Proceedings of Machine Learning Research vol 75:1–26, 2018 31st Annual Conference on Learning Theory

Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints

Yanjun Han YJHAN@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Ayfer Özgür AOZGUR@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Tsachy Weissman TSACHY@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

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 under squared ℓ_2 loss for a large class of 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, and dense/sparse Gaussian location models 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.

Keywords: Distributed estimation; Minimax lower bound; High-dimensional geometry; Blackboard communication protocol; Strong data processing inequality

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 Boyd et al. (2011); Balcan et al. (2012); Daume III et al. (2012); Daumé et al. (2012); Dekel et al. (2012).

In this paper, we focus on the impact of a finite-communication budget per sample on the performance of several statistical estimation problems. 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 $\theta \in \Theta \subset \mathbb{R}^d$ under squared ℓ_2 loss. In most examples throughout, we will assume that P_θ enjoys a product structure

$$P_\theta = p_{\theta_1} \times p_{\theta_2} \times \cdots \times p_{\theta_d}, \qquad \theta = (\theta_1, \cdots, \theta_d) \in \mathbb{R}^d.$$

Unlike the traditional setting where $X_1, \cdots, X_n$ are available to the estimator as they are, we consider a distributed setting where each observation X_i is available at a different sensor and has to be communicated to a central estimator by using a communication budget of k bits. We consider the blackboard communication protocol Π_{BB} Kushilevitz and Nisan (1997): 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.

Under both models, the central sensor needs to produce an estimate $\hat{\theta}$ 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 the blackboard communication protocol Π_{BB} and the estimator $\hat{\theta}(\cdot)$ so as to minimize the worst case squared ℓ_2 risk, i.e., to characterize

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2.$$

Distributed parameter estimation and function computation has been considered in many recent works; we refer to Duchi et al. (2013); Zhang et al. (2013); Shamir (2014); Garg et al. (2014); Braverman et al. (2016); Xu and Raginsky (2017) and the references therein for an overview. Most of these works focus on the Gaussian location model and strong/distributed data processing inequalities appear as the key technical step in developing converse results. A more recent work Diakonikolas et al. (2017) studied the high-dimensional distribution estimation problem under the blackboard model without using strong data processing inequalities. However, a complete characterization of the minimax risk for this problem with general (n, d, k) is still missing.

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 and under both independent and interactive models, where d is the dimension of the parameter to be estimated. Moreover, as opposed to the linear dependence on k in prior works, our new approach enables us to show that the penalty is at most exponential in k, which turns out to be tight in high-dimensional distribution estimation.

3. Our new approach recovers the linear dependence on k when some sub-Gaussian structure is available, e.g., in the Gaussian location model. This result builds on a geometric inequality for the Gaussian measure, which may be of independent interest.

Notations: 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; standard notations from information theory are used: $I(X;Y)$ denotes the mutual information, and $D(P\|Q)$ denotes the Kullback–Leibler (KL) divergence between probability measures P and Q; $\mathsf{Multi}(n; P)$ denotes the multinomial model which observes n independent samples from P; 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 $b_n \lesssim a_n$.

2. Main Results

2.1. Assumptions

We first consider the distributed estimation problem of θ in a general statistical model $(P_\theta)_{\theta\in\Theta\subset\mathbb{R}^d}$. Choose an interior point $\theta_0 \in \Theta$, we consider the following regularity assumptions on $(P_\theta)_{\theta\in\Theta}$ for θ near θ_0:

Assumption 1 *The statistical model $(P_\theta)_{\theta\in\Theta}$ is differentiable in quadratic mean at $\theta = \theta_0$, with score function S_{θ_0} and non-singular Fisher information matrix $I(\theta_0)$.*

Assumption 2 *Let $\delta > 0$ and $\theta_0 + \delta[-1, 1]^d \subset \Theta$. There exist constants $\delta_0, c_0 > 0$ such that if $\delta < \delta_0(B^2 d)^{-\frac{1}{4}}$, then*

$$\mathbb{E}\left[\mathbb{E}_U\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(X) - 1\right)\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(X') - 1\right) - \delta^2 S_{\theta_0}(X)^T S_{\theta_0}(X')\right]^2 \le (c_0\delta^4 B^2 d)^2, \quad (1)$$

where $U \sim \mathsf{Unif}(\{\pm 1\}^d)$, random variables $X, X' \sim P_{\theta_0}$ are independent, and B is the maximum of all diagonal elements of $I(\theta_0)$.

Assumption 3 *Let $\delta > 0$ and $\theta_0 + \delta[-1, 1]^d \subset \Theta$. Let $\mathcal{X}$ be the sample space, and $\mathcal{X}_0 \subset \mathcal{X}$ satisfy $\inf_{\|\theta-\theta_0\|_\infty \le \delta} P_\theta(\mathcal{X}_0) \ge 1 - d^{-5}$. Define $(Q_\theta)_{\theta\in\Theta}$ to be the conditional probability measure obtained by restricting $(P_\theta)_{\theta\in\Theta}$ on $\mathcal{X}_0$, i.e., $Q_\theta(\cdot) = \frac{P_\theta(\cdot\cap\mathcal{X}_0)}{P_\theta(\mathcal{X}_0)}$. There exist constants $\delta_1, \delta_2, c_1, c_2 > 0$ such that if $\delta < \delta_1(B^2d^2 + B^3 d)^{-\frac{1}{4}}$, then*

$$\mathbb{E}_{Q_{\theta_0}}\left(\frac{dQ_{\theta_0+\delta u}}{dQ_{\theta_0}}(X) - 1\right)^4 \le c_1^2\left(B^2d^2 + B^3 d\right)\delta^4 \quad (2)$$

holds for any $u \in \{\pm 1\}^d$, and if $\delta < \delta_2(B^2 d\log d)^{-\frac{1}{4}}$, then

$$\begin{aligned}\mathbb{E}_U\left(\frac{dQ_{\theta_0+\delta U}}{dQ_{\theta_0}}(x) - 1\right)\left(\frac{dQ_{\theta_0+\delta U}}{dQ_{\theta_0}}(x') - 1\right) + 1 - \exp(\delta^2 S_{\theta_0}(x)^T S_{\theta_0}(x')) \\ \le c_2(\delta^4 B^2 d\log d + \sqrt{\delta^4 B^2 d\log d}\cdot\exp(\delta^2 S_{\theta_0}(x)^T S_{\theta_0}(x')))\end{aligned} \quad (3)$$

holds for any $x, x' \in \mathcal{X}_1$ with $Q_{\theta_0}(\mathcal{X}_1) \ge 1 - d^{-5}$, where U, B are the same as Assumption 2.

Assumption 1 is a standard regularity condition commonly used in asymptotic statistics Ibragimov and Has'minskii (2013). Assumptions 2 and 3 roughly correspond to the product measure case where $P_{\theta_0} = p_{\theta_1} \times p_{\theta_2} \times \cdots \times p_{\theta_d}$, and control the remainder term in different ways. The insights behind Assumptions 2 and 3 are that, based on local expansion of $(P_\theta)_{\theta\in\Theta}$ around $\theta \approx \theta_0$, for small δ we have

$$\mathbb{E}_U \left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x) - 1\right)\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x') - 1\right) \approx \exp(\delta^2 S_{\theta_0}(x)^T S_{\theta_0}(x')) - 1 \approx \delta^2 S_{\theta_0}(x)^T S_{\theta_0}(x').$$

Roughly speaking, Assumption 2 corresponds to an approximate product measure for general statistical models and 3 imposes additional sub-Gaussian structure. We can choose $\mathcal{X}_0 = \mathcal{X}_1 = \mathcal{X}$ in Assumption 3 under some models, while sometimes we use $\mathcal{X}_0$ to deal with the unbounded support of (P_θ) and avoid the assumptions of bounded likelihood ratios, which were assumed in some previous works Braverman et al. (2016). The next proposition shows that, these assumptions hold for many commonly used statistical models.

Proposition 1 *Assumptions 1 and 2 hold for the Gaussian location model $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ with any $\theta_0 \in \mathbb{R}^d$, the product Bernoulli model $P_\theta = \prod_{i=1}^d \mathsf{Bern}(\theta_i)$ with $\theta_0 = (p, p, \cdots, p)$ for $p \in (0, 1)$, and the Multinomial model $P_\theta = \mathsf{Multi}(1; \theta)$ for any probability measure θ over $d + 1$ elements. In particular, for the Gaussian location model and the product Bernoulli model above, Assumption 3 also holds.*

2.2. Main Theorems

We present the following main theorem for the distributed inference of θ in general statistical models $(P_\theta)_{\theta\in\Theta\subset\mathbb{R}^d}$:

Theorem 2 (General lower bound) *Let $P_\theta = \prod_{i=1}^d p_{\theta_i}$, and Assumptions 1 and 2 be fulfilled for $(P_\theta)_{\theta\in\Theta}$ at θ_0, with $P_{\theta_0} = p_{\theta_0}^{\otimes d}$. Let $s_0(x), I_0$ be the score function and Fisher information of (p_θ) at $\theta = \theta_0$, respectively. Then for any $k \in \mathbb{N}, n \geq \frac{d^2}{2^k \wedge d}$, we have*

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \geq C \cdot \frac{d^2}{n(2^k \wedge d)\mathsf{Var}_{p_{\theta_0}}(s_0(X))} = C \cdot \frac{d^2}{n(2^k \wedge d) I_0}$$

where the infimum is taken over all possible estimators $\hat{\theta} = \hat{\theta}(Y^n)$ and blackboard protocols with k-bit communication constraint, and the constant $C > 0$ is independent of n, d, k, I_0.

