On The Memory Complexity of Uniformity Testing

Tomer Berg   TOMERBERG@MAIL.TAU.AC.IL
School of Electrical Engineering, Tel Aviv University

Or Ordentlich   OR.ORDENTLICH@MAIL.HUJI.AC.IL
School of Computer Science and Engineering, Hebrew University of Jerusalem

Ofer Shayevitz   OFERSHA@ENG.TAU.AC.IL
School of Electrical Engineering, Tel Aviv University

Abstract

In this paper we consider the problem of uniformity testing with limited memory. We observe a sequence of independent identically distributed random variables drawn from a distribution p over $[n]$, which is either uniform or is ε-far from uniform under the total variation distance, and our goal is to determine the correct hypothesis. At each time point we are allowed to update the state of a finite-memory machine with S states, where each state of the machine is assigned one of the hypotheses, and we are interested in obtaining an asymptotic probability of error at most $0 < \delta < 1/2$ uniformly under both hypotheses. The main contribution of this paper is deriving upper and lower bounds on the number of states S needed in order to achieve a constant error probability δ, as a function of n and ε, where our upper bound is $O(\frac{n \log n}{\varepsilon})$ and our lower bound is $\Omega(n + \frac{1}{\varepsilon})$. Prior works in the field have almost exclusively used collision counting for upper bounds, and the Paninski mixture for lower bounds. Somewhat surprisingly, in the limited memory with unlimited samples setup, the optimal solution does not involve counting collisions, and the Paninski prior is not hard. Thus, different proof techniques are needed in order to attain our bounds.

Keywords: Memory complexity, uniformity testing

1. Introduction

During the past couple of decades, distribution testing has developed into a mature field. One of the most fundamental tasks in this field is the problem of *uniformity testing*. In this problem, the objective is to design a sample-optimal algorithm that, given i.i.d samples from a discrete distribution p over $[n]$ and some parameter $\varepsilon > 0$, can distinguish (with probability $0 < \delta < 1/2$) between the case where p is uniform and the case where p is ε-far from uniform in total variation. The common benchmark for characterizing the statistical hardness of uniformity testing is the *minimax sample complexity*, namely the number of independent samples from the distribution one needs to see in order to guarantee that the expected $0 - 1$ loss is at most δ. However, as the amount of available data is constantly increasing, collecting enough samples for accurate estimation is becoming less of a problem, and it is the dedicated computational resources that constitute the main bottleneck on the performance that a testing algorithm can attain.

As a result, the topic of testing under computational constraints is currently drawing considerable attention, and in particular, the problem of testing under memory constraints has been recently studied in various different setups, as we elaborate in Section 2. In order to single out the effects limited memory has on our algorithms, we let the number of samples we process be arbitrarily large.

Perhaps surprisingly, the bounded memory problem does not become trivial in the unbounded sample regime. Moreover, the optimal solution to the problem without memory constraints does not translate to the memory constrained case, as we shall later see.

Let us formally define the problem of interest: $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables drawn from a distribution p over $[n]$, which is either uniform or is ε-far from uniform under the total variation distance. Let $\mathcal{H}_0$ denote the hypothesis $p = u$, and $\mathcal{H}_1$ the composite hypothesis $p \in \Theta$, where $\Theta \triangleq \{q : d_{\mathsf{TV}}(q, u) > \varepsilon\}$, where u is the uniform distribution over $[n]$. An *S-state uniformity tester* consists of two functions: f, and d, where $f : [S] \times [n] \to [S]$ is a state transition (or memory update) function, and $d : [S] \to \{\mathcal{H}_0, \mathcal{H}_1\}$ is a decision function. Letting M_t denote the state of the memory at time t, this finite-state machine evolves according to the rule

$$M_0 = s_{\text{init}}, \tag{1}$$
$$M_t = f(M_{t-1}, X_t) \in [S], \tag{2}$$

for some predetermined initial state $s_{\text{init}} \in [S]$. If the machine is stopped at time t, it outputs the decision $d(M_t)$. Note that here the memory update function f is not allowed to depend on time. The restriction to time-invariant algorithms is operationally appealing, since storing the time index necessarily incurs a memory cost. Furthermore, since the number of samples is unbounded, simply storing the code generating a time-varying algorithm may require unbounded memory. We define the asymptotic expected $0 - 1$ loss under each hypothesis (also known as the type-1 and type-2 probabilities of error) as

$$\mathsf{P_e}(f, d|\mathcal{H}_0) = \lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} \Pr(d(M_i) = 1 | X_1, X_2, \ldots \overset{iid}{\sim} u), \tag{3}$$

$$\mathsf{P_e}(f, d|\mathcal{H}_1) = \sup_{p \in \Theta} \lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} \Pr(d(M_i) = 0 | X_1, X_2, \ldots \overset{iid}{\sim} p). \tag{4}$$

We are interested in the *minimax memory complexity* of the tester, $S^*(n, \varepsilon)$, defined as the smallest integer s for which there exist (f, d) such that $\mathsf{P_e}(f, d|\mathcal{H}_i) \leq \delta$, for any $i \in \{0, 1\}$. In previous works dealing with memory limited testing/estimation problems, the memory complexity was taken to be either the number of bits or the number of memory words needed to solve the problem, up to multiplicative constants. In those cases, a large gap between, for example, $S = n$ and $S = n^2$ states, would correspond only to a factor of 2 in the number of bits we need to store. These definitions of memory complexity are natural in the large memory regime. However, in this work, we consider the small memory regime, so the definition in terms of the number of states seems more appropriate to us. In such regime, one might not be able to increase the number of available memory bits even by a constant factor. In other words, we are interested in deriving additive constant bounds on the number of bits stored, rather than constant factor bounds.

The sample complexity of uniformity testing is well known, as a result of the works of Goldreich and Ron (2000) and Paninski (2008), and is $\Theta(\sqrt{n}/\varepsilon^2)$. The lower bound is established using χ^2 divergences, where the input is drawn from the Paninski mixture. The upper bound is based on collision estimators, since $\sqrt{n}$ samples will result in 1 expected collision on average under the uniform distribution, and $1 + \Omega(\varepsilon^2)$ expected collisions on average under the alternative, and the

2

variances are small enough to distinguish between the hypotheses. Collision estimators have henceforth become the default algorithmic tool whenever uniformity testers are involved, and most upper bounds in the field rely on some variant of the collision estimator. In a similar fashion, the Paninski prior has become a recurring ingredient in lower bounds for the problem, since it is difficult to identify with limited samples. As it turns out, when memory is the more stringent restriction, these approaches are lacking. This can be observed in a recent work of Diakonikolas et al. (2019) that is concerned with the sample complexity under memory constraints of uniformity testing, and in which the memory $S >> n^{1/\varepsilon^2}$ considered by the authors is substantially greater than what our results imply, even though the authors work with the more lenient streaming algorithms, that allow for time-varying functions. The increased required memory in Diakonikolas et al. (2019) mainly stems from the fact that the algorithm considered there is based on collision estimation, which is more wasteful in general. On the other hand, the sample complexity attained by that algorithm is much smaller than the number of samples the algorithm proposed here requires, but this reduced sample size comes at the price of increased memory (we discuss this in Section 4).

Our main results are twofold. The first is showing that reduction of the uniformity testing problem to sequential binary hypothesis testing results in a substantial saving in memory, with respect to collision estimation. The performance of a scheme based on this idea is analyzed, and gives rise to an upper bound on the memory complexity (Theorem 1). The second result consists of two lower bounds on the memory complexity. For the first lower bound, we use a member of the Paninski prior family to show that $S^*(n, \varepsilon) = \Omega(1/\varepsilon)$. However, even for the full Paninski mixture, we show that $O(1/\varepsilon)$ states are sufficient for reliably testing against the uniform distribution, when the number of available samples is large. Consequently, we must choose a different approach for our $\Omega(n)$ lower bound. Through analysis of stationary distributions of Markov processes, we show that for any uniformity tester with insufficient memory there is some distribution q with $d_{\mathsf{TV}}(q, u) > \varepsilon$ which induces the same stationary distribution on its memory states as the uniform distribution, hence is indistinguishable from it (Theorem 2). This is in contrast to the standard approach of finding a suitable prior on Θ under which the Bayesian problem is hard.

