Effective Memory Shrinkage in Estimation

Ayush Jain* Himanshu Tyagi†

***Abstract*—It is known that a processor with limited memory consisting of an m-state machine can distinguish two coins with biases that differ by $1/m$. On the other hand, the best additive accuracy with which the same processor can estimate the bias of a coin is only $1/\sqrt{m}$. We demystify this apparent shrinkage in memory by showing that for any such estimator using an m-state machine, there exist two values of the bias that are $1/\sqrt{m}$ apart but for which the effective number of states available to resolve them is only $O(\sqrt{m})$. Building on this result, we show that the number of bits of memory required to estimate a bias in the interval $(a, a2^{\alpha})$ with a multiplicative accuracy of $2^{\pm\delta}$ is $\log(\alpha/\delta^2)$, up to an additive constant. In fact, we show that the lower bound is attained by a *Gaussian counter*, namely a probabilistic counter whose stationary distribution has a Gaussian form. This gives a precise characterization of memory-complexity of bias estimation along with a heuristically appealing family of optimal estimators. Underlying our results are new bounds for estimation of the natural parameter of a discrete exponential family, which maybe of independent interest.**

I. Introduction

How much memory is required to estimate the bias of a coin? When the bias can only take one of two values differing by ϵ, a fundamental result of Cover and Hellman [7] shows that one needs a finite-state machine with at least $O(1/\epsilon)$ states. On the other hand, when the bias can take any value in the interval $[0, 1]$, a seminal work of Leighton and Rivest [8] shows that to estimate the bias up to an additive accuracy of ϵ, one needs at least $O(1/\epsilon^2)$ states. Thus, there is a shrinkage in the effective memory available for estimation as the domain of possible values of the bias increases. Furthermore, in both these works the optimal estimators are probabilistic counters, albeit of seemingly different forms.

We demystify the memory shrinkage phenomenon by showing that when estimating the unknown bias p from the interval, say, $[1/3, 1/2]$ using an m-state machine, the effective number of states available is only $O(\sqrt{m})$. Specifically, we show that for any estimator with m states, there exist values of p and p' of the bias such that $|p - p'|$ is more than $1/\sqrt{m}$, but the effective support sets of their equilibrium distributions are contained in a set of cardinality $O(\sqrt{m})$. In fact, the result we derive is much stronger and gives a bound for effective memory available as a function of the size of the uncertainty interval of the bias.

The memory shrinkage result further implies a lower bound for the memory (in bits) required to estimate an unknown bias in the interval $[a, a2^{\alpha}]$ up to a desired multiplicative accuracy. Interestingly, we show that this lower bound is attained by a class of simple probabilistic counters, termed the *Gaussian counter*, where the equilibrium distribution has a discrete Gaussian form. We depict a canonical probabilistic counter for our problem in Figure 1; the specific Gaussian counter that optimally solves our proposed problem is obtained by using a_is given in (14) below.

The counters prescribed in the works of Cover and Hellman and Leighton and Rivest can be recovered as special cases of our general estimator. Furthermore, we strengthen those results by providing an estimator that provides a desired multiplicative accuracy (in contrast to the additive accuracy guarantee considered in these works) and can incorporate the knowledge of the domain of the unknown bias.

Our treatment is based on a reduction of the memory limited estimation problem to a problem of estimation of the natural parameter of an exponential family. Unlike the standard version of this problem studied in the classic statistics literature, we are now allowed to choose any constants for the exponential family. This reduction is based on the Markov chain tree theorem of [8] which gives a closed-form expression for the equilibrium distribution of a Markov chain and reveals the exponential form we exploit.

Memory limited estimation has a long history and was studied in both information theory and computer science communities, perhaps with different motivations (see [7], [8], [5], [6], [4]). While this classic thread seems to have faded, related problems have been considered in the context of streaming algorithms (see, for instance, [1]) and in the context of communication constrained distributed inference ($cf.$ [9]). We revisit this classic field motivated by applications arising in IoT where the end sensor devices are memory starved with only 100s of KB of RAM available to them. Our work here is a first step towards realizing machine intelligence, at least in part, on such edge devices.

