Distributed Statistical Estimation of High-Dimensional and Nonparametric Distributions

Yanjun Han, Pritam Mukherjee, Ayfer Özgür and Tsachy Weissman
Department of Electrical Engineering, Stanford University
Email: {yjhan, pritamm, aozgur, tsachy}@stanford.edu

***Abstract*—We consider the problem of estimating high-dimensional and nonparametric distributions in distributed networks, where each sensor in the network observes an independent sample from the underlying distribution and can communicate it to a central processor by writing at most k bits on a public blackboard. We obtain matching upper and lower bounds for the minimax risk of estimating the underlying distribution under L_1 loss. Our results reveal that the minimax risk reduces exponentially in k. Instead of relying on strong data processing inequalities for the converse as commonly done in the literature, we build on a new representation of the communication constraint, which leads to a tight characterization of the problem.**

I. Introduction and Main Results

Consider the following density estimation model

$$X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} f$$

where we would like to estimate the unknown density f under L_1 loss. Unlike the traditional nonparametric setting where samples $X_1, \cdots, X_n$ are available to the estimator as they are, we consider a distributed setting where each observation X_i is available at a different sensor and has to be communicated to a central processor by using k bits. We consider two different communication protocols:

1) Independent communication protocol Π_{Ind}: each sensor outputs a k-bit string $Y_i = b_i(X_i) \in [2^k]$ simultaneously (independent of the other sensots) to the central processor, and the final transcript is $Y = (Y_1, \cdots, Y_n)$;
2) Blackboard communication protocol Π_{BB} [1]: all sensors communicate via a publicly shown blackboard while the total number of bits each sensor can write in the final transcript Y is limited by k. Note that when one sensor writes a message (bit) on the blackboard, all other sensors can see the content of the message. We assume that public randomness is available in the blackboard communication protocol.

Upon receiving the transcript Y, the central processor produces an estimate $\hat{f}$ of density f based on the transcript Y and known procotol Π.

As is typical in nonparametric statistics [2], we assume that f possesses some regularity conditions. For the sake of simplicity, suppose that the density f is supported on $[0, 1]$ and has Hölder smoothness $s \in (0, 1]$, though the generalization to unbounded support, higher dimension and higher smoothness is straightforward. Hölder smoothness $s \in (0, 1]$ implies that there exists some constant $L > 0$ such that

$$\sup_{x \neq y \in [0,1]} \frac{|f(x) - f(y)|}{|x - y|^s} \leq L. \tag{1}$$

We denote by $\mathcal{H}^s(L)$ the set of all densities satisfying (1).

The first main result of this paper is as follows:

Theorem 1. *Let $s \in (0, 1], L > 0, k \in \mathbb{N}$. There exist positive constants $c = c(s, L), C = C(s, L)$ such that*

$$\begin{aligned} c\left[(n \cdot 2^k)^{-\frac{s}{2(s+1)}} + n^{-\frac{s}{2s+1}}\right] &\leq \inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{f}} \sup_{f \in \mathcal{H}^s(L)} \mathbb{E}_f \|\hat{f} - f\|_1 \\ &\leq C\left[(n \cdot 2^k)^{-\frac{s}{2(s+1)}} + n^{-\frac{s}{2s+1}}\right] \end{aligned} \tag{2}$$

where the infimum is taken over all possible blackboard communication protocols Π_{BB} and estimators $\hat{f} = \hat{f}(Y)$. Moreover, there exists an independent communication protocol under which the upper bound is attained.

Since it is well-known that the minimax L_1 risk of density estimation over $\mathcal{H}^s(L)$ is $\Theta(n^{-\frac{s}{2s+1}})$ [2], the following corollary is immediate.

