Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints

Yanjun Han, Ayfer Özgür, Tsachy Weissman*

July 23, 2021

Abstract

We consider parameter estimation in distributed networks, where each sensor in the network observes an independent sample from an underlying distribution and has k bits to communicate its sample to a centralized processor which computes an estimate of a desired parameter. We develop lower bounds for the minimax risk of estimating the underlying parameter for a large class of losses and distributions. Our results show that under mild regularity conditions, the communication constraint reduces the effective sample size by a factor of d when k is small, where d is the dimension of the estimated parameter. Furthermore, this penalty reduces at most exponentially with increasing k, which is the case for some models, e.g., estimating high-dimensional distributions. For other models however, we show that the sample size reduction is re-mediated only linearly with increasing k, e.g. when some sub-Gaussian structure is available. We apply our results to the distributed setting with product Bernoulli model, multinomial model, Gaussian location models, and logistic regression which recover or strengthen existing results.

Our approach significantly deviates from existing approaches for developing information-theoretic lower bounds for communication-efficient estimation. We circumvent the need for strong data processing inequalities used in prior work and develop a geometric approach which builds on a new representation of the communication constraint. This approach allows us to strengthen and generalize existing results with simpler and more transparent proofs.

Contents

*Y. Han, A. Özgür, and T. Weissman are with the Department of Electrical Engineering, Stanford University, email: {yjhan,aozgur,tsachy}@stanford.edu

1 Introduction

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 [BPC+11, BBFM12, DIPSV12, DPSV12, DGBSX12]. A notable example is the federated learning [MMR+17], where multiple entities collaborate in solving a machine learning problem under the coordination of a central server, and communication could be a primary bottleneck since wireless links and other end-user internet connections typically operate at low rates and can be potentially expensive and unreliable; see [KMA+19] for an overview.

In this paper, we consider general distributed statistical estimation problems under communication constraints, and focus on the impact of a finite-communication budget per sample on the final estimation accuracy. More formally, consider the following parameter estimation problem

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

where we would like to estimate $\theta \in \Theta \subseteq \mathbb{R}^d$ under some general loss function L such as the ℓ_1 or ℓ_2^2 loss. Unlike the traditional setting where $X_1, \cdots, X_n$ are directly available to the estimator, we consider a distributed setting where each observation X_i is available at a different sensor and has to be communicated to a central estimator by using a communication budget of k bits. We consider a general interactive communication model known as the blackboard communication protocol

Π_{BB} [KN97]: 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. The main motivation for considering a blackboard communication protocol is that it models arbitrary interaction between the nodes. Impossibility results proven under this assumption provide insights about whether communication protocols that make (better) use of interaction can potentially lead to better performance than those achieved by simple schemes, e.g. simultaneous protocols. We note that impossibility results are strongest when proven under this most flexible communication model. We also consider a weaker family of protocols called the simultaneous message passing protocols (denoted by Π_{SMP}), where each sensor independently sends k bits to the centralized processor.

Under both models, the central sensor needs to produce an estimate $\widehat{\theta}$ of the underlying parameter θ from the the k-bit observations Y^n it collects at the end of the communication. Our goal is to jointly design a communication protocol in Π (which is either Π_{BB} or Π_{SMP}) and the estimator $\widehat{\theta}(\cdot)$ so as to minimize the worst case risk, i.e. to characterize the following *distributed minimax risk*

$$R^\star(n, k, \Theta, \Pi) \triangleq \inf_{\Pi} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \widehat{\theta})].$$

In this paper, lower bounds of the distributed minimax risk will typically be shown under the stronger blackboard communication protocol, while upper bounds of the same order will be attainable under the weaker simultaneous message passing protocol.

The main contributions of our paper are as follows:

1. For a large class of statistical models, we develop a novel geometric approach that builds on a new representation of the communication constraint to establish information-theoretic lower bounds for distributed parameter estimation problems. Our approach circumvents the need for strong data processing inequalities, and relate the experimental design problem directly to an explicit optimization problem in high-dimensional geometry.

2. Based on our new approach, we show that the communication constraint reduces the effective sample size from n to n/d for $k = 1$ under mild regularity conditions, where d is the dimension of the parameter to be estimated. Moreover, for general communication budget k, our new approach enables us to show that the penalty is at most exponential in k.

3. Our new approach also reveals that the tight dependence of the distributed minimax risk on k is determined by different geometric inequalities in different statistical models. Our result recovers the linear dependence on k when some sub-Gaussian structure is available, e.g., in the Gaussian location model. However, in models with heavier tails such as the high-dimensional distribution estimation model, we show that the exponential dependence on k becomes tight.

1.1 Related Work

Distributed parameter estimation and learning, under communication or privacy constraints, have been considered in many recent works. Early work [ZDJW13, Sha14, GMN14, BGM+16, XR17] established strong data-processing inequalities to prove tight lower bounds of distributed minimax risk under communication constraints. In particular, they showed that for communication-constrained Gaussian mean estimation, the distributed minimax risk depends linearly on k under the blackboard communication protocol. Similar approaches were also used to obtain minimax risks under privacy constraints [DJW13, KBR16, DJW18].

Due to the nature of strong data-processing inequalities, the above procedure typically leads to lower bounds linear in k (or the squared privacy parameter). However, for certain statistical estimation problems such as the high-dimensional distribution estimation, this dependence may not be tight. For example, a tight exponential dependence on the privacy parameter was established in [YB18] for discrete distribution estimation, and [DGL+17] established the tight total communication budget for this problem under the communication constraint[1]. A complete characterization of the distributed minimax risk for discrete distributione estimation and general (n, k) was obtained in [HMÖW18] under the simultaneous message passing protocol, with a tight exponential dependence on k. Generalization of [HMÖW18] to general statistical models with varying dependence on k was studied in an earlier version of this work [HÖW18], but both papers require the usage of the simultaneous message passing procotol (and there was a technical mistake in handling the blackboard communication protocol). All the above work modeled the communication/privacy constraint directly and did not use the strong data-processing inequality; along this direction, a flourishing line of recent research has studied different distributed estimation [ASZ19, AS19, ACT20b, ACT20c, CKO20] and testing [ACT20b, ACT20c, ACF+21, ACH+20] problems for the discrete distribution model under the simultaneous message passing protocol. A similar Gaussian identity testing problem was also studied in [ACT20a].

However, although it is relatively easy to extend the strong data-processing inequality based approach to blackboard communication protocols, it is more difficult to extend the approach based on direct modeling to interactive protocols Π_{BB}. We review some recent work which dealt with interactive communication protocols. Duchi and Rogers [DR19] established lower bounds for general interactive communication protocols based on machinery in the communication complexity literature, given a total privacy constraint. Barnes et al. [BHÖ20] studied a quantized Fisher information under blackboard communication protocols and proved a Bayesian lower bound using a continuous prior and the van Trees inequality, which typically requires the usage of the ℓ_2^2 loss. This Fisher information based approach has been extended to distributed estimation under local differential privacy constraints in [BCÖ20] recovering the results of [DR19] in the case of the ℓ_2^2 loss. For the identity testing, there are three recent papers Amin et al. [AJM20], Berrett and Butucea [BB20], and Acharya et al. [ACL+20] which established the sharp statistical rates under interactive communication protocols, where they focused on the discrete distribution estimation model and sequential interactive protocols. Sequential communication protocols are stronger than Π_{SMP} but weaker than Π_{BB}, as samples are encoded in a sequential fashion and the encoding of the sample i can decode on the messages transmitted by sensors $1, \ldots, i-1$. In this paper, we extend the results of [HÖW18] to blackboard communication protocols Π_{BB} (fixing the mistake in [HMÖW18] and [HÖW18]) via a generalization of the idea in [ACL+20], and therefore extending the approach of [ACL+20] to a broader family of statistical models and the blackboard communication protocol.

We also compare with a recent paper [ACT21] which appears after our submission. Both papers build upon the framework presented for the discrete setting in [ACL+20], and have similar assumptions and results on the high-dimensional estimation problem with communication constraints. We also point out some differences. In terms of scope, [ACT21] also studied the privacy constraints, and its latest version analyzed sparse estimation models in more detail. In terms of assumptions, an exact orthogonality assumption of likelihood ratios is required in [ACT21], whereas our likelihood ratio condition could be viewed as an approximate version which enables us to study the logistic regression model as well. Finally, in terms of the communication protocol, our blackboard communication protocol is more general than the sequential communication protocol studied in [ACT21], requiring additional effort in handling the tree-based communication protocol.

[1] However, no full version of [DGL+17] with complete proofs is available online at the time of writing.

1.2 Notation

For a finite set A, let $|A|$ denote its cardinality; $[n] \triangleq \{1, 2, \cdots, n\}$; for a measure μ, let $\mu^{\otimes n}$ denote its n-fold product measure; lattice operations $\wedge, \vee$ are defined as $a \wedge b = \min\{a, b\}, a \vee b = \max\{a, b\}$; throughout the paper, logarithms $\log(\cdot)$ are in the natural base; $\|P - Q\|_{\mathsf{TV}}$ and $D_{\mathsf{KL}}(P\|Q)$ denote the total variation (TV) distance and Kullback–Leibler (KL) divergence between probability measures P and Q, respectively; $\mathsf{Multi}(n; P)$ denotes the multinomial model which observes n independent samples from P; for a matrix $A \in \mathbb{R}^{m\times n}$, $\|A\|_{\mathsf{op}} = \max_{x\in\mathbb{R}^n:\|x\|_2=1} \|Ax\|_2$ denotes the operator norm; 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 $b_n \lesssim a_n$.

1.3 Organization

The rest of the paper is organized as follows. Section 2 presents our assumptions on the statistical model and two main lower bounds on the distributed minimax risk, which lead to new results or recover existing results in distributed estimation. In Section 3 we introduce the tree representation of the blackboard communication protocol, and sketch the lower bound proof based on this representation. Section 4 is devoted to the proof of Theorems 1 and 2, where the key steps are two geometric inequalities. Further discussions are in Section 5, and auxiliary lemmas and the proof of main lemmas are in the appendices.

2 Main Results

2.1 Assumptions

To derive meaningful results in the general minimax formulation, proper assumptions are necessary for the statistical model $(P_\theta)_{\theta\in\Theta\subseteq\mathbb{R}^d}$ and the loss function L. To this end, we begin with the standard regularity condition on $(P_\theta)_{\theta\in\Theta}$.

Assumption 1. *The statistical model $(P_\theta)_{\theta\in\Theta}$ is differentiable in quadratic mean at every $\theta \in \Theta$, with the score function S_θ and the Fisher information matrix I_θ.*

Note that Assumption 1 is a mild condition commonly used in classical asymptotic statistics [IH13], which leads to the asymptotically tight Cramér–Rao lower bound for centralized estimation. However, to obtain finite-sample results we need additional assumptions requiring the following notations. For a binary vector $u \in \{\pm 1\}^m$ and $j \in [m]$, let $u^{\oplus j}$ be the vector after flipping the j-th coordinate of u. Also, for two binary vectors $u, u' \in \{\pm 1\}^m$, let $d_{\mathsf{Ham}}(u, u') = \sum_{j=1}^m \mathbb{1}(u_j \neq u'_j)$ be their Hamming distance. In addition, let $\mathcal{X} \subseteq \mathbb{R}^d$ be the common support of all probability measures $(P_\theta)_{\theta\in\Theta}$, and $\mathcal{A}$ be a generic action space in which the estimator $\widehat{\theta}$ takes value. Finally, by a loss function L we mean a generic non-negative (measurable) function $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$. The next assumption is a refinement of Assumption 1 which concerns the finite-sample property of $(P_\theta)_{\theta\in\Theta}$ and L.

Assumption 2. *There exist $1 \le d_0 \le d$ and a subset of parameters $(\theta_u)_{u\in\{\pm1\}^{d_0}} \subseteq \Theta$ such that the following conditions hold:*

1. ***Regular grid condition**: For each $u \in \{\pm 1\}^{d_0}$, the $d \times d_0$ matrix M_u with columns $\theta_{u^{\oplus j}} - \theta_u$ ranging over $j \in [d_0]$ has an operator norm at most 2δ.*

2. ***Separation condition***: *for any* $u, u' \in \{\pm 1\}^{d_0}$, *it holds that*

$$\inf_{a \in \mathcal{A}} [L(\theta_u, a) + L(\theta_{u'}, a)] \geq \kappa \cdot d_{\mathsf{Ham}}(u, u'). \tag{1}$$

3. ***Likelihood ratio condition***: *for any* $u \in \{\pm 1\}^{d_0}$ *and* $j \in [d_0]$, *it holds that*

$$\mathbb{E}_{X \sim P_{\theta_u}} \left[\left| \frac{dP_{\theta_{u^{\oplus j}}}}{dP_{\theta_u}}(X) - 1 - (\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(X) \right|^2 \right] \leq \varepsilon^2. \tag{2}$$

In addition, it holds that $dP_{\theta_{u^{\oplus j}}} / dP_{\theta_u}(x) \geq 1/2$ *for all* $x \in \mathcal{X}$.