Our main results, including the reduction to the exponential family and the memory shrinkage theorem, are given in the next section. Section III gives a proof of the memory shrinkage theorem and the resulting lower bound. Section IV gives the optimal estimator and, in particular, shows the optimality of Gaussian counters. The final section contains discussion on ongoing work and extensions.

II. Formulation and main results

Let $\{X_t\}_{t=1}^{\infty}$ denote an independent and identically distributed sequence with each X_i a $\texttt{Ber}(p)$ random variable, where $p \in [0, 1]$ is unknown. A memory-limited estimator

*Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA. Email: ayjain@ucsd.edu

†Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India. Email: htyagi@iisc.ac.in

for p consists of a time-invariant Markov chain with a finite state space $\mathcal{M}$ where the state transition at time t can depend on X_t. Formally, an *m-state estimator* for p is a tuple $\mathcal{E} = (\Pi_0, T_0, T_1, \hat{p})$ where Π_0 is a distribution on the state space $\mathcal{M} = \{1, ..., m\}$, T_1 and T_2 are two transition probability matrices (TPMs), and $\hat{p}$ is a mapping from $\mathcal{M}$ to $[0, 1]$. To estimate p using $\mathcal{E}$, we start with the initial distribution Π_0 on $\mathcal{M}$, and make a transition from state i to j at time t with probability $T_{X_t}(i, j)$; the overall process is a Markov chain M_p with TPM $T_p = pT_1 + (1 - p)T_0$. At time t, we can obtain an estimate of p as $\hat{p}(M_p(t))$ where $M_p(t)$ is the state of Markov chain at time t.

We are interested in the limiting behavior of this estimator as t goes to infinity. This limiting behavior is determined by the long term transition probabilities of the Markov chain T_p, which in turn are given by the matrix

$$\overline{T}_p = \lim_{t\to\infty} \frac{1}{t} \sum_{i\in[t]} T_p^{i-1}.$$

Specifically, the limiting behavior of $\mathcal{E}$ is captured by the estimate $\hat{p}(M_p(\infty))$ of p, where $M_p(\infty)$ is a random variable with distribution $\Pi_0\overline{T}_p$.

In the section below, we note that the optimal estimator has a very simple structure, roughly that of a counter. Using this structural result, we next observe that the memory constraint estimation problem is essentially equivalent to that of estimating the natural parameter of a discrete exponential family with support size m, and designing an m-state estimator is tantamount to defining the constants of the exponential family.

A. Memory constraint estimation to parametric estimation

We closely follow the approach in [8]; in particular, we rely on the *Markov chain tree theorem* of [8] (see [2] for an alternative proof) to obtain a closed-form expression for $\pi_p = \Pi_0\overline{T}_p$. This result states that $\pi_p(i)$ is proportional to the sum of weights of the spanning trees of the directed graph representing the Markov transition matrix with root at i, where the weight of a tree is the product of probabilities on the edges of the tree. An application of the Markov chain tree theorem yields the following form for π_p (*cf.* [8])

$$\pi_p(j) = \frac{\sum\limits_{i\in[m]} a_{ij} p^{i-1}(1-p)^{m-i}}{\sum\limits_{i\in[m]}\sum\limits_{j\in[m]} a_{ij} p^{i-1}(1-p)^{m-i}}, \tag{1}$$

where $a_{i,j} \geq 0$ for all $i, j \in \mathcal{M}$. Note that states with $a_i := \sum\limits_{j\in[m]} a_{ij} = 0$ have probability $\pi_p(i)$, and so, we can assume[1] $a_i > 0$ for all $i \in \mathcal{M}$.

