Proceedings of Machine Learning Research vol 75:1–25, 2018 31st Annual Conference on Learning Theory

Testing Symmetric Markov Chains From a Single Trajectory

Constantinos Daskalakis[*] COSTIS@MIT.EDU
EECS, MIT

Nishanth Dikkala[†] NISHANTHD@CSAIL.MIT.EDU
EECS, MIT

Nick Gravin [‡] NIKOLAI@MAIL.SHUFE.EDU.CN
ITCS, SHUFE

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Classical distribution testing assumes access to i.i.d. samples from the distribution that is being tested. We initiate the study of Markov chain testing, assuming access to a *single trajectory of a Markov Chain.* In particular, we observe a single trajectory $X_0, \ldots, X_t, \ldots$ of an unknown, symmetric, and finite state Markov Chain $\mathcal{M}$. We do not control the starting state X_0, and we cannot restart the chain. Given our single trajectory, the goal is to test whether $\mathcal{M}$ is identical to a model Markov Chain $\mathcal{M}'$, or far from it under an appropriate notion of difference.

We propose a measure of difference between two Markov chains, motivated by the early work of Kazakos (1978), which captures the scaling behavior of the total variation distance between trajectories sampled from the Markov chains as the length of these trajectories grows. We provide efficient testers and information-theoretic lower bounds for testing identity of symmetric Markov chains under our proposed measure of difference, which are tight up to logarithmic factors if the hitting times of the model chain $\mathcal{M}'$ is $\tilde{O}(n)$ in the size of the state space n.

1. Introduction

We formulate theories about the laws that govern physical phenomena by making observations and testing them against our hypotheses. A common scenario is when our observations can be reasonably modeled as i.i.d. samples from a distribution that we are trying to understand. This is the setting tackled by most classical work in Statistics. Of course, having access to i.i.d. samples from a distribution is rare and quite commonly an approximation of reality. We typically only have access to approximate samples from a stationary distribution, sampled by a stochastic process whose description is unknown to us. For instance, the stochastic process might be a Markov chain whose transition matrix/kernel is unknown to us and which can only be observed for some finite time horizon. In fact, to the best of our knowledge, the underlying Markov chain may not even be rapidly mixing, so there is no guarantee that we will ever see samples that are approximately distributed according to the stationary distribution.

[*] Supported by a Microsoft Research Faculty Fellowship, and NSF Award CCF-1551875, CCF-1617730 and CCF-1650733.
[†] Supported by NSF Award CCF-1551875, CCF-1617730 and CCF-1650733.
[‡] Supported by NSF Award CCF-1551875, CCF-1617730 and CCF-1650733.

These issues are exacerbated in high-dimensional settings, e.g. when observing the configurations of a deck of cards where the state space consists of $52!$ permutations, or a weather system, where it may also be completely impractical to work with the high-dimensional stationary distribution itself. Moreover, several different processes may generate the same stationary distribution. For all these considerations, it may be both more interesting and more practical to understand the "mechanics" of the process that generates our observations, namely the transition matrix/kernel of the Markov chain whose evolution we get to observe.

Motivated by these considerations, in this paper we initiate the study of testing identity of Markov chains, and as a first step we focus on the case of finite and symmetric[1] Markov Chains. In our setting, we are given access to a *single* trajectory $X_0, X_1, \ldots, X_t, \ldots$ of some *unknown* symmetric Markov chain $\mathcal{M}$ over some finite state space $[n]$, and we want to test the identity of $\mathcal{M}$ to some *given* symmetric Markov chain $\mathcal{M}'$ over the same state space. Importantly, we do not get to control the distribution of the starting state X_0, and we can only observe a single trajectory of $\mathcal{M}$, i.e. *we cannot restart the Markov chain*. Such situations are plenty in nature. For instance, consider Markov models used to study the weather of a city, population growth of a species, the exchange rate of currencies, or the price of a stock where one cannot control the evolution of the chain and moreover cannot ask for restarts of the chain. What could we hope to achieve in such a situation?

If there is any difference in the transition matrices of $\mathcal{M}$ and $\mathcal{M}'$, one would think that we would *ultimately* be able to identify it by observing a sufficiently long trajectory. However, whether we can identify the difference or not depends on the connectivity properties of the chain. We can certainly identify the difference (*ultimately*) if the transition matrices of the two chains differ at a state that belongs to the essential communicating class (see Definition 2) of $\mathcal{M}$ where X_0 lies. However, it is, in general, not always necessary that one be able to observe such a difference. For instance, consider the following simple example.

The Two Communicating Classes Example: Suppose that $\mathcal{M}$ is a chain on states $\{1, 2, \ldots, 7\}$ whose transition matrix is the random walk matrix on a graph that is the disjoint union of a square on nodes $\{1, \ldots, 4\}$ and a triangle on nodes $\{5, 6, 7\}$, while $\mathcal{M}'$'s transition matrix is the random walk matrix on a graph that is the disjoint union of a clique on nodes $\{1, \ldots, 4\}$ and a triangle on nodes $\{5, 6, 7\}$. If our observed trajectory of $\mathcal{M}$ lies in the strong connected component defined by states $\{1, \ldots, 4\}$ (which forms an essential communicating class), we will easily identify its difference to $\mathcal{M}'$. On the other hand, if our observed trajectory of $\mathcal{M}$ lies in the essential communicating class defined by states $\{5, 6, 7\}$, we will not be able to identify that we are not observing a trajectory of $\mathcal{M}'$, no matter how long the trajectory is.

For some notion of difference, $\mathrm{Dist}\,(\mathcal{M}, \mathcal{M}')$, between Markov chains, we would like to quantify *how long* a trajectory $X_0, \ldots, X_\ell$ from an *unknown* chain, $\mathcal{M}$, we need to observe to be able to distinguish, with probability at least $1 - \delta$:

$$\mathcal{M} = \mathcal{M}' \;\text{ versus }\; \mathrm{Dist}\left(\mathcal{M}, \mathcal{M}'\right) > \epsilon, \tag{1}$$

for some given parameters $\delta \in (0, 1)$ and $\epsilon > 0$. Let us call this problem *single-sample goodness-of-fit (or identity) testing for Markov chains*. We will study it taking $\delta = 1/3$, with the understanding that this probability can be boosted to any small constant at the cost of a $O(\log(1/\delta))$-multiplicative factor in the length ℓ of the observed trajectory.

1. We also get a few observations for general asymmetric case that may be used as a foundation for future studies.

What notion of difference between Markov chains is the right one to use to study the afore-described goodness-of-fit testing problem? Here are some desiderata for such a notion of difference:

1. First, as our simple example above illustrates, under a worst-case starting state X_0, we may not be able to identify that $\mathcal{M} \neq \mathcal{M}'$ from a single trajectory. So, we would like to identify a notion of difference that takes a value $\mathrm{Dist}\left(\mathcal{M}, \mathcal{M}'\right) = 0$, whenever chains $\mathcal{M}$ and $\mathcal{M}'$ are indistinguishable from a single trajectory starting at a worst-case starting state.[2] Obviously, if the chains are irreducible, this constraint is immaterial.

2. Whenever $\mathcal{M}$ and $\mathcal{M}'$ *are* distinguishable from a single trajectory, whose starting state we do not get to control, i.e. from any starting state, we would like that our difference measure quantifies *how different* the chains are. Clearly, our notion of difference could not just be a combinatorial property of the connectivity of the state space of $\mathcal{M}$ and $\mathcal{M}'$, since the combinatorial structure won't reflect the magnitude of the differences in the chains.

One of our main contributions is to identify a meaningful measure of difference between Markov Chains capturing the above properties.

A Difference Measure Between Markov Chains. Total Variation (TV) is a standard distance between distributions used in the property/distribution testing literature. One reason for this is that it captures precisely our ability to distinguish two distributions p and q by observing a single sample from one of them.[3] Similarly, given two product measures $p^{\otimes \ell}$ and $q^{\otimes \ell}$, outputting a vector of ℓ i.i.d. samples drawn from p and q respectively, our ability to distinguish between them using a single sample is captured by $\mathrm{d}_{\mathrm{TV}}\left(p^{\otimes \ell}, q^{\otimes \ell}\right)$. Unfortunately, it is analytically difficult to relate $\mathrm{d}_{\mathrm{TV}}\left(p^{\otimes \ell}, q^{\otimes \ell}\right)$ to $\mathrm{d}_{\mathrm{TV}}\left(p, q\right)$ to study how our distinguishing ability improves with ℓ. For this reason, other distances are often employed when studying high-dimensional distributions. One such distance which will be of interest to us is the Hellinger distance $\mathrm{d}_{\mathrm{Hel}}\left(p^{\otimes \ell}, q^{\otimes \ell}\right)$.[4] Generalizing from product measures to Markov Chains, a natural notion of difference between two chains $\mathcal{M}$ and $\mathcal{M}'$ is the total variation distance, $\mathrm{d}_{\mathrm{TV}}\left(\mathcal{W}^{\ell}_{\mathcal{M}}, \mathcal{W}^{\ell}_{\mathcal{M}'}\right)$, between ℓ-step trajectories (a.k.a. *words*) $\mathcal{W}^{\ell}_{\mathcal{M}} \stackrel{\text{def}}{=} X_0 X_1 \cdots X_\ell$ and $\mathcal{W}^{\ell}_{\mathcal{M}'} \stackrel{\text{def}}{=} Y_0 Y_1 \cdots Y_\ell$ sampled from the two chains starting at some state $X_0 = s_0 = Y_0$. But due to the analytical difficulties presented by the TV distance for high-dimensional distributions we look towards the Hellinger distance as noted above. The usage of Hellinger square distance for capturing the difference between two high-dimensional distributions, for instance as was proposed in the early work of Kazakos (1978) and the more recent work of Daskalakis and Pan (2017) [5], is well known. Hence, we study the Hellinger distance $\mathrm{d}_{\mathrm{Hel}}\left(\mathcal{W}^{\ell}_{\mathcal{M}}, \mathcal{W}^{\ell}_{\mathcal{M}'}\right)$ between two trajectories, which satisfies a precise recurrence formula stated as Lemma 5 in Section 3. The relation between Hellinger and TV distances allows us to provide upper and lower bounds on the latter in terms of the former.

2. The worst-case starting state assumption is a choice also made when defining mixing time. It is also worth noting that in this scenario, since the chains are reducible they will not converge to the stationary distribution and hence the mixing time is infinite.
3. Formally, consider a guessing game where p or q is chosen uniformly at random (or by an adversary), then one sample is generated from the chosen distribution, and we must guess which one it is. The optimal error for this guessing game is precisely $0.5 \cdot (1 - \mathrm{d}_{\mathrm{TV}}\left(p, q\right))$.
4. Indeed, it enjoys the precise recurrence relation $1 - \mathrm{d}^2_{\mathrm{Hel}}\left(p^{\otimes \ell}, q^{\otimes \ell}\right) = \left[1 - \mathrm{d}^2_{\mathrm{Hel}}\left(p, q\right)\right]^{\ell}$. Moreover, there is a tight relationship between TV and Hellinger distances, see (4), so one can derive upper and a lower bound on $\mathrm{d}_{\mathrm{TV}}\left(p^{\otimes \ell}, q^{\otimes \ell}\right)$ based on $\mathrm{d}_{\mathrm{Hel}}\left(p^{\otimes \ell}, q^{\otimes \ell}\right)$. See Section 2.
5. For more discussion on this, see the related work section.

