Binary Hypothesis Testing with Deterministic Finite-Memory Decision Rules

Tomer Berg
Tel Aviv University
tomerberg@mail.tau.ac.il

Or Ordentlich
Hebrew University of Jerusalem
or.ordentlich@mail.huji.ac.il

Ofer Shayevitz
Tel Aviv University
ofersha@eng.tau.ac.il

***Abstract*—In this paper we consider the problem of binary hypothesis testing with finite memory systems. Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed Bernoulli random variables, with expectation p under $\mathcal{H}_0$ and q under $\mathcal{H}_1$. 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 we let the process run for a very long time ($n \to \infty$), and then make our decision according to some mapping from the state space to the hypothesis space. The main contribution of this paper is a lower bound on the Bayes error probability P_e of any such machine. In particular, our findings show that the ratio between the maximal exponential decay rate of P_e with S for a deterministic machine and for a randomized one, can become unbounded, complementing a result by Hellman.**

I. Introduction

Consider the following binary hypothesis testing problem: $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables drawn according to either the $\mathsf{Bern}(p)$ distribution, under hypothesis $\mathcal{H}_0$, or the $\mathsf{Bern}(q)$ distribution, under hypothesis $\mathcal{H}_1$, for $0 < q < p < 1$. For simplicity, we assume throughout that the prior probabilities of both hypothesis are given and are equal. A finite memory decision rule for this problem is a triplet (S, f, d) where S is the number of states used by the machine, $f : [S] \times \{0, 1\} \to [S]$ is the state transition function, and $d : [S] \to \{\mathcal{H}_0, \mathcal{H}_1\}$ is the decision function. In contrast to much of the prior work, where randomized state-transition functions f were allowed, here we restrict our attention to *deterministic* f.

Letting M_n denote the state of the memory at time n, the finite state machine evolves according to the rule

$$M_0 = s, \tag{1}$$

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

for some $s \in [S]$. If the machine is stopped at time n, it outputs the decision $d(M_n)$.

Conditioned on $\mathcal{H}_0$, the process $\{M_n\}$, induced by the function f, is a Markov chain with stochastic transition matrix

$$\mathbf{P}(p) = [\Pr(f(i, X) = j|\mathcal{H}_0)] = [p_{ij}(p)], \tag{3}$$

for all $i, j \in [S]$. Similarly, under $\mathcal{H}_1$, the induced Markov chain has stochastic transition matrix $\mathbf{P}(q) = [\Pr(f(i, X) = j|\mathcal{H}_1)] = [p_{ij}(q)]$. Following [1], we define the asymptotic probability of error of an algorithm as

$$\mathsf{P_e}(S, f, d) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Pr(e_i = 1), \tag{4}$$

where $e_i = \mathbb{1}_{\{d(M_i) \neq \mathcal{H}_t\}}$, and $\mathcal{H}_t$ is the true hypothesis. Arguably, a more natural definition of error probability is

$$\mathsf{P_e}(S, f, d) = \limsup_{n\to\infty} \Pr(d(M_n) \neq \mathcal{H}_t). \tag{5}$$

However, as (5) is always larger than (4), by a factor of at most S, the two definitions are equivalent for the purposes of this study.

The focus of this paper is the quantity

$$\mathsf{P}^*_{\mathsf{e}}(S) = \min_{\text{deterministic } f,d} \mathsf{P_e}(f, d) \tag{6}$$

where the minimum is taken over all S-state machines with *deterministic* transition functions f. We are specifically interested in the asymptotics of the error exponent with regards to S,

$$\overline{\mathsf{E}}(p, q) = -\liminf_{S\to\infty} \frac{1}{S} \log \mathsf{P}^*_{\mathsf{e}}(S), \tag{7}$$

$$\underline{\mathsf{E}}(p, q) = -\limsup_{S\to\infty} \frac{1}{S} \log \mathsf{P}^*_{\mathsf{e}}(S). \tag{8}$$

A. Related work

It seems that interest in the limited memory binary hypothesis testing problem was sparked by the work of Robbins [2] on the Two-Armed Bandit problem: A player is given two coins, with parameters unknown to him, and is required to maximize the long-run proportions of "heads" obtained, by successively choosing which coin to flip at any moment. Robbins proposed an algorithm that works with limited memory S. Cover [3] discovered a time-varying finite memory algorithm that achieves the maximum with $S = 2$, and in a subsequent paper addressing the binary hypothesis problem [4] described a time-varying finite memory machine that has probability of error approaching zero with $S = 4$. Due to the unlimited memory that is needed to implement a time-varying machine, Hellman and Cover [1] addressed the problem of binary hypothesis testing within the class of time-invariant finite memory machines. They have studied the quantity

$$\mathsf{P}^*_{\mathsf{e}_{\text{rand}}}(S) = \inf_{\text{randomized } f,d} \mathsf{P_e}(f, d), \tag{9}$$