Our first observation is that it suffices to focus on *probabilistic counters*, namely estimators for which the underlying Markov chain, when at state i, either moves to $(i + 1)$ with some probability p_i on observing heads or $(i - 1)$ with some probability q_i on observing tails. To state our result, we recall a notion of Blackwell [3] which allows us to compare two experiments. Specifically, an experiment described by a parametric family $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ on $\mathcal{X}$ is *better* than $\mathcal{Q} = \{\mathrm{Q}_\theta, \theta \in \Theta\}$ on $\mathcal{Y}$ iff we can find a channel $W : \mathcal{X} \to \mathcal{Y}$ such that Q_θ is the distribution of the output of the channel when the input is distributed as P_θ. When $\mathcal{P}$ is better than $\mathcal{Q}$, a decision based on $\mathcal{P}$ will outperform that based on $\mathcal{Q}$ for any loss function.

[1]Performance corresponding to $a_i = 0$ can be attained by putting an setting $a_i = \eta$ and making η arbitrarily small.

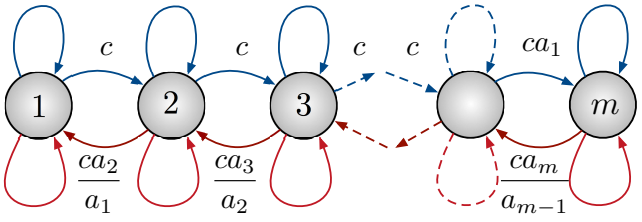

Fig. 1. A probabilistic counter with π_p of the form (3). Here c denotes $1/\max\{a_1, a_2/a_1, ..., a_m/a_{m-1}, 1\}$. The counter is increased probabilistically on observing heads or decreased probabilistically on observing tails.

Proposition 1 (Counters are optimal)**.** *Given an estimator $\mathcal{E}$, we can find a probabilistic counter $\mathcal{E}_c$ such that $\mathcal{E}_c$ is better than $\mathcal{E}$.*

To prove this result, we need the following characterization of π_p for probabilistic counters.

Lemma 2. *For a probabilistic counter $\mathcal{E}$, the distribution π_p is of the form*

$$\pi_p(i) = \frac{a_i p^{i-1}(1-p)^{m-i}}{\sum\limits_{j\in[m]} a_j p^{j-1}(1-p)^{m-j}}, \quad \forall\, i \in \mathcal{M}, \tag{2}$$

where $a_i > 0$ for all i.

Conversely, for every pmf P *of the form* (2)*, we can find a probabilistic counter $\mathcal{E}$ such that $\pi_p = \mathrm{P}$.*

The proof follows from (1) – a probabilistic counter for which π_p has the form in (2) is given in Figure 1 .

Proof of Proposition 1. For a given estimator $\mathcal{E}$, let π_p be given by (1). Let channel $W : \mathcal{M} \to \mathcal{M}$ be given by $W(j|i) = a_{ij}/a_i$. Then, π_p is the output distribution of W when the input distribution is given by (2), whereby the latter family of distributions is better than the former. But by Lemma 2 the distribution in (2) corresponds to the probabilistic counter given in Figure 1, which proves the claim. □

Therefore, it suffices to search for an optimal estimator in the class of probabilistic counters. Note that upon substituting $\theta = p/(1-p)$, the parametric family in (2) can be re-expressed as

$$\pi_p(i) = \frac{a_i \theta^{i-1}}{\sum\limits_{j\in[m]} a_j \theta^{j-1}}, \tag{3}$$

namely a discrete exponential family with natural parameter $\log\theta$. Furthermore, for $\theta(p) = p/(1-p)$,

$$\frac{1}{2}\left|\log\frac{\theta(\hat{p})}{\theta}\right| \leq \max\left\{\left|\log\frac{\hat{p}}{p}\right|, \left|\log\frac{1-p}{1-\hat{p}}\right|\right\} \leq \left|\log\frac{\theta(\hat{p})}{\theta}\right|.$$

Thus, in view of Proposition 1 and Lemma 2, the problem of estimating the bias of a coin with multiplicative accuracy using an m-state estimator is essentially equivalent to that of estimating with additive accuracy the natural parameter of the discrete exponential family (3) on $[m]$. In particular, we need to find the exponential family in this class (design constants a_is) that yields the least approximation error for a fixed m. From here on, we simply consider this alternative problem.

