Statistical Windows in Testing for the Initial Distribution of a Reversible Markov Chain

Quentin Berthet
Statistical Laboratory
DPMMS, University of Cambridge

Varun Kanade
Department of Computer Science
University of Oxford

Abstract

We study the problem of hypothesis testing between two discrete distributions, where we only have access to samples after the action of a known reversible Markov chain, playing the role of noise. We derive instance-dependent minimax rates for the sample complexity of this problem, and show how its dependence in time is related to the spectral properties of the Markov chain. We show that there exists a wide *statistical window*, in terms of sample complexity for hypothesis testing between different pairs of initial distributions. We illustrate these results in several concrete examples.

1 INTRODUCTION

Random walks on graphs, or Markov chains more generally, have long served as a natural model for data observed through a noisy channel. In many applications, one wishes to determine the origin of a random walk, after a certain number of steps: Where has a rumor started in a social network? What is the distribution of a deck of cards before shuffling? What are the ancestors of current species before evolution? What is the initial configuration of spins in Glauber dynamics? In these cases, the initial distribution is considered as the true information, and observations are made after the action of a few steps of the Markov chain. In this work, we consider the problem of testing for the initial distribution: determining which of two candidate distributions μ and μ' is the initial distribution, based on an i.i.d. sample after t steps. The mixing time of the chain can be interpreted as the time at which "all" starting information is lost. However, depending on the two candidate distributions over the starting states, this total loss of information may occur at a time much sooner than the mixing time. In this work, we characterize precisely this rate of loss of information in terms of spectral properties of the transition matrix: we give a theoretical guarantee on the performance of a test in terms of necessary sample size as a function of all the parameters of the problem, and show that this dependency is tight, using information-theoretic tools. In particular, we show that the sample complexity of the hypothesis testing problem between μ and μ', and its dependency in time t, depends critically on the pair (μ, μ') in an explicit manner. We call this wide range of sample complexities the *statistical window*. Pairs of distributions that exhibit behaviour at the extreme ends of this statistical window can be explicitly constructed using the spectrum of the associated Markov chain and we illustrate this phenomenon on several concrete examples.

Recovering information about a discrete distribution with access to samples is one of the central problems of statistical theory, going back at least to Laplace (1812). This essential problem has attracted much attention in the modern treatment of learning theory, on problems related to learning, testing and estimation. Recently, there has been renewed focus on learning discrete distributions from a sample—typically, it is assumed that the (unknown) distribution satisfies certain properties such as k-modality or monotonicity, which, in certain cases, allows significantly improved sample complexity over the basic approach of using the empirical distribution (see e.g. Chan et al. (2013), Daskalakis et al. (2012), Kamath et al. (2015), Diakonikolas et al. (2014), Daskalakis et al. (2015), Diakonikolas (2016), Diakonikolas and Kane (2016), Diakonikolas et al. (2017b,a), Valiant and Valiant (2014, 2016)). A related area of research is that of property testing, i.e. to test whether a distribution satisfies some property such as uniformity or monotonicity, from a sample (see e.g. Batu et al. (2000), Valiant (2011), Chan et al. (2014), Diakonikolas et al. (2015),

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Canonne (2017)). Yet another area of interest has been estimating quantities related to a discrete distribution, such as support size or entropy (see e.g. Valiant and Valiant (2011), Acharya et al. (2014), Wu and Yang (2016a,b), Orlitsky et al. (2016)).

In much of this literature, it is often assumed that one has direct access to independent samples from the true unknown distribution of interest μ.[1]

As stated above, we consider in this work a setting where we only have access to these samples after a reversible Markov chain has acted on them $t \geq 0$ times. In doing so, we allow for the introduction of noise, in the form of the action of a known Markov chain with transition matrix P. Formally, this is equivalent to learning about μ, with access to $\mu_t = \mu P^t$, and can be seen as a statistical inverse problem. In our setting, as t increases and μ_t approaches the stationary distribution, more information is lost and the statistical problem becomes more difficult. This is in stark contrast to the usual applications of Markov chains in statistical learning, in particular for Markov Chain Monte Carlo methods, where the stationary distribution π is the quantity of interest, from which it is difficult to sample, and a large t is desirable. In our setting, t can be understood as a way to measure the amount of noise, and we seek to understand how the difficulty increases with it. This is a common point of view in some continuous settings: if $\mu = \delta_x$ for some $x \in \mathbf{R}$, the action of the heat kernel for time t leads to a distribution $\mu_t = \mathcal{N}(x, t)$, and the impact of $\sigma^2 = t$ on the statistical difficulty of recovering x is clear. We transfer this idea to discrete distributions, and the action of a Markov chain is the most natural way to introduce noise.

Formally, we focus here on the fundamental problem of hypothesis testing between two known distributions μ and μ', based on samples from $\mu_t = \mu P^t$ or $\mu'_t = \mu' P^t$ (cf. Fig. 1). This choice allows us to illustrate, for many natural examples, the impact of the pair μ, μ' on the instance-dependent sample complexity of this problem, the required sample size $n^*_{\mu,\mu',t}$ to solve this problem with a small probability of error, and in particular its dependency in time. There is much focus in the field of mathematical statistics on the analysis of the maximum probability of error in hypothesis testing. It is well-understood that it is equivalent to studying the *total variation distance* $d_{\mathsf{TV}}(\mu_t^{\otimes n}, \mu_t'^{\otimes n})$ between the two distributions of samples of size n. This quantity grows in n and decays in t, and the goal of our analysis is to describe the trade-off between these two phenomena: we establish how large

[1] There have been some recent advances where some of these problems can be solved even when some (small) fraction of the data has been adversarially tampered with.

