of such chains with the parameters $t_{\mathrm{MIX}}^{(n)} = (1+o(1))t_M(n)$ and $t_{\mathrm{REL}}^{(n)} = (1+o(1))t_R(n)$ that has a cutoff window of $(t_{\mathrm{MIX}}^{(n)}\cdot t_{\mathrm{REL}}^{(n)})^{1/2}$. In other words, no better bound on the cutoff window can be given without exploiting additional information on the chains.

Indeed, there are examples where additional attributes of the chain imply a cutoff window of order smaller than $\sqrt{t_{\mathrm{MIX}}\cdot t_{\mathrm{REL}}}$. For instance, the cutoff window has size t_{REL} for the Ehrenfest urn (see, e.g., [9]) and for the magnetization chain in the mean field Ising Model at high temperature (see [12]).

Theorem 3.1, given in Section 3, extends Corollary 2 to the case of δ-lazy discrete-time chains. We note that this is in fact the setting that corresponds to the magnetization chain in the mean-field Ising Model (see, e.g., [20]).

Following is the continuous-time version of Corollary 2.

Theorem 3. *Let $(X_t^{(n)})$ be a sequence of continuous-time birth-and-death chains. Then $(X_t^{(n)})$ exhibits cutoff in total-variation iff $t_{\text{REL}}^{(n)} = o(t_{\text{MIX}}^{(n)})$, and the cutoff window size is at most $\sqrt{t_{\text{MIX}}^{(n)}(\frac{1}{4}) \cdot t_{\text{REL}}^{(n)}}$.*

By combining our results with those of [10] (while bearing in mind the relation between the mixing-times in total-variation and in separation), one can relate worst-case total-variation cutoff in any continuous-time family of irreducible birth-and-death chains, to cutoff in separation started from 0. This suggests that total-variation cutoff should be equivalent to separation cutoff in such chains under the original definition of the worst starting point (as opposed to fixing the starting point at one of the endpoints). Indeed, it turns out that for any lazy or continuous-time birth-and-death chain, the separation is always attained by the two endpoints, as formulated by the next proposition.

Proposition 4. *Let (X_t) be a lazy (or continuous-time) birth-and-death chain with stationary distribution π. Then for every integer (resp. real) $t > 0$, the separation $1 - \mathbf{P}_x(X_t = y)/\pi(y)$ is maximized when x, y are the endpoints.*

That is, for such chains, the maximal separation from π at time t is simply $1 - P^t(0,n)/\pi(n)$ (for the lazy chain with transition kernel P) or $1 - H_t(0,n)/\pi(n)$ (for the continuous-time chain with heat kernel H_t). As we later show, this implies the following corollary:

Corollary 5. *For any continuous-time family of irreducible birth-and-death chains, cutoff in worst-case total-variation distance is equivalent to cutoff in worst-case separation.*

Note that, clearly, the above equivalence is in the sense that one cutoff implies the other, yet the cutoff locations need not be equal (and sometimes indeed are not equal, e.g., the Bernoulli-Laplace models, surveyed in [10, Section 7]).

The rest of this paper is organized as follows. The proofs of Theorem 1 and Corollary 2 appear in Section 2. Section 3 contains the proofs of the variants of Theorem 1 for the continuous-case (Theorem 3) and the δ-lazy case. In Section 4, we discuss separation in general birth-and-death chains, and provide the proofs of Proposition 4 and Corollary 5. The final section, Section 5, is devoted to concluding remarks and open problems.

2. Cutoff in lazy birth-and-death chains

In this section we prove the main result, which shows that the condition $\mathtt{gap} \cdot t_{\text{MIX}} \to \infty$ is necessary and sufficient for total-variation cutoff in lazy birth-and-death chains.

2.1. **Proof of Corollary 2.** The fact that any family of lazy irreducible birth-and-death chains satisfying $t_{\text{MIX}} \cdot \mathsf{gap} \to \infty$ exhibits cutoff, follows by definition from Theorem 1, as does the bound $\sqrt{t_{\text{REL}} \cdot t_{\text{MIX}}}$ on the cutoff window size.

It remains to show that this condition is necessary for cutoff; this is known to hold for any family of reversible Markov chains, using a straightforward and well known lower bound on t_{MIX} in terms of t_{REL} (cf., e.g., [21]). We include its proof for the sake of completeness.

Lemma 2.1. *Let (X_t) denote a reversible Markov chain, and suppose that $t_{\text{REL}} \geq 1 + \theta t_{\text{MIX}}(\frac{1}{4})$ for some fixed $\theta > 0$. Then for any $0 < \varepsilon < 1$*

$$t_{\text{MIX}}(\varepsilon) \geq t_{\text{MIX}}(\tfrac{1}{4}) \cdot \theta \log(1/2\varepsilon) \,. \tag{2.1}$$

In particular, $t_{\text{MIX}}(\varepsilon)/t_{\text{MIX}}(\frac{1}{4}) \geq K$ for all $K > 0$ and $\varepsilon < \frac{1}{2}\exp(-K/\theta)$.

Proof. Let P denote the transition kernel of X, and recall that the fact that X is reversible implies that P is a symmetric operator with respect to $\langle \cdot, \cdot \rangle_\pi$ and $\mathbf{1}$ is an eigenfunction corresponding to the trivial eigenvalue 1.

Let λ denote the largest absolute-value of all nontrivial eigenvalues of P, and let f be the corresponding eigenfunction, $Pf = \pm\lambda f$, normalized to have $\|f\|_\infty = 1$. Finally, let r be the state attaining $|f(r)| = 1$. Since f is orthogonal to $\mathbf{1}$, it follows that for any t,

$$\begin{aligned} \lambda^t = \left|(P^t f)(r) - \langle f, \mathbf{1}\rangle_\pi\right| &\leq \max_{x\in\Omega} \left| \sum_{y\in\Omega} P^t(x,y) f(y) - \pi(y) f(y)\right| \\ &\leq \|f\|_\infty \max_{x\in\Omega} \|P^t(x,\cdot) - \pi\|_1 = 2 \max_{x\in\Omega} \|P^t(x,\cdot) - \pi\|_{\text{TV}} \,. \end{aligned}$$

Therefore, for any $0 < \varepsilon < 1$ we have

$$t_{\text{MIX}}(\varepsilon) \geq \log_{1/\lambda}(1/2\varepsilon) \geq \frac{\log(1/2\varepsilon)}{\lambda^{-1} - 1} = (t_{\text{REL}} - 1)\log(1/2\varepsilon) \,, \tag{2.2}$$

and (2.1) immediately follows. ∎

This completes the proof of Corollary 2. ∎

2.2. **Proof of Theorem 1.** The essence of proving the theorem lies in the treatment of the regime where t_{REL} is much smaller than $t_{\text{MIX}}(\frac{1}{4})$.

Theorem 2.2. *Let (X_t) denote a lazy irreducible birth-and-death chain, and suppose that $t_{\text{REL}} < \varepsilon^5 \cdot t_{\text{MIX}}(\frac{1}{4})$ for some $0 < \varepsilon < \frac{1}{16}$. Then*

$$t_{\text{MIX}}(4\varepsilon) - t_{\text{MIX}}(1 - 2\varepsilon) \leq (6/\varepsilon)\sqrt{t_{\text{REL}} \cdot t_{\text{MIX}}(\tfrac{1}{4})} \,.$$

Proof of Theorem 1. To prove Theorem 1 from Theorem 2.2, let $\varepsilon > 0$, and suppose first that $t_{\text{REL}} < \varepsilon^5 \cdot t_{\text{MIX}}(\frac{1}{4})$. If $\varepsilon < \frac{1}{64}$, then the above theorem clearly implies that (1.2) holds for $c_\varepsilon = 24/\varepsilon$. Since that the left-hand-side

of (1.2) is monotone decreasing in ε, this result extends to any value of $\varepsilon < \frac{1}{2}$ by choosing

$$c_1 = c_1(\varepsilon) = 24 \max\{1/\varepsilon, 64\} .$$

It remains to treat the case where $t_{\text{REL}} \geq \varepsilon^5 \cdot t_{\text{MIX}}(\frac{1}{4})$. In this case, the sub-multiplicativity of the mixing-time (see, e.g., [4, Chapter 2]) gives

$$t_{\text{MIX}}(\varepsilon) \leq t_{\text{MIX}}(\tfrac{1}{4}) \lceil \tfrac{1}{2} \log_2(1/\varepsilon) \rceil \quad \text{for any } 0 < \varepsilon < \tfrac{1}{4} . \tag{2.3}$$

In particular, for $\varepsilon < \frac{1}{4}$ our assumption on t_{REL} gives

$$t_{\text{MIX}}(\varepsilon) - t_{\text{MIX}}(1-\varepsilon) \leq t_{\text{MIX}}(\varepsilon) \leq \varepsilon^{-5/2} \log_2(1/\varepsilon) \sqrt{t_{\text{REL}} \cdot t_{\text{MIX}}(\tfrac{1}{4})} .$$

Therefore, a choice of

$$c_2 = c_2(\varepsilon) = \max\{\log_2(1/\varepsilon)/\varepsilon^{5/2}, 64\}$$

gives (1.2) for any $\varepsilon < \frac{1}{2}$ (the case $\varepsilon > \frac{1}{4}$ again follows from monotonicity).