A Scale-Free Measure of Difference Between Markov Chains. Both the distance measures $\mathrm{d}_{\mathrm{TV}}\left(\mathcal{W}_{\mathcal{M}}^{\ell}, \mathcal{W}_{\mathcal{M}'}^{\ell}\right)$ and $\mathrm{d}_{\mathrm{Hel}}\left(\mathcal{W}_{\mathcal{M}}^{\ell}, \mathcal{W}_{\mathcal{M}'}^{\ell}\right)$ depend on (1): the length ℓ of the trajectory and (2): the starting state s_0. We would like, instead, a parameter-free and scale-free notion of difference between Markov Chains satisfying the above desiderata. A popular way of tackling such a parameter dependency in Markov Chain literature is to study the inverse dependency of the length ℓ of a trajectory required to achieve a certain threshold value for some quantity, e.g. mixing time is defined as the minimum number of steps ℓ needed so that the distribution of the ℓ-th state of a trajectory starting at any state s_0 is no more than $1/4$ away from the stationary distribution. Similarly, in our case, we propose to analyze the minimum number of steps ℓ required so that $\mathrm{d}_{\mathrm{Hel}}\left(\mathcal{W}_{\mathcal{M}}^{\ell}, \mathcal{W}_{\mathcal{M}'}^{\ell}\right)$ is at least some constant (we choose 0.5):[6]

$$\min_{\ell>0} \quad \ell: \qquad \forall s_0 \in [n] \qquad \mathrm{d}_{\mathrm{Hel}}\left(\mathcal{W}_{\mathcal{M}}^{\ell}, \mathcal{W}_{\mathcal{M}'}^{\ell} \mid X_0 = Y_0 = s_0\right) \geq \delta. \tag{2}$$

The above definition assumes a worst-case starting state s_0 which reflects our desiderata stated above that we do not get to control the starting state and we cannot restart the chain. Moreover, it is the choice made in the definition of mixing time. In Section 3 we show a tight relationship between the above definition and an appropriate "average-case" version.

Clearly, the answer to (2) depends on the *scaling behavior*, as $\ell \to \infty$, of the following quantity:

$$\delta(\ell) \stackrel{\text{def}}{=} \min_{s_0} \mathrm{d}_{\mathrm{Hel}}\left(\mathcal{W}_{\mathcal{M}}^{\ell}, \mathcal{W}_{\mathcal{M}'}^{\ell} \mid X_0 = Y_0 = s_0\right). \tag{3}$$

Interestingly, as we discuss in Section 2, this scaling behavior is tightly captured by the following matrix:

$$[P, Q]_{\surd} \stackrel{\text{def}}{=} \left[\sqrt{P_{ij} \cdot Q_{ij}}\right]_{ij \in [n \times n]},$$

where P and Q are the transition matrices of the two chains, i.e. P_{ij} and Q_{ij} denote the probabilities of transitioning from state i to state j in the two chains. In Lemma 5, we state a recursive decomposition that allows us to exactly express the square Hellinger similarity, $1 - \mathrm{d}_{\mathrm{Hel}}^2\left(\mathcal{W}_{\mathcal{M}}^{\ell}, \mathcal{W}_{\mathcal{M}'}^{\ell}\right)$ of ℓ-length words sampled from the two chains in terms of the ℓ-th power of the above matrix, and the distribution of the starting states X_0 and Y_0 in the two words.

To identify a word-length independent measure of difference between the two chains based on (2), we employ a spectral approach. We show that the scaling behavior (w.r.t. ℓ) of the Hellinger square distance between $\mathcal{W}_{\mathcal{M}}^{\ell}$ and $\mathcal{W}_{\mathcal{M}'}^{\ell}$ is captured by the largest eigenvalue $\lambda_1 = \rho([P, Q]_{\surd})$ of matrix $[P, Q]_{\surd}$. We show that always $\lambda_1 \leq 1$ (Claim 1), and that $\lambda_1 = 1$ if and only if the two chains have an identical essential communicating class (Claim 1), in which case we would be unable to identify the difference between the two chains from a single trajectory which starts at a state in the essential communicating class which is identical in the two chains (see the two communicating classes example above). These statements hold even for asymmetric chains. For *symmetric* Markov chains, ℓ in (2) is almost proportional to $\frac{1}{\varepsilon}$ ($\ell = \widetilde{\Theta}\left(\frac{1}{\varepsilon}\right)$ up to a $\log n$ factor, see Claim 2) where $\varepsilon = 1 - \rho([P, Q]_{\surd})$.[7]. The latter estimation on ℓ also holds for the case when initial state in P and Q is chosen uniformly at random.

6. Note that a trajectory of this length also satisfies $\mathrm{d}_{\mathrm{TV}}\left(\mathcal{W}_{\mathcal{M}}^{\ell}, \mathcal{W}_{\mathcal{M}'}^{\ell}\right) \geq 0.25$.

7. For non symmetric Markov chains, one can show that the slowest (with respect to the choice of the starting state) that the square Hellinger similarity (defined as $1 - \mathrm{d}_{\mathrm{Hel}}^2$) of the two chains can drop as a function of the length ℓ is λ_1^{ℓ}, up

Given these properties, we propose the use of

$$\mathrm{Dist}\left(\mathcal{M}, \mathcal{M}'\right) = 1 - \rho([P, Q]_{\surd})$$

as a scale-free and meaningful measure of difference between Markov chains. Figure 1 illustrates how $\mathrm{Dist}\left(\mathcal{M}, \mathcal{M}'\right)$ behaves for different pairs of Markov chains $\mathcal{M}$ and $\mathcal{M}'$.

Our Results. Using our proposed measure of difference between Markov chains we provide algorithms for goodness-of-fit testing of Markov chains, namely Problem (1), where $\mathrm{Dist}\left(\mathcal{M}, \mathcal{M}'\right) = 1-\rho([P, Q]_{\surd})$, where P and Q are the transition matrices of chains $\mathcal{M}$ and $\mathcal{M}'$. We study this problem when $\mathcal{M}$ and $\mathcal{M}'$ are both *symmetric*, and provide upper and lower bounds for the minimum length ℓ of a trajectory from the unknown chain $\mathcal{M}$ that is needed to determine the correct answer with probability at least $2/3$. In particular, Theorems 9 and 10 combined show that the length of the required trajectory from $\mathcal{M}$ to answer Problem (1) is n/ε, where n is the size of the state space, up to logarithmic factors and an additive term that does not depend on ε or $\mathcal{M}$. Our upper bound is established via an information-efficient reduction from single-sample identity testing for Markov chains with n states to the classical problem of identity testing of distributions over $O(n^2)$ elements, from i.i.d. samples. A naive attempt to obtain such a reduction is to look at every $\mathrm{MixT}_{\mathcal{M}'}$-th step of the trajectory of $\mathcal{M}$, where $\mathrm{MixT}_{\mathcal{M}'}$ is the mixing time of chain $\mathcal{M}'$, pretending that these transitions are i.i.d. samples from the distribution $\{\frac{1}{n}P_{ij}\}_{ij\in[n^2]}$. This incurs an unnecessary blow-up of a factor of $\mathrm{MixT}_{\mathcal{M}'}$ in the required length of the observed trajectory and also requires some additional work of checking the mixing time of the the unknown Markov chain $\mathcal{M}'$. On the other hand, we cannot simply wait while we collect a predetermined small number of samples per every row of the transition matrix and treat them as i.i.d. samples. Indeed, the fact that certain states are visited can create dependencies among transitions from the other states[8]. We show how to avoid these issues via a more subtle approach, which also exchanges the mutliplicative dependence on the mixing time of $\mathcal{M}'$ with an additive term that is nearly-linear in the hitting time of $\mathcal{M}'$.

Related Work. Testing goodness-of-fit for distributions has a long history in Statistics; for some old and more recent references see, e.g., Pearson (1900); Fisher (1935); Rao and Scott (1981); Agresti (2012). In this literature the emphasis has been on the asymptotic analysis of tests, pinning down their error exponents as the number of samples tends to infinity Agresti (2012); Tan et al. (2010). In the last two decades or so, distribution testing has also piqued the interest of theoretical computer scientists Batu et al. (2001); Paninski (2008); Levi et al. (2013); Valiant and Valiant (2014); Chan et al. (2014); Acharya et al. (2015); Canonne et al. (2016); Diakonikolas and Kane (2016); Daskalakis et al. (2013); Canonne et al. (2014); Rubinfeld (2012); Goldreich (2011); Canonne (2015), where the emphasis, in contrast, has been on minimizing the number of samples required to test hypotheses with a strong control for both type I and type II errors. A few recent

to factors that do not depend on ℓ; this follows from (5) and (7). That is, the slowest that the square Hellinger distance of the two chains can increase is $1 - O(\lambda_1^{\ell})$. However, the dependency on the starting state is more significant than in the symmetric case, and the dependency in the worst-case may be not as smooth as for the symmetric $\mathcal{M}$ and $\mathcal{M}'$. (See Figure 1 for examples of irregular behavior of certain non-symmetric MC.)

8. Consider for example a symmetric Markov chain $\mathcal{M}$ with two cliques of size $\frac{n}{2}$ connected by a single edge $A - B$ (the transition probability of the edge is $\frac{2}{n}$). The expected cover time of $\mathcal{M}$ is $\Theta(n^2)$, however we can get lucky and finish the pass over all states early in, say, $n\sqrt{n}$ steps (then it is also likely that every state was visited $\Omega(\sqrt{n})$ times). In this case, we know that the bridge $A - B$ was necessarily used, which is actually unlikely event if we make $2\sqrt{n}$ i.i.d. transition from each of A and B states.

works have identified tight upper and lower bounds on the sample complexities of various testing problems Paninski (2008); Valiant and Valiant (2014); Acharya et al. (2015); Diakonikolas and Kane (2016). All of the papers in this vast body of literature assume access to i.i.d. samples from the underlying distribution.

Some work in Statistics has considered the problem of testing with dependent samples. For instance, Bartlett (1951); Moore et al. (1982); Gleser et al. (1983); Molina et al. (2002) and the references therein study goodness-of-fit testing under Markov dependences. These works study how the classical tests used to perform goodness-of-fit testing with independent samples, perform when there are Markovian dependencies among the samples. Tavare and Altham (1983) and more recently Barsotti et al. (2016) study the problem of testing the stationary distribution of Markov chains. Kazakos (1978) studies the problem of asymptotically perfect detection (APD) between two Markov chains. All these works focus on the asymptotic regime where the length of the observed trajectories tends to infinity, and study the conditions under which hypothesis testing can be performed successfully or focus on pinning down the error exponents. In the computer science literature, Batu et al. (2013) considered the problem of testing whether a Markov chain is fast mixing or not. They defined a notion of closeness between two random walks starting at different states of the *same* chain, which is different in spirit to the distance notion we define in this work. In particular, their distance is based on the L_1 norm of the state distributions attained by starting at two different states u and v and running the chain for t steps. This ignores any differences in trajectory seen along the way and it is apt for their setting as they focus on mixing time which is a trajectory independent property of the chain. Moreover, they assume the chain can be restarted any number of times making it fundamentally different from our setting.

There is a large body of statistical literature on estimating properties and parameters of Markov chains. Mixing time is one such important and well studied parameter (see, e.g., Hsu et al. (2015) and the references therein), as it is useful in designing MCMC algorithms. The question of mixing time estimation is related to but different than the goodness-of-fit kernel testing that we perform here.

Our notion of distance is derived from Lemma 5 but the work of Kazakos (1978) or more recent works of Daskalakis and Pan (2017) or Diakonikolas et al. (2016) don't consider the spectral behavior as we do in this work.

Organization We start in Section 2 with a description of the notational conventions we use and provide all necessary formal details for our difference measure in Section 3. In Section 4, we study the problem of testing identity of symmetric Markov chains and present our tester. We give a sample complexity lower bound for this problem in Section 5.