If all above conditions hold, we call this statistical estimation problem $(d, d_0, \delta, \kappa, \varepsilon)$***-regular***.

Assumption 2 will be best understood via an important special case. Consider $d_0 = d$, and for each $u \in \{\pm 1\}^d$, let $\theta_u = \theta_0 + \delta u$ be a local perturbation of some $\theta_0 \in \Theta$. Then the regularity grid condition clearly holds as the matrix M_u is diagonal with diagonal entries being $\pm 2\delta$. Therefore, this condition essentially says that the parameters $(\theta_u)_{u \in \{\pm 1\}^{d_0}}$ look like the vertices of a cube with side length δ. The separation condition is standard in applying Assouad or Fano-type arguments to a cube-like hypothesis class [Yu97], and is fulfilled for many natural loss functions with $\kappa = \kappa(\delta)$ a function of δ. For example, $\kappa(\delta) = 2\delta^p$ when $L = \ell_p^p$, with important special cases including the ℓ_1 loss when $p = 1$, and the mean squared error when $p = 2$. The last likelihood ratio condition is motivated by the local expansion

$$\frac{dP_{\theta + t \cdot h}}{dP_\theta}(x) = \exp\left(t \cdot h^\top S_\theta(x) - \frac{t^2}{2} \cdot h^\top I_\theta h + o_{P_\theta}(t^2) \right), \quad \forall h \in \mathbb{R}^d$$

and $e^x = 1 + x + x^2/2 + o(x^2)$, as well as the identity $\mathbb{E}_{X \sim P_\theta}[S_\theta(X) S_\theta(X)^\top] = I_\theta$. In other words, (2) is a quantitative way to approximate the local likelihood ratio by score functions, with the approximation error $\varepsilon = \varepsilon(\delta)$ typically growing with δ. This quantitative condition will help us to show the indistinguishability among the locally perturbed statistical models. We also remark that the likelihood ratio is computed only between two neighboring vertices and often not growing with the dimensionality d, thus the lower bound assumption on the local likelihood ratio is not restrictive in high dimensions. Finally, to show that Assumption 2 holds for a certain statistical model, typically we first choose a suitable perturbation distance δ and then work out the parameters κ and ε as functions of δ.

While the first two conditions of Assumption 2 are relatively easier to hold, the last condition may fail for statistical models with an unbounded support (e.g. $\mathcal{X} = \mathbb{R}^d$). To mitigate this drawback, we propose a slightly weaker assumption which requires that the likelihood ratio condition holds with a high probability.

Assumption 3. *Assume the same conditions in Assumption 2, except that in the likelihood ratio condition, there exists some* $\mathcal{X}_0 \subseteq \mathcal{X}$ *such that*

$$\mathbb{E}_{X \sim P_{\theta_u}} \left[\left| \frac{dP_{\theta_{u^{\oplus j}}}}{dP_{\theta_u}}(X) - 1 - (\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(X) \right|^2 \cdot \mathbb{1}(X \in \mathcal{X}_0) \right] \leq \varepsilon^2, \tag{3}$$

and $dP_{\theta_{u^{\oplus j}}} / dP_{\theta_u}(x) \geq 1/2$ *for all* $x \in \mathcal{X}_0$. *Moreover, we require that* $P_{\theta_u}(\mathcal{X}_0) \geq 1 - \alpha$ *for all* $u \in \{\pm 1\}^{d_0}$. *If the above condition holds, we call this statistical estimation problem* $(d, d_0, \delta, \kappa, \varepsilon, \alpha)$***-approximately-regular***.

In all the models considered in this paper, we always have $\alpha = o(n^{-1})$, so the likelihood ratio condition is only violated with a tiny probability, which suffices for our main Theorems 1 and 2. The next proposition shows that many common statistical models are regular or approximately regular.

Proposition 1. *For $L = \ell_p^p$ with $p \in [1, \infty)$, the following statements hold:*

- *The product Bernoulli model $P_\theta = \prod_{j=1}^d \mathsf{Bern}(\theta_j)$ with $\Theta = [0,1]^d$ is $(d, d, \delta, \kappa, \varepsilon)$-regular with any $\delta \in (0, 1/6)$, and $\kappa(\delta) = 2\delta^p, \varepsilon(\delta) \equiv 0$.*
- *The product Bernoulli model $P_\theta = \prod_{j=1}^d \mathsf{Bern}(\theta_j)$ or the Multinomial model $P_\theta = \mathsf{Multi}(1; \theta)$ with $\Theta = \{(\theta_1, \cdots, \theta_d) : \sum_{j=1}^d \theta_j = 1\}$ is $(d, d/2, \delta, \kappa, \varepsilon)$-regular with any $\delta \in (0, 1/(2d))$, and $\kappa(\delta) = 2^{2-p}\delta^p$, $\varepsilon(\delta) \equiv 0$.*
- *The Gaussian location model $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ with $\Theta = \mathbb{R}^d$ is $(d, d, \delta, \kappa, \varepsilon, \alpha)$-approximately-regular with any $\delta \in (0, c\sigma/\sqrt{\log(nd)})$ for some small constant $c > 0$, and $\kappa(\delta) = 2\delta^p$, $\varepsilon(\delta) = O(\delta^2/\sigma^2)$, $\alpha = o(n^{-1})$.*
- *Consider the following logistic regression model P_θ with random design: the observation vector is $X = (z, y)$, with feature $z \sim \mathcal{N}(0, I_d)$ and label $y \sim \mathsf{Bern}(1/(1 + \exp(-\theta^\top z)))$ given z. This model with $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}$ is $(d, d, \delta, \kappa, \varepsilon, \alpha)$-approximately-regular with any $\delta \in (0, 1/\sqrt{d})$, and $\kappa(\delta) = 2\delta^p$, $\varepsilon(\delta) = O(\delta^2)$, $\alpha = o(n^{-1})$.*

2.2 Main Theorems

Although the previous assumptions give that the local likelihood ratio could be approximated by a linear form of the score function, they do not impose any assumption on the score function itself. As we recall from the classical asymptotic theory that the score function and the Fisher information play central roles in the estimation error, additional properties on the score function will be required to state the minimax lower bound. Our first and general lower bound states that, if the score function has a finite variance along any direction, then the estimation error in the distributed case decays at most exponentially with k.

Theorem 1 (General lower bound I)**.** *Let the statistical problem be $(d, d_0, \delta, \kappa, \varepsilon, \alpha)$-approximately-regular with*

$$I_0 \triangleq \max_{u \in \{\pm 1\}^{d_0}} \max_{v \in \mathbb{R}^d : \|v\|_2 = 1} \mathbb{E}_{\theta_u}[(v^\top S_{\theta_u}(X))^2] < \infty.$$

Then it holds that

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \widehat{\theta})] \ge c\kappa d_0 \left[\exp\left(-Cn\left(\frac{2^k \wedge d}{d_0} \cdot I_0 \delta^2 + \varepsilon^2\right)\right) - 4n\alpha\right],$$

where the infimum is taken over all possible estimators $\widehat{\theta} = \widehat{\theta}(Y^n)$ and blackboard protocols with k-bit communication constraint, and $c, C > 0$ are absolute constants independent of $(n, d, k, I_0, d_0, \delta, \kappa, \varepsilon, \alpha)$.

We first show how Theorem 1 could be used to give a meaningful lower bound. Since $\kappa = \kappa(\delta)$ and $\varepsilon = \varepsilon(\delta)$ are typically increasing in δ, as δ increases, the leading coefficient will be larger while the exponential term will be smaller. To handle this tradeoff, we will choose the largest $\delta > 0$ such that the statistical problem remains to be (approximately-)regular, and $\delta^2 = O(d_0/(nI_0(2^k \wedge d)))$. Meanwhile, for this choice of δ, we expect that $\varepsilon = \varepsilon(\delta)$ is at most $O(n^{-1/2})$, which holds in many

examples in the next subsection. Finally, for this choice of δ, we conclude that the minimax risk is lower bounded by $\Omega(d_0\kappa(\delta))$.

We provide several intuitive implications of Theorem 1. Assume for simplicity that $L = \ell_2^2$, $d_0 = d$, $\kappa = \kappa(\delta) = 2\delta^2$ and $\varepsilon = \varepsilon(\delta) \equiv 0$, then the choice of $\delta^2 \asymp d/(nI_0(2^k \wedge d))$ in Theorem 1 leads to the lower bound $\Omega(d^2/(nI_0(2^k \wedge d)))$. In the centralized case without any communication constraints, we have $k = \infty$ and therefore the lower bound $\Omega(d/(nI_0))$ for the mean squared error. Since the Fisher information matrix I_θ satisfies $I_\theta = \mathbb{E}_\theta[S_\theta(X)S_\theta(X)^\top]$, an equivalent expression of I_0 is

$$I_0 = \max_{u\in\{\pm 1\}^{d_0}} \lambda_{\max}(I_{\theta_u}),$$

where $\lambda_{\max}$ denotes the largest eigenvalue. As a comparison, the standard Cramér–Rao lower bound for the mean squared error is $\Omega(\mathrm{trace}(I_\theta^{-1})/n)$ for any $\theta \in \Theta$ [Háj72]. Consequently, Theorem 1 reduces to a weaker but non-asymptotic version of the Cramér–Rao lower bound in the centralized case, which often remains rate-optimal when P_θ is of a product structure.

Now what happens when there are communication constraints? Using the above result, in the most communication-starved case $k = 1$, we have an effective sample size reduction from n to n/d. This bound is intuitively achievable by a simple grouping idea: the sensors are splitted into n/d groups, and all d sensors in one group "simulate" a full d-dimensional observation with each sensor working on one coordinate (see, e.g. Proposition 3). Therefore, we expect that the dependence on n, d of our lower bound to be tight for $k = 1$. When $k > 1$, the lower bound $\Omega(d^2/(nI_0(2^k \wedge d)))$ shows that the dependence of the squared ℓ_2 risk on k cannot be faster than 2^{-k}, i.e., the penalty incurred by the distributed setting reduces at most exponentially in k. In the next subsection we will see examples where this exponential reduction is indeed tight.

A natural question is that whether or not the exponential dependence on k is always tight. The answer turns out to be *negative*: the above penalty will reduce at most linearly in k when the score function has a sub-Gaussian tail along any direction. Recall that the ψ_2-norm of a random variable X is defined by

$$\|X\|_{\psi_2(P)} = \inf\left\{t > 0 : \mathbb{E}_P\left[\exp\left(\frac{X^2}{t^2}\right)\right] \le 2\right\},$$

which is the Orlicz norm of X associated with the Orlicz function $\psi_2(x) = \exp(x^2) - 1$ [BO31]. There are some equivalent definitions of the ψ_2-norm, and $\|X\|_{\psi_2(P)} \le \sigma$ if and only if X is sub-Gaussian under P with parameter $\Theta(\sigma)$ [Ver10]. The following theorem shows another lower bound when the score function has a finite ψ_2-norm along any direction.

Theorem 2 (Lower bound with sub-Gaussian structure)**.** *Let the statistical problem be $(d, d_0, \delta, \kappa, \varepsilon, \alpha)$-approximately-regular with*

$$\Sigma_0 \triangleq \max_{u\in\{\pm 1\}^{d_0}} \max_{v\in\mathbb{R}^d:\|v\|_2=1} \|v^\top S_{\theta_u}(X)\|^2_{\psi_2(P_{\theta_u})} < \infty.$$

Then it holds that

$$\inf_{\Pi_{\mathrm{BB}}} \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta[L(\theta, \widehat{\theta})] \ge c\kappa d_0 \left[\exp\left(-Cn\left(\frac{k\wedge d}{d_0}\cdot\Sigma_0\delta^2 + \varepsilon^2\right)\right) - 4n\alpha\right],$$

where the infimum is taken over all possible estimators $\widehat{\theta} = \widehat{\theta}(Y^n)$ and blackboard protocols with k-bit communication constraint, and $c, C > 0$ are absolute constants independent of $(n, d, k, \Sigma_0, d_0, \delta, \kappa, \varepsilon, \alpha)$.