B. Estimation of natural parameter

Distributions corresponding to 3, can be expressed as an exponential family $\mathcal{P} = \{P_\theta, \theta > 0\}$ with support $\{1, ..., m\}$ where

$$\mathrm{P}_\theta(i) = c(\theta) a_i \theta^i, \tag{4}$$

for some $a_i > 0$ and $c(\theta)$ denoting the normalizing constant (that depends on a_is). Suppose that we know a priori that θ belongs to an interval $\mathcal{I}_{a,\alpha} = [a, a2^\alpha]$ for $a > 0$ and $\alpha > 0$. We consider a *probably approximately correct* (PAC) formulation for estimating θ in $\mathcal{I}_{a,\alpha}$ by observing a sample from P_θ. The parameter of the family that we get to design is the vector $(a_1, ..., a_m)$. Specifically, denoting by $M \in [1, m]$ a sample from P_θ, the minimax probability of error for estimating θ is given by[2]

$$\epsilon(m, \alpha, \delta) = \inf_{(a_1,...,a_m)} \sup_{\theta \in \mathcal{I}_{a,\alpha}} \mathbb{P}\left(\left| \log \frac{\hat{\theta}(M)}{\theta} \right| \geq \delta \right).$$

We are interested in the quantity $C(\epsilon, \alpha, \delta)$ defined to be the least value of $\log m$ such that $\epsilon(m, \alpha, \delta) \leq \epsilon$. Note that in view of the discussion of the previous section, $C(\epsilon, \alpha, \delta)$ represents the *memory-complexity* of estimating p, namely memory required in bits estimating p.

We now prove a *memory shrinkage theorem*, but we describe it in the context of the exponential family given in (4).

To describe the result, we define the *t-essential upper bound* for a random variable X by

$$u_X(t) = \sup\left\{x : \mathbb{P}(X \geq x) \geq 2^{-t}\right\},$$

and the *t-essential lower bound bound* for X by

$$l_X(t) = \inf\left\{x : \mathbb{P}(X \leq x) \geq 2^{-t}\right\}.$$

Note that the effective support-size of X is $u_X(t) - l_X(t)$, or equivalently we can talk about the size of the support of distribution of X. For a distribution P_θ from the exponential family (4) with fixed constants $(a_1, ..., a_m)$, let $u_\theta(t)$ and $l_\theta(t)$, respectively, denote[3] the t-essential upper and lower bounds for P_θ.

Theorem 3 (Memory shrinkage theorem)**.** *Given an interval $\mathcal{I}_{a,\alpha}$ and an exponential family of the form* (4)*, for every $0 < \delta < \sqrt{\alpha/m}$ there exists a $\theta \in \mathcal{I}_{a,\alpha}$ such that*

$$\max\{u_{2^\delta\theta}(t), u_\theta(t)\} - \min\{l_{2^\delta\theta}(t), l_\theta(t)\} \leq (2t+7)\sqrt{\frac{m}{\alpha}}. \tag{5}$$

In particular, there exist $\theta \in [\frac{1}{2}, 1]$ and $\theta' > \theta + \frac{1}{\sqrt{m}}$ such that

$$\max\{u_{\theta'}(t), u_\theta(t)\} - \min\{l_{\theta'}(t), l_\theta(t)\} \leq (2t+7)\sqrt{m}. \tag{6}$$

Remark 1. It follows from (6) that there exist two biases p, p' in $(1/3, 1/2)$ that are at least $1/\sqrt{m}$ apart for which the supports of $\pi_{p'}$ and π_p are contained in a set of cardinality $O(\sqrt{m})$.

As a corollary, we obtain the following bound for the divergence between P_θ and $P_{2^\delta\theta}$.

Corollary 4. *Given an interval $\mathcal{I}_{a,\alpha}$, an exponential family of the form* (4)*, and $0 < \delta < \sqrt{\alpha/m}$, there exists $\theta \in \mathcal{I}_{a,\alpha}$ such that $\theta_1 = 2^\delta\theta$ satisfies*

$$D(P_\theta \| P_{\theta_1}) \leq c \cdot \frac{m\delta^2}{\alpha},$$

where $c > 0$ is a constant independent m, a, α, δ, and the constants $(a_1, ..., a_m)$ of the exponential family.