Theorem 1

$$S^*(n, \varepsilon) = O\left(\frac{n \log n}{\varepsilon}\right). \tag{5}$$

It is worth noting that the algorithm achieving this upper bound is deterministic. Our lower bound holds for randomized algorithms.

Theorem 2

$$S^*(n, \varepsilon) = \Omega\left(n + \frac{1}{\varepsilon}\right). \tag{6}$$

We note that both bounds hold also for the hypothesis testing problem of the collision probability, in which we test between collision probability that is either $\frac{1}{n}$ or greater than $\frac{1+\varepsilon^2}{n}$.

2. Related work

The study of learning and estimation under memory constraints has been initiated in the late 1960s by Cover and Hellman (with a precursor by Robbins (1956)) and remained an active research area

for a decade or so. Most of the old work on learning with finite memory has been focused on the hypothesis testing problem. For the problem of deciding whether an i.i.d. sequence was drawn from Bern(p) or Bern(q), Cover et al. (1969) described a time-varying finite-state machine with only $S = 4$ states, whose error probability approaches zero with the sequence length. As time-varying procedures suffer from the shortcomings described earlier, Hellman and Cover (1970) addressed the same binary hypothesis testing problem within the class of *time-invariant randomized* procedures. They have found an *exact* characterization of the smallest attainable error probability for this problem. In a recent paper, Berg et al. (2020) derived a lower bound on the error probability attained by any S-state deterministic procedure, showing that while the smallest attainable error probability decreases exponentially fast with S in both the randomized and the deterministic setups, the base of the exponent can be arbitrarily larger in the randomized case.

The work on the uniformity testing problem has been initiated by Goldreich and Ron (2000), who proposed a simple and natural uniformity tester that relies on the *collision probability* of the unknown distribution, which is the probability that two samples drawn according to p are equal, and succeeds after drawing $O(\sqrt{n}/\epsilon^4)$ samples. In subsequent work, Paninski (2008) showed an information-theoretic lower bound of $\Omega(\sqrt{n}/\epsilon^2)$ on the sample complexity and provided a matching upper bound that holds under some assumption on ε, which has been later shown to be unnecessary Diakonikolas et al. (2014). It was recently shown in Goldreich (2016) that the more general problem of distribution testing can be reduced to uniformity testing with only a constant factor loss in sample complexity. In Acharya et al. (2019), the authors addressed the distributed variant of the problem: each player receives one sample from the distribution, about which they can only provide limited information to a central referee.

The space-complexity (which is the minimal memory in bits needed for the algorithm) of estimating the empirical collision probability is well-studied for worst case data streams of a given length, dating back to the seminal work of Alon et al. (1999). In Crouch et al. (2016) the authors studied the trade-off between sample complexity and space complexity for the problem of estimating the collision probability. The trade-off between sample complexity and space/communication complexity for the distribution testing problem has recently been addressed by Diakonikolas et al. (2019), who gave upper and lower bounds on the sample complexity under memory constraints. On a broader level, the problem of estimating statistics with bounded memory is receiving more and more attention in the machine learning literature lately, see, e.g., Chien et al. (2010); McGregor et al. (2012); Kontorovich (2012); Steinhardt and Duchi (2015); Steinhardt et al. (2016); Raz (2018); Dagan and Shamir (2018); Dagan et al. (2019); Sharan et al. (2019). Another closely related active line of work is that of estimating statistics under limited communication, e.g., Zhang et al. (2013); Garg et al. (2014); Braverman et al. (2016); Xu and Raginsky (2017); Jordan et al. (2018); Han et al. (2018a,b); Barnes et al. (2018); Acharya et al. (2018); Hadar et al. (2019); Hadar and Shayevitz (2019); Acharya et al. (2020).

3. Preliminaries

In this section, we introduce the mathematical notation and background necessary to state and prove our results in the following sections.

3.1. Notations

We write $[n]$ to denote the set $\{1, \ldots, n\}$, and consider discrete distributions over $[n]$. We use the notation p_i to denote the probability of element i in distribution p, and the notation $p_i^{(i,j)}$ to denote the probability of element i restricted to (i, j), that is, $p_i^{(i,j)} = \frac{p_i}{p_i + p_j}$. The ℓ^k norm of a distribution is $||p||_k = \sqrt[k]{\sum_{i=1}^{n} |p_i|^k}$. The total variation distance between distributions p and q is defined as half their ℓ^1 distance, i.e., $d_{\mathsf{TV}}(p, q) = \frac{1}{2}||p - q||_1 = \frac{1}{2}\sum_{i=1}^{n} |p_i - q_i|$. The collision probability of a distribution p is defined as the squared ℓ_2 norm of p, that is, $\mathrm{C}_{\mathrm{oll}}(p) = ||p||_2^2 = \sum_{i=1}^{n} p_i^2$.

3.2. Equivalence to collision estimation

As we mentioned, the problem of uniformity testing with limited samples and unlimited memory is equivalent to the collision testing problem, due to the following two reasons:

1. $d_{\mathsf{TV}}(p, u) > \varepsilon$ implies $\mathrm{C}_{\mathrm{oll}}(p) > \frac{1+\varepsilon^2}{n}$, since, using the Cauchy-Schwarz inequality:

$$
\begin{aligned}
\mathrm{C}_{\mathrm{oll}}(p) &= \sum_{i=1}^{n} p_i^2 \\
&= \frac{1}{n} + \sum_{i=1}^{n} \left(p_i - \frac{1}{n} \right)^2 \\
&\geq \frac{1}{n} + \frac{1}{n} \left(\sum_{i=1}^{n} \left| p_i - \frac{1}{n} \right| \right)^2 \qquad (7) \\
&= \frac{1 + d_{\mathsf{TV}}(p, u)^2}{n}, \qquad (8)
\end{aligned}
$$

and we can solve the latter problem with $O(\sqrt{n}/\varepsilon^2)$ samples, hence we can also solve the former with the same number of samples.

2. The lower bound on the sample complexity is proven for the Paninski prior, $q^{\mathrm{Pan}}(z)$, whose value at each coordinate $i \in [n]$ is determined by the vector $z \in \{-1, 1\}^{n/2}$, and is given by

$$
q_i^{\mathrm{Pan}}(z) = \begin{cases} \frac{1 + \varepsilon z_{i/2}}{n}, & i \text{ even} \\ \frac{1 - \varepsilon z_{(i+1)/2}}{n}, & i \text{ odd.} \end{cases} \qquad (9)
$$

It is not difficult to see that $d_{\mathsf{TV}}(u, q^{\mathrm{Pan}}) = \varepsilon$ and $\mathrm{C}_{\mathrm{oll}}(q^{\mathrm{Pan}}(z)) = \frac{1+\varepsilon^2}{n}$, for all $z \in \{-1, 1\}^{n/2}$. Now taking $z_1, \ldots z_{n/2}$ to be Rademacher random variables, we obtain a distribution on Θ, for which the sample complexity in the Bayesian setting is $\Omega(\sqrt{n}/\varepsilon^2)$. This implies the same lower bound for the composite hypothesis testing problem.

4. Upper Bound

Following the above equivalence, it seem only natural to try and obtain upper bounds by utilizing collision-based estimators in our bounded memory framework as well. Naively implementing the collision counting decision rule of Goldreich and Ron (2000); Paninski (2008) requires $\Omega(n^{\sqrt{n}/\varepsilon^2})$ states. However, since the number of samples is unbounded, this algorithm is extremely wasteful.