Using the similar intuitive analysis, Theorem 2 roughly shows a lower bound $\Omega(d^2/(n\Sigma_0(k\wedge d)))$ for the mean squared error. When the coordinates of the score function $S_\theta(X)$ are independent (which holds when P_θ is a product distribution), the quantity Σ_0 is essentially the maximum ψ_2 norm for each coordinate. Compared with the lower bound $\Omega(d^2/(nI_0(2^k\wedge d)))$ in Theorem 1, the new lower bound has a better dependence on k when the score function not only admits a finite variance but also behaves like a Gaussian random variable. However, neither of these bounds is better than the other in general, for it is possible that $I_0 \ll \Sigma_0$; the next subsection will provide examples where each of these bounds is tight. We also remark that the different dependence on k in Theorems 1 and 2 is due to the nature of different geometric inequalities (cf. Lemma 3 and Lemma 4) satisfied by general probability distributions and sub-Gaussian distributions.

2.3 Applications

Next we apply Theorems 1 and 2 to some concrete statistical estimation examples. The first corollary concerns the discrete distribution estimation model.

Corollary 1 (Discrete distribution estimation)**.** *Let $P_\theta = \mathsf{Multi}(1;\theta)$ with Θ being the probability simplex over d elements. For $k\in\mathbb{N}$, $p\in[1,\infty)$, and $n\geq d^2/(2^k\wedge d)$, we have*

$$\inf_{\Pi_{\mathsf{BB}}}\inf_{\widehat{\theta}}\sup_{\theta\in\Theta}\mathbb{E}_\theta\|\widehat{\theta}-\theta\|_p^p \geq C_p\cdot\frac{d}{(n(2^k\wedge d))^{p/2}},$$

where $C_p>0$ is an absolute constant independent of n,k,d.

Using the construction of θ_u's in Proposition 1, after some algebra one could verify that $I_0 = O(d)$ in Theorem 1. Consequently, choosing $\delta = c/\sqrt{n(2^k\wedge d)}$ for a small enough constant $c>0$ ensures that $\delta<1/(2d)$ (which is required in Proposition 1), with $\kappa\asymp\delta^p$ and $\varepsilon=0$, giving the result of Corollary 1. For $p\in\{1,2\}$, there is a matching upper bound in [HMÖW18], showing the tightness of this minimax lower bound. This result also improves over the total communication budget in [DGL+17]. Under the sequential communication protocol, the recent paper [ACT21] established the same lower bound for $n\geq d^2/(2^k\wedge d)$, as well as a different lower bound $\Omega((n(2^k\wedge d))^{-(p-1)/2})$ for $n<d^2/(2^k\wedge d)$, the tightness of which is currently unclear (this lower bound also follows from Corollary 1 via replacing d by a smaller quantity $d_{\min}=\sqrt{n(2^k\wedge d)}$ such that $n\geq d_{\min}^2/(2^k\wedge d_{\min})$). Note that in this case, the tight dependence of the minimax risk on k is exponential.

The next corollary characterizes the distributed minimax risk of mean estimation in the Gaussian location model.

Corollary 2 (Gaussian location model)**.** *Let $P_\theta=\mathcal{N}(\theta,\sigma^2 I_d)$ with $\Theta=\mathbb{R}^d$. For $k\in\mathbb{N}$, $p\in[1,\infty)$ and $n\geq d^2/(k\wedge d)^2 + d\log d/(k\wedge d)$, we have*

$$\inf_{\Pi_{\mathsf{BB}}}\inf_{\widehat{\theta}}\sup_{\theta\in\Theta}\mathbb{E}_\theta\|\widehat{\theta}-\theta\|_p^p \geq C_p\cdot d\left(\frac{d\sigma^2}{n(d\wedge k)}\right)^{\frac{p}{2}},$$

where $C_p>0$ is an absolute constant independent of n,k,d,σ^2.

For the Gaussian location model, the score function is $S_\theta(x)=(x-\theta)/\sigma^2$, and therefore the assumption of Theorem 2 is fulfilled with $\Sigma_0=O(1/\sigma^2)$. Consequently, in Theorem 2 we may choose $\delta\asymp\sigma\sqrt{d/(n(d\wedge k))}$ (which satisfies the constraint of Proposition 1 by the choice of n) with $\kappa\asymp\delta^p$ and $\varepsilon\asymp\delta^2/\sigma^2 = O(n^{-1/2})$ to derive Corollary 2. Note that for $p=2$, Corollary 2 recover the results in [ZDJW13, GMN14], without logarithmic factors in the risk. Also, in this model, the tight dependence of the minimax risk on k is linear.

The above two models have different tight dependence on k: in Corollary 1, when $2^k < d$, we see an effective sample size reduction from n to $n2^k/d$; in Corollary 2, when $k < d$, we see an effective sample size reduction from n to nk/d. This phenomenon may be better illustrated using the following example:

Corollary 3 (Product Bernoulli model). *Let $P_\theta = \prod_{i=1}^d \mathsf{Bern}(\theta_i)$. If $\Theta = [0,1]^d$ and $n \geq \frac{d}{d \wedge k}$, we have*

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_2^2 \asymp \frac{d^2}{nk} \vee \frac{d}{n}.$$

If $\Theta \triangleq \{(\theta_1, \cdots, \theta_d) \subseteq [0,1]^d : \sum_{i=1}^d \theta_i = 1\}$ and $n \geq \frac{d^2}{d \wedge 2^k}$, we have

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_2^2 \asymp \frac{d}{n2^k} \vee \frac{1}{n}.$$

The first lower bound follows from Theorem 2, and it was also obtained in [ZDJW13] under the independent protocol with a matching upper bound. The same lower bound was also obtained in recent papers [ACL+20, ACT21]. The second lower bound follows from Theorem 1, and the upper bound could be obtained using the "simulate-and-infer" procedure in [ACT20c]. Note that the dependence of the squared ℓ_2 risk on k is significantly different under these two scenarios, even if both of them are product Bernoulli models: the dependence is linear in k when $\Theta = [0,1]^d$, while it is exponential in k when Θ is the probability simplex. We remark that this is due to the different behaviors of the score function: in the first case, we have $I_0 \asymp \Sigma_0 = \Theta(1)$; in the second case, we have $I_0 \asymp d \ll d^2 \asymp \Sigma_0$. Hence, Theorem 2 utilizes the sub-Gaussian nature and gives a better lower bound in the first case, and Theorem 1 becomes better in the second case where the tail of the score function is essentially not sub-Gaussian.

In addition to mean estimation, our main theorems also provide the following lower bound for parameter estimation in logistic regression.

Corollary 4 (Logistic regression). *Consider the logistic regression model with random design formulated in Proposition 1. For $k \in \mathbb{N}$, $p \in [1, \infty)$ and $n \geq d^2/(d \wedge k)$, we have*

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_p^p \geq C_p \cdot d \left(\frac{d}{n(d \wedge k)} \right)^{\frac{p}{2}},$$

where $C_p > 0$ is an absolute constant independent of n, k, d.

The proof of Corollary 4 follows from Proposition 1 and Theorem 2. Specifically, in logistic regression, the score function at $x = (z, y)$ is given by

$$S_\theta(x) = S_\theta(y, z) = \left(y - \frac{1}{e^{-\theta^\top z} + 1} \right) z,$$

which satisfies the sub-Gaussian condition of Theorem 2 with $\Sigma_0 = O(1)$ as the scalar parameter of z always lies in $[-1, 1]$. Consequently, choosing $\delta \asymp \sqrt{d/(n(d \wedge k))}$, as well as the quantities $\kappa(\delta) = 2\delta^p$, $\varepsilon(\delta) = O(\delta^2) = O(n^{-1/2})$, and $\alpha = o(n^{-1})$ given by Proposition 1, in Theorem 2 proves Corollary 4. Note that the above argument only requires the random feature vector z to be sub-Gaussian. For $p = 2$, the same result was also proved in [BO19] using the van Trees inequality, with a matching upper bound when $z \sim \mathsf{Unif}(\{\pm 1\}^d)$. A similar lower bound for logistic regression under privacy constraint was obtained in [DR19, Corollary 4]: although they studied the excess risk instead of the ℓ_2^2 loss, in the proof they essentially lower bounded the excess risk by the ℓ_2^2 loss.

For $p = 2$, their overhead compared with the centralized case is $d/(\varepsilon \wedge \varepsilon^2)$ with average privacy budget ε, while ours is $d/(d \wedge k)$. Also, while the lower bound under the privacy constraint could be attained using private gradient updates (see [BDF+18, Corollary 3.2]), it is unknown whether a similar approach works under the communication constraint.

Finally we look at the distributed mean estimation problem for sparse Gaussian location models.

Theorem 3 (Sparse Gaussian location model)**.** *Let $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ with $\Theta = \{\theta \in \mathbb{R}^d : \|\theta_0\| \le s\}$ with $s \le d/2$. For $k \in \mathbb{N}$ and $n \ge sd^2 \log(d/s)/(k \wedge d)^2$, we have*

$$\inf_{\Pi_{\mathsf{SMP}}} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_2^2 \ge C \cdot \left(\frac{sd \log(d/s)}{nk} \vee \frac{s \log(d/s)}{n} \right) \sigma^2$$

where $C > 0$ is an absolute constant independent of n, d, s, k, σ^2, and Π_{SMP} represents the family of simultaneous message passing protocols.

Under a different notion of communication cost, [BGM+16] proved a lower bound $\Omega(sd\sigma^2/(nk))$, *without* the logarithmic factor $\log(d/s)$, under blackboard communication protocols. Moreover, under the above notion of communication and Π_{SMP}, an upper bound $O(sd\sigma^2 \log d/(nk))$, *with* the logarithmic factor, was obtained in [GMN14]. Interestingly, the recent paper [ACT21] showed that an upper bound of $O(sd\sigma^2/(nk))$, *without* the logarithmic factor, is indeed achievable under the sequential communication protocol. Therefore, Theorem 3 shows that the logarithmic factor is unavoidable under the non-interactive communication protocols, so there is a strict separation between the interactive and non-interactive protocols. The existence/non-existence of the logarithmic factor in constrained sparse estimation is an interesting research topic, and has drawn several recent attentions such as [AKLS20, CKÖ21].

Ignoring the issues on the logarithmic factor, we see that as opposed to the logarithmic dependence on the ambient dimension d in the centralized setting, the number of nodes required to achieve a vanishing error in the distributed setting must scale with d. Hence, the sparse mean estimation problem becomes much harder in the distributed case, and the dimension involved in the effective sample size reduction (from n to nk/d) is the ambient dimension d instead of the effective dimension s.

3 Representations of Blackboard Communication Protocol

The centralized lower bounds without communication constraints simply follows from the classical asymptotics [Háj70, Háj72], thus we devote our analysis to the communication constraints. In this section, we establish an equivalent tree representation of the blackboard communication protocol, and prove the statistical lower bound based on this representation.

3.1 Tree representation of blackboard communication protocol

Assume first that there is no public/private randomness, which will be revisited in the next subsection, and thus the protocol is deterministic. In this case, the blackboard communication protocol Π_{BB} can be viewed as a binary tree [KN97], 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. 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; sensor 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 [BYJKS04] for the blackboard communication protocol:

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

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

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:

Lemma 2. *For each $i \in [n]$ and $\{x_j\}_{j=1}^n \in \mathcal{X}^n$, the following equalities hold:*

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

3.2 Minimax lower bound

This subsection is devoted to setting up the proof of the minimax lower bound in Theorems 1 and 2. To this end, we apply the standard testing argument with the Assouad's lemma [Ass83] to the cube-like distribution family $(P_{\theta_u})_{u\in\{\pm1\}^{d_0}}$ in Assumptions 2 and 3, and arrive at a target quantity to be upper bounded in Section 4. This subsection is devoted exclusively to $(d, d_0, \delta, \kappa, \varepsilon)$-regular problems to reflect the main ideas, while the modification to handle approximately regular problems is postponed to the next subsection.