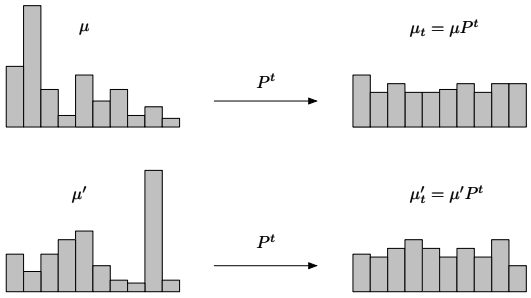

Figure 1: The hypothesis testing problem between μ and μ' is easier if we have direct sample access to these distributions. It is harder with the action of P^t, as the distributions look increasingly alike. For very large t, the distribution is very close to stationarity and all initial information is lost.

n and how small t need to be for the total variation distance to be bounded away from 0, for testing to be possible. We are particularly interested in showing that the behaviour in t is far from universal: we exhibit pairs (μ, μ') for which it is very different, as well as a systematic manner to construct them.

Our analysis bridges two fields where studying this quantity is a central problem. On the one hand, for fixed t, given μ_t and μ'_t, understanding the growth in n of this distance is one of the central problems of mathematical statistics and information theory. It is difficult to establish the growth of $d_{\mathsf{TV}}(\mu_t^{\otimes n}, \mu_t^{\otimes n})$ in terms of n and $d_{\mathsf{TV}}(\mu_t, \mu_t)$ directly (see, e.g. Berthet, 2014, Berthet and Ellenberg, 2019).

In our analysis, this quantity is controlled by comparison to other notions of "distances". On the other hand, one of the most important questions in the study of Markov chains is the behavior of μ_t, and in particular the decay of its total variation distance to the stationary distribution, as a function of t. One key quantity is that of mixing time, which describes the time t at which the total variation goes to e^{-1}.

After mixing, the total variation distance to the stationary distribution π decays exponentially. The mixing time represents the time at which the sample complexity of the hypothesis testing problem explodes, *for all* pairs of initial distributions μ, μ'. What emerges from our analysis in Section 3 is that for t less than the mixing time, this is far from universal over all pairs μ, μ', and that in many natural Markov chains, the sample complexity can vary dramatically. In particular, asking when μ_t and μ'_t are close is a very different question from asking when μ_t is close to the stationary distribution π. In particular, we construct explicitly examples of pairs of initial distributions (μ, μ')

whose sample complexity has a very different behavior in time.

Our main results are in Section 3.3, where we show that the instance sample complexity is

$$n^*_{\mu,\mu',t} \asymp 1/\sum_{i=2}^{d} \lambda_i^{2t}(\alpha_i - \alpha'_i)^2 \,,$$

where the λ_i are the eigenvalues of P and the α_i, α'_i are components of μ and μ' along its eigenvectors. In particular, a very rich structure emerges, where all the eigenvalues play a role: depending on the values of $(\alpha_i - \alpha'_i)$, the sample complexity as a function of time can be driven by terms of order λ_i^{2t} for any i, and does not only depend on the spectral gap. At the extremes, we describe the *statistical window*, the ratio of sample complexities $n^*_{\mu,\mu',t}/n^*_{\gamma,\gamma',t}$, for pairs μ, μ' and γ, γ' of initial distributions with comparable initial total sample complexity. We show that

$$\frac{n^*_{\mu,\mu',t}}{n^*_{\gamma,\gamma',t}} \asymp \frac{n^*_{\mu,\mu',0}}{n^*_{\gamma,\gamma',0}} \left(\frac{\lambda_{[2]}}{\lambda_{[d]}}\right)^{2t} \,,$$

is governed by the ratio between the eigenvalues of the Markov chain with the largest and smallest absolute value (see Theorem 8). We illustrates these findings by deriving the sample complexity for several concrete examples of Markov chains in Section 4. All original proofs are given in the appendix.

Related Work. There has been considerable work regarding reconstruction of a signal observed through noisy channels. An area of particular interest has been information flow on (rooted) trees, where the label (color) of the root of a (possibly infinite) tree is chosen according to a discrete distribution. Each edge of the tree acts a noisy channel, given by the transition matrix of a Markov chain P. The goal is to reconstruct the signal at the root, given the information at the leaves. Sharp results are known for several cases depending on the branching factor of the tree and the eigenvalues of P (Evans et al., 2000, Mossel et al., 2003, Mossel, 2001). Although, the noise model is similar to the one we consider in this paper, the goal is significantly different—their focus is on reconstruction of a single signal from several (possibly correlated) corrupted observations. Other problems based on similar principles include population recovery (Dvir et al., 2012, Polyanskiy et al., 2017) and trace reconstruction (Levenshtein, 2001, McGregor et al., 2014, Hartung et al., 2018, Holden et al., 2018). Our setting can also shed light on problems where the sample complexity of learning problems is affected by communication or privacy constraints (Han et al., 2018, Acharya et al., 2018, Gupta et al., 2018).

A recent line of work has focused on testing whether a Markov chain P is identical to a fixed chain P' or sufficiently far from it, given a single trajectory $X_0, \ldots, X_t$ generated from P (Daskalakis et al., 2017). Their work does not assume the knowledge of P. In contrast, for the testing problem considered in this work, it is easy to see that in the absence of some knowledge of P, the testing problem is impossible, e.g. one can easily construct pairs of starting distribution and transition matrix, (μ, P) and (μ', P'), such that the distributions μP and $\mu' P'$ are identical. Also, there is little to be gained by observing the trajectory in our setting as all the relevant information is contained in the first observation of the trajectory. Other statistical problems based on the observations from a Markov Chain include learning a graphical model from Glauber dynamics (Bresler et al., 2014), the mixing time (Hsu et al., 2015), or the entropy rate (Kamath and Verdú, 2016). The notions of statistical distances in relation with Markov chains, in particular with respect to their stationary distributions have also been explored in (Bresler and Nagaraj, 2017). The problem of how initial information is lost with the action of a Markov chain is also considered in (Goldfeld et al., 2018), who study the case of Glauber dynamics in the context of information storage.