We compare Theorem 2 with the centralized case. When there is no communication constraints, classical Hájek–Le Cam theory Hájek (1972) tells that we can achieve a squared ℓ_2 risk $1/(nI_0)$ asymptotically for each coordinate, which sums up to $d/(nI_0)$ for the entire vector. Compared with Theorem 2, we see an effective sample size reduction from n to $n/(2^{-k}d \vee 1)$ if each sensor can only transmit k bits. The following corollary is immediate for $k = 1$.

Corollary 3 (General lower bound for $k = 1$) *When each sensor can only transmit one bit (i.e., $k = 1$), under the conditions of Theorem 2, for $n \geq d^2$ we have*

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \geq C \cdot \frac{d^2}{nI_0}.$$

Corollary 3 shows that when $k = 1$, we have an effective sample size reduction from n to n/d. This bound can also possibly be achieved by a simple grouping idea: the sensors are splitted into d groups, and all n/d sensors in one group contribute to estimating only one coordinate of θ. Hence, we expect that the dependence on n, d of our lower bound to be tight for $k = 1$.

When $k > 1$, Theorem 2 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. The next theorem shows that, when the score function $s_0(X)$ has a sub-Gaussian tail, the above penalty will reduce at most linearly in k. Recall that the ψ_2-norm of a random variable X is defined by

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

which is the Orlicz norm of X associated with the Orlicz function $\psi_2(x) = \exp(x^2) - 1$ Birnbaum and Orlicz (1931). There are also some equivalent definitions of ψ_2-norm, and $\|X\|_{\psi_2} \le \sigma$ if and only if X is sub-Gaussian with parameter $C_0\sigma$ for some absolute constant $C_0 > 0$ Vershynin (2010).

Theorem 4 (Lower bound with sub-Gaussian structure) *Let Assumptions 1 and 3 be fulfilled for $(P_\theta)_{\theta\in\Theta}$ at θ_0, with $P_\theta = \prod_{i=1}^d p_{\theta_i}, Q_\theta = \prod_{i=1}^d q_{\theta_i}, Q_{\theta_0} = q_{\theta_0}^{\otimes d}$. Let $s_0(x)$ be the score function of (p_θ) at $\theta = \theta_0$, $R \triangleq \sup_{x,x'\in\mathcal{X}_0} \max_{i\in[d]} |s_0(x_i) - s_0(x_i')|$ be the diameter of $\mathcal{X}_0$ in Assumption 3 in terms of the ℓ_∞ norm, and $\sigma^2 \triangleq \|s_0(X)\|^2_{\psi_2(q_{\theta_0})} \le d$ be the sub-Gaussian parameter of the score function under q_{θ_0}. Then for any $k \ge (R/\sigma)^2 \vee \log d$ and $n \ge \frac{d^2}{k\wedge d}$, we have*

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat\theta} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat\theta - \theta\|_2^2 \ge C \cdot \frac{d^2}{n(k\wedge d)\sigma^2},$$

where the constant $C > 0$ is independent of n, d, k, σ.

Theorem 4 improves over Theorem 2 in scenarios where $s_0(X)$ not only admits a finite variance but also behaves like a Gaussian random variable. We remark that the different dependence on k in Theorems 2 and 4 is due to the nature of different geometric inequalities (cf. Lemma 13 and Lemma 14) satisfied by general probability distributions and a sub-Gaussian distribution. Since typically $(R/\sigma)^2 \lesssim \log d$, compared with the phase transition threshold $k \asymp d$, the condition $k \ge (R/\sigma)^2$ is mild; we believe this condition can be removed using better technical arguments.

2.3. Applications

Next we apply Theorems 2 and 4 to some concrete examples.

Corollary 5 (Distribution estimation) *Let $P_\theta = \mathsf{Multi}(1;\theta)$ with $\Theta = \mathcal{M}_d$ being the probability simplex over d elements. For $k \in \mathbb{N}$ and $n \ge \frac{d^2}{2^k\wedge d}$, we have*

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat\theta} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat\theta - \theta\|_2^2 \ge C \cdot \left(\frac{d}{n2^k} \vee \frac{1}{n}\right)$$

where $C > 0$ is a universal constant independent of n, k, d.

Corollary 6 (Gaussian location model) *Let $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ with $\Theta = \mathbb{R}^d$. Under any blackboard communication protocol, for $k \geq \log d$ and $n \geq \frac{d^2}{k \wedge d}$, we have*

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \geq C \cdot \left(\frac{d^2}{nk} \vee \frac{d}{n} \right) \sigma^2$$

where $C > 0$ is a universal constant independent of n, k, d, σ^2.

Corollarys 5 and 6 follow from Theorems 2 and 4, respectively. Corollary 5 completely characterizes the minimax risk for distribution estimation under general (n, k, d) Han et al. (2018), which improves over Diakonikolas et al. (2017). Corollary 6 recover the results in Zhang et al. (2013); Garg et al. (2014) (without logarithmic factors in the risk) under a mild technical condition $k \geq \log d$. Note that these two models have different tight dependence on k: in Corollary 5, when $2^k < d$, we see an effective sample size reduction from n to $n2^k/d$; in Corollary 6, 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:

Proposition 7 (Product Bernoulli model) *Let $P_\theta = \prod_{i=1}^d \mathsf{Bern}(\theta_i)$. If $\Theta = [0,1]^d$ and $n \geq \frac{d^2}{d \wedge k}$, we have*

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

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

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

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: if $\theta_0 = \frac{1}{2}$, we have $\mathsf{Var}(s_0(X)) \asymp \|s_0(X)\|_{\psi_2}^2 = \Theta(1)$; if $\theta_0 = d^{-1}$, then $\mathsf{Var}(s_0(X)) \asymp d \ll d^2 \asymp \|s_0(X)\|_{\psi_2}^2$. Hence, Theorem 4 utilizes the sub-Gaussian nature and gives a better lower bound in the first case, and Theorem 2 becomes better in the second case where the tail of the score function is essentially not sub-Gaussian.

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

Theorem 8 (Sparse Gaussian location model) *Let $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ with $\Theta = \{\theta \in \mathbb{R}^d : \|\theta_0\| \leq s \leq \frac{d}{2}\}$. For $k \geq \log d$ and $n \geq \frac{sd \log(d/s)}{k \wedge d}$, we have*

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

where $C > 0$ is a universal constant independent of n, d, s, k, σ^2.

Theorem 8 improves over Braverman et al. (2016) under a slightly different framework, with tight logarithmic factors matching the upper bound in Garg et al. (2014). We see from Theorem 8 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.

The rest of the paper is organized as follows. In Section 3 we introduce the tree representation of the blackboard communication protocol, and sketch the lower bound proof based on the previous representation. Section 4 is devoted to the proof of Theorems 2 and 4, 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.

3. Representations of Blackboard Communication Protocol

The centralized lower bounds without communication constraints simply follows from the classical asymptotics Hájek (1970, 1972), 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 Kushilevitz and Nisan (1997), 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 Bar-Yossef et al. (2004) for the blackboard communication protocol:

Lemma 9 *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 10 *For each $i \in [n]$ and $\{x_j\}_{j=1}^n \in \mathcal{X}^n$, the following equalities hold: $\sum_{y\in\{0,1\}^{nk}} \prod_{j=1}^{n} p_{j,y}(x_j) = 1$ and $\sum_{y\in\{0,1\}^{nk}} \prod_{j\neq i} p_{j,y}(x_j) = 2^k$.*

3.2. Minimax lower bound

This subsection is devoted to setting up the proof of the minimax lower bound in Theorems 2 and 4. We apply the standard testing argument: first, we construct a class of hypotheses and relate the minimax risk to some mutual information via a distance-based Fano's inequality; second, we derive a universal upper bound for the mutual information which holds for any blackboard communication protocol $\{a_v(\cdot)\}$.

Let $U \sim \mathsf{Unif}(\{\pm 1\}^d)$. For each $u \in \{\pm 1\}^d$, we associate with a product probability measure P_u given by $P_u \triangleq p_{\theta_0+\delta u_1} \times p_{\theta_0+\delta u_2} \times \cdots \times p_{\theta_0+\delta u_d}$, where $\delta > 0$ is some parameter to be specified later. We also denote by P_0 the product distribution $P_{\theta_0} = p_{\theta_0}^{\otimes d}$ for brevity. We will assume that

$$0 < \delta < \min\{\delta_0, \delta_1\} \cdot \frac{1}{\sqrt{dI_0}} \tag{4}$$

throughout the proof (with δ_0, δ_1 appearing in Assumptions 2 and 3), and will get back to it when we specify δ in the end.