The work of Tomer Berg was supported by the ISF under Grant 1791/17 and the ERC under Grant 639573. The work of Or Ordentlich was supported by the ISF under Grant 1791/17. The work of Ofer Shayevitz was supported by the ERC under Grant 639573.

where $\mathsf{P}_\mathsf{e}(f,d)$ is as defined in (4), and the infimum is over all time-invariant S-state machines with *randomized* transition functions f. It was shown in [1] that $\mathsf{P}^*_{\mathsf{e}_{\text{rand}}}(S) \geq \left(1+\gamma^{\frac{S-1}{2}}\right)^{-1}$ where $\gamma = \frac{p(1-q)}{q(1-p)}$, and that this value can be approached arbitrarily closely using a randomized algorithm.

To demonstrate the important role randomization plays in approaching this value, the same authors show in [5] that for any memory size $S < \infty$ and $\delta > 0$ there exists problems such that any S-state deterministic machine has probability of error $\mathsf{P}_\mathsf{e} \geq \frac{1}{2} - \delta$, while the randomized machine from [1] has $\mathsf{P}_\mathsf{e} \leq \delta$. When no external source of randomness is available, one can use some of the samples of $\{X_n\}$ for randomness extraction, e.g., using von Neumann extraction [6]. However, the extracted random bits must be stored, which could result in a substantial increase in memory [7].

In [8] (see also [9]) it is shown that $\underline{\mathsf{E}}(p,q)$, as defined in (8), is positive for all $p \neq q$.[1] Thus, recalling that $\mathsf{P}^*_{\mathsf{e}_{\text{rand}}}(S) \geq \left(1+\gamma^{\frac{S-1}{2}}\right)^{-1}$, we see that whenever $\gamma < \infty$, i.e., for any $0 < p, q < 1$, there exists some integer $1 \leq C = C(p,q) < \infty$ such that $\mathsf{P}^*_\mathsf{e}(S \cdot C) \leq \mathsf{P}^*_{\mathsf{e}_{\text{rand}}}(S)$, *for all* S. Our main result, stated in Theorem 1 below, may be interpreted as a lower bound on the required $C(p,q)$. Moreover, our Corollary 1 below shows that $C(p,q)$ grows unbounded for fixed $q < 1/2$ and $p \to 1$.

Finally, we note that after being abandoned for decades, the problem of learning under memory constraints is again attracting considerable attention in the machine learning literature, see, e.g., [11]–[16]. Another closely related active line of work is that of learning under communication constraints [17]–[22].

II. Main Result

We are now ready to present our main result.

Theorem 1. *Define*

$$d(p,q) \triangleq -\frac{\log(\min\{p,1-p\}) \cdot \log(\min\{q,1-q\})}{\log(\min\{p,1-p\}) + \log(\min\{q,1-q\})}. \quad (10)$$

Then

$$\overline{\mathsf{E}}(p,q) \leq d(p,q). \quad (11)$$

As it turns out, for extreme values of p (resp. q), the bound is tight. To show that, we need the following theorem.

Theorem 2. *Define*

$$r(p,q) \triangleq \frac{\log p \log(1-q) - \log q \log(1-p)}{\log p(1-p) + \log q(1-q)}. \quad (12)$$

Then for every $p > q$,

$$\underline{\mathsf{E}}(p,q) \geq r(p,q). \quad (13)$$

This lower bound on the error exponent is not tight in general, and in particular, for the symmetric case $p = 1-q$ it is worse then the exponent derived in [10]. We introduce it for the sole purpose of showing the tightness of our converse in certain regimes. The following corollary shows that in the limit of fixed $q < \frac{1}{2}$ (resp. $p > \frac{1}{2}$) and $p \to 1$ (resp. $q \to 0$) our upper and lower bounds coincide.

Corollary 1. *For any fixed* $q < \frac{1}{2}$,

$$\lim_{p\to 1} \overline{\mathsf{E}}(p,q) = \lim_{p\to 1} \underline{\mathsf{E}}(p,q) = -\log q. \quad (14)$$

Similarly, For any fixed $p > \frac{1}{2}$,

$$\lim_{q\to 0} \overline{\mathsf{E}}(p,q) = \lim_{q\to 0} \underline{\mathsf{E}}(p,q) = -\log(1-p). \quad (15)$$

Our converse, though in general not tight, demonstrates the gap between the error exponent for deterministic machines, and that of randomized ones, which was derived in [1]. Recalling that for any $q < 1/2$, the error exponent for randomized machines grows unbounded in the limit of $p \to 1$, Corollary 1 reveals that the restriction to deterministic machines may arbitrarily degrade the error exponent.

