Deterministic Finite-Memory Bias Estimation

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

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

In this paper we consider the problem of estimating a Bernoulli parameter using finite memory. Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed Bernoulli random variables with expectation θ, where $\theta \in [0, 1]$. Consider a finite-memory deterministic machine with S states, that updates its state $M_n \in \{1, 2, \ldots, S\}$ at each time according to the rule $M_n = f(M_{n-1}, X_n)$, where f is a deterministic time-invariant function. Assume that the machine outputs an estimate at each time point according to some fixed mapping from the state space to the unit interval. The quality of the estimation procedure is measured by the asymptotic risk, which is the long-term average of the instantaneous quadratic risk. The main contribution of this paper is an upper bound on the smallest worst-case asymptotic risk any such machine can attain. This bound coincides with a lower bound derived by Leighton and Rivest, to imply that $\Theta(1/S)$ is the minimax asymptotic risk for deterministic S-state machines. In particular, our result disproves a longstanding $\Theta(\log S/S)$ conjecture for this quantity, also posed by Leighton and Rivest.

Keywords: Learning with Memory Constraints, Parametric Estimation, Minimax Estimation.

1. Introduction

The statistical hardness of a parametric estimation problem has been traditionally characterized by the number of independent samples from the distribution P_θ one needs to see in order to accurately estimate θ. However, as the amount of available data is constantly increasing, collecting enough samples for accurate estimation is becoming less of a problem, provided that the parameter θ is of a relatively low dimension. In this regime, it is the computational resources dedicated to the estimation task, rather than the number of samples, that constitute the main bottleneck determining the quality of estimation one can attain.

As a result, the topic of estimation / learning under computational constraints is currently drawing considerable attention; in particular, the problem of estimation / learning under memory constraints has been recently studied in various different setups, as we further elaborate in Subsection 1.1. Despite this ongoing effort, there are still substantial gaps in the understanding of the effects memory limitations can have on the minimal possible estimation error. This work addresses such a gap in arguably the simplest setup possible: estimation of a single parameter $\theta \in [0, 1]$ from an infinite number of independent samples from P_θ, using a finite-memory learning algorithm.

Specifically, we consider the bias estimation problem, defined as follows: $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables drawn according to the $\mathsf{Bern}(\theta)$ distribution. An *S-state estimation procedure* for this problem consists of two functions: f, and $\hat{\theta}$,

where $f : [S] \times \{0,1\} \to [S]$ is a deterministic state transition (or memory update) function (here $[S] = \{1, \ldots, S\}$), and $\hat{\theta} : [S] \to [0,1]$ is the estimate function. Letting M_n denote the state of the memory at time n, this finite-state machine evolves according to the rule

$$M_0 = s_{\text{init}}, \tag{1}$$
$$M_n = f(M_{n-1}, X_n) \in [S], \tag{2}$$

for some predetermined initial state $s_{\text{init}} \in [S]$. If the machine is stopped at time n, it outputs the estimation $\hat{\theta}(M_n)$. We define the (asymptotic) quadratic risk attained by this estimation procedure, given that the true value of the parameter is θ, to be[1]

$$R_\theta(f, \hat{\theta}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\hat{\theta}(M_i) - \theta\right)^2. \tag{3}$$

We are interested in the *minimax risk* of the estimation problem, defined as

$$R(S) \triangleq \min_{f,\hat{\theta}} \max_{\theta\in[0,1]} R_\theta(f, \hat{\theta}), \tag{4}$$

where the minimum is taken over all S-state estimation procedures. This paper derives an upper bound on the minimax risk, which together with a known lower bound, establishes the exact behavior of the minimax risk with S.

Note that here the memory update function f is not allowed to depend on time. First, as our focus here is on upper bounds, it is always desirable to use the weakest possible model. Moreover, 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, just storing the code generating a time-varying algorithm may require unbounded memory.

Besides memory, another resource that plays an important role here is randomness. While allowing the use of randomization in the estimation procedure may certainly help, this resource has a cost. Even if one has access to unlimited randomness (which is the case in our setting, since randomness can be easily extracted from the i.i.d. sequence $X_1, X_2, \ldots$), storing this randomness places a toll on one's memory budget, which needs to be taken into account in our deterministic setup. One can nevertheless define the *randomized minimax risk* of the estimation problem, to be the smallest asymptotic risk that can be uniformly guaranteed when randomized state-transition functions f are allowed, i.e.,

$$R_{\mathsf{rand}}(S) \triangleq \min_{\text{randomized } f,\hat{\theta}} \max_{\theta\in[0,1]} R_\theta(f, \hat{\theta}), \tag{5}$$

We emphasize that in the above, in contrast to the deterministic setup we consider in this paper, randomness is "free" and not counted toward the memory budget. Our main result is that, in contrast to what was conjectured by Leighton and Rivest (1986), $R(S)$ and $R_{\mathsf{rand}}(S)$ are equal up to constants independent of S.

Let us be more precise. Prior to this work, it was known that $R_{\mathsf{rand}}(S) = \Theta(1/S)$. The upper bound was proved by Samaniego (1973), who constructed an S-state randomized estimation

1. It is not difficult to show that the limit exists due to the underlying finite-state structure and the independence of the samples.

procedure that asymptotically induces a $\mathrm{Binomial}(S-1,\theta)$ stationary distribution on the memory state space. The lower bound was established a decade later by Leighton and Rivest (1986), using the Markov chain tree theorem. In the same paper, Leighton and Rivest (1986) further constructed a deterministic S-state estimation procedure by de-randomizing Samaniego's construction, and as a result showed that $R(S) = O(\log S/S)$. They then conjectured that this is the best possible asymptotic minimax risk any deterministic S-state estimation procedure can attain, and further stated the problem of proving or disproving this conjecture as the first out of five open problems left for future research. A nice interpretation of their conjecture is the naturally appealing claim that an optimal deterministic algorithm can be obtained by de-randomizing the optimal random algorithm. In their deterministic algorithm, which they believed to be optimal, randomness is extracted from the measurements by augmenting each state with $O(\log(S))$ additional states, which increases the overall MSE (see Section III of Leighton and Rivest (1986)). Surprisingly, we show that such a de-randomization is suboptimal, thereby disproving the conjecture of Leighton and Rivest.

Theorem 1

$$R(S) = O\left(\frac{1}{S}\right). \tag{6}$$

Since deterministic S-state estimation procedures are a subset of the class of S-state randomized estimation procedures, we clearly have $R(S) \geq R_{\mathsf{rand}}(S) = \Omega(1/S)$, where the lower bound is due to Leighton and Rivest (1986). Consequently:

Corollary 2

$$R(S) = \Theta\left(\frac{1}{S}\right). \tag{7}$$

1.1. 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. It has then been largely abandoned, but recently it has been again enjoying much attention, due to the reasons described above, and many works have addressed different aspects of the learning under memory constraints problem over the last few years. See, e.g., Steinhardt and Duchi (2015); Steinhardt et al. (2016); Raz (2018); Dagan and Shamir (2018); Dagan et al. (2019); Sharan et al. (2019) for a far from exhaustive list of recent works.