Altogether, a choice of $c_\varepsilon = \max\{c_1, c_2\}$ completes the proof. ∎

In the remainder of this section, we provide the proof of Theorem 2.2. To this end, we must first establish several lemmas.

Let $X = X(t)$ be the given (lazy irreducible) birth-and-death chain, and from now on, let $\Omega_n = \{0, \ldots, n\}$ denote its state space. Let P denote the transition kernel of X, and let π denote its stationary distribution. Our first argument relates the mixing-time of the chain, starting from various starting positions, with its hitting time from 0 to certain quantile states, defined next.

$$Q(\varepsilon) := \min\left\{k : \sum_{j=0}^{k} \pi(j) \geq \varepsilon\right\}, \quad \text{where } 0 < \varepsilon < 1 . \tag{2.4}$$

Similarly, one may define the hitting times from n as follows:

$$\tilde{Q}(\varepsilon) := \max\left\{k : \sum_{j=k}^{n} \pi(j) \geq \varepsilon\right\}, \quad \text{where } 0 < \varepsilon < 1 . \tag{2.5}$$

Remark. Throughout the proof, we will occasionally need to shift from $Q(\varepsilon)$ to $\tilde{Q}(1-\varepsilon)$, and vice versa. Though the proof can be written in terms of $Q, \tilde{Q}$, for the sake of simplicity it will be easier to have the symmetry

$$Q(\varepsilon) = \tilde{Q}(1-\varepsilon) \text{ for almost any } \varepsilon > 0 . \tag{2.6}$$

This is easily achieved by noticing that at most n values of ε do not satisfy (2.6) for a given chain $X(t)$ on n states. Hence, for any given countable family of chains, we can eliminate a countable set of all such problematic values of ε and obtain the above mentioned symmetry.

Recalling that we defined $\mathbf{P}_k$ to be the probability on the event that the starting position is k, we define $\mathbf{E}_k$ and Var_k analogously. Finally, here and in what follows, let τ_k denote the hitting-time of the state k, that is, $\tau_k := \min\{t : X(t) = k\}$.

Lemma 2.3. *For any fixed $0 < \varepsilon < 1$ and lazy irreducible birth-and-death chain X, the following holds for any t:*

$$\|P^t(0,\cdot) - \pi\|_{\mathrm{TV}} \le \mathbf{P}_0(\tau_{Q(1-\varepsilon)} > t) + \varepsilon\,, \tag{2.7}$$

and for all $k \in \Omega$,

$$\|P^t(k,\cdot) - \pi\|_{\mathrm{TV}} \le \mathbf{P}_k(\max\{\tau_{Q(\varepsilon)}, \tau_{Q(1-\varepsilon)}\} > t) + 2\varepsilon\,. \tag{2.8}$$

Proof. Let X denote an instance of the lazy birth-and-death chain starting from a given state k, and let $\tilde{X}$ denote another instance of the lazy chain starting from the stationary distribution. Consider the following *no-crossing* coupling of these two chains: at each step, a fair coin toss decides which of the two chains moves according to its original (non-lazy) rule. Clearly, this coupling does not allow the two chains to cross one another without sharing the same state first (hence the name for the coupling). Furthermore, notice that by definition, each of the two chains, given the number of coin tosses that went its way, is independent of the other chain. Finally, for any t, $\tilde{X}(t)$, given the number of coin tosses that went its way until time t, has the stationary distribution.

In order to deduce the mixing-times bounds, we show an upper bound on the time it takes X and $\tilde{X}$ to coalesce. Consider the hitting time of X from 0 to $Q(1-\varepsilon)$, denoted by $\tau_{Q(1-\varepsilon)}$. By the above argument, $\tilde{X}(\tau_{Q(1-\varepsilon)})$ enjoys the stationary distribution, hence by the definition of $Q(1-\varepsilon)$,

$$\mathbf{P}\left(\tilde{X}(\tau_{Q(1-\varepsilon)}) \le X(\tau_{Q(1-\varepsilon)})\right) \ge 1 - \varepsilon\,.$$

Therefore, by the property of the no-crossing coupling, X and $\tilde{X}$ must have coalesced by time $\tau_{Q(1-\varepsilon)}$ with probability at least $1-\varepsilon$. This implies (2.7), and it remains to prove (2.8). Notice that the above argument involving the no-crossing coupling, this time with X starting from k, gives

$$\mathbf{P}\left(\tilde{X}(\tau_{Q(\varepsilon)}) \ge X(\tau_{Q(\varepsilon)})\right) \ge 1 - \varepsilon\,,$$

and similarly,

$$\mathbf{P}\left(\tilde{X}(\tau_{Q(1-\varepsilon)}) \le X(\tau_{Q(1-\varepsilon)})\right) \ge 1 - \varepsilon\,.$$

Therefore, the probability that X and $\tilde{X}$ coalesce between the times $\tau_{Q(\varepsilon)}$ and $\tau_{Q(\varepsilon)}$ is at least $1-2\varepsilon$, completing the proof. ∎

Corollary 2.4. *Let $X(t)$ be a lazy irreducible birth-and-death chain on Ω_n. The following holds for any $0 < \varepsilon < \frac{1}{16}$:*

$$t_{\mathrm{MIX}}(\tfrac{1}{4}) \le 16 \max\left\{\mathbf{E}_0\tau_{Q(1-\varepsilon)}, \mathbf{E}_n\tau_{Q(\varepsilon)}\right\}\,. \tag{2.9}$$

Proof. Clearly, for any source and target states $x, y \in \Omega$, at least one of the endpoints $s \in \{0, n\}$ satisfies $\mathbf{E}_s\tau_y \geq \mathbf{E}_x\tau_y$ (by the definition of the birth-and-death chain). Therefore, if T denotes the right-hand-side of (2.9), then

$$\mathbf{P}_x(\max\{\tau_{Q(\varepsilon)}, \tau_{Q(1-\varepsilon)}\} \geq T) \leq \mathbf{P}_x(\tau_{Q(\varepsilon)} \geq T) + \mathbf{P}_x(\tau_{Q(1-\varepsilon)} \geq T) \leq \frac{1}{8}\,,$$

where the last transition is by Markov's inequality. The proof now follows directly from (2.8). ∎

Remark. The above corollary shows that the order of the mixing time is at most $\max\{\mathbf{E}_0\tau_{Q(1-\varepsilon)}, \mathbf{E}_n\tau_{Q(\varepsilon)}\}$. It is in fact already possible (and not difficult) to show that the mixing time has this order *precisely*. However, our proof only uses the order of the mixing-time as an upper-bound, in order to finally deduce a stronger result: this mixing-time is asymptotically equal to the above maximum of the expected hitting times.

Having established that the order of the mixing-time is at most the expected hitting time of $Q(1-\varepsilon)$ and $Q(\varepsilon)$ from the two endpoints of Ω, assume here and in what follows, without loss of generality, that $\mathbf{E}_0\tau_{Q(1-\varepsilon)}$ is at least $\mathbf{E}_n\tau_{Q(\varepsilon)}$. Thus, (2.9) gives

$$t_{\text{MIX}}(\tfrac{1}{4}) \leq 16 \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)} \quad \text{for any } 0 < \varepsilon < \tfrac{1}{16}\,. \tag{2.10}$$

A key element in our estimation is a result of Karlin and McGregor [16, Equation (45)], reproved by Keilson [17], which represents hitting-times for birth-and-death chains in continuous-time as a sum of independent exponential variables (see [13],[9], [14] for more on this result). The discrete-time version of this result was given by Fill [13, Theorem 1.2].

Theorem 2.5 ([13])**.** *Consider a discrete-time birth-and-death chain with transition kernel P on the state space $\{0, \ldots, d\}$ started at 0. Suppose that d is an absorbing state, and suppose that the other birth probabilities p_i, $0 \leq i \leq d-1$, and death probabilities q_i, $1 \leq i \leq d-1$, are positive. Then the absorption time in state d has probability generating function*

$$u \mapsto \prod_{j=0}^{d-1} \Big[\frac{(1-\theta_j)u}{1-\theta_j u}\Big]\,, \tag{2.11}$$

where $-1 \leq \theta_j < 1$ are the d non-unit eigenvalues of P. Furthermore, if P has nonnegative eigenvalues then the absorption time in state d is distributed as the sum of d independent geometric random variables whose failure probabilities are the non-unit eigenvalues of P.