As a warm-up, we first propose the following variation, which uses the unlimited number of samples to save memory, by only counting consecutive collisions. Note that the probability that two consecutive samples are equal under p is exactly $\sum_{i=1}^{n} p_i^2 = \mathrm{C}_{\mathrm{oll}}(p)$. We define a *collision machine* to be a machine that, at each time point, outputs a 1 if the current sample is the same as the preceding sample and 0 otherwise. This can be implemented using $O(n)$ states, since we are storing an element of the domain (the preceding sample) to check for collisions, and results in a Bernoulli process with parameter $p = \frac{1}{n}$ under the uniform distribution, or a Bernoulli process with parameter $p > \frac{1+\varepsilon^2}{n}$ under the alternative (due to (8)). Thus, if we can successfully distinguish between two i.i.d Bernoulli processes with parameters that are roughly ε^2/n apart, our uniformity tester will separate the hypotheses successfully. To that end, we appeal to Berg et al. (2021), where we defined the following machine which, given $X_1, X_2, \ldots \sim \mathsf{Bern}(\theta)$, can distinguish between the composite hypotheses $\theta < q$ and $\theta > p$, for $p > q$. [1]

Definition 3 RUNS(N, p, q) *is the machine with $N \geq 4$ states depicted in Figure 1, designed to decide between the hypotheses $\mathcal{H}_0 : \{\theta > p\}$ vs. $\mathcal{H}_1 : \{\theta < q\}$, for some $0 < q < p < 1$. The machine is initialized at state s and evolves according to the sequence of input bits $X_1, X_2, \ldots$. If the machine observes a run of $N - s$ ones before observing a run of $s - 1$ zeros, it decides $\mathcal{H}_0$, otherwise it decides $\mathcal{H}_1$. The initial state of the machine is $s = f(N, p, q)$, where*

$$f(N, p, q) \triangleq 2 + \left\lceil \frac{\log pq}{\log p(1-p) + \log q(1-q)}(N - 3) \right\rceil, \tag{10}$$

is an integer between 2 and $N - 1$. We denote the (worst case) error probability of the machine by $\mathsf{P_e}^{\mathsf{RUNS}(N,p,q)} = \max\left\{p_1^0, p_0^1\right\}$, *where*

$$p_1^0 = \sup_{\theta < q} \Pr_{X_1, X_2 \ldots \overset{i.i.d.}{\sim} \mathsf{Bern}(\theta)} \left(\mathsf{RUNS}(N, p, q) \text{ decides } \mathcal{H}_0\right), \tag{11}$$

$$p_0^1 = \sup_{\theta > p} \Pr_{X_1, X_2 \ldots \overset{i.i.d.}{\sim} \mathsf{Bern}(\theta)} \left(\mathsf{RUNS}(N, p, q) \text{ decides } \mathcal{H}_1\right). \tag{12}$$
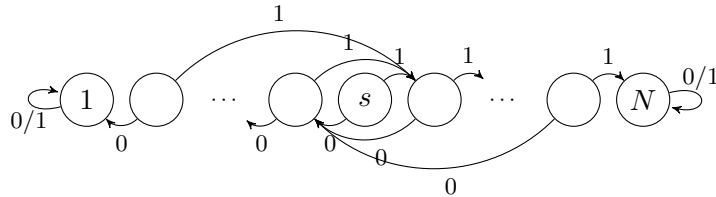

Figure 1: RUNS(N, p, q) - Deterministic Binary Hypothesis Testing Machine

In the same paper, we have shown that with $N = O(K)$ states, the machine RUNS(N, p, q) can decide whether $\theta > p$ or $\theta < q = p - 1/K$ with constant error probability $\delta < 1/2$.

---

1. Their machine was designed to solve the *simple* binary hypothesis test $\mathcal{H}_0 : \{\theta = p\}$ vs. $\mathcal{H}_1 : \{\theta = q\}$, but the difference between the two problems is not significant.

Lemma 4 (Berg et al. (2021)) *For any $\frac{2}{K} \leq p \leq 1 - \frac{1}{K}$, $q = p - \frac{1}{K}$ and $0 < \delta < 1/2$, let*

$$N = N(\delta, p, K) \triangleq 3 + \left\lceil K \cdot 6 \log \frac{2}{\delta \cdot \left(p - \frac{1}{K}\right)(1-p)} \right\rceil. \tag{13}$$

Then

$$\mathsf{P_e}^{\mathsf{RUNS}(N,p,q)} < \delta. \tag{14}$$

We can thus separate the two hypotheses with $O(n/\varepsilon^2)$ states, so overall we need the product of the two parts, which is $O(n^2/\varepsilon^2)$ states, a memory size that is *quadratic* in n/ε. We note that in order to distinguish hypotheses that are $1/K$ apart, we must have $\Omega(K)$ states [Hellman and Cover (1970), Theorem 3], therefore once the reduction to pairwise collisions has been done, using RUNS is essentially the best possible, up to constants. It is worth mentioning that the effective sample complexity of this machine (the expected number of samples needed for the algorithm to converge) is $O(n/\varepsilon^2)$, which is indeed greater then the $O(\sqrt{n}/\varepsilon^2)$ sample complexity without memory constraints. This follows since n samples will result in 1 collision on average under the uniform distribution, and $1 + \Omega(\varepsilon^2)$ collisions on average under the alternative. However, the memory used is substantially smaller than n^{1/ε^2}, the smallest considered by Diakonikolas et al. (2019). As it turns out, we can find good uniformity testers with even smaller memory. In the following, we describe a uniformity tester with a memory size that is only *linear* in n/ε, up to logarithmic factors.

4.1. Proof of Theorem 1

In a nutshell, our algorithm reduces uniformity testing (and collision estimation) to a sequence of binary tests. The intuition here is that if $d_{\mathsf{TV}}(p, u) > \varepsilon$ then there must be some symbols $(i, j) \in [n] \times [n]$ such that either $\Pr(X = i | X \in \{i, j\})$ or $\Pr(X = j | X \in \{i, j\})$ are sufficiently biased, where by sufficiently we mean with some bias $\Omega(\varepsilon)$. We could then clearly run binary tests over all n^2 pairs of symbols, but it turns out we can reduce this significantly to just $2n$ tests. One way to do this is to test, for all $i > 1$, whether $\Pr(X = 1 | X \in \{1, i\})$ or $\Pr(X = i | X \in \{1, i\})$ are sufficiently biased. If $p = u$, none of these probabilities are sufficiently biased, whereas if $d_{\mathsf{TV}}(p, u) > \varepsilon$, at least one of them is (we prove this in the sequel). Of course, this machine can get stuck for an arbitrarily long time if $\Pr(X \in \{1, i\})$ is extremely small (or even zero) for some $i > 1$, as it will be waiting for an arbitrarily low probability event. To circumvent this issue, we ensure that the algorithm terminates at some point by first checking that $\Pr(X = 1)$ is not too small. Specifically, we check that it is not much smaller than n^{-1}. The reason is since p can only be uniform, in which case all probabilities are n^{-1}, or ε-far from uniform. Thus, if $\Pr(X = 1)$ is much smaller than n^{-1}, we can confidently assert that p is not uniform. Otherwise, we continue to the previously mentioned testing of pairs.