Corollary 1. *For nonparametric density estimation over $\mathcal{H}^s(L)$, it is necessary and sufficient to have $k \geq \frac{1}{2s+1} \log_2 n - O(1)$ to achieve the centralized performance without communication constraints.*

Based on parametric reduction, the nonparametric density estimation problem is closely related to the following problem of estimating high-dimensional distributions: let $P = (p_1, \cdots, p_S)$ be a discrete distribution, and $X_1, \cdots, X_n$ be n i.i.d. samples drawn from P. Given the k-bit communication constraints, the problem is to find a rate-optimal estimator $\hat{P}$ of P. By "high-dimensional" we mean that the support size S of the underlying distribution may be comparable to the sample size n. Our main result for distributed high-dimensional distribution estimation is summarized in the following theorem:

Theorem 2. *Let $k \in \mathbb{N}$ and $n \gtrsim S \vee 2^{-k} S^2$. There exist constants c, C independent of n, S, k such that*

$$\frac{cS}{\sqrt{n(2^k \wedge S)}} \leq \inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \leq \frac{CS}{\sqrt{n(2^k \wedge S)}} \tag{3}$$

where $\mathcal{M}_S$ denotes the probability simplex over S elements, and the infimum is taken over all possible blackboard communication protocols Π_{BB} and estimators $\hat{f} = \hat{f}(Y)$. Moreover, there exists an independent communication protocol under which the upper bound is attained.

In the centralized case, it has been shown in [3]–[5] that the minimax L_1 risk of discrete distribution estimation is $\Theta(\sqrt{\frac{S}{n}})$. As a result, we have the following corollary:

Corollary 2. *For discrete distribution estimation over $\mathcal{M}_S$, it is necessary and sufficient to have $k \geq \log_2 S - O(1)$ to achieve the centralized performance.*

Note that achievability in the case $k = \log_2 S$ is in fact trivial: each sensor can transmit its observation X_i as it is by using $\log_2 S$ bits, which yields the centralized performance. Corollary 2 states that the centralized performance could not be achieved with fewer bits (by a potentially smarter scheme) even if interactions and public randomness are allowed; it is essentially necessary to fully communicate the observations to achieve the centralized performance.

Statistical estimation in distributed settings has gained increasing popularity motivated by the fact that modern data sets are often distributed across multiple machines and processors, and bandwidth and energy limitations in networks and within multiprocessor systems often impose significant bottlenecks on the performance of algorithms. There are also an increasing number of applications in which data is generated in a distributed manner and it (or features of it) are communicated over bandwidth-limited links to central processors [6]. Distributed estimation and function computation has been considered in [7]–[11], where strong/distributed data processing inequalities appear as the key technical step in developing converse results. Our converse approach is significantly different. We propose a new representation of the communication constraint, which circumvents the need for strong data processing inequalities. As we discuss in Section IV, this approach can be generalized to other settings and yields stronger results with respect to prior work. A more recent work [12] studied the same discrete distribution estimation problem, and obtained Corollary 2. We, on the other hand, provide a complete characterization as function of the communication constraint k.

Notations: for a finite set A, let $|A|$ denote its cardinality; $[n] \triangleq \{1, 2, \cdots, n\}$; $a \wedge b \triangleq \min\{a, b\}, a \vee b \triangleq \max\{a, b\}$; for non-negative sequences $\{a_n\}$ and $\{b_n\}$, the notation $a_n \lesssim b_n$ (or $b_n \gtrsim a_n, a_n = O(b_n), b_n = \Omega(a_n)$) means $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$, and $a_n \ll b_n$ ($b_n \gg a_n, a_n = o(b_n), b_n = \omega(a_n)$) means $\limsup_{n\to\infty} \frac{a_n}{b_n} = 0$, and $a_n \asymp b_n$ (or $a_n = \Theta(b_n)$) is equivalent to both $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

II. Achievability: the Grouping Idea

In this section, we prove the achievability parts of Theorem 1 and 2 under independent communication protocol Π_{Ind} via the grouping idea. We consider the discrete case first, which serves as a key step of estimating nonparametric densities.