The above theorem provides means of establishing the concentration of the passage time from left to right of a chain, where the target (right end) state is turned into an absorbing state. Since we are interested in the hitting

time from one end to a given state (namely, from 0 to $Q(1-\varepsilon)$), it is clearly equivalent to consider the chain where the target state is absorbing. We thus turn to handle the hitting time of an absorbing end of a chain starting from the other end. The following lemma will infer its concentration from Theorem 2.5.

Lemma 2.6. *Let $X(t)$ be a lazy irreducible birth-and-death chain on the state space $\{0,\ldots,d\}$, where d is an absorbing state, and let* gap *denote its spectral gap. Then* $\mathrm{Var}_0\,\tau_d \le (\mathbf{E}_0\tau_d)\,/\mathsf{gap}$.

Proof. Let $\theta_0 \ge \ldots \ge \theta_{d-1}$ denote the d non-unit eigenvalues of the transition kernel of X. Recalling that X is a lazy irreducible birth-and-death chain, $\theta_i \ge 0$ for all i, hence the second part of Theorem 2.5 implies that $\tau_d \sim \sum_{i=0}^{d-1} Y_i$, where the Y_i-s are independent geometric random variables with means $1/(1-\theta_i)$. Therefore,

$$\mathbf{E}_0\tau_d = \sum_{i=0}^{d-1} \frac{1}{1-\theta_i}\,, \quad \mathrm{Var}_0\,\tau_d = \sum_{i=0}^{d-1} \frac{\theta_i}{(1-\theta_i)^2}\,, \tag{2.12}$$

which, using the fact that $\theta_0 \ge \theta_i$ for all i, gives

$$\mathrm{Var}_0\,\tau_d \le \frac{1}{1-\theta_0}\sum_{i=0}^{d-1}\frac{1}{1-\theta_i} = \frac{\mathbf{E}_0\tau_d}{\mathsf{gap}}\,,$$

as required. ∎

As we stated before, the hitting time of a state in our original chain has the same distribution as the hitting time in the modified chain (where this state is set to be an absorbing state). In order to derive concentration from the above lemma, all that remains is to relate the spectral gaps of these two chains. This is achieved by the next lemma.

Lemma 2.7. *Let $X(t)$ be a lazy irreducible birth-and-death chain, and* gap *be its spectral gap. Set $0<\varepsilon<1$, and let $\ell = Q(1-\varepsilon)$. Consider the modified chain $Y(t)$, where ℓ is turned into an absorbing state, and let* $\mathsf{gap}|_{[0,\ell]}$ *denote its spectral gap. Then* $\mathsf{gap}|_{[0,\ell]} \ge \varepsilon\cdot\mathsf{gap}$.

Proof. By [4, Chapter 3, Section 6], we have

$$\mathsf{gap} = \min_{\substack{f\,:\,\mathbf{E}_\pi f=0\\ f\not\equiv 0}} \frac{\langle (I-P)f, f\rangle_\pi}{\langle f, f\rangle_\pi} = \min_{\substack{f\,:\,\mathbf{E}_\pi f=0\\ f\not\equiv 0}} \frac{1}{2}\frac{\sum_{i,j}(f(i)-f(j))^2\,P(i,j)\pi(i)}{\sum_i f(i)^2\pi(i)}\,. \tag{2.13}$$

Observe that $\mathsf{gap}|_{[0,\ell]}$ is precisely $1-\lambda$, where λ is the largest eigenvalue of $P|_\ell$, the principal sub-matrix on the first ℓ rows and columns, indexed by

$\{0, \ldots, \ell - 1\}$ (notice that this sub-matrix is strictly sub-stochastic, as X is irreducible). Being a birth-and-death chain, X is reversible, that is,

$$P_{ij}\pi(i) = P_{ji}\pi(j) \text{ for any } i, j \,.$$

Therefore, it is simple to verify that $P|_\ell$ is a symmetric operator on $\mathbb{R}^\ell$ with respect to the inner-product $\langle \cdot, \cdot \rangle_\pi$; that is, $\langle P|_\ell x, y\rangle_\pi = \langle x, P|_\ell y\rangle_\pi$ for every $x, y \in \mathbb{R}^\ell$, and hence the Rayleigh-Ritz formula holds (cf., e.g., [15]), giving

$$\lambda = \max_{\substack{x\in\mathbb{R}^\ell \\ x\neq 0}} \frac{\langle P|_\ell x, x\rangle_\pi}{\langle x, x\rangle_\pi} \,.$$

It follows that

$$\begin{aligned} \mathsf{gap}|_{[0,\ell]} = 1 - \lambda &= \min_{\substack{f \,:\, f\not\equiv 0 \\ f(k)=0 \ \forall k\geq \ell}} \frac{\sum_{i=0}^n \left(f(i) - \sum_{j=0}^n P(i,j) f(j)\right) f(i)\pi(i)}{\sum_{i=0}^n f(i)^2\pi(i)} \\ &= \min_{\substack{f \,:\, f\not\equiv 0 \\ f(k)=0 \ \forall k\geq \ell}} \frac{1}{2} \frac{\sum_{0\leq i,j\leq n} (f(i) - f(j))^2 \, P(i,j)\pi(i)}{\sum_{i=0}^n f(i)^2\pi(i)} \,, \end{aligned} \tag{2.14}$$

where the last equality is by the fact that P is stochastic.

Observe that (2.13) and (2.14) have similar forms, and for any f (which can also be treated as a random variable) we can write $\tilde{f} = f - \mathbf{E}_\pi f$ such that $\mathbf{E}_\pi \tilde{f} = 0$. Clearly,

$$(f(i) - f(j))^2 \, P(i,j)\pi(i) = (\tilde{f}(i) - \tilde{f}(j))^2 P(i,j)\pi(i),$$

hence in order to compare gap and $\mathsf{gap}|_{[0,\ell]}$, it will suffice to compare the denominators of (2.13) and (2.14). Noticing that

$$\mathrm{Var}_\pi(f) = \sum_i \tilde{f}(i)^2\pi(i) \,, \quad \text{and } \mathbf{E}_\pi f^2 = \sum_i f(i)^2\pi(i) \,,$$

we wish to bound the ratio between the above two terms. Without loss of generality, assume that $\mathbf{E}_\pi f = 1$. Then every f with $f(k) = 0$ for all $k \geq \ell$ satisfies

$$\frac{\mathbf{E}_\pi f^2}{\pi(f \neq 0)} = \mathbf{E}_\pi \left[f^2 \mid f \neq 0\right] \geq \left(\mathbf{E}_\pi \left[f \mid f \neq 0\right]\right)^2 = \left(\pi\left(f \neq 0\right)\right)^{-2} \,,$$

and hence

$$\frac{1}{\mathbf{E}_\pi f^2} \leq \pi(f \neq 0) \leq 1 - \varepsilon \,, \tag{2.15}$$

where the last inequality is by the definition of ℓ as $Q(1 - \varepsilon)$. Once again, using the fact that $\mathbf{E}_\pi f = 1$, we deduce that

$$\frac{\mathrm{Var}_\pi f}{\mathbf{E}_\pi f^2} = 1 - \frac{1}{\mathbf{E}_\pi f^2} \geq \varepsilon \,. \tag{2.16}$$

Altogether, by the above discussion on the comparison between (2.13) and (2.14), we conclude that $\mathsf{gap}|_{[0,\ell]} \geq \varepsilon \cdot \mathsf{gap}$. ∎

Combining Lemma 2.6 and Lemma 2.7 yields the following corollary:

Corollary 2.8. *Let $X(t)$ be a lazy irreducible birth-and-death chain on Ω_n, let* gap *denote its spectral-gap, and $0 < \varepsilon < 1$. The following holds:*

$$\mathrm{Var}_0\, \tau_{Q(1-\varepsilon)} \leq \frac{\mathbf{E}_0 \tau_{Q(1-\varepsilon)}}{\varepsilon \cdot \mathsf{gap}} \,. \tag{2.17}$$

Remark. The above corollary implies the following statement: whenever the product $\mathsf{gap} \cdot \mathbf{E}_0 \tau_{Q(1-\varepsilon)}$ tends to infinity with n, the hitting-time $\tau_{Q(1-\varepsilon)}$ is concentrated. This is essentially the case under the assumptions of Theorem 2.2 (which include a lower bound on $\mathsf{gap} \cdot t_{\text{MIX}}(\frac{1}{4})$ in terms of ε), as we already established in (2.10) that $\mathbf{E}_0 \tau_{Q(1-\varepsilon)} \geq \frac{1}{16} t_{\text{MIX}}(\frac{1}{4})$.