Let us now describe the particular structure of the proposed machine. In our construction, the state space $[S]$ is partitioned into $K = 2n + 1$ disjoint sets denoted by $\mathcal{S}_1, \ldots, \mathcal{S}_K$, which we will refer to as *mini-chains*. Each mini-chain performs a binary hypothesis test, and we move sequentially from one mini-chain to the next, until we become confident of our decision, in which case we enter an absorbing state. $\mathcal{S}_1$ is a $\mathsf{RUNS}(N, p, q)$ machine for testing whether the probability of the symbol 1, namely p_1, is sufficiently large, and its parameters are $q = \frac{1}{4n}$, $p = \frac{1}{2n}$, and

$$N_1 = N\left(\frac{\delta}{2}, \frac{1}{2n}, \frac{1}{4n}\right) \triangleq 3 + \left\lceil 6 \cdot 4n \log \frac{4 \cdot 4n}{\delta \cdot \left(1 - \frac{1}{2n}\right)} \right\rceil \tag{15}$$

states. If it decides in favor of q it outputs $\mathcal{H}_1$ and terminates, whereas if it decides in favor of p it continues to $\mathcal{S}_2$. Mini-chains $\mathcal{S}_2$ to $\mathcal{S}_{2n}$ test consecutively for the conditional distributions $p_1^{(1,2)}, p_2^{(1,2)}$ up to $p_1^{(1,n)}, p_n^{(1,n)}$ between $q = \frac{1}{2} + \frac{\tilde{\varepsilon}}{2}$ and $p = \frac{1}{2} + \tilde{\varepsilon}$, for $\tilde{\varepsilon} = \frac{\varepsilon}{8+4\varepsilon}$. To that end, we use identical $\mathsf{RUNS}(N, p, q)$ machines which, when testing for the pair $(1, i)$, change their state only upon observing symbols 1 or i. Let $\mathsf{P_e} \triangleq \frac{\delta/2}{2n}$, and construct the machines to attain worst case error probability of at most $\mathsf{P_e}$. Thus, the number of states needed is, for all $2 \le j \le 2n + 1$,

$$N_j = N\left(\mathsf{P_e}, \frac{1}{2} + \tilde{\varepsilon}, \frac{\tilde{\varepsilon}}{2}\right) \triangleq 3 + \left\lceil \frac{12}{\tilde{\varepsilon}} \log \frac{4}{\mathsf{P_e} \cdot (1 + \tilde{\varepsilon})\,(1/2 - \tilde{\varepsilon})} \right\rceil \tag{16}$$

states. Here, if the machine decides in favor of q it moves to the next mini-chain, whereas if it decides in favor of p it outputs $\mathcal{H}_1$ and terminates. The construction of $\mathcal{S}_{2n+1}$ is similar, with the exception that if it decides in favor of q it outputs $\mathcal{H}_0$ and terminates. Thus, the machine only outputs $\mathcal{H}_0$ after all binary tests are completed, and none of the conditional probabilities are found to be too biased. The binary hypothesis testing machines are concatenated left to right, such that the $2i$-th machine, for $i \in [n]$, is as depicted in Figure 2.
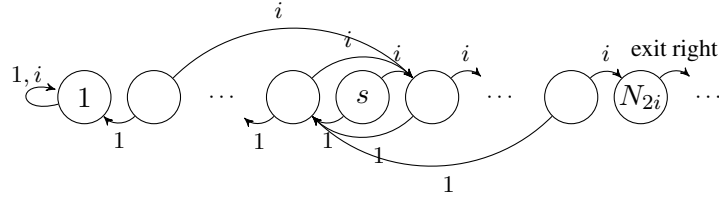

Figure 2: Illustration of mini-chain $\mathcal{S}_{2i}$

The structure of the $(2i + 1)$-th machine, for $1 \le i \le n - 1$, is similar, except that the arrow labels 1 and i are swapped, since we test for $p_i^{(1,i)}$ instead of $p_1^{(1,i)}$. The only difference in the $2n + 1$ machine is the last state, which is an absorbing one. The total number of states is:

$$S = \sum_{j=1}^{2n+1} N_j = N_1 + 2n \cdot N_2 \tag{17}$$

$$\le 4(2n+1) + 24n \cdot \log \frac{4 \cdot 4n}{\delta\left(1 - \frac{1}{2n}\right)} + \frac{24n}{\tilde{\varepsilon}} \cdot \log \frac{4 \cdot 4n}{\delta(1 + \tilde{\varepsilon})\,(1/2 - \tilde{\varepsilon})} \tag{18}$$

$$\le 4(2n+1) + 24n\left(1 + \frac{1}{\tilde{\varepsilon}}\right) \cdot \log \frac{16n}{\delta(1 + \tilde{\varepsilon})\,(1/2 - \tilde{\varepsilon})} \tag{19}$$

$$\le 9n + \frac{10 \cdot 24n}{\varepsilon} \cdot \log \frac{40n}{\delta}. \tag{20}$$

We now show that this is indeed an (S, δ) uniformity tester. Consider first the case $p = u$, which implies $p_i = \frac{1}{2}$ for all i. The probability of outputting $\mathcal{H}_1$ in $\mathcal{S}_1$ is upper bounded by $\frac{\delta}{2}$. The probability of outputting $\mathcal{H}_1$ in any of the consecutive binary tests is the probability we make an error in *at least one* binary test, and is, at most

$$1 - (1 - \mathsf{P_e})^{2n} = 1 - \left(1 - \frac{\delta/2}{2n}\right)^{2n} \le 1 - e^{-\delta/2} \le \frac{\delta}{2}. \tag{21}$$

Hence, by the union bound, the total error probability is at most δ. For the case $p \in \Theta$, we show that if $d_{\mathsf{TV}}(p, u) > \varepsilon$, then the probability of outputting $\mathcal{H}_0$ is smaller than δ. To that end, it is sufficient to show that there is some i for which $\max\{p_1^{(i,1)}, p_i^{(i,1)}\} > \frac{1}{2} + \tilde{\varepsilon}$, since that would imply the probability that either $\mathcal{S}_{2i}$ or $\mathcal{S}_{2i+1}$ output $\mathcal{H}_1$ is at least $1 - \mathsf{P_e}$, which in turn implies that the overall error probability of the algorithm is at most $\mathsf{P_e} = \frac{\delta/2}{2n} < \delta$. Thus, we prove the following lemma in the appendix.

Lemma 5 *If $d_{\mathsf{TV}}(p, u) > \varepsilon$, then there exist some $i \in [n]$ such that*

$$\max\{p_1^{(i,1)}, p_i^{(i,1)}\} > \frac{1}{2} + \tilde{\varepsilon}. \tag{22}$$

While we leave the investigation of the optimal memory-sample complexity trade off to future work, it is worth mentioning that the effective sample complexity of the above algorithm is $O\left(n^2 e^{O(1/\varepsilon)}\right)$. This follows since there are $O(n)$ mini-chains that we traverse in sequence, each is activated every $O(n)$ samples on average, and the mixing time of each one is $e^{O(1/\varepsilon)}$. Thus, if ε is taken to be a constant and n is the dominating factor, we can loosely say that, with respect to the simpler collision tester proposed earlier, a quadratic saving in memory states is translated to a quadratic increase in the number of expected samples needed for testing.

5. Lower Bound

In this section we prove Theorem 2. We prove the $\Omega\left(1/\varepsilon\right)$ bound via reduction to simple binary hypothesis testing. The proof of the $\Omega(n)$ bound is separated to the ergodic and non-ergodic cases, corresponding to irreducible and reducible machines respectively: in the first part we prove the ergodic case using linear algebra tools, and in the second part we show that any good uniformity tester that induces a non-ergodic Markov chain can be reduced to a good uniformity tester that induces an ergodic Markov chain.

5.1. Proof of the $\Omega\left(1/\varepsilon\right)$ bound

In this proof, we use the fact that a good uniformity tester is also a good binary hypothesis tester for hypotheses that are ε apart. Assume we have an (S, δ) uniformity tester. We use this machine to construct a randomized binary hypothesis testing machine that decides whether an independent identically distributed sequence of bits $B_1, B_2, \ldots$ is either $\mathsf{Bern}\left(\frac{1}{2}\right)$ distributed or $\mathsf{Bern}\left(\frac{1-\varepsilon}{2}\right)$ distributed as follows:

1. Let $U_i \stackrel{iid}{\sim} \mathsf{Unif}\left(\left[\frac{n}{2}\right]\right)$ and let $B_i \stackrel{iid}{\sim} \mathsf{Bern}(\theta)$, where θ is either $\frac{1}{2}$ or $\frac{1-\varepsilon}{2}$.

