Journal of Machine Learning Research 25 (2024) 1-63 Submitted 3/21; Revised 12/22; Published 2/24

Distributed Gaussian Mean Estimation under Communication Constraints: Optimal Rates and Communication-Efficient Algorithms

T. Tony Cai TCAI@WHARTON.UPENN.EDU
Department of Statistics and Data Science
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104, USA

Hongji Wei WEIHONGJI1206@GMAIL.COM
Department of Statistics and Data Science
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: Andrea Montanari

Abstract

Distributed estimation of a Gaussian mean under communication constraints is studied in a decision theoretical framework. Minimax rates of convergence, which characterize the tradeoff between communication costs and statistical accuracy, are established under the independent protocols. Communication-efficient and statistically optimal procedures are developed. In the univariate case, the optimal rate depends only on the total communication budget, so long as each local machine has at least one bit. However, in the multivariate case, the minimax rate depends on the specific allocations of the communication budgets among the local machines.

Although optimal estimation of a Gaussian mean is relatively simple in the conventional setting, it is quite involved under communication constraints, both in terms of the optimal procedure design and the lower bound argument. An essential step is the decomposition of the minimax estimation problem into two stages, localization and refinement. This critical decomposition provides a framework for both the lower bound analysis and optimal procedure design. The optimality results and techniques developed in the present paper can be useful for solving other problems such as distributed nonparametric function estimation and sparse signal recovery.

Keywords: Communication constraints, distributed learning, Gaussian mean, minimax lower bound, optimal rate of convergence

1. Introduction

In the conventional statistical decision theoretical framework, the focus is on the centralized setting where all the data are collected together and directly available. The main goal is to develop optimal (estimation, testing, detection, ...) procedures, where optimality is understood with respect to the sample size and parameter space. Communication/computational costs are not part of the consideration.

In the age of big data, communication/computational concerns associated with a statistical procedure are becoming increasingly important in contemporary applications. One of the difficulties for analyzing large datasets is that data are distributed, instead of in a single centralized location. This setting arises naturally in many statistical applications.

- **Large datasets.** When the datasets are too large to be stored on a single computer or data center, it is necessary to divide the whole dataset into multiple computers or data centers, each assigned a smaller subset of the full dataset. Such is the case for a wide range of applications.

- **Privacy and security.** Privacy and security concerns can also cause the decentralization of the datasets. For example, medical and financial institutions often collect datasets that contain sensitive and valuable information. For privacy and security reasons, the data cannot be released to a third party for a centralized analysis and need to be stored in different and secure places while performing data analysis.

Distributed learning, which aims to learn from distributed datasets, has attracted much recent attention. For example, Google AI proposed a machine learning setting called "Federated Learning" (McMahan and Ramage, 2017), which develops a high-quality centralized model while the training data remain distributed over a large number of clients. Figure 1a provides a simple illustration of a distributed learning network. In addition to advances on architecture design for distributed learning in practice, there is also an increasing amount of literature on distributed learning theories, including Jordan et al. (2019), Battey et al. (2018), Dobriban and Sheng (2018), and Fan et al. (2019) in statistics, computer science, and information theory communities. Several distributed learning procedures with some theoretical properties have been developed in recent works. However, they do not impose any communication constraints on the proposed procedures thus fail to characterize the relationship between the communication costs and statistical accuracy. Indeed, in a decision theoretical framework, if no communication constraints are imposed, one can always output the original data from the local machines to the central machine and treat the problem same as in the conventional centralized setting.

For large-scale data analysis, communications between machines can be slow and expensive and limitation on bandwidth and communication sometimes becomes the main bottleneck on statistical efficiency. It is therefore necessary to take communication constraints into consideration when constructing statistical procedures. When the communication budget is limited, the algorithm must carefully "compress" the information contained in the data as efficiently as possible, leading to a trade-off between communication costs and statistical accuracy. The precisely quantification of this trade-off is an important and challenging problem.

Estimation of a Gaussian mean occupies a central position in parametric statistical inference. In the present paper we consider distributed Gaussian mean estimation under the communication constraints in both the univariate and multivariate settings. Although optimal estimation of a Gaussian mean is a relatively simple problem in the conventional setting, this problem is quite involved under the communication constraints, both in terms of the construction of the rate optimal distributed estimator and the lower bound argument. Optimal distributed estimation of a Gaussian mean also serves as a starting point for investigating other more complicated statistical problems in distributed learning including distributed nonparametric function estimation, distributed high-dimensional linear regression, and distributed large-scale multiple testing.

1.1 Problem formulation

We begin by giving a formal definition of **transcript**, **distributed estimator**, and **independent distributed protocol**. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric family of distributions supported on space $\mathcal{X}$, where $\theta \in \Theta \subseteq \mathbb{R}^d$ is the parameter of interest. Suppose there are m local machines and a central machine, where each local machine contains n i.i.d observations and the central machine produces the final estimator of θ under the communication constraints between the local and central machines. More precisely, suppose we observe i.i.d. random samples drawn from a distribution $P_\theta \in \mathcal{P}$:

$$X_{i,j} \overset{\text{iid}}{\sim} P_\theta, \quad i = 1, \ldots, m;\ j = 1, \ldots, n$$

where the i-th local machine has access to $X_{i,1}, X_{i,2}, ..., X_{i,n}$ only. We denote $\tilde{X}_i = (X_{i,1}, X_{i,2}, ..., X_{i,n})$ as the set of data on the i-th local machine.

For $i = 1, ..., m$, let $b_i \geq 1$ be a positive integer and the i-th local machine can only transmit b_i bits to the central machine. That is, the observation $\tilde{X}_i$ on the i-th local machine needs to be processed to a binary string of length b_i by a (possibly random) function $\Pi_i : \mathcal{X}^n \to \{0,1\}^{b_i}$. The resulting string $Z_i \triangleq \Pi_i(\tilde{X}_i)$, which is called the **transcript** from the i-th machine, is then transmitted to the central machine. Finally, a **distributed estimator** $\hat{\theta}$ is constructed on the central machine based on the transcripts $Z_1, Z_2, ..., Z_m$,

$$\hat{\theta} = \hat{\theta}(Z_1, Z_2, ..., Z_m).$$

The above scheme to obtain a distributed estimator $\hat{\theta}$ is called an **independent distributed protocol**, or independent protocol.