Let $U \sim \mathsf{Unif}(\{\pm1\}^{d_0})$, and write P_u as a shorthand of P_{θ_u} throughout this section. Given the i.i.d. observations $X_1, \cdots, X_n \sim P_u$ and a communication protocol Π, let Q_u be the probability distribution of the final transcript $Y \in \{0,1\}^{nk}$. As the final estimator $\widehat{\theta} = \widehat{\theta}(Y)$ is a function of Y, the standard separation condition (1) with the Assouad's lemma [Ass83] (see also [Han19, Theorem 5]) gives that

$$\mathbb{E}_U \mathbb{E}_{Q_U}[L(\theta_U, \widehat{\theta}(Y))] \geq \frac{d_0\kappa}{2}\left(1 - \frac{1}{d_0}\sum_{j=1}^{d_0} \mathbb{E}_U \|Q_U - Q_{U^{\oplus j}}\|_{\mathsf{TV}}\right). \tag{4}$$

As [Tsy08, Lemma 2.6] shows that $\|P-Q\|_{\mathsf{TV}} \le 1-\exp(-D_{\mathsf{KL}}(P\|Q))/2$, the above inequality (4) together with the convexity of $x \mapsto \exp(-x)$ implies that

$$\begin{aligned}\mathbb{E}_U \mathbb{E}_{Q_U}[L(\theta_U, \widehat{\theta}(Y))] &\ge \frac{d_0\kappa}{2} \cdot \frac{1}{2d_0} \sum_{j=1}^{d_0} \mathbb{E}_U \exp\left(-D_{\mathsf{KL}}(Q_U \| Q_{U^{\oplus j}})\right) \\ &\ge \frac{d_0\kappa}{4} \cdot \exp\left(-\mathbb{E}_U \left[\frac{1}{d_0} \sum_{j=1}^{d_0} D_{\mathsf{KL}}(Q_U \| Q_{U^{\oplus j}}) \right] \right). \end{aligned} \tag{5}$$

The usage of (a slightly different form of) the inequality (5) is motivated by [ACL+20], which studies discrete distribution estimation models under the sequential communication protocol. In the sequel, we extend this approach to generic statistical models and fully interactive (blackboard) communication protocols.

Next we upper bound the average KL divergence in (5) for each given $U \in \{\pm 1\}^{d_0}$. To this end, first we note that it suffices to assume no private/public randomness due to the data-processing property of the KL divergence $D_{\mathsf{KL}}(P\|Q) \le \mathbb{E}_R[D_{\mathsf{KL}}(P_{|R}\|Q_{|R})]$. Then by Lemma 1, we have

$$Q_U(y) = \prod_{i=1}^n \mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)]$$

for each transcript $y \in \{0,1\}^{nk}$. Consequently, for each $j \in [d_0]$,

$$\begin{aligned} D_{\mathsf{KL}}(Q_U \| Q_{U^{\oplus j}}) &= \sum_{i=1}^n \sum_{y \in \{0,1\}^{nk}} \left(\prod_{s=1}^n \mathbb{E}_{X_s \sim P_U}[p_{s,y}(X_s)] \right) \cdot \log \frac{\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)]}{\mathbb{E}_{X_i \sim P_{U^{\oplus j}}}[p_{i,y}(X_i)]} \\ &\overset{\text{(a)}}{\le} \sum_{i=1}^n \sum_{y \in \{0,1\}^{nk}} \left(\prod_{s=1}^n \mathbb{E}_{X_s \sim P_U}[p_{s,y}(X_s)] \right) \cdot \left(\frac{\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)]}{\mathbb{E}_{X_i \sim P_{U^{\oplus j}}}[p_{i,y}(X_i)]} - 1 \right) \\ &\overset{\text{(b)}}{=} \sum_{i=1}^n \sum_{y \in \{0,1\}^{nk}} \left(\prod_{s \ne i} \mathbb{E}_{X_s \sim P_U}[p_{s,y}(X_s)] \right) \cdot \left(\frac{(\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)] - \mathbb{E}_{X_i \sim P_{U^{\oplus j}}}[p_{i,y}(X_i)])^2}{\mathbb{E}_{X_i \sim P_{U^{\oplus j}}}[p_{i,y}(X_i)]} \right) \\ &\overset{\text{(c)}}{\le} 2 \sum_{i=1}^n \sum_{y \in \{0,1\}^{nk}} \left(\prod_{s \ne i} \mathbb{E}_{X_s \sim P_U}[p_{s,y}(X_s)] \right) \cdot \left(\frac{(\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)] - \mathbb{E}_{X_i \sim P_{U^{\oplus j}}}[p_{i,y}(X_i)])^2}{\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)]} \right) \\ &\overset{\text{(d)}}{=} 2 \sum_{i=1}^n \sum_{y \in \{0,1\}^{nk}} \left(\prod_{s \ne i} \mathbb{E}_{X_s \sim P_U}[p_{s,y}(X_s)] \right) \cdot \frac{(\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)(1 - dP_{U^{\oplus j}}/dP_U(X_i))])^2}{\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)]} \end{aligned}$$

where (a) is due to the inequality $\log x \le x - 1$, (b) follows from the identity

$$\sum_{y \in \{0,1\}^{nk}} \prod_{s=1}^n \mathbb{E}_{X_s \sim P_U}[p_{s,y}(X_s)] = \sum_{y \in \{0,1\}^{nk}} \mathbb{E}_{X_i \sim P_{U^{\oplus j}}}[p_{i,y}(X_i)] \cdot \prod_{s \ne i} \mathbb{E}_{X_s \sim P_U}[p_{s,y}(X_s)] = 1$$

given by Lemma 2, (c) follows from the likelihood ratio condition in Assumption 2, and (d) is due to a simple change of measure. Consequently, for each realization of U we have

$$\frac{1}{d_0} \sum_{j=1}^{d_0} D_{\mathrm{KL}}(Q_U \| Q_{U^{\oplus j}}) \le \frac{2}{d_0} \sum_{i=1}^n \sum_{y \in \{0,1\}^{nk}} \left(\prod_{s \ne i} \mathbb{E}_{X_s \sim P_U}[p_{s,y}(X_s)] \right) \cdot \frac{\|\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i) s_U(X_i)]\|_2^2}{\mathbb{E}_{X_i \sim P_U}[p_{i,y}(X_i)]}, \tag{6}$$

where $s_U(x)$ is a d_0-dimensional vector of likelihood ratios:

$$s_U(x) \triangleq \left(1 - \frac{dP_{U\oplus 1}}{dP_U}(x), \cdots, 1 - \frac{dP_{U\oplus d_0}}{dP_U}(x)\right).$$

To deal with $s_U(x)$, we use the likelihood ratio condition in Assumption 2 to write that

$$s_U(x) = -M_u^\top S_{\theta_U}(x) + \varepsilon(x),$$

where M_u is the matrix appearing in the regular grid condition in Assumption 2, and $\varepsilon(x)$ is some remainder term satisfying that $\mathbb{E}[\|\varepsilon(X)\|_2^2] \le d_0\varepsilon^2$ for all $X \sim P_U$. Consequently, for the remainder term we have

$$\begin{aligned}
&\sum_{y\in\{0,1\}^{nk}} \left(\prod_{s\neq i} \mathbb{E}_{X_s\sim P_U}[p_{s,y}(X_s)]\right) \cdot \frac{\|\mathbb{E}_{X_i\sim P_U}[p_{i,y}(X_i)\varepsilon(X_i)]\|_2^2}{\mathbb{E}_{X_i\sim P_U}[p_{i,y}(X_i)]} \\
&\overset{\text{(a)}}{\le} \sum_{y\in\{0,1\}^{nk}} \left(\prod_{s\neq i} \mathbb{E}_{X_s\sim P_U}[p_{s,y}(X_s)]\right) \cdot \mathbb{E}_{X_i\sim P_U}[p_{i,y}(X_i)\|\varepsilon(X_i)\|_2^2] \\
&\overset{\text{(b)}}{=} \mathbb{E}_{X_1,\cdots,X_n\sim P_U}\left[\sum_{y\in\{0,1\}^{nk}} \left(\prod_{s=1}^n p_{s,y}(X_s)\right) \cdot \|\varepsilon(X_i)\|_2^2\right] \\
&\overset{\text{(c)}}{=} \mathbb{E}_{X_i\sim P_U}[\|\varepsilon(X_i)\|_2^2] \le d_0\varepsilon^2,
\end{aligned}$$

where (a) is due to Cauchy–Schwarz, (b) swaps the expectation and sum, and (c) is due to the first identity of Lemma 2. As for the main term, since $\|Ax\|_2 \le \|A\|_{\mathsf{op}}\|x\|_2$, the regular grid assumption in Assumption 2 gives

$$\frac{\|\mathbb{E}_{X_i\sim P_U}[p_{i,y}(X_i)\cdot M_u^\top S_{\theta_U}(X_i)]\|_2^2}{\mathbb{E}_{X_i\sim P_U}[p_{i,y}(X_i)]} \le 4\delta^2 \cdot \frac{\|\mathbb{E}_{X_i\sim P_U}[p_{i,y}(X_i)S_{\theta_U}(X_i)]\|_2^2}{\mathbb{E}_{X_i\sim P_U}[p_{i,y}(X_i)]}.$$

Consequently, by the triangle inequality $\|x+y\|_2^2 \le 2(\|x\|_2^2 + \|y\|_2^2)$ and Lemma 2, the above inequalities together with (5) and (6) imply that

$$\mathbb{E}_U\mathbb{E}_{Q_U}[L(\theta_U, \widehat{\theta}(Y))] \ge \frac{d_0\kappa}{4} \cdot \exp\left(-\frac{16nS}{d_0}\cdot\delta^2 - 4n\varepsilon^2\right), \tag{7}$$

where

$$S \triangleq \max_{u\in\{\pm1\}^{d_0}} \max_{i\in[n]} \sum_{y\in\{0,1\}^{nk}} \left(\prod_{s\neq i} \mathbb{E}_{X_s\sim P_u}[p_{s,y}(X_s)]\right) \cdot \frac{\|\mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i)S_{\theta_u}(X_i)]\|_2^2}{\mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i)]}. \tag{8}$$

Hence, to obtain the final minimax lower bound, it suffices to provide upper bounds of the quantity S in (8). This is the main focus of Section 4.

3.3 Approximately regular problems

In this subsection we show how to modify the above arguments to work for approximately regular problems. Note that when $\alpha = \omega(n^{-1})$, the lower bounds in Theorems 1 and 2 are negative and thus trivial; in the sequel we always assume that $\alpha = O(n^{-1}) = o(1)$. For each $u \in \{\pm1\}^{d_0}$, let

$\widetilde{P}_u(\cdot) = P_u(\cdot \mid \mathcal{X}_0)$ be the restriction of P_u to the set $\mathcal{X}_0$, and $\widetilde{Q}_u$ be the distribution of the transcript Y under $X_1, \cdots, X_n \sim \widetilde{P}_u$. Then by the property of $\mathcal{X}_0$ in Assumption 3, we have

$$\begin{aligned}
\max_{u \in \{\pm 1\}^{d_0}} \|\widetilde{Q}_u - Q_u\|_{\mathsf{TV}} &\le \max_{u \in \{\pm 1\}^{d_0}} \|\widetilde{P}_u^{\otimes n} - P_u^{\otimes n}\|_{\mathsf{TV}} \\
&\le n \cdot \max_{u \in \{\pm 1\}^{d_0}} \|\widetilde{P}_u - P_u\|_{\mathsf{TV}} \\
&= n \cdot \max_{u \in \{\pm 1\}^{d_0}} P_u(\mathcal{X}_0^c) \\
&\le n\alpha.
\end{aligned}$$

Consequently, applying the triangle inequality to the TV distance in (4) gives that

$$\mathbb{E}_U \mathbb{E}_{Q_U}[L(\theta_U, \widehat{\theta}(Y))] \ge \frac{d_0 \kappa}{2} \left(1 - \frac{1}{d_0} \sum_{j=1}^{d_0} \mathbb{E}_U \|\widetilde{Q}_U - \widetilde{Q}_{U^{\oplus j}}\|_{\mathsf{TV}} - 2n\alpha\right).$$

Hence, if we could show that the new statistical model $(\widetilde{P}_u)_{u \in \{\pm 1\}^{d_0}}$ is regular with essentially the same parameters in Assumption 2, and that the quantities I_0 and Σ_0 in Theorems 1 and 2 does not change much as we move from P_u to $\widetilde{P}_u$, we could repeat the same analysis in the previous subsection with (P_U, Q_U) replaced by $(\widetilde{P}_U, \widetilde{Q}_U)$ and arrive at the same results.