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 $\mathsf{Bern}(p)$ or $\mathsf{Bern}(q)$, Cover (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 above, 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. To demonstrate the important role randomization plays in approaching this value, the same authors show in Hellman and Cover (1971) that for any memory size $S < \infty$ and $\delta > 0$, there exists problems such that any S-state deterministic procedure has probability of error $\mathsf{P_e} \geq \frac{1}{2} - \delta$, while their randomized procedure

from Hellman and Cover (1970) has $\mathsf{P_e} \leq \delta$. Note that one can simulate a randomized procedure with a deterministic one by using some of the samples of $\{X_n\}$ for randomness extraction, e.g., using Von Neumann (1951) extraction. However, the extracted random bits must be stored, which could result in a substantial increase in memory, see Chandrasekaran (1970). 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 deterministic case.

One of the earlier works on estimation with finite memory is due to Roberts and Tooley (1970), who tackled the problem of estimation under quadratic risk for a random variable with additive noise. Hellman (1974) studied the problem of estimating the mean θ of a Gaussian distribution and discovered a S-state estimation procedure that asymptotically achieves the same Bayesian quadratic risk as the optimal S-level quantizer $Q(\theta)$ for θ, where $Q : \mathbb{R} \to [S]$. As already described above, Samaniego (1973) and Leighton and Rivest (1986) have showed that $R_{\mathsf{rand}}(S) = \Theta(1/S)$. Meron and Feder (2004); Ingber and Feder (2006); Dar and Feder (2014) studied the subject of finite-memory universal prediction of sequences using randomized/deterministic machines. More recently, Jain and Tyagi (2018) studied the shrinkage in memory between the hypothesis testing and the estimation problem, namely the interesting fact that a machine with S states can distinguish between two coins with biases that differ by $1/S$, whereas the best additive accuracy it can achieve in estimating the bias is only $1/\sqrt{S}$. We further note that the problem of estimating statistics with bounded memory is attracting considerable attention in the machine learning literature lately, see, e.g., Chien et al. (2010); Kontorovich (2012); McGregor et al. (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).

2. Proof of Theorem 1

We now proceed to prove Theorem 1. We will describe our deterministic S-state estimation procedure and show that it attains quadratic risk of $O(1/S)$ uniformly for all $\theta \in [0, 1]$. In this section we provide the entire proof, but for clarity we rely on several technical claims whose proofs are relegated to the next section or to the Appendix.

Recall from (1) and (2) that any deterministic S-state estimation procedure corresponds to a finite-state machine with S states, with at most two outgoing edges from each state, one for $X_i = 0$ and one for $X_i = 1$. Running this machine on an i.i.d. $\mathsf{Bern}(\theta)$ input sequence $X_1, X_2, \ldots$, generates a Markov chain $\{M_n\}_{n=1}^{\infty}$, where M_n denotes the state of the machine at time n. We emphasize that the distribution of the process $\{M_n\}$ depends on θ, which is the parameter we are trying to estimate. To lighten notation, we nevertheless leave this dependence implicit. The construction we describe below trivially achieves $R_\theta(f, \hat{\theta}) = O(1/S)$ for $\theta = 0$ and $\theta = 1$, and thus in the remainder of the paper we assume without loss of generality that $\theta \in (0, 1)$.

The high-level idea underlying our scheme is to break down the memory-constrained estimation task into a sequence of memory-constrained (composite) binary hypothesis testing sub-problems. In each such sub-problem, the goal is to decide whether the true underlying parameter θ satisfies

$\{\theta < q\}$ or $\{\theta > p\}$, for some $0 < q < p < 1$. Those decisions are then used in order to traverse an induced Markov chain in a way that enables us to accurately estimate θ.

Let us now describe the particular structure of the proposed machine. In our construction, the state space $[S]$ is partitioned into K disjoint sets denoted by $\mathcal{S}_1, \ldots, \mathcal{S}_K$, where the estimation function value is the same inside each $\mathcal{S}_k$, i.e.,

$$\hat{\theta}(s) = \hat{\theta}_k, \quad \forall s \in \mathcal{S}_k,\ k \in [K]. \tag{8}$$

The goal is to design a machine for which the stationary distribution of $\{M_n\}$ corresponding to the parameter θ will concentrate on states that belong to classes $\mathcal{S}_k$ for which $(\theta - \hat{\theta}_k)^2$ is the smallest. This goal is in general easier to achieve when each set consists of a large number of states, which corresponds to small K (as the total number of states S is fixed). On the other hand, the quadratic risk such a machine can attain is obviously limited by the number of different sets K, and in particular is $\Omega(1/K^2)$, as there must exist some $\theta \in [0,1]$ at distance $\Omega(1/K)$ from all points $\hat{\theta}_1, \ldots, \hat{\theta}_K$. Thus, the choice of K should balance the tension between these two contrasting goals; specifically, we will see that the choice $K = \Theta(\sqrt{S})$ is suitable to that end.

Since the estimator $\hat{\theta}$ depends on $\{M_n\}$ only through its class, it is natural to define the *quantized process* $\{Q_n\}_{n=1}^{\infty}$ obtained by the deterministic scalar mapping

$$Q_n = \phi(M_n), \quad n = 1, 2, \ldots, \tag{9}$$

where $\phi : [S] \to [K]$ maps each state to its set label (namely: $\phi(s) = k$ iff $s \in \mathcal{S}_k$). The process $\{Q_n\}$, as well as any process on a finite alphabet, consists of *runs* of the same letter. We can therefore decompose it as $\{S_1, \tau_1\}, \{S_2, \tau_2\}, \ldots$, where S_i denotes the first letter in the ith run, and τ_i denotes its length. We refer to the process $\{S_i\}_{i=1}^{\infty}$, supported on $[K]$ as the *sampled process*, and to $\{\tau_i\}_{i=1}^{\infty}$, supported on $\mathbb{N}$, as the *holding times* process. Note that both $\{S_i\}$ and $\{\tau_i\}$ are deterministically determined by $\{Q_n\}$ and hence, by the original process $\{M_n\}$. In general, the sampled process can be complicated; however, in our construction, we impose a particular structure that ensures that the sampled process $\{S_n\}$ is also a Markov process. Specifically, for each $k \in [K]$ there is an *entry state* $s_k \in \mathcal{S}_k$, such that all edges going out of a state $\ell \notin \mathcal{S}_k$ to the set $\mathcal{S}_k$, go into the entry state $s_k \in \mathcal{S}_k$. In other words, whenever M_n enters the set $\mathcal{S}_k$ from a different set, it does so through the designated entry state only. This feature guarantees that at the beginning of the ith run, the state of the original process $\{M_n\}$ is determined by S_i, and consequently $\{S_i\}$ is indeed a Markov process itself. Furthermore, conditioned on S_i, the holding time τ_i is independent of the entire past. We denote the conditional distribution of τ_i conditioned on the event $S_i = k$, by P_{T_k}. It will be convenient to also define the random variables $T_k \sim P_{T_k}$, for $k \in [K]$. In our construction, we further guarantee that any set $\mathcal{S}_k$ is accessible from any other set $\mathcal{S}_j$, $j \neq k$. This ensures that the underlying Markov process $\{M_n\}$ is ergodic, and as a result, so is the sampled process $\{S_n\}$. We refer to the structure described here, i.e., all sets are accessible from one another and have entry states, as a *nested Markov structure*.