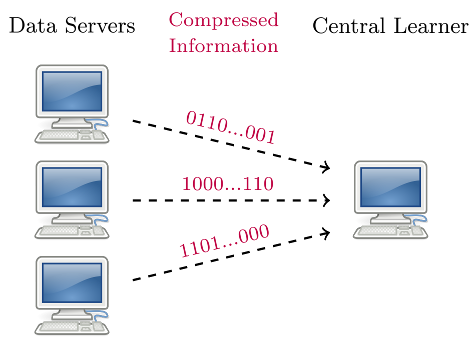
In addition to the independent protocol, there are other more general and interactive distributed protocols including the sequential protocol and blackboard protocol, which are two popular communication protocols considered in the literature (Zhang et al., 2013a; Barnes et al., 2019). We shall only focus on the independent protocol in the present work.

The class of independent distributed protocols with communication budgets $b_1, b_2, ..., b_m$ is defined as

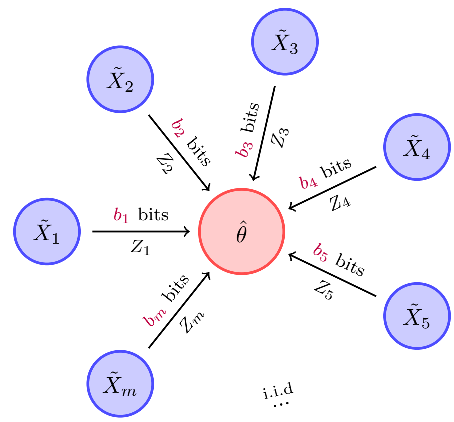
$$\begin{aligned}\mathcal{A}_{ind}(b_1, b_2, ..., b_m) = \{(\hat{\theta}, \Pi_1, \Pi_2, ..., \Pi_m) : \Pi_i : \mathcal{X}^n \to \{0,1\}^{b_i},\ i = 1, 2, ..., m,\\ \hat{\theta} = \hat{\theta}(\Pi_1(\tilde{X}_1), ..., \Pi_m(\tilde{X}_m))\}.\end{aligned}$$

We use $b_{1:m}$ as a shorthand for $(b_1, b_2, ..., b_m)$ and denote $\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})$ for $(\hat{\theta}, \Pi_1, ..., \Pi_m) \in \mathcal{A}_{ind}(b_{1:m})$. We shall always assume $b_i \geq 1$ for all $i = 1, 2, ..., m$, i.e. each local machine can transmit at least one bit to the central machine. Otherwise, if no communication is allowed from any of the local machines, one can just exclude those local machines and treat the problem as if there are fewer local machines available. Figure 1b gives a simple illustration for the distributed protocols.

As usual, the estimation accuracy of a distributed estimator $\hat{\theta}$ is measured by the mean squared error (MSE), $\mathbb{E}_{P_\theta} \|\hat{\theta} - \theta\|_2^2$, where the expectation is taken over the randomness in



(a) Distributed learning network



(b) Independent distributed protocol

Figure 1: (a) Left panel: An illustration of a distributed learning network. Communication between the data servers and the central learner is necessary in order to learn from distributed datasets. (b) Right panel: An illustration of independent distributed protocol. The i-th machine can only transmit a b_i bits transcript to the central machine. The transcript Z_i only depends on observations $\tilde{X}_i$.

both the data and construction of the transcripts and estimator. As in the conventional decision theoretical framework, a quantity of particular interest in distributed learning is the minimax risk for the distributed protocols

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{P_\theta \in \mathcal{P}} \mathbb{E}_{P_\theta} \|\hat{\theta} - \theta\|_2^2,$$

which characterizes the difficulty of the distributed learning problem under the communication constraints $b_{1:m}$. As mentioned earlier, in a rigorous decision theoretical formulation of distributed learning, the communication constraints are essential. Without the constraints, one can always output the original data from the local machines to the central machine and the problem is then reduced to the usual centralized setting.

1.2 Distributed estimation of a univariate Gaussian mean

We first consider distributed estimation of a univariate Gaussian mean under the communication constraints $b_{1:m}$, where $P_\theta = N(\theta, \sigma^2)$ with $\theta \in [0, 1]$ and the variance σ^2 known. Set $\sigma_n = \sigma/\sqrt{n}$. Note that by a sufficiency argument, one can estimate θ based on the sample means $X_i \triangleq \frac{1}{n}\sum_{j=1}^n X_{i,j}$ on the local machines, and the problem is the same as if one only observes $X_i \sim N(\theta, \sigma_n^2)$ on the i-th machine, for $i = 1, \ldots, m$. Throughout the paper we will focus on the case $\sigma_n < 1$, since the case $\sigma_n \geq 1$ can be considered easy and the minimax rate of convergence has been already derived in the literature, see Zhang et al. (2013b); Garg et al. (2014); Braverman et al. (2016); Barnes et al. (2019).

Assume $\sigma_n < 1$ and $b_i \geq 1$ for $i = 1, 2, ..., m$, our analysis in Section 2 establishes the following minimax rate of convergence for distributed univariate Gaussian mean estimation

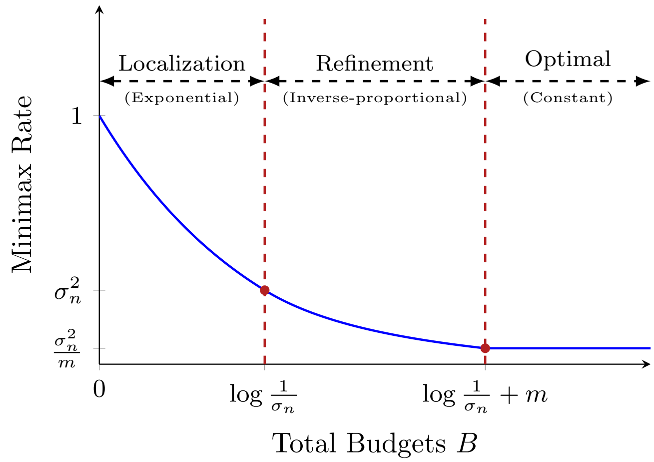


Figure 2: The minimax rate of univariate Gaussian mean estimation under communication constraints has 3 phases: localization, refinement and optimal-rate.