To verify Assumption 2, note that the regular grid assumption and separation condition do not depend on the statistical model and thus hold under $\widetilde{P}_u$ as well. For the likelihood ratio condition, note that for all $x \in \mathcal{X}_0$, we have

$$\frac{d\widetilde{P}_{u^{\oplus j}}}{d\widetilde{P}_u}(x) = \frac{dP_{u^{\oplus j}}}{dP_u}(x) \cdot \frac{P_u(\mathcal{X}_0)}{P_{u^{\oplus j}}(\mathcal{X}_0)}.$$

Therefore, the lower bound on the likelihood ratio could be replaced by $(1-\alpha)/2$, only slightly smaller than $1/2$. Moreover, as $dP_{u^{\oplus j}}/dP_u(x) \le 2$ for all $x \in \mathcal{X}_0$, by triangle inequality

$$\begin{aligned}
&\mathbb{E}_{X \sim \widetilde{P}_u} \left[\left| \frac{d\widetilde{P}_{\theta_{u^{\oplus j}}}}{d\widetilde{P}_{\theta_u}}(X) - 1 - (\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(X) \right|^2 \right] \\
&\le \frac{2}{1-\alpha} \mathbb{E}_{X \sim P_u} \left[\left| \frac{dP_{\theta_{u^{\oplus j}}}}{dP_{\theta_u}}(X) - 1 - (\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(X) \right|^2 \cdot \mathbb{1}(X \in \mathcal{X}_0) \right] + 2 \cdot \left(\frac{2\alpha}{1-\alpha} \right)^2,
\end{aligned}$$

therefore the condition (3) implies that the parameter ε in (2) could simply be replaced by $O(\varepsilon/\sqrt{1-\alpha} + \alpha)$. As $\alpha = o(1)$, the new statistical model becomes regular with essentially the same parameters.

To compute the new I_0 and Σ_0 under new models, note that for any non-negative function f, it holds that

$$\mathbb{E}_{\widetilde{P}_u}[f(X)] \le \frac{1}{P_u(\mathcal{X}_0)} \cdot \mathbb{E}_{P_u}[f(X)] \le \frac{1}{1-\alpha} \cdot \mathbb{E}_{P_u}[f(X)].$$

Consequently, the quantities I_0 and Σ_0 in Theorems 1 and 2 for regular problems could be replaced by slightly larger quantities $I_0/(1-\alpha)$ and $\Sigma_0/(1-\alpha)$, respectively.

Combining the above points, the minimax lower bounds for approximately regular problems could be argued in an entirely similar manner as regular problems.

4 Lower Bounds via Geometric Inequalities

In this section, we upper bound the quantity S in (8) using two different geometric inequalities, and complete the proof of main Theorems 1 and 2.

4.1 Proof of Theorem 1 via Geometric Inequality I

Note that under a deterministic protocol, each function $p_{i,y}$ only takes value in $\{0,1\}$. Therefore, if we write $\mathcal{X}_{i,y} = \{x \in \mathcal{X} : p_{i,y}(x) = 1\}$, then

$$\frac{\|\mathbb{E}_{X_i \sim P_u}[p_{i,y}(X_i) S_{\theta_u}(X_i)]\|_2^2}{\mathbb{E}_{X_i \sim P_u}[p_{i,y}(X_i)]} = P_u(\mathcal{X}_{i,y}) \cdot \|\mathbb{E}_{P_u}[S_{\theta_u}(X) \mid \mathcal{X}_{i,y}]\|_2^2.$$

Therefore, a quantity of interest is the ℓ_2-norm of the conditional mean of a random vector $S_{\theta_u}(X)$ restricted to some set $\mathcal{X}_{i,y}$. This motivates us to ask the following general question:

Question 1. *For a random vector $X \sim P$ and a target probability $P(A) = t \in (0,1)$, which subset $A \subseteq \mathcal{X}$ maximizes the ℓ_2 norm of the vector $\mathbb{E}[X \mid A]$? What is the corresponding maximum ℓ_2 norm?*

The following lemma presents an answer to Question 1 under the assumption that X has a finite second moment along any direction.

Lemma 3 (Geometric Inequality I)**.** *Assume that $\mathbb{E}[(u^\top X)^2] \le I_0$ for all unit vector $u \in \mathbb{R}^d$. Then for any set $A \subseteq \mathcal{X}$, the following inequality holds:*

$$\|\mathbb{E}[X \mid A]\|_2^2 \le I_0 \cdot \frac{1}{P(A)}.$$

Moreover, the RHS could be improved to $I_0 \cdot (1 - P(A))/P(A)$ if $\mathbb{E}[X] = 0$.

Note that Lemma 3 is a dimension-free result: the LHS depends on the dimensionality d, while the RHS does not. For a comparison, if we trivially use $\|\mathbb{E}[X \mid A]\|_2^2 \le \mathbb{E}[\|X\|_2^2 \mid A]$, there would be an additional factor of d on the RHS. The key observation in the dimensionality reduction is that the "independence" between coordinates of X is implied by the condition and needs to be exploited.

Now we have all necessary tools for the proof of Theorem 1. Applying Lemma 3 to the score function $S_{\theta_u}(x)$ in Theorem 1, we have

$$\frac{\|\mathbb{E}_{X_i \sim P_u}[p_{i,y}(X_i) S_{\theta_u}(X_i)]\|_2^2}{\mathbb{E}_{X_i \sim P_u}[p_{i,y}(X_i)]} \le I_0.$$

Consequently, by Lemma 2, it holds that

$$S \le I_0 \cdot \max_{u \in \{\pm 1\}^{d_0}} \max_{i \in [n]} \sum_{y \in \{0,1\}^{nk}} \left(\prod_{s \neq i} \mathbb{E}_{X_s \sim P_u}[p_{s,y}(X_s)] \right) = I_0 \cdot 2^k,$$

and plugging this upper bound of S into the minimax lower bound (7) completes the proof of one lower bound of Theorem 1. For the other lower bound independent of k, an alternative upper

bound of S could be used:

$$
\begin{aligned}
S &\overset{\text{(a)}}{\leq} \max_{u\in\{\pm 1\}^{d_0}} \max_{i\in[n]} \sum_{y\in\{0,1\}^{nk}} \left(\prod_{s\neq i} \mathbb{E}_{X_s\sim P_u}[p_{s,y}(X_s)]\right) \cdot \mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i)\cdot \|S_{\theta_u}(X_i)\|_2^2] \\
&\overset{\text{(b)}}{=} \max_{u\in\{\pm 1\}^{d_0}} \max_{i\in[n]} \mathbb{E}\left[\sum_{y\in\{0,1\}^{nk}} \left(\prod_{s=1}^{n} p_{s,y}(X_s)\right)\cdot \|S_{\theta_u}(X_i)\|_2^2\right] \\
&\overset{\text{(c)}}{=} \max_{u\in\{\pm 1\}^{d_0}} \max_{i\in[n]} \mathbb{E}_{X_i\sim P_u}[\|S_{\theta_u}(X_i)\|_2^2] \overset{\text{(d)}}{\leq} dI_0,
\end{aligned}
$$

where (a) is due to Cauchy–Schwarz, (b) follows from swapping the expectation and the sum with the expectation taken over i.i.d. $X_1, \cdots, X_n \sim P_u$, (c) is due to Lemma 2, and (d) follows from choosing $u = e_1, \cdots, e_d$, the canonical vectors, in the assumption of Theorem 1.

4.2 Proof of Theorem 2 via Geometric Inequality II

In this section, we provide another upper bound on $\|\mathbb{E}[X \mid A]\|_2^2$ when X is a sub-Gaussian random variable along any direction.

Lemma 4 (Geometric Inequality II)**.** *Assume that $\|u^\top X\|_{\psi_2}^2 \leq \Sigma_0$ for all unit vector $u \in \mathbb{R}^d$. Then for any set $A \subseteq \mathcal{X}$, the following inequality holds:*

$$
\|\mathbb{E}[X \mid A]\|_2^2 \leq \Sigma_0 \cdot \log \frac{2}{P(A)}.
$$

Note that lemma 4 presents a dimension-free upper bound again. Compared with Lemma 3, Lemma 4 improves the upper bound from $O(\Sigma_0)$ to $O(\Sigma_0 t \log \frac{1}{t})$ for sub-Gaussian random vector X, where $t = P(A)$ is the volume of the set A and Σ_0 is the sub-Gaussian parameter. Lemma 4 could be derived from standard arguments of the Talagrand's transportation-cost inequality [Led05, Chapter 6], but for completeness we provide two proofs of Lemma 4 in the appendix. The first proof directly reduces the problem to one dimension and then makes use of the Orlicz norm condition. The second proof is more geometric when X is exactly Gaussian, where tight constants are obtained for $X \sim \mathsf{Unif}(\{\pm 1\}^d)$ via information-theoretic inequalities, and then the "tensor power trick" is applied to prove the Gaussian case.

To move from Lemma 4 to an upper bound of the quantity S and therefore Theorem 2, note that the assumption of Theorem 2 and Lemma 4 show that

$$
\frac{\|\mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i) S_{\theta_u}(X_i)]\|_2^2}{\mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i)]} \leq \Sigma_0 \cdot \mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i)] \log \frac{2}{\mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i)]}.
$$

Therefore,

$$
\begin{aligned}
S &\leq \Sigma_0 \cdot \max_{u\in\{\pm 1\}^{d_0}} \max_{i\in[n]} \sum_{y\in\{0,1\}^{nk}} \left(\prod_{s=1}^{n} \mathbb{E}_{X_s\sim P_u}[p_{s,y}(X_s)]\right) \log \frac{2}{\mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i)]} \\
&\overset{\text{(a)}}{\leq} \Sigma_0 \cdot \max_{u\in\{\pm 1\}^{d_0}} \max_{i\in[n]} \log \left(\sum_{y\in\{0,1\}^{nk}} \left(\prod_{s=1}^{n} \mathbb{E}_{X_s\sim P_u}[p_{s,y}(X_s)]\right) \cdot \frac{2}{\mathbb{E}_{X_i\sim P_u}[p_{i,y}(X_i)]}\right) \\
&\overset{\text{(b)}}{=} \Sigma_0 \cdot \max_{u\in\{\pm 1\}^{d_0}} \max_{i\in[n]} \log(2^{k+1}) = (k+1)\Sigma_0 \log 2,
\end{aligned}
$$

where (a) is due to the first identity of Lemma 2 as well as the concavity of $x \mapsto \log x$, and (b) is due to the second identity of Lemma 2. Plugging this upper bound into the minimax lower bound (7) completes the proof of Theorem 2 (the other independent-of-k upper bound of S could be obtained analogously to the last section).

5 Discussions

5.1 Some Applications of Geometric Inequalities

The inequalities in Lemmas 3 and 4 have some other combinatorial applications related to geometry. We consider the following combinatorial problem on the binary Hamming cube $\Omega = \{\pm 1\}^d$:

1. Suppose we pick half of the vectors in Ω and compute the mean $\bar{v} \in \mathbb{R}^d$, i.e., $\bar{v} = |A|^{-1} \sum_{v \in A} v$ for some $A \subseteq \Omega, |A| = 2^{d-1}$, what is the maximum possible ℓ_2 norm $\|\bar{v}\|_2$?
2. Suppose we pick 2^{dR} vectors in Ω and compute the mean $\bar{v} \in \mathbb{R}^d$, where $R \in (0, 1)$, what is the dependence of the maximum possible ℓ_2 norm $\|\bar{v}\|_2$ on d and R?

This geometric problem is closely related to the optimal data compression in multiterminal statistical inference [Ama11]. We prove the following proposition:

Proposition 2. *Under the previous setting, we have*

$$\max_{A \subseteq \Omega : |A| = 2^{d-1}} \left\| \frac{1}{|A|} \sum_{v \in A} v \right\|_2 = 1,$$

$$\max_{A \subseteq \Omega : |A| = 2^{dR}} \left\| \frac{1}{|A|} \sum_{v \in A} v \right\|_2 = \sqrt{d}(1 - 2h_2^{-1}(R)) \cdot (1 + o_d(1)),$$

where $h_2(\cdot)$ is the binary entropy function defined in Lemma 6.

Proposition 2 gives the exact maximum ℓ_2 norm when $|A| = 2^{d-1}$ and its asymptotic behavior on d and R as $d \to \infty$ when $|A| = 2^{dR}$. We see that for $|A| = 2^{d-1}$, the maximum ℓ_2 norm is attained when A is the half space (or the $d-1$ dimensional sub-cube), i.e., $A = \{x \in \Omega : x_1 = 1\}$. However, for relatively small $|A| = 2^{dR}$, the maximum ℓ_2 norm is nearly attained at spherical caps, i.e., $A = \{x \in \Omega : d_{\mathsf{Ham}}(x, x_0) \le t\}$ for any fixed $x_0 \in \Omega$ and a proper radius t such that $|A| = 2^{dR}$. Hence, there are different behaviors for dense and sparse sets A.