The proof of Theorem 3 and Corollary 4 are given in Section III.

The corollary above yields a difficult pair of hypothesis with a small distance between the observed distributions. Combining this with Le Cam's two point method, we obtain a lower bound for $C(\epsilon, \alpha, \delta)$. Note that we need an $\Omega(\sqrt{\alpha/m})$ lower bound for δ, and therefore, we can make the assumption $\delta < \sqrt{\alpha/m}$ needed for Corollary 4.

In fact, we can find an estimator attaining this lower bound up to a constant gap (depending mildly only on ϵ). The following result is obtained.

Theorem 5. *For every $\epsilon \in (0, 1/4)$,$0 < a$, $0 < \alpha$, and $0 < \delta$,*

$$C(\epsilon, \alpha, \delta) = \log \frac{\alpha}{\delta^2} + O(\log\log \frac{1}{\epsilon}).$$

The optimal choice of $(a_1, ..., a_m)$ and the analysis for the corresponding family (4) is given in Section IV. Interestingly, for the corresponding exponential family, the distribution P_θ has a Gaussian form – we term the corresponding probabilistic counter a Gaussian counter. Thus, Gaussian counters attain the bound promised in Theorem 5 and are order-wise optimal.

III. The proof of memory shrinkage theorem and lower bounds

Fix the exponential family $\mathcal{P}$ to be that given in (4). All our proofs in This section rely on the following property of $\mathcal{P}$:

$$\frac{P_{\theta_2}(i)}{P_{\theta_1}(i)} = \frac{\theta_2}{\theta_1}^{i-j} \frac{P_{\theta_2}(j)}{P_{\theta_1}(j)}. \tag{7}$$

In fact, we only need the lower bound on left-side implied by (7), along with the following property of monotonicity of family in the first-order stochastic dominance.

[2]We omit the dependence on a in our notation since the minimax error doesn't depend on it.

[3]For brevity, we have omitted the dependence of u_θ and l_θ on $(a_1, ..., a_m)$ from our notation.

Lemma 6 (Monotonicity)**.** *For any fixed $k \in [m]$, $P_\theta(M \geq k)$ is a nondecreasing function of θ.*

We skip the proof, which uses (7). As a simple corollary of this monotonicity property, we obtain the monotonicity of t-essential upper and lower bounds $u_\theta(t)$ and $l_\theta(t)$, respectively.

Lemma 7. *For any fixed $t \geq 1$, $l_\theta(t)$ and $u_\theta(t)$ are nondecreasing functions of θ.*

Note that $u_\theta(1) = l_\theta(1)$ corresponds to the median of M under P_θ. As we increase t, $u_\theta(t)$ and $l_\theta(t)$ deviate and satisfy $P_\theta(M \in [l_\theta(t), u_\theta(t)])$ is at least $1 - 2/2^t$. The next lemma provides a bound on how fast $u_\theta(t)$ and $l_\theta(t)$ move away from the median of P_θ – the speed is logarithmic in t. In fact, we show a stronger bound, which we need, where we bound the deviation from the medians of $P_{\theta'}$ where θ' can be a factor γ away from θ.

Lemma 8 (Spread of quantiles)**.** *For a $\gamma > 1$, we have*

$$u_\theta(t) \leq u_{\theta\gamma}(1) + \frac{t+2}{\log(\gamma)}, \text{ and} \tag{8}$$

$$l_\theta(t) \geq u_{\theta/\gamma}(1) - \frac{t+2}{\log(\gamma)}. \tag{9}$$

Proof. For any $\theta, \theta' \in [a, a2^\alpha]$,

$$\frac{P_{\theta'}(M \leq u_{\theta'}(1))}{P_\theta(M \leq u_{\theta'}(1))} \geq 1/2.$$