Notation. Throughout the paper, δ is the probability of error and $\varepsilon \in (0, 1)$ is a measure of *non-degeneracy* of distributions. The notation $\asymp$ indicates equality up to factors that may depend only on δ and ε. Standard notions from information theory which we use are defined in Appendix

2 PROBLEM DESCRIPTION

Let $\mathcal{X}$ be a set of size d, and P the transition matrix of a known irreducible reversible Markov chain on $\mathcal{X}$. We observe $X_1, \ldots, X_n$ that are i.i.d. draws from this Markov chain after $t \geq 0$ steps, with an unknown *initial distribution* ν. The distribution of the X_i is therefore $\nu_t = \nu P^t$. Our objective is to determine, given two distributions μ or μ' on $\mathcal{X}$, which one of these is the initial distribution ν, based on the observation of a sample $(X_i)_{i\in[n]}$. This is equivalent to a hypothesis testing problem between $\mu_t = \mu P^t$ and $\mu'_t = \mu' P^t$, for known t and P, based on an i.i.d. sample of size n.

For any test $\psi : \mathcal{X}^n \to \{\mu, \mu'\}$, its performance is measured in terms of its *maximum probability of error*

$$\max_{\nu_0 \in \{\mu,\mu'\}} \mathbf{P}^{\otimes n}_{\nu_t}(\psi \neq \nu_0) = \mathbf{P}^{\otimes n}_{\mu_t}(\psi \neq \mu) \vee \mathbf{P}^{\otimes n}_{\mu'_t}(\psi \neq \mu') \,.$$

In this work, we analyze the *sample complexity* of this problem, i.e. the required sample size $n^*_{\mu,\mu',t}$ to have a small probability of error. Formally, for $n \gtrsim n^*_{\mu,\mu',t}$, we have that the probability of error is smaller than δ for

some test, and for $n \lesssim n^*_{\mu,\mu',t}$, that it is greater than $1/2-\delta$ for all tests, for a fixed probability $\delta \in (0, 1/4)$, up to multiplicative constants. If $\mu_t = \mu'_t$ for some t, the statistical problem is impossible and we write $n^*_{\mu,\mu',t} = \infty$.

As discussed in the introduction, the maximum probability of error is related to the total variation distance $d_{\mathsf{TV}}(\mu_t^{\otimes n}, \mu_t'^{\otimes n})$, and our analysis relies on understanding the behavior of this quantity. We analyze its growth in n using tools of mathematical statistics, and its decay in t through the lens of spectral analysis for reversible Markov chains in Section 3.3.

We use the following notion of two distributions having a bounded-likelihood ratio and state an immediate consequence below.

Definition 1 (Bounded likelihood-ratio)**.** *Two distributions μ and μ' have an ε-bounded likelihood-ratio if for all $x \in \mathcal{X}$, $\varepsilon \le \mu_x/\mu'_x \le 1/\varepsilon$.*

Proposition 2. *If μ, μ' have an ε-bounded likelihood-ratio, so do $\mu_t = \mu P^t$ and $\mu'_t = \mu' P^t$*

We make the assumption in some of our results, that the three distributions, π, the stationary distribution of the Markov chain P, and μ, μ', the initial distributions, pairwise have an ε-bounded likelihood-ratio; this is to avoid some pathological cases where the statistical complexity is very small, e.g. if μ_t and μ'_t do not have the same support, some observations suffice to solve the testing problem with probability of error 0. This assumption therefore ensures that it cannot be a trivial problem. For the Markov chains that we consider, this property is eventually satisfied for t larger than some fixed quantity, provided P is aperiodic.

Our main interest is to exhibit the statistical window phenomenon: we exhibit pairs of initial distributions for which the hypothesis testing problem has very different sample complexities. These distributions all satisfy the assumption above. Further, we show in Section 3.4 that these results can be extended to more general cases: if the distributions μ, μ' do not satisfy the bounded likelihood-ratio assumption (Definition 1), the hypothesis testing question can be rephrased in terms of distributions that do, up to a loss in multiplicative factors.

3 SAMPLE COMPLEXITY

3.1 Reversible Markov chains

In this problem, we have access to a sample of size n from either μ_t of μ'_t. It is therefore important to understand the behavior of these two distributions as t increases, and in particular how quickly they become similar, depending on the initial starting points μ and μ'. We state some basic properties of reversible Markov chains and their spectra without proof; these can be found in standard texts on Markov chains (e.g. Levin et al. (2008)). Recall that P is the transition matrix of an irreducible reversible Markov Chain on a finite state space $\mathcal{X}$ with stationary distribution π. We adopt the convention that P_{ij} is the probability of transitioning from state i to j; as P is reversible we have $\pi_i P_{ij} = \pi_j P_{ji}$ for all i, j. Denote by Π the diagonal matrix with $\Pi_{ii} = \pi_i$ and consider the matrix $Q := \Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}$. As P is reversible and π is the stationary distribution, $Q_{ij} = \sqrt{\frac{\pi_i}{\pi_j}} P_{ij} = \sqrt{\frac{\pi_j}{\pi_i}} P_{ji} = Q_{ji}$, and as a result Q is symmetric. The following proposition holds for reversible Markov Chains.

Proposition 3. *Let $\nu_1, \dots, \nu_d$ be the eigenvectors[2] and $1 = \lambda_1 > \lambda_2 \ge \lambda_3 \ge \cdots \ge \lambda_d \ge -1$ the corresponding eigenvalues of Q. Then,*
– $v_i = \Pi^{-\frac{1}{2}} \nu_i$ is a right eigenvector of P with eigenvalue λ_i; for $\lambda_1 = 1$, $v_1 = (1, \cdots, 1)^\top$.
– $u_i = \Pi^{\frac{1}{2}} \nu_i = \Pi v_i$ is a left eigenvector of P with eigenvalue λ_i; for $\lambda_1 = 1$, $u_1 = \pi$.