under the communication constraints $b_{1:m}$,

$$\inf_{\hat{\theta}\in\mathcal{A}_{ind}(b_{1:m})} \sup_{\theta\in[0,1]} \mathbb{E}(\hat{\theta}-\theta)^2 \asymp \begin{cases} 2^{-2B} & \text{if } B < \log_2 \frac{1}{\sigma_n} + 2 \\ \frac{\sigma_n^2}{(B-\log_2 \frac{1}{\sigma_n})} & \text{if } \log_2 \frac{1}{\sigma_n} + 2 \le B < \log_2 \frac{1}{\sigma_n} + m \\ \frac{\sigma_n^2}{m} & \text{if } B \ge \log_2 \frac{1}{\sigma_n} + m \end{cases}, \tag{1}$$

where $B = \sum_{i=1}^m b_i$ is the total communication budget, and $a \asymp b$ denotes $cb \le a \le Cb$ for some constants $c, C > 0$.

The above minimax rate characterizes the trade-off between the communication costs and statistical accuracy for univariate Gaussian mean estimation. An illustration of the minimax rate is shown in Figure 2.

The minimax rate (1) is interesting in several aspects. First, the optimal rate of convergence only depends on the total communication budget $B = \sum_{i=1}^m b_i$, but not the specific allocation of the communication budgets among the m local machines, as long as each machine has at least one bit. Second, the rate of convergence has three different phases:

1. Localization phase. When $B < \log_2 \frac{1}{\sigma_n} + 2$, as a function of B, the minimax risk decreases fast at an exponential rate. In this phase, having more communication budget is very beneficial in terms of improving the estimation accuracy. Also, in this phase, the statistical accuracy is the same as noiseless case (i.e. $\sigma_n \approx 0$) where statistical risk mainly comes from approximation error under communication constraints.

2. Refinement phase. When $\log_2 \frac{1}{\sigma_n} + 2 \le B < \log_2 \frac{1}{\sigma_n} + m$, as a function of B, the minimax risk decreases relatively slowly and is inverse-proportional to the total communication budget B. This can be also viewed as an intermediate phase between localization and optimal-rate phase.

3. Optimal-rate phase. When $B \geq \log_2 \frac{1}{\sigma_n} + m$, the minimax rate does not depend on B, and is the same as in the centralized setting where all the data are combined and no communication constraints are present.

An essential technique for solving this problem is the decomposition of the minimax estimation problem into two steps, *localization* and *refinement*. This critical decomposition provides a framework for both the lower bound analysis and optimal procedure design. In the lower bound analysis, the statistical error is decomposed into "localization error" and "refinement error". It is shown that one of these two terms is inevitably large under the communication constraints. In our optimal procedure called MODGAME, bits of the transcripts are divided into three types: crude localization bits, finer localization bits, and refinement bits. They compress the local data in a way that both the localization and refinement errors can be optimally reduced. More technical details and discussion are presented in Section 2. Furthermore, it will be shown that MODGAME is also robust against departures from Gaussianity. See Section 4 for a detailed discussion.

1.3 Distributed estimation of a multivariate Gaussian mean

We then consider the multivariate case under the communication constraints $b_{1:m}$, where $P_\theta = N_d(\theta, \sigma^2 I_d)$ with $\theta \in [0,1]^d$ and the noise level σ is known. As in the univariate case, by a sufficiency argument, it is equivalent to consider distributed estimation where each local machine only observes a local sample mean vector $X_i \sim N_d(\theta, \sigma_n^2 I_d)$, with $\sigma_n = \sigma/\sqrt{n}$. The goal is to optimally estimate the mean vector θ under the squared error loss.

Assume $\sigma_n < 1$ and $b_i \geq 1$ for $i = 1, 2, ..., m$, the construction and the analysis given in Section 3 show that the minimax rate of convergence in this case is given by

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{\theta \in [0,1]^d} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \asymp \begin{cases} 2^{-2B/d} d & \text{if } B/d < \log_2 \frac{1}{\sigma_n} + 2 \\ \frac{d\sigma_n^2}{(B/d - \log_2 \frac{1}{\sigma_n})} & \text{if } \log_2 \frac{1}{\sigma_n} + 2 \leq B/d < \log_2 \frac{1}{\sigma_n} + \max\{m', 2\} \\ d \min \frac{\sigma_n^2}{m'} & \text{if } B/d \geq \log_2 \frac{1}{\sigma_n} + \max\{m', 2\} \end{cases} \tag{2}$$

where $B = \sum_{i=1}^m b_i$ is the total communication budgets and $m' = \sum_{i=1}^m \min\left\{\frac{b_i}{d}, 1\right\}$ is the "effective sample size".

The minimax rate in the multivariate case (2) is an extension of its univariate counterpart (1), but it also has its distinct features, both in terms of the estimation procedure and lower bound argument. Intuitively, the total communication budgets B are evenly divided into d parts so that roughly B/d bits can be used to estimate each coordinate. Because there are d coordinates, the risk is multiplied by d. The effective sample size m' is a special and interesting quantity in multivariate Gaussian mean estimation. This quantity suggests that even when the total communication budget is sufficient, the rate of convergence must be larger than the benchmark $d \min\left\{\frac{\sigma_n^2}{m'}, 1\right\}$. There is a gap between the distributed optimal rate and centralized optimal rate if $m' \ll m$. See Section 3 for further technical details and discussion.

1.4 Related literature

The study on how the communication constraints compromise the estimation accuracy in the distributed settings has a long history. Dating back to 1980's, Zhang and Berger (1988) proposed an asymptotically unbiased distributed estimator and calculated its variance. In recent years, there has been emerging literature focusing on the theoretical properties of distributed estimation under communication constraints. Among them, distributed Gaussian mean estimation has been intensively studied. We divide the discussion into two parts – lower bound and upper bound.