The ergodicity of $\{M_n\}$ immediately implies the ergodicity of the quantized process $\{Q_n\}$, by (9). Denote by π_k the stationary probability of state k for the process $\{Q_n\}$. We therefore have that for a machine $f, \hat{\theta}$ of the type described above,

$$R_\theta = R_\theta(f, \hat{\theta}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\hat{\theta}(M_i) - \theta\right)^2 = \sum_{k=1}^{K} \pi_k \left(\hat{\theta}_k - \theta\right)^2. \tag{10}$$

The next lemma determines the stationary distribution $\{\pi_k\}_{k\in[K]}$ of the quantized process $\{Q_n\}$, in terms of the stationary distribution $\{\mu_k\}_{k\in[K]}$ of the sampled process $\{S_n\}$ and the expected holding times $\{\mathbb{E}[T_k]\}_{k\in[K]}$.

Lemma 3 *The unique stationary probability of state k under the process $\{Q_n\}$ is*

$$\pi_k = \frac{\mathbb{E}[T_k]\mu_k}{\sum_{j=1}^{M}\mathbb{E}[T_j]\mu_j}. \tag{11}$$

Combining Lemma 3 with (10), we have that the risk of such machine is

$$R_\theta = \sum_{k=1}^{K}\frac{\mathbb{E}[T_k]\mu_k}{\sum_{j=1}^{M}\mathbb{E}[T_j]\mu_j}\left(\hat{\theta}_k-\theta\right)^2. \tag{12}$$

It is immediately evident from (12) that the asymptotic risk attained by a machine with the nested Markov structure defined above depends only on the stationary distribution of the sampled process $\{S_n\}$ and the expected holding times. Ideally, we would like to construct this machine such that two things would happen for every θ:

(i) $\{\mu_k\}$ would be concentrated on states whose corresponding estimate $\hat{\theta}_k$ is close to θ;

(ii) The expected holding times for these states will be at least as large as those of other states.

We now describe how our machine is designed to achieve the desired behaviour (i) of $\{\mu_k\}$. Later, we will see that the desired behavior (ii) of $\{\mathbb{E}[T_k]\}$ follows more or less automatically.

First, we set our estimators to be[2]

$$\hat{\theta}_k = \frac{k}{K+2}, \quad k\in[K]. \tag{13}$$

We then design our machine such that the sampled process $\{S_n\}$ is a random walk, that moves either one state left or one state right from each state (except for the extreme states 1 and K that behave slightly differently). In particular, the kth state in $\{S_n\}$ is connected only to states $k+1$ and $k-1$ for all $k\in\{2,\dots,K-1\}$. The precise diagram for the sampled process $\{S_n\}$ is shown in Figure 1, where the transition probabilities $\{p_k, q_k=1-p_k\}_{k\in[K]}$ will depend on θ through the construction of the original machine generating the original Markov chain $\{M_n\}$. We design the machine in a way that guarantees that the random walk $\{S_n\}$ has a strong *drift* towards the state k whose corresponding estimator is closest to θ. In particular, if $\theta > \frac{k+1}{K+2}$ then $p_k > 1-\epsilon$ and conversely, if $\theta < \frac{k}{K+2}$ then $p_k < \epsilon$, for some $\epsilon < 1/2$ and all states $k\in\{2,\dots,K-1\}$.

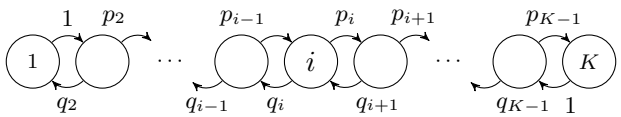

Figure 1: A sampled chain of K states.

2. The denominator is set to $K+2$ rather than K for minor technical reasons, in order to avoid dealing with probabilities on the boundary of the simplex in the analysis.

The desired drift behavior is enabled by constructing the sets $\mathcal{S}_1, \ldots, \mathcal{S}_K$ as *mini-chains*, where the kth mini-chain consists of N_k states, and is designed to solve the composite binary hypothesis testing problem: $\mathcal{H}_0 : \left\{\theta > \frac{k+1}{K+2}\right\}$ vs. $\mathcal{H}_1 : \left\{\theta < \frac{k}{K+2}\right\}$. Each mini-chain $\mathcal{S}_k$ is initialized in its entry state s_k, and eventually moves to the entry state s_{k+1} of mini-chain $\mathcal{S}_{k+1}$ if it decided in favor of hypothesis $\mathcal{H}_0$, or to the entry state s_{k-1} of mini-chain $\mathcal{S}_{k-1}$ if it decided in favor of hypothesis $\mathcal{H}_1$. The time it takes it to "make a decision" is the random holding time with some distribution P_{T_k}. Note that if the error probability of the machine is smaller than $\epsilon < 1/2$ under both hypotheses, we will indeed attain the desired drift behavior. Our goal now is to design mini-chains that attain small error probabilities with as few states as possible. To that end, we appeal to Berg et al. (2020), where the authors defined the following machine.[3]

Definition 4 $\mathsf{RUNS}(N, p, q)$ *is the machine with* $N \geq 4$ *states depicted in Figure 2, 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* $\mathcal{H}_0$ *and exists right. Otherwise, it decides* $\mathcal{H}_1$ *and exists left. The initial state of the machine is* $s = f(N, p, q)$*, where*

$$f(N, p, q) \triangleq 2 + \left\lceil \frac{\log pq}{\log p(1-p) + \log q(1-q)} (N - 3) \right\rceil, \tag{14}$$

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*

$$p_1^0 = \sup_{\theta < q} \Pr_{X_1, X_2 \ldots \overset{i.i.d.}{\sim} \mathsf{Bern}(\theta)} \left(\mathsf{RUNS}(N, p, q) \textit{ decides } \mathcal{H}_0\right), \tag{15}$$

$$p_0^1 = \sup_{\theta > p} \Pr_{X_1, X_2 \ldots \overset{i.i.d.}{\sim} \mathsf{Bern}(\theta)} \left(\mathsf{RUNS}(N, p, q) \textit{ decides } \mathcal{H}_1\right). \tag{16}$$

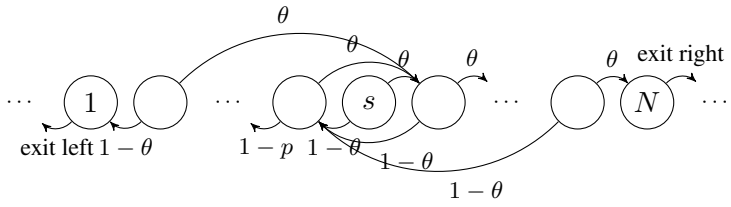

Figure 2: $\mathsf{RUNS}(N, p, q)$ - Deterministic Binary Hypothesis Testing Machine

The next lemma demonstrates that with $N = O(K)$ states, the machine $\mathsf{RUNS}(N, p, q)$ can decide whether $\theta > p$ or $\theta < q = p - 1/K$ with constant error probability $\epsilon < 1/2$. Thus, the desired drift can be attained by mini-chains of $O(K)$ states.