5.2 Comparison with Strong Data Processing Inequalities (SDPI)

We compare our techniques with existing ones in establishing the lower bound for distributed parameter estimation problem. By Fano's inequality, the key step 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 the communication constraint $I(X; Y) \le k$. While trivially $I(U; Y) \le I(U; X)$ and $I(U; Y) \le I(X; Y)$, neither of these two inequalities are typically sufficient to obtain a good lower bound. A strong data processing inequality (SDPI)

$$I(U; Y) \le \gamma^*(U, X) I(X; Y), \qquad \forall p_{Y|X} \tag{9}$$

with $\gamma^*(U, X) < 1$ can be desirable. The SDPI may take different forms (e.g., for f-divergences), and it is applied in most works on distributed estimation, e.g., [ZDJW13, BGM+16, XR17]. The SDPI-based approach turns out to be tight in certain models (e.g., the Gaussian model [ZDJW13, BGM+16]), while it is also subject to some drawbacks:

1. The tight constant $\gamma^*(U, X)$ is hard to obtain in general;
2. The linearity of (9) in $I(X;Y)$ can only give a linear dependence of $I(U;Y)$ on k, which may not be tight. For example, in Corollary 1 the optimal dependence on k is exponential;
3. The conditional distribution $p_{Y^*|X}$ achieving the equality in (9) typically leads to $I(X;Y^*) \to 0$, and (9) may be loose for $I(X;Y) = k$;
4. The operational meaning of (9) is not clear, which may not result in a valid encoding scheme from X to Y.

In contrast to the linear dependence on k using SDPI, our technique implies that the dependence on k is closely related to the tail of the score function. It would be an interesting future direction to explore other dependence on k (instead of linear or exponential) in other statistical models.

6 Acknowledgements

We are grateful to Jayadev Acharya, Clément Canonne, Himanshu Tyagi, and Rishabh Dudeja for spotting an error in handling interactive communication protocols in the earlier version [HÖW18], and communicating with us. In particular, we would like to thank Jayadev Acharya, Clément Canonne, and Himanshu Tyagi for pointing out their recent paper [ACL+20] studying discrete distribution estimation models under the sequential communication protocol, which motivates the current proof technique to fix the error and handle the blackboard communication protocol for general models.

A Auxiliary Lemmas

Lemma 5. *[MU05] For $X \sim \mathsf{Poi}(\lambda)$ or $X \sim \mathsf{B}(n, \frac{\lambda}{n})$ and any $\delta > 0$, we have*

$$\mathbb{P}(X \geq (1+\delta)\lambda) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\lambda} \leq \exp\left(-\frac{(\delta^2 \wedge \delta)\lambda}{3}\right),$$
$$\mathbb{P}(X \leq (1-\delta)\lambda) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\lambda} \leq \exp\left(-\frac{\delta^2\lambda}{2}\right).$$

Lemma 6. *[Wyn73] For the binary entropy function $h_2(x) \triangleq -x\log_2 x - (1-x)\log_2(1-x)$ on $[0, \frac{1}{2}]$, let $h_2^{-1}(y)$ be its inverse for $y \in [0,1]$. Then the function*

$$f(y) = (1 - 2h_2^{-1}(y))^2$$

is a decreasing concave function, with $f(y) \leq 2\log 2 \cdot (1-y)$ for all $y \in [0,1]$.

B Proof of Main Lemmas

B.1 Proof of Lemma 2

We prove a stronger result: for any strategy $\{a_v(\cdot)\}$, if each path from the root to any leaf node visits exactly k_i internal nodes with label i for each $i \in [n]$, then

$$\sum_{y \in \{0,1\}^{\sum_{i=1}^n k_i}} \prod_{v \in \tau(y), l_v \neq i} b_{v,y}(x_{l_v}) = 2^{k_i} \tag{10}$$

for any $\{x_j\}_{j\neq i}$. Clearly (10) implies the lemma (i.e., with $k_i = 0$ and $k_i = k$, respectively).

We prove (10) by induction on the depth $D = \sum_{i=1}^n k_i$ of the binary tree. The base case $D = 0$ is obvious. To move from D to $D+1$, distinguish into two cases and apply the induction hypothesis to the left/right tree of the root:

1. If the root node is labeled as i, then (10) follows from $2^{k_i} = 2^{k_i-1} + 2^{k_i-1}$;
2. If the root node is not labeled as i, then (10) follows from $2^{k_i} = 2^{k_i} a_{\text{root}}(x_i) + 2^{k_i}(1 - a_{\text{root}}(x_i))$.

B.2 Proof of Lemma 3

As $\|x\|_2 = \max_{u:\|u\|_2=1} u^\top x$, it suffices to prove the same upper bounds of $\mathbb{E}[u^\top X \mid A]^2$ for any unit vector $u \in \mathbb{R}^d$. First, by the Cauchy–Schwarz inequality, we have

$$\mathbb{E}[u^\top X \mid A]^2 \le \mathbb{E}[(u^\top X)^2 \mid A] \le \frac{\mathbb{E}[(u^\top X)^2]}{P(A)} \le \frac{I_0}{P(A)},$$

establishing the first inequality. The improved inequality when $\mathbb{E}[X] = 0$ is due to

$$\begin{aligned}
I_0 &\ge \mathbb{E}[(u^\top X)^2] \\
&= \mathbb{E}[(u^\top X)^2 \mathbb{1}(X \in A)] + \mathbb{E}[(u^\top X)^2 \mathbb{1}(X \in A^c)] \\
&\overset{\text{(a)}}{\ge} \frac{\mathbb{E}^2[(u^\top X)\mathbb{1}(X \in A)]}{P(A)} + \frac{\mathbb{E}^2[(u^\top X)\mathbb{1}(X \in A^c)]}{1 - P(A)} \\
&\overset{\text{(b)}}{=} \frac{\mathbb{E}^2[(u^\top X)\mathbb{1}(X \in A)]}{P(A)} + \frac{\mathbb{E}^2[(u^\top X)\mathbb{1}(X \in A)]}{1 - P(A)} \\
&= \frac{P(A)}{1 - P(A)} \cdot \mathbb{E}[u^\top X \mid A]^2,
\end{aligned}$$

where (a) is due to Cauchy–Schwarz, and (b) follows from the assumption $\mathbb{E}[X] = 0$.

B.3 Proof of Lemma 4

By the definition of the Orlicz ψ_2-norm, for any unit vector $u \in \mathbb{R}^d$ we have

$$\begin{aligned}
2 &\ge \mathbb{E}\left[\exp\left(\frac{(u^\top X)^2}{\Sigma_0}\right)\right] \\
&\ge P(A) \cdot \mathbb{E}\left[\exp\left(\frac{(u^\top X)^2}{\Sigma_0}\right) \middle| A\right] \\
&\ge P(A) \cdot \exp\left(\frac{(u^\top \mathbb{E}[X \mid A])^2}{\Sigma_0}\right),
\end{aligned}$$

where the last inequality follows from the convexity of $x \mapsto \exp(cx^2)$ for any $c > 0$. Consequently, we have $u^\top \mathbb{E}[X \mid A] \le \Sigma_0 \log(2/P(A))$ for all unit vectors $u \in \mathbb{R}^d$, and the result follows.

B.4 Another Proof of Lemma 4 in Gaussian Case

We prove the following lemma:

Lemma 7. *For $X \sim \mathcal{N}(0, I_d)$ and any measurable $A \subseteq \mathbb{R}^d$, we have*

$$\|\mathbb{E}[X \mid A]\|_2^2 \le 2 \cdot \log \frac{1}{\mathbb{P}(A)}.$$

We split the proof into two steps: we first consider the uniform distribution on the binary hypercube, and then use the tensor power trick to reduce to the Gaussian case.

B.4.1 Geometric inequality on binary hypercube

We prove the following lemma:

Lemma 8. *For $X \sim \mathsf{Unif}(\{\pm 1\}^d)$ and any non-negative function $a(\cdot) \in [0,1]$, we have*

$$\left\| \frac{\mathbb{E}[Xa(X)]}{\mathbb{E}[a(X)]} \right\|_2^2 \le 2 \cdot \log \frac{1}{\mathbb{E}[a(X)]}$$

Moreover, the dimension-free constant 2 cannot be improved.

Proof. Define a new probability measure $Q(\cdot)$ on the binary hypercube $\{\pm 1\}^d$ with $Q(y) \propto a(y)$, and let $Y \sim Q$. Let $p_i \triangleq \mathbb{P}(Y_i = 1)$ for $i \in [d]$, then

$$\left\| \frac{\mathbb{E}[Xa(X)]}{\mathbb{E}[a(X)]} \right\|_2^2 = \|\mathbb{E}Y\|_2^2 = \sum_{i=1}^d (\mathbb{E}Y_i)^2 = \sum_{i=1}^d (1-2p_i)^2.$$

Recall the definition of $h_2(\cdot)$ in Lemma 6. Define $q_i \triangleq h_2(p_i)$, the concavity in Lemma 6 gives

$$\left\| \frac{\mathbb{E}[Xa(X)]}{\mathbb{E}[a(X)]} \right\|_2^2 = \sum_{i=1}^d (1-2h_2^{-1}(q_i))^2 \le d\left(1 - 2h_2^{-1}\left(\frac{1}{d}\sum_{i=1}^d q_i\right)\right)^2.$$

On the other hand, by the subadditivity of Shannon entropy,

$$\begin{aligned}\sum_{i=1}^d q_i = \frac{1}{\log 2}\sum_{i=1}^d H(Y_i) \ge \frac{H(Y)}{\log 2} &= d - \mathbb{E}\left[\log_2 \frac{a(Y)}{\mathbb{E}[a(X)]}\right] \\ &\ge d - \mathbb{E}\left[\log_2 \frac{1}{\mathbb{E}[a(X)]}\right] = d - \log_2 \frac{1}{\mathbb{E}[a(X)]}.\end{aligned}$$

Hence, applying the decreasing property and the last inequality in Lemma 6, we have

$$\begin{aligned}\left\| \frac{\mathbb{E}[Xa(X)]}{\mathbb{E}[a(X)]} \right\|_2^2 &\le d\left(1 - 2h_2^{-1}\left(1 - \frac{1}{d}\log_2 \frac{1}{\mathbb{E}[a(X)]}\right)\right)^2 \\ &\le d \cdot 2\log 2 \cdot \frac{1}{d}\log_2 \frac{1}{\mathbb{E}[a(X)]} \\ &= 2\log \frac{1}{\mathbb{E}[a(X)]}.\end{aligned}$$

To show that 2 is the best possible constant, pick $a(x) = \mathbb{1}_B(x)$ where B is the Hamming ball with center $\mathbf{1}$ and radius ϵd. Direct computation gives the constant 2 as $d \to \infty$ and $\epsilon \to 0$. □

B.4.2 Tensor Power Trick

Next we make use of Lemma 8 to prove the Gaussian case. We apply the so-called *tensor power trick*: we lift the dimension by making B independent copies, and apply CLT to move to the Gaussian case as $B \to \infty$. This idea has been widely used in harmonic analysis and high-dimensional geometry, e.g., to prove the isoperimetric inequality for the Gaussian measure [Led05].