Recalling the definition of cutoff and the relation between the mixing time and hitting times of the quantile states, we expect that the behaviors of $\tau_{Q(\varepsilon)}$ and $\tau_{Q(1-\varepsilon)}$ would be roughly the same; this is formulated in the following lemma.

Lemma 2.9. *Let $X(t)$ be a lazy irreducible birth-and-death chain on Ω_n, and suppose that for some $0 < \varepsilon < \frac{1}{16}$ we have $t_{\text{REL}} < \varepsilon^4 \cdot \mathbf{E}_0 \tau_{Q(1-\varepsilon)}$. Then for any fixed $\varepsilon \leq \alpha < \beta \leq 1 - \varepsilon$:*

$$\mathbf{E}_{Q(\alpha)} \tau_{Q(\beta)} \leq \frac{3}{2\varepsilon} \sqrt{t_{\text{REL}} \cdot \mathbf{E}_0 \tau_{Q(\frac{1}{2})}} \,. \tag{2.18}$$

Proof. Since by definition, $\mathbf{E}_{Q(\varepsilon)} \tau_{Q(1-\varepsilon)} \geq \mathbf{E}_{Q(\alpha)} \tau_{Q(\beta)}$ (the left-hand-side can be written as a sum of three independent hitting times, one of which being the right-hand-side), it suffices to show (2.18) holds for $\alpha = \varepsilon$ and $\beta = 1 - \varepsilon$.

Consider the random variable ν, distributed according to the restriction of the stationary distribution π to $[Q(\varepsilon)] := \{0, \ldots, Q(\varepsilon)\}$, that is:

$$\nu(k) := \frac{\pi(k)}{\pi([Q(\varepsilon)])} \mathbf{1}_{\{[Q(\varepsilon)]\}} \,, \tag{2.19}$$

and let $w \in \mathbb{R}^\Omega$ denote the vector $w := \mathbf{1}_{\{[Q(\varepsilon)]\}} / \pi([Q(\varepsilon)])$. As X is reversible, the following holds for any k:

$$P^t(\nu, k) = \sum_i P^t(i,k) \pi(i) w(i) = (P^t w)(k) \cdot \pi(k) \,.$$

Thus, by the definition of the total-variation distance (for a finite space):

$$\begin{aligned} \|P^t(\nu, \cdot) - \pi(\cdot)\|_{\text{TV}} &= \frac{1}{2} \sum_{k=0}^{n} \pi(k) \left| (P^t w)(k) - 1 \right| = \frac{1}{2} \|P^t(w - \mathbf{1})\|_{L^1(\pi)} \\ &\leq \frac{1}{2} \|P^t(w - \mathbf{1})\|_{L^2(\pi)} \,, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality. As $w - \mathbf{1}$ is orthogonal to $\mathbf{1}$ in the inner-product space $\langle \cdot, \cdot \rangle_{L^2(\pi)}$, we deduce that

$$\|P^t(w - \mathbf{1})\|_{L^2(\pi)} \leq \lambda_2^t \|w - \mathbf{1}\|_{L^2(\pi)} ,$$

where λ_2 is the second largest eigenvalue of P. Therefore,

$$\|P^t(v, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq \frac{1}{2}\lambda_2^t \|w - \mathbf{1}\|_{L^2(\pi)} = \frac{1}{2}\lambda_2^t \sqrt{(1/\pi([Q(\varepsilon)])) - 1} \leq \frac{\lambda_2^t}{2\sqrt{\varepsilon}} ,$$

where the last inequality is by the fact that $\pi([Q(\varepsilon)]) \geq \varepsilon$ (by definition). Recalling that $t_{\mathrm{REL}} = \mathsf{gap}^{-1} = 1/(1 - \lambda_2)$, define

$$t_\varepsilon = 2 \log(1/\varepsilon) t_{\mathrm{REL}} .$$

Since $\log(1/x) \geq 1 - x$ for all $x \in (0, 1]$, it follows that $\lambda_2^{t_\varepsilon} \leq \varepsilon^2$, thus (with room to spare)

$$\left\| P^{t_\varepsilon}(v, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} \leq \varepsilon/2 .$$

We will next use a second moment argument to obtain an upper bound on the expected commute time. By (2.20) and the definition of the total-variation distance,

$$\mathbf{P}_v(\tau_{Q(1-\varepsilon)} \leq t_\varepsilon) \geq \varepsilon - \|P^{t_\varepsilon}(v, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \geq \varepsilon/2 , \tag{2.20}$$

whereas the definition of v as being supported by the range $[Q(\varepsilon)]$ gives

$$\mathbf{P}_v(\tau_{Q(1-\varepsilon)} \leq t_\varepsilon) \leq \mathbf{P}_{Q(\varepsilon)}(\tau_{Q(1-\varepsilon)} \leq t_\varepsilon) \leq \frac{\mathrm{Var}_{Q(\varepsilon)} \tau_{Q(1-\varepsilon)}}{\left| \mathbf{E}_{Q(\varepsilon)} \tau_{Q(1-\varepsilon)} - t_\varepsilon \right|^2} .$$

Combining the two,

$$\mathbf{E}_{Q(\varepsilon)} \tau_{Q(1-\varepsilon)} \leq t_\varepsilon + \sqrt{\frac{2}{\varepsilon} \mathrm{Var}_{Q(\varepsilon)} \tau_{Q(1-\varepsilon)}}. \tag{2.21}$$

Recall that starting from 0, the hitting time to point $Q(1 - \varepsilon)$ is exactly the sum of the hitting time from 0 to $Q(\varepsilon)$ and the hitting time from $Q(\varepsilon)$ to $Q(1 - \varepsilon)$, where both these hitting times are independent. Therefore,

$$\mathrm{Var}_{Q(\varepsilon)} \tau_{Q(1-\varepsilon)} \leq \mathrm{Var}_0 \tau_{Q(1-\varepsilon)} . \tag{2.22}$$

By (2.21) and (2.22) we get

$$\begin{aligned} \mathbf{E}_{Q(\varepsilon)} \tau_{Q(1-\varepsilon)} &\leq t_\varepsilon + \sqrt{\frac{2}{\varepsilon} \mathrm{Var}_0 \tau_{Q(1-\varepsilon)}} \\ &\leq 2 \log(1/\varepsilon) t_{\mathrm{REL}} + (1/\varepsilon) \sqrt{2 t_{\mathrm{REL}} \cdot \mathbf{E}_0 \tau_{Q(1-\varepsilon)}} , \end{aligned} \tag{2.23}$$

where the last inequality is by Corollary 2.8.

We now wish to rewrite the bound (2.23) in terms of $t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}$ using our assumptions on t_{REL} and $\mathbf{E}_0\tau_{Q(\frac{1}{2})}$. First, twice plugging in the fact that $t_{\text{REL}} < \varepsilon^4 \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)}$ yields

$$\begin{aligned}\mathbf{E}_{Q(\varepsilon)}\tau_{Q(1-\varepsilon)} &\le \left(2\varepsilon^3 \log(1/\varepsilon) + \sqrt{2}\right)\varepsilon \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)} \\ &\le \tfrac{3}{2}\varepsilon \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)}\,, \end{aligned} \tag{2.24}$$

where in the last inequality we used the fact that $\varepsilon < \frac{1}{16}$. In particular,

$$\mathbf{E}_0\tau_{Q(1-\varepsilon)} \le \mathbf{E}_0\tau_{Q(\frac{1}{2})} + \mathbf{E}_{Q(\varepsilon)}\tau_{Q(1-\varepsilon)} \le \mathbf{E}_0\tau_{Q(\frac{1}{2})} + \tfrac{3}{2}\varepsilon \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)}\,,$$

and after rearranging,

$$\mathbf{E}_0\tau_{Q(1-\varepsilon)} \le \left(\mathbf{E}_0\tau_{Q(\frac{1}{2})}\right) / \left(1 - \tfrac{3}{2}\varepsilon\right)\,. \tag{2.25}$$

Plugging this result back in (2.23), we deduce that

$$\mathbf{E}_{Q(\varepsilon)}\tau_{Q(1-\varepsilon)} \le 2\log(1/\varepsilon) \cdot t_{\text{REL}} + \frac{1}{\varepsilon}\sqrt{\frac{2t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}}{1 - \frac{3}{2}\varepsilon}}\,.$$

A final application of the fact $t_{\text{REL}} < \varepsilon^4 \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)}$, together with (2.25) and the fact that $\varepsilon < \frac{1}{16}$, gives

$$\begin{aligned}\mathbf{E}_{Q(\varepsilon)}\tau_{Q(1-\varepsilon)} &\le \left(\frac{2\varepsilon^2 \log(1/\varepsilon) + \sqrt{2}/\varepsilon}{\sqrt{1 - \frac{3}{2}\varepsilon}}\right)\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}} \\ &\le \frac{3}{2\varepsilon}\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}}\,, \end{aligned} \tag{2.26}$$

as required. ∎

We are now ready to prove the main theorem.