Thus, we can find $j \leq u_{\theta'}(1)$ such that $\dfrac{P_{\theta'}(j)}{P_\theta(j)} \geq 1/2$. Then, setting $\theta' = \gamma\theta$ and using (7), for every $i \geq u_{\theta'}(1) + l$ we get $\dfrac{P_{\theta'}(i)}{P_\theta(i)} \geq \gamma^l/2$, whereby

$$P_\theta(M \geq u_{\theta'}(1) + l) \leq \frac{2}{\gamma^l} P_{\theta'}(M \geq u_{\theta'}(1) + l).$$

In particular, upon setting $l = (t+2)/\log\gamma$ we get (8); the bound in (9) can be shown similarly. □

Proof of Theorem 3. We have form Lemma 8 that

$$u_{\gamma\theta}(t) - l_\theta(t) \leq u_{\gamma^2\theta}(1) - u_{\theta/\gamma}(1) + \frac{2t+4}{\log(\gamma)}.$$

Thus, in view of Lemma 7 and the inequality above, Theorem 3 will be obtained upon substituting $\gamma = 2^{\sqrt{\alpha/m}}$ if we can exhibit a $\theta \in [a, a2^\alpha]$ such that

$$u_{\gamma^2\theta}(1) - u_{\theta/\gamma}(1) \leq \frac{3m\log\gamma}{\alpha}. \tag{10}$$

Indeed, such a θ can be found. Let $\theta_0 = a$ and $\theta_i = \gamma^{3i}$, $i = 1, ..., N$ where $N = \lfloor \alpha/3\log\gamma \rfloor$. The required θ is found upon noting that

$$\frac{1}{N-1}\sum_{i=1}^{N-1} u_{\theta_{i+1}}(1) - u_{\theta_i}(1) = \frac{u_{\theta_N}(1) - u_{\theta_1}(1)}{N-1} \leq \frac{m-1}{N-1}.$$

To prove (6), fix $a = 1/2$, $\alpha = 1$, and $\gamma = 1/\sqrt{m}$ and note that $(2^{1/\sqrt{m}} - 1)$ is $O(1/\sqrt{m})$, whereby $2^\gamma\theta - \theta$ is $O(1/\sqrt{m})$. □

Proof of Corollary 4. We will show a weaker upper bound of $O(\sqrt{m\delta^2/\alpha})$; the square improvement claimed in Corollary 4 is obtained using a Taylor series approximation of KL divergence and is skipped here due to lack of space.

To that end, we show that for a θ satisfying (5) and $i \in [l_\theta(t), u_{\theta\gamma}(t)]$

$$\log\frac{P_\theta(i)}{P_{2^\delta\theta}(i)} < (2t+29)\sqrt{\frac{m\delta^2}{\alpha}}. \tag{11}$$

Then,

$$P_\theta\left(\left\{i : \log\frac{P_\theta(i)}{P_{\beta\theta}(i)} > (2t+29)\sqrt{\frac{m\delta^2}{\alpha}}\right\}\right) \leq \frac{1}{2^{t-1}},$$

and so,

$$D(P_{\beta\theta}||P_\theta) = \mathbb{E}_{M\sim P_\theta}\left\{\ln\frac{P_\theta(M)}{P_{\beta\theta}(M)}\right\} = O\left(\sqrt{\frac{m\delta^2}{\alpha}}\right),$$

which proves Corollary 4.

It remains to prove (11), which will follow upon showing that

$$\frac{P_\theta(l_\theta(1))}{P_{2^\delta\theta}(l_\theta(1))} < 2^{22\sqrt{\frac{m\delta^2}{\alpha}}}. \tag{12}$$

Indeed, since every $i \in [l_\theta(t), u_{2^\delta\theta}(t)]$ satisfies $|i - l_\theta(1)| < (2t+7)\sqrt{m/\alpha}$, by (7) we get

$$\begin{aligned}\frac{P_\theta(i)}{P_{2^\delta\theta}(i)} &\leq 2^{\delta(2t+7)\sqrt{m/\alpha}}\frac{P_\theta(l_\theta(1))}{P_{2^\delta\theta}(l_\theta(1))} \\ &\leq 2^{\delta(2t+29)\sqrt{m/\alpha}}.\end{aligned} \tag{13}$$