III. Achievability

Before we proceed to the proof of Theorem 1, which is our main result, we start with upper bounding $\mathsf{P}^*_\mathsf{e}(S)$ by analyzing various machines. It may be instructive to review some intuitive algorithms first, in order of increasing complexity, and evaluate their respective error probabilities.

A. Storing Sequences

Assume S is a power of 2, such that $k = \log(S)$, and store $X_1, \ldots X_k$. With this strategy, the problem reduces to the standard binary hypothesis testing for which the error probability is given by $2^{-kD^*(1+o(1))}$, where D^* is the Chernoff information between the two hypotheses [23]. Therefore, the error probability is polynomially decreasing in S.

B. Counting Ones

The flaw in the above storage mechanism is that it wastes a tremendous amount of memory by storing all sequences, where it is sufficient to keep track of the number of ones in the sequence.

Claim 1. *Let* S^* *be the minimal number of states required to determine whether or not a sequence of length* k *contains at least* $tk-1$ *ones, for some* $0 < t < 1$ *such that* $tk \in \mathbb{Z}$. *Then*

$$\frac{1}{2}\min\{t^2,(1-t)^2\}k^2 \leq S^* \leq tk^2. \quad (16)$$

The (straightforward) proof is omitted. From the claim we conclude that we can attain P_e that decreases exponentially in $\sqrt{S}$.

C. Proof of Theorem 2 - Detecting Discriminating Sequences

We begin by providing some high-level intuition guiding our construction. Since the sequence length is unbounded, one can afford to wait for the events that most sharply distinguish between the hypotheses, even if these events are arbitrarily rare. A reasonable choice for such events is a long consecutive run of either zeros or ones. We choose integers a and b such that $S = a + b + 1$. If we observe a run of a consecutive ones before a run of b consecutive zeros we decide $\mathcal{H}_0$, and if we observe a run of b consecutive zeros before a run

[1] For the symmetric setting, where $p = 1-q$, Shubert et al. [10] have also derived an upper bound on $\mathsf{P}^*_\mathsf{e}(S)$ that yields a positive error exponent $\underline{\mathsf{E}}(p,q)$.

of a consecutive ones, we decide $\mathcal{H}_1$. This algorithm can be implemented using the finite-state machine with S states depicted in Figure 1, for which $a = S - s$ and $b = s - 1$ (the probabilities on the arrows correspond to $\mathcal{H}_0$), where s is the initial state.

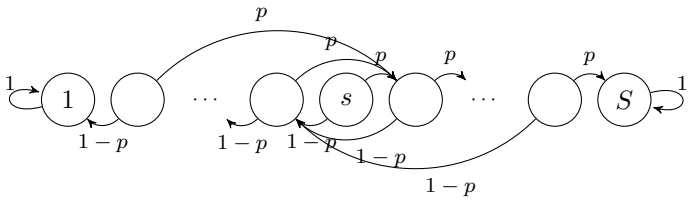

Figure 1: Counting consecutive runs of heads or tails

According to ([24], chapter VIII) the probability of observing a run of a consecutive ones before a run of b consecutive zeros under $\mathcal{H}_0$, which corresponds to the probability of absorption in state S when starting in state s for the machine of Figure 1, is

$$p_0^0(s) \triangleq \frac{1-(1-p)^b}{1+\frac{(1-p)^{b-1}}{p^{a-1}}-(1-p)^{b-1}} \tag{17}$$

$$= \frac{1-(1-p)^{s-1}}{1+\frac{(1-p)^{s-2}}{p^{S-s-1}}-(1-p)^{s-2}}. \tag{18}$$

Consequently, the probability of absorption in state 1 when starting in state s under the same hypothesis is

$$p_0^1(s) \triangleq \frac{1-p^a}{1+\frac{p^{a-1}}{(1-p)^{b-1}}-p^{a-1}} \tag{19}$$

$$= \frac{1-p^{S-s}}{1+\frac{p^{S-s-1}}{(1-p)^{s-2}}-p^{S-s-1}}. \tag{20}$$

Similarly, the respective probabilities under $\mathcal{H}_1$ are

$$p_1^0(s) \triangleq \frac{1-(1-q)^{s-1}}{1+\frac{(1-q)^{s-2}}{q^{S-s-1}}-(1-q)^{s-2}}, \tag{21}$$

$$p_1^1(s) \triangleq \frac{1-q^{S-s}}{1+\frac{q^{S-s-1}}{(1-q)^{s-2}}-q^{S-s-1}}. \tag{22}$$