Proof of Theorem 2.2. Recall our assumption (without loss of generality)

$$\mathbf{E}_0\tau_{Q(1-\varepsilon)} \ge \mathbf{E}_n\tau_{Q(\varepsilon)}\,, \tag{2.27}$$

and define what would be two ends of the cutoff window:

$$\begin{cases} t^- = t^-(\gamma) := \mathbf{E}_0\tau_{Q(\frac{1}{2})} - \gamma\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}}\,, \\ t^+ = t^+(\gamma) := \mathbf{E}_0\tau_{Q(\frac{1}{2})} + \gamma\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}}\,. \end{cases}$$

For the lower bound, let $0 < \varepsilon < \frac{1}{16}$; combining (2.10) with the assumption that $t_{\text{REL}} \le \varepsilon^5 \cdot t_{\text{MIX}}(\frac{1}{4})$ gives

$$t_{\text{REL}} \le 16\varepsilon^5 \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)} < \varepsilon^4 \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)}\,. \tag{2.28}$$

Thus, we may apply Lemma 2.9 to get

$$\mathbf{E}_0\tau_{Q(\varepsilon)} \geq \mathbf{E}_0\tau_{Q(\frac{1}{2})} - \mathbf{E}_{Q(\varepsilon)}\tau_{Q(1-\varepsilon)} \geq \mathbf{E}_0\tau_{Q(\frac{1}{2})} - \frac{3}{2\varepsilon}\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}}\,.$$

Furthermore, recalling Corollary 2.8, we also have

$$\text{Var}_0\,\tau_{Q(\varepsilon)} \leq \frac{1}{1-\varepsilon} t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\varepsilon)} \leq 2t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}\,.$$

Therefore, by Chebyshev's inequality, the following holds for any $\gamma > \frac{3}{2\varepsilon}$:

$$\|P^{t^-}(0,\cdot) - \pi\|_{\text{TV}} \geq 1 - \varepsilon - \mathbf{P}_0(\tau_{Q(\varepsilon)} \leq t^-) \geq 1 - \varepsilon - 2\left(\gamma - \frac{3}{2\varepsilon}\right)^{-2}\,,$$

and a choice of $\gamma = 2/\varepsilon$ implies that (with room to spare, as $\varepsilon < \frac{1}{16}$)

$$t_{\text{MIX}}(1-2\varepsilon) \geq \mathbf{E}_0\tau_{Q(\frac{1}{2})} - (2/\varepsilon)\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}}\,. \tag{2.29}$$

The upper bound will follow from a similar argument. Take $0 < \varepsilon < \frac{1}{16}$ and recall that $t_{\text{REL}} < \varepsilon^4 \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)}$. Applying Corollary 2.8 and Lemma 2.9 once more (with (2.25) as well as (2.27) in mind) yields:

$$\begin{aligned}
\mathbf{E}_n\tau_{Q(\varepsilon)} &\leq \mathbf{E}_0\tau_{Q(1-\varepsilon)} \leq \mathbf{E}_0\tau_{Q(\frac{1}{2})} + \frac{3}{2\varepsilon}\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}}\,,\\
\text{Var}_0\,\tau_{Q(1-\varepsilon)} &\leq (1/\varepsilon)t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)} \leq \frac{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}}{\varepsilon(1-\frac{3}{2}\varepsilon)} \leq (2/\varepsilon)t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}\,,\\
\text{Var}_n\,\tau_{Q(\varepsilon)} &\leq (1/\varepsilon)t_{\text{REL}} \cdot \mathbf{E}_n\tau_{Q(\varepsilon)} \leq (2/\varepsilon)t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}\,.
\end{aligned}$$

Hence, combining Chebyshev's inequality with (2.8) implies that for all k and $\gamma > \frac{3}{2\varepsilon}$,

$$\begin{aligned}
\|P^{t^+}(k,\cdot) - \pi\|_{\text{TV}} &\leq 2\varepsilon + \mathbf{P}_0(\tau_{Q(1-\varepsilon)} > t^+) + \mathbf{P}_n(\tau_{Q(\varepsilon)} > t^+)\\
&\leq 2\varepsilon + \frac{4}{\varepsilon}\left(\gamma - \frac{3}{2\varepsilon}\right)^{-2}\,.
\end{aligned}$$

Again choosing $\gamma = 3/\varepsilon$ we conclude that (with room to spare)

$$t_{\text{MIX}}(4\varepsilon) \leq \mathbf{E}_0\tau_{Q(\frac{1}{2})} + (3/\varepsilon)\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0\tau_{Q(\frac{1}{2})}} \tag{2.30}$$

We have thus established the cutoff window in terms of t_{REL} and $\mathbf{E}_0\tau_{Q(\frac{1}{2})}$, and it remains to write it in terms of t_{REL} and t_{MIX}. To this end, recall that (2.25) implies that

$$t_{\text{REL}} < \varepsilon^4 \cdot \mathbf{E}_0\tau_{Q(1-\varepsilon)} \leq \frac{\varepsilon^4}{1-\frac{3}{2}\varepsilon}\mathbf{E}_0\tau_{Q(\frac{1}{2})}\,,$$

hence (2.29) gives the following for any $\varepsilon < \frac{1}{16}$:

$$t_{\text{MIX}}\left(\tfrac{1}{4}\right) \geq \left(1 - \frac{2\varepsilon}{\sqrt{1-\frac{3}{2}\varepsilon}}\right) \cdot \mathbf{E}_0 \tau_{Q(\frac{1}{2})} \geq \frac{5}{6} \mathbf{E}_0 \tau_{Q(\frac{1}{2})} \,. \tag{2.31}$$

Altogether, (2.29), (2.30) and (2.31) give

$$t_{\text{MIX}}(4\varepsilon) - t_{\text{MIX}}(1-2\varepsilon) \leq (5/\varepsilon)\sqrt{t_{\text{REL}} \cdot \mathbf{E}_0 \tau_{Q(\frac{1}{2})}} \leq (6/\varepsilon)\sqrt{t_{\text{REL}} \cdot t_{\text{MIX}}(\tfrac{1}{4})} \,,$$

completing the proof of the theorem. ∎

2.3. **Tightness of the bound on the cutoff window.** The bound $\sqrt{t_{\text{MIX}} \cdot t_{\text{REL}}}$ on the size of the cutoff window, given in Corollary 2, is essentially tight in the following sense. Suppose that $t_M(n)$ and $t_R(n) \geq 2$ are the mixing-time $t_{\text{MIX}}(\frac{1}{4})$ and relaxation-time t_{REL} of a family $(X_t^{(n)})$ of lazy irreducible birth-and-death chains that exhibits cutoff. For any fixed $\varepsilon > 0$, we construct a family $(Y_t^{(n)})$ of such chains satisfying

$$\begin{cases} (1-\varepsilon)t_M \leq t_{\text{MIX}}^{(n)}(\frac{1}{4}) \leq (1+\varepsilon)t_M \,, \\ |t_{\text{REL}}^{(n)} - t_R| \leq \varepsilon \,, \end{cases} \tag{2.32}$$

and in addition, having a cutoff window of size $(t_{\text{MIX}}^{(n)} \cdot t_{\text{REL}}^{(n)})^{1/2}$.

Our construction is as follows: we first choose n reals in $[0,1)$, which would serve as the nontrivial eigenvalues of our chain: any such sequence can be realized as the nontrivial eigenvalues of a birth-and-death chain with death probabilities all zero, and an absorbing state at n. Our choice of eigenvalues will be such that $t_{\text{MIX}}^{(n)} = (\frac{1}{2} + o(1))t_M$, $t_{\text{REL}}^{(n)} = \frac{1}{2}t_R$ and the chain will exhibit cutoff with a window of $\sqrt{t_M \cdot t_R}$. Finally, we perturb the chain to make it irreducible, and consider its lazy version to obtain (2.32).

First, notice that $t_R = o(t_M)$ (a necessary condition for the cutoff, as given by Corollary 2). Second, if a family of chains has mixing-time and relaxation-time t_M and t_R respectively, then the cutoff point is without loss of generality the expected hitting time from 0 to some state m (namely, for $m = Q(\frac{1}{2})$); let h_m denote this expected hitting time. Theorem 2.5, combined with Lemma 2.7, asserts that $h_m \leq m \cdot t_R$.