A. Achievability of Theorem 2

Without loss of generality we assume that $k < \log_2 S$. Note that k bits can describe 2^k distinct symbols. We partition the alphabet $[S]$ into $\approx \frac{S}{2^k}$ groups of size $\approx 2^k$ each, and let each sensor be responsible for one group. Specifically, let $S_1, S_2, \cdots, S_m$ (each of size $2^k - 1$) be a partition of the alphabet $[S]$, where without loss of generality we assume that $m = \frac{S}{2^k - 1}$ is an integer. For $j \in [m]$, fix a labeling to the symbols in $S_j = \{a_{j,1}, \cdots, a_{j,2^k-1}\}$, and consider the following encoding function:

$$b_j(s) = \begin{cases} \ell & \text{if } s = a_{j,\ell} \in S_j \\ 2^k & \text{if } s \notin S_j \end{cases} \in [2^k].$$

Next we also partition n sensors in the network to m groups $N_1, \cdots, N_m$ of size $\frac{n}{m}$ each (also assume that $\frac{n}{m}$ is an integer since $n \gtrsim 2^{-k} S^2$), and apply the encoding function $b_j(\cdot)$ to sensors in j-th group N_j. The crucial observation is that, for $s = a_{j,\ell} \in S_j$ and $X \sim P$, we have

$$\mathbb{P}(b_j(X) = \ell) = p_s.$$

Hence, for any $s \in [S]$ with $s = a_{j,\ell}$, the statistic

$$\hat{p}_s = \frac{1}{|N_j|} \sum_{i \in N_j} \mathbb{1}(Y_i = \ell) = \frac{m}{n} \sum_{i \in N_j} \mathbb{1}(b_j(X_i) = \ell)$$

satisfies $\frac{n}{m}\hat{p}_s \sim \mathsf{B}(\frac{n}{m}, p_s)$. By the binomial nature, $\hat{p}_s$ is the natural unbiased estimator for p_s, and the resulting estimator for P is $\hat{P} = (\hat{p}_1, \cdots, \hat{p}_S)$. To analyze the performance of $\hat{P}$, note that

$$\mathbb{E}|\hat{p}_s - p_s| \leq \sqrt{\mathbb{E}(\hat{p}_s - p_s)^2} = \sqrt{\frac{m}{n} p_s (1 - p_s)}$$

and thus by the concavity of $x \mapsto \sqrt{x}$ and $\sum_{s=1}^S p_s = 1$,

$$\mathbb{E}\|\hat{P} - P\|_1 \leq \sum_{s=1}^S \sqrt{\frac{m}{n} p_s} \leq \sqrt{\frac{mS}{n}} = \frac{S}{\sqrt{n(2^k - 1)}}$$

completing the proof of the achievability part of Theorem 2.

B. Achievability of Theorem 1

To estimate the nonparametric density f, a parametric reduction is used. Specifically, we consider some bandwidth $h > 0$ with $S \triangleq h^{-1}$, and

$$f_h(x) = \sum_{j=1}^S \frac{p_j}{h} \mathbb{1}(x \in I_j)$$

where $I_j \triangleq [(j-1)h, jh)$ and $p_j \triangleq \int_{I_j} f(x)dx$. Note that $f \in \mathcal{H}^s(L)$ has Hölder smoothness $s \in (0, 1]$, it is well-known [2] that the piecewise approximation f_h of f satisfies $\|f_h - f\|_1 \leq Ch^s$ with constant $C > 0$ depending on L only.

Moreover, defining $Z_i \in \{1, 2, \cdots, S\}$ to be the index such that $X_i \in I_{Z_i}$, one may verify that $Z_1, \cdots, Z_n \overset{i.i.d}{\sim} P = (p_1, \cdots, p_S)$. As a result, the estimator

$$\hat{f}_h(x) \triangleq \sum_{j=1}^S \frac{\hat{p}_j}{h} \mathbb{1}(x \in I_j)$$

with the vector $\hat{P} = (\hat{p}_1, \cdots, \hat{p}_S)$ given by the achievability part of Theorem 2 satisfies

$$\begin{aligned}\mathbb{E}\|\hat{f}_h - f\|_1 &\le \mathbb{E}\|\hat{f}_h - f_h\|_1 + \|f_h - f\|_1 \\ &= \mathbb{E}\|\hat{P} - P\|_1 + \|f_h - f\|_1 \\ &\le \frac{1}{h\sqrt{n(2^k \wedge h^{-1})}} + Ch^s.\end{aligned}$$

Now the choice of the optimal bandwidth

$$h^* = (n \cdot 2^k)^{-\frac{1}{2(s+1)}} \vee n^{-\frac{1}{2s+1}}$$

completes the proof of the achievability part of Theorem 1.