2. Preliminaries

We list the general notational conventions used in this paper. We denote vectors by small letters such as $\boldsymbol{v}$ and matrices by capital letters such as A,B,P,Q. The i^{th} entry of vector $\boldsymbol{v}$ is denoted by v_i or $v[i]$ and the $(ij)^{th}$ entry of matrix A (i^{th} row, j^{th} column) is denoted by A_{ij} or $A[ij]$; $\boldsymbol{e}_i$ denotes the standard basis vector with 1 in its i^{th} coordinate and 0 elsewhere; $\mathbb{1}$ denotes the vector of all ones. The "entrywise" L_1 and L_2 norms of a matrix A are respectively denoted as $\|A\|_1 = \sum_{i,j} |A_{ij}|$ and $\|A\|_2 = \sqrt{\sum_{i,j} A_{ij}^2}$; $\rho(A)$ denotes the spectral radius of matrix A, i.e., the maximum absolute eigenvalue of A. The eigenvalues of A are denoted by $\lambda_1, \ldots, \lambda_i, \ldots, \lambda_n$ and the respective right

eigenvectors by $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i, \ldots, \boldsymbol{v}_n$ (left eigenvectors by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$)[9]; for symmetric matrix A we assume that $\lambda_1 \geq \cdots \geq \lambda_i \geq \cdots \geq \lambda_n$.

Two popular notions of distance between distributions will be used heavily in this paper. We state their formal definitions below and also specify the relation between them.

Definition 1 *The* total variation *and* Hellinger *distances between distributions p, q over $[n]$ are defined as :* $\mathrm{d}_{\mathrm{TV}}(p,q) \stackrel{def}{=} \frac{1}{2}\sum_{i\in[n]} |p_i - q_i|$; $\mathrm{d}^2_{\mathrm{Hel}}(p,q) \stackrel{def}{=} \frac{1}{2}\sum_{i\in[n]} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 = 1 - \sum_{i\in[n]} \sqrt{p_i \cdot q_i}$; *The following relation between these notions of distance is well known (see, e.g., Gibbs and Su (2002)):*

$$\sqrt{2} \cdot \mathrm{d}_{\mathrm{Hel}}(p,q) \geq \mathrm{d}_{\mathrm{TV}}(p,q) \geq \mathrm{d}^2_{\mathrm{Hel}}(p,q). \tag{4}$$

2.1. Markov Chains

A *discrete-time* Markov chain is a stochastic process $\{X_t\}_{t\in\{0,1,\ldots\}}$ over a state space S which satisfies the Markov property: the probability of being in state s at time $t+1$ depends only on the state at previous time t. In this paper, we only consider Markov chains with the *finite state space* $[n]$. Such Markov chains can be completely specified by a $n \times n$ transition matrix (*kernel*) that contains probabilities of transitioning from state i to state j in the i^{th} row and j^{th} column, and a description of the distribution of their starting state. The transition matrix has non-negative entries and is a stochastic matrix. We use capital letters P, Q, M to represent Markov chains as well as their respective transition matrices. The stationary distribution π of a Markov chain P is a distribution over the state space S such that it satisfies $\pi^\top \cdot P = \pi^\top$. Another important parameter is the distribution of the starting state s_0 which we denote by $\boldsymbol{p}$ (for the Markov chain P). It may or not may not be the stationary distribution.
The state space of a Markov chain can be partitioned into communicating classes which are groups of states reachable from each other with positive probability. The formal definition of essential communicating classes is as follows.

Definition 2 (Essential Communicating Classes) *Given a Markov chain M over the state space $[n]$, we define $x \to y$ if there exists an integer $r > 0$ such that $M^r(x,y) > 0$. Similarly, we define equivalence relation $x \leftrightarrow y$ iff $x \to y$ and $y \to x$. The equivalence classes under relation $\leftrightarrow$ are called communicating classes. Any communicating class C with the property that y must be in C for any $x \in C$ and $x \to y$ is said to be an* essential *communicating class*[10].

2.1.1. HITTING TIMES AND MIXING TIMES

Two commonly studied random variables associated with Markov chains which are relevant to this paper are their mixing times and hitting times.

9. If matrix A is not symmetric, we allow $\lambda_i \in \mathbb{C}$ and $\boldsymbol{v}_i, \boldsymbol{u}_i \in \mathbb{C}^n$. Then, we will only use $\lambda_1 \in \mathbb{R}$ and $\boldsymbol{v}_1, \boldsymbol{u}_1 \in \mathbb{R}^n$.
10. An essential communicating class can be intuitively thought of as a strong connected component of the underlying directed graph with no outgoing edges.

Definition 3 (Hitting Time HitT_P of a Markov chain P) *Given a Markov chain P over a state space $[n]$, let s_t denote the state at time t. The hitting time HitT_P is*

$$\mathrm{HitT}_P = \max_{r,s\in[n]} \{\mathbb{E}\left[\min\{t \geq 0 : s_t = r \text{ given } s_0 = s\}\right]\}$$

Definition 4 (Mixing Time MixT_P of a Markov chain P) *Given a Markov chain P with a stationary distribution π and a starting state distribution $\boldsymbol{p}$,*

$$\mathrm{MixT}_P = \max_{\boldsymbol{p}} \ \min\{t \geq 0 : \left\|P^t\boldsymbol{p} - \pi\right\|_1 \leq 1/4\}$$

3. Deriving a Notion of Distance between Markov Chains

Given two Markov chains P and Q, we want to come up with a distance notion which captures how easy it is to distinguish which Markov chain P or Q a word $w = s_0 \to s_1 \cdots \to s_\ell$ of certain length ℓ was generated from (while being agnostic to the distribution of s_0). This distinguishability is precisely captured by the TV distance $\mathrm{d}_{\mathrm{TV}}\left(\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell\right)$ between *word distributions* $\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell$ for words of length ℓ generated by Markov chains P and Q respectively. It is more convenient in our setting to use, instead of total variation distance, the square of the Hellinger distance $\mathrm{d}^2_{\mathrm{Hel}}\left(\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell\right)$ or the closely related Bhattacharya coefficient[11], which is useful for studying divergence of non-stationary and continuous Markov chains as was observed in Kazakos (1978). Kazakos (1978) establishes nice recurrence relations for the Bhattacharya coefficient of two word distributions, which is captured by the matrix $[P,Q]_{\surd} \stackrel{\text{def}}{=} \left[\sqrt{P_{ij} \cdot Q_{ij}}\right]_{i,j\in[n\times n]}$.

Lemma 5 (Kazakos (1978)) *Suppose P and Q are Markov Chains over states $[n]$, $\boldsymbol{p}$ and $\boldsymbol{q}$ are probability distributions of the initial state. Let $\mathcal{W}_P^\ell$, $\mathcal{W}_Q^\ell$ be the distributions denoting a length ℓ trajectory of Markov Chains P (resp. Q) starting at a random node s_0 sampled from $\boldsymbol{p}$ (resp. $\boldsymbol{q}$). Moreover, define the vector $[\boldsymbol{p},\boldsymbol{q}]_{\surd} \stackrel{\text{def}}{=} \left[\sqrt{p_s \cdot q_s}\right]_{s\in[n]}$ and the matrix $[P,Q]_{\surd} \stackrel{\text{def}}{=} \left[\sqrt{P_{ij} \cdot Q_{ij}}\right]_{i,j\in[n\times n]}$. Then:*

$$1 - \mathrm{d}^2_{\mathrm{Hel}}\left(\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell\right) = [\boldsymbol{p},\boldsymbol{q}]_{\surd}^\top \cdot \left([P,Q]_{\surd}\right)^\ell \cdot \mathbb{1}, \tag{5}$$

There are two important parameters which affect the expression given by Kazakos (1978). The first is the distributions of the starting states of the Markov chains $(\boldsymbol{p},\boldsymbol{q})$ and the second is the length of the word (l). We want a notion of distance which is a scale-free non-negative real number. To achieve this, we study next how to eliminate the dependencies on the starting state distributions $(\boldsymbol{p},\boldsymbol{q})$ and the word length (l).

Assumption on the starting state. We study two scenarios for the choice of the starting state: (i) a **worst-case** scenario where both P and Q begin from the same state i chosen in adversarial manner to make P and Q look as much alike as possible; (ii) an **average-case** scenario, where the initial distributions $\boldsymbol{p} = \boldsymbol{q}$ for P and Q either are given to us, or are related to P and Q in some

11. Hellinger distance is tightly related to the Bhattacharya coefficient between two distributions which is defined as $BC(p,q) = \sum_{i\in[k]} \sqrt{p_i \cdot q_i}$. It captures similarity of two distributions and lies in $[0,1]$.

natural way[12]. Given the assumption on the starting state we want to answer the question of what ℓ to pick, so that $\mathcal{W}_P^\ell$ and $\mathcal{W}_Q^\ell$ are far apart in squared Hellinger distance (say ≥ 0.5). Formally, we have the following respectively for the worst-case and average-case scenarios listed above:

$$\min_{\ell>0} \quad \ell: \qquad \forall i \in [n] \qquad 0.5 \geq 1 - \mathrm{d}_{\mathrm{Hel}}^2\left(\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell\right) = \boldsymbol{e}_i{}^\top \cdot \left([P,Q]_{\surd}\right)^\ell \cdot \mathbb{1}. \tag{6}$$

$$\min_{\ell>0} \quad \ell: \qquad\qquad 0.5 \geq 1 - \mathrm{d}_{\mathrm{Hel}}^2\left(\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell\right) = [\boldsymbol{p},\boldsymbol{q}]_{\surd}^\top \cdot \left([P,Q]_{\surd}\right)^\ell \cdot \mathbb{1}$$

Due to the relation between Hellinger and total variation distances, an inequality similar to (6) holds for $1 - \mathrm{d}_{\mathrm{TV}}\left(\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell\right)$ as well but with a different constant on the left.

We call the minimal ℓ that satisfies $\mathrm{d}_{\mathrm{TV}}\left(\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell\right) \geq \frac{2}{3}$ for all starting states $i \in [n]$ (or for fixed starting distributions $\boldsymbol{p} = \boldsymbol{q}$) the *minimal distinguishing length*. We note that (6) gives us an estimate on ℓ up to a constant factor.

Next we argue that when ℓ is large, the behavior of the RHS of (6) is governed by *the largest eigenvalue* $\lambda_1 = \rho\left([P,Q]_{\surd}\right)$ of $[P,Q]_{\surd}$. In particular, by Perron-Frobenius theorem, we have that the largest eigenvalue of $[P,Q]_{\surd}$ is non-negative and the corresponding left eigenvector $\boldsymbol{u}_1$: $\boldsymbol{u}_1^\top \cdot [P,Q]_{\surd} = \lambda_1 \cdot \boldsymbol{u}_1^\top$ has non-negative coordinates. In particular, if we choose initial distributions $\boldsymbol{p} = \boldsymbol{q}$ proportional to $\boldsymbol{u}_1$, then

$$\boldsymbol{p}^\top \cdot \left([P,Q]_{\surd}\right)^\ell \cdot \mathbb{1} = \lambda_1^\ell \cdot \langle \boldsymbol{p}, \mathbb{1}\rangle = \lambda_1^\ell. \tag{7}$$

Claim 1 *It is always true that* $\lambda_1 = \rho\left([P,Q]_{\surd}\right) \leq 1$. *Moreover,* $\lambda_1 = 1$ *iff* P *and* Q *have an identical essential communicating class.*

Proof of Claim 1 is defered to Appendix B.

We propose the use of the quantity $1 - \rho\left([P,Q]_{\surd}\right)$ as a distance measure between Markov chains P and Q.

Definition: $\mathrm{Dist}\,(P,Q) \stackrel{\text{def}}{=} 1 - \rho\left([P,Q]_{\surd}\right)$.