Since all states on the chain are transient apart from $\{1, S\}$, when n is large the machine converges to one of these states with probability one. Hence, the error probability is

$$\mathsf{P_e}(s) = \frac{1}{2}(p_1^0(s) + p_0^1(s)). \tag{23}$$

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

$$\frac{\log pq}{\log p(1-p)+\log q(1-q)} S + \log\left(\frac{\frac{(1-q)^2}{q}\log q(1-q)}{\frac{p}{(1-p)^2}\log p(1-p)}\right), \tag{24}$$

rounded to the nearest integer, we have

$$\mathsf{P_e}(s^*) \le \max\left\{\frac{p^{1+c}}{(1-p)^{2-c}}, \frac{(1-q)^{1+c}}{q^{2-c}}\right\} \cdot 2^{-r(p,q)(S-1)} \tag{25}$$

where $c = \log \frac{(1-p)^2(1-q)^2 \log q(1-q)}{pq \log p(1-p)}$ and the result follows.

IV. Converse

The converse of Hellman and Cover implicitly assumes that the transition probabilities between states can be as small as desired, which is true when local randomness is an unlimited resource. In deterministic machines, however, the transition probabilities can only be as small as $\min(p, 1-p)$ under $\mathcal{H}_0$, or $\min(q, 1-q)$ under $\mathcal{H}_1$, a fact that plays a crucial role in the proof of our converse result. We note that any finite-state machine induces a Markov chain, and proceed to prove Theorem 1 in steps, first for ergodic Markov chains, and then for non-ergodic ones. For brevity, we denote $\mathsf{P_e} = \mathsf{P_e}(f, d)$.

A. Ergodic Markov chains

Assume the finite state machine is irreducible and aperiodic, such that the induced Markov chain is ergodic under both hypotheses. We note that, due to irreducibility, the average fraction of time spent in each state converges to a unique stationary distribution. Thus, the proof below still holds for periodic chains.

Denote by μ_i^p (resp. μ_i^q) the stationary probability of state i in the chain, under hypothesis $\mathcal{H}_0$ (resp. $\mathcal{H}_1$). Due to the equal prior on the hypotheses, the decision rule d that minimizes (4) maps each state to the hypothesis with the larger stationary probability. We show that there must exist a state $i \in [S]$ for which both μ_i^p and μ_i^q are large, and that this forces $\mathsf{P_e}$ to be large as well. We now proceed to formalize this idea.

Lemma 1. *Let $\{\mu_i^p\}_{i=1}^S$ be the stationary probabilities corresponding to $\mathbf{P}(p)$, and let $\{\mu_i^q\}_{i=1}^S$ be the stationary probabilities corresponding to $\mathbf{P}(q)$. Then*

$$\mathsf{P_e} \ge \frac{1}{2}\max_i \min\{\mu_i^p, \mu_i^q\} \triangleq \mathsf{P}_{\min}\left(\{\mu_i^p\}, \{\mu_i^q\}\right). \tag{26}$$

Proof. Since the prior on the hypotheses is uniform, the decision rule d that minimizes (4) is of the form $d(i) = \mathbb{1}(\mu_i^q \ge \mu_i^p)$. Hence

$$\mathsf{P_e} = \frac{1}{2}\sum_i \mu_i^p \,\mathbb{1}(\mu_i^q \ge \mu_i^p) + \frac{1}{2}\sum_i \mu_i^q \,\mathbb{1}(\mu_i^p > \mu_i^q) \tag{27}$$

$$= \frac{1}{2}\sum_i \min\{\mu_i^p, \mu_i^q\} \tag{28}$$

$$\ge \frac{1}{2}\max_i \min\{\mu_i^p, \mu_i^q\}. \tag{29}$$

□

Lemma 2. *Let $\{\mu_i^{\downarrow p}\}_{i=1}^S$ be an arrangement of $\{\mu_i^p\}$ in non-increasing order and $\{\mu_i^{\uparrow q}\}_{i=1}^S$ be an arrangement of $\{\mu_i^q\}$ in non-decreasing order. Then $\mathsf{P}_{\min}\left(\{\mu_i^p\}, \{\mu_i^q\}\right) \ge \mathsf{P}_{\min}\left(\{\mu_i^{\downarrow p}\}, \{\mu_i^{\uparrow q}\}\right)$.*