III. Proof of the Converse

In this section, we prove the converse results of Theorem 1 and 2, showing that the simple grouping idea is in fact optimal. We first prove the lower bound for discrete distribution estimation, which serves as an intermediate step for establishing the lower bound for nonparametric density estimation. The key starting point in our proof is a convenient representation of the communication constrained blackboard protocol Π_{BB}.

A. The Blackboard Communication Protocol

Assume first that there is no public/private randomness, which will be revisited at the end of Section III-B. In this case, the blackboard communication protocol Π_{BB} can be viewed as a binary tree [1], where each internal node v of the tree is assigned a deterministic label $l_v \in [n]$ indicating the identity of the sensor to write the next bit on the blackboard if the protocol reaches node v. The left and right edges departing from v correspond to the two possible values of this bit and are labeled by 0 and 1 respectively. Note that because all bits written on the blackboard up to the current time are observed by all nodes, the sensors can keep track of the progress of the protocol in the binary tree. The value of the bit written by node l_v (when the protocol is at node v) can depend on the sample X_{l_v} observed by this node (and implicitly on all bits previously written on the blackboard encoded in the position of the node v in the binary tree). Therefore, this bit can be represented by a binary function $a_v(x) \in \{0, 1\}$, which we associate with the node v; node l_v evaluates this function on its sample X_{l_v} to determine the value of its bit.

Note that the k-bit communication constraint for each node can be viewed as a labeling constraint for the binary tree; for each $i \in [n]$, each possible path from the root node to a leaf node can visit exactly k internal nodes with label i. In particular, the depth of the binary tree is nk and there is one-to-one correspondance between all possible transcripts $y \in \{0, 1\}^{nk}$ and paths in the tree. Note that a proper labeling of the binary tree together with the collection of functions $\{a_v(\cdot)\}$ (where v ranges over all internal nodes) completely characterizes all possible (deterministic) communication strategies for the sensors. Under this protocol model, the distribution of the transcript Y is

$$\mathbb{P}_{X_1,\cdots,X_n\sim P}(Y = y) = \mathbb{E}_{X_1,\cdots,X_n\sim P} \prod_{v\in\tau(y)} b_{v,y}(X_{l_v})$$

where $v \in \tau(y)$ ranges over all internal nodes in the path $\tau(y)$ corresponding to $y \in \{0, 1\}^{nk}$, and $b_{v,y}(x) = a_v(x)$ if the path $\tau(y)$ goes through the right child of v and $b_{v,y}(x) = 1 - a_v(x)$ otherwise. Due to the independence of $X_1, \cdots, X_n$, we have the following lemma which is similar to the "cut-paste" property [13] for the blackboard communication protocol:

Lemma 1. *The distribution of the transcript Y can be written as follows: for any $y \in \{0, 1\}^{nk}$,*

$$\mathbb{P}_{X_1,\cdots,X_n\sim P}(Y = y) = \prod_{i=1}^{n} \mathbb{E}_{X\sim P} p_{i,y}(X)$$

where $p_{i,y}(x) \triangleq \prod_{v\in\tau(y), l_v=i} b_{v,y}(x)$.

The k-bit communication constraint results in the following important property (see [14] for the proof):

Lemma 2. *For each $i \in [n]$ and $\{x_j\}_{j=1}^n$,*

$$\sum_{y\in\{0,1\}^{nk}} \prod_{j=1}^{n} p_{j,y}(x_j) = 1, \qquad \sum_{y\in\{0,1\}^{nk}} \prod_{j\neq i} p_{j,y}(x_j) = 2^k.$$

B. Proof of Theorem 2

The lower bound $\Omega(\sqrt{\frac{S}{n}})$ for the centralized case has been established, so it suffices to prove the lower bound $\Omega(\frac{S}{\sqrt{n\cdot 2^k}})$. We will establish the lower bound via a testing argument. First, we construct a class of hypotheses on the binary hypercube and relate the minimax risk to some mutual information via a distance-based Fano's inequality. Second, we derive a universal upper bound for the mutual information which holds for any blackboard communication procotol Π_{BB} and encoding strategy $\{a_v(\cdot)\}$.