Now the setting is as follows: the observations $X_1, \cdots, X_n$ are drawn from P_U, then sensors output the transcript $Y \in \{0,1\}^{nk}$ according to the blackboard communication protocol, and finally an estimator $\hat{\theta}(Y)$ is used to estimate θ. By a standard testing argument Tsybakov (2008), we have

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat{\theta}(Y) - \theta\|_2^2 \geq \frac{d\delta^2}{10} \inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{U}} \mathbb{P}\left(d_{\mathsf{Ham}}(\hat{U}, U) \geq \frac{d}{5}\right)$$

where $d_{\mathsf{Ham}}(x,y) \triangleq \sum_{i=1}^{d} \mathbb{1}(x_i \neq y_i)$ denotes the Hamming distance. To lower bound $\mathbb{P}(d_{\mathsf{Ham}}(\hat{U}, U) \geq \frac{d}{5})$ for any estimator $\hat{U}$ under any blackboard communication protocol, we use the following distance-based Fano's inequality:

Lemma 11 *(Duchi and Wainwright, 2013, Corollary 1) Let random variables V and $\hat{V}$ take value in $\mathcal{V}$, V be uniform on some finite alphabet $\mathcal{V}$, and $V - X - \hat{V}$ form a Markov chain. Let d be any metric on $\mathcal{V}$, and for $t > 0$, define $N_{\max}(t) \triangleq \max_{v\in\mathcal{V}} |v' \in V : d(v,v') \leq t|$, $N_{\min}(t) \triangleq \min_{v\in\mathcal{V}} |v' \in V : d(v,v') \leq t|$. If $N_{\max}(t) + N_{\min}(t) < |\mathcal{V}|$, the following inequality holds:*

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

Applying Lemma 11 to the Markov chain $U - Y - \hat{U}$ with Hamming distance $d_{\mathsf{Ham}}(\cdot,\cdot)$ and $t = \frac{d}{5}$, we have

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat{\theta}(Y) - \theta\|_2^2 \geq \frac{d\delta^2}{10}\left(1 - \frac{I(U;Y) + \log 2}{d/8}\right) \tag{5}$$

where Chernoff bound (cf. Lemma 17) implies $\frac{N_{\max}(t)}{|\mathcal{V}|} \le \exp(-\frac{d}{8})$. Hence, to establish the minimax lower bound, it suffices to upper bound the mutual information $I(U;Y)$, where Lemma 9 plays a central role in characterizing the distribution of Y given any $U = u$. Specifically,

$$I(U;Y) \overset{\text{(a)}}{\le} \mathbb{E}_U D(P_{Y|U} \| P_{Y|X\sim P_0}) \overset{\text{(b)}}{=} \mathbb{E}_U \mathbb{E}_{Y|U} \sum_{i=1}^n \log \frac{\mathbb{E}_{P_U} p_{i,Y}(X_i)}{\mathbb{E}_{P_0} p_{i,Y}(X_i)} \overset{\text{(c)}}{\le} \mathbb{E}_U \mathbb{E}_{Y|U} \sum_{i=1}^n \left(\frac{\mathbb{E}_{P_U} p_{i,Y}(X_i)}{\mathbb{E}_{P_0} p_{i,Y}(X_i)} - 1 \right)$$
$$\overset{\text{(d)}}{=} \mathbb{E}_U \sum_{i=1}^n \sum_{y\in\{0,1\}^{nk}} \left(\prod_{j\neq i} \mathbb{E}_{P_U} p_{j,y}(X_j) \right) \cdot \frac{(\mathbb{E}_{P_U} p_{i,y}(X_i) - \mathbb{E}_{P_0} p_{i,y}(X_i))^2}{\mathbb{E}_{P_0} p_{i,y}(X_i)}, \tag{6}$$

where (a) follows from the variational representation of mutual information $I(X;Y) = \inf_{Q_Y} \mathbb{E}_X D(P_{Y|X} \| Q_Y)$, inequality (b) follows from Lemma 9, (c) is due to $\log x \le x - 1$, and (d) follows from Lemma 9 and the first equality of Lemma 10.

Before we further upper bound $I(U;Y)$, we make some remarks. First, we show that it suffices to consider deterministic protocols: this is due to the joint convexity of the KL divergence $D(P\|Q) \le \mathbb{E}_R D(P_{\cdot|R} \| Q_{\cdot|R})$, and thus we can always condition on the randomness and prove an upper bound on the mutual information in the deterministic case. Second, if Assumption 3 holds, we may apply the previous analysis to (Q_θ) instead of (P_θ). In fact, note that the total variation distance between Q_θ and P_θ is $\mathbb{P}_\theta(\mathcal{X}_0^c) \le d^{-5}$ for $\|\theta - \theta_0\|_\infty \le \delta$, applying the testing arguments to Q_θ will only affect the test error by an additive factor of nd^{-5}, which is negligible compared to the $\Omega(1)$ test error we aim to obtain. With a slight abuse of notation we still write Q_θ as P_θ in the sequel for notational simplicity.

The main ingredient to upper bound $I(U;Y)$ is summarized in the following lemma:

Lemma 12 *Fix any $i \in [n]$ and $\{x_j\}_{j\neq i} \in \mathcal{X}^{n-1}$, and define $w_{i,y} \triangleq \prod_{j\neq i} p_{j,y}(x_j)$. Let $S_0(X) \triangleq (s_0(X_1), \cdots, s_0(X_d))$ be the d-dimensional score function, and I_0 be the Fisher information of the 1D model $(p_\theta)_{\theta\in\Theta}$ at $\theta = \theta_0$. The following inequalities hold:*

1. *Under Assumptions 1 and 2, we have*

$$\sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \mathbb{E}_U \frac{(\mathbb{E}_{P_U} p_{i,y}(X_i) - \mathbb{E}_{P_0} p_{i,y}(X_i))^2}{\mathbb{E}_{P_0} p_{i,y}(X_i)} \le S_1 + c_0 I_0^2 \cdot 2^k \delta^4 d$$

where c_0 is the constant appearing in Assumption 2, and

$$S_1 \triangleq \delta^2 \sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \frac{\|\mathbb{E}_{P_0} S_0(X) p_{i,y}(X)\|_2^2}{\mathbb{E}_{P_0} p_{i,y}(X)}.$$

2. *Under Assumptions 1 and 3, if $I_0 \le d$, we have*

$$\sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \mathbb{E}_U \frac{(\mathbb{E}_{P_U} p_{i,y}(X_i) - \mathbb{E}_{P_0} p_{i,y}(X_i))^2}{\mathbb{E}_{P_0} p_{i,y}(X_i)}$$
$$\le S_2 + 3(2c_1 I_0 \cdot \delta^2 + d^{-1}) + c_2(2I_0^2 \delta^4 d \log d + S_2 \cdot \sqrt{I_0^2 \delta^4 d \log d})$$

where c_1, c_2 are the constants appearing in Assumption 3, and

$$S_2 \triangleq \sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \frac{\mathbb{E}_{P_0}[p_{i,y}(X)p_{i,y}(X')(\exp(\delta^2 S_0(X)^T S_0(X')) - 1)]}{\mathbb{E}_{P_0} p_{i,y}(X)}$$

with X' an independent copy of X.

The next section will upper bound the leading terms S_1, S_2 via geometric inequalities.

4. Lower Bounds via Geometric Inequalities

In this section, we upper bound S_1, S_2 in Lemma 12 using two different geometric inequalities, and complete the proof of main Theorems 2 and 4.

4.1. Proof of Theorem 2 via Geometric Inequality I

Note that under a deterministic protocol, each summand of S_1 has the following geometric interpretation: since $p_{i,y}(X) = \mathbb{1}(A_y)$ must be an indicator function, then we may write S_1 as

$$S_1 = \delta^2 \sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \mathbb{P}(A_y)\|\mathbb{E}[S_0(X)|A_y]\|_2^2$$

where $\|\mathbb{E}[S_0(X)|A_y]\|_2$ is the ℓ_2 norm of the mean score function vector $S_0(X)$ conditioning on the set A_y. Hence, we ask the following question:

Question 1 *Given $P_0(A) = t \in (0,1)$, which set A maximizes the ℓ_2 norm of the vector $\mathbb{E}_{P_0}[S_0(X)|A]$? What is the corresponding maximum ℓ_2 norm?*

The following lemma presents an answer to Question 1:

Lemma 13 (Geometric Inequality I) *For any set $A \subset \mathcal{X}$, the following inequality holds:*

$$\|\mathbb{E}_{P_0}[S_0(X)|A]\|_2^2 \le I_0 \cdot \frac{1 - P_0(A)}{P_0(A)}.$$

Note that Lemma 13 is a dimension-free result: the LHS depends on the dimensionality d, while the RHS does not. For a comparison, if we directly apply Cauchy–Schwartz inequality to the LHS, we will lose a multiplicative factor of d. The key observation in the dimensionality reduction is that the independence between coordinates of $S_0(X)$ needs to be exploited.