Lower bound: Zhang et al. (2013a) introduced general technical tools to derive lower bounds for several distributed estimation problems. Specifically, for d-dimensional Gaussian mean estimation with independent protocols, the lower bound is of order $\frac{\sigma_n^2 d^2}{(\sum_{i=1}^m b_i \wedge d)\log m}$. Garg et al. (2014) studied distributed estimation of the mean of a high-dimensional Gaussian distribution. A lower bound of order $\min\{\frac{\sigma_n^2 d^2}{B}, d\}$ is established for the mean squared error of any independent protocol. Braverman et al. (2016) applied a strong data processing inequality to obtain lower bounds for distributed estimation with blackboard protocols. A lower bound for sparse Gaussian mean estimation is derived. Han et al. (2018); Barnes et al. (2019) proposed non-information theoretic approaches to obtain lower bounds for distributed estimation. In the case of Gaussian mean estimation, it was shown in Barnes et al. (2019) that a lower bound of order $\sigma_n^2 \max\{\frac{d^2}{B}, \frac{d}{m}\}$ holds for any independent, sequential or blackboard protocols.

Upper bound: Garg et al. (2014) proposed a blackboard distributed protocol with the communication cost $O(md)$ which estimates the mean vector up to a squared loss of $O(\frac{d\sigma_n^2}{m})$. Braverman et al. (2016) introduced an independent distributed protocol for Gaussian mean estimation. If $\log(md/\sigma_n) = o(m)$, the protocol achieves the mean squared error $O(\frac{\sigma_n^2 d}{\alpha m})$ with the communication cost $C = \alpha dm$.

In summary, the known minimax rate for distributed Gaussian mean estimation is $\frac{\sigma_n^2 d^2}{B}$ when $\log(md/\sigma_n) = o(m)$. So the tight bound has been already obtained when $\sigma_n \geq 1$. However, when n is large such that $\log(1/\sigma_n)/m$ is bounded away from zero, the optimal rate is still unknown.

In addition to the above closely related literature, Szabó and van Zanten (2018); Zhu and Lafferty (2018) considered distributed nonparametric regression with Gaussian noise and derived an optimal rate of convergence up to a logarithmic factor. The optimal rate is divided into three phases, namely insufficient regime, intermediate regime, and sufficient regime. Current best results for distributed nonparametric regression also suffer from a logarithmic gap, which in our opinion is due to the incomplete understanding of distributed Gaussian mean estimation with a small variance. Other related results can be found in the literature, see, for example, Zhang et al. (2013b); Shamir (2014); Diakonikolas et al. (2017); Han et al. (2018); Lee et al. (2017); Chang et al. (2017); Guo et al. (2017); Kipnis and Duchi (2019); Hadar and Shayevitz (2019); Mücke et al. (2022); Dobriban and Sheng (2018); Mücke and Blanchard (2018); Szabó and van Zanten (2019, 2020).

1.5 Our contribution

Although the interplay between communication costs and statistical accuracy has drawn increasing recent attention, to the best of our knowledge, the present paper is the first to establish a sharp minimax rate for distributed Gaussian mean estimation that holds for all values of the parameters d, m, σ_n and in all communication budget regimes under independent protocol. Two rate-optimal estimation procedures – MODGAME for the univariate case and multi-MODGAME for the multivariate case – are developed and are shown to be robust against departures from Gaussianity.

In particular, the unified minimax rate applies to the case $\sigma_n < 1$. In comparison, when $\sigma_n < 1$, the previous results are not sharp even in the high communication budget regime (i.e. refinement phase and optimal-rate phase). See Remarks 9 and 11 for detailed comparison with previous results. This is an important case that arises in many statistical applications including distributed nonparametric regression and sparse signal recovery. Establishing a sharp and complete minimax rate is not only important for distributed Gaussian mean estimation itself, but also fundamental for solving these related problems.

This paper also develops a key technique – the decomposition of the minimax estimation problem into two steps, *localization* and *refinement*. We provide a general framework and techniques to study the optimal trade-off between the localization and refinement errors. This is reflected in both the construction of the MODGAME procedure and in the lower bound argument. In contrast, the previous literature focused exclusively on the refinement error, and failed to consider the localization error. As a result, the existing results are sharp only when the communication costs for localization are negligible. We believe the technique for understanding the interplay between the localization and refinement errors is of independent interest as it can be used to solve other distributed estimation problems.

1.6 Organization of the paper

We finish this section with notation and definitions that will be used in the rest of the paper. Section 2 studies distributed estimation of a univariate Gaussian mean under communication constraints with independent protocols and Section 3 considers the multivariate case. Section 4 considers the robustness of the proposed procedures against departures from Gaussianity. The numerical performance of the proposed distributed estimators is investigated in Section 5 and further research directions are discussed in Section 6. For reasons of space, we prove the lower bound for the univariate case in Section 7 and defer the proofs of the other main results and the technical lemmas to the Appendix.

1.7 Notation and definitions

For any $a \in \mathbb{R}$, let $\lfloor a \rfloor$ denote the floor function (the largest integer not larger than a). Unless otherwise stated, we shorthand $\log a$ as the base 2 logarithmic of a. For any $a, b \in \mathbb{R}$, let $a \wedge b \triangleq \min\{a, b\}$ and $a \vee b \triangleq \max\{a, b\}$. For any vector a, we will use $a^{(k)}$ to denote the k-th coordinate of a, and denote by $\|a\| \triangleq \sqrt{\sum_k \left(a^{(k)}\right)^2}$ its l_2 norm. For any set S, let $S^k \triangleq S \times S \times ... \times S$ be the Cartesian product of k copies of S. Let $\mathbb{I}_{\{.\}}$ denote the indicator function taking values in $\{0, 1\}$.