We consider $1 = \lambda_{[1]}, \lambda_{[2]}, \dots, \lambda_{[d]}$ as the ordering of the eigenvalues by their absolute values, i.e. $1 = |\lambda_{[1]}| \ge |\lambda_{[2]}| \ge \cdots \ge |\lambda_{[d]}| \ge 0$. It is worth pointing out that $\lambda_{[2]} = \lambda_d = -1$ is possible if and only if P is periodic with period 2, i.e. the underlying graph on $\mathcal{X}$ is bipartite. It is common to consider *lazy* chains, where one considers the transition matrix $(1-q)P + q\mathbf{I}$. An eigenvalue λ of the original chain yields an eigenvalue $(1-q)\lambda + q$ for the lazy chain. In particular, for $q \ge 1/2$, all eigenvalues become non-negative. We discuss the impact of adding *laziness* to the problem of testing initial distributions in Section 3.3.

Inner Product and Norm with respect to π. As P is irreducible, $\pi_x > 0$ for every $x \in \mathcal{X}$. Thus, we can define an inner product over $\mathbf{R}^d$ as follows:

$$\langle u, u' \rangle_\pi = \sum_x \frac{u_x u'_x}{\pi_x} \tag{1}$$

We denote the associated norm as $\|u\|_\pi = \sqrt{\langle u, u \rangle_\pi}$. The following lemma states that the left eigenvectors form an orthonormal basis with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$.

Lemma 4. *The left eigenvectors $u_1, \dots, u_d$ of P form an orthonormal basis with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$. Furthermore, for any $u \in \mathbf{R}^d$ with $\sum_x u_x = 1$, we have $\langle u, \pi \rangle_\pi = 1$; in particular for $u_1 = \pi$, we have $\|\pi\|_\pi = 1$. By orthogonality, we also have that for $i \ge 2$, $\sum_x u_{i,x} = \langle u_i, \pi \rangle_\pi = 0$; henceforth without*

[2] As Q is symmetric the left and right eigenvectors are the same.

loss of generality, we will assume that they form an orthonormal *basis, i.e.* $\|u_i\|_\pi = 1$ *for each* i.

The distance $\|\mu - \mu'\|_\pi$ might seem strange at first sight. Note that $\|\mu-\mu'\|_\pi^2 = \sum_x \frac{(\mu_x-\mu'_x)^2}{\pi_x}$. Let us consider the case when one of the two distributions, say μ' is the stationary distribution π: then, it is simply the χ^2 divergence, $D_{\chi^2}(\mu, \pi)$ between μ and π. In our analysis, we compare this distance to other notions of distances between μ_t and μ'_t, to obtain guarantees on the sample complexity of this problem. We rely on the spectral properties of the transition matrix to understand the temporal evolution of an initial distribution over $\mathcal{X}$.

3.2 Guarantees for likelihood-ratio test

For this testing problem, we show guarantees for the performance of the likelihood ratio test. It is based on the log-likelihood ratio statistic L_n between μ_t and μ'_t, given by

$$L_n = \sum_{x\in\mathcal{X}} \hat{\mu}_{t,x} \log(\mu_{t,x}/\mu'_{t,x}) = \langle \ell, \hat{\mu}_t \rangle \,,$$

where $\hat{\mu}_t$ is the empirical or observed distribution of the $(X_i)_{i\in[n]}$, and $\ell_x = \log(\mu_{t,x}/\mu'_{t,x})$.

Definition 5. *The likelihood ratio test* ψ_{LR} *takes* $(X_i)_{i\in[n]} \in \mathcal{X}^n$ *as input and outputs* μ *or* μ' *such that*

$$\psi_{\mathsf{LR}} = \begin{cases} \mu & \text{if } L_n > 0\,, \\ \mu' & \text{if } L_n \leq 0\,. \end{cases}$$

The Kullback-Leibler divergences between μ_t and μ'_t are naturally associated to the quantity L_n, as its expected value under these distributions. This divergence is a statistical measure of divergence that captures well the sample complexity of the problem, and also appears in large deviations, in the description of the asymptotic behavior of $\hat{\mu}_t$. The connections between notions of "distances" between distributions and sample complexity have been extensively studied, see e.g. (Polyanskiy and Wu, 2017) and references therein, and (Berend et al., 2014).

3.3 Sample Complexity Guarantees

Theorem 6. *For two initial distributions* μ, μ'*, with* μ, μ', π *all pairwise having* ε*-bounded likelihood-ratios and* P *reversible, for some* $\varepsilon \in (0,1)$*, the likelihood-ratio test* ψ_{LR} *has probability of error less than* δ *if*

$$n \geq C(\varepsilon,\delta) / \sum_{i=2}^{d} \lambda_i^{2t} (\langle u_i, \mu\rangle_\pi - \langle u_i, \mu'\rangle_\pi)^2 \,,$$

for $C(\varepsilon,\delta) = 16\varepsilon^{-5/2}\log(1/\delta)$.

This result is not obtained by focusing on the behavior of the random variable L_n and using concentration inequalities, as is usually the case in such problems, but directly by analyzing and linking several notion of distances between the distributions $\mu_t^{\otimes n}$ and $\mu_t'^{\otimes n}$. Indeed, this allows us to understand simultaneously the growth in n and convergence in t of these distances and to give guarantees on how large n needs to be for any fixed t. We present the alternate point of view, more common in the analysis of Markov chains, in Section 3.5 below. Furthermore, using a different analysis of other measures of statistical distance between distributions, we show that this guarantee on the performance of the likelihood ratio test is optimal: up to constants, it is tight for *all* tests depending only on the observation of a sample of size n.