1) Distance-based Fano's inequality: Let $T \triangleq \frac{S}{2}$ be an integer, and random vector U be uniformly distributed on the binary hypercube $\{\pm 1\}^T$. For each $u \in \{\pm 1\}^T$, we associate with a probability vector $P_u \in \mathcal{M}_S$ given by

$$P_u \triangleq \left(\frac{1}{S} + \delta u_1, \cdots, \frac{1}{S} + \delta u_T, \frac{1}{S} - \delta u_1, \cdots, \frac{1}{S} - \delta u_T\right)$$

where $\delta > 0$ is some parameter to be specified later. To ensure that P_u is a probability vector, we will assume that

$$\delta \in \left(0, \frac{1}{S}\right) \tag{4}$$

throughout the proof, and will get back to it when we specify δ in the end. This construction of P_u is known as the *Paninski's construction* [15].

Now by a standard testing argument [16], we have

$$\inf_{\hat{P}} \sup_{P\in\mathcal{M}_S} \mathbb{E}_P\|\hat{P} - P\|_1 \ge \frac{S\delta}{10} \inf_{\hat{U}} \mathbb{P}\left(d_{\mathrm{H}}(\hat{U}, U) \ge \frac{T}{5}\right)$$

where $d_{\mathrm{H}}(\cdot,\cdot)$ denotes the (unnormalized) Hamming distance. To lower bound $\mathbb{P}\left(d_{\mathrm{H}}(\hat{U}, U) \ge \frac{T}{5}\right)$ for any estimator $\hat{U}$, we use the following distance-based Fano's inequality:

Lemma 3. *[17, Corollary 1] Let random variables V and $\hat{V}$ take value in $\mathcal{V}$, V be uniform on some finite $\mathcal{V}$, and $V - X - \hat{V}$*

form a Markov chain. Let d be any metric on $\mathcal{V}$, and for $t > 0$, define $N_{\max}(t) \triangleq \max_{v\in\mathcal{V}} |v' \in V : d(v, v') \le t|, N_{\min}(t) \triangleq \min_{v\in\mathcal{V}} |v' \in V : d(v, v') \le t|$.

If $N_{\max}(t)+N_{\min}(t) < |\mathcal{V}|$, the following inequality holds:

$$\mathbb{P}(d(V, \hat{V}) > t) \ge 1 - \frac{I(V; X) + \ln 2}{\ln \frac{|\mathcal{V}|}{N_{\max}(t)}}.$$

Applying Lemma 3 to the Markov chain $U - Y - \hat{U}$ with Hamming distance $d_{\mathrm{H}}(\cdot, \cdot)$ and $t = \frac{T}{4}$, we have

$$\inf_{\hat{P}} \sup_{P\in\mathcal{M}_S} \mathbb{E}_P\|\hat{P} - P\|_1 \ge \frac{S\delta}{8}\left(1 - \frac{I(U; Y) + \ln 2}{T/8}\right) \quad (5)$$

where Chernoff bound implies $\frac{|N_{\max}(t)|}{|\mathcal{V}|} \le \exp(-\frac{T}{8})$. Now it remains to upper bound the mutual information $I(U; Y)$.

2) Upper bound of $I(U; Y)$: Let P_0 be the uniform distribution over $[S]$, we have

$$\begin{aligned}
I(U; Y) &\overset{(a)}{\le} \mathbb{E}_U D(P_{Y|U}\|P_{Y|X\sim P_0}) \\
&\overset{(b)}{=} \mathbb{E}_U\mathbb{E}_{Y|U} \sum_{i=1}^{n} \log \frac{\mathbb{E}_{P_U} p_{i,Y}(X)}{\mathbb{E}_{P_0} p_{i,Y}(X)} \\
&\overset{(c)}{\le} \mathbb{E}_U\mathbb{E}_{Y|U} \sum_{i=1}^{n} \left(\frac{\mathbb{E}_{P_U} p_{i,Y}(X)}{\mathbb{E}_{P_0} p_{i,Y}(X)} - 1\right) \\
&\overset{(d)}{=} \mathbb{E}_U \sum_{i=1}^{n} \sum_{y\in\{0,1\}^{nk}} \left(\prod_{j\neq i} \mathbb{E}_{P_U} p_{j,y}(X)\right) \\
&\qquad \cdot \frac{(\mathbb{E}_{P_U} p_{i,y}(X) - \mathbb{E}_{P_0} p_{i,y}(X))^2}{\mathbb{E}_{P_0} p_{i,y}(X)}
\end{aligned}$$