For any discrete random variables X, Y supported on $\mathcal{X}, \mathcal{Y}$, the entropy $H(X)$, conditional entropy $H(X|Y)$, and mutual information $I(X;Y)$ are defined as

$$\begin{aligned} H(X) &\triangleq -\sum_{x\in\mathcal{X}} \mathbb{P}(X=x)\log\mathbb{P}(X=x), \\ H(X|Y) &\triangleq -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mathbb{P}(X=x,Y=y)\log\mathbb{P}(X=x|Y=y), \\ I(X;Y) &\triangleq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mathbb{P}(X=x,Y=y)\log\frac{\mathbb{P}(X=x|Y=y)}{\mathbb{P}(X=x)}. \end{aligned}$$

2. Distributed Univariate Gaussian Mean Estimation

In this section we consider distributed estimation of a univariate Gaussian mean, where one observes on m local machines i.i.d. random samples:

$$X_i \overset{\text{iid}}{\sim} N(\theta, \sigma_n^2), \quad i = 1, \ldots, m,$$

under the constraints that the i-th machine has access to X_i only and can transmit b_i bits only to the central machine. We denote by $\mathcal{P}^1_{\sigma_n}$ the Gaussian location family

$$\mathcal{P}^1_{\sigma_n} = \left\{N(\theta,\sigma_n^2) : \theta \in [0,1]\right\},$$

where $\theta \in [0,1]$ is the mean parameter of interest and the variance σ_n^2 is known. For given communication budgets $b_{1:m}$ with $b_i \geq 1$ for $i = 1, \ldots, m$, the goal is to optimally estimate the mean θ under the squared error loss. A particularly interesting quantity is the minimax risk under the communication constraints, i.e., the minimax risk for the independent distributed protocol $\mathcal{A}_{ind}(b_{1:m})$:

$$R_1(b_{1:m}) = \inf_{\hat{\theta}\in\mathcal{A}_{ind}(b_{1:m})} \sup_{\theta\in[0,1]} \mathbb{E}(\hat{\theta}-\theta)^2,$$

which characterizes the difficulty of the estimation problem with independent protocols under the communication constraints.

We first introduce an estimation procedure and provide an upper bound for its performance and then establish a matching lower bound on the minimax risk for the case $\sigma_n < 1$. The upper and lower bounds together establish the minimax rate of convergence and the optimality of the proposed estimator. Then we will briefly discuss the case $\sigma_n \geq 1$ for completeness.

2.1 Estimation procedure - MODGAME

We begin with the construction of an estimation procedure under the communication constraints and provide a theoretical analysis of the proposed procedure. The procedure, called MODGAME (Minimax Optimal Distributed GAussian Mean Estimation), is a deterministic procedure that generates a distributed estimator $\hat{\theta}_D$ under the distributed protocol $\mathcal{A}_{ind}(b_{1:m})$. We assume $\sigma_n < 1$ is known in MODGAME procedure.

High-level intuition of MODGAME procedure: MODGAME consists of two steps: localization and refinement.

1. Roughly speaking, the first step utilizes $\log \frac{1}{\sigma_n} + o(B - \log \frac{1}{\sigma_n})$ bits, out of the total budget $B = \sum_{i=1}^{m} b_i$ bits, for localization to roughly locate where θ is, up to $O(\sigma_n)$ error. This step is similar to binary search - we make use of $\log \frac{1}{\sigma_n}$ bits to locate θ to an interval whose length is $O(\sigma_n)$. This binary search could be trivial if $\sigma_n = 0$, i.e. at each machine we observe exact value of θ, where we could just query the first $\log \frac{1}{\sigma_n}$ digits of θ and do the job. However, if observations X_i are noisy Gaussian random variables, the task is not as easy. We use Gray codes with some additional denoising techniques (as shown in finer localization step) so that we achieve the same communication cost-efficiency as if there is no noise in observations.

2. Building on the location information, after first step, we are supposed to have located θ in an interval of length $O(\sigma_n)$. However, there is still a gap to the centralized rate $O(\sigma_n/\sqrt{m})$. The remaining $B - \log \frac{1}{\sigma_n}$ bits are used for refinement to further increase the accuracy of the estimator, trying to bridge the gap. In this refinement step, we query one bit from each machine (up to the communication budget), trying to tell the central machine whether the observation X_i is on the "left side" of the interval or the "right side" of the interval. Then based on these one-bit information from local machines, we generate a maximum likelihood estimate as the final MODGAME estimate $\hat{\theta}$.

Before describing the MODGAME procedure in detail, we define several useful functions that will be used to generate the transcripts. For any interval $[L, R]$, let $\tau_{[L,R]} : \mathbb{R} \to [L, R]$ be the truncation function defined by

$$\tau_{[L,R]}(x) = \begin{cases} L & \text{if} \quad x \leq L \\ x & \text{if} \quad L < x < R \\ R & \text{if} \quad x \geq R \end{cases} . \tag{3}$$

For any integer $k \geq 0$, denote $g_k : \mathbb{R} \to \{0, 1\}$ be the k-th Gray function defined by

$$g_k(x) \triangleq \begin{cases} 0 & \text{if} \quad \lfloor 2^k \tau_{[0,1]}(x) \rfloor \bmod 4 = 0 \text{ or } 3 \\ 1 & \text{if} \quad \lfloor 2^k \tau_{[0,1]}(x) \rfloor \bmod 4 = 1 \text{ or } 2. \end{cases}$$

Similarly we denote by $\bar{g}_k : \mathbb{R} \to \{0, 1\}$ the k-th conjugate Gray function defined by

$$\bar{g}_k(x) \triangleq \begin{cases} 0 & \text{if} \quad \lfloor 2^k \tau_{[0,1]}(x) \rfloor \bmod 4 = 0 \text{ or } 1 \\ 1 & \text{if} \quad \lfloor 2^k \tau_{[0,1]}(x) \rfloor \bmod 4 = 2 \text{ or } 3. \end{cases}$$

To unify the notation we set $g_k(x) \equiv \bar{g}_k(x) \equiv 0$ if $k < 0$.