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

Lemma 5 *For any* $\frac{2}{K} \leq p \leq 1 - \frac{1}{K}$, $q = p - \frac{1}{K}$ *and* $0 < \epsilon < 1/2$, *let*[4]

$$N = N(\epsilon, p, K) \triangleq 3 + \left\lceil K \cdot 6 \log \frac{2}{\epsilon \cdot \left(p - \frac{1}{K}\right)(1-p)} \right\rceil. \tag{17}$$

Then

$$\mathsf{P_e}^{\mathsf{RUNS}(N,p,q)} < \epsilon. \tag{18}$$

We therefore take the kth mini-chain $\mathcal{S}_k$ as the machine $\mathsf{RUNS}(N_k, p, q)$ with $q = \frac{k}{K+2}$, $p = q + \frac{1}{K+2} = \frac{k+1}{K+2}$, and $N_k = N(\epsilon, \frac{k+1}{K+2}, K+2)$. The total number of states in our machine is therefore (see calculation in the appendix)

$$S = \sum_{k=1}^{K} N_k = \sum_{k=1}^{K} N\left(\epsilon, \frac{k+1}{K+2}, K+2\right) \leq 6(K+2)^2 \log\left(\frac{2e}{\epsilon}\right), \tag{19}$$

and the sampled chain $\{S_n\}$ indeed satisfies the desired drift property: for all $2 \leq k \leq K-1$ we have that if $\theta > \frac{k+1}{K+2}$ then $p_k > 1 - \epsilon$ whereas if $\theta < \frac{k}{K+2}$ then $p_k < \epsilon$. Note that we did not quantify p_k for the case where $\theta \in \left[\frac{k}{K+2}, \frac{k+1}{K+2}\right]$, but as will become apparent below, it is indeed not needed for our analysis. Also note that whenever the sampled chain reaches state 1 it will immediately move back to state 2, and whenever it reaches state K it will immediately move back to state $K-1$ (that is, $p_1 = 1$ and $p_K = 0$), but the holding times in those states are nevertheless random (and may be very large if the underlying θ is very close to 0 or 1, and dictated by the time it takes for the corresponding $\mathsf{RUNS}(N, p, q)$ mini-chains $\mathcal{S}_1$ and $\mathcal{S}_K$ to reach a decision). The next lemma shows that the drift property implies that if $\theta \in \left[\frac{k}{K+2}, \frac{k+1}{K+2}\right]$, then the stationary probability μ_j of the jth state in the sampled chain decreases exponentially with the "distance" $|j-k|$.

Lemma 6 *Assume that* $\theta \in \left[\frac{k}{K+2}, \frac{k+1}{K+2}\right]$. *Then, the stationary distribution of the sampled process* $\{S_n\}$ *induced by the machine described above satisfies*

$$\mu_{k-i} \leq \mu_{k-1} \left(\frac{\epsilon}{1-\epsilon}\right)^{i-1} \tag{20}$$

for $1 \leq i \leq k-1$, *and*

$$\mu_{k+i} \leq \mu_{k+1} \left(\frac{\epsilon}{1-\epsilon}\right)^{i-1} \tag{21}$$

for $1 \leq i \leq K-k$.

This shows that the stationary distribution of the sampled chain $\{S_n\}$ is indeed concentrated on the desired states. The next lemma deals with the expected holding times, and lower bounds the ratio between the expected holding time in the "correct state" k and the expected holding time in any other state of the sampled chain.

4. Logarithms in this paper are taken to base 2.

Lemma 7 *If $\theta < \frac{j}{K+2}$, then the expected holding time in state i satisfies*

$$\mathbb{E}[T_j] \geq (1-\epsilon)\,\mathbb{E}[T_i] \tag{22}$$

for all $i > j$. Similarly, if $\theta > \frac{j+1}{K+2}$, then the expected holding time in state i satisfies

$$\mathbb{E}[T_j] \geq (1-\epsilon)\,\mathbb{E}[T_i] \tag{23}$$

for all $i < j$.

We now combine (12) with Lemma 6 and Lemma 7 in order to upper bound the asymptotic risk attained by our machine, and establish the claim $R_\theta = O(1/S)$ for all $\theta \in (\frac{1}{K+2}, \frac{K+1}{K+2})$. The cases where $\theta \in [0, \frac{1}{K+2})$ and $\theta \in (\frac{K+1}{K+2}, 1]$ then follow easily from minor adjustments, and are treated in the appendix.

Assume that $\frac{k}{K+2} \leq \theta \leq \frac{k+1}{K+2}$ for some $k \in [K]$. From (3), the asymptotic risk is then

$$R_\theta = \sum_{i=1}^{K} \frac{\mathbb{E}[T_i]\mu_i}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\left(\frac{i}{K+2}-\theta\right)^2 \tag{24}$$

$$= \sum_{i=1}^{k-1} \frac{\mathbb{E}[T_i]\mu_i}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\left(\frac{i}{K+2}-\theta\right)^2 + \frac{\mathbb{E}[T_k]\mu_k}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\left(\frac{k}{K+2}-\theta\right)^2$$

$$+ \sum_{i=k+1}^{K} \frac{\mathbb{E}[T_i]\mu_i}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\left(\frac{i}{K+2}-\theta\right)^2 \tag{25}$$

$$\leq \frac{1}{1-\epsilon}\sum_{i=1}^{k-1} \frac{\mathbb{E}[T_{k-1}]\mu_{k-1}}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\frac{\mu_i}{\mu_{k-1}}\left(\frac{i}{K+2}-\theta\right)^2 + \frac{1}{(K+2)^2}$$

$$+ \frac{1}{1-\epsilon}\sum_{i=k+1}^{K} \frac{\mathbb{E}[T_{k+1}]\mu_{k+1}}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\frac{\mu_i}{\mu_{k+1}}\left(\frac{i}{K+2}-\theta\right)^2 \tag{26}$$

$$\leq \frac{1}{1-\epsilon}\sum_{i=1}^{k-1}\left(\frac{\epsilon}{1-\epsilon}\right)^{i-1}\left(\frac{i+1}{K+2}\right)^2 + \frac{1}{(K+2)^2} + \frac{1}{1-\epsilon}\sum_{i=1}^{K-k}\left(\frac{\epsilon}{1-\epsilon}\right)^{i-1}\left(\frac{i+1}{K+2}\right)^2 \tag{27}$$

$$\leq \frac{1}{(K+2)^2}\cdot\frac{1}{1-\epsilon}\left(2\cdot\sum_{i=1}^{\infty}\left(\frac{\epsilon}{1-\epsilon}\right)^{i-1}(i+1)^2+1\right) \tag{28}$$

$$\leq \frac{6\log\left(\frac{2e}{\epsilon}\right)}{S}\left(\frac{2\epsilon}{(1-2\epsilon)^3}+\frac{8(1-\epsilon)}{(1-2\epsilon)^2}+\frac{1}{1-\epsilon}\right), \tag{29}$$

where (26) follows from Lemma 7, (27) follows from Lemma 6 and since $\frac{\mathbb{E}[T_j]\mu_j}{\sum_{k=1}^{M}\mathbb{E}[T_k]\mu_k} \leq 1$, (28) is since we only add positive terms, and (29) is due to the identity $\sum_{i=0}^{\infty} q^i(i+2)^2 = \frac{q(1+q)+4(1-q)}{(1-q)^3}$ and by substituting (19). Finally, substituting $\epsilon = 1/100$ into (29) gives $R_\theta < \frac{600}{S}$.

3. Proofs of Technical Claims

The following simple lemma will be useful for the proofs of Lemma 3 and Lemma 6.