Now we have all necessary tools for the proof of Theorem 2. By Lemma 13,

$$S_1 \le \delta^2 \sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot I_0 = 2^k \cdot \delta^2 I_0 \tag{7}$$

where the last identity is due to Lemma 10. Combining (5), (6), (7) and Lemma 12, we have

$$\inf_{\hat\theta} \sup_{\theta\in\Theta} \mathbb{E}_\theta\|\hat\theta - \theta\|_2^2 \ge \frac{d\delta^2}{10}\left(1 - \frac{n2^k I_0(\delta^2 + c_0 I_0 \cdot \delta^4 d) + \log 2}{d/8}\right).$$

Now choosing $\delta^2 = c\frac{d}{n2^k I_0}$, the condition $n2^k \ge d^2$ ensures that $\delta^2 \le \frac{c}{dI_0}$. Hence, by choosing $c > 0$ sufficiently small, the condition (4) is satisfied, and thus

$$\inf_{\Pi_{\text{BB}}} \inf_{\hat\theta} \sup_{\theta\in\Theta} \mathbb{E}_\theta\|\hat\theta - \theta\|_2^2 \gtrsim \frac{d^2}{n2^k \cdot I_0}.$$

4.2. Proof of Theorem 4 via Geometric Inequality II

To upper bound S_2, we first note that when δ is small, S_2 coincides with S_1 up to first-order Taylor expansion. Hence, we may ask the following similar question:

Question 2 *Suppose $\|s_0(X)\|_{\psi_2} \le \sigma$. Given $P_0(A) = t \in (0,1)$, which set $A \subset \mathbb{R}^d$ maximizes the ℓ_2 norm of the conditional mean vector $\mathbb{E}_{P_0}[S_0(X)|A]$ in A? What is the maximum ℓ_2 norm?*

An upper bound on the ℓ_2 norm is given in the following lemma.

Lemma 14 (Geometric Inequality II) *Assume that $\|s_0(X)\|_{\psi_2} \le \sigma$. Then for any $A \subset \mathcal{X}$,*

$$\|\mathbb{E}_{P_0}[S_0(X)|A]\|_2^2 \le \sigma^2 \cdot \log \frac{2}{P_0(A)}.$$

Note that lemma 14 presents a dimension-free upper bound again. Compared with Lemma 13, for sub-Gaussian score function $S_0(X)$, Lemma 14 improves the upper bound from $O(\sigma^2)$ to $O(\sigma^2 t \log \frac{1}{t})$, where $t = P_0(A)$ is the "volume" of the set A. We provide two proofs of Lemma 14 in the appendix. The first proof first reduces the problem to 1D and then makes use of the sub-Gaussian tail. The second proof is more geometric when $S_0(X)$ is exactly Gaussian: information-theoretic inequalities can be used to obtain a tight inequality for $X \sim \mathsf{Unif}(\{\pm 1\}^d)$, and then the "tensor power trick" is applied to prove the Gaussian case.

Although Lemma 14 only upper bounds the first-order Taylor expansion of S_2 when δ is small, it serves as the key step in establishing the upper bound of S_2:

Lemma 15 *Assume that $|s_0(X_i)| \le R$ almost surely for any $i \in [n]$ under p_{θ_0}, and $\delta^2 d R^2 \le 1$. Then if $\|s_0(X_i)\|_{\psi_2} \le \sigma$, there exists some constant $C > 0$ independent of δ, d, R, k, σ such that*

$$S_2 \le C\delta^2 \left(k\sigma^2 + R^2\right).$$

Now we prove Theorem 4. Combining (5), (6), Lemma 12 and Lemma 15, we have

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \ge \frac{d\delta^2}{10} \left(1 - \frac{n(C\delta^2(k\sigma^2 + R^2) + 3(2c_1 I_0 \delta^2 + d^{-1}) + c_2 C I_0^2 \delta^4 d \log d) + \log 2}{d/8}\right)$$

and choosing $\delta^2 \asymp \frac{d}{nk\sigma^2}$ completes the proof (note that $k \ge (R/\sigma)^2 \vee \log d$, and $I_0 \le \sigma^2 \le d$).

5. Discussions

5.1. Some Applications of Geometric Inequalities

The inequalities in Lemmas 13 and 14 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 $A \subset \Omega$, $|A| = 2^{d-1}$, 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 Amari (2011). We prove the following proposition:

Proposition 16 *Under the previous setting, we have*

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

where $h_2(\cdot)$ is the binary entropy function defined in Lemma 18.

Proposition 16 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)

$$I(U;Y)\le\gamma^*(U,X)I(X;Y),\qquad\forall p_{Y|X}\tag{8}$$

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., Zhang et al. (2013); Braverman et al. (2016); Xu and Raginsky (2017). The SDPI-based approach turns out to be tight in certain models (e.g., the Gaussian model Zhang et al. (2013); Braverman et al. (2016)), 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 (8) 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 5 the optimal dependence on k is exponential;
3. The conditional distribution $p_{Y^*|X}$ achieving the equality in (8) typically leads to $I(X;Y^*)\to 0$, and (8) may be loose for $I(X;Y)=k$;
4. The operational meaning of (8) 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.

References

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Distributed simulation and distributed inference. *arXiv preprint arXiv:1804.06952*, 2018.

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

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.

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.

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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

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.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions. In *Information Theory, 2018. ISIT 2018. IEEE International Symposium on*. IEEE, 2018.

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

E Kushilevitz and N Nisan. Communication complexity. cambridge university press, 1997.

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

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

Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 1987.

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.

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

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

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

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.

Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.

Appendix A. Auxiliary Lemmas

Lemma 17 *Mitzenmacher and Upfal (2005) For $X \sim \mathsf{Poi}(\lambda)$ or $X \sim \mathsf{B}(n, \frac{\lambda}{n})$ and any $\delta > 0$, we have*

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

Lemma 18 *Wyner (1973) 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 $y \in [0,1]$.

Appendix B. Proof of Main Lemmas

B.1. Proof of Lemma 10

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

$$\sum_{y \in \{0,1\}^{\sum_{i=1}^n k_i}} \prod_{v \in \tau(y), l_v \neq i} b_{v,y}(x_{l_v}) = 2^{k_i} \tag{9}$$

for any $\{x_j\}_{j \neq i}$. Clearly (9) implies the lemma (i.e., with $k_i = 0$ and $k_i = k$, respectively).

We prove (9) 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 (9) follows from $2^{k_i} = 2^{k_i - 1} + 2^{k_i - 1}$;
2. If the root node is not labeled as i, then (9) 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 12

We first assume that Assumptions 1 and 2 hold. By Fubini's theorem,

$$\begin{aligned}
\mathbb{E}_U(\mathbb{E}_{P_U} p_{i,y}(X) - \mathbb{E}_{P_0} p_{i,y}(X))^2 &= \mathbb{E}_U \left(\mathbb{E}_{P_0}\left(\frac{dP_U}{dP_0}(X) - 1\right) p_{i,y}(X)\right)^2 \\
&= \mathbb{E}_U \mathbb{E}_{P_0}\left(\frac{dP_U}{dP_0}(X) - 1\right)\left(\frac{dP_U}{dP_0}(X') - 1\right) p_{i,y}(X) p_{i,y}(X') \\
&= \mathbb{E}_{P_0}\left[p_{i,y}(X) p_{i,y}(X') \mathbb{E}_U \left(\frac{dP_U}{dP_0}(X) - 1\right)\left(\frac{dP_U}{dP_0}(X') - 1\right)\right]
\end{aligned}$$

where X' is an independent copy of X. By (1), we write

$$\mathbb{E}_U\left(\frac{dP_U}{dP_0}(X)-1\right)\left(\frac{dP_U}{dP_0}(X')-1\right)=\delta^2 S_0(X)^T S_0(X')+r_1(X,X')$$

with $\mathbb{E}[r_1(X,X')^2]^{\frac{1}{2}} \le c_0 I_0 \cdot \delta^4 d$. Note that

$$\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot\delta^2 S_0(X)^T S_0(X')\right]=\delta^2\|\mathbb{E}_{P_0}S_0(X)p_{i,y}(X)\|_2^2 \tag{10}$$

and by Cauchy–Schwartz,

$$\begin{aligned}\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot r_1(X,X')\right] &\le \sqrt{\mathbb{E}_{P_0}[p_{i,y}(X)^2 p_{i,y}(X')^2]}\cdot\sqrt{\mathbb{E}_{P_0}r_1(X,X')^2}\\ &\le \mathbb{E}_{P_0}[p_{i,y}(X)^2]\cdot c_0 I_0^2\delta^4 d.\end{aligned}$$

Hence, the sum of the remainder terms can be upper bounded as

$$\begin{aligned}\sum_{y\in\{0,1\}^{nk}} w_{i,y}\cdot\frac{\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot r_1(X,X')\right]}{\mathbb{E}_{P_0}[p_{i,y}(X)]} &\le c_0 I_0^2\delta^4 d\sum_{y\in\{0,1\}^{nk}} w_{i,y}\cdot\frac{\mathbb{E}_{P_0}[p_{i,y}(X)^2]}{\mathbb{E}_{P_0}[p_{i,y}(X)]}\\ &\le c_0 I_0^2\delta^4 d\sum_{y\in\{0,1\}^{nk}} w_{i,y}\\ &= c_0 I_0^2\cdot 2^k\delta^4 d\end{aligned} \tag{11}$$

where we have used $p_{i,y}(\cdot)\in[0,1]$ and the identity $\sum_y w_{i,y}=2^k$ in Lemma 10. Combining (10) and (11) completes the proof of the first inequality of Lemma 12.