It is worth mentioning that these Gray functions mimic the behavior of the Gray codes (for reference see Savage (1997)). Fix $K \geq 1$, if we treat $(g_1(x), g_2(x), ..., g_K(x))$ as a string of code for any source $x \in [0, 1]$, then those x within the interval $[2^{-K}(s-1), 2^{-K}s)$ where s is a integer will match the same code. Moreover, the code for adjacent intervals only differs by one bit, which is also a key feature for the Gray codes. This key feature guarantees the robustness of the Gray codes, where the Gray codes satisfy the so-called stability property for coding/decoding: There exist small numbers $\epsilon, \delta > 0$ such that, for any X and its

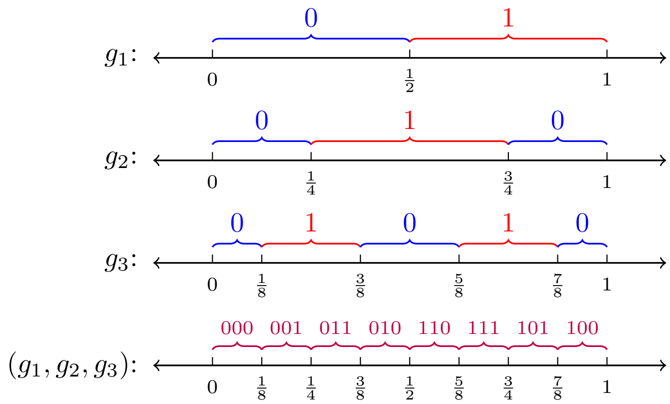


Figure 3: An illustration of the Gray functions and Gray codes.

perturbed observations $X_1, X_2, ..., X_k$ satisfying $|X_k - X| < \epsilon$ for all $k = 1, ..., K$. If $\tilde{X}$ makes $g_k(\tilde{X}) = g_k(X_k), k = 1, ..., m$, then we must have $|\tilde{X} - X| < \delta$.

Such stability property makes the Gray functions very useful for distributed estimation An example for $K = 3$ is shown in Figure 3 to better illustrate the behavior of the Gray functions.

Along with the figure, we also provide a simple example to show why the Gray codes are robust to stochastic errors. Suppose X_1, X_2, and X_3 are three i.i.d random variables with mean $1/4 + \epsilon$ and a small variance that is slightly larger than ϵ^2. The goal is to estimate their mean by one-bit measurement of each observation. By using the Gray codes, $(g_1(X_1), g_2(X_2), g_3(X_3))$ is equal to (001) or (011) with large probability, whose decoded interval $(1/8, 1/4)$ or $(1/4, 3/8)$ is close to $1/4$. As a contrast, if one uses the binary codes, the result will be unstable due to the stochastic error of X_2. In the MODGAME procedure, the Gray codes are used to help crudely "locate" the final estimator $\hat{\theta}_D$ to an interval of length $O(\sigma_n)$.

Define the refinement function $h(x) : \mathbb{R} \to \{0, 1\}$ and the conjugate refinement function $\bar{h}(x) : \mathbb{R} \to \{0, 1\}$ by

$$h(x) \triangleq \lfloor 2^{(\lfloor \log \frac{1}{\sigma_n} \rfloor - 7)} x \rfloor \bmod 2 \quad \text{and} \quad \bar{h}(x) \triangleq \lfloor 2^{(\lfloor \log \frac{1}{\sigma_n} \rfloor - 7)} x - \frac{1}{2} \rfloor \bmod 2. \tag{4}$$

For any function f, define the convolution function

$$\Phi_f(x) \triangleq \mathbb{E}_{X \sim N(x, \sigma_n^2)} f(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(y-x)^2}{2\sigma_n^2}} f(y) dy.$$

The above refinement functions and convolution function are used to accurately estimate the mean of the Gaussian observations. In the MODGAME procedure, the central machine collects one-bit measurements of some observations, say $h(X_1), h(X_2), ...h(X_u)$. By definition, the mean of those one-bit measurements is exactly $\Phi_h(\theta)$. Note that $\Phi_h(x)$ is a periodic wave-shape function, therefore after locating θ to a short interval of length $O(\sigma_n)$ during the preliminary steps, the central machine obtains a good estimate for θ by solving estimating equation $\Phi_h(\theta) = u^{-1} \sum_{i=1}^{u} h(X_i)$. A similar communication strategy is also adopted in Braverman et al. (2016).

For any $K \geq 1$, let $\mathrm{Dec}_K(y_1, y_2, ..., y_K) : \{0,1\}^K \to 2^{[0,1]}$ be the decoding function defined by

$$\mathrm{Dec}_K(y_1, y_2, ..., y_K) \triangleq \{x \in [0,1] : g_k(x) = y_k \quad \text{for } k = 1, 2, ..., K\}.$$

Last, we define the distance between a point $x \in \mathbb{R}$ and a set $S \subseteq \mathbb{R}$ as

$$d(x, S) \triangleq \min_{y \in S} |x - y|.$$

We are now ready to introduce the MODGAME procedure in detail. Again, we divide into three cases.

Case 1: $B < \log \frac{1}{\sigma_n} + 2$**.** In this case, the output is the values of the first B localization bits from the local machines, where the k-th localization bit is defined as the value of the function $g_k(\cdot)$ evaluated on the local sample. The procedure can be described as follows.

Step 1: *Generate transcripts on local machines.* Define $s_0 = 0$ and $s_i = \sum_{j=1}^{i} b_j$ for $i = 1, \ldots, m$. On the i-th machine, the transcript Z_i is concatenated by the $(s_{i-1}+1)$-th, $(s_{i-1}+2)$-th, ..., $(s_{i-1}+b_i)$-th Gray functions evaluated at X_i. That is,

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, ..., U_{s_{i-1}+b_i}),$$

where $U_{s_{i-1}+k} \triangleq g_{s_{i-1}+k}(X_i)$ for $k = 1, 2, ..., b_i$.

Step 2: *Construct distributed estimator $\hat{\theta}_D$.* Now we collect the bits $U_1, U_2, ..., U_B$ from the transcripts $Z_1, Z_2, ..., Z_m$. Note that U_k is the k-th Gray function evaluate at a random sample drawn from $N(\theta, \sigma_n^2)$, one may reasonably "guess" that $U_k \approx g_k(\theta)$. By this intuition, we set $\hat{\theta}_D$ to be the minimum number in the interval $\mathrm{Dec}_B(U_1, U_2, ..., U_B)$, i.e.

$$\hat{\theta}_D = \min\{x : x \in \mathrm{Dec}_B(U_1, U_2, ..., U_B)\}.$$

Case 2: $\log \frac{1}{\sigma_n} + 2 \leq B \leq \log \frac{1}{\sigma_n} + m$. Let

$$u \triangleq \max\left\{ s \in \mathbb{N} : \lfloor \log s \rfloor^2 + 2s \leq B - \lfloor \log \frac{1}{\sigma_n} \rfloor \right\}, \tag{5}$$

and define finer localization functions:

$$\begin{aligned} f_1(x) &\triangleq g_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 2}(x), \\ f_2(x) &\triangleq \bar{g}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 2}(x), \\ f_k(x) &\triangleq g_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 4 + k}(x) \text{ for } k \geq 3. \end{aligned} \tag{6}$$

In this case the total communication budget is divided into 3 parts: crude localization bits (roughly $\lfloor \log \frac{1}{\sigma_n} \rfloor$ bits), finer localization bits ($\lfloor \log u \rfloor^2$ bits), and refinement bits ($2u$ bits). The crude localization bits are the values of the functions $g_1(\cdot), g_2(\cdot), ..., g_{\lfloor \log \frac{1}{\sigma_n} \rfloor}(\cdot)$, each evaluated on a local sample. We denote those resulting binary bits by $U_1, U_2, ..., U_{\lfloor \log \frac{1}{\sigma_n} \rfloor}$.