2. We generate the random variables $X_i \in [n]$ by setting $X_i = 2U_i - B_i$. Note that if $\theta = \frac{1}{2}$, then $X_i \sim u$. On the other hand, if $\theta = \frac{1-\varepsilon}{2}$, then

$$X_i \sim p = \frac{1}{n} \cdot [1 - \varepsilon \ \ 1 + \varepsilon \ \ 1 - \varepsilon \ \ 1 + \varepsilon \ \ \ldots], \tag{23}$$

   so that $d_{\mathsf{TV}}(p, u) = \varepsilon$ (p in this case is one of the priors in the Paninski family of (9)).

3. Thus, our randomized machine can distinguish between $\mathsf{Bern}\left(\frac{1}{2}\right)$ and $\mathsf{Bern}\left(\frac{1-\varepsilon}{2}\right)$. Appealing to Hellman and Cover (1970)[Theorem 3], it therefore must have $S = \Omega(1/\varepsilon)$ states.

5.2. Proof of the $\Omega(n)$ bound for ergodic machines

As we mentioned earlier, the Paninski prior is not a good choice for lower bounds in the memory limited regime, for the following reason: Any distribution in the Paninski family has either $p_1^{(1,2)} = \frac{1+\varepsilon}{2}$ or $p_2^{(1,2)} = \frac{1+\varepsilon}{2}$. Thus, we can test these two options with two $\mathsf{RUNS}(N, p = \frac{1+\varepsilon}{2}, q = p - \frac{\varepsilon}{2})$ machines with $O(1/\varepsilon)$ states, hence the Paninski mixture cannot give us a lower bound that depends on the dimension n. The reason for this is that since the number of samples is unlimited, we can afford to wait for events that sharply distinguish between the hypotheses, and, in particular relevance to the above argument, to wait for appearances of the first two symbols alone. When one cares about sample complexity only, then this approach is obviously very wasteful since many samples are ignored. But when one is concerned with memory complexity only, then waiting for rare binary events turns out to be the right approach, since reducing the dimension of the testing problem is of most importance.

We first consider the case of ergodic machines, corresponding to irreducible Markov chains. The method of proof is the following: Any finite-state machine induces a Markov chain, that is characterized by some transition matrix $\mathbf{P}$. Our method of proof is based on analysis of stationary distributions and properties of Markov chain transition matrices, specifically their null space. Firstly, we show that for any finite-state machine with $S = O(\sqrt{n})$ states, there must exist some distribution p over $[n]$ with $d_{\mathsf{TV}}(p, u) > \epsilon$ that induces the same transition matrix as u. Secondly, we tighten this result by showing that when $S = O(n)$ there is some distribution p over $[n]$ with $d_{\mathsf{TV}}(p, u) > \epsilon$ that induces the same stationary distribution as u (but not necessarily the same transition matrix), thus p is indistinguishable from u when viewed through the state sequence of the machine.

Lemma 6 *Denote the transition matrix induced by a distribution p as $\mathbf{P}(p)$. Then for any finite-state machine with $S = O(\sqrt{n})$ there must exist some distribution p with $d_{\mathsf{TV}}(p, u) > \epsilon$ such that $\mathbf{P}(u) = \mathbf{P}(p)$.*

Proof Fix a finite state machine with S states, and define the matrix $A \in [0, 1]^{S^2 \times n}$ whose rows are indexed by $(i, j) \in [S] \times [S]$. The entry $A_{(i,j),m}$ specifies the probability of moving to state j from state i when the input is $m \in [n]$. Note that for deterministic machines the values of A can only be 0 or 1, but our lemma in fact holds for the more general case of randomized machines. Let $p(u) = Au$, where $u = [1/n \ldots 1/n]^T$, be the vectorized form of the transition matrix induced on the machine by the uniform distribution, that is, $p(u)$ is generated by stacking the columns of $\mathbf{P}(u)$ on top of each other. Now, if $S^2 < n$, the matrix A must have a nontrivial kernel of rank at least $n - S^2$. Let V be the subspace spanned by $\ker(A) \cap \ker([1 \ldots 1])$. We have that $\dim(V) \geq n - 1 - S^2 \triangleq k$. Before we continue, we need the following lemma, due to David E. Speyer.[2]

Lemma 7 *Let V be a k-dimensional linear subspace of $\mathbb{R}^n$. Define its $p \to q$ norm as*

$$\|V\|_{p \to q} = \max_{x \in V \setminus \{0\}} \frac{\|x\|_q}{\|x\|_p}. \tag{24}$$

Then, for all $q \geq 1$,

$$\|V\|_{\infty \to q} \geq \sqrt[q]{k}. \tag{25}$$

---

2. https://math.stackexchange.com/questions/1850488

Proof Let Q be the hypercube $[-1, 1]^n$. Then $V \cap Q$ is a bounded nonempty polytope, thus it must have a vertex. Denote that vertex as $x = (x_1, \ldots, x_n)$. Without loss of generality, we may assume that there is some integer $0 < r < n$ such that $x_i = 1$ for all $1 \leq i \leq r$ and $|x_i| < 1$ for all $r + 1 \leq i \leq n$. This is since Q is symmetric, thus we can always rearrange the coordinates of V without affecting its norms. We claim that $r \geq k$. Assume towards contradiction that it is not the case. Then there must be some nonzero vector $v \in V$ with $v_i = 0$ for all $1 \leq i \leq r$, such that for a small enough constant ε, both $x + \epsilon v$ and $x - \varepsilon v$ would be in $V \cap Q$, contradicting that x is a vertex. Thus $\|x\|_\infty = 1$ and $\|x\|_q \geq \sqrt[q]{r} \geq \sqrt[q]{k}$. ∎

Proof of Lemma 6, continued: Lemma 7 implies that there must be a vector $x \in V$ with $\|x\|_\infty = 1/n$ and $\|x\|_1 \geq k/n$. For this x, we can define $q = u - x$. Recalling that V is orthogonal to $[1 \ldots 1]$, we have $\sum q_i = \sum u_i - \sum x_i = 1$, thus q is a valid probability distribution and $d_{\mathsf{TV}}(u, q) = \frac{1}{2}\|x\|_1 \geq k/2n$. Since $x \in \ker(A)$, we have that the transition probability q induces on the machine is

$$p(q) = Aq = A(u - x) = Au = p(u). \tag{26}$$

Thus, for any finite state machine with $S = ((1 - 2\varepsilon)n - 1)^{1/2}$ states, there exists a distribution q with $d_{\mathsf{TV}}(u, q) \geq \varepsilon$ such that $\mathbf{P}(u) = \mathbf{P}(q)$. ∎

Proof of Theorem 2, irreducible case : Recall that the stationary row vector satisfies $\pi_u = \pi_u \mathbf{P}(u)$. This can be written as $\pi_u = Mp(u) = MAu$, where $M \in [0, 1]^{S \times S^2}$ is of the form

$$M = \begin{pmatrix} \pi_u & 0 & \ldots & 0 \\ 0 & \pi_u & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \pi_u \end{pmatrix} \tag{27}$$

Denote $B = MA \in [0, 1]^{S \times n}$. Similarly to Lemma 6, we have that if $S < n$, the matrix B must have a nontrivial kernel of rank at least $n - S$. Letting V_1 be the subspace spanned by $\ker(B) \cap \ker([1 \ldots 1])$, we have that $\dim(V_1) \geq n - 1 - S \triangleq k$. Appealing to Lemma 7, there is some valid probability distribution $q = u - x$ with $d_{\mathsf{TV}}(u, q) \geq k/2n$ and $x \in \ker(B)$. We thus have that $Bq = Bu = \pi_u$. But $Bq = \pi_u$ are simply the stationary equations for the chain under q, thus $\pi_q = \pi_u$. This implies that for any finite state machine with $S = (1 - 2\varepsilon)n - 1$ states, there exists a distribution q with $d_{\mathsf{TV}}(u, q) \geq \varepsilon$ such that $\pi_u = \pi_q$. ∎

5.3. Proof of the $\Omega(n)$ bound for non-ergodic machines

We now prove the case of non-ergodic machines, corresponding to reducible Markov chains. Specifically, we show that any (S, δ) non-ergodic uniformity tester can be reduced to an $(S, \delta + \gamma)$ ergodic uniformity tester, where γ can be made arbitrarily small. We start by considering a simple non-ergodic machine with only two absorbing states.