Next we assume that Assumptions 1 and 3 hold. By (3), we write

$$\mathbb{E}_U\left(\frac{dP_U}{dP_0}(X)-1\right)\left(\frac{dP_U}{dP_0}(X')-1\right)=\exp(\delta^2 S_0(X)^T S_0(X'))-1+r_2(X,X')$$

where $|r_2(X,X')|\le c_2 I_0\cdot(1+\exp(\delta^2 S_0(X)^T S_0(X')))\cdot\delta^4 d\log d$ almost surely conditioning on $X,X'\in\mathcal{X}_1$. Define $Z\triangleq\mathbb{1}(X,X'\in\mathcal{X}_1)$, we split the remainder term into two parts:

$$\begin{aligned}&\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot r_2(X,X')\right]\\ &=\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot r_2(X,X')Z\right]+\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot r_2(X,X')(1-Z)\right]\\ &\triangleq A_{1,y}+A_{2,y}.\end{aligned}$$

For the first term $A_{1,y}$, since $r_2(X,X')Z$ is upper bounded and $p_{i,y}(\cdot)\ge 0$, we have

$$\begin{aligned}A_{1,y}&\le c_2\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot(I_0^2\delta^4 d\log d+\sqrt{I_0^2\delta^4 d\log d}\cdot\exp(\delta^2 S_0(X)^T S_0(X')))\right]\\ &=c_2\left(I_0^2\delta^4 d\log d\cdot(\mathbb{E}_{P_0}p_{i,y}(X))^2+\sqrt{I_0^2\delta^4 d\log d}\cdot\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot\exp(\delta^2 S_0(X)^T S_0(X'))\right]\right)\end{aligned}$$

and thus the sum can be upper bounded as

$$\begin{aligned}&\sum_{y\in\{0,1\}^{nk}} w_{i,y}\cdot\frac{A_{1,y}}{\mathbb{E}_{P_0}[p_{i,y}(X)]}\\ &\le c_2\left(I_0^2\delta^4 d\log d+\sqrt{I_0^2\delta^4 d\log d}\sum_{y\in\{0,1\}^{nk}} w_{i,y}\cdot\frac{\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X')\cdot\exp(\delta^2 S_0(X)^T S_0(X'))\right]}{\mathbb{E}_{P_0}[p_{i,y}(X)]}\right)\end{aligned} \tag{12}$$

where we have used the identity $\sum_y w_{i,y}\mathbb{E}_{P_0}p_{i,y}(X) = 1$ in Lemma 10.

As for the second term $A_{2,y}$, note that

$$r_2(X, X') \le r_3(X, X') \triangleq \mathbb{E}_U \left(\frac{dP_U}{dP_0}(X) - 1\right)\left(\frac{dP_U}{dP_0}(X') - 1\right) + 1.$$

We further write the indicator function

$$1 - Z = \mathbb{1}(X \in \mathcal{X}_1, X' \notin \mathcal{X}_1) + \mathbb{1}(X \notin \mathcal{X}_1, X' \in \mathcal{X}_1) + \mathbb{1}(X \notin \mathcal{X}_1, X' \notin \mathcal{X}_1)$$

as the sum of three indicators functions on rectangles. For the first rectangle, by Fubini's theorem we have

$$\begin{aligned}
&\mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X') \cdot r_3(X, X')\mathbb{1}(X \in \mathcal{X}_1, X' \notin \mathcal{X}_1)\right] \\
&= \left(\mathbb{E}_{P_0}\left[p_{i,y}(X)\left(\frac{dP_U}{dP_0}(X) - 1\right)\mathbb{1}(X \in \mathcal{X}_1)\right]\right)\left(\mathbb{E}_{P_0}\left[p_{i,y}(X)\left(\frac{dP_U}{dP_0}(X) - 1\right)\mathbb{1}(X \notin \mathcal{X}_1)\right]\right) \\
&\quad + \left(\mathbb{E}_{P_0}\left[p_{i,y}(X)\mathbb{1}(X \in \mathcal{X}_1)\right]\right)\left(\mathbb{E}_{P_0}\left[p_{i,y}(X)\mathbb{1}(X \notin \mathcal{X}_1)\right]\right). \qquad (13)
\end{aligned}$$

To deal with the above terms, we define

$$\begin{aligned}
f(X) &\triangleq \left(\frac{dP_U}{dP_0}(X) - 1\right)\mathbb{1}(X \in \mathcal{X}_1) \\
g(X) &\triangleq \left(\frac{dP_U}{dP_0}(X) - 1\right)\mathbb{1}(X \notin \mathcal{X}_1) \\
e_y(X) &\triangleq \sqrt{\frac{w_{i,y}}{\mathbb{E}_{P_0}[p_{i,y}(X)]}} \cdot p_{i,y}(X), \qquad y \in \{0,1\}^{nk}.
\end{aligned}$$

Consider the inner product $\langle u, v\rangle \triangleq \mathbb{E}_{P_0}[u(X)v(X)]$ for $u, v \in L^2(P_0)$, the sum of the first term of (13) can be written as $\sum_y \langle f, e_y\rangle\langle g, e_y\rangle$. Since we are considering a deterministic protocol, we have $w_{i,y} \in \{0,1\}, p_{i,y}(X) \in \{0,1\}$. As a result, $\{e_y(\cdot)\}$ are orthogonal to each other, and $\|e_y\| \le 1$. Hence,

$$\left|\sum_y \langle f, e_y\rangle\langle g, e_y\rangle\right| = \left|\left\langle f, \sum_y e_y\langle g, e_y\rangle\right\rangle\right| \le \|f\| \cdot \left\|\sum_y e_y\langle g, e_y\rangle\right\| \le \|f\| \cdot \|g\|$$

where the first inequality is due to Cauchy–Schwartz, and the second inequality is Bessel's inequality due to orthogonality. By inequality (2) in Assumption 3, we further have

$$\begin{aligned}
\|f\| \cdot \|g\| &\le \sqrt{\mathbb{E}_{P_0}\left(\frac{dP_U}{dP_0}(X) - 1\right)^2} \cdot \sqrt{\mathbb{E}_{P_0}\left(\frac{dP_U}{dP_0}(X) - 1\right)^2 \mathbb{1}(X \notin \mathcal{X}_1)} \\
&\le \sqrt{\mathbb{E}_{P_0}\left(\frac{dP_U}{dP_0}(X) - 1\right)^4} \cdot [\mathbb{E}_{P_0}\mathbb{1}(X \notin \mathcal{X}_1)]^{\frac{1}{4}} \\
&\le c_1(I_0 \cdot d\delta^2 + I_0^{\frac{3}{2}} \cdot \sqrt{d}\delta^2) \cdot d^{-\frac{5}{4}} \\
&\le 2c_1 I_0 \cdot \delta^2
\end{aligned}$$

when $I_0 \leq d$. Using similar arguments to deal with the second term of (13), we arrive at

$$\sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \mathbb{E}_{P_0}\left[p_{i,y}(X)p_{i,y}(X') \cdot r_3(X,X')\mathbb{1}(X \in \mathcal{X}_1, X' \notin \mathcal{X}_1)\right] \leq 2c_1 I_0 \cdot \delta^2 + d^{-1}. \quad (14)$$

We handle the other two rectangles analogously, and the proof of Lemma 12 is completed using (12), (14).

B.3. Proof of Lemma 13

Consider the Hilbert space $\mathcal{H}$ consisting of all squared integrable random variables X under P_0, with inner product $\langle X, Y\rangle_{\mathcal{H}} \triangleq \mathbb{E}_{P_0}[XY]$. Since $(s_0(X_1), \cdots, s_0(X_d))$ is an i.i.d random vector with

$$\mathbb{E}_{P_0}[s_0(X_1)] = 0, \qquad \mathbb{E}_{P_0}[s_0(X_1)^2] = I_0$$

we conclude that the constant 1 and $I_0^{-1/2}(s_0(X_1), \cdots, s_0(X_d))$ constitute an orthonormal system in $\mathcal{H}$. Now for the element $\mathbb{1}_A(X) \in \mathcal{H}$, Bessel's inequality Rudin (1987) gives

$$\|\mathbb{1}_A\|_{\mathcal{H}}^2 \geq \langle \mathbb{1}_A(X), 1\rangle_{\mathcal{H}}^2 + \sum_{i=1}^{d} \langle \mathbb{1}_A(X), I_0^{-\frac{1}{2}} s_0(X_i)\rangle_{\mathcal{H}}^2.$$

A rearrangement of this inequality gives the desired result.