In particular in (6) if $\boldsymbol{p} = \boldsymbol{q}$ is proportional to $\boldsymbol{u}_1$, then $\ell \cdot \ln(1-\varepsilon) \leq \ln 0.5 \implies \ell \geq \frac{\ln 2}{2\varepsilon}$. This shows that in the worst-case we need to observe a trajectory of length at least $\Omega(1/\varepsilon)$ before we can satisfactorily distinguish the two chains. Note however that, in general, ℓ might need to be larger than $\Omega(\frac{1}{\varepsilon})$ as is illustrated in Example 2. However, we will see that in the case of symmetric Markov chains we observe a more regular behavior. In the remainder of this section and the following sections we only consider symmetric Markov chains that avoid such irregular behavior and dependency on the starting state.

12. For example $\boldsymbol{p}$ and $\boldsymbol{q}$ could be respective stationary distributions of P and Q. However, we still assume identical initial distributions for P and Q, i.e. $\boldsymbol{p} = \boldsymbol{q}$, as otherwise there might be a simpler trivial strategy to distinguish P and Q by observing only one initial sample from $\boldsymbol{p}$. Example 3 illustrates how two Markov chains can produce very similar distributions of words $\mathcal{W}_P^\ell, \mathcal{W}_Q^\ell$ starting from any state for some large ℓ, and yet have vastly different stationary distributions.

Word distance between Symmetric Markov Chains. The stationary distribution for any symmetric Markov chain is the uniform distribution over all states. In this case the most natural starting distributions for the average-case part of equation (6) are $\boldsymbol{p} = \boldsymbol{q} = \frac{1}{n}\mathbb{1}$. In this setting of symmetric Markov chains, we can provide sharp bounds on the minimal distinguishing length ℓ.

Claim 2 *The necessary and sufficient distinguishing length ℓ, which allows to distinguish P vs. Q with high probability, is $\widetilde{\Theta}\left(\frac{1}{\varepsilon}\right)$ (up to a $\log n$ factor), where $\varepsilon = 1 - \rho\left([P,Q]_{\surd}\right)$ under both worst-case and average-case (we assume $\boldsymbol{p} = \boldsymbol{q} = \frac{1}{n}\mathbb{1}$) scenarios for the starting state.*

Proof of Claim 2 is given in Appendix B.

We note that, if one could pick the starting state instead of working with average-case or worst-case assumptions of Claim 2, then ℓ can be much smaller (see Example 5). Claim 2 gives a strong evidence that $1 - \rho\left([P,Q]_{\surd}\right)$ is a meaningful and important parameter that captures closeness between P and Q. In the following section we will use it as analytical proxy for the distance between Markov Chains[13].

4. Identity Testing of Symmetric Markov Chains

After understanding the problem of distinguishing between two given distributions, a next fundamental question is the identity testing problem where the goal is to test whether an unknown distribution p from which we see a stream of samples, coincides with a given hypothesis distribution q. In this section, we study identity testing of symmetric Markov chains and provide an efficient algorithm (Theorem 9). We begin by giving below a formal statement of the problem:

Input: $\varepsilon > 0$; explicit description of a symmetric Markov chain Q; a trajectory $s_1 \cdots s_m$ of length m from a symmetric Markov Chain P.
Output: $P = Q$, or $P \neq Q$ if $1 - \rho\left([P,Q]_{\surd}\right) > \varepsilon$.

Our approach. Identity testing problem with i.i.d. samples, is a well studied problem in the distribution testing literature. The problem is quite non trivial[14] and to achieve tight sample complexity one needs to do careful estimations of collisions in observed samples. Markov chain identity testing appears to be at least as hard as the i.i.d. identity testing problem with the added complication of dependent samples. To avoid involved analysis of collisions among dependent samples we will instead try to find a black-box reduction of the MC testing problem to identity testing with i.i.d. samples. A naive attempt at such a reduction proceeds by waiting for a period of mixing time MixT_Q of the known Markov Chain Q to get one (potential) i.i.d. sample from the stationary distribution of P (in case P has mixed). If the empirical distribution for the number of visits is far from the uniform

13. In general this notion of distance should be used with care. For instance, note that $\mathrm{Dist}\,(P,Q) = 1 - \rho\left([P,Q]_{\surd}\right)$, is not a metric. In particular, $\mathrm{Dist}\,(P,Q)$ violates the triangle inequality ($\mathrm{Dist}\,(M_1,M_2) = \mathrm{Dist}\,(M_2,M_3) = 0$, but $\mathrm{Dist}\,(M_1,M_3) > 0$ for some M_1, M_2, M_3) as is illustrated by Example 1. We note that this problem can only appear for reducible chains, as is shown in Claim 1. Also it is not always possible to extend the sharp bounds on ℓ of Claim 2 from symmetric Markov chains to non-symmetric Markov chains, even if both MC have the uniform distribution as their stationary distribution (see Example 4)
14. It is studied by a number of works. For instance, see Batu et al. (2001); Paninski (2008); Valiant and Valiant (2014); Acharya et al. (2015); Diakonikolas and Kane (2016) (This is not an exhaustive list).

distribution, we can immediately reject P (since if $P = Q$, then P is a symmetric chain and will have the uniform distribution as the stationary) and if it is not, then we would have attained multiple transitions from a sizeable set of nodes and one could hope they contain sufficient signal to distinguish P from Q. It is non-trivial to extract this signal as the mere fact that we have seen multiple samples from a single node within a short length of the trajectory introduces dependencies in our samples. That is, two samples from the same node are not independent samples from the transition distribution of that node, if it took only a little time to return to this node. Moreover, this attempt, if it works, will incur a *multiplicative loss* of MixT_Q in the sample complexity.

We take a more subtle and involved approach to achieve a successful reduction to the classical setting with i.i.d. samples. Moreover, our reduction yields an algorithm that suffers only an *additive loss* of $\widetilde{O}\left(\mathrm{HitT}_Q \cdot \log\left(\mathrm{HitT}_Q\right)\right)$ in sample complexity. We reduce the Markov chain problem to the classical identity testing problem with respect to squared Hellinger distance of distributions supported on a domain of size n^2. Our result is always as good as the naive approach. Indeed, for symmetric chains, the hitting time cannot be larger than mixing time by more than a $c.n$ factor (where c is a constant), but usually it is much smaller (in fact hitting time can be even smaller than mixing time). We note that many broad classes of graphs and Markov chains have close to linear hitting times, e.g., expanders, d-dimensional grids (which are not expanders). Below we describe how we map samples from a Markov chain to i.i.d. samples from the appropriate distribution.

A Mapping From Infinite Words. Consider a mapping $\mathcal{K}_{\boldsymbol{k}}$ from words of infinite length $w \in W_M^\infty$ of an irreducible Markov chain M on the state space $[n]$ to $\prod_{i=1}^n [n]^{k_i}$, where $\boldsymbol{k} = (k_1, \cdots, k_n)$ is a vector of n non negative integers, as follows. For each infinite word $w = s_1 s_2 \cdots$ and each state $i \in [n]$ we look at the first k_i visits to state i (i.e., at times $t = t_1, \ldots, t_{k_i}$ with $s_t = i$) and write down the corresponding transitions in w, i.e., s_{t+1}. We note that every state is visited almost surely in w, since M is an irreducible finite-state Markov chain. Therefore, mapping $\mathcal{K}_{\boldsymbol{k}}$ defines a probability distribution on $\prod_{i=1}^n [n]^{k_i}$. Now, crucially, this distribution is independent across all different states and/or independent for a particular state i because of the Markov property of Markov chains. Furthermore, a specific transition from a copy of the state i is distributed according to the i-th row of the transition matrix M.

In Lemma 6, we show that even for a finite length trajectory with length $m = \widetilde{O}\left(\mathrm{HitT}_Q \log\left(\mathrm{HitT}_Q\right) + \frac{n}{\varepsilon}\right)$ [15] and $k_i = O(\mathbf{E}[\#\text{ visits to } i]) = O(\frac{m}{n})$ the mapping $\mathcal{K}_{\boldsymbol{k}}$ is well defined for all but a small fraction (in probability) of the words from the distribution $\mathcal{W}_M^m$. This effectively allows us, with high probability, to generate a large number of independent samples from the following distribution supported over $[n] \times [n]$: pick uniformly at random a state $i \in [n]$ and then observe a transition from i according to transition probabilities of row P_i. Indeed, to this end, we first simulate $m' = \Theta\left(\frac{m}{\log^2(n/\varepsilon)}\right)$ i.i.d. samples from $[n]$. These samples describe how many visits an independent sampler would make to state $i \in [n]$. Let $\boldsymbol{k}$ be the histogram of these m' samples (note that $\max_i k_i \leq O(m' \log n/n)$ with high probability). We apply $\mathcal{K}_{\boldsymbol{k}}$ mapping to our stream of m consecutive samples of Markov chain P, which is well defined with high probability. Apart from some small probability events ($\max_i k_i$ is too large, or $\mathcal{K}_{\boldsymbol{k}}$ is not defined for our choice of m) we obtain the desired m' i.i.d. samples.

15. in this paper, $\widetilde{O}$ always hides poly $\log(n/\varepsilon)$ factors, but not HitT_Q.

Lemma 6 *Given an irreducible Markov chain M and the mapping from infinite words $\mathcal{W}_M^\infty$ described above, for $m = \widetilde{O}\left(\log\left(\mathrm{HitT}_Q\right)\mathrm{HitT}_Q\right)$, then $\mathbf{Pr}[\exists$ state i s.t. $|\{t : i = s_t \in w\}| < \frac{m}{8e\cdot n}] \leq \frac{\varepsilon^2}{n}$ where the probability is over the sampling of $\boldsymbol{k}$ and word w.*

The proof is deferred to Appendix C.

In the following we present our algorithm for Markov Chain identity testing and provide an upper bound on its sampling complexity.

Algorithm 1: Independent Edges Sampler.

$\boldsymbol{k} \leftarrow$ Histogram $\left(\Theta\left(\frac{m}{\log^2(n/\varepsilon)}\right)\right.$ i.i.d. `Uniform` [n] samples) **for** $t \leftarrow 1$ **to** $m-1$ **do**
  if $|\mathsf{Samples}[s_t]| < \boldsymbol{k}[s_t]$ **then** Add $(s_t \to s_{t+1})$ to $\mathsf{Samples}[s_t]$;
end
if $\exists i$, *s.t.,* $|\mathsf{Samples}[i]| < \boldsymbol{k}[i]$ **then**
  return REJECT;
else
  $\mathsf{Samples} \leftarrow \mathsf{Samples}[1] \cup \cdots \cup \mathsf{Samples}[n]$ **return** `IdentityTestIID` $(\varepsilon, \{q_{ij} = \frac{1}{n} \cdot Q_{ij}\}_{i,j\in[n]}, \mathsf{Samples})$
end

Algorithm 1 uses as a black-box the tester of Algorithm 1 of Daskalakis et al. (2017). The following Lemma follows from Theorem 1 of Daskalakis et al. (2017).

Lemma 7 *Given a discrete distribution q supported on $[n]$ and access to i.i.d. samples from a discrete distribution p on the same support, there is a tester which can distinguish whether $p = q$ or $\mathrm{d}_{\mathrm{Hel}}(p,q) \geq \varepsilon$ with probability $\geq 2/3$ using $O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples.*

As a corollary of Lemma 7, we get a test that can distinguish whether $P = Q$, or $\mathrm{d}^2_{\mathrm{Hel}}\left(\frac{1}{n}P, \frac{1}{n}Q\right) \geq \varepsilon$ using $m = O\left(\frac{n}{\varepsilon}\right)$ i.i.d samples from $\frac{1}{n}P$, which can be viewed as a distribution on a support of size n^2. Lemma 8 shows that the required distance condition for the i.i.d. sampler is implied by our input guarantee.