Lemma 8 *Consider a Markov chain with initial state s and two absorbing states 0 and 1, such that the absorption probabilities are $p_\infty^j = [\Pr(s \to 0|\mathcal{H}_j), \Pr(s \to 1|\mathcal{H}_j)]$ under hypothesis $\mathcal{H}_j$. Further let $\mathsf{T}_{\mathsf{hit}}$ be the maximal expected hitting time of state i under $\mathcal{H}_j$, given that we absorb in i,*

where the maximum is over all $i, j \in \{0, 1\}$. Let $M > 0$ be some positive constant. Then for any $j \in \{0, 1\}$, the ergodic chain that results by connecting each absorbing state to s with probability $\delta = \frac{1}{M \, \mathsf{T}_{\mathsf{hit}}}$ has stationary probabilities π_0^j, π_1^j such that

$$|\pi_i^j - p_\infty^j(i)| \le \frac{1}{M+1} \cdot \max_{i \in \{0,1\}} p_\infty^j(i) \le \frac{1}{M+1}. \tag{28}$$

Proof Consider the ergodic Markov chain obtained by connecting each of the absorbing states 0 and 1 to s with probability $\delta = \frac{1}{M \, \mathsf{T}_{\mathsf{hit}}}$. The transition rules of the transient states are left unchanged. Assume w.l.o.g that the correct hypothesis is $\mathcal{H}_0$, and for simplicity let $p_\infty = p_\infty^0, \pi_i = \pi_i^0$. Appealing to Kac (1947) [Theorem 2 - Kac's return time], the stationary distribution of state $i \in \{0, 1\}$ is $\pi_i = \frac{1}{\mathbb{E}[\tau_i]}$, where $\mathbb{E}[\tau_i]$ is the expected first return time to state i. The expected return time to state 0 (under $\mathcal{H}_0$) can be upper bounded by the following argument: with probability $1 - \delta$ we return in one step, otherwise we play the following game. We flip a $\mathsf{Bern}(1 - p_\infty(0))$ coin that determines whether we end in state 0 or not. If the coin value is 1, we travel the chain at most $\mathsf{T}_{\mathsf{hit}}$ time units on average and then another δ^{-1} time units on average 'stuck' in state 1 before returning to state s and flipping the coin again. If the coin value is 0, we arrive at the desired state and terminate the game after at most $\mathsf{T}_{\mathsf{hit}}$ time units on average. This is since the limiting distribution $p_\infty(0)$ is simply the probability of arriving at state 0 before arriving at state 1.

Formally, let W denote the number of rounds in the above game, i.e., the number of times we return to s before hitting state 0. Thus $\Pr(W = m) = (1 - p_\infty(0))^{m-1} p_\infty(0)$, and furthermore, $\mathbb{E}[\tau_0 | W = m] \le \mathsf{T}_{\mathsf{hit}} + m \, \mathsf{T}_{\mathsf{hit}} + m \delta^{-1}$. We then have

$$\mathbb{E}[\tau_0] = (1 - \delta) + \delta \cdot \sum_{m=0}^{\infty} \Pr(W = m) \, \mathbb{E}[\tau_0 | W = m] \tag{29}$$

$$= (1 - \delta) + \delta \cdot p_\infty(0) \sum_{m=0}^{\infty} (1 - p_\infty(0))^m \left(\mathsf{T}_{\mathsf{hit}} + m \, \mathsf{T}_{\mathsf{hit}} + m \delta^{-1} \right) \tag{30}$$

$$= (1 - \delta) + \delta \cdot p_\infty(0) \left(\frac{\mathsf{T}_{\mathsf{hit}}}{p_\infty(0)} + \left(\mathsf{T}_{\mathsf{hit}} + \delta^{-1} \right) \sum_{m=0}^{\infty} m(1 - p_\infty(0))^m \right) \tag{31}$$

$$= (1 - \delta) + \delta \, \mathsf{T}_{\mathsf{hit}} + (\delta \, \mathsf{T}_{\mathsf{hit}} + 1) \cdot \frac{1 - p_\infty(0)}{p_\infty(0)} \tag{32}$$

$$\le 1 + \frac{\delta \, \mathsf{T}_{\mathsf{hit}} + 1 - p_\infty(0)}{p_\infty(0)} = \frac{1 + \delta \, \mathsf{T}_{\mathsf{hit}}}{p_\infty(0)}. \tag{33}$$

Substituting $\delta = \frac{1}{M \, \mathsf{T}_{\mathsf{hit}}}$ we have $p_\infty(0) - \pi_0 \le p_\infty(0) \cdot \frac{1}{M+1}$, which implies $\pi_1 - p_\infty(1) \le p_\infty(0) \cdot \frac{1}{M+1}$. Applying the above argument to $\mathbb{E}[\tau_1]$ concludes the proof. ∎

Proof of Theorem 2, reducible case: Consider a Markov chain comprised of K recurrent classes $\mathcal{R}_1, \dots, \mathcal{R}_K$, and a set $\mathcal{T}$ of transient states with initial state s. Assume w.l.o.g that $s \in \mathcal{T}$, as otherwise the machine is in fact irreducible. Let $\mathsf{T}_{\mathsf{hit}}$ be the maximum of the sum of the expected hitting time of class $\mathcal{R}_k$ and the mixing time of the class under $\mathcal{H}_j$, given that we absorb in $\mathcal{R}_k$, where the maximum is over all $k \in [K], j \in \{0, 1\}$. Motivated by the construction of Lemma 8 we connect all recurrent states (that is, all states not in $\mathcal{T}$) to s with probability $\delta = \frac{1}{M \, \mathsf{T}_{\mathsf{hit}}}$, and with probability $1 - \delta$ leave their transition rules unchanged, thus generating an ergodic Markov chain.

Assume w.l.o.g that the correct hypothesis is $\mathcal{H}_0$, and to simplify notations let $\mathsf{P_e} = \mathsf{P_e}(f, d|\mathcal{H}_0)$ and $\Pr(s \to \mathcal{R}_k) = \Pr(s \to \mathcal{R}_k|\mathcal{H}_0)$ for all $k \in [K]$. For each k, consider a chain with $|\mathcal{T}| + 2$ states, obtained from the original chain by merging the states in $\mathcal{R}_k$ and $\{\mathcal{R}_j\}_{j \neq k}$ into two respectively absorbing states. By applying Lemma 8, the stationary probability of the class $\mathcal{R}_k$ for all $k \in [K]$, has

$$|\pi(\mathcal{R}_k) - \Pr(s \to \mathcal{R}_k)| \leq \frac{1}{M+1}, \tag{34}$$

where $\pi(\mathcal{R}_k)$ is the stationary probability of the class $\mathcal{R}_k$. We now bound the stationary probability of state $i \in \mathcal{R}_k$ given class $\mathcal{R}_k$, for all $k \in [K]$. To that end, let $\tilde{\pi}_i = \frac{\pi_i}{\pi(\mathcal{R}_k)}$ for $i \in \mathcal{R}_k$, and let π_i^∞ be the limiting distribution of state i in the original chain, given that we absorb in $\mathcal{R}_k$. We show the following lemma, whose proof is relegated to the appendix.