Setting $\varepsilon > 0$, we may assume that $t_R \geq 2(1+\varepsilon)$ (since $t_R \geq 2$, and a small additive error is permitted in (2.32)). Set $K = \frac{1}{2}h_m/t_R$, and define the following sequence of eigenvalues $\{\lambda_i\}$: the first $\lfloor K \rfloor$ eigenvalues will be equal to $\lambda := 1 - 2/t_R$, and the remaining eigenvalues will all have the value λ', such that the sum $\sum_{i=1}^{n} 1/(1-\lambda_i)$ equals $\frac{1}{2}h_m$ (our choice of K and the fact that $h_m \leq nt_R$ ensures that $\lambda' \leq \lambda$). By Theorem 2.5, the birth-and-death

chain with absorbing state in n which realizes these eigenvalues satisfies:

$$\begin{cases} t_{\text{MIX}}^{(n)} = (1+o(1))\mathbf{E}_0\tau_n = (\frac{1}{2}+o(1))t_M \,, \\ t_{\text{REL}}^{(n)} = \frac{1}{2}t_R \,, \\ \text{Var}_0\,\tau_n \geq \lfloor K \rfloor \frac{\lambda}{(1-\lambda)^2} \geq \frac{\varepsilon+o(1)}{8(1+\varepsilon)} t_M \cdot t_R \,, \end{cases}$$

where in the last inequality we merely considered the contribution of the first K geometric random variables to the variance. Continuing to focus on the sum of these K i.i.d. random variables, and recalling that $K \to \infty$ with n (by the assumption $t_R = o(t_M)$), the Central-Limit-Theorem implies that

$$\mathbf{P}_0(\tau_n - \mathbf{E}_0\tau_n > \gamma\sqrt{t_M \cdot t_R}) \geq c(\gamma,\varepsilon) > 0 \quad \text{for any } \gamma > 0 \,.$$

Hence, the cutoff window of this chain has order at least $\sqrt{t_M \cdot t_R}$.

Clearly, perturbing the transition kernel to have all death-probabilities equal some ε' (giving an irreducible chain), shifts every eigenvalue by at most ε' (note that τ_n from 0 has the same distribution if n is an absorbing state). Finally, the lazy version of this chain has twice the values of $\mathbf{E}_0\tau_n$ and t_{REL}, giving the required result (2.32).

3. Continuous-time chains and δ-lazy discrete-time chains

In this section, we discuss the versions of Corollary 2 (and Theorem 2.2) for the cases of either continuous-time chains (Theorem 3), or δ-lazy discrete-time chains (Theorem 3.1). Since the proofs of these versions follow the original arguments almost entirely, we describe only the modifications required in the new settings.

3.1. **Continuous-time birth-and-death chains.**

In order to prove Theorem 3, recall the definition of the heat-kernel of a continuous-time chain as $H_t(x,y) := \mathbf{P}_x(X_t = y)$, rewritten in matrix-representation as $H_t = \mathrm{e}^{t(P-I)}$ (where P is the transition kernel of the chain).

It is well known (and easy) that if $H_t, \widetilde{H}_t$ are the heat-kernels corresponding to the continuous-time chain and the lazy continuous-time chain, then $H_t = \widetilde{H}_{2t}$ for any t. This follows immediately from the next simple and well-known matrix-exponentiation argument shows:

$$H_t = \mathrm{e}^{t(P-I)} = \mathrm{e}^{2t(\frac{P+I}{2}-I)} = \widetilde{H}_{2t} \,. \tag{3.1}$$

Hence, it suffices to show cutoff for the lazy continuous-time chains. We therefore need to simply adjust the original proof dealing with lazy irreducible chains, from the discrete-time case to the continuous-time case.

The first modification is in the proof of Lemma 2.3, where a no-crossing coupling was constructed for the discrete-time chain. Clearly, no such coupling is required for the continuous case, as the event that the two chains cross one another at precisely the same time now has probability 0.

To complete the proof, one must show that the statement of Corollary 2.8 still holds; indeed, this follows from the fact that the hitting time $\tau_{Q(1-\varepsilon)}$ of the discrete-time chain is concentrated, combined with the concentration of the sum of the exponential variables that determine the timescale of the continuous-time chain.

3.2. **Discrete-time δ-lazy birth-and-death chains.**

Theorem 3.1. *Let $(X_t^{(n)})$ be a family of discrete-time δ-lazy birth-and-death chains, for some fixed $\delta > 0$. Then $(X_t^{(n)})$ exhibits cutoff in total-variation iff $t_{\text{REL}}^{(n)} = o(t_{\text{MIX}}^{(n)})$, and the cutoff window size is at most $\sqrt{t_{\text{MIX}}^{(n)}(\frac{1}{4}) \cdot t_{\text{REL}}^{(n)}}$.*

Proof. In order to extend Theorem 2.2 and Corollary 2 to δ-lazy chains, notice that there are precisely two locations where their proof rely on the fact that the chain is lazy. The first location is the construction of the no-crossing coupling in the proof of Lemma 2.3. The second location is the fact that all eigenvalues are non-negative in the application of Theorem 2.5.

Though we can no longer construct a no-crossing coupling, Lemma 2.3 can be mended as follows: recalling that $P(x, x) \geq \delta$ for all $x \in \Omega$, define $P' = \frac{1}{1-\delta}(P - \delta I)$, and notice that P' and P share the same stationary distribution (and hence define the same quantile states $Q(\varepsilon)$ and $Q(1-\varepsilon)$ on Ω). Let X' denote a chain which has the transition kernel P', and X denote its coupled appropriate lazy version: the number of steps it takes X to perform the corresponding move of X' is an independent geometric random variable with mean $1 - \delta$.

Set $p = 1 - \delta(1 - 2\varepsilon)$, and condition on the path of the chain X', from the starting point and until this chain completes $T = \log_p \varepsilon$ rounds from $Q(\varepsilon)$ to $Q(1-\varepsilon)$, back and forth. As argued before, as X follows this path, upon completion of each commute time from $Q(\varepsilon)$ to $Q(1-\varepsilon)$ and back, it has probability $1 - 2\varepsilon$ to cross $\tilde{X}$. Hence, by definition, in each such trip there is a probability of at least $\delta(1 - 2\varepsilon)$ that X and $\tilde{X}$ coalesce. Crucially, these events are independent, since we pre-conditioned on the trajectory of X'. Thus, after T such trips, the X and $\tilde{X}$ have a probability of at least $1 - \varepsilon$ to meet, as required.

It remains to argue that the expressions for the expectation and variance of the hitting-times, which were derived from Theorem 2.5, remain unchanged when moving from the $\frac{1}{2}$-lazy setting to δ-lazy chains. Indeed, this follows directly from the expression for the probability-generating-function, as given in (2.11). ∎

4. Separation in birth-and-death chains

In this section, we provides the proofs for Proposition 4 and Corollary 5.

Let (X_t) be an ergodic birth-and-death chain on $\Omega = \{0, \ldots, n\}$, with a transition kernel P and stationary distribution π. Let $d_{\text{sep}}(t; x)$ denote the separation of X, started from x, from π, that is

$$d_{\text{sep}}(t; x) := \max_{y \in \Omega} \left(1 - P^t(x, y)/\pi(y)\right) .$$

According to this notation, $d_{\text{sep}}(t) := \max_{x \in \Omega} d_{\text{sep}}(t; x)$ measures separation from the worst starting position.

The chain X is called *monotone* iff $P_{i,i+1} + P_{i+1,i} \leq 1$ for all $i < n$. It is well known (and easy to show) that if X is monotone, then the likelihood ratio $P^t(0, k)/\pi(k)$ is monotone decreasing in k (see, e.g., [8]). An immediate corollary of this fact is that the separation of such a chain from the the stationary distribution is the same for the two starting points $\{0, n\}$. We provide the proof of this simple fact for completeness.