Lemma 8 *Consider two symmetric Markov chains P and Q on a finite state space $[n]$. Denote by $\frac{1}{n}P$ the distribution over n^2 elements obtained by scaling down every entry of the transition matrix P by a factor $1/n$. We have,*

$$\frac{1}{2}\sum_{i,j\in[n]}\left(\sqrt{\frac{P_{ij}}{n}} - \sqrt{\frac{Q_{ij}}{n}}\right)^2 = \mathrm{d}^2_{\mathrm{Hel}}\left(\frac{1}{n}P, \frac{1}{n}Q\right) \geq \mathrm{Dist}\left(P,Q\right) \stackrel{def}{=} 1 - \rho\left([P,Q]_{\sqrt{}}\right). \quad (8)$$

The proof of Lemma 8 is given in Appendix C.

Finally, the following Theorem 9 gives an upper bound on sampling complexity of Algorithm 1. We note that $\widetilde{O}\left(\mathrm{HitT}_Q\right)$ samples are necessary for a reduction approach to work. Indeed, if we are to simulate $n\log n$ i.i.d. samples ($v \to u$, where $v \sim$ `Uniform`$[n]$ and $u \sim P_v$), then we shall see all states $v \in [n]$ at least once with high probability. I.e., the random walk must visit all the states, which would require at the very least HitT_Q steps in the random walk. On the other hand, our bound of $\widetilde{O}\left(\mathrm{HitT}_Q \cdot \log\left(\mathrm{HitT}_Q\right) + \frac{n}{\varepsilon}\right)$ is always better than a naive bound of $\mathrm{MixT}_Q \cdot \frac{n}{\varepsilon}$, since $\mathrm{HitT}_Q < n \cdot \mathrm{MixT}_Q$ and, in fact, for most of the reasonable MC HitT_Q is much less than that.

Theorem 9 *Given the description of a symmetric Markov chain Q and access to a single trajectory of length m from another symmetric Markov chain P, Algorithm 1 distinguishes between the cases $P = Q$ versus $1-\rho\left([P,Q]_{\sqrt{}}\right) > \varepsilon$ with probability at least $2/3$, for $m = \widetilde{O}\left(\mathrm{HitT}_Q \cdot \log\left(\mathrm{HitT}_Q\right) + \frac{n}{\varepsilon}\right)$.*

Proof In the case $P = Q$, the probability that Algorithm 1 proceeds to IID tester, i.e., it does not reject P, because of small number of visits to a state, is at least $\mathbf{Pr}[\forall i \in [n] \; |\{t : i = s_t \in w\}| > \frac{m}{8e\cdot n}] \cdot \mathbf{Pr}[\forall i : \frac{m}{8e\cdot n} > k_i] \geq \left(1 - \frac{\varepsilon^2}{n}\right) \cdot \left(1 - \frac{\varepsilon^2}{n}\right) \geq 1 - \frac{2\varepsilon^2}{n}$. In the previous estimate, we used Lemma 6 to bound $\mathbf{Pr}[\forall i \in [n] \; |\{t : s_t \in w, s_t = i\}| > \frac{m}{8e\cdot n}]$, the fact that $\mathbf{Pr}[\frac{m}{8e\cdot n} \leq k_i] \leq \frac{\varepsilon^2}{n^2}$ (follows from a Chernoff bound), and a union bound. IID tester then correctly accepts $P = Q$ with probability at least $4/5$. Hence, the error probability is at most $1/5 + \frac{2\varepsilon^2}{n} < 1/3$.

For the case $P \neq Q$, Lemma 8 says that if $1 - \rho\left([P,Q]_{\sqrt{}}\right) > \varepsilon$, then distributions passed down to the IID tester $\{p : p_{ij} = \frac{1}{n}P_{ij}\}$ and $\{q : q_{ij} = \frac{1}{n}Q_{ij}\}$ are at least ε far in Hellinger-squared distance. A black-box application of Lemma 7 implies a $O\left(\frac{n}{\varepsilon}\right)$ sampling complexity for the IID tester in our case. Furthermore, random mapping $\mathcal{K}_{\boldsymbol{k}} : \mathcal{W}_P^\infty \to p$ (where $\boldsymbol{k}$ is a histogram of $m' = \Theta\left(\frac{m}{\log^2(n/\varepsilon)}\right)$ i.i.d. uniform samples from $[n]$) produces m' i.i.d. samples from p. Hence, if Algorithm 1 has sufficient samples from P to define the mapping $\mathcal{K}_{\boldsymbol{k}}$, it would be able to distinguish p and q with probability at least $2/3$. On the other hand, if Algorithm 1 gets finite number of samples which are not sufficient to define the mapping $\mathcal{K}_{\boldsymbol{k}}$, then it correctly rejects P before even running the IID tester.

Thus in both cases the probability of error is at most $1/3$. ■

5. A Lower Bound for Identity Testing of Symmetric Markov Chains

In this section we provide an information theoretic lower bound to the identity testing problem on Markov chains defined in Section 4.

Theorem 10 *There exists a constant $c > 0$ and an instance of the identity testing problem for symmetric Markov chains such that any tester on this instance requires a word of length at least $c\frac{n}{\varepsilon}$ as input to produce the correct output with probability > 0.99.*

The full proof of Theorem 10 is given in Appendix D. The high level idea is to construct a Markov chain Q and a family of chains $\mathcal{P}$ such that it is hard to distinguish Q from a randomly chosen $\bar{P} \in \mathcal{P}$ by only looking at trajectories of length $o(n/\varepsilon)$. The chain Q and the family $\mathcal{P}$ we work with are described below (we think of symmetric Markov chains as weighted undirected graphs with multi-edges allowed).

Markov Chain Q: complete double graph on n vertices with uniform weights, i.e.,

$$\forall \quad i \neq j \qquad (ij)_1, (ij)_2 \in E \qquad Q_{(ij)_1} = Q_{(ij)_2} = \frac{1}{2(n-1)}.$$

Family $\mathcal{P}$: for any pair of vertices $i \neq j$ there are two bidirectional edges $(ij)_1$, $(ij)_2$ with weights randomly (and independently for each pair of (i,j)) chosen to be either

$$P_{(ij)_1}, P_{(ij)_2} = \frac{1 \pm \sqrt{8\varepsilon}}{2(n-1)}, \qquad \text{or} \qquad P_{(ij)_1}, P_{(ij)_2} = \frac{1 \mp \sqrt{8\varepsilon}}{2(n-1)}.$$

From the construction above it is clear that one needs to observe a number of collisions to distinguish Q from a randomly chosen member of $\mathcal{P}$. The proof proceeds by a careful analysis of these collision probabilities to bound the TV distance between words of length k from Q and from a randomly chosen $\bar{P} \in \mathcal{P}$.

6. Open Questions

In this paper, we proposed a new framework for studying property testing questions on Markov chains. There seem to be multiple avenues for future research and abundant number of open problems arising from this framework. We first list some questions which may be of interest here.

1. What is the optimal sample complexity for identity testing on symmetric Markov chains? In this paper, we show an upper bound of $\widetilde{O}\left(\mathrm{HitT}_Q \cdot \log\left(\mathrm{HitT}_Q\right) + \frac{n}{\varepsilon}\right)$ samples (Theorem 9). We conjecture that $\Theta\left(\frac{n}{\varepsilon}\right)$ (same as our lower bound) is the right sample complexity for this problem and an explicit dependence on the hitting time of chain Q may not be necessary. It is implicitly captured to an extent by the guarantee we get from the parameter ε.

2. As there is a natural operation of taking a convex combination of Markov chains, it is natural to ask how our spectral definition of distance $1 - \rho\left([P, Q]_{\sqrt{}}\right)$ between two symmetric chains changes if we substitute either P or Q with a convex combination of P and Q. How does the distance now relate to the original value?

3. Given $\varepsilon_2 \geq \varepsilon_1$, and access to words from each of two chains, can we distinguish whether the two chains are $\leq \varepsilon_1$-close or $\geq \varepsilon_2$-far? This problem, known as two-sample testing in literature, is another interesting direction using our framework.

References

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3591–3599, 2015. URL http://papers.nips.cc/paper/5839-optimal-testing-for-properties-of-distributions.

Alan Agresti. *Categorical Data Analysis*. Wiley, 2012.

Flavia Barsotti, Anne Philippe, and Paul Rochet. Hypothesis testing for Markovian models with random time observations. *Journal of Statistical Planning and Inference*, 173:87–98, 2016.

Maurice S Bartlett. The frequency goodness of fit test for probability chains. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 47, pages 86–95. Cambridge Univ Press, 1951.

Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '01, pages 442–451, Washington, DC, USA, 2001. IEEE Computer Society.

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):4, 2013.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 22, pages 1–9, 2015.

Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1174–1192, 2014. doi: 10.1137/1.9781611973402.87. URL http://dx.doi.org/10.1137/1.9781611973402.87.

Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In *Proceedings of the 33rd Symposium on Theoretical Aspects of Computer Science*, STACS '16, pages 25:1–25:14, 2016.

Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203. Society for Industrial and Applied Mathematics, 2014.

Constantinos Daskalakis and Qinxuan Pan. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, 2017.

Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1833–1852, 2013. doi: 10.1137/1.9781611973105.131. URL http://dx.doi.org/10.1137/1.9781611973105.131.

Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? *arXiv preprint arXiv:1708.00002*, 2017.

Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 685–694, Washington, DC, USA, 2016. IEEE Computer Society.

Ilias Diakonikolas, Daniel Kane, and Alistair Stewart. Robust learning of fixed-structure bayesian networks. *arXiv preprint arXiv:1606.07384*, 2016.

Ronald A. Fisher. *The Design of Experiments*. Macmillan, 1935.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Leon J Gleser, David S Moore, et al. The effect of dependence on chi-squared and empiric distribution tests of fit. *The Annals of Statistics*, 11(4):1100–1108, 1983.

Oded Goldreich. A brief introduction to property testing., 2011.

Daniel J Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. In *Advances in neural information processing systems*, pages 1459–1467, 2015.

Dimitri Kazakos. The Bhattacharyya distance and detection between Markov chains. *IEEE Trans. Information Theory*, 24(6):747–754, 1978. URL http://dx.doi.org/10.1109/TIT.1978.1055967.

Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.

I Molina, D Morales, L Pardo, and I Vajda. On size increase for goodness of fit tests when observations are positively dependent. *Statistics & Risk Modeling*, 20(1-4):399–414, 2002.

David S Moore et al. The effect of dependence on chi squared tests of fit. *The Annals of Statistics*, 10(4):1163–1171, 1982.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

Jon N.K. Rao and Alastair J. Scott. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the Americal Statistical Association*, 76(374):221–230, 1981.

Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24–28, 2012.

Vincent Y.F. Tan, Animashree Anandkumar, and Alan S. Willsky. Error exponents for composite hypothesis testing of Markov forest distributions. In *Proceedings of the 2010 IEEE International Symposium on Information Theory*, ISIT '10, pages 1613–1617, Washington, DC, USA, 2010. IEEE Computer Society.

Simon Tavare and Patricia M. E. Altham. Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. *Biometrika*, 70(1):139–144, 1983. ISSN 00063444. URL http://www.jstor.org/stable/2335951.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society.