where (a) follows from the variational representation of mutual information $I(X; Y) = \inf_{Q_Y} \mathbb{E}_X D(P_{Y|X}\|Q_Y)$, (b) follows from Lemma 1, (c) is due to $\log x \le x - 1$, and (d) follows from Lemma 1 and the first equality of Lemma 2.

Since $X \sim P_U$ can only take S distinct values, the function $p_{i,y}(\cdot)$ is a length-S vector, and $\mathbb{E}_P p_{i,y}(X) = p_{i,y}^T P$. Hence,

$$\begin{aligned}
\mathbb{E}_U \frac{(\mathbb{E}_{P_U} p_{i,y}(X) - \mathbb{E}_{P_0} p_{i,y}(X))^2}{\mathbb{E}_{P_0} p_{i,y}(X)} &= \frac{\mathbb{E}_U[p_{i,y}^T(P_U - P_0)]^2}{p_{i,y}^T \mathbf{1}/S} \\
&\overset{(e)}{\le} 2\delta^2 S \cdot \frac{p_{i,y}^T p_{i,y}}{p_{i,y}^T \mathbf{1}} \le 2\delta^2 S
\end{aligned}$$

where in (e) we have used $\mathbb{E}_U(P_U - P_0)(P_U - P_0)^T \preceq 2\delta^2 I$, and the last inequality follows from the fact that $p_{i,y}(\cdot)$ is a binary function.

Finally, combining the previous inequalities and invoking Lemma 2, we have

$$I(U; Y) \le 2n\delta^2 S \cdot 2^k. \quad (6)$$

Combining (5) and (6), and choosing $\delta = \frac{c}{\sqrt{n2^k}}$ with constant $c > 0$ small enough, we arrive at

$$\inf_{\hat{P}} \sup_{P\in\mathcal{M}_S} \mathbb{E}_P\|\hat{P} - P\|_1 \gtrsim \frac{S}{\sqrt{n2^k}}$$

which is the desired lower bound. Moreover, since it is assumed that $n \gtrsim \frac{S^2}{2^k}$ in this case, we indeed have $\delta \in (0, \frac{1}{S})$ by choosing c small enough, i.e., (4) holds, as desired.

3) Public/Private randomness: We show that the previous lower bound remains valid if some randomness is available. Let R denote any (private or public) randomness, then

$$\begin{aligned}
I(U; Y) &\le \mathbb{E}_U D(P_{Y|U}\|P_{Y|X\sim P_0}) \\
&\le \mathbb{E}_U\mathbb{E}_R D(P_{Y|R,U}\|P_{Y|R,X\sim P_0}) \\
&= \mathbb{E}_R\mathbb{E}_U D(P_{Y|R,U}\|P_{Y|R,X\sim P_0})
\end{aligned}$$

where the second step follows from the joint convexity of the KL divergence. Note that everything becomes deterministic conditioning on R, therefore we can upper bound $\mathbb{E}_U D(P_{Y|R,U}\|P_{Y|R,X\sim P_0})$ as before and arrive at the same minimax lower bound.

C. Proof of Theorem 1

Similar to the achievability proof, we prove the lower bound of nonparametric density estimation via a standard parametric reduction [2]. Let g be some non-negative smooth function on $\mathbb{R}$ vanishing outside $[0, 1]$ with $\|g\|_1 = 1$. Fix some bandwidth $h > 0$ with integer $S \triangleq h^{-1}$, and choose

$$f_P(x) = 1 + \sum_{j=1}^{S} \frac{p_j - h}{h} \cdot g\left(\frac{x - x_j}{h}\right), \quad x \in [0, 1]$$

where $P = (p_1, \cdots, p_S) \in \mathcal{M}_S, x_j \triangleq (j - 1)h$. As long as $|p_j - h| \le Ch^{s+1}$ for any $j \in [S]$, then f_P is a valid density on $[0, 1]$ for any $P \in \mathcal{M}_S$, and f_P has Hölder smoothness s.