Here the trick goes: fix any dimension d and any function $a(\cdot) \in [0,1]$ defined on $\mathbb{R}^d$. By a suitable approximation we may assume that $a(\cdot)$ is continuous. Now for any $B > 0$, we define a new function $\tilde{a}(\cdot)$ on $\{\pm 1\}^{dB}$ as follows:

$$\tilde{a}(X) = \tilde{a}(\{X_{i,j}\}_{i\in[d],j\in[B]}) \triangleq a\left(\frac{\sum_{j=1}^n X_{1,j}}{\sqrt{B}}, \cdots, \frac{\sum_{j=1}^n X_{d,j}}{\sqrt{B}}\right).$$

By symmetry, we have

$$\|\mathbb{E}[X\tilde{a}(X)]\|_2^2 = \sum_{i=1}^d \left(\mathbb{E}\left[\frac{\sum_{j=1}^B X_{i,j}}{\sqrt{B}} a\left(\frac{\sum_{j=1}^n X_{1,j}}{\sqrt{B}}, \cdots, \frac{\sum_{j=1}^n X_{d,j}}{\sqrt{B}}\right)\right]\right)^2.$$

Moreover, by Lemma 8, we have

$$\left\|\frac{\mathbb{E}[X\tilde{a}(X)]}{\mathbb{E}[\tilde{a}(X)]}\right\|_2^2 \le 2 \cdot \log \frac{1}{\mathbb{E}[\tilde{a}(X)]}. \tag{11}$$

Let $Z \sim \mathcal{N}(0, I_d)$, then CLT gives $\|\mathbb{E}[X\tilde{a}(X)]\|_2^2 \to \|\mathbb{E}[Za(Z)]\|_2^2$ and $\mathbb{E}[\tilde{a}(X)] \to \mathbb{E}[a(Z)]$ as $B \to \infty$. Hence, as $B \to \infty$, (11) becomes

$$\left\|\frac{\mathbb{E}[Za(Z)]}{\mathbb{E}[a(Z)]}\right\|_2^2 \le 2 \cdot \log \frac{1}{\mathbb{E}[a(Z)]}. \tag{12}$$

Note that (12) holds for all d and $a(\cdot)$, the proof of Lemma 7 is complete by choosing $a(\cdot) = \mathbb{1}_A(\cdot)$.

C Proof of Propositions

C.1 Proof of Proposition 1

Product Bernoulli models. We begin with the first product Bernoulli model. For any $\delta \in (0, 1/6)$, we choose $\theta_u = (1/2, \cdots, 1/2) + \delta u \in \Theta$ for all $u \in \{\pm 1\}^d$. Clearly the regular grid condition and the separation condition hold. For the likelihood ratio condition, note that

$$\frac{dP_{\theta_{u\oplus j}}}{dP_{\theta_u}}(x) = \frac{1-2\delta u_j}{1+2\delta u_j} \cdot \mathbb{1}(x_j = 1) + \frac{1+2\delta u_j}{1-2\delta u_j} \cdot \mathbb{1}(x_j = 0) \ge \frac{1}{2}$$

by the choice of δ. Moreover, the j-th component of the score function at θ_u is

$$[S_{\theta_u}(x)]_j = \frac{2}{1+2\delta u_j} \cdot \mathbb{1}(x_j = 1) - \frac{2}{1-2\delta u_j} \cdot \mathbb{1}(x_j = 0),$$

therefore (2) is satisfied with $\varepsilon \equiv 0$.

Multinomial models. For $d_0 = d/2$, consider the following construction known as the Paninski's construction [Pan08]:

$$\theta_u = \left(\frac{1}{d} - \frac{\delta u_1}{2}, \frac{1}{d} + \frac{\delta u_1}{2}, \cdots, \frac{1}{d} - \frac{\delta u_{d_0}}{2}, \frac{1}{d} + \frac{\delta u_{d_0}}{2}\right).$$

After proper permutation of the rows, it is easy to see that the matrix M_u is $\delta \cdot [\mathrm{diag}(v)\ \mathrm{diag}(v)]^\top$ for some $v \in \{\pm 1\}^{d_0}$. Consequently, the operator norm of this matrix is $\sqrt{2}\delta$, which is smaller than

2δ. Also, after simple algebra, the separation condition (1) is fulfilled with $\kappa(\delta) = 2^{2-p}\delta^p$. For the likelihood ratio condition, note that

$$\frac{dP_{\theta_{u^{\oplus j}}}}{dP_{\theta_u}}(x) = \frac{2 + d\delta u_j}{2 - d\delta u_j} \cdot \mathbb{1}(x = 2j-1) + \frac{2 - d\delta u_j}{2 + d\delta u_j} \cdot \mathbb{1}(x = 2j) \geq \frac{1}{2}$$

as $\delta \in (0, 1/(2d))$. Moreover, although there is some ambiguity in defining the score function for the Multinomial model (depending on the choice of free parameters), the inner product $(\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(x)$ is well-defined and expressed as

$$(\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(x) = \frac{\delta u_j}{1/d - \delta u_j/2} \cdot \mathbb{1}(x = 2j-1) - \frac{\delta u_j}{1/d + \delta u_j/2} \cdot \mathbb{1}(x = 2j).$$

Therefore, (2) holds with $\varepsilon \equiv 0$. The product Bernoulli model is handled analogously.

Gaussian location models. Choose $\theta_u = \delta u \in \mathbb{R}^d$ for all $u \in \{\pm 1\}^d$, then clearly the regular grid condition and the separation condition hold. Let $\mathcal{X}_0 = \{x \in \mathbb{R}^d : \|x\|_\infty \leq (C\sqrt{\log(nd)} + 1)\sigma\}$, then for a large enough constant $C > 0$, using the Gaussian tail and the union bound gives that $P_{\theta_u}(\mathcal{X}_0) \geq 1 - o(n^{-1})$ for any $u \in \{\pm 1\}^d$ and $\delta < \sigma$. For the likelihood ratio condition, we have

$$\frac{dP_{\theta_{u^{\oplus j}}}}{dP_{\theta_u}}(x) = \exp\left(-\frac{2\delta u_j x_j}{\sigma^2}\right) \geq \frac{1}{2}, \quad \forall x \in \mathcal{X}_0$$

as $|\delta| \leq c\sigma/\sqrt{\log(nd)}$ for a small enough constant $c > 0$ and $|x_j| \leq (C\sqrt{\log(nd)} + 1)\sigma$. Moreover, $S_{\theta_u}(x) = (x - \theta_u)/\sigma^2$, and therefore

$$\begin{aligned}\left|\frac{dP_{\theta_{u^{\oplus j}}}}{dP_{\theta_u}}(x) - 1 - (\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(x)\right| &= \left|\exp\left(-\frac{2\delta u_j x_j}{\sigma^2}\right) - 1 - \frac{2\delta u_j(\delta u_j - x_j)}{\sigma^2}\right| \\ &= \left|\exp\left(-\frac{2\delta^2}{\sigma^2} - \frac{2\delta u_j z_j}{\sigma}\right) - 1 + \frac{2\delta u_j z_j}{\sigma}\right|,\end{aligned}$$

with $x_j \triangleq \sigma z_j + \delta u_j$. Note that when $x \sim P_{\theta_u}$, we have $z_j \sim \mathcal{N}(0,1)$, and therefore the above term has an explicit second moment as

$$\mathbb{E}_{X \sim P_{\theta_u}}\left[\left|\frac{dP_{\theta_{u^{\oplus j}}}}{dP_{\theta_u}}(X) - 1 - (\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(X)\right|^2\right] = \exp\left(\frac{4\delta^2}{\sigma^2}\right) - 1 - \frac{4\delta^2}{\sigma^2},$$

which is ε^2 with $\varepsilon = O(\delta^2/\sigma^2)$ as $\delta = O(\sigma)$. Hence, the Gaussian location model is approximately regular with $\varepsilon(\delta) = O(\delta^2/\sigma^2)$.

Logistic regression models with random design. Choose $\theta_u = \delta u \in \mathbb{R}^d$ for all $u \in \{\pm 1\}^d$, then clearly the regular grid condition and the separation condition hold. For the likelihood ratio condition, we choose

$$\mathcal{X}_0 = \{(z, y) : \|z\|_\infty \leq C\sqrt{\log(nd)}, y \in \{0, 1\}\},$$

and for a large enough constant $C > 0$ we have $P_\theta(\mathcal{X}_0) \geq 1 - \alpha$ with $\alpha = o(n^{-1})$ for any $\theta \in \mathbb{R}^d$. We first show that for any fixed $y \in \{0, 1\}$, taking only the expectation with respect to $z \sim \mathcal{N}(0, I_d)$ satisfies (3). By symmetry, we shall only consider the case $y = 1$, where $X = (z, 1)$ and

$$\left|\frac{dP_{\theta_{u^{\oplus j}}}}{dP_{\theta_u}}(z,1) - 1 - (\theta_{u^{\oplus j}} - \theta_u)^\top S_{\theta_u}(z,1)\right| = \left|\frac{e^{-\theta^\top z} + 1}{e^{-\theta^\top z} \cdot e^{2\delta u_j z_j} + 1} - 1 + 2\delta u_j z_j \cdot \frac{e^{-\theta^\top z}}{e^{-\theta^\top z} + 1}\right|.$$

We will prove that for any $A \geq 0$ and $t \in [-1/2, 1/2]$, it holds that

$$\left| \frac{A+1}{Ae^t+1} - 1 + \frac{At}{A+1} \right| \leq 2t^2. \tag{13}$$

In fact, if (13) holds, the choice of $\delta \in (0, 1/\sqrt{d})$ satisfies $|2\delta u_j z_j| \leq 1/2$ for any $X = (z, y) \in \mathcal{X}_0$ with large d. Now choosing $A = e^{-\theta^\top z}$ and $t = 2\delta u_j z_j \in [-1/2, 1/2]$ in (13) gives the desired inequality (3) with $\varepsilon = O(\delta^2)$.

Next we prove the inequality (13). After simple algebra, it is equivalent to prove that

$$\frac{A}{A+1} \cdot \left| \frac{[(t-1)e^t + 1]A + (1 + t - e^t)}{Ae^t + 1} \right| \leq 2t^2.$$

Since $|t| \leq 1/2$, it is easy to verify that $|(t-1)e^t + 1| \leq t^2$ and $|1 + t - e^t| \leq t^2$, and clearly $e^t \geq 1/2$. Consequently, the above inequality holds, and so is (13).

Finally, we verify $dP_{\theta_{u \oplus j}} / dP_{\theta_u}(x) \geq 1/2$ for all $x \in \mathcal{X}_0$. Using the above notations $A = e^{-\theta^\top z}$ and $t = 2\delta u_j z_j \in [-1/2, 1/2]$ again, this quantity is

$$\frac{A+1}{Ae^t+1} \geq \min\left\{1, e^{-t}\right\} \geq \frac{1}{2},$$

as desired.

C.2 Proof of Proposition 2

Let X follow the uniform distribution on Ω, then $\bar{v} = \mathbb{E}[X \mid A]$. As X has independent coordinates each of which has a unit second moment, the assumption of Lemma 3 is fulfilled with $I_0 = 1$. By Lemma 3, for $|A| = 2^{d-1}$ we have

$$\|\mathbb{E}[X \mid A]\|_2 \leq 1 \cdot \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)} = 1,$$

establishing the first inequality. Similarly, the second inequality follows from Lemma 8 (and its proof).

D Proof of Theorem 3

As the hypothesis class for sparse Gaussian models is typically not cube-like, we use the following Fano's inequality instead of the Assouad's lemma to establish the lower bound. The present form is taken from [DW13, Corollary 1]; see also [CGZ16] and [Han19, Theorem 8] for a general statement.