B.4. Proof of Lemma 14

Using $\|u\|_2 = \sup_{v:\|v\|_2=1}\langle u, v\rangle$, it suffices to prove that $\mathbb{E}_{P_0}[\langle X, v\rangle|A]^2 \leq C_1 \log \frac{C_2}{P_0(A)}$ for any unit vector v. Note that the random vector $S_0(X)$ consists of i.i.d sub-Gaussian components, the inner product $\langle S_0(X), v\rangle$ is also sub-Gaussian with

$$\|\langle S_0(X), v\rangle\|_{\psi_2}^2 \leq \sum_{i=1}^{d} v_i^2 \|s_0(X_i)\|_{\psi_2}^2 \leq \sigma^2.$$

Hence, we may always reduce to the 1D case and assume that $S_0(X)$ is sub-Gaussian with $\|S_0(X)\|_{\psi_2} \leq \sigma$. Now by the convexity of $x \mapsto \exp(\frac{x^2}{\sigma^2})$,

$$2 \geq \mathbb{E}_{P_0}[\exp(\frac{[S_0(X)]^2}{\sigma^2})] \geq \mathbb{E}_{P_0}[\exp(\frac{[S_0(X)]^2}{\sigma^2})\mathbb{1}_A(X)] \geq P_0(A) \cdot \exp(\frac{\mathbb{E}_{P_0}[S_0(X)|A]^2}{\sigma^2})$$

which gives the desired lemma.

B.5. Another Proof of Lemma 14 in Gaussian Case

We prove the following lemma:

Lemma 19 *For $X \sim \mathcal{N}(0, I_d)$ and any measurable $A \subset \mathbb{R}^d$, we have*

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

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.5.1. GEOMETRIC INEQUALITY ON BINARY HYPERCUBE

We prove the following lemma:

Lemma 20 *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 18. Define $q_i \triangleq h_2(p_i)$, the concavity in Lemma 18 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 18, 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.5.2. TENSOR POWER TRICK

Next we make use of Lemma 20 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 Ledoux (2005).

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 20, we have

$$\left\|\frac{\mathbb{E}X\tilde{a}(X)}{\mathbb{E}\tilde{a}(X)}\right\|_2^2 \leq 2 \cdot \log \frac{1}{\mathbb{E}[\tilde{a}(X)]}. \tag{15}$$

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 $B \to \infty$, (15) becomes

$$\left\|\frac{\mathbb{E}Za(Z)}{\mathbb{E}a(Z)}\right\|_2^2 \leq 2 \cdot \log \frac{1}{\mathbb{E}[a(Z)]}. \tag{16}$$

Note that (16) holds for all d and $a(\cdot)$, the proof of Lemma 19 is complete by choosing $a(\cdot) = \mathbb{1}_A(\cdot)$.

B.6. Proof of Lemma 15

We use the notation $\langle u, v\rangle \triangleq u^T v$ to denote the inner product between two vectors. Moreover, for the sake of notational simplicity, we write $y = S_0(x), y' = S_0(x')$. We have the Taylor expansion

$$\exp(\delta^2\langle S_0(x), S_0(x')\rangle) - 1 = \sum_{m=1}^\infty \frac{\delta^{2m}\langle y, y'\rangle^m}{m!} = \sum_{m=1}^\infty \frac{\delta^{2m}\langle y^{\otimes m}, (y')^{\otimes m}\rangle}{m!}.$$

Hence,

$$S_2 = \sum_{m=1}^\infty \frac{\delta^{2m}}{m!} \sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \frac{\|\mathbb{E}_{P_0} Y^{\otimes m} p_{i,y}(X)\|_2^2}{\mathbb{E}_{P_0} p_{i,y}(X)}. \tag{17}$$

We upper bound $\|\mathbb{E}_{P_0} Y^{\otimes m} p_{i,y}(X)\|_2^2$ for each $m \geq 1$. The tensor $Y^{\otimes m}$ has dimension d^m, and each coordinate of $Y^{\otimes m}$ takes the form $y_{i_1} y_{i_2} \cdots y_{i_m}$ for $i_1, i_2, \cdots, i_m \in [d]$. We split the entire d^m indices into several groups:

1. If there are repeated elements in $i_1, \cdots, i_m$, we define d groups $\mathcal{G}_{i_1,i_2,\cdots,i_m} = \{(i_1, \cdots, i_m)\}$ for each $i_m \in [d]$. Each group $\mathcal{G}$ only has one element.

2. If there is no repeated element in $i_1, \cdots, i_{m-1}$, we define one group $\mathcal{H}_{i_1,i_2,\cdots,i_{m-1}} = \{(i_1, \cdots, i_m) : i_m \notin \{i_1, \cdots, i_{m-1}\}\}$. Each group $\mathcal{H}$ has $d - m + 1$ elements.

It's obvious that all possible $\mathcal{G}$'s and $\mathcal{H}$'s constitute a partition of $[d]^m$. Let $Y_{\mathcal{G}}$ be the shortened vector consisting of indices in $\mathcal{G}$ only, we have

$$\|\mathbb{E}_{P_0} Y^{\otimes m} p_{i,y}(X)\|_2^2 = \sum_{\mathcal{G}} \|\mathbb{E}_{P_0} Y_{\mathcal{G}} p_{i,y}(X)\|_2^2 + \sum_{\mathcal{H}} \|\mathbb{E}_{P_0} Y_{\mathcal{H}} p_{i,y}(X)\|_2^2.$$

Next we upper bound each term of the RHS separately.

For each $\mathcal{G}$-group $\mathcal{G}_{i_1,\cdots,i_m}$, the restriction $Y_{\mathcal{G}_{i_1,\cdots,i_m}}$ is in fact a scalar, and thus by the boundedness assumption of Y, we have $\|\mathbb{E}_{P_0} Y_{\mathcal{G}} p_{i,y}(X)\|_2^2 \le R^{2m}(\mathbb{E}_{P_0} p_{i,y}(X))^2$. The total number of $\mathcal{G}$-groups is at most $d^{m-1} \cdot (m-1)$, and thus

$$\sum_{\mathcal{G}} \|\mathbb{E}_{P_0} Y_{\mathcal{G}} p_{i,y}(X)\|_2^2 \le d^{m-1}(m-1) \cdot R^{2m}(\mathbb{E}_{P_0} p_{i,y}(X))^2. \tag{18}$$

For each $\mathcal{H}$-group $\mathcal{H}_{i_1,\cdots,i_{m-1}}$, the restriction $Y_{\mathcal{H}_{i_1,\cdots,i_{m-1}}}$ is a vector in $\mathbb{R}^{d-m+1}$. Moreover, for any unit vector $v \in \mathbb{R}^{d-m+1}$, the inner product

$$\langle Y_{\mathcal{H}_{i_1,\cdots,i_{m-1}}}, v\rangle = y_{i_1} y_{i_2} \cdots y_{i_{m-1}} \langle (y_{i_m})_{i_m \notin \{i_1,\cdots,i_{m-1}\}}, v\rangle$$

has squared ψ_2 norm at most $R^{2(m-1)}\sigma^2$, where we have used $|y_{i_l}| \le R$ for any $l \in [m-1]$, the sub-Gaussian assumption of each y_{i_m}, and the independence between coordinates of Y. As a result, using the same argument in Lemma 14, we have

$$\|\mathbb{E}_{P_0} Y_{\mathcal{H}_{i_1,\cdots,i_{m-1}}} p_{i,y}(X)\|_2^2 \le R^{2(m-1)}\sigma^2 \cdot (\mathbb{E}_{P_0} p_{i,y}(X))^2 \log \frac{2}{\mathbb{E}_{P_0} p_{i,y}(X)}.$$

The total number of $\mathcal{H}$-groups is $\binom{d}{m-1} \le d^{m-1}$, and thus

$$\sum_{\mathcal{H}} \|\mathbb{E}_{P_0} Y_{\mathcal{H}} p_{i,y}(X)\|_2^2 \le d^{m-1} \cdot R^{2(m-1)}\sigma^2 (\mathbb{E}_{P_0} p_{i,y}(X))^2 \log \frac{2}{\mathbb{E}_{P_0} p_{i,y}(X)}. \tag{19}$$

Combining (18) and (19), we have

$$\begin{aligned}
&\sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \frac{\|\mathbb{E}_{P_0} Y^{\otimes m} p_{i,y}(X)\|_2^2}{\mathbb{E}_{P_0} p_{i,y}(X)} \\
&\quad \le (dR^2)^{m-1} \sum_{y\in\{0,1\}^{nk}} w_{i,y} \left((m-1)R^2 \cdot \mathbb{E}_{P_0} p_{i,y}(X) + \sigma^2 \cdot \mathbb{E}_{P_0} p_{i,y}(X) \log \frac{2}{\mathbb{E}_{P_0} p_{i,y}(X)} \right) \\
&\quad = (dR^2)^{m-1} \cdot \left((m-1)R^2 + \sigma^2 \sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \mathbb{E}_{P_0} p_{i,y}(X) \log \frac{2}{\mathbb{E}_{P_0} p_{i,y}(X)} \right)
\end{aligned} \tag{20}$$

where we have used Lemma 10 in the last equality. For the remaining sum, we apply Lemma 10 and Jensen's inequality to the concave function $x \mapsto x \log \frac{1}{x}$ to obtain

$$\begin{aligned}
&\sum_{y\in\{0,1\}^{nk}} w_{i,y} \cdot \mathbb{E}_{P_0} p_{i,y}(X) \log \frac{2}{\mathbb{E}_{P_0} p_{i,y}(X)} \\
&\quad = 2^k \sum_{y\in\{0,1\}^{nk}} \frac{w_{i,y}}{2^k} \cdot \mathbb{E}_{P_0} p_{i,y}(X) \log \frac{2}{\mathbb{E}_{P_0} p_{i,y}(X)} \\
&\quad \le 2^k \left(\sum_{y\in\{0,1\}^{nk}} \frac{w_{i,y}}{2^k} \mathbb{E}_{P_0} p_{i,y}(X) \right) \log \frac{2}{\left(\sum_{y\in\{0,1\}^{nk}} \frac{w_{i,y}}{2^k} \mathbb{E}_{P_0} p_{i,y}(X) \right)} \\
&\quad = 2^k \cdot \frac{1}{2^k} \log \frac{2}{2^{-k}} = k+1. \qquad (21)
\end{aligned}$$