Proof. Since $\mathsf{P}_{\min}\left(\{\mu_i^p\}, \{\mu_i^q\}\right)$ is invariant to relabeling of the states, without loss of generality, we may assume $\{\mu_i^p\} = \{\mu_i^{\downarrow p}\}$. It suffices to show that if $\mu_j^q \le \mu_i^q$ for $j > i$, then swapping μ_j^q with μ_i^q cannot increase the maxmin in (26). Let $j > i$ and let $(\mu_i^p, \mu_i^q) = (a, c)$, $\left(\mu_j^p, \mu_j^q\right) = (b, d)$, where

$a \geq b, c \geq d$. The restriction of the maxmin to the nodes (i,j) is given by

$$\max\left(\min\{a,c\},\min\{b,d\}\right) \geq \min\{a,c\}. \tag{30}$$

Replacing μ_j^q with μ_i^q changes this value to

$$\begin{aligned}\max\left(\min\{a,d\},\min\{b,c\}\right) &\leq \max\left(\min\{a,c\},\min\{a,c\}\right)\\ &= \min\{a,c\},\end{aligned} \tag{31}$$

which clearly cannot increase the maxmin. □

The next lemma exploits the restriction to deterministic machines.

Lemma 3. *Let $\{\mu_i^{\downarrow p}\}_{i=1}^S$ be an arrangement of $\{\mu_i^p\}$ in non-increasing order. Then:*

$$\mu_{i+1}^{\downarrow p} \geq \mu_i^{\downarrow p} \cdot \min\{p, 1-p\}. \tag{32}$$

Proof. Without loss of generality, we may relabel the states such that $\mu_i^{\downarrow p} = \mu_i^p$, for all i. Let $A = \{1,\ldots,i\}$ and consider the partition of S to $S = A \cup A^c$. Since the chain is irreducible, there is some $j \in A^c$ that is accessible from some $j' \in A$ in one step. Then

$$\begin{aligned}\mu_{i+1}^{\downarrow p} \geq \mu_j^{\downarrow p} &\geq \mu_{j'}^{\downarrow p} \cdot \min\{p,1-p\} &(33)\\ &\geq \mu_i^{\downarrow p} \cdot \min\{p,1-p\}. &(34)\end{aligned}$$

□

Proof of Theorem 1 for ergodic Markov chains:

A repeated application of Lemma 3 implies that

$$\begin{aligned}\mu_i^{\downarrow p} &\geq \mu_1^{\downarrow p} \min\{p,1-p\}^{i-1} &(35)\\ &\geq \frac{1}{S}\min\{p,1-p\}^{i-1}, &(36)\end{aligned}$$

as well as

$$\begin{aligned}\mu_i^{\uparrow q} &\geq \mu_S^{\uparrow q} \min\{q,1-q\}^{S-i} &(37)\\ &\geq \frac{1}{S}\min\{q,1-q\}^{S-i}, &(38)\end{aligned}$$

where we used the fact that the largest stationary probability among all states must be at least $\frac{1}{S}$. From Lemma 1 and Lemma 2, by ordering μ_i^p in decreasing order and μ_i^q in increasing order, we get the following lower bound on the error probability,

$$\mathsf{P_e} \geq \frac{1}{S} \cdot \max_i \min\left\{\min\{p,1-p\}^{i-1}, \min\{q,1-q\}^{S-i}\right\}. \tag{39}$$

Since both functions are monotone in $1 \leq i \leq S$, one is decreasing from 1 and the other is increasing to 1, the maximum over $i \in [1,S]$ is attained for i such that $\min\{p,1-p\}^{i-1} = \min\{q,1-q\}^{S-i}$, namely, for

$$i = \frac{\log\min\{q,1-q\}}{\log\left(\min\{p,1-p\}\min\{q,1-q\}\right)} S + \log\min\{p,1-p\}. \tag{40}$$

As i must be an integer, the expression above should be rounded up or down. However, asymptotically this has no effect on the bound. Substituting (40) into (39), the theorem follows for the ergodic case.