Lemma 9. *Let random variables V and $\widehat{V}$ take value in $\mathcal{V}$, V be uniform on some finite alphabet $\mathcal{V}$, and $V - X - \widehat{V}$ form a Markov chain. Let d be any metric on $\mathcal{V}$, and for $t > 0$, define*

$$N_{\max}(t) \triangleq \max_{v \in \mathcal{V}} |v' \in V : d(v, v') \leq t|,$$
$$N_{\min}(t) \triangleq \min_{v \in \mathcal{V}} |v' \in V : d(v, v') \leq t|.$$

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

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

We construct the following family of hypotheses: let $U \in \mathbb{R}^d$ be uniformly distributed on the finite set

$$\mathcal{U} = \{\theta \in \{0, \pm 1\}^d : \|\theta\|_0 = s\}.$$

Clearly $|\mathcal{U}| = 2^s \binom{d}{s}$. For $u \in \mathcal{U}$ we associate with the Gaussian distribution $P_u \triangleq \mathcal{N}(\delta u, \sigma^2 I_d)$, draw n i.i.d. observations $X = (X_1, \cdots, X_n)$ from P_u, and obtain the transcript $Y = (Y_1, \cdots, Y_n) \in \{0,1\}^{nk}$, where $Y_i \in \{0,1\}^k$ denotes the transcript from node i under the simultaneous message passing protocol. Choosing $t = s/5$ in Lemma 9, we have

$$\left|\left\{u' \in \mathcal{U} : d_{\mathsf{Ham}}(u, u') \le \frac{s}{5}\right\}\right| = \sum_{u+v \le \frac{s}{5}} \binom{s}{u}\binom{s-u}{v}\binom{d-s}{v} \le \left(\frac{s}{5}+1\right)^2 \cdot \binom{s}{s/5}^2 \binom{d}{s/5}.$$

As a result, we have $\log \frac{|\mathcal{U}|}{N_{\max}(s/5)} \ge cs \log \frac{d}{s}$ for some constant $c > 0$, and Lemma 9 gives

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_2^2 \ge \frac{s\delta^2}{10}\left(1 - \frac{I(U;Y) + \log 2}{cs \log(d/s)}\right). \tag{14}$$

To lower bound (14), we seek an upper bound of $I(U; Y_i)$ for each $i \in [n]$. Under the simultaneous message passing protocol, the communication strategy of node i could be represented by a family of non-negative functions $p_{i,y}(\cdot)$ with $y \in \{0,1\}^k$, where

$$p_{i,y}(x) \triangleq \mathbb{P}[Y_i = y \mid X_i = x].$$

Clearly $\sum_{y \in \{0,1\}^k} p_{i,y}(x) = 1$ for all $x \in \mathbb{R}^d$. Moreover,

$$\mathbb{P}[Y_i = y \mid U = u] = \mathbb{E}_{X_i \sim P_u}[p_{i,y}(X_i)].$$

Let $P_0 \triangleq \mathcal{N}(0, \sigma^2 I_d)$, we could upper bound the mutual information as

$$\begin{aligned}
I(U; Y_i) &\overset{\text{(a)}}{\le} \mathbb{E}_U[D_{\mathsf{KL}}(P_{Y_i|U} \| P_{Y_i|U=0})] \\
&\overset{\text{(b)}}{\le} \mathbb{E}_U[\chi^2(P_{Y_i|U} \| P_{Y_i|U=0})] \\
&\overset{\text{(c)}}{=} \sum_{y \in \{0,1\}^k} \mathbb{E}_U\left[\frac{\mathbb{E}^2_{X_i \sim P_0}[p_{i,y}(X_i) s_U(X_i)]}{\mathbb{E}_{X_i \sim P_0}[p_{i,y}(X_i)]}\right] \\
&\overset{\text{(d)}}{\le} \frac{2\delta^2}{\sigma^4} \sum_{y \in \{0,1\}^k} \mathbb{E}_U\left[\frac{\mathbb{E}^2_{X_i \sim P_0}[p_{i,y}(X_i) \cdot U^\top X_i]}{\mathbb{E}_{X_i \sim P_0}[p_{i,y}(X_i)]}\right] + 2 \sum_{y \in \{0,1\}^k} \mathbb{E}_U\left[\frac{\mathbb{E}^2_{X_i \sim P_0}[p_{i,y}(X_i) \varepsilon_U(X_i)]}{\mathbb{E}_{X_i \sim P_0}[p_{i,y}(X_i)]}\right],
\end{aligned}$$

where (a) is due to the variational representation of the mutual information

$$I(X;Y) = \min_{Q_Y} \mathbb{E}_X[D_{\mathsf{KL}}(P_{Y|X} \| Q_Y)],$$

(b) uses the fact that the KL divergence is upper bounded by the χ^2 divergence, (c) follows from simple algebra with

$$s_U(x) \triangleq \frac{dP_U}{dP_0}(x) - 1 = \exp\left(\frac{\delta \cdot U^\top x}{\sigma^2} - \frac{\delta^2 s}{2\sigma^2}\right) - 1,$$

and (d) uses the triangle inequality $(a+b)^2 \le 2(a^2+b^2)$ with

$$\varepsilon_U(x) \triangleq \exp\left(\frac{\delta \cdot U^\top x}{\sigma^2} - \frac{\delta^2 s}{2\sigma^2}\right) - 1 - \frac{\delta \cdot U^\top x}{\sigma^2}.$$

Next we upper bound each term separately. For the remainder term, in view of the identity

$$\mathbb{E}_{X\sim P_0}[\varepsilon_U(X)^2] = \exp\left(\frac{\delta^2 s}{\sigma^2}\right) - 1 - \frac{\delta^2 s}{\sigma^2} = O\left(\frac{\delta^4 s^2}{\sigma^4}\right)$$

as long as $\delta = O(\sigma/\sqrt{s})$, the Cauchy–Schwarz inequality gives

$$\sum_{y\in\{0,1\}^k} \mathbb{E}_U\left[\frac{\mathbb{E}^2_{X_i\sim P_0}[p_{i,y}(X_i)\varepsilon_U(X_i)]}{\mathbb{E}_{X_i\sim P_0}[p_{i,y}(X_i)]}\right] \le \mathbb{E}_U \mathbb{E}_{X_i\sim P_0}\left[\sum_{y\in\{0,1\}^k} p_{i,y}(X_i)\varepsilon_U(X_i)^2\right] = O\left(\frac{\delta^4 s^2}{\sigma^4}\right). \quad (15)$$

As for the main term, we have

$$\begin{aligned}
\sum_{y\in\{0,1\}^k} \mathbb{E}_U\left[\frac{\mathbb{E}^2_{X_i\sim P_0}[p_{i,y}(X_i)\cdot U^\top X_i]}{\mathbb{E}_{X_i\sim P_0}[p_{i,y}(X_i)]}\right] &\overset{\text{(a)}}{=} \frac{s}{d}\sum_{y\in\{0,1\}^k} \frac{\|\mathbb{E}_{X_i\sim P_0}[p_{i,y}(X_i)X_i]\|_2^2}{\mathbb{E}_{X_i\sim P_0}[p_{i,y}(X_i)]} \\
&\overset{\text{(b)}}{\le} \frac{2s\sigma^2}{d}\sum_{y\in\{0,1\}^k} \mathbb{E}_{X_i\sim P_0}[p_{i,y}(X_i)]\log\frac{1}{\mathbb{E}_{X_i\sim P_0}[p_{i,y}(X_i)]} \\
&\overset{\text{(c)}}{\le} \frac{2s\sigma^2}{d}\cdot k, \quad (16)
\end{aligned}$$

where (a) follows from $\mathbb{E}[UU^\top] = (s/d)I_d$, (b) follows from Lemma 4 (or more precisely, Lemma 7), and (c) uses the concavity of $x \mapsto \log x$. Now combining (15) and (16) gives the following upper bound on the mutual information:

$$I(U;Y_i) = O\left(\frac{sk\delta^2}{d\sigma^2} + \frac{s^2\delta^4}{\sigma^4}\right).$$

Without loss of generality we may assume that there is no public randomness (otherwise we use $I(U;Y) \le I(U;Y|R)$ for external randomness R and repeat the previous arguments). Consequently, $(Y_1,\cdots,Y_n)$ are conditionally independent given U, and therefore

$$I(U;Y) \le \sum_{i=1}^n I(U;Y_i) = O\left(\frac{nsk\delta^2}{d\sigma^2} + \frac{ns^2\delta^4}{\sigma^4}\right). \quad (17)$$

Finally, choosing $\delta^2 \asymp d\sigma^2\log(d/s)/(nk)$ in (14) and (17) completes the proof of Theorem 3 for $k \le d$ (also recall our choice of n). The case $k > d$ simply follows from the centralized minimax risk and is thus omitted.

References

[ACF+21] Jayadev Acharya, Clément L Canonne, Cody Freitag, Ziteng Sun, and Himanshu Tyagi. Inference under information constraints iii: Local privacy constraints. *IEEE Journal on Selected Areas in Information Theory*, 2(1):253–267, 2021.

[ACH+20] Jayadev Acharya, Clément L Canonne, Yanjun Han, Ziteng Sun, and Himanshu Tyagi. Domain compression and its application to randomness-optimal distributed goodness-of-fit. In *Conference on Learning Theory*, pages 3–40, 2020.

[ACL+20] Jayadev Acharya, Clément L Canonne, Yuhan Liu, Ziteng Sun, and Himanshu Tyagi. Interactive inference under information constraints. *arXiv preprint arXiv:2007.10976*, 2020.

[ACT20a] Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Distributed signal detection under communication constraints. In *Conference on Learning Theory*, pages 41–63. PMLR, 2020.

[ACT20b] Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints i: Lower bounds from chi-square contraction. *IEEE Transactions on Information Theory*, 66(12):7835–7855, 2020.

[ACT20c] 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.

[ACT21] Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Unified lower bounds for interactive high-dimensional estimation under information constraints. *arXiv preprint arXiv:2010.06562v4*, 2021.

[AJM20] Kareem Amin, Matthew Joseph, and Jieming Mao. Pan-private uniformity testing. In *Conference on Learning Theory*, pages 183–218. PMLR, 2020.

[AKLS20] Jayadev Acharya, Peter Kairouz, Yuhan Liu, and Ziteng Sun. Estimating sparse discrete distributions under local privacy and communication constraints. *arXiv preprint arXiv:2011.00083*, 2020.

[Ama11] Shun-ichi Amari. On optimal data compression in multiterminal statistical inference. *IEEE Transactions on Information Theory*, 57(9):5577–5587, 2011.

[AS19] Jayadev Acharya and Ziteng Sun. Communication complexity in locally private distribution estimation and heavy hitters. In *International Conference on Machine Learning*, pages 51–60. PMLR, 2019.

[Ass83] Patrice Assouad. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

[ASZ19] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1120–1129, 2019.

[BB20] Thomas Berrett and Cristina Butucea. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. In *Advances in Neural Information Processing Systems*, volume 33, pages 3164–3173, 2020.

[BBFM12] Maria Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. In *Conference on Learning Theory*, pages 26–1, 2012.

[BCÖ20] Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3):645–659, 2020.

[BDF+18] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

[BGM+16] 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.

[BHÖ20] Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. Lower bounds for learning distributions under communication constraints via Fisher information. *Journal of Machine Learning Research*, 21(236):1–30, 2020.

[BO31] Z Birnbaum and W-f Orlicz. Über die verallgemeinerung des begriffes der zueinander konjugierten potenzen. *Studia Mathematica*, 3(1):1–67, 1931.

[BO19] Leighton Pate Barnes and Ayfer Ozgur. Minimax bounds for distributed logistic regression. *arXiv preprint arXiv:1910.01625*, 2019.

[BPC+11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[BYJKS04] Ziv Bar-Yossef, Thathachar S Jayram, Ravi Kumar, and D Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[CGZ16] Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On bayes risk lower bounds. *The Journal of Machine Learning Research*, 17(1):7687–7744, 2016.

[CKO20] Wei-Ning Chen, Peter Kairouz, and Ayfer Ozgur. Breaking the communication-privacy-accuracy trilemma. In *Advances in Neural Information Processing Systems*, volume 33, pages 3312–3324, 2020.

[CKÖ21] Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Breaking the dimension dependence in sparse distribution estimation under communication constraints. *arXiv preprint arXiv:2106.08597*, 2021.

[DGBSX12] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

[DGL+17] Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, and Ludwig Schmidt. Communication-efficient distributed learning of discrete distributions. In *Advances in Neural Information Processing Systems*, pages 6394–6404, 2017.

[DIPSV12] Hal Daume III, Jeff Phillips, Avishek Saha, and Suresh Venkatasubramanian. Protocols for learning classifiers on distributed data. In *Artificial Intelligence and Statistics*, pages 282–290, 2012.

[DJW13] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 429–438. IEEE, 2013.

[DJW18] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[DPSV12] Hal Daumé, Jeff M Phillips, Avishek Saha, and Suresh Venkatasubramanian. Efficient protocols for distributed classification and optimization. In *International Conference on Algorithmic Learning Theory*, pages 154–168. Springer, 2012.

[DR19] John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pages 1161–1191, 2019.

[DW13] John C Duchi and Martin J Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

[GMN14] Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014.

[Háj70] Jaroslav Hájek. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(4):323–330, 1970.

[Háj72] Jaroslav Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194, 1972.

[Han19] Yanjun Han. Lecture 8: multiple hypothesis testing: tree, fano, and assouad. 2019.

[HMÖW18] Yanjun Han, Pritam Mukherjee, Ayfer Özgür, and Tsachy Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 506–510. IEEE, 2018.

[HÖW18] Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. *Conference on Learning Theory (COLT)*, pages 3163–3188, 2018.

[IH13] Ildar Abdulovich Ibragimov and Rafail Z Has'minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.

[KBR16] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444, 2016.

[KMA+19] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[KN97] E Kushilevitz and N Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[Led05] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2005.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Sha14] Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*, pages 163–171, 2014.

[Tsy08] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, 2008.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Wyn73] A Wyner. A theorem on the entropy of certain binary sequences and applications–ii. *IEEE Transactions on Information Theory*, 19(6):772–777, 1973.

[XR17] Aolin Xu and Maxim Raginsky. Information-theoretic lower bounds on Bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600, 2017.

[YB18] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018.

[Yu97] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[ZDJW13] 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.