Lemma 8 *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*

$$\Pr(X_n \in \mathcal{C}, X_{n+1} \in \mathcal{C}') = \Pr(X_n \in \mathcal{C}', X_{n+1} \in \mathcal{C}). \tag{30}$$

Proof For any disjoint partition $\mathcal{C} \cup \mathcal{C}' = \mathcal{S}$ we have

$$\Pr(X_{n+1} \in \mathcal{C}') = \Pr(X_n \in \mathcal{C}') = \Pr(X_n \in \mathcal{C}', X_{n+1} \in \mathcal{C}') + \Pr(X_n \in \mathcal{C}', X_{n+1} \in \mathcal{C}). \tag{31}$$

Subtracting $\Pr(X_n \in \mathcal{C}', X_{n+1} \in \mathcal{C}')$ from both sides, establishes the claim. ■

Proof of Lemma 3 : The proof is very similar to the derivation of the invariant measure of a continuous-time Markov chain. Let $\{M'_n\}$ be the process defined as follows:

1. Draw M'_0 according to the stationary distribution of M_n.
2. For $n > 0$, draw $M'_{n+1}|M'_n \sim W$, where W is the Markov kernel of our chain.
3. For $n < 0$, draw $M'_{n-1}|M'_n \sim W'$, where W' is the reverse Markov kernel corresponding to the stationary distribution.

Clearly, $\{M'_n\}$ is a stationary ergodic process with marginal distribution equal to the stationary distribution of $\{M_n\}$. Let $Q'_n = \phi(M'_n)$ where ϕ is the mapping to the set label (similar to Q_n). Clearly, $\{Q'_n\}$ is a stationary process as well, and $\{Q_n\}$ converges to the marginal distribution of $\{Q'_n\}$. Recall that $\{Q'_n\}$ is composed of runs of consecutive letters of $[K]$, and that the length of each run is independent of all past runs. The run-length random variables do depend on the letter $k \in [K]$ of the run, and we denote by $T_k \sim P_{T_k}$ a generic random variable corresponding to a run of the letter k. Furthermore, we denote by $A_k(t)$ the event that a new run in $\{Q'_n\}$ of letters k started at time t, and let the integer random variable $Z_t \geq 1$ denote the number of symbols left in the current run at time t (including the one at time t). If $Q'_0 = k$, this means that a run of letters k started at some time $-t$, and its corresponding Z_{-t} was greater than t. We can therefore write

$$\begin{aligned}
\pi_k &= \Pr\left(Q'_0 = k\right) && (32)\\
&= \sum_{t=0}^{\infty} \Pr\left(A_k(-t), Z_{-t} > t\right) && (33)\\
&= \sum_{t=0}^{\infty} \Pr\left(A_k(-t)\right) \Pr\left(T_k > t\right) && (34)\\
&= \Pr\left(A_k(0)\right) \sum_{t=0}^{\infty} \Pr\left(T_k > t\right) && (35)\\
&= \Pr\left(A_k(0)\right) \mathbb{E}\left(T_k\right), && (36)
\end{aligned}$$

where (34) follows since given that $A_k(t)$ occurred, Z_t is independent of everything that happened before this run began and has the distribution P_{T_k}, (35) is from stationarity, and (36) is due to the

identity $\sum_{t=0}^{\infty} \Pr(T_k > t) = \mathbb{E}(T_k)$ for a non-negative random variable. Thus, from stationarity, for each t we have

$$\Pr(A_k(t)) = \Pr(A_k(0)) = \frac{\pi_k}{\mathbb{E}(T_k)}. \tag{37}$$

Now, denote by $B_k(t)$ the event that a run of letters k ended at time t. Note that since $\{Q'_n\}$ is stationary, Lemma 8 suggests that the probability it enters a state k is equal to the probability it leaves a state k at any given time, namely

$$\Pr(B_k(t)) = \Pr(A_k(t)) = \frac{\pi_k}{\mathbb{E}(T_k)}, \quad k \in [k]. \tag{38}$$

Now consider the sampled Markov chain $\{S_n\}$, and denote its stationary distribution for state j by μ_j, and its transition probability from state j to state k by P_{jk}. We have

$$\Pr(A_k(t+1)) = \sum_{j \neq k} \Pr(B_j(t)) P_{jk}. \tag{39}$$

Substituting (37) into (39), we have

$$\frac{\pi_k}{\mathbb{E}(T_k)} = \sum_{j \neq k} \frac{\pi_j}{\mathbb{E}(T_j)} P_{jk}. \tag{40}$$

Thus, the stationary distribution $\{\pi_k\}_{k\in[K]}$ of $\{Q_n\}$ must satisfy (40). Since $\{\mu_k\}_{k\in[K]}$ is the unique stationary distribution of $\{S_n\}$, we have that

$$\pi^*_j = \frac{\mathbb{E}[T_j]\mu_j}{\sum_{k=1}^{M} \mathbb{E}[T_k]\mu_k}, \quad j \in [K], \tag{41}$$

is the unique distribution satisfying (40), and is consequently the stationary distribution of $\{Q_n\}$, as claimed. ■

Proof of Lemma 6 : By construction, $\{S_n\}$ follows the transition probability law plotted in Figure 1. For all $i \in \{2, \ldots, K-2\}$, we have from Lemma 8 that by choosing the partition $\mathcal{C} = \{1, \ldots, i-1\}, \mathcal{C}' = \{i, \ldots, K\}$ and noting from Figure 1 that only adjacent states are connected, $\mu_{i-1}p_{i-1} = \mu_i q_i$, or equivalently

$$\mu_{i-1} = \frac{q_i}{p_{i-1}} \mu_i. \tag{42}$$

By construction of the mini-chains $\mathcal{S}_i$ and by Lemma 5, we have that $q_i < \epsilon$ and $p_i > 1-\epsilon$ for $i < k$. Thus, repeated application of (42) yields

$$\mu_{k-i} = \prod_{j=1}^{i} \frac{q_{k-j+1}}{p_{k-j}} \mu_k \leq \left(\frac{\epsilon}{1-\epsilon}\right)^{i-1} \mu_{k-1}, \tag{43}$$

for $2 \leq i \leq k-1$. Similarly, since $p_i < \epsilon$ and $q_i > 1-\epsilon$ for $i > k$, we have

$$\mu_{k+i} = \prod_{j=1}^{i} \frac{p_{k+j-1}}{q_{k+j}} \mu_k \leq \left(\frac{\epsilon}{1-\epsilon}\right)^{i-1} \mu_{k+1}, \tag{44}$$

for $1 \leq i \leq K-1-k$. For the extreme states 1 and K, by appealing to Lemma 8 and recalling that $p_1 = 1$ and $q_K = 1$, we have

$$\mu_1 = q_2\mu_2 \leq \epsilon\mu_2 < \frac{\epsilon}{1-\epsilon} \cdot \mu_2, \tag{45}$$

and

$$\mu_K = p_{K-1}\mu_{K-1} \leq \epsilon\mu_{K-1} < \frac{\epsilon}{1-\epsilon} \cdot \mu_{K-1}. \tag{46}$$