Appendix A. Examples

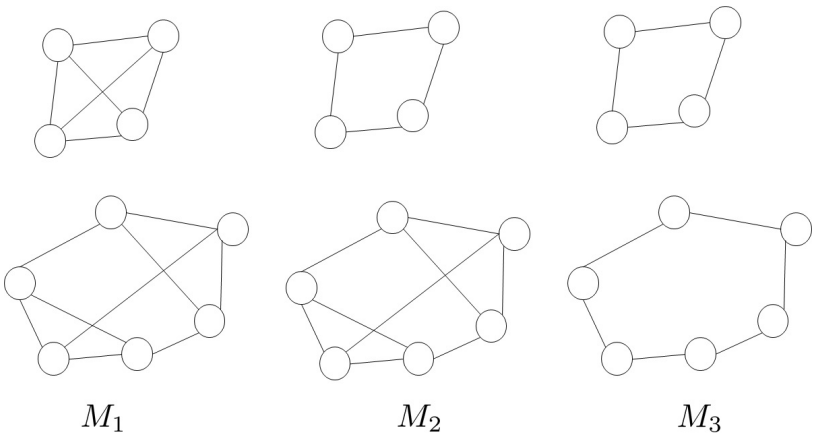

Figure 1: Dist $(M_1, M_2) = 1 - \rho\left([M_1, M_2]_{\vee}\right)$ is not a metric. Dist $(M_1, M_2) =$ Dist $(M_2, M_3) = 0$, but Dist $(M_1, M_3) > 0$.

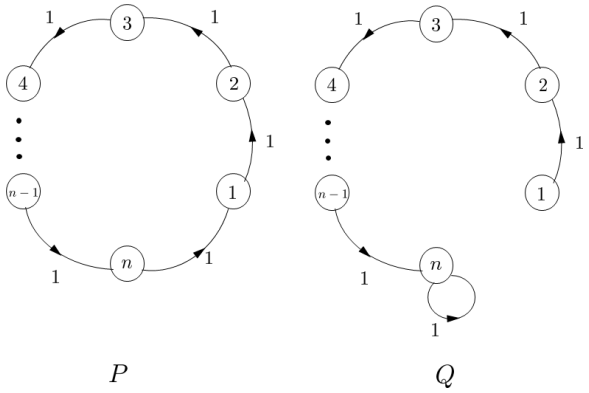

Figure 2: To distinguish P vs. Q walking from a random state we need $\Omega(n)$ steps, but Dist $(P, Q) = 1$.

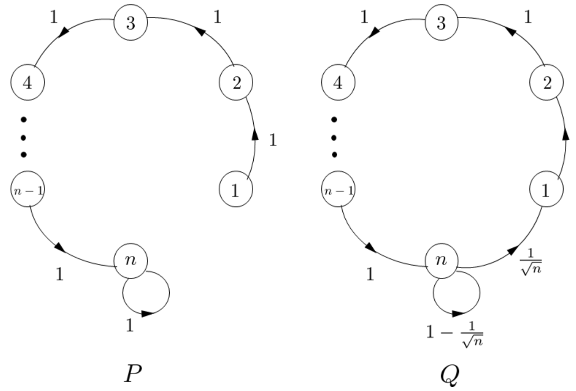

Figure 3: $\text{Dist}\,(P, Q) = o(1)$, stationary distributions $\boldsymbol{q}_0, \boldsymbol{p}_0$ are different: $\mathrm{d}_{\mathrm{TV}}\,(\boldsymbol{q}_0, \boldsymbol{p}_0) = 1 - o(1)$.

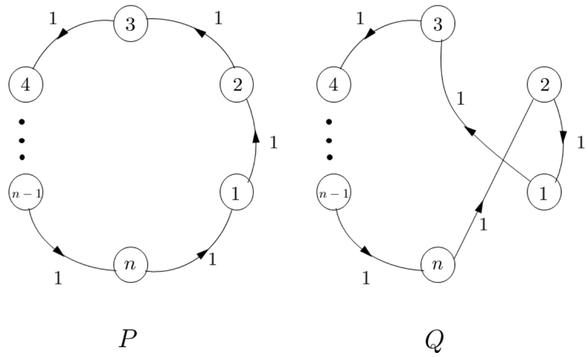

Figure 4: $\text{Dist}\,(P, Q) = 1$. Uniform is stationary for both P and Q. On average $\Omega(n)$ steps to tell $P \neq Q$.

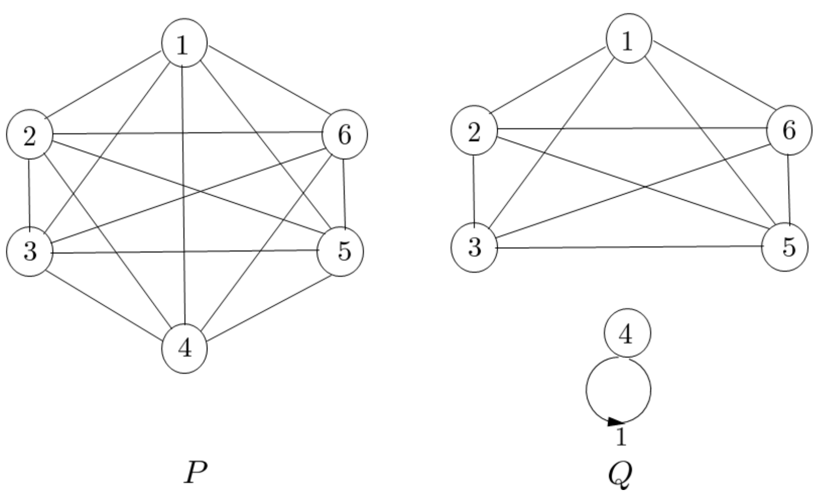

Figure 5: After one step from state 4, we would know if $w \sim P$, or $w \sim Q$. If w starts from any other state $s_0 \neq 4$, it would take many steps.

	Description
Example 1	Two disjoint connected components.
Example 5	Q – clique K_n; P – clique K_{n-1} and disjoint vertex. Eigenvalue of $[P,Q]_{\surd}$: $\lambda_1 = \sqrt{\frac{n-1}{n}} = 1 - o(1)$, $\lambda_2 = \sqrt{\frac{1}{n}}$, $\lambda_3 = \cdots = \lambda_n = 0$
Example 2	P – oriented cycle, Q – cycle with one link substituted by a loop.
Example 3	P – oriented cycle with edge $e = (v_1 v_2)$ substituted by a loop at v_1; Q is almost like P, but e has weight $\frac{1}{\sqrt{n}}$, loop at v_n has weight $1 - \frac{1}{\sqrt{n}}$. Stationary distributions: $\boldsymbol{p}_0 = (1, 0, \cdots, 0)^\top$ and $\boldsymbol{q}_0 = (\frac{\sqrt{n}}{n+\sqrt{n}-1}, \frac{1}{n+\sqrt{n}-1}, \ldots, \frac{1}{n+\sqrt{n}-1})^\top$. $\rho\left([P,Q]_{\surd}\right) = \sqrt{1 - \frac{1}{\sqrt{n}}}$.
Example 4	Two oriented cycles $P \stackrel{\text{def}}{=} s_1 \to s_2 \to \cdots \to s_n \to s_1$ and $Q \stackrel{\text{def}}{=} s_1 \to s_3 \to s_4 \cdots \to s_n \to s_2 \to s_1$.

Table 1: Examples.

Appendix B. Missing proofs from Section 3

Proof of Lemma 5:

$$1 - \mathrm{d}^2_{\mathrm{Hel}}\left(\mathcal{W}^\ell_P, \mathcal{W}^\ell_Q\right) = \sum_{w=s_0\ldots s_\ell} \sqrt{\mathop{\mathbf{Pr}}_P[w]\mathop{\mathbf{Pr}}_Q[w]} = \left[\sum_{\substack{w=s_0\ldots s_\ell \\ s_\ell = s}} \sqrt{\mathop{\mathbf{Pr}}_P[w]\mathop{\mathbf{Pr}}_Q[w]}\right]^\top_{s\in[n]} \cdot \mathbb{1}$$

$$= \left[\sum_{r\in[n]} \sqrt{\mathop{\mathbf{Pr}}_P[r\to s]\mathop{\mathbf{Pr}}_Q[r\to s]} \sum_{\substack{w=s_0\ldots s_{\ell-1} \\ s_{\ell-1}=r}} \sqrt{\mathop{\mathbf{Pr}}_P[w]\mathop{\mathbf{Pr}}_Q[w]}\right]^\top_{s\in[n]} \cdot \mathbb{1}$$

$$= \left[\sum_{\substack{w=s_0\ldots s_{\ell-1} \\ s_{\ell-1}=r}} \sqrt{\mathop{\mathbf{Pr}}_P[w]\mathop{\mathbf{Pr}}_Q[w]}\right]^\top_{r\in[n]} \cdot \begin{bmatrix} \vdots & \\ \cdots & \sqrt{P_{rs}\cdot Q_{rs}} \quad \cdots \\ \vdots & \end{bmatrix}_{r,s\in[n\times n]} \cdot \mathbb{1}$$

$$= \left[\sum_{\substack{w=s_0\ldots s_{\ell-1} \\ s_{\ell-1}=r}} \sqrt{\mathop{\mathbf{Pr}}_P[w]\mathop{\mathbf{Pr}}_Q[w]}\right]^\top_{r\in[n]} \cdot [P,Q]_{\surd} \cdot \mathbb{1} = [\boldsymbol{p},\boldsymbol{q}]^\top_{\surd} \cdot \left([P,Q]_{\surd}\right)^\ell \cdot \mathbb{1},$$

■

Proof of Claim 1: Note that $\frac{P+Q}{2}$ is a stochastic matrix that entry-wise dominates matrix $[P,Q]_{\surd}$ with non-negative entries. Therefore, $\lambda_1 \cdot \langle \boldsymbol{u}_1, \mathbb{1}\rangle = \boldsymbol{u}_1^\top \cdot [P,Q]_{\surd} \cdot \mathbb{1} \le \boldsymbol{u}_1^\top \cdot \left[\frac{P+Q}{2}\right] \cdot \mathbb{1} = \boldsymbol{u}_1^\top \cdot \mathbb{1} = \langle \boldsymbol{u}_1, \mathbb{1}\rangle$, where $\mathbb{1}$ is vector with all 1 entries. We get $\lambda_1 \le 1$, since $\langle \boldsymbol{u}_1, \mathbb{1}\rangle > 0$.

For the case of equality, if P and Q have the same essential communicating class C, then matrix $[P,Q]_{\surd}$ has the same transition probabilities as Markov chains P and Q restricted to the vertices of

C. We note that C is a stochastic matrix and, therefore, its largest positive eigenvalue is one. Hence, $\rho\left([P,Q]_{\surd}\right) \geq \rho(C) = 1$.