B. Non-Ergodic Markov chains

Consider first the case where we have only two absorbing states, one for each hypothesis, i.e., assume that we decide $\mathcal{H}_0$ if the process is absorbed in state S and $\mathcal{H}_1$ if the process is absorbed in state 1 . Define X_0 and X_1 as the independent random walks under $\mathcal{H}_0$ and $\mathcal{H}_1$. Then X_0 (resp. X_1) is a stochastic process over the alphabet $[S]$ that starts at s and evolves according to the stochastic matrix $\mathbf{P}(p)$ (resp. $\mathbf{P}(q)$). Define the conditional error probabilities:

$$\begin{aligned}p_0 &= \Pr(1 \in X_0), &(41)\\ p_1 &= \Pr(S \in X_1), &(42)\end{aligned}$$

and hence $\mathsf{P_e} = \frac{1}{2}(p_0 + p_1)$. Define the total distance of a state u to be the smallest sum of lengths of two simple paths from u to 1 and from u to S, and denote it by $\mathsf{td}(u)$. Furthermore, define the occupancy of a state u to be the minimal probability that one of the random walks will visit it, i.e., $\mathsf{occ}(u) \triangleq \min_i \Pr(u^* \in X_i)$. A simple bound on the error probability of any system is the probability of the shortest path from s to the incorrect absorbing state under either hypothesis. However, such a bound may not be tight, since s itself can only be guaranteed to have $\mathsf{td}(s) \leq 2S$. To see this, consider that the shortest path to each state cannot be larger than S, and is exactly S for the linear graph that splits at the last node to either absorbing state. On the other hand, the best possible guarantee we can hope for is total distance of S, which corresponds to a chain in which the shortest paths are non-intersecting. This motivates us to find a state with the smallest possible total distance and a non-negligible occupancy.

Lemma 4. *There exists a state u^* with $\mathsf{td}(u^*) \leq S$ and*

$$\mathsf{occ}(u^*) \geq \frac{1-\max\{p_0,p_1\}}{S}, \tag{43}$$

where p_0 and p_1 are as in (41)*,* (42)*.*

Proof. Let $\mathcal{A}$ (resp. $\mathcal{B}$) denote the collection of all simple paths that start at s and terminate at 1 (resp. S). Let $\mathcal{C}$ be the set of all vertices $v \in [S]$, for which there exist two simple paths $a \in \mathcal{A}$ and $b \in \mathcal{B}$, where v is the last vertex in a that also appears in b. This implies that the sum of path lengths from any $v \in \mathcal{C}$ to 1 and S is smaller than S, i.e., $\forall v \in \mathcal{C}$ we have $\mathsf{td}(v) \leq S$. Define $\tilde{X}_0$ (resp. $\tilde{X}_1$) to be a stochastic process with the distribution of X_0 (resp. X_1) conditioned on the event that X_0 terminated at S (resp. 1). Define U to be the last state on $\tilde{X}_0$ that also appears on $\tilde{X}_1$. Then by definition $\Pr(U \in \mathcal{C}) = 1$, so there must be a state $u^* \in S$ such that

$$\Pr(U = u^*) \geq \frac{1}{|\mathcal{C}|} \geq \frac{1}{S}. \tag{44}$$

This in particular implies that $\Pr(u^* \in \tilde{X}_0) \geq \frac{1}{S}$ and $\Pr(u^* \in \tilde{X}_1) \geq \frac{1}{S}$. Now, the probability of the unconditioned walk X_1,

to pass through u^* is lower bounded by

$$\Pr(u^* \in X_1) \geq \Pr(1 \in X_1) \Pr(u^* \in X_1 | 1 \in X_1) \quad (45)$$
$$= \Pr(1 \in X_1) \Pr(u^* \in \tilde{X}_1) \quad (46)$$
$$\geq \frac{1}{S}(1 - p_1). \quad (47)$$

Similarly bounding $\Pr(u^* \in X_0)$, the lemma follows. □

Proof of Theorem 1 for two absorbing states:

Without loss of generality, we may assume that $\max\{p_0, p_1\} < 1/2$ as otherwise the theorem is trivially true. Furthermore, from Lemma 4 there is some state u^* with $\mathsf{td}(u^*) \leq S$ and $\mathsf{occ}(u^*) \geq \frac{1-\max\{p_0,p_1\}}{S}$. Write

$$\mathsf{P_e} \geq \frac{1}{2} \Pr(u^* \in X_0) \Pr(1 \in X_0 | u^* \in X_0) \quad (48)$$
$$+ \frac{1}{2} \Pr(u^* \in X_1) \Pr(S \in X_1 | u^* \in X_1) \quad (49)$$
$$\geq \frac{1 - \max\{p_0, p_1\}}{2S} (\Pr(1 \in X_0 | u^* \in X_0) + \Pr(S \in X_1 | u^* \in X_1)). \quad (50)$$

Let m_{u^*} be the length of the shortest path from u^* to 1, and recall that we must have a path from u^* to S of length smaller than $S - m_{u^*}$, since that $\mathsf{td}(u^*) \leq S$. Thus,

$$\Pr(1 \in X_0 | u^* \in X_0) + \Pr(S \in X_1 | u^* \in X_1) \quad (51)$$
$$\geq (\min\{p, 1-p\})^{m_{u^*}} + (\min\{q, 1-q\})^{S - m_{u^*}}. \quad (52)$$

Minimizing the lower bound with respect to $m_{u^*} \in [0, S]$ yields

$$m_{u^*} = \frac{\log \min\{q, 1-q\}}{\log \min\{p, 1-p\} + \log \min\{q, 1-q\}} \cdot S, \quad (53)$$

and substituting (53) into (52) implies the theorem for the case of two absorbing states.