Assume by contradiction that there is some estimator $\hat{f}$ which achieves $\sup_{f\in\mathcal{H}^s(L)} \mathbb{E}_f\|\hat{f} - f\|_1 \ll (n \cdot 2^k)^{-\frac{s}{2(s+1)}} + n^{-\frac{s}{2s+1}}$, then for $\hat{p}_j = \int_{(j-1)h}^{jh} \hat{f}(x)dx$, the estimator $\hat{P} = (\hat{p}_1, \cdots, \hat{p}_S)$ satisfies

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\|\hat{P} - P\|_1 \ll (n \cdot 2^k)^{-\frac{s}{2(s+1)}} + n^{-\frac{s}{2s+1}} \quad (7)$$

where $\mathcal{P} \triangleq \{P \in \mathcal{M}_S : |p_j - h| \le Ch^{s+1}, j \in [S]\}$.

On the other hand, restricting to the parametric submodel $f \in \{f_P : P \in \mathcal{P}\}$, the statistics $Z_1, \cdots, Z_n \in [S]$ with $X_i \in [(Z_i - 1)h, Z_i h)$ become sufficient, and

$$Z_1, \cdots, Z_n \overset{i.i.d.}{\sim} P. \quad (8)$$

Hence, model reduction by sufficiency reduces the parametric submodel $f \in \{f_P : P \in \mathcal{P}\}$ to the discrete Multinomial model (8), and by the proof of Theorem 2, as long as $[n(2^k \wedge S)]^{-\frac{1}{2}} \lesssim h^{s+1}$, we have

$$\inf_{\hat{P}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\|\hat{P} - P\|_1 \gtrsim \frac{S}{\sqrt{n(2^k \wedge S)}}. \quad (9)$$

Finally, choosing $h = (n \cdot 2^k)^{-\frac{1}{2(s+1)}} \vee n^{-\frac{1}{2s+1}}$ yields to the desired contradiction between inequalities (7) and (9).

IV. Discussions

The common approach for proving lower bounds in communication constrained settings in the literature is to develop so called strong data processing inequalities. Our approach on the other hand directly incorporates the k-bit communication constraint in the representation of the blackboard protocol. In

this section, we compare our proof technique with strong data processing inequalities and show that our technique can indeed be generalized to other statistical models.

A. Comparison with Strong Data Processing

By Fano's inequality, the key step in proving the converse is to upper bound the mutual information $I(U;Y)$ under the Markov chain $U - X - Y$, where the link $U - X$ is dictated by the statistical model, and the link $X - Y$ is subject to a k-bit communication constraint, which in turn implies that $I(X;Y) \leq k$. While trivially $I(U;Y) \leq I(U;X)$ and $I(U;Y) \leq I(X;Y)$, neither of these two inequalities are typically sufficient to obtain a good lower bound, and most works [8], [10], [11] rely on strong data processing inequalities of the form

$$I(U;Y) \leq \gamma^*(U,X) I(X;Y), \qquad \forall p_{Y|X} \tag{10}$$

with $\gamma^*(U,X) < 1$, which turns out to be tight in certain models (e.g., the Gaussian model [8], [10]). However, this approach is also subject to some drawbacks:

1) The tight constant $\gamma^*(U,X)$ is hard to obtain in general;
2) The conditional distribution $p_{Y^*|X}$ achieving the equality in (10) typically leads to $I(X;Y^*) \to 0$, and (10) may be loose for $I(X;Y) = k$;
3) The linearity of (10) in $I(X;Y)$ can only give a linear dependence of $I(U;Y)$ on k, which may not be tight. For example, in our case the optimal dependence on k is exponential;
4) The operational meaning of (10) is not clear, which may not result in a valid encoding scheme from X to Y. In contrast, the functions $\{a_v(\cdot)\}$ in our approach have clear operational meanings, and lead directly to an encoding scheme.