Theorem 7. *For two initial distributions* μ, μ'*, with* μ, μ', π *all pairwise having* ε*-bounded likelihood-ratios for some* $\varepsilon \in (0,1)$ *and* P *reversible, all tests have probability of error greater or equal to* $1/2 - \delta$ *if*

$$n \leq c(\varepsilon,\delta) / \sum_{i=2}^{d} \lambda_i^{2t} (\langle u_i, \mu\rangle_\pi - \langle u_i, \mu'\rangle_\pi)^2 \,,$$

for $c(\varepsilon,\delta) = 8\varepsilon\delta^2$.

When the sample size is smaller than this bound, no test can accurately identify the correct distribution, and significantly outperform a coin flip. Together, these results give a complete picture of the statistical complexity of the hypothesis testing problem defined in Section 2. The sample complexity of this problem is of order

$$n^*_{\mu,\mu',t} \asymp 1 / \sum_{i=2}^{d} \lambda_i^{2t} (\langle u_i, \mu\rangle_\pi - \langle u_i, \mu'\rangle_\pi)^2 \,.$$

This expression gives a very clear understanding of how the initial information is lost over time. The component of the difference between μ and μ' (seen as vectors in $\mathbf{R}^{\mathcal{X}}$) aligned with eigenvectors with eigenvalues close to 0 will be lost fast, while that along those with eigenvalues close to -1 and 1 will be retained longer. As a consequence, different pairs of initial distributions have very different statistical complexities. The results above allow us to describe exactly this phenomenon. The range of sample complexities for this problem can therefore be very large, and is governed by the spectral properties of the matrix.

Theorem 8. *For* P *reversible, there are pairs of initial distributions* μ, μ' *and* γ, γ'*, with* μ, μ', π *and* γ, γ', π *pairwise having* ε*-bounded likelihood-ratios for some* $\varepsilon \in (0,1)$*, such that*

$$\frac{n^*_{\mu,\mu',t}}{n^*_{\gamma,\gamma',t}} \asymp \frac{n^*_{\mu,\mu',0}}{n^*_{\gamma,\gamma',0}} \left(\frac{\lambda_{[2]}}{\lambda_{[d]}}\right)^{2t} .$$

In particular, if the initial statistical complexities $n^*_{\mu,\mu',0}$ and $n^*_{\gamma,\gamma',0}$ are similar, the ratio scales like $(\lambda_{[2]}/\lambda_{[d]})^{2t}$, and we refer to it as the *statistical window*. If there is an eigenvalue that is negative, we can arbitrarily increase the size of the statistical window with laziness, moving the eigenvalue closer to 0, or even make it infinite. We analyze this window for examples of Markov chains in the following section, describing as well the extremal pairs of initial distributions at the two ends of this statistical window: the hypothesis testing problems that become hard quickly, and those which are the least affected by the action of P^t. In many applications, this provides an intuitive understanding of the type of questions that become hard, for several natural random processes, through the lens of loss of information.

3.4 Guarantees without likelihood-ratio bounds

Our main message is that there *exist* pairs of distributions whose associated hypothesis testing problems have vastly different sample complexity, and in particular that this phenomenon can be exhibited by taking distributions satisfying a bounded likelihood-ratio assumption. To analyze the sample complexity for two distributions that do not satisfy this assumption, one can reduce the problem to the case of two distributions that do.

Definition 9. *For any $\eta \in (0,1)$, μ, μ' distributions on $\mathcal{X}$ and a reversible Markov chain P with stationary distribution π, we consider*

$$\beta = (\mu + \mu' + \pi)/3\,,$$

the average of μ, μ', and π. The centered versions $\tilde{\mu}$ and $\tilde{\mu}'$ are defined for any $\eta \in (0,1)$ as

$$\tilde{\mu} = (1-\eta)\mu + \eta\beta\,, \quad \tilde{\mu}' = (1-\eta)\mu' + \eta\beta\,.$$

Considering the hypothesis testing problem between $\tilde{\mu}$ and $\tilde{\mu}'$ only makes the statistical problem harder: it can be interpreted as drawing each sample point from β instead of either μ or μ', with probability η. Note that $\tilde{\mu}, \tilde{\mu}'$, and π all pairwise having $\eta/3$-bounded likelihood-ratios. Using these distributions, we generalize Theorem 6 as follows.

Theorem 10. *For two initial distributions μ, μ' and P reversible, the likelihood-ratio test ψ_{LR} has probability of error less than δ if*

$$n \geq c \log(1/\delta) / \sum_{i=2}^{d} \lambda_i^{2t} (\langle u_i, \mu\rangle_\pi - \langle u_i, \mu'\rangle_\pi)^2\,,$$

for a universal constant $c > 0$.

This result rests on the proof of Theorem 6 describing the sample complexity for the pair $(\tilde{\mu}, \tilde{\mu}')$, and controlling the difference in sample complexity with testing for the pair (μ, μ').

However, the sample complexity could be much smaller: if μ and μ' have different supports, if this property still holds for μ_t and μ'_t, the sample complexity can be of the order of a constant. Outside of such degenerate cases, if they have full supports, the lower bound of Theorem 7 can be recovered up to a multiplicative factor of the maximum of $\pi_x/\mu'_{t,x}$ - which is always finite, by following the same proof. Both upper and lower bounds are therefore valid up to constants if one of the two distributions is π, and the other has the same support. This last case allows also to showcase the full width of the statistical window, by taking distributions μ such that $\mu - \pi$ is aligned along different left eigenvectors of P.