Lemma 9 *For any $k \in [K]$ and $i \in \mathcal{R}_k$,*

$$\tilde{\pi}_i \geq \pi_i^\infty - \frac{|\mathcal{R}_k|}{M}\left(1 + \frac{1}{\pi(\mathcal{R}_k)}\right). \tag{35}$$

Let $\Pr(\mathcal{C}) = \sum_{k=1}^{K} \Pr(s \to \mathcal{R}_k)\Pr(\mathcal{C}|\mathcal{R}_k)$ be the asymptotic probability of success in the original chain, and let $\Pr(\mathcal{C}^{\mathrm{erg}})$ be the asymptotic probability of success in the new ergodic chain. Then

$$\Pr(\mathcal{C}^{\mathrm{erg}}) \geq \sum_{k=1}^{K} \pi(\mathcal{R}_k)\Pr(\mathcal{C}^{\mathrm{erg}}|\mathcal{R}_k) \tag{36}$$

$$\geq \sum_{k=1}^{K} \Pr(s \to \mathcal{R}_k)\Pr(\mathcal{C}^{\mathrm{erg}}|\mathcal{R}_k) - \sum_{k=1}^{K} \frac{1}{M+1} \tag{37}$$

$$\geq \Pr(\mathcal{C}) - \sum_{k=1}^{K} \frac{|\mathcal{R}_k|^2}{M}\left(1 + \frac{1}{\pi(\mathcal{R}_k)}\right) - \frac{K}{M+1} \tag{38}$$

$$\geq \Pr(\mathcal{C}) - \frac{S^2\left(1 + 1/\pi(\mathcal{R}_k)\right) + K}{M}, \tag{39}$$

where (37) follows from (34), (38) follows from $\Pr(\mathcal{C}^{\mathrm{erg}}|\mathcal{R}_k) = \sum_{i:d(i)=\mathcal{H}_0} \tilde{\pi}_i$ and by substituting Lemma 9, and (39) follows from $\sum_{k=1}^{K} |\mathcal{R}_k|^2 \leq S^2$. We conclude by recalling that Theorem 2 for the irreducible case implies that any irreducible machine with probability of error smaller than $1/2$ must have $\Omega(n)$ states, therefore, by choosing M to be sufficiently large, the bound

$$\mathsf{P_e}^{\mathrm{erg}} \leq \mathsf{P_e} + \frac{S^2(1 + 1/\pi(\mathcal{R}_k)) + K}{M} \tag{40}$$

implies that so should any reducible machine. ∎

Acknowledgments

References

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Distributed simulation and distributed inference. *arXiv preprint arXiv:1804.06952*, 2018.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints: Lower bounds from chi-square contraction. In *Conference on Learning Theory*, pages 3–17. PMLR, 2019.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. A geometric characterization of Fisher information from quantized samples with applications to distributed statistical estimation. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 16–23. IEEE, 2018.

Tomer Berg, Or Ordentlich, and Ofer Shayevitz. Binary hypothesis testing with deterministic finite-memory decision rules. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1259–1264. IEEE, 2020.

Tomer Berg, Or Ordentlich, and Ofer Shayevitz. Deterministic finite-memory bias estimation. In *Conference on Learning Theory*, pages 566–585. PMLR, 2021.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020. ACM, 2016.

Steve Chien, Katrina Ligett, and Andrew McGregor. Space-efficient estimation of robust statistics and distribution testing. In *ICS*, pages 251–265. Citeseer, 2010.

Thomas M Cover et al. Hypothesis testing with finite statistics. *The Annals of Mathematical Statistics*, 40(3):828–835, 1969.

Michael Crouch, Andrew McGregor, Gregory Valiant, and David P Woodruff. Stochastic streams: Sample complexity vs. space complexity. In *24th Annual European Symposium on Algorithms (ESA 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

Yuval Dagan and Ohad Shamir. Detecting correlations with little memory and communication. In *Conference On Learning Theory*, pages 1145–1198, 2018.

Yuval Dagan, Gil Kur, and Ohad Shamir. Space lower bounds for linear prediction in the streaming model. In *Conference on Learning Theory*, pages 929–954, 2019.

Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1841–1854. SIAM, 2014.

Ilias Diakonikolas, Themis Gouleakis, Daniel M Kane, and Sankeerth Rao. Communication and memory efficient testing of discrete distributions. In *Conference on Learning Theory*, pages 1070–1106. PMLR, 2019.

Ankit Garg, Tengyu Ma, and Huy L Nguyen. On communication cost of distributed statistical estimation and dimensionality. *arXiv preprint arXiv:1405.1665*, 2014.

Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 23, page 1, 2016.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 20, 2000.

Uri Hadar and Ofer Shayevitz. Distributed estimation of gaussian correlations. *IEEE Transactions on Information Theory*, 65(9):5323–5338, 2019.

Uri Hadar, Jingbo Liu, Yury Polyanskiy, and Ofer Shayevitz. Communication complexity of estimating correlations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 792–803, 2019.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory*, pages 3163–3188. PMLR, 2018a.

Yanjun Han, Ayfer Ozgur, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. *Proceedings of Machine Learning Research*, 75, 2018b.

Martin E Hellman and Thomas M Cover. Learning with finite memory. *The Annals of Mathematical Statistics*, pages 765–782, 1970.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

Mark Kac. On the notion of recurrence in discrete stochastic processes. *Bulletin of the American Mathematical Society*, 53(10):1002–1010, 1947.

Leonid Aryeh Kontorovich. Statistical estimation with bounded memory. *Statistics and Computing*, 22(5):1155–1164, 2012.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Andrew McGregor, A Pavan, Srikanta Tirthapura, and David Woodruff. Space-efficient estimation of statistics over sub-sampled streams. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, pages 273–282, 2012.

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

Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. *Journal of the ACM (JACM)*, 66(1):3, 2018.

Herbert Robbins. A sequential decision problem with a finite memory. *Proceedings of the National Academy of Sciences of the United States of America*, 42(12):920, 1956.

Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Memory-sample tradeoffs for linear regression with small error. In *Symposium on Theory of Computing (STOC)*, 2019.

Jacob Steinhardt and John Duchi. Minimax rates for memory-bounded sparse linear regression. In *Conference on Learning Theory*, pages 1564–1587, 2015.

Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory*, pages 1490–1516, 2016.

A. Xu and M. Raginsky. Information-theoretic lower bounds on Bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600, March 2017.

Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.

Appendix A. Proof of Lemma 5

Proof Assume toward contradiction that, for all i,

$$p_1^{(i,1)} \leq \frac{1}{2} + \tilde{\varepsilon} \implies p_1 \leq \frac{1 + 2\tilde{\varepsilon}}{1 - 2\tilde{\varepsilon}} \cdot p_i, \tag{41}$$

$$p_i^{(i,1)} \leq \frac{1}{2} + \tilde{\varepsilon} \implies p_i \leq \frac{1 + 2\tilde{\varepsilon}}{1 - 2\tilde{\varepsilon}} \cdot p_1. \tag{42}$$

(41) and (42) imply that

$$1 = \sum_{i=1}^{n} p_i \leq n p_1 \cdot \frac{1 + 2\tilde{\varepsilon}}{1 - 2\tilde{\varepsilon}}, \tag{43}$$

$$1 = \sum_{i=1}^{n} p_i \geq n p_1 \cdot \frac{1 - 2\tilde{\varepsilon}}{1 + 2\tilde{\varepsilon}}. \tag{44}$$

Thus we have

$$\frac{1 - 2\tilde{\varepsilon}}{1 + 2\tilde{\varepsilon}} \cdot \frac{1}{n} \leq p_1 \leq \frac{1 + 2\tilde{\varepsilon}}{1 - 2\tilde{\varepsilon}} \cdot \frac{1}{n}, \tag{45}$$

and, as a result, for all $i > 1$,

$$\left(\frac{1 - 2\tilde{\varepsilon}}{1 + 2\tilde{\varepsilon}} \right)^2 \cdot \frac{1}{n} \leq p_i \leq \left(\frac{1 + 2\tilde{\varepsilon}}{1 - 2\tilde{\varepsilon}} \right)^2 \cdot \frac{1}{n}. \tag{46}$$