B. Generalization to other problems

Since our representations for the blackboard model and sensors' strategy are valid for general models, we can generalize our result to other distributed estimation problems with communication constraints. Consider the model $X_1, \cdots, X_n \overset{i.i.d.}{\sim} P_{\theta_1} \times \cdots \times P_{\theta_d}$ in the same setting and we would like to estimate the parameter vector $\theta = (\theta_1, \cdots, \theta_d)$ (say, under the ℓ_1 loss). We can show that, under mild regularity conditions, the optimal centralized performance in a local minimax sense [18] is $\Theta(\frac{d}{\sqrt{nI(\theta_0)}})$, while it is $\Omega(\frac{d^{3/2}}{\sqrt{n2^k I(\theta_0)}})$ in the distributed setting, where $I(\theta_0)$ denotes the Fisher information of the model (P_θ) at some target θ_0. Theorem 2 roughly corresponds to $P_\theta = \mathsf{Bern}(\theta)$, $\theta_0 = \frac{1}{S}$ and $d = S$, so it is implied by the general result with $I(\theta_0) \asymp S$. Moreover, the general result shows that the dependence on k is at most exponential, which is tight for distribution estimation.

One may wonder whether our technique would always yield an exponential dependence on k, which may not be tight in some models. The answer is actually *no*: in addition to solely applying the second equality in Lemma 2 to prove Theorem 2, the combination of both equalities of Lemma 2 may be useful in other models. We show via a geometric isoperimetric inequality that in Gaussian location models, new constraints give the final linear dependence on k, recovering the result of [8]. We refer to [14] for details.

V. Acknowledgement

We thank the anonymous reviewers of this paper and that of [14] for raising the question of whether our initial result can be generalized to the current blackboard communication protocol. This work was supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

References

[1] E. Kushilevitz and N. Nisan, "Communication complexity. cambridge university press," 1997.
[2] A. Nemirovski, "Topics in non-parametric statistics," *Ecole d'Eté de Probabilités de Saint-Flour*, vol. 28, p. 85, 2000.
[3] I. Diakonikolas, "Beyond histograms: Structure and distribution estimation," in *Workshop of the 46th ACM Symposium on Theory of Computing*, 2014.
[4] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6343–6354, 2015.
[5] S. Kamath, A. Orlitsky, V. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Proceedings of The 28th Conference on Learning Theory*, 2015, pp. 1066–1100.
[6] M. F. Balcan, A. Blum, S. Fine, and Y. Mansour, "Distributed learning, communication complexity and privacy," in *Conference on Learning Theory*, 2012, pp. 26–1.
[7] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS),*. IEEE, 2013, pp. 429–438.
[8] 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*, 2013, pp. 2328–2336.
[9] A. Garg, T. Ma, and H. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Advances in Neural Information Processing Systems*, 2014, pp. 2726–2734.
[10] 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*. ACM, 2016, pp. 1011–1020.
[11] A. Xu and M. Raginsky, "Information-theoretic lower bounds on Bayes risk in decentralized estimation," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1580–1600, 2017.
[12] I. Diakonikolas, E. Grigorescu, J. Li, A. Natarajan, K. Onak, and L. Schmidt, "Communication-efficient distributed learning of discrete distributions," in *Advances in Neural Information Processing Systems*, 2017, pp. 6394–6404.
[13] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar, "An information statistics approach to data stream and communication complexity," *Journal of Computer and System Sciences*, vol. 68, no. 4, pp. 702–732, 2004.
[14] Y. Han, A. Özgür, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," *to appear in Conference on Learning Theory (COLT)*, 2018.
[15] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4750–4755, 2008.
[16] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.
[17] J. C. Duchi and M. J. Wainwright, "Distance-based and continuum Fano inequalities with applications to statistical estimation," *arXiv preprint arXiv:1311.2669*, 2013.
[18] J. Hájek, "Local asymptotic minimax and admissibility in estimation," in *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1972, pp. 175–194.