3.5 Statistical time guarantees

Our results are presented in a fixed time, fixed probability setting, and we give guarantees in terms of how large the sample size n needs to be. However, in most of the literature on Markov chains, there is no sample size, and results are given in terms of guarantees on the time t. Some of our results can be formulated in a similar manner: given a fixed sample size n, what is the *statistical time* $t^*_{\mu,\mu',n}$ such that the testing problem is possible when $t \leq t^*_{\mu,\mu',n}$ and becomes impossible when $t \geq t^*_{\mu,\mu',n}$, up to terms involving only δ and ε. There is no general expression for this statistical time, however it can be made explicit for many pairs of initial distributions μ, μ', also a direct consequence of Theorem 6 and 7 (cf. Fig. 2).

Theorem 11. *For any reversible Markov chain P, and for every $i \in [d]$, there exists a pair of initial distributions μ, μ' such that*

$$t^*_{\mu,\mu',n} = t_{[i],n} \asymp \frac{1}{2} \frac{\log(n/n_0)}{\log(1/\lambda_{[i]})}\,,$$

for some initial sample complexity n_0, i.e. the sample complexity to distinguish μ and μ' without the action of P.

Yet again, this establishes that there is not only one important time, such as mixing time, describing the loss of information in this problem, but that this can happen at different timescales; alternatively, the loss of information is not only driven by the spectral gap, but may depend on all the eigenvalues. Our results rely on the reversibility of the Markov chain, as it is expressed through its spectral properties.

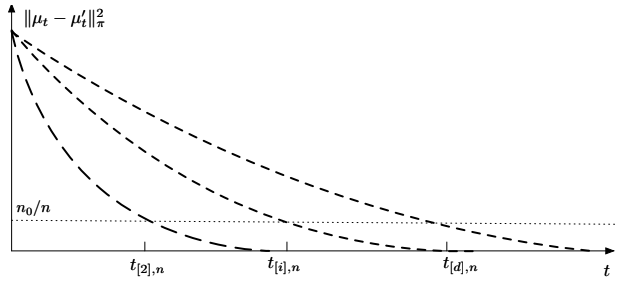

Figure 2: For different cases of starting distributions, the time at which the squared distance between the two distributions is much smaller than n_0/n, and the hypothesis testing problem becomes statistically impossible, can vary greatly.

4 APPLICATIONS

In this section, we consider several concrete examples of Markov Chains on commonly studied graph topologies. Additional examples are included in Appendix

Let $u_{[2]}$ and $u_{[d]}$ denote the left eigenvectors corresponding to the second largest and smallest eigenvalues when ordered by absolute values, i.e. $\lambda_{[2]}$ and $\lambda_{[d]}$, respectively. For a sufficiently small α, we can consider the pairs of distributions given by (μ, μ') and (γ, γ'), where $\mu = \pi + \alpha u_{[2]}$, $\mu' = \pi - \alpha u_{[2]}$, $\gamma = \pi + \alpha u_{[d]}$ and $\gamma' = \pi - \alpha u_{[d]}$; α is chosen to be small enough so that $\mu, \mu', \gamma, \gamma'$ are all valid probability distributions and μ, μ', π pairwise satisfy the ε-bounded likelihood condition, as do the distributions γ, γ', π. Note that by Theorems 6 and 7, the sample complexity of distinguishing between the pairs (μ, μ') and (γ, γ') is the same up to factors depending only on ε and δ—in each case it is given by $1/4\alpha^2 = \|\mu - \mu'\|_\pi^{-2} = \|\gamma - \gamma'\|_\pi^{-2}$. However, when considering the statistical complexity of distinguishing between (μ_t, μ'_t), the sample complexity grows as $\lambda_{[2]}^{-2t}$, whereas for distinguishing between (γ_t, γ'_t) it grows as $\lambda_{[d]}^{-2t}$. Thus, despite having roughly the same sample complexity at time $t = 0$, at a later time the ratio of sample complexities grows to be as large as $(\lambda_{[2]}/\lambda_{[d]})^{2t}$. In each of the concrete examples below, we discuss lower bounds on $\lambda_{[2]}$ and upper bounds on $\lambda_{[d]}$ and the ratio of these bounds gives a lower bound on the statistical window. In some of the examples, we explicitly derive the eigenvectors associated with these eigenvalues and discuss the associated pairs of initial distributions.

4.1 Random Walk on a Bipartite Clique

We start with the simple example of a random walk on a bipartite clique. Let $\mathcal{X} = L \cup R$ be an equipartition and let $E = \{\{x, x'\} | x \in L, x' \in R\}$, then the transition matrix P is given by:

$$\mathbf{P}[X_s = x | X_{s-1} = x'] = 2/d \qquad \text{if } \{x, x'\} \in E$$

The transition matrix P has exactly two non-zero eigenvalues, 1 and -1; thus $\lambda_{[2]} = -1$ and $\lambda_{[d]} = 0$. The eigenvector corresponding to the eigenvalue -1 has negative entries on one side of the bi-partite graph and positive entries on the other. Thus, the component of this eigenvector in the distribution controls the imbalance of the distribution between the two sides. If the initial distributions satisfy $\mu(L) = \mu'(L)$, they become indistinguishable after just one step of the Markov Chain. On the other hand, the difference $|\mu(L) - \mu'(L)|$ remains unaffected by the Markov chain. Thus, the problem at any time $t \geq 1$ remains exactly as hard as the problem at time $t = 1$, i.e. all initial information except for the *starting side* is lost in exactly one time step. So the *statistical window* is infinite in this case, for $t \geq 1$.

4.2 Random Walk on the Cycle

Definition 12 (Random Walk on the d-Cycle)**.** *Let $\mathcal{X} = \{0, 1, \ldots, d-1\}$ be the d nodes of the cycle. Let P be the Markov chain on $\mathcal{X}$, where,*

$$\mathbf{P}(X_s = i | X_{s-1} = j) = \begin{cases} \frac{1}{2} & \text{if } i \equiv j \pm 1 (\text{mod } d) \\ 0 & \text{otherwise} \end{cases}$$

The spectral properties of P are well known (see e.g. (Levin et al., 2008, Chap. 12.3)); we summarize them in the following lemma.