If $\rho\left([P,Q]_{\surd}\right) = 1$, we apply Perron-Frobinius theorem to $[P,Q]_{\surd}$ to get that the largest eigenvalue $\lambda_1 = \rho\left([P,Q]_{\surd}\right) = 1$ has corresponding (left) eigenvector $\boldsymbol{u}_1$ with non-negative entries. We observe that $\boldsymbol{u}_1^\top \cdot \left(\frac{P+Q}{2} - [P,Q]_{\surd}\right) \cdot \mathbb{1} = 0$, and all entries of the matrix in this expression are non-negative. This implies that $P_{ij} = Q_{ij}$ for every strictly positive coordinates i of the eigenvector $\boldsymbol{u}_1$ and any $j \in [n]$. Since $\boldsymbol{u}_1^\top \cdot [P,Q]_{\surd} = \boldsymbol{u}_1^\top$, we also have $P_{ij} = Q_{ij} = 0$ for any positive coordinate i and zero coordinate j of eigenvector $\boldsymbol{u}_1$. Therefore, the set of vertices corresponding to positive coordinates of $\boldsymbol{u}_1$ form a component (which might have more than one connected component of P and Q) such that $P = Q$ on these vertices. ■

Proof of Claim 2: We first consider the average-case model for the starting state. Note that $[P,Q]_{\surd}$ is a symmetric matrix. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be normalized orthogonal eigenvectors of $[P,Q]_{\surd}$, corresponding to real $\lambda_1 \geq \cdots \geq \lambda_n$ eigenvalues. Then for RHS of (6) we have

$$\frac{1}{n}\mathbb{1}^\top \cdot \left([P,Q]_{\surd}\right)^\ell \cdot \mathbb{1} = \frac{1}{n}\mathbb{1}^\top \cdot \left(\sum_{i=1}^n \lambda_i \cdot \boldsymbol{v}_i \cdot \boldsymbol{v}_i^\top\right)^\ell \cdot \mathbb{1} = \sum_{i=1}^n \lambda_i^\ell \cdot \frac{1}{n}\langle \mathbb{1}, \boldsymbol{v}_i\rangle^2 = (*) \tag{9}$$

Now we can write an upper and lower bound on $(*)$ in terms of λ_1^ℓ (assuming that ℓ is even):

$$\frac{\lambda_1^\ell}{n} = \frac{\lambda_1^\ell}{n}\|\boldsymbol{v}_1\|_2^2 \leq \lambda_1^\ell \cdot \frac{1}{n}\|\boldsymbol{v}_1\|_1^2 \leq (*) \leq \sum_{i=1}^n \lambda_i^\ell \cdot \frac{1}{n}\|\boldsymbol{v}_i\|_1^2 \leq \sum_{i=1}^n \lambda_i^\ell \cdot \|\boldsymbol{v}_i\|_2^2 = \sum_{i=1}^n \lambda_i^\ell \leq n \cdot \lambda_1^\ell,$$

where in the second inequality we used Perron-Frobenius theorem stating that all coordinates of $\boldsymbol{v}_1$ are non negative. Consequently, these bounds imply that $\ell = \Theta\left(\frac{1}{\varepsilon}\right)$ up to a $\log n$ factor, if $\rho\left([P,Q]_{\surd}\right) = \lambda_1 = 1 - \varepsilon$. I.e., $\ell = \widetilde{\Theta}\left(\frac{1}{\varepsilon}\right)$.

For the worst-case assumption on the starting state, it is sufficient to show an upper bound $\ell = O\left(\frac{\log n}{\varepsilon}\right)$. In this case (9) becomes

$$\boldsymbol{e}_i^\top \cdot \left([P,Q]_{\surd}\right)^\ell \cdot \mathbb{1} = \sum_{i=1}^n \lambda_i^\ell \cdot \langle \boldsymbol{e}_i, \boldsymbol{v}_i\rangle \cdot \langle \mathbb{1}, \boldsymbol{v}_i\rangle \leq \sum_{i=1}^n |\lambda_i|^\ell \cdot \|\boldsymbol{v}_i\|_\infty \cdot \|\boldsymbol{v}_i\|_1 \leq \sum_{i=1}^n |\lambda_i|^\ell \cdot \sqrt{n} \leq n^{1.5} \cdot \lambda_1^\ell,$$

since $\|\boldsymbol{v}_i\|_1 \leq \sqrt{n}\,\|\boldsymbol{v}_i\|_2 = \sqrt{n}$, and $\|\boldsymbol{v}_i\|_\infty \leq \|\boldsymbol{v}_i\|_2 = 1$. ■

Appendix C. Missing proofs from Section 4

Proof of Lemma 6: To simplify notations we denote by $\Delta \stackrel{\text{def}}{=} 2\mathrm{HitT}_Q$. By union bound over all states i it is enough to show that $\mathbf{Pr}[|\{t : i = s_t \in w\}| < \frac{m}{8e \cdot n}] \leq \frac{\varepsilon^2}{n^2}$ for each fixed state i. We can make sure that in the first $\frac{m}{2}$ steps state i is visited at least once with probability at least $1 - \frac{\varepsilon^2}{n^3}$. Once we visited state i, instead of hitting time for state i we can analyze the *return time* Return_i

for i. Note that for symmetric Markov chains $\frac{1}{n}\mathbb{1}$ (uniform distribution) is a stationary distribution. Therefore, every state appears at average once in every n steps in an infinite word from $\mathcal{W}_Q^\infty$. In other terms, the expectation of Return_i for each state i is exactly n. By definition of hitting time we have that in $\frac{\Delta}{2}$ steps the probability of reaching a particular state i from any other state j is greater than $1-1/e$ (or any other given constant). It implies that $\mathbf{Pr}[\mathrm{Return}_i \geq \frac{\Delta}{2}\cdot C] \leq e^{-C}$ for any $C \in \mathbb{N}$. Indeed, one can show this by induction on parameter C. Notice that if the random walk did not return to i after $C-1$ steps it has stopped at some state $j \neq i$. Then for any choice of j by definition of the hitting time the random walk will return to i with probability at least $1/e$ in the next $\frac{\Delta}{2}$ steps. It is not hard to get a similar bound $\mathbf{Pr}[\mathrm{Return}_i \geq \Delta \cdot C] \leq e^{-C}$ for any $C \geq 1, C \in \mathbb{R}$. To simplify notations we use X to denote the random variable Return_i and $X_1, \ldots, X_\ell$ to denote ℓ i.i.d. samples of X. We have

$$X \geq 0 \quad \text{and} \quad \forall C \in \mathbb{R}_{\geq 1}, \mathbf{Pr}\left[X \geq \Delta \cdot C\right] \leq e^{-C} \quad \text{and} \quad \mathbf{E}\left[X\right] = n. \tag{10}$$

We only need to show that $\mathbf{Pr}[X_1 + \cdots + X_\ell > m/2] \leq \frac{\varepsilon^2}{n^2}$ for $\ell = \frac{m}{8e\cdot n}$. To this end, we use a standard technique for large deviations and apply Markov's inequality to the moment generating function of $X_1 + \cdots + X_\ell$,

$$\mathbf{Pr}\left[X_1 + \cdots + X_\ell > m/2\right] = \mathbf{Pr}\left[e^{\theta\cdot(X_1+\cdots+X_\ell)} > e^{\theta\cdot m/2}\right] \leq \frac{\mathbf{E}[e^{\theta\cdot(X_1+\cdots+X_\ell)}]}{e^{\theta\cdot m/2}} = \frac{\mathbf{E}[e^{\theta X}]^\ell}{e^{\theta\cdot m/2}} \tag{11}$$

We note that given restrictions (10) on X maximum of $\mathbf{E}[e^{\theta X}]$ for any fixed $\theta > 0$ is attained at

$$X^* \sim \begin{cases} \Delta\cdot x & x \in [C_0, \infty) \text{ with probability density function } e^{-x} \\ 0 & \text{with remaining probability } 1 - e^{-C_0}, \end{cases}$$

where constant $C_0 > 1$ is such that $\mathbf{E}[X^*] = n$. Indeed, distribution X^* maximizes (11) due to simple variational inequality: $\epsilon\cdot e^{\theta\cdot a} + \epsilon\cdot e^{\theta\cdot b} < \epsilon\cdot e^{\theta\cdot(a-c)} + \epsilon\cdot e^{\theta\cdot(b+c)}$ for any $b \geq a \geq c > 0$, and probability mass $\epsilon > 0$. This inequality allows us to increase $\mathbf{E}[e^{\theta\cdot X}]$ and not change $\mathbf{E}[X]$ by tweaking density function $f(x)$ of X $f'(a-c) = f(a-c)+\epsilon$, $f'(a) = f(a)-\epsilon$, $f'(b) = f(b)-\epsilon$, $f'(b+c) = f'(b+c)+\epsilon$, ($f'(x) = f(x)$ for all other x), for some $c \leq a$. The only time we cannot apply this incremental change is when $X = X^*$.

We have

$$\mathbf{E}\left[X^*\right] = \Delta(C_0+1)e^{-C_0} = n. \tag{12}$$

We set $\theta \stackrel{\text{def}}{=} \frac{1}{2\Delta\log\Delta}$ in (11). Now we are ready to estimate $\mathbf{E}[e^{\theta\cdot X}]$. To simplify notations we denote $\gamma \stackrel{\text{def}}{=} \frac{1}{2\log\Delta}$.

$$\begin{aligned}\mathbf{E}\left[e^{\theta\cdot X}\right] &= 1 - e^{-C_0} + \int_{C_0}^{\infty} e^{\theta\cdot\Delta\cdot x}\cdot e^{-x}\,\mathrm{d}x = 1 - e^{-C_0} + \int_{C_0}^{\infty} e^{-x\cdot(1-\frac{1}{2\log\Delta})}\,\mathrm{d}x \\ &= 1 - e^{-C_0} + \frac{e^{-C_0(1-\gamma)}}{1-\gamma} = 1 + e^{-C_0}\left(\frac{e^{C_0\gamma}}{1-\gamma} - 1\right). \end{aligned} \tag{13}$$

We notice that $\gamma C_0 < 1$, since from (12) we can conclude that $\frac{e^{C_0}}{C_0+1} = \frac{\Delta}{n} \implies C_0 < 2\log\Delta = 1/\gamma$. The last implication can be obtained as follows: for $C_0 > 2.52$, we have $C_0 - \frac{C_0}{2} \leq C_0 -$

$\ln(1+C_0) = \ln(\frac{\Delta}{n})$. Now, we can estimate $e^{\gamma C_0} \le 1 + e \cdot \gamma C_0$ in (13). Furthermore, since $\gamma < 1/2$ we have the term $\frac{e^{C_0\gamma}}{1-\gamma} - 1$ in (13) to be at most $2e\gamma(C_0+1)$. With this estimate we continue (13)

$$\mathbf{E}\left[e^{\theta \cdot X}\right] \le 1 + e^{-C_0} 2e\gamma(C_0+1) = 1 + \frac{e \cdot n}{\Delta \log \Delta}. \tag{14}$$

We apply estimate (14) and formula $\theta = \frac{1}{2\Delta \log \Delta}$ to (11) to obtain

$$\mathbf{Pr}\left[X_1 + \cdots + X_\ell > m/2\right] \le \frac{\left(1 + \frac{e \cdot n}{\Delta \log \Delta}\right)^\ell}{e^{m/4\Delta \log \Delta}} \le \frac{e^{m/8\Delta \log \Delta}}{e^{m/4\Delta \log \Delta}} = e^{\frac{-m}{8\Delta \log \Delta}} < \frac{\varepsilon^2}{n^2},$$

where in the second inequality we used the fact $\left(1 + \frac{e \cdot n}{\Delta \log \Delta}\right)^{\frac{\Delta \log \Delta}{e \cdot n}} < e$, and to get the last inequality we used $m = \widetilde{\Omega}(\Delta \log \Delta)$ (where in $\widetilde{\Omega}$ the hidden dependency is only on $\log \varepsilon$ and $\log n$). ■

Proof of Lemma 8: We note that, as P and Q are symmetric matrices, so is $[P,Q]_{\surd}$. Thus we have

$$1 - \varepsilon = \rho\left([P,Q]_{\surd}\right) = \max_{\|\boldsymbol{v}\|_2 = 1} \boldsymbol{v}^\top \cdot [P,Q]_{\surd} \cdot \boldsymbol{v}. \tag{15}$$

If we use a particular $\boldsymbol{v} = \frac{1}{\sqrt{n}}\mathbb{1}$ in (15), then we get the following inequality.

$$1 - \varepsilon \ge \frac{1}{\sqrt{n}}\mathbb{1}^\top \cdot [P,Q]_{\surd} \cdot \frac{1}{\sqrt{n}}\mathbb{1} = \frac{1}{n}\sum_{i,j}\sqrt{P_{ij} \cdot Q_{ij}} = 1 - \mathrm{d}^2_{\mathrm{Hel}}\left(\frac{1}{n}P, \frac{1}{n}Q\right),$$

which implies $\mathrm{d}^2_{\mathrm{Hel}}\left(\frac{1}{n}P, \frac{1}{n}Q\right) \ge \varepsilon$. ■

Appendix D. Proof of the Lower Bound

Here we present the proof of Theorem 10.