Finally, we complete the proof by showing (12). We note some basic properties of the intervals I_t defined by

$$I(t) = [l_\theta(t), u_{2^\delta\theta}(t)].$$

Specifically, note that

(i) The sequence of intervals $\{I(t), t \geq 1\}$ is is nondecreasing, *i.e.*, $I(t-1) \subset I(t), \quad t \geq 2$;

(ii) $I(t)^c$ has exponentially small mass, *i.e.*,

$$P_\theta(I(t)) \geq 1 - 2^{-t}, \quad t \geq 2;$$

(iii) by (13),

$$P_{2^\delta\theta}(I(t) \setminus I(t-1)) \geq f(t) P_\theta(I(t) \setminus I(t-1)),$$

where

$$f(t) = \frac{P_\theta(l_\theta(1))}{P_{2^\delta\theta}(l_\theta(1))} 2^{-(2t+7)\sqrt{m\delta^2/\alpha}}.$$

Note that $f(t)$ is decreasing in t. From these properties, we derive an upper bound for $c = \dfrac{P_\theta(l_\theta(1))}{P_{2^\delta\theta}(l_\theta(1))}$ as follows:

$$\begin{aligned}1 &= P_{2^\delta\theta}(I(2)) + \sum_{t\geq 2} P_{2^\delta\theta}(I(t) \setminus I(t-1)) \\ &\geq f(2)P_\theta(I(2)) + \sum_{t>2} f(t) P_\theta(I(t) \setminus I(t-1)).\end{aligned}$$

Furthermore, for every $t > 2$,

$$\begin{aligned} & f(t)P_\theta(I(t) \setminus I(t-1)) \\ &= \frac{f(t)}{2^t} + f(t)[P_\theta(I(t)) - 2^{-t}] \\ &\qquad - f(t)[P_\theta(I(t-1)) - 2^{-t-1}] \\ &\geq \frac{f(t)}{2^t} + f(t)[P_\theta(I(t)) - 2^{-t}] \\ &\qquad - f(t-1)[P_\theta(I(t-1)) - 2^{-t-1}]. \end{aligned}$$

On combining the two bounds above, we get

$$\begin{aligned} 1 &\geq \sum_{t\geq 2} \frac{f(t)}{2^{t-1}} \\ &= c \sum_{t\geq 2} 2^{-(2t+7)\sqrt{m\delta^2/\alpha}-t+1} \\ &= c \frac{2^{-11\sqrt{m\delta^2/\alpha}-1}}{1 - 2^{-2\sqrt{m\delta^2/\alpha}-1}}, \end{aligned}$$

whereby

$$c \leq 2.2^{11\sqrt{m\delta^2/\alpha}} - 2^{9\sqrt{m\delta^2/\alpha}} \leq 2^{22\sqrt{m\delta^2/\alpha}},$$

where the previous inequality uses $2x - 1 \leq x^2$. Thus, we have established (12), and the proof is complete. □

IV. Optimal estimator

Define $\Delta = (2/\delta)\log(4/\epsilon)$ and assume it to be an integer for simplicity. For i in the set $\{1, ..., m\}$, let

$$a_i = \begin{cases} 2^{-\frac{\delta(i-1)^2}{2\Delta}} a^{-(i-1)}, & i \mod \Delta = 1 \\ \eta, \quad \text{otherwise}, \end{cases} \tag{14}$$

where η is chosen to be arbitrarily small. Then, for $\theta_k = a2^{k\delta}$:

$$P_{\theta_k}(j\Delta + 1) \propto \left(\frac{\epsilon}{4}\right)^{j^2 - 2kj} \propto \left(\frac{\epsilon}{4}\right)^{(j-k)^2}.$$