Proof of Theorem 1 for the general reducible case:

Consider a Markov chain with K recurrent classes $\mathcal{R}_1, \ldots, \mathcal{R}_K$, and a set $\mathcal{T}$ of transient states with initial state s. Note that if $s \notin \mathcal{T}$ the chain is essentially an ergodic one, hence we consider only $s \in \mathcal{T}$. Define X_0 and X_1 as before, and denote the probability that X_i ends up in class $\mathcal{R}_j$ as

$$\Pr(X_i \to \mathcal{R}_j), \ i = 0, 1, \ j = 1, \ldots, K. \quad (54)$$

We further denote the probability of error under hypothesis $\mathcal{H}_i$ if the initial state were in class $\mathcal{R}_j$ as $\mathsf{P_e}(\mathcal{R}_j | \mathcal{H}_i)$. Consider first the case where the probability of error under $\mathcal{H}_0$ is larger than the probability of error under $\mathcal{H}_1$ in every recurrent class. Then

$$\mathsf{P_e} \geq \frac{1}{2} \min_{1 \leq j \leq K} \mathsf{P_e}(\mathcal{R}_j | \mathcal{H}_0) \quad (55)$$
$$\geq \frac{1}{2} \cdot 2^{-\max_{1 \leq j \leq K} |\mathcal{R}_j| \cdot (d(p,q) + o(1))} \quad (56)$$
$$\geq 2^{-S \cdot (d(p,q) + o(1))}, \quad (57)$$

where $d(p, q)$ was defined in (10) and $o(1)$ is relative to S. Note that in (55) we bound the error probability under $\mathcal{H}_0$ with the smallest error probability across classes, and in (56) we used the fact that the error probability under $\mathcal{H}_0$ in class $\mathcal{R}_j$ is larger than the average error probability, and then used Theorem 1 for the ergodic case.

For the second case, we define the non-empty sets

$$\mathcal{C}_1 = \{\mathcal{R}_k : \mathsf{P_e}(\mathcal{R}_k | \mathcal{H}_0) \geq \mathsf{P_e}(\mathcal{R}_k | \mathcal{H}_1)\}, \quad (58)$$
$$\mathcal{C}_0 = \{\mathcal{R}_k : \mathsf{P_e}(\mathcal{R}_k | \mathcal{H}_0) < \mathsf{P_e}(\mathcal{R}_k | \mathcal{H}_1)\}. \quad (59)$$

For any $k \in \mathcal{C}_1$, we have

$$\mathsf{P_e}(\mathcal{R}_k | \mathcal{H}_0) \geq 2^{-|\mathcal{R}_k| \cdot (d(p,q) + o(1))}, \quad (60)$$

and for any $k \in \mathcal{C}_0$ we have

$$\mathsf{P_e}(\mathcal{R}_k | \mathcal{H}_1) \geq 2^{-|\mathcal{R}_k| \cdot (d(p,q) + o(1))}, \quad (61)$$

according to Theorem 1 for the ergodic case. Now, write

$$\mathsf{P_e} \geq \frac{1}{2} \Pr(X_0 \to \mathcal{C}_1) \min_{k \in \mathcal{C}_1} \mathsf{P_e}(\mathcal{R}_k | \mathcal{H}_0) \quad (62)$$
$$+ \frac{1}{2} \Pr(X_1 \to \mathcal{C}_0) \min_{k \in \mathcal{C}_0} \mathsf{P_e}(\mathcal{R}_k | \mathcal{H}_1) \quad (63)$$
$$\geq \frac{1}{2} (\Pr(X_0 \to \mathcal{C}_1) + \Pr(X_1 \to \mathcal{C}_0)) \times 2^{-\max\{\max_{k \in \mathcal{C}_1} |\mathcal{R}_k|, \max_{k' \in \mathcal{C}_0} |\mathcal{R}_{k'}|\} \cdot (d(p,q) + o(1))} \quad (64)$$
$$= \frac{1}{2} (\Pr(X_0 \to \mathcal{C}_1) + \Pr(X_1 \to \mathcal{C}_0)) \times 2^{-\max_k |\mathcal{R}_k| \cdot (d(p,q) + o(1))}. \quad (65)$$