■

Proof of Lemma 7 : Fix θ, and recall that each state i in the sampled chain corresponds to a $\mathsf{RUNS}\left(N_i, \frac{i}{K+2}, \frac{i+1}{K+2}\right)$ mini-chain in the original chain, where $N_i = N(\epsilon, \frac{i}{K+2}, K+2)$ is as defined in (17). Restricting our attention to that mini-chain, denote by $s_i = f\left(N_i, \frac{i}{K+2}, \frac{i+1}{K+2}\right)$ its initial state, and denote by T_i^1 the first time a run of $N_i - s_i$ consecutive ones is observed, and T_i^0 as the first time a run of $s_i - 1$ consecutive zeros is observed. We exit the mini-chain when either a run of $N_i - s_i$ consecutive ones or a run of $s_i - 1$ consecutive zeros is observed, so we have that the exit time T_i satisfies $T_i \leq T_i^1$ and $T_i \leq T_i^0$, which implies

$$\mathbb{E}[T_i] \leq \mathbb{E}[T_i^1], \tag{47}$$

$$\mathbb{E}[T_i] \leq \mathbb{E}[T_i^0]. \tag{48}$$

Next, we observe that $i \mapsto s_i$ is monotonically non-increasing and $i \mapsto N_i - s_i$ is monotonically non-decreasing. These facts can be verified from the formulas (17) and (14) for $N(\epsilon, \frac{i}{K+2}, K+2)$ and $f\left(N_i, \frac{i}{K+2}, \frac{i+1}{K+2}\right)$, respectively. Thus the expected time to observe a run of $N_i - s_i$ consecutive ones is also non-decreasing and we have

$$\mathbb{E}\left[T_1^1\right] \leq \mathbb{E}\left[T_2^1\right] \leq \ldots \leq \mathbb{E}\left[T_j^1\right], \tag{49}$$

and similarly

$$\mathbb{E}\left[T_S^0\right] \leq \mathbb{E}\left[T_{S-1}^0\right] \leq \ldots \leq \mathbb{E}\left[T_j^0\right]. \tag{50}$$

Let $\{W_n^j(\theta)\}$ be a random walk in $\mathsf{RUNS}\left(N_j, \frac{j}{K}, \frac{j+1}{K}\right)$ under θ, and let $W_n^j(\theta) \to 1$ (resp. $W_n^j(\theta) \to 0$) denote the event that $\{W_n^j(\theta)\}$ exits right (resp. exits left). We have

$$T_j^1 = T_j + \left(T_j^1 - T_j\right) \mathbb{1}(W_n^j(\theta) \to 0). \tag{51}$$

By taking the expectation of both sides, we have

$$\mathbb{E}\left[T_j^1\right] = \mathbb{E}\left[T_j\right] + \mathbb{E}\left[\left(T_j^1 - T_j\right) \mathbb{1}(W_n^j(\theta) \to 0)\right] \tag{52}$$

$$= \mathbb{E}\left[T_j\right] + \Pr(W_n^j(\theta) \to 0)\, \mathbb{E}\left[T_j^1 - T_j | W_n^j(\theta) \to 0\right] \tag{53}$$

$$= \mathbb{E}\left[T_j\right] + \Pr(W_n^j(\theta) \to 0)\, \mathbb{E}\left[T_j^1\right], \tag{54}$$

due to

$$\mathbb{E}\left[T_j^1 - T_j | W_n^j(\theta) \to 0\right] = \sum_{t=1}^{\infty} \Pr(T_j = t | W_n^j(\theta) \to 0)\, \mathbb{E}\left[T_j^1 - T_j | T_j = t, W_n^j(\theta) \to 0\right] \quad (55)$$

$$= \sum_{t=1}^{\infty} \Pr(T_j = t | W_n^j(\theta) \to 0)\, \mathbb{E}\left[T_j^1 - t | T_j^1 > t, W_t^j(\theta) = 1\right] \quad (56)$$

$$= \sum_{t=1}^{\infty} \Pr(T_j = t | W_n^j(\theta) \to 0)\, \mathbb{E}\left[T_j^1\right] \quad (57)$$

$$= \mathbb{E}\left[T_j^1\right], \quad (58)$$

where (56) is since no run of $N_j - s_j$ ones was observed until time t and the last bit was $X_t = 0$, and (57) follows from the memoryless property of the chain. Thus,

$$\mathbb{E}\left[T_j\right] = \Pr(W_n^j(\theta) \to 1)\, \mathbb{E}\left[T_j^1\right], \quad (59)$$

$$\mathbb{E}\left[T_j\right] = \Pr(W_n^j(\theta) \to 0)\, \mathbb{E}\left[T_j^0\right]. \quad (60)$$

Equation (22) now follows by recalling that $\Pr(W_n^j(\theta) \to 0) \geq 1 - \epsilon$ for $\theta < \frac{j}{K}$ and by appealing to (50) and (48). (23) is proven similarly by appealing to (49). ■

Acknowledgments

This work was supported by the ISF under Grants 1791/17 and 1495/18.

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 ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877, 2020.

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.

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.

Balakrishnan Chandrasekaran. Finite-memory hypothesis testing–a critique (corresp.). *IEEE Transactions on Information Theory*, 16(4):494–496, 1970.

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

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.

Ronen Dar and Meir Feder. Finite-memory prediction as well as the empirical mean. *IEEE transactions on information theory*, 60(8):4526–4543, 2014.

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

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. Finite-memory algorithms for estimating the mean of a gaussian distribution (corresp.). *IEEE Transactions on Information Theory*, 20(3):382–384, 1974.

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

Martin E Hellman and Thomas M Cover. On memory saved by randomization. *The Annals of Mathematical Statistics*, 42(3):1075–1078, 1971.

Amir Ingber and Meir Feder. Prediction of individual sequences using universal deterministic finite state machines. In *2006 IEEE International Symposium on Information Theory*, pages 421–425. IEEE, 2006.

Ayush Jain and Himanshu Tyagi. Effective memory shrinkage in estimation. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1071–1075. IEEE, 2018.

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

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

F. Thomson Leighton and Ronald Rivest. Estimating a probability using finite memory. *IEEE Transactions on Information Theory*, 32(6):733–742, 1986.

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.

Eado Meron and Meir Feder. Finite-memory universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 50(7):1506–1523, 2004.

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.

Richard Roberts and J Tooley. Estimation with finite memory. *IEEE Transactions on Information Theory*, 16(6):685–691, 1970.

Francisco J. Samaniego. Estimating a binomial parameter with finite memory. *IEEE Transactions on Information Theory*, 19(5):636–643, 1973.

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.

John Von Neumann. 13. various techniques used in connection with random digits. *Appl. Math Ser*, 12(36-38):5, 1951.

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. Bound on the error probability

To establish Lemma 5, we need to prove two supporting lemmas. First, we show that p_0^1 is achieved by $\theta = q$, and p_1^0 is achieved by $\theta = p$. Denote the probability of deciding $\mathcal{H}_0$ under θ as $p_0(\theta)$, and the probability of deciding $\mathcal{H}_1$ under θ as $p_1(\theta)$.

Lemma 9 *For* $\mathsf{RUNS}(N, p, q)$*, if* $\theta > p$*, then* $p_1(\theta) \leq p_1(p)$*. Similarly, if* $\theta < q$*, then* $p_0(\theta) \leq p_0(q)$*.*

Proof We prove the first part of the claim, and the second follows symmetrically. To that end, we use a coupling argument. Denote by $\{W_n^p\}$ the random walk on $\mathsf{RUNS}(N, p, q)$ under p and $\{W_n^\theta\}$ as the random walk on $\mathsf{RUNS}(N, p, q)$ under θ, where here we assume the extreme states 1 and N are absorbing, such that once the random walk reaches one of these states, it stays there forever. We couple the two processes using the following joint distribution for $\left(\{W_n^p\}, \{W_n^\theta\}\right)$: Let $\{W_n^p\}$ be the standard walk on the chain under the $\mathsf{Bern}(p)$ sequence. For any n, if W_n^p goes one step to the right, W_n^θ goes one step to the right as well. If W_n^p goes one step to the left, we flip an independent $\mathsf{Bern}\left(\frac{\theta-p}{1-p}\right)$ coin, and W_n^θ goes one step to the right upon seeing 1 or one step to the left upon seeing 0. It is easy to see that the marginal distribution under $\{W_n^\theta\}$ corresponds to the chain under the $\mathsf{Bern}(\theta)$ distribution, and this is therefore a valid coupling. Our claim now immediately follows from the observation that under this coupling, W_n^θ is never to the left of W_n^p. ■

Second, we prove the following lemma, which bounds the error probability of $\mathsf{RUNS}(N, p, q)$ when the hypotheses are $\frac{1}{K}$ apart.