Lemma 4.1. *Let P be the transition kernel of a monotone birth-and-death chain on $\Omega = \{0, \ldots, n\}$. If $f : \Omega \to \mathbb{R}$ is a monotone increasing (decreasing) function, so is Pf. In particular,*

$$P^t(k, 0) \geq P^t(k + 1, 0) \text{ for any } t \geq 0 \text{ and } 0 \leq k < n . \tag{4.1}$$

Proof. Let $\{p_i\}$, $\{q_i\}$ and $\{r_i\}$ denote the birth, death and holding probabilities of the chain respectively, and for convenience, let $f(x)$ be 0 for any $x \notin \Omega$. Assume without loss of generality that f is increasing (otherwise, one may consider $-f$). In this case, the following holds for every $0 \leq x < n$:

$$\begin{aligned} Pf(x) &= q_x f(x - 1) + r_x f(x) + p_x f(x + 1) \\ &\leq (1 - p_x) f(x) + p_x f(x + 1) , \end{aligned}$$

and

$$\begin{aligned} Pf(x + 1) &= q_{x+1} f(x) + r_{x+1} f(x + 1) + p_{x+1} f(x + 2) \\ &\geq (1 - q_{x+1}) f(x) + q_{x+1} f(x + 1) . \end{aligned}$$

Therefore, by the monotonicity of f and the fact that $p_x + q_{x+1} \leq 1$ we obtain that $Pf(x) \leq Pf(x + 1)$, as required.

Finally, the monotonicity of the chain implies that $P^t(\cdot, 0)$ is monotone decreasing for $t = 1$, hence the above argument immediately implies that this is the case for any integer $t \geq 1$. ∎

By reversibility, the following holds for any monotone birth-and-death chain with transition kernel P and stationary distribution π:

$$\frac{P^t(0, k)}{\pi(k)} \geq \frac{P^t(0, k + 1)}{\pi(k + 1)} \text{ for any } t \geq 0 \text{ and } 0 \leq k < n . \tag{4.2}$$

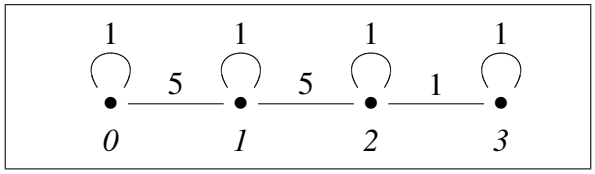

FIGURE 1. A monotone irreducible birth-and-death chain where worst separation may not involve the endpoints. Edge weights denote the conductances (see Example 4.3).

In particular, the maximum of $1 - P^t(0, j)/\pi(j)$ is attained at $j = n$, and the separation is precisely the same when starting at either of the two endpoints:

Corollary 4.2. *Let (X_t) be a monotone irreducible birth-and-death chain on $\Omega = \{0, \ldots, n\}$ with transition kernel P and stationary distribution π. Then for any integer t,*

$$\operatorname{sep}(P^t(0,\cdot),\pi) = 1 - \frac{P^t(0,n)}{\pi(n)} = 1 - \frac{P^t(n,0)}{\pi(0)} = \operatorname{sep}(P^t(n,\cdot),\pi) \ .$$

Since lazy chains are a special case of monotone chains, the relation (3.1) between lazy and non-lazy continuous-time chains gives an analogous statement for continuous-time irreducible birth-and-death chains. That is, for any real $t > 0$,

$$\operatorname{sep}(H_t(0,\cdot),\pi) = 1 - \frac{H_t(0,n)}{\pi(n)} = 1 - \frac{H_t(n,0)}{\pi(0)} = \operatorname{sep}(H_t(n,\cdot),\pi) \ ,$$

where H_t is the heat-kernel of the chain, and π is its stationary distribution.

Unfortunately, when attempting to generalize Lemma 4.1 (and Corollary 4.2) to an arbitrary starting point, one finds that it is no longer the case that the worst separation involves one of the endpoints, even if the chain is monotone and irreducible. This is demonstrated next.

Example 4.3. Let P and π denote the transition kernel and stationary distribution of the birth-and-death chain on the state space $\Omega = \{0, 1, 2, 3\}$, given in Figure 1. It is easy to verify that this chain is monotone and irreducible, and furthermore, that the following holds:

$$\min_{y\in\Omega} \frac{P^2(1,y)}{\pi(y)} \text{ is attained solely at } y = 2,$$
$$\min_{x,y\in\Omega} \frac{P^3(x,y)}{\pi(y)} \text{ is attained solely at } x = y = 1.$$

Thus, when starting from an interior point, the worst separation might not be attained by an endpoint, and in addition, the overall worst separation may not involve the endpoints at all.

However, as we next show, once we replace the monotonicity requirement with the stricter assumption that the chain is *lazy*, it turns out that the above phenomenon can no longer occur.

The approach that led to the following result relied on *maximal couplings* (see, e.g., [19], [22] and [18], and also [23, Chapter III.3]). We provide a straightforward proof for it, based on an inductive argument.

Lemma 4.4. *Let P be the transition kernel of a lazy birth-and-death chain. Then for any unimodal non-negative $f : \Omega \to \mathbb{R}^+$, the function Pf is also unimodal. In particular, for any integer t, all columns of P^t are unimodal.*

Proof. Let $\{p_i\}$, $\{q_i\}$ and $\{r_i\}$ be the birth, death and holding probabilities of the chain respectively, and for convenience, define $f(i)$ to be 0 for $i \in \mathbb{N} \setminus \Omega$. Let $m \in \Omega$ be a state achieving the global maximum of f, and set $g = Pf$.

For every $0 < x < m$, the unimodality of f implies that

$$\begin{aligned} g(x) &= q_x f(x-1) + r_x f(x) + p_x f(x+1) \\ &\geq q_x f(x-1) + (1-q_x) f(x) \,, \end{aligned}$$

and similarly,

$$\begin{aligned} g(x-1) &= q_{x-1} f(x-2) + r_{x-1} f(x-1) + p_{x-1} f(x) \\ &\leq (1-p_{x-1}) f(x-1) + p_{x-1} f(x) \,. \end{aligned}$$

Therefore, by the monotonicity of the chain, we deduce that $g(x) \geq g(x-1)$. The same argument shows that for every $m < y < n$ we have $g(y) \geq g(y+1)$.

As g is increasing on $\{0, \ldots, m-1\}$ and decreasing on $\{m+1, \ldots, n\}$, unimodality will follow from showing that $g(m) \geq \min\{g(m-1), g(m+1)\}$ (the global maximum of g would then be attained at $m' \in \{m-1, m, m+1\}$). To this end, assume without loss of generality that $f(m-1) \geq f(m+1)$. The following holds:

$$\begin{aligned} g(m) &= q_m f(m-1) + r_m f(m) + p_m f(m+1) \\ &\geq r_m f(m) + (1-r_m) f(m+1) \,, \end{aligned}$$

and

$$\begin{aligned} g(m+1) &= q_{m+1} f(m) + r_{m+1} f(m+1) + p_{m+1} f(m+2) \\ &\leq q_{m+1} f(m) + (1-q_{m+1}) f(m+1) \,. \end{aligned}$$

Thus, the laziness of the chain implies that $g(m) \geq g(m+1)$, as required. ∎

By reversibility, Lemma 4.4 has the following corollary:

Corollary 4.5. *Let (X_t) be a lazy and irreducible birth-and-death chain on the state space $\Omega = \{0, \ldots, n\}$, with transition kernel P and stationary distribution π. Then for any $s \in \Omega$ and any integer $t \geq 0$, the function $f(x) := P^t(s,x)/\pi(x)$ is unimodal.*

Remark. The maximum of the unimodal function $f(x)$ in Corollary 4.5 need not be located at $x = s$, the starting point of the chain. This can be demonstrated, e.g., by the biased random walk.

Proposition 4 will immediately follow from the above results.

Proof of Proposition 4. We begin with the case where (X_t) is a lazy birth-and-death chain, with transition kernel P. Let $s \in \Omega$ be a starting position which maximizes $d_{\text{sep}}(t)$. Then by Corollary 4.5, $d_{\text{sep}}(t)$ is either equal to $1 - P^t(s,0)/\pi(0)$ or to $1 - P^t(s,n)/\pi(n)$. Consider the first case (the second case is treated by the exact same argument); by reversibility,

$$d_{\text{sep}}(t) = 1 - \frac{P^t(0,s)}{\pi(s)} \le 1 - \frac{P^t(0,n)}{\pi(n)} ,$$

where the last inequality is by Lemma 4.1. Therefore, the endpoints of X assume the worst separation distance at every time t.

To show that $d_{\text{sep}}(t) = 1 - H_t(0,n)/\pi(n)$ in the continuous-time case, recall that

$$H_t(x,y) = \mathbf{P}_x(X_t = y) = \mathbf{E}\left[P^{N_t}(x,y)\right] = \sum_k P^k(x,y)\mathbf{P}(N_t = k) ,$$

where P is the transition kernel of the corresponding discrete-time chain, and N_t is a Poisson random variable with mean t. Though P^k has unimodal columns for any integer k, a linear combination of the matrices P^k does not necessarily maintain this property. We therefore consider a variant of the process, where N_t is approximated by an appropriate binomial variable.