Consider a chain with $|\mathcal{T}|+2$ states, obtained from the original chain by merging the states in $\mathcal{C}_0$ and $\mathcal{C}_1$ into two respectively absorbing states. Then Lemma 4 holds, with

$$p_0 = \Pr(X_0 \to \mathcal{C}_1), \quad (66)$$
$$p_1 = \Pr(X_1 \to \mathcal{C}_0). \quad (67)$$

According to (65), we may assume that $\max_i p_i < 1/2$ as otherwise the theorem is trivially true. Now, repeating the same arguments as in the proof of the two absorbing states, one can show that

$$\Pr(X_0 \to \mathcal{C}_1) + \Pr(X_1 \to \mathcal{C}_0) \quad (68)$$
$$\geq \frac{1 - \max_i p_i}{|\mathcal{T}| + 2} \cdot 2^{-(|\mathcal{T}|+2) \cdot (d(p,q) + o(1))}. \quad (69)$$

The proof follows by noting that $\max_k |R_k| + |\mathcal{T}| \leq S - 1$.

REFERENCES

[1] M. E. Hellman and T. M. Cover, "Learning with finite memory," *The Annals of Mathematical Statistics*, pp. 765–782, 1970.

[2] H. Robbins, "A sequential decision problem with a finite memory," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 42, no. 12, p. 920, 1956.

[3] T. M. Cover, "A note on the two-armed bandit problem with finite memory," *Information and Control*, vol. 12, no. 5, pp. 371–377, 1968.

[4] T. M. Cover *et al.*, "Hypothesis testing with finite statistics," *The Annals of Mathematical Statistics*, vol. 40, no. 3, pp. 828–835, 1969.

[5] M. E. Hellman and T. M. Cover, "On memory saved by randomization," *The Annals of Mathematical Statistics*, vol. 42, no. 3, pp. 1075–1078, 1971.

[6] J. Von Neumann, "13. various techniques used in connection with random digits," *Appl. Math Ser*, vol. 12, no. 36-38, p. 5, 1951.

[7] B. Chandrasekaran, "Finite-memory hypothesis testing–a critique (corresp.)," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 494–496, 1970.

[8] M. Hellman, "The effects of randomization on finite-memory decision schemes," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 499–502, 1972.

[9] M. Hellman and T. Cover, "A review of recent results on learning with finite memory," in *2nd International Symposium on Information Theory*, pp. 289–294, 1973.

[10] B. Shubert and C. Anderson, "Testing a simple symmetric hypothesis by a finite-memory deterministic algorithm," *IEEE Transactions on Information Theory*, vol. 19, no. 5, pp. 644–647, 1973.

[11] J. Steinhardt and J. Duchi, "Minimax rates for memory-bounded sparse linear regression," in *Conference on Learning Theory*, pp. 1564–1587, 2015.

[12] J. Steinhardt, G. Valiant, and S. Wager, "Memory, communication, and statistical queries," in *Conference on Learning Theory*, pp. 1490–1516, 2016.

[13] R. Raz, "Fast learning requires good memory: A time-space lower bound for parity learning," *Journal of the ACM (JACM)*, vol. 66, no. 1, p. 3, 2018.

[14] Y. Dagan and O. Shamir, "Detecting correlations with little memory and communication," in *Conference On Learning Theory*, pp. 1145–1198, 2018.

[15] Y. Dagan, G. Kur, and O. Shamir, "Space lower bounds for linear prediction in the streaming model," in *Conference on Learning Theory*, pp. 929–954, 2019.

[16] V. Sharan, A. Sidford, and G. Valiant, "Memory-sample tradeoffs for linear regression with small error," in *Symposium on Theory of Computing (STOC)*, 2019.

[17] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems*, pp. 2328–2336, 2013.

[18] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. 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*, pp. 1011–1020, ACM, 2016.

[19] A. Xu and M. Raginsky, "Information-theoretic lower bounds on Bayes risk in decentralized estimation," *IEEE Transactions on Information Theory*, vol. 63, pp. 1580–1600, March 2017.

[20] Y. Han, A. Ozgur, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," *Proceedings of Machine Learning Research*, vol. 75, 2018.

[21] J. Acharya, C. L. Canonne, and H. Tyagi, "Distributed simulation and distributed inference," *arXiv preprint arXiv:1804.06952*, 2018.

[22] L. P. Barnes, Y. Han, and A. 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)*, pp. 16–23, IEEE, 2018.

[23] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[24] W. Feller, "An introduction to probability theory and its applications, vol. 2," 1968.