Lemma 13. *For any $d \geq 3$, the eigenvalues of P are given by $\cos(2\pi i/d)$ for $i \in \{0, \ldots, d-1\}$; the (right and left) eigenvector $u_i = (u_{i,0}, \ldots, u_{i,d-1})$ corresponding to eigenvalue $\cos(2\pi i/d)$, is given by $u_{i,k} = \cos(2\pi ik/d)$.*

Let us first consider a cycle of length d with $d \equiv 0 (\text{mod } 4)$. In this case, $\lambda_{[2]} = -1$ and $\lambda_{[d]} = 0$. The associated eigenvectors, $u_{[2]}$ and $u_{[d]}$, are (up to scaling) given by: $u_{[2]}(i) = 1$ for even i and -1 for odd i; $u_{[d]}(i) = 1$ for $i \equiv 0 (\text{mod } 4)$, -1 for $i \equiv 2 (\text{mod } 4)$ and 0 for $i \equiv 1 (\text{mod } 2)$. Now consider the pairs μ, μ', where $\mu = \pi + \alpha v_{[2]}$ and $\mu' = \pi - \alpha v_{[2]}$, and γ, γ', where $\gamma = \pi + \alpha v_{[d]}$ and $\gamma' = \pi - \alpha v_{[d]}$. In the first case, it is easy to see that the probability mass on *odd* and *even* nodes is noticeably different under μ and μ', and this will remain so in perpetuity. On the other hand, starting from γ or γ', stationarity is achieved in one step. Observe that in this case $n^*_{\mu,\mu',0} \asymp n^*_{\gamma,\gamma',0}$, so initially the two problems are roughly equally hard; however, as t increases (in the simple case of cycle

k	ℓ	γ_k	$\|.\|_\pi$ orthonormal eigenvectors
$k=1$		$\gamma_1 = \beta_0+\beta_1+\beta_2+\beta_3 = \lambda_1 = 1$	$v_1 = (1,1,1,1,1,1,1,1)/8$
$k=2$	$\ell=3$	$\gamma_2 = \beta_0+\beta_1+\beta_2-\beta_3 = \lambda_2$	$v_2 = (1,1,1,1,-1,-1,-1,-1)/8$
$k=3$	$\ell=2$	$\gamma_3 = \beta_0+\beta_1-\beta_2 = \lambda_3 = \lambda_4$	$v_3 = (1,1,-1,-1,0,0,0,0)/2^{5/2}$ $v_4 = (0,0,0,0,1,1,-1,-1)/2^{5/2}$
$k=4$	$\ell=1$	$\gamma_4 = \beta_0-\beta_1 = \lambda_5 = \lambda_6 = \lambda_7 = \lambda_8$	$v_5 = (1,-1,0,0,0,0,0,0)/4$ $v_6 = (0,0,1,-1,0,0,0,0)/4$, etc.

Table 1: Eigenvalues and eigenvectors of the Pachinko random walk for $r = 3$.

lengths being multiples of 4, even for $t = 1$) the difference between the statistical hardness of these problems differs dramatically. This behavior will be approximately replicated with cycles of any length provided d is large enough; in particular we always have $|\lambda_{[2]}| = 1 - O(1/d^2)$ and $\lambda_{[d]} = O(1/d)$, and so the statistical window is of size d^{2t}.

4.3 Pachinko random walk

We introduce the following random walk inspired by the Japanese pinball game of *Pachinko*, on $\mathcal{X} = [d]$ where $d = 2^r$ and the space $\mathcal{X}$ is understood as the leaves of a dyadic tree of height r. It allows to further illustrate the statistical window phenomenon.

Definition 14 (Pachinko Random Walk)**.** *Let P be the Markov chain on the $d = 2^r$ leaves of a dyadic tree such that for two leaves i and j with first common ancestor at height ℓ between 0 and r, we have*

$$\mathbf{P}(X_s = i | X_{s-1} = j) = p_\ell = \beta_\ell / 2^{\ell-1}\,,$$

where $\beta_0 > \ldots > \beta_r$ are positive real numbers that sum to 1.

This can be understood as a random walk on a graph, with a large amount of structure, that can be a consequence of an underlying geometry: at every level, each half of the vertices is "very far" from the other half, and jumping from one half to the other is less probable than staying in the same half.

Proposition 15. *For each k between 2 and $r+1$, there exists an eigenvalue γ_k of multiplicity 2^{k-2}, associated to the height $\ell = r+2-k$. It is given by $\gamma_k = \beta_0 + \ldots + \beta_{r+1-k} - \beta_{r+2-k} = \lambda_i$, for $2^{k-2}+1 \le i \le 2^{k-1}$. The 2^{k-2} vectors associated to the 2^{k-2} nodes at height $\ell = r+2-k$ - with coefficients equal to 1 for their left descendants, -1 for their right descendants, and 0 otherwise - are eigenvectors with eigenvalue γ_k. For $k = 1$, $\gamma_1 = \lambda_1 = 1$ is an eigenvalue with associated simplex eigenvector $\pi = \mathbf{1}/d$. The eigenvalues satisfy $\gamma_1 > \gamma_2 > \ldots > \gamma_{r+1} > 0$.*

This situation is summarized in the case $r = 3$ of Figure 3 in Table 1.