Note that in the limit as η goes to 0, $P_\theta(M \neq k\Delta + 1) \leq \epsilon$. Consider the estimate $\hat{\theta}$ that simply declares the estimate of θ as $a2^{(i-1)\delta/\Delta}$ when Δ divides $(i-1)$ and an arbitrary estimate, say a, otherwise. Then, when $\theta = \theta_k$, $P_{\theta_k}(\hat{\theta}(M) = \theta) \geq 1 - \epsilon$. Also, using the monotonicity property given in Lemma 6, for $\theta < \theta_k$

$$P_\theta(M \geq k\Delta + 2) \leq P_{\theta_k}(M \geq k\Delta + 2) \leq \frac{\epsilon}{2},$$

which implies that

$$P_\theta(\hat{\theta}(M) > \theta_k) \leq \frac{\epsilon}{2},$$

Similarly, when $\theta > \theta_{k-1}$, $P_\theta(M \leq (k-1)\Delta) \leq \frac{\epsilon}{2}$ which implies that

$$P_\theta(\hat{\theta}(M) < \theta_{k-1}) \leq \frac{\epsilon}{2}.$$

Therefore, for every $\theta \in [\theta_{k-1}, \theta_k]$, our proposed estimator outputs either θ_{k-1} or θ_k with probability greater than $1 - \epsilon$.

Also, the number of states m required for this estimator to work for all $\theta \in [a, a2^\alpha]$ must satisfy

$$\frac{m}{\Delta} \geq \frac{\alpha}{\delta}. \tag{15}$$

Thus, this estimator needs $m = O(\frac{\alpha \log(1/\epsilon)}{\delta^2})$ and matches the bound claimed in Theorem 5.

Note that the probability of state i under θ_k has the Gaussian form

$$\log P_{\theta_k}(i) = -c_1(i-k)^2 + c_2, \quad (i-1) \mod \Delta = 0.$$

Thus, we call the counter corresponding to (14) a Gaussian counter (see Figure 1 for a possible choice for the optimal probabilistic counter).

We close with the remark that this seemingly ad-hoc probabilistic counter has been constructed in an attempt to ensure that the inequality in (10) is satisfied with equality for all θ.

V. Discussion

While Theorem 5 provides a rather complete characterization of the memory required for estimating the bias of an unknown bit, a nagging gap that remains is obtaining a lower bound that reflects the optimal dependence on ϵ. Indeed, we conjecture that Gaussian counters outperform any other probabilistic counter and, therefore, expect the $O(\log\log\epsilon)$ dependence on ϵ is tight as well.

It is of interest to consider the general problem of estimating a discrete pmf on a k-ary alphabet within a desired total variation distance. Our preliminary work suggests that the shrinkage phenomenon is restricted to the problem of estimating the bias of a bit, and once we account for that, the memory scales as $(k-1)\log 1/\delta$. Showing this result formally is work in progress.

References

[1] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. ACM, 1996, pp. 20–29.

[2] V. Anantharam and P. Tsoucas, "A proof of the markov chain tree theorem," *Stochastic and Probability letters*, pp. 189–192, June 1989.

[3] D. Blackwell *et al.*, "Comparison of experiments," in *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. The Regents of the University of California, 1951.

[4] T. Cover and M. Hellman, "The two-armed-bandit problem with time-invariant finite memory," *IEEE Transactions on Information Theory*, vol. 16, no. 2, pp. 185–195, 1970.

[5] T. M. Cover, "Hypothesis testing with finite statistics," *The Annals of Mathematical Statistics*, pp. 828–835, 1969.

[6] M. Hellman, "Finite-memory algorithms for estimating the mean of a gaussian distribution (corresp.)," *IEEE Transactions on Information Theory*, vol. 20, no. 3, pp. 382–384, 1974.

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

[8] F. Leighton and R. Rivest, "Estimating a probability using finite memory," *IEEE Transactions on Information Theory*, vol. 32, no. 6, pp. 733–742, 1986.

[9] O. Shamir, "Fundamental limits of online and distributed algorithms for statistical learning and estimation," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'14, 2014, pp. 163–171.