Lemma 10 *For any* $\frac{2}{K} \leq p \leq 1 - \frac{1}{K}$*,* $q = p - \frac{1}{K}$ *and* $N \geq 3 + \left\lceil K \cdot 6 \log \frac{2}{p_{\min}} \right\rceil$*, it holds that*

$$\mathsf{P_e}^{\mathsf{RUNS}(N,p,q)} \leq \frac{2}{p_{\min}} \cdot \exp_2 \left\{ - \frac{\left(1 - \frac{1}{K \cdot p}\right)\left(\frac{1}{K} H_b(p) - \frac{1}{K^2} \log p\right)(N-3)}{\frac{1}{K}\left(1 - 2\left(p - \frac{1}{K}\right)\right) - 2\left(p - \frac{1}{K}\right)\left(1 - p + \frac{1}{K}\right) \log p(1-p)} \right\}, \tag{61}$$

where $p_{\min} = \min\{p(1-p), q(1-q)\}$ *and* $H_b(p) \triangleq -p \log p - (1-p)\log(1-p)$ *is the binary entropy of* p*.*

Proof In Berg et al. (2020), the authors showed that for initial state s, we have

$$p_1(p) = \frac{1 - p^{N-s}}{1 + \frac{p^{N-s-1}}{(1-p)^{s-2}} - p^{N-s-1}} \tag{62}$$

$$\leq \frac{(1-p)^{s-2}}{p^{N-s-1}} \cdot \frac{1}{1 - (1-p)^{s-2}}, \tag{63}$$

and

$$p_0(q) = \frac{1 - (1-q)^{s-1}}{1 + \frac{(1-q)^{s-2}}{q^{N-s-1}} - (1-q)^{s-2}} \tag{64}$$

$$\leq \frac{q^{N-s-1}}{(1-q)^{s-2}} \cdot \frac{1}{1 - q^{N-s-1}}. \tag{65}$$

Choosing $s = s^*$, where s^* is

$$2 + \frac{\log pq}{\log p(1-p) + \log q(1-q)}(N-3), \tag{66}$$

we get

$$\frac{(1-p)^{s^*-2}}{p^{N-s^*-1}} = \frac{q^{N-s^*-1}}{(1-q)^{s^*-2}} = 2^{-r(p,q)(N-3)}, \tag{67}$$

where

$$r(p,q) \triangleq \frac{\log p \log(1-q) - \log q \log(1-p)}{\log p(1-p) + \log q(1-q)}. \tag{68}$$

We therefore have, for $s = s^*$,

$$\max\{p_0(q), p_1(p)\} \leq 2^{-r(p,q)(N-3)} \cdot \max\left\{\frac{1}{1-(1-p)^{s^*-2}}, \frac{1}{1-q^{N-s^*-1}}\right\}. \tag{69}$$

Recall that s is a state in the chain so it must be an integer, whereas s^* may not be. Thus, we need to round s^* either up or down, in which case, both ratios in (67) $\frac{(1-p)^{s^*-2}}{p^{N-s^*-1}}$, and $\frac{q^{N-s^*-1}}{(1-q)^{s^*-2}}$, will increase by at most $\frac{1}{p_{\min}}$, where $p_{\min} = \min\{p(1-p), q(1-q)\}$. Furthermore, for our choice of N, $\frac{2}{K} \leq p \leq 1 - \frac{1}{K}$ and $q = p - \frac{1}{K}$, we have that $3 < s* < N-2$ and the rightmost part of (69) is always upper bounded by 2. Combining this with Lemma 9, we therefore get the bound

$$\mathsf{P_e}^{\mathsf{RUNS}(N,p,q)} = \max\{p_1^0, p_0^1\} = \max\{p_0(q), p_1(p)\} \leq \frac{2}{p_{\min}} \cdot 2^{-r(p,q)(N-3)}. \tag{70}$$

Setting $p - q = \delta > 0$, we have

$$r(p, p-\delta) = \frac{\log p \log(1-p+\delta) - \log(p-\delta)\log(1-p)}{\log p(1-p) + \log(p-\delta)(1-p+\delta)} \tag{71}$$

$$= \frac{\log p\left(\log(1-p) + \log\left(1+\frac{\delta}{1-p}\right)\right) - \left(\log p + \log\left(1-\frac{\delta}{p}\right)\right)\log(1-p)}{\log p(1-p) + \log(1-p) + \log\left(1+\frac{\delta}{1-p}\right) + \log p + \log\left(1-\frac{\delta}{p}\right)} \tag{72}$$

$$\geq \frac{\frac{\delta}{1-p+\delta}\log p + \frac{\delta}{p}\log(1-p)}{2\log p(1-p) + \frac{\epsilon}{1-p+\delta} - \frac{\delta}{p-\delta}} \tag{73}$$

$$= -\frac{p-\delta}{p} \cdot \frac{\delta p \log p + \delta(1-p+\delta)\log(1-p)}{2(p-\delta)(1-p+\delta)\log p(1-p) - \delta(1-2(p-\delta))} \tag{74}$$

$$= \left(1-\frac{\delta}{p}\right) \cdot \frac{\delta H_b(p) - \delta^2 \log(1-p)}{\delta(1-2(p-\delta)) - 2(p-\delta)(1-p+\delta)\log p(1-p)}, \tag{75}$$

where (73) follows from $\frac{x}{x+1} \leq \log(1+x) \leq x$ and (75) follows from the definition of the binary entropy. The claim follows by substituting $\delta = \frac{1}{K}$. ■

Proof of Lemma 5 : Let $N = 3 + \lceil c \cdot K \rceil$, for some $c \geq 6 \log \frac{2}{p_{\min}}$. From Lemma 10,

$$\mathsf{P_e}^{\mathsf{RUNS}\left(N,p,p-\frac{1}{K}\right)} \leq \frac{2}{p_{\min}} \cdot \exp_2 \left\{ - \frac{c\left(1-\frac{1}{K\cdot p}\right)\left(H_b(p) - \frac{1}{K}\log(1-p)\right)}{\frac{1}{K}\left(1-2\left(p-\frac{1}{K}\right)\right) - 2\left(p-\frac{1}{K}\right)\left(1-p+\frac{1}{K}\right)\log p(1-p)} \right\} \tag{76}$$

In order to guarantee $\mathsf{P_e}^{\mathsf{RUNS}\left(N,p,p-\frac{1}{K}\right)} \leq \epsilon$, it is sufficient to choose c to be

$$\frac{\frac{1}{K}\left(1-2\left(p-\frac{1}{K}\right)\right) - 2\left(p-\frac{1}{K}\right)\left(1-p+\frac{1}{K}\right)\log p(1-p)}{\left(1-\frac{1}{K\cdot p}\right)\left(H_b(p) - \frac{1}{K}\log(1-p)\right)} \cdot \log \frac{2}{\epsilon p_{\min}}. \tag{77}$$