The finer localization bits are the values of the functions $f_1(\cdot), f_2(\cdot), ..., f_{\lfloor \log u \rfloor}(\cdot)$, each function is evaluated on $\lfloor \log u \rfloor$ different local samples. The function values of $f_k(\cdot)$ are denoted by $W_{k,1}, W_{k,2}, ..., W_{k,\lfloor \log u \rfloor}$. The refinement bits are the values of the function $h(\cdot)$, evaluated on u local samples; and the values of the function $\bar{h}(\cdot)$, evaluated on u different local samples. The resulting binary bits are denoted by $V_1, V_2, ..., V_n$ and $\bar{V}_1, \bar{V}_2, ..., \bar{V}_n$ respectively.

These three types of bits are assigned to local machines by the following way: (1) Among all m machines, there are $\lfloor \log u \rfloor^2$ local machines who will output transcript consisting of 1 finer localization bit and $b_i - 1$ crude localization bits. (2) Among all m machines, there are $2u$ local machines who will output transcript consist of 1 refinement bit and $b_i - 1$ crude localization bits. (3) The remain $m - (\lfloor \log u \rfloor^2 + 2u)$ machines will output transcript consist of b_i crude localization bits. The above assignment is feasible because

$$\lfloor \log u \rfloor^2 + 2u \leq B - \lfloor \log \frac{1}{\sigma_n} \rfloor \leq m.$$

It is worth mentioning that every finer localization bits and every refinement bits are assigned to different machines. Intuitively, this is because we need these bits to be independent so that we can gain enough information for estimation. See Figure 4 for an overview of the MODGAME procedure. The procedure can be summarized as follows:

Step 1: *Generate transcripts on local machines.* Define $s_i = \sum_{j=1}^{i}(b_j - \mathbb{I}_{\{j \leq \lfloor \log u \rfloor^2 + 2u\}})$ and $s_0 = 0$. On the i-th machine:

- If $(j-1)\lfloor \log u \rfloor + 1 \leq i \leq j\lfloor \log u \rfloor$ for some integer $1 \leq j \leq \lfloor \log u \rfloor$, output

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, ..., U_{s_{i-1}+b_i-1}, W_{j,i-(j-1)\lfloor \log u \rfloor});$$

  (If $b_i = 1$, just output $Z_i = W_{j,i-(j-1)\lfloor \log u \rfloor}$.)
- If $\lfloor \log u \rfloor^2 + 1 \leq i \leq \lfloor \log u \rfloor^2 + u$, output

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, ..., U_{s_{i-1}+b_i-1}, V_{i-\lfloor \log u \rfloor^2});$$

  (If $b_i = 1$, just output $Z_i = V_{i-\lfloor \log u \rfloor^2}$.)
- If $\lfloor \log u \rfloor^2 + u + 1 \leq i \leq \lfloor \log u \rfloor^2 + 2u$, output

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, ..., U_{s_{i-1}+b_i-1}, \bar{V}_{i-(\lfloor \log u \rfloor^2+u)});$$

  (If $b_i = 1$, just output $Z_i = \bar{V}_{i-(\lfloor \log u \rfloor^2+u)}$.)
- If $i \geq \lfloor \log u \rfloor^2 + 2u + 1$, output

$$Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, ..., U_{s_{i-1}+b_i}).$$

The above binary bits are calculated by

$$\begin{aligned}
U_{s_{i-1}+k} &\triangleq g_{s_{i-1}+k}(X_i) \quad \text{for} \quad i = 1, 2, ..., m;\ k = 1, 2, ..., b_i; \\
W_{j,i-(j-1)\lfloor \log u \rfloor} &\triangleq f_j(X_i) \quad \text{for} \quad j = 1, 2, ..., \lfloor \log u \rfloor - 1; \\
&\qquad\qquad\qquad\quad i = (j-1)\lfloor \log u \rfloor + 1, (j-1)\lfloor \log u \rfloor + 2, ..., j\lfloor \log u \rfloor; \\
V_{i-\lfloor \log u \rfloor^2} &\triangleq h(X_i) \quad \text{for} \quad i = \lfloor \log u \rfloor^2 + 1, \lfloor \log u \rfloor^2 + 2, ..., \lfloor \log u \rfloor^2 + u; \\
\bar{V}_{i-(\lfloor \log u \rfloor^2+u)} &\triangleq \bar{h}(X_i) \quad \text{for} \quad i = \lfloor \log u \rfloor^2 + u + 1, ..., \lfloor \log u \rfloor^2 + 2u.
\end{aligned}$$