Finally, a combination (17), (20) and (21) yields

$$\begin{aligned}
S_2 &\le \sum_{m=1}^{\infty} \frac{\delta^{2m}}{m!} \cdot (dR^2)^{m-1} \left((m-1)R^2 + (k+1)\sigma^2 \right) \\
&\le \delta^2 R^2 \sum_{m=1}^{\infty} \frac{(\delta^2 dR^2)^{m-1}}{(m-1)!} + (k+1)\delta^2\sigma^2 \sum_{m=1}^{\infty} \frac{(\delta^2 dR^2)^{m-1}}{m!} \\
&\le \delta^2 \exp(\delta^2 dR^2) \cdot \left((k+1)\sigma^2 + R^2 \right) \\
&\le C\delta^2 \left(k\sigma^2 + R^2 \right)
\end{aligned}$$

where the last step used the assumption $\delta^2 dR^2 \le 1$. The proof is complete.

Appendix C. Proof of Propositions and Theorem 8

C.1. Proof of Proposition 1

For notational simplicity, let $r_1(x,x'), r_2(x,x')$ denote the remainder terms (inside the expectation over X, X') appearing in (1) and (3), respectively.

For the Multinomial distribution with probability measure θ over $d+1$ elements, we consider the free parameter $(\theta_1, \cdots, \theta_d)$ with $\theta_{d+1} = 1 - \sum_{i=1}^{d} \theta_i$. In this model we have

$$S_{\theta_0}(X)_i = \frac{\mathbb{1}(X=i)}{\theta_i} - \frac{\mathbb{1}(X=d+1)}{\theta_{d+1}}, \qquad I(\theta_0)_{i,j} = \frac{\mathbb{1}(i=j)}{\theta_i} + \frac{1}{\theta_{d+1}}, \qquad i,j \in [d].$$

Moreover,

$$\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(X) - 1 = \sum_{i=1}^{d} \frac{\delta u_i}{\theta_i} \mathbb{1}(X=i) - \frac{\delta \sum_{i=1}^{d} u_i}{\theta_{d+1}} \mathbb{1}(X=d+1) = \delta \cdot S_{\theta_0}(X)^T U.$$

Since $\mathbb{E}[UU^T] = I$, for any $x, x' \in \mathcal{X}$, we have

$$\mathbb{E}_U \left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x) - 1 \right) \left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x') - 1 \right) = \delta^2 \cdot S_{\theta_0}(x)^T S_{\theta_0}(x')$$

i.e., (1) is satisfied. Hence, the Multinomial distribution satisfies Assumptions 1 and 2.

Next we consider the product Bernoulli distribution $\prod_{i=1}^{d} \mathsf{Bern}(\theta_i)$ with $\theta_0 = (p, p, \cdots, p)$, where $p \in (0,1)$. Assume that $\mathcal{X} = \{-1, 1\}$, in this model we have

$$S_{\theta_0}(X)_i = \frac{X_i + (1-2p)}{2p(1-p)}, \qquad I(\theta_0)_{i,j} = \frac{\mathbb{1}(i=j)}{p(1-p)}, \qquad i, j \in [d].$$

Moreover,

$$\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(X) - 1 = \prod_{i=1}^{d} \left(1 + \delta u_i \cdot \frac{X_i + (1-2p)}{2p(1-p)}\right) - 1 = \prod_{i=1}^{d} (1 + \delta u_i \cdot S_{\theta_0}(X_i)) - 1.$$

As a result,

$$\mathbb{E}_U \left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x) - 1\right)\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x') - 1\right) = \prod_{i=1}^{d} (1 + \delta^2 S_{\theta_0}(x_i) S_{\theta_0}(x_i')) - 1.$$

Since $S_{\theta_0}(X_i), i \in [d]$ are i.i.d. Bernoulli random variables, we may calculate the second moment of the remainder term $r_1(X, X')$ explicitly as (note that $B = \frac{1}{p(1-p)}$)

$$\mathbb{E} r_1(X, X')^2 = (1 + \delta^4 B^2)^d - 1 - \delta^4 B^2 d \le (c_0 \cdot \delta^4 B^2 d)^2$$

as long as $\delta^4 B^2 d = O(1)$, establishing (1). As for Assumption 3, we choose $\mathcal{X}_0 = \mathcal{X}_1 = \mathcal{X} = \{\pm 1\}^d$. Since

$$r_2(x, x') = \prod_{i=1}^{d} (1 + \delta^2 S_{\theta_0}(x_i) S_{\theta_0}(x_i')) - \exp(\delta^2 S_{\theta_0}(x)^T S_{\theta_0}(x')) \le 0,$$

inequality (3) holds. As for inequality (2), for any $u \in \{\pm 1\}^d$ we have

$$\begin{aligned}
\mathbb{E}\left(\frac{dP_{\theta_0+\delta u}}{dP_{\theta_0}}(X) - 1\right)^4 &= \sum_{\ell=0}^{4} (-1)^\ell \binom{4}{\ell} \prod_{i=1}^{d} \left(p\left(1 + \frac{\delta u_i}{p}\right)^\ell + (1-p)\left(1 - \frac{\delta u_i}{1-p}\right)^\ell\right) \\
&= \sum_{\ell=0}^{4} (-1)^\ell \binom{4}{\ell} \prod_{i=1}^{d} \left(1 + \binom{\ell}{2} B\delta^2 + \binom{\ell}{3} u_i^3 \left(\frac{1}{p^2} - \frac{1}{(1-p)^2}\right)\delta^3 + O(B^3\delta^4)\right) \\
&= \sum_{\ell=0}^{4} (-1)^\ell \binom{4}{\ell} \left(1 + \sum_{i=1}^{d} \left(\binom{\ell}{2} B\delta^2 + \binom{\ell}{3} u_i^3 \left(\frac{1}{p^2} - \frac{1}{(1-p)^2}\right)\delta^3\right)\right. \\
&\qquad \left. + O(B^2\delta^4 d^2 + B^3\delta^4 d)\right) \\
&= O(B^2\delta^4 d^2 + B^3\delta^4 d)
\end{aligned}$$

if $B^2\delta^4 d^2 + B^3\delta^4 d = O(1)$. Hence, Assumptions 1, 2, 3 hold for product Bernoulli models with any $p \in (0,1)$.

Finally we consider the Gaussian location model $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ for any $\theta \in \mathbb{R}^d, \sigma > 0$. By translation and scaling properties, it suffices to consider the case where $\theta_0 = (0, 0, \cdots, 0), \sigma^2 = 1$. In this model we have

$$S_{\theta_0}(X)_i = X_i, \qquad I(\theta_0)_{i,j} = \mathbb{1}(i=j), \qquad i, j \in [d].$$

Moreover,

$$\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(X)-1=\exp\left(\delta\sum_{i=1}^d u_iX_i-\frac{\delta^2 d}{2}\right)-1.$$

As a result,

$$\begin{aligned}
&\mathbb{E}_U\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x)-1\right)\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x')-1\right)\\
&=\exp(-\delta^2 d)\prod_{i=1}^d\cosh(\delta(x_i+x_i'))-\exp(-\delta^2 d/2)\left(\prod_{i=1}^d\cosh(\delta x_i)+\prod_{i=1}^d\cosh(\delta x_i')\right)+1\\
&=\exp(-\delta^2 d)\left(\prod_{i=1}^d\cosh(\delta(x_i+x_i'))-\prod_{i=1}^d\cosh(\delta x_i)\prod_{i=1}^d\cosh(\delta x_i')\right)\\
&\quad+\left(\exp\left(-\frac{\delta^2 d}{2}\right)\prod_{i=1}^d\cosh(\delta x_i)-1\right)\left(\exp\left(-\frac{\delta^2 d}{2}\right)\prod_{i=1}^d\cosh(\delta x_i')-1\right). \qquad (22)
\end{aligned}$$

Note that $\cosh(x)=\exp(\frac{x^2}{2}+O(x^4))$, and $\sum_{i=1}^d X_i^2=d+O_P(\sqrt{d}),\sum_{i=1}^d X_i^4=O_P(d)$ for $X\sim\mathcal{N}(0,I_d)$, we have

$$\exp\left(-\frac{\delta^2 d}{2}\right)\prod_{i=1}^d\cosh(\delta X_i)-1=\exp\left(\frac{\delta^2(\sum_{i=1}^d X_i^2-d)}{2}+O(\delta^4)\cdot\sum_{i=1}^d X_i^4\right)-1=O_P(\delta^2\sqrt{d})$$

as long as $\delta^4 d=O(1)$. Hence, the second term in (22) is of the order $O_P(\delta^4 d)$. Similarly, the first term in (22) is of the order $(1+O_P(\delta^2\sqrt{d}))\cdot(\exp(\delta^2 X^TX')-1)$. Note that $X^TX'=O_P(\sqrt{d})$, we have $\exp(\delta^2 X^TX')-1=\delta^2 X^TX'+O_P(\delta^4 d)$, and therefore

$$\mathbb{E}_U\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(X)-1\right)\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(X')-1\right)=\delta^2 X^TX'+O_P(\delta^4 d)$$

establishing (1).