Upper bounding the first term in the brackets, we get

$$\frac{\frac{1}{K}\left(1-2\left(p-\frac{1}{K}\right)\right) - 2\left(p-\frac{1}{K}\right)\left(1-p+\frac{1}{K}\right)\log p(1-p)}{\left(1-\frac{1}{K\cdot p}\right)\left(H_b(p) - \frac{1}{K}\log(1-p)\right)} \tag{78}$$

$$\leq \frac{1}{1-\frac{1}{K\cdot p}} \cdot \frac{\frac{1}{K} + 2\left(H_b(p) - \frac{1}{K}\log(1-p)\right)}{H_b(p) - \frac{1}{K}\log(1-p)} \tag{79}$$

$$= \frac{1}{1-\frac{1}{K\cdot p}} \left(2 + \frac{1}{K\cdot H_b(p) - \log(1-p)}\right) \tag{80}$$

$$\leq \frac{3}{1-\frac{1}{K\cdot p}} \tag{81}$$

$$\leq 6, \tag{82}$$

where (79), (81) and (82) follows since $p \geq \frac{2}{K}$ implies

(i) $H_b(p) - \frac{1}{K}\log(1-p) \geq -\left(p-\frac{1}{K}\right)\left(1-p+\frac{1}{K}\right)\log p(1-p)$,

(ii) $K \cdot H_b(p) - \log(1-p) \geq 1$,

(iii) $\frac{1}{1-\frac{1}{K\cdot p}} \leq 2$.

Combining (82) and (77), noting that $\min\left\{p(1-p), \left(p-\frac{1}{K}\right)\left(1-p+\frac{1}{K}\right)\right\} \geq \left(p-\frac{1}{K}\right)(1-p)$, and choosing

$$c = c_{\epsilon,p} = 6 \log \frac{2}{\epsilon\left(p-\frac{1}{K}\right)(1-p)}, \tag{83}$$

the proof is concluded. ■

Appendix B. Calculation of number of states S in (19)

Using the expression in (17) for $N(\epsilon, p, K)$ we obtain

$$S = \sum_{k=1}^{K} N_k \tag{84}$$

$$= \sum_{k=1}^{K} N\left(\epsilon, \frac{k+1}{K+2}, K+2\right) \tag{85}$$

$$\leq 4K + 6(K+2)\sum_{k=1}^{K} \log \frac{2}{\epsilon\left(\frac{k}{K+2}\cdot\frac{K-k+1}{K+2}\right)} \tag{86}$$

$$= 4K + 6K(K+2)\log\left(\frac{2}{\epsilon}\right) - 6(K+2)\cdot 2\sum_{k=1}^{\frac{K}{2}} \log\left(\frac{k}{K+2}\cdot\frac{K-k+1}{K+2}\right) \tag{87}$$

$$\leq 4K + 6K(K+2)\log\left(\frac{2}{\epsilon}\right) - 6(K+2)\cdot 2\sum_{k=1}^{\frac{K}{2}} \log\left(\frac{k}{K+2}\right) - 6K(K+2) \tag{88}$$

$$\leq 4K + 6K(K+2)\log\left(\frac{2}{\epsilon}\right) - 6K(K+2)\log\left(\frac{K}{2e(K+2)}\right) - 6K(K+2) \tag{89}$$

$$\leq 4K + 6K(K+2)\log\left(\frac{2e}{\epsilon}\right) + 12(K+2) \leq 6(K+2)^2\log\left(\frac{2e}{\epsilon}\right), \tag{90}$$

where (87) follows from the symmetry of $\left(\frac{k}{K+2}\cdot\frac{K-k+1}{K+2}\right)$ around $k = \frac{K}{2}$, (88) from $\frac{K-k+1}{K+2} \geq \frac{1}{2}$ for all $1 \leq k \leq \frac{K}{2}$, (89) is from $n! \geq \left(\frac{n}{e}\right)^n$ and (90) follows from $\log(1+x) \geq \frac{x}{x+1}$.

Appendix C. Proof of $R_\theta = O(1/S)$ for $\theta \in \left[0, \frac{1}{K+2}\right)$ and $\theta \in \left(\frac{K+1}{K+2}, 1\right]$

We shall prove the case $\theta \leq \frac{1}{K+2}$. The case of $\theta \geq 1 - \frac{1}{K+2}$ follows from a symmetric argument. We show how previous results imply that for very small θ the stationary distribution is concentrated on the two leftmost states of the sampled chain. From there, the proof is similar (yet not identical) to the proof of the general case. Let us go step by step:

- Firstly, Lemma 5 implies that $p_k < \epsilon$ for all $k > 1$ in the chain of Figure 1.
- Now, a simplified (one-sided) version of Lemma 6 shows the stationary distribution is exponentially decreasing for all states ≥ 2. This follows since eq. (44) still holds with $k = 1$,

$$\mu_{i+1} \leq \left(\frac{\epsilon}{1-\epsilon}\right)^{i-1} \mu_2, \tag{91}$$

  for $1 \leq i \leq K-1$.
- Applying Lemma 7, eq. (22) states that $\mathbb{E}[T_j] > (1-\epsilon)\,\mathbb{E}[T_i]$ for all $j \in [n]$ and $i > j$.
- Calculate the risk R_θ.

$$R_\theta = \sum_{i=1}^{K} \frac{\mathbb{E}[T_i]\mu_i}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\left(\frac{i}{K+2}-\theta\right)^2 \tag{92}$$

$$= \frac{\mathbb{E}[T_1]\mu_1}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\left(\frac{1}{K+2}-\theta\right)^2 + \sum_{i=2}^{K} \frac{\mathbb{E}[T_i]\mu_i}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\left(\frac{i}{K+2}-\theta\right)^2 \tag{93}$$

$$\leq \frac{1}{(K+2)^2} + \frac{1}{1-\epsilon}\sum_{i=2}^{K} \frac{\mathbb{E}[T_2]\mu_2}{\sum_{j=1}^{K}\mathbb{E}[T_j]\mu_j}\frac{\mu_i}{\mu_2}\left(\frac{i}{K+2}-\theta\right)^2 \tag{94}$$

$$\leq \frac{1}{(K+2)^2} + \frac{1}{1-\epsilon}\sum_{i=1}^{K-1} \left(\frac{\epsilon}{1-\epsilon}\right)^{i-1}\left(\frac{i+1}{K+2}\right)^2 \tag{95}$$

$$\leq \frac{1}{(K+2)^2}\cdot\frac{1}{1-\epsilon}\left(\sum_{i=1}^{\infty}\left(\frac{\epsilon}{1-\epsilon}\right)^{i-1}(i+1)^2+1\right) \tag{96}$$

$$\leq \frac{6\log\left(\frac{2e}{\epsilon}\right)}{S}\left(\frac{\epsilon}{(1-2\epsilon)^3}+\frac{4(1-\epsilon)}{(1-2\epsilon)^2}+\frac{1}{1-\epsilon}\right), \tag{97}$$

where (94) follows from Lemma 7, (95) follows from Lemma 6 and since $\frac{\mathbb{E}[T_j]\mu_j}{\sum_{k=1}^{M}\mathbb{E}[T_k]\mu_k} \leq 1$, (96) is since we only add positive terms, and (97) is due to the identity $\sum_{i=0}^{\infty} q^i(i+2)^2 = \frac{q(1+q)+4(1-q)}{(1-q)^3}$ and by substituting (19). Finally, substituting $\epsilon = 1/100$ into (97) gives $R_\theta < \frac{300}{S}$.