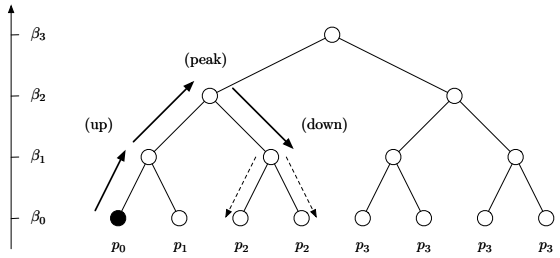

Figure 3: Analogously to the game of Pachinko, a ball starts at a leaf, goes *(up)* in the dyadic tree, *(peaks)* with probability β_ℓ at height ℓ, and goes *(down)* on the other side of the highest point, going left or right independently with probability 1/2 at each further descendant node on the way down, thus stopping uniformly at random in one of the $2^{\ell-1}$ leaves. In this figure $r = 3$, and we represent a trajectory that peaks at height $\ell = 2$.

As a consequence, if between height $\ell = r+1-k$ and $\ell+1 = r+2-k$, there is a large gap between the probabilities β_ℓ and $\beta_{\ell+1}$ (i.e., it is much harder for a particle to "jump" to height $\ell+1$ than to height ℓ), we have that

$$\gamma_k - \gamma_{k+1} = 2\beta_\ell - \beta_{\ell+1} > \beta_\ell - \beta_{\ell+1} > 0\,,$$

which implies a large gap between the eigenvalues of eigenvectors associated to height superior to ℓ and those associated to a height less or equal to ℓ. From a statistical point of view, in light of Theorem 6 and similarly to Theorem 8, this implies that difference between μ and μ' which is observable at height $\ell+1$ or above, i.e. difference of mass between two sides of nodes at these height) will be statistically observable for much larger t than difference at lower levels.

Acknowledgments

This work was supported in part by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating rényi entropy. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1855–1869. SIAM, 2014.

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Distributed simulation and distributed inference. 04 2018. URL https://arxiv.org/abs/1804.06952.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269, 2000.

D. Berend, P. Harremoës, and A. Kontorovich. Minimum kl-divergence on complements of l_1 balls. *IEEE Transactions on Information Theory*, 2014.

Q. Berthet and J. Ellenberg. Detection of planted solutions for flat satisfiability problems. *AIStats 2019*, 2019.

Quentin Berthet. Optimal testing for planted satisfiability problems. *Electron. J. Stat.*, 2014.

Guy Bresler and Dheeraj M. Nagaraj. Stein's method for stationary distributions of markov chains and application to ising models. 12 2017. URL https://arxiv.org/abs/1712.05743.

Guy Bresler, David Gamarnik, and Devavrat Shah. Learning graphical models from the glauber dynamics. *2014 52nd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2014*, 10 2014.

Clément Canonne. A survey on distribution testing. Technical report, Columbia University, 2017.

Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1380–1394. Society for Industrial and Applied Mathematics, 2013.

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*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics. ISBN 978-1-611973-38-9.

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A Servedio. Learning k-modal distributions via testing. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1371–1385. SIAM, 2012.

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A Servedio. Learning poisson binomial distributions. *Algorithmica*, 72(1):316–357, 2015.

Constantinos Daskalakis, Nishanth Dikkala, and Nick Gravin. Testing symmetric markov chains from a single trajectory. Arxiv preprint, 2017.

Ilias Diakonikolas. Learning structured distributions. *Handbook of Big Data*, 267, 2016.

Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. *FOCS 2016*, 01 2016.

Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. *SODA 2015*, 2014.

Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1183–1202. IEEE, 2015.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. *Arxiv preprint*, 2017a.

Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Near-optimal closeness testing of discrete histogram distributions. *Arxiv preprint*, 2017b.

Zeev Dvir, Anup Rao, and Avi Wigderson. Restriction access. In ACM, editor, *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 19–33, 2012.

William Evans, Claire Kenyon, Yuval Peres, and Leonard J Schulman. Broadcasting on trees and the ising model. *Annals of Applied Probability*, pages 410–433, 2000.

Z. Goldfeld, G. Bresler, and Y. Polyanskiy. Information storage in the stochastic ising model. *Preprint*, 2018.

Samarth Gupta, Gauri Joshi, and Osman Yağan. Active distribution learning from indirect samples. 08 2018. URL https://arxiv.org/abs/1808.05334.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. 02 2018. URL https://arxiv.org/abs/1802.08417.

Lisa Hartung, Nina Holden, and Yuval Peres. Trace reconstruction with varying deletion probabilities. *2018 Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, 2018.

Nina Holden, Robin Pemantle, and Yuval Peres. Subpolynomial trace reconstruction for random strings

and arbitrary deletion probability. *Arxiv preprint*, 2018.

Daniel Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. *roceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

Sudeep Kamath and Sergio Verdú. Estimation of entropy rate and rényi entropy rate for markov chains. *Information Theory (ISIT), 2016 IEEE International Symposium on*, 2016.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1066–1100, 2015.

Pierre-Simon Laplace. *Théorie Analytique des Probabilités.* Courcier, Paris, 1812.

V. I. Levenshtein. Efficient reconstruction of sequences. *IEEE Trans. Inf. Theor.*, 47(1):2–22, 2001.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

Andrew McGregor, Eric Price, and Sofya Vorotnikova. Trace reconstruction revisited. In Andreas S. Schulz and Dorothea Wagner, editors, *Algorithms - ESA 2014*, pages 689–700, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg. ISBN 978-3-662-44777-2.

Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, pages 285–300, 2001.

Elchanan Mossel, Yuval Peres, et al. Information flow on trees. *The Annals of Applied Probability*, 13(3): 817–844, 2003.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. 2017.

Yury Polyanskiy, Ananda Theertha Suresh, and Yihong Wu. Sample complexity of population recovery. *Proceedings of the 2017 Conference on Learning Theory (COLT)*, 2017.

Gregory Valiant and Paul Valiant. Estimating the unseen: An n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 685–694. ACM, 2011.

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

Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 142–155. ACM, 2016.

Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theor.*, 2016a.

Yihong Wu and Pengkun Yang. Sample complexity of the distinct elements problem. *Arxiv preprint*, 2016b.