As for Assumption (3), choose $\mathcal{X}_0=[-C\sqrt{\log d},C\sqrt{\log d}]^d\subset\mathcal{X}=\mathbb{R}^d$. By Gaussian tail, by choosing C large enough we have $\mathbb{P}(\mathcal{X}_0)\geq 1-d^{-5}$. Also, choosing $\mathcal{X}_1=\{x\in\mathbb{R}^d:|\sum_{i=1}^d x_i^2-d|\leq C\sqrt{d\log d},\sum_{i=1}^d x_i^4\leq Cd\}$, for C large enough we have $\mathbb{P}(\mathcal{X}_1)\geq 1-d^{-5}$. For $x,x'\in\mathcal{X}_0\cap\mathcal{X}_1$ and $\delta^4 d\log d=O(1)$, applying $\cosh(x)=\exp(\frac{x^2}{2}+O(x^4))$ in (22) yields

$$\mathbb{E}_U\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x)-1\right)\left(\frac{dP_{\theta_0+\delta U}}{dP_{\theta_0}}(x')-1\right)=(1+O(\delta^2\sqrt{d\log d}))\cdot(\exp(\delta^2 x^Tx')-1)+O(\delta^4 d\log d)$$

establishing (3).

Finally, for any $u\in\{\pm1\}^n$ and $\delta=O(d^{-\frac{1}{2}})$, we have

$$\begin{aligned}
\mathbb{E}\left(\frac{dP_{\theta_0+\delta u}}{dP_{\theta_0}}(X)-1\right)^4&=\mathbb{E}\left[\exp\left(\delta\sum_{i=1}^d u_iX_i-\frac{\delta^2 d}{2}\right)-1\right]^4\\
&=\exp(6\delta^2 d)-4\exp(3\delta^2 d)+6\exp(\delta^2 d)-3\\
&=O(\delta^4 d^2)
\end{aligned}$$

where the last step follows from Taylor expansion. Hence, the Gaussian location model satisfies all assumptions.

C.2. Proof of Proposition 7

For the first result, the lower bound follows from Zhang et al. (2013) (or Theorem 4), with a matching upper bound in Zhang et al. (2013). The lower bound of the second result follows from Theorem 2 (or along the same line of Han et al. (2018)), and it suffices to prove an upper bound for the case where $\sum_{i=1}^{d} \theta_i = 1$.

We apply a slightly different "simulate-and-infer" procedure in Acharya et al. (2018). Specifically, for $2^k \leq d$, we split $[d]$ into $m \triangleq \frac{d}{2^k-2}$ (assumed to be an integer) groups $\mathcal{G}_1, \cdots, \mathcal{G}_m$ of size $2^k - 2$ each, and also split the sensors $[n]$ into $N = \frac{n}{2m}$ (also assumed to be an integer) groups $\mathcal{H}_1, \cdots, \mathcal{H}_N$ of size $2m$ each. For each group $\mathcal{H}_j$ of $2m$ sensors, consider the following protocol: for any $\ell \in [m]$,

1. sensor $(2\ell - 1)$ sends $a_0 \in [2^k]$ if $\sum_{i \in \mathcal{G}_\ell} X_{2\ell-1,i} = 0$, sends $a_1 \in [2^k]$ if $\sum_{i \in \mathcal{G}_\ell} X_{2\ell-1,i} \geq 2$, and sends the unique $i^* \in \mathcal{G}_\ell$ with $X_{2\ell-1,i^*} = 1$ in the remaining $2^k - 2$ cases;
2. sensor 2ℓ first looks at the message that sensor $(2\ell - 1)$ transmits. If the message is a_0 or a_1, sensor 2ℓ can transmit an arbitrary message; otherwise, sensor $(2\ell - 1)$ must have transmitted a unique location $i \in [d]$, and then sensor $2l$ tranmits the one-bit message $X_{2\ell,i}$.

Clearly the communication constraints are satisfied here. For each group $\mathcal{H}_j$ of sensors, we call this group *succeeds* if and only if:

1. there exists a unique $\ell^* \in [m]$ such that sensor $(2\ell^* - 1)$ does not send a_0 or a_1, and any sensor $(2\ell - 1)$ sends a_0 for $\ell \neq \ell^*$;
2. for the index ℓ^* above, sensor $2\ell^*$ sends zero.

If this group succeeds, the centralizer records the index $i^* \in [d]$ sent by sensor $(2\ell^* - 1)$ above. We show that:

1. conditioning on the event that the group succeeds, $i^* \sim \mathsf{Multi}(1; \theta)$;
2. any group succeeds with probability $\Omega(1)$.

To establish the first result, note that the probability for any fixed group to succeed *and* $i^* = i \in \mathcal{G}_\ell$ is

$$p_i = \prod_{\ell' \neq \ell} \prod_{i' \in \mathcal{G}_{\ell'}} (1 - \theta_{i'}) \cdot \theta_i \prod_{i' \in \mathcal{G}_\ell, i' \neq i} (1 - \theta_{i'}) \cdot (1 - \theta_i) = \theta_i \prod_{i'=1}^{d} (1 - \theta_{i'}).$$

Hence, the probability of that group to succeed is $p = \sum_{i=1}^{d} p_i = \prod_{i'=1}^{d} (1 - \theta_{i'})$, and thus the conditional distribution of i^* is exactly $\mathsf{Multi}(1; \theta)$. The second result is established using the same arguments as (Acharya et al., 2018, Theorem 4.7), while replacing one sensor by two sensors if necessary.

Hence, we have N groups, each of which succeeds independently with probability $\Omega(1)$. Let M be the number of successful groups, Lemma 17 yields $\mathbb{P}(M \geq \Omega(1) \cdot N) \geq 1 - e^{-\Omega(N)}$. Moreover, we observe M i.i.d. observations from the discrete distribution $(\theta_1, \theta_2, \cdots, \theta_d)$, where the empirical distribution has squared ℓ_2 risk at most

$$\frac{1}{M} \lesssim \frac{1}{N} \asymp \frac{d}{n2^k}$$

which completes the proof for the case where $2^k \leq d$.

When $2^k > d$, we simply apply the previous protocol again with 2^k replaced by d (and $m = 1$), then any group of two sensors has $\Omega(1)$ probability to generate a random sample from the discrete distribution $(\theta_1, \cdots, \theta_d)$. As a result, the squared ℓ_2 risk of the empirical distribution is at most $O(\frac{1}{n})$ with probability at least $1 - e^{-\Omega(n)}$, as desired.

C.3. Proof of Proposition 16

Let X follow the uniform distribution on Ω, then $\bar{v} = \mathbb{E}[X|A]$. Choosing $S_0(X) = X$ in Lemmas 13 and 14, each coordinate of X has variance 1 and is 1-sub-Gaussian. By Lemma 13, for $|A| = 2^{d-1}$ we have

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

establishing the first inequality.

Similarly, the second inequality follows from Lemma 20 (and its proof).

C.4. Proof of Theorem 8

We construct a new 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 $P_u \triangleq \mathcal{N}(\delta u, I_d)$, and

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

As a result, we have $\log \frac{|\mathcal{U}|}{N_{\max}(s/5)} \geq cs \log \frac{d}{s}$ for some constant $c > 0$, and Lemma 11 gives

$$\inf_{\Pi_{\mathsf{BB}}} \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \geq \frac{s\delta^2}{10}\left(1 - \frac{I(U;Y) + \log 2}{cs \log(d/s)}\right). \tag{23}$$

By construction, conditioning on the support T of U, the restriction U_T is uniform on $\{\pm 1\}^d$. Hence, by Proposition 1, Assumption (3) still holds with d replaced by s, $d \log d$ replaced by $s \log d$, and inner product between score functions replaced by the expected inner product between score functions restricted on T, where the expectation is taken over the random support T with $|T| = s$. Hence, by the same argument as in the proof of Theorem 4, we arrive at

$$I(U;Y) \leq \frac{Cns}{d}\left(\frac{\delta^2}{\sigma^2}(k + \log d) + \frac{\delta^4 s \log d}{\sigma^4}\right)$$

for some universal constant $C > 0$. Now choosing $\delta^2 \asymp \frac{d \log(d/s)}{nk}\sigma^2$ in (23) completes the proof.