Fix $t > 0$, and for any integer $m \ge 2t$ let $N'_t(m)$ be a binomial random variable with parameters $\text{Bin}(m, t/m)$. Since $N'_t(m)$ converges in distribution to N_t, it follows that $H'_t(m) := \mathbf{E}\left[P^{N'_t(m)}\right]$ converges to H_t as $m \to \infty$. Writing $N'_t(m)$ as a sum of independent indicators $\{B_i : i = 1, \dots, m\}$ with success probabilities t/m, and letting $Q := \left(1 - \frac{t}{m}\right) I + \frac{t}{m} P$, we have

$$H'_t(m) = \mathbf{E}\left[P^{\sum_{i=1}^m B_i}\right] = Q^m .$$

Note that for every $m \ge 2t$, the transition kernel Q corresponds to a lazy birth-and-death chain, thus Lemma 4.4 ensures that $H'_t(m)$ has unimodal columns for every such m. In particular, $H_t = \lim_{m\to\infty} H'_t(m)$ has unimodal columns. This completes the proof. ∎

Proof of Corollary 5. By Theorem 3, total-variation cutoff (from the worst starting position) occurs iff $t_{\text{REL}} = o(t_{\text{MIX}}(\frac{1}{4}))$. Combining Proposition 4 with [10, Theorem 5.1] we deduce that separation cutoff (from the worst starting point) occurs if and only if $t_{\text{REL}} = o(t_{\text{sep}}(\frac{1}{4}))$ (where $t_{\text{sep}}(\varepsilon) = \max_x t_{\text{sep}}(\varepsilon; x)$ is the minimum t such that $\max_x \text{sep}(H_t(x,\cdot), \pi) \le \varepsilon$).

Therefore, the proof will follow from the well known fact that $t_{\text{sep}}(\frac{1}{4})$ and $t_{\text{MIX}}(\frac{1}{4})$ have the same order. One can obtain this fact, for instance, from Lemma 7 of [4, Chapter 4], which states that (as the chain is reversible)

$$\bar{d}(t) \le d_{\text{sep}}(t)\,, \quad \text{and} \quad d_{\text{sep}}(2t) \le 1 - \left(1 - \bar{d}(t)\right)^2\,,$$

where $\bar{d}(t) := \max_{x,y\in\Omega} \left\|\mathbf{P}_x(X_t \in \cdot) - \mathbf{P}_y(X_t \in \cdot)\right\|_{\text{TV}}$. Combining this with the sub-multiplicativity of $\bar{d}(t)$, and the fact that $d(t) \le \bar{d}(t) \le 2d(t)$ (see Definition 3.1 in [4, Chapter 4]), we obtain that for any t,

$$d(t) \le d_{\text{sep}}(t)\,, \quad \text{and} \quad d_{\text{sep}}(8t) \le 2\bar{d}(4t) \le 32\,(d(t))^4\,.$$

This in turn implies that $\frac{1}{8}t_{\text{sep}}(\frac{1}{4}) \le t_{\text{MIX}}(\frac{1}{4}) \le t_{\text{sep}}(\frac{1}{4})$, as required. ∎

5. Concluding remarks and open problems

- As stated in Corollary 5, our results on continuous-time birth-and-death chains, combined with those of [10], imply that cutoff in total-variation distance is equivalent to separation cutoff for such chains. This raises the following question:

  Question 5.1. Let $(X_t^{(n)})$ denote a family of irreducible reversible Markov chains, either in continuous-time or in lazy discrete-time. Is it true that there is cutoff in separation iff there is cutoff in total-variation distance (where the distance in both cases is measured from the worst starting position)?

- One might assume that the cutoff-criterion (1.4) also holds for close variants of birth-and-death chains. For that matter, we note that Aldous's example of a family of reversible Markov chains, which satisfies $t_{\text{REL}}^{(n)} = o(t_{\text{MIX}}^{(n)}(\frac{1}{4}))$ and yet does not exhibit cutoff, can be written so that each of its chains is a biased random walk on a cycle. In other words, it suffices that a family of birth-and-death chains permits the one extra transition between states 0 and n, and already the cutoff criterion (1.4) ceases to hold.

- Finally, it would be interesting to characterize the cutoff criterion in additional natural families of ergodic Markov chains.

  Question 5.2. Does (1.4) hold for the family of lazy simple random walks on vertex transitive bounded-degree graphs?

Acknowledgments

We thank Persi Diaconis, Jim Fill, Jim Pitman and Laurent Saloff-Coste for useful comments on an early draft.

REFERENCES

[1] D. Aldous, *Random walks on finite groups and rapidly mixing Markov chains*, Seminar on probability, XVII, 1983, pp. 243–297.

[2] D. Aldous, American Institute of Mathematics (AIM) research workshop "Sharp Thresholds for Mixing Times" (Palo Alto, December 2004). Summary available at `http://www.aimath.org/WWN/mixingtimes`.

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

[4] D. Aldous and J. A. Fill, *Reversible Markov Chains and Random Walks on Graphs*. In preparation, `http://www.stat.berkeley.edu/~aldous/RWG/book.html`.

[5] G.-Y. Chen, *The cut-off phenomenon for finite Markov chains*, Ph.D. dissertation, Cornell University (2006).

[6] G.-Y. Chen and L. Saloff-Coste, *The cutoff phenomenon for ergodic Markov processes*, Electronic Journal of Probability **13** (2008), 26–78.

[7] P. Diaconis, *The cutoff phenomenon in finite Markov chains*, Proc. Nat. Acad. Sci. U.S.A. **93** (1996), no. 4, 1659–1664.

[8] P. Diaconis and J. A. Fill, *Strong stationary times via a new form of duality*, Ann. Probab. **18** (1990), no. 4, 1483–1522.

[9] P. Diaconis and L. Miclo, *On times to quasi-stationarity for birth and death processes*. preprint.

[10] P. Diaconis and L. Saloff-Coste, *Separation cut-offs for birth and death chains*, Ann. Appl. Probab. **16** (2006), no. 4, 2098–2122.

[11] P. Diaconis and M. Shahshahani, *Generating a random permutation with random transpositions*, Z. Wahrsch. Verw. Gebiete **57** (1981), no. 2, 159–179.

[12] J. Ding, E. Lubetzky, and Y. Peres, *The mixing time evolution of Glauber dynamics for the Mean-field Ising Model*. preprint.

[13] J. A. Fill, *The passage time distribution for a birth-and-death chain: Strong stationary duality gives a first stochastic proof*. preprint.

[14] J. A. Fill, *On hitting times and fastest strong stationary times for skip-free chains*. preprint.

[15] P. R. Halmos, *Finite-dimensional vector spaces*, Springer-Verlag, New York, 1974.

[16] S. Karlin and J. McGregor, *Coincidence properties of birth and death processes*, Pacific J. Math. **9** (1959), 1109–1140.

[17] J. Keilson, *Markov chain models – rarity and exponentiality*, Applied Mathematical Sciences, vol. 28, Springer-Verlag, New York, 1979.

[18] S. Goldstein, *Maximal coupling*, Z. Wahrsch. Verw. Gebiete **46** (1978/79), no. 2, 193–204.

[19] D. Griffeath, *A maximal coupling for Markov chains*, Z. Wahrsch. Verw. Gebiete **31** (1975), 95–106.

[20] D. A. Levin, M. Luczak, and Y. Peres, *Glauber dynamics for the Mean-field Ising Model: cut-off, critical power law, and metastability*. to appear.

[21] D. A. Levin, Y. Peres, and E. Wilmer, *Markov Chains and Mixing Times*, 2007. In preparation.

[22] J. W. Pitman, *On coupling of Markov chains*, Z. Wahrsch. Verw. Gebiete **35** (1976), no. 4, 315–322.

[23] T. Lindvall, *Lectures on the coupling method*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1992. A Wiley-Interscience Publication.

[24] Y. Peres, American Institute of Mathematics (AIM) research workshop "Sharp Thresholds for Mixing Times" (Palo Alto, December 2004). Summary available at http://www.aimath.org/WWN/mixingtimes.
[25] L. Saloff-Coste, *Random walks on finite groups*, Probability on discrete structures, 2004, pp. 263–346.

JIAN DING
DEPARTMENT OF STATISTICS, UC BERKELEY, BERKELEY, CA 94720, USA.
E-mail address: jding@stat.berkeley.edu

EYAL LUBETZKY
MICROSOFT RESEARCH, ONE MICROSOFT WAY, REDMOND, WA 98052-6399, USA.
E-mail address: eyal@microsoft.com

YUVAL PERES
MICROSOFT RESEARCH, ONE MICROSOFT WAY, REDMOND, WA 98052-6399, USA.
E-mail address: peres@microsoft.com