Finally, we have

$$d_{\mathsf{TV}}(p, u) = \sum_{i=1}^{n} \left| p_i - \frac{1}{n} \right| \leq \max \left\{ \frac{8\tilde{\varepsilon}}{(1 - 2\tilde{\varepsilon})^2}, \frac{8\tilde{\varepsilon}}{(1 + 2\tilde{\varepsilon})^2} \right\} = \frac{8\tilde{\varepsilon}}{(1 - 2\tilde{\varepsilon})^2}, \tag{47}$$

and by substituting $\tilde{\varepsilon} = \frac{\varepsilon}{8 + 4\varepsilon}$ we arrive at $d_{\mathsf{TV}}(p, u) \leq \varepsilon$, a contradiction. Thus, there must be some i for which either $p_1^{(i,1)}$ or $p_i^{(i,1)}$ are greater than $\frac{1}{2} + \tilde{\varepsilon}$. ∎

Appendix B. Proof of Lemma 9

Proof Assume w.l.o.g that $\mathcal{R}_k = [m]$. Let π_i be the stationary distribution of state i in the ergodic chain, and for all $i \in \mathcal{R}_k$, let π_i^{∞} be the limiting distribution of state i in the original chain, given that we absorb in $\mathcal{R}_k$. Furthermore, let p_{ij} be the transition probability from state i to state j in the original chain. The stationary equations for π_j^{∞} and π_j, for any $j \in \mathcal{R}_k$, are

$$\pi_j^{\infty} = \sum_{i \in \mathcal{R}_k} \pi_i^{\infty} p_{ij}, \tag{48}$$

$$\pi_j = \sum_{i \in \mathcal{R}_k} \pi_i p_{ij}(1 - \delta) + \sum_{i \in \mathcal{T}} \pi_i p_{ij}. \tag{49}$$

Thus we have

$$\tilde{\pi}_j = \sum_{i \in \mathcal{R}_k} \tilde{\pi}_i p_{ij}(1 - \delta) + \frac{1}{\pi(\mathcal{R}_k)} \sum_{i \in \mathcal{T}} \pi_i p_{ij} \tag{50}$$

$$= \sum_{i \in \mathcal{R}_k} \tilde{\pi}_i p_{ij} + \underbrace{\frac{1}{\pi(\mathcal{R}_k)} \sum_{i \in \mathcal{T}} \pi_i p_{ij} - \delta \sum_{i \in \mathcal{R}_k} \tilde{\pi}_i p_{ij}}_{e_j}. \tag{51}$$

Denote $\tilde{\boldsymbol{\pi}} = [\tilde{\pi}_1, \ldots, \tilde{\pi}_m], \boldsymbol{\pi}^{\infty} = [\pi_1^{\infty}, \ldots, \pi_m^{\infty}], \boldsymbol{e} = [e_1, \ldots, e_m]$. Further Let $\mathbf{P}$ denote the transition matrix of the original chain restricted to $\mathcal{R}_k$, and define $\boldsymbol{d} \triangleq \tilde{\boldsymbol{\pi}} - \boldsymbol{\pi}^{\infty}$. Note that since both $\boldsymbol{\pi}^{\infty}$ and $\tilde{\boldsymbol{\pi}}$ are valid probability distributions over $[m]$, we have that $\boldsymbol{d} \perp \mathbf{1}$. Writing the above equations in matrix form, we have

$$\boldsymbol{d} = \boldsymbol{d}\mathbf{P} + \boldsymbol{e}, \tag{52}$$

$$\boldsymbol{d}(\boldsymbol{I} - \mathbf{P}) = \boldsymbol{e}. \tag{53}$$

Let $\lambda_{\min}(\boldsymbol{I} - \mathbf{P})$ be the minimal eigenvalue $\lambda \neq 0$ of $\boldsymbol{I} - \mathbf{P}$, and $\lambda_{\max}(\mathbf{P})$ be the maximal eigenvalue $\lambda \neq 1$ of $\mathbf{P}$. Thus,

$$\|\boldsymbol{e}\|_2 = \|\boldsymbol{d}(\boldsymbol{I} - \mathbf{P})\|_2 \geq \|\boldsymbol{d}\|_2 \cdot \lambda_{\min}(\boldsymbol{I} - \mathbf{P}) \tag{54}$$

$$= \|\boldsymbol{d}\|_2(1 - \lambda_{\max}(\mathbf{P})), \tag{55}$$

where (54) follows since for any matrix $\boldsymbol{A}$ operating on vector space V, its minimal eigenvalue has $\lambda_{\min} = \min_{\boldsymbol{v} \in V} \frac{\|\boldsymbol{A}\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2}$, and (55) follows since the eigenvalues of $\boldsymbol{I} - \mathbf{P}$ are $1 - \lambda_i(\mathbf{P})$ (recall that all the

eigenvalues of a Markov transition matrix has $|\lambda_i| \leq 1$). The reason we dispense with $\lambda = 0$, which is indeed an eigenvalue of $I - P$ with eigenvector $\mathbf{1}$, is that $d \perp \mathbf{1}$, hence the largest contraction it can undergo is by $\lambda_{\min}(I - P)$. On the other hand,

$$\|e\|_2 \leq m \cdot \|e\|_\infty \leq m \cdot \delta \left(1 + \frac{1}{\pi(\mathcal{R}_k)}\right), \tag{56}$$

where the first inequality is a standard norm inequality, and the second is a result of the following lemma, which we introduce without the (simple) proof:

Lemma 10 *Let $\{X_n\}$ be a stationary process over some alphabet $\mathcal{S}$. Then for any disjoint partition $\mathcal{C} \cup \mathcal{C}' = \mathcal{S}$, it holds that*

$$\Pr(X_n \in \mathcal{C}, X_{n+1} \in \mathcal{C}') = \Pr(X_n \in \mathcal{C}', X_{n+1} \in \mathcal{C}). \tag{57}$$

Appealing to the lemma with $\mathcal{C} = \mathcal{T}, \mathcal{C}' = \bigcup_{k=1}^{K} \mathcal{R}_k$, we have

$$\sum_{k=1}^{K} \sum_{j \in \mathcal{R}_k} \sum_{i \in \mathcal{T}} \pi_i p_{ij} = \sum_{k=1}^{K} \sum_{i \in \mathcal{R}_k} \sum_{j \in \mathcal{T}} \pi_i p_{ij} = \sum_{k=1}^{K} \sum_{i \in \mathcal{R}_k} \pi_i \cdot \delta \leq \delta, \tag{58}$$

since the only transition from any $i \in \mathcal{R}_k$ to $\mathcal{T}$ is transition to s with probability δ. Thus we have, for any $j \in [K]$,

$$|e_j| \leq \frac{1}{\pi(\mathcal{R}_k)} \sum_{i \in \mathcal{T}} \pi_i p_{ij} + \delta \sum_{i \in \mathcal{R}_k} \tilde{\pi}_i p_{ij} \tag{59}$$

$$\leq \frac{1}{\pi(\mathcal{R}_k)} \sum_{k=1}^{K} \sum_{j \in \mathcal{R}_k} \sum_{i \in \mathcal{T}} \pi_i p_{ij} + \delta \tag{60}$$

$$\leq \frac{\delta}{\pi(\mathcal{R}_k)} + \delta, \tag{61}$$

and (56) is established. Combining (55) and (56), and recalling that $\|d\|_2 \geq \|d\|_\infty$, we conclude

$$\|d\|_\infty \leq \frac{m \cdot \delta}{1 - \lambda_{\max}(\mathbf{P})} \left(1 + \frac{1}{\pi(\mathcal{R}_k)}\right) \tag{62}$$

$$\leq m \cdot \delta(T_{\mathrm{mix}}^k + 1) \left(1 + \frac{1}{\pi(\mathcal{R}_k)}\right) \tag{63}$$

$$\leq \frac{m}{M} \left(1 + \frac{1}{\pi(\mathcal{R}_k)}\right), \tag{64}$$

where T_{mix}^k is the mixing time of the Markov chain with transition matrix $\mathbf{P}$, (63) follows from Levin and Peres (2017)[Theorem 12.5], and (64) follows from substituting $\delta = \frac{1}{M \, \mathsf{T}_{\mathrm{hit}}}$. ∎