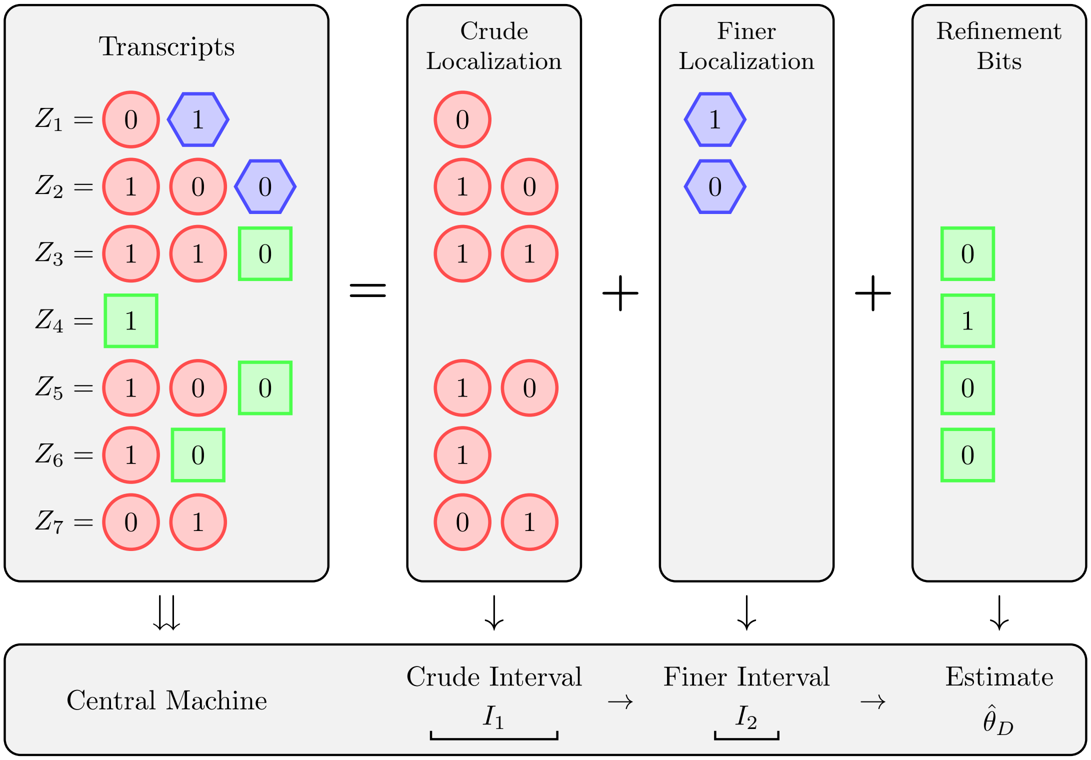


Figure 4: An illustration of MODGAME. The bits in the transcripts are transmitted to the central machine and divided into three types: crude localization bits, finer localization bits, and refinement bits. One then constructs on the central machine a crude interval I_1, a finer interval I_2, and the final estimate $\hat{\theta}_D$ step by step.

Step 2: *Construct distributed estimator $\hat{\theta}_D$.* From transcripts $Z_1, Z_2, ..., Z_m$, we can collect (a) crude localization bits $U_1, U_2, ..., U_{\lfloor \log \frac{1}{\sigma_n} \rfloor}$; (b) finer localization bits $W_{1,1}, W_{1,2}, ...,$ $W_{\lfloor \log u \rfloor, \lfloor \log u \rfloor}$; (c) refinement bits $V_1, V_2, ..., V_u$ and $\bar{V}_1, \bar{V}_2, ..., \bar{V}_u$.

Step 2.1: First, we use crude localization bits $U_1, U_2, ...U_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3}$ to roughly locate θ. The "crude interval" I_1 will be obtained in this step.

(a) If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \leq 3$, just set $I_1 = I_1' = [0, 1]$.

(b) If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \geq 4$, let

$$I_1' \triangleq \mathrm{Dec}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3}(U_1, U_2, ..., U_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3}). \tag{7}$$

Then we further stretch I_1' to a larger interval I_1 so that I_1 will double the length of I_1':

$$I_1 \triangleq \left\{ x : d(x, I_1') \leq 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 2)} \right\}. \tag{8}$$

Step 2.2: Then, we use finer localization bits to locate θ to a smaller interval of length roughly $O(\sigma_n)$. A "finer interval" I_2 will be generated in this step. For any

$1 \le k \le \lfloor \log u \rfloor$, let

$$W_k = \mathbb{I}_{\{\sum_{j=1}^{\lfloor \log u \rfloor} W_{k,j} \ge \frac{1}{2}\lfloor \log u \rfloor\}}$$

be the majority voting summary statistic for $W_{k,1}, W_{k,2}, ..., W_{k,\lfloor \log u \rfloor}$.

(a) If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \le 3$, and $\lfloor \log \frac{1}{\sigma_n} \rfloor \le 4$, let

$$I_2 = I_2' = [0, 1].$$

(b) If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \le 3$, and $\lfloor \log \frac{1}{\sigma_n} \rfloor \ge 5$, let

$$I_2' \triangleq \mathrm{Dec}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - 4}(W_{\lfloor \log u \rfloor - \lfloor \log \frac{1}{\sigma_n} \rfloor + 5}, W_{\lfloor \log u \rfloor - \lfloor \log \frac{1}{\sigma_n} \rfloor + 6}, ..., W_{\lfloor \log u \rfloor}). \tag{9}$$

Then we further stretch I_2' to a larger interval I_2 so that I_2 will double the length of I_2':

$$I_2 \triangleq \left\{ x : d(x, I_2') \le 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 3)} \right\}.$$

(c) If $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \ge 4$, let

$$I_2' \triangleq \{ x \in I_1 : f_k(x) = W_k \text{ for all } k = 1, 2, ..., \lfloor \log u \rfloor \}. \tag{10}$$

Lemma 21 in the Appendix shows I_2' is an interval. Then we further stretch I_2' to a larger interval I_2 so that I_2 will double the length of I_2':

$$I_2 \triangleq \left\{ x : d(x, I_2') \le 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 3)} \right\}.$$

Step 2.3: Finally, we use refinement bits $V_1, V_2, ..., V_u$ and $\bar{V}_1, \bar{V}_2, ..., \bar{V}_u$ to get an accurate estimate $\hat{\theta}_D$. Lemma 22 in the Appendix shows that one of the following two conditions must hold:

$$I_2 \subseteq \left[(2j - \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, (2j + \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)} \right] \text{ for some } j \in \mathbb{Z}$$

or

$$I_2 \subseteq \left[(2j + \frac{1}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, (2j + \frac{7}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)} \right] \text{ for some } j \in \mathbb{Z}.$$

So we can divide the procedure into the following two cases.

(a) If $I_2 \subseteq [(2j - \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, (2j + \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}]$ for some $j \in \mathbb{Z}$. Then $\Phi_h(x)$ is a strictly monotone function on I_2 (proved in Lemma 22 in the Appendix. Denote

$$L_I \