Proof of Theorem 10: We use Le Cam's two point method and construct a symmetric Markov chain Q and a class of symmetric Markov chains $\mathcal{P}$ s.t. (i) every $P \in \mathcal{P}$ is at least ε far from Q for a given constant ε. That is $1 - \rho\left([P,Q]_{\surd}\right) \ge \varepsilon$ for any $P \in \mathcal{P}$; (ii) there is a constant $c > 0$, s.t. it is impossible to distinguish a word of length m generated by a randomly chosen Markov chain $\bar{P} \sim \mathcal{P}$, from a word of length m produced by Q with probability equal to or greater than $\frac{99}{100}$ for $m \le \frac{cn}{\varepsilon}$. To prove (ii) we show that the total variation distance between the m-word distributions obtained from the two processes, Q and $\bar{P}$, is small when $m < \frac{cn}{\varepsilon}$ for some constant c. We denote distribution of length m words obtained from Q by $\mathcal{W}^m_Q$, and from MC $\bar{P} \sim \mathcal{P}$ by $\mathcal{W}^m_{\mathcal{P}}$. We represent symmetric MC as undirected weighted graphs $G = (V, E)$. We allow graph to have multi-edges (this is helpful to provide an intuitive understanding of the lower bound construction and is not essential). We can ultimately remove all multi-edges and give a construction with only simple edges by doubling the number of states.

Markov Chain Q: complete double graph on n vertices with uniform weights, i.e.,

$$\forall \quad i \neq j \qquad (ij)_1, (ij)_2 \in E \qquad Q_{(ij)_1} = Q_{(ij)_2} = \frac{1}{2(n-1)}.$$

Family $\mathcal{P}$: for any pair of vertices $i \neq j$ there are two bidirectional edges $(ij)_1$, $(ij)_2$ with weights randomly (and independently for each pair of (i, j)) chosen to be either

$$P_{(ij)_1}, P_{(ij)_2} = \frac{1 \pm \sqrt{8\varepsilon}}{2(n-1)}, \qquad \text{or} \qquad P_{(ij)_1}, P_{(ij)_2} = \frac{1 \mp \sqrt{8\varepsilon}}{2(n-1)}.$$

To make this instance a simple graph with at most one bidirectional edge between any pair of vertices we apply a standard graph theoretic transformation: we make a copy i' for each vertex i; for each pair of double edges $e_1 = (ij)_1, e_2 = (ij)_2$ construct 4 edges $(ij), (ij'), (i'j), (i'j')$ with weights $w(ij) = w(i'j') = w(e_1)$ and $w(ij') = w(i'j) = w(e_2)$.

As all Markov chains Q and $P \in \mathcal{P}$ are symmetric with respect to the starting state, we can assume without loss of generality that word w starts from the state $i = 1$. First, we observe that for the simple graph $2n$-state representation

Lemma 11 *Every Markov chain $P \in \mathcal{P}$ is at least ε-far from Q.*

Proof For any $P \in \mathcal{P}$, it can be seen that

$$[P, Q]_{\surd} \cdot \mathbb{1} = \left(\frac{\sqrt{1 + \sqrt{8\varepsilon}} + \sqrt{1 - \sqrt{8\varepsilon}}}{2} \right) \cdot \mathbb{1}.$$

By Perron-Frobenius theorem $\mathbb{1}$ is the unique eigenvector corresponding to the largest absolute value eigenvalue. Hence, $\rho\left([P, Q]_{\surd}\right) = \frac{\sqrt{1+\sqrt{8\varepsilon}}+\sqrt{1-\sqrt{8\varepsilon}}}{2}$ which by Taylor series expansion implies $1 - \rho\left([P, Q]_{\surd}\right) \geq \varepsilon + \frac{5}{2}\varepsilon^2 + o(\varepsilon^2) \geq \varepsilon$ for any $\varepsilon < \frac{1}{8}$. ■

We say that a given word $w = s_1 \dots s_m$ from a Markov chain P represented as a multi-edge graph on n states has a (ij) *collision*, if any state transition between states i and j (in any direction along any of the edges $(ij)_1, (ij)_2$) occurs more than once in w. We now state and prove the following claims about the Markov chain family $\mathcal{P}$.

Lemma 12 *Consider a word w of length m drawn from Q. The expected number of collisions in w is at most $O\left(\frac{m^2}{n^2}\right) = O\left(\frac{1}{\varepsilon^2}\right)$.*

Proof of Lemma 12: Let $I_w(t_1, t_2, (ij))$ indicate the event that in the multi-edge interpretation of the Markov chain P, the transition along (ij) edge occurs at times $t_1 < t_2$ in w. First, we observe that $\mathbf{Pr}[s_{t_1} = s | s_{t_1-1} = x] \leq \frac{1}{n-1}$ and $\mathbf{Pr}[s_{t_2} = s | s_{t_1-1} = x] \leq \frac{1}{n-1}$ for all x and both $s = i$ or $s = j$. Thus for any $t_2 \geq t_1 + 2$ by a union bound for all four possible cases of $s_{t_1}, s_{t_1+1}, s_{t_2}, s_{t_2+1} \in \{i, j\}$ we have

$$\mathbb{E}\left[I_w(t_1, t_2, (ij))\right] \leq \frac{4}{(n-1)^4}.$$

Similarly, for the case $t_2 = t_1 + 1$ we can obtain

$$\mathbb{E}\left[I_w(t_1, t_2, (ij))\right] \leq \frac{2}{(n-1)^3}.$$

Let X denote the random variable which is equal to the total number of collisions in the word w. Then,

$$\begin{aligned}\mathbf{E}\left[X\right] &= \sum_{t_2 \geq t_1+2} \sum_{i \neq j} \mathbf{E}\left[I_w(t_1, t_2, (ij))\right] + \sum_{t_1=1}^{m-1} \sum_{i \neq j} \mathbf{E}\left[I_w(t_1, t_1+1, (ij))\right] \\ &\leq \frac{4}{(n-1)^4} \cdot \frac{m^2}{2} \cdot \frac{n(n-1)}{2} + \frac{2}{(n-1)^3} \cdot m \cdot \frac{n(n-1)}{2} = O\left(\frac{m^2}{n^2}\right)\end{aligned}$$

■

We also consider 3-way collisions which are collisions where there was at least 3 different transition between a pair of states i and j in the word w.

Lemma 13 *Consider a word w of length m drawn from Q. The probability of w having a 3-way collision is at most $O(\frac{m^3}{n^4}) = o(1)$.*

Proof of Lemma 13: Similar to the proof of Lemma 12 we can give a sharp upper bound on the expected number of 3-way collisions with the most significant term being $\frac{8m^3}{6(n-1)^6} \cdot \frac{n(n-1)}{2}$, i.e., the expected number of 3-way collisions is $O\left(\frac{m^3}{n^4}\right)$. By Markov inequality we obtain the required bound on the probability of a 3-way collision. ■

Now consider a typical word w generated by Q. As we know from Lemma 13 it has no 3-way collisions and by Markov inequality and Lemma 12 has at most $O(\frac{1}{\varepsilon^2})$ collisions with high probability. As we show next a typical word w has similar probabilities under Q or $\bar{P} \sim \mathcal{P}$ models.

Lemma 14 *For $m = O(\frac{n}{\varepsilon})$ at least $\frac{1}{2}$ fraction of words $w = s_1 \cdots s_m$ generated by Q satisfy*

$$\frac{1}{2} \cdot \mathbf{Pr}_Q\left[w\right] < \mathbf{Pr}_{\bar{P} \sim \mathcal{P}}\left[w\right] < 2 \cdot \mathbf{Pr}_Q\left[w\right]$$

Proof of Lemma 14: For each feasible word w in Q, i.e., w such that $\mathbf{Pr}_Q[w] > 0$

$$\mathop{\mathbf{Pr}}_{Q}\left[w\right] = \left(\frac{1}{2(n-1)}\right)^{m-1} \qquad \mathop{\mathbf{Pr}}_{\bar{P} \sim \mathcal{P}}\left[w\right] = \prod_{j>i} \sum_{\bar{P}_{(ij)_1} = \frac{1 \pm \sqrt{8\varepsilon}}{2(n-1)}} \bar{P}_{(ij)_1}^{|\{(ij)_1 \in w\}|} \cdot \bar{P}_{(ij)_2}^{|\{(ij)_2 \in w\}|}$$

First, if w has only one transition along edge (ij), then the corresponding term in $\mathbf{Pr}_{\bar{P} \sim \mathcal{P}}[w]$

$$\sum_{\bar{P}_{(ij)_1}} \bar{P}_{(ij)_1}^{|\{(ij)_1 \in w\}|} \cdot \bar{P}_{(ij)_2}^{|\{(ij)_2 \in w\}|} = \frac{1}{2}\left(\frac{1+\sqrt{8\varepsilon}}{2(n-1)} + \frac{1-\sqrt{8\varepsilon}}{2(n-1)}\right) = \frac{1}{2(n-1)}.$$

From Lemma 13, we know that probability of a 3-way collision in w is $o(1)$ under Q model. We observe that for a 2-way collision (ij) (a collision which is not a 3-way collision), the corresponding term in $\mathbf{Pr}_{\bar{P}\sim\mathcal{P}}[w]$ for the case of transition along two different edges $(ij)_1$ and $(ij)_2$ is

$$\sum_{\bar{P}_{(ij)_1}} \bar{P}_{(ij)_1}^{|\{(ij)_1\in w\}|} \cdot \bar{P}_{(ij)_2}^{|\{(ij)_2\in w\}|} = \frac{1+\sqrt{8\varepsilon}}{2(n-1)} \cdot \frac{1-\sqrt{8\varepsilon}}{2(n-1)} = \frac{(1-8\varepsilon)}{4(n-1)^2}.$$

We call this type of collision *type I* collision. For the other case (*type II* collisions) of transition along the same edges the respective probability is $\frac{(1+8\varepsilon)}{4(n-1)^2}$. By Lemma 12 and by Markov inequality the total number of collisions is $O(\frac{1}{\varepsilon^2})$ with probability $\frac{3}{4}$. We can also make sure that out of these collisions number of type I and type II collisions is roughly the same. More precisely, the difference between numbers of type I and type II collisions is at most $O(\frac{1}{\varepsilon})$ with probability of at least $\frac{3}{4}$. Indeed, the choice of edge collision type in w is uniform between type I and type II, and is independent across all collision edges. Now, for small enough m we can make sure that at least $\frac{1}{2}$ fraction of words w has number of collisions at most $\frac{c_1}{\varepsilon^2}$ and the difference between number of type I and II collisions is at most $\frac{c_2}{\varepsilon}$, for some small constants $c_1, c_2 > 0$. Thus the corresponding density functions can be related as follows.

$$2 > (1+8\varepsilon)^{\frac{c_2}{\varepsilon}} > \frac{\mathbf{Pr}_{\bar{P}\sim\mathcal{P}}[w]}{\mathbf{Pr}_Q[w]} > \left(1-64\varepsilon^2\right)^{\frac{c_1}{2\varepsilon^2}} \cdot (1-8\varepsilon)^{\frac{c_2}{\varepsilon}} > 1/2$$

■

Lemma 14 shows that $\mathrm{d}_{\mathrm{TV}}\left(\mathcal{W}_Q^m, \mathcal{W}_{\mathcal{P}}^m\right) \leq \frac{3}{4}$, which implies that no algorithm can successfully distinguish Q from the family $\mathcal{P}$ with probability greater than $\frac{3}{4}$ for some $m = \Omega(\frac{n}{\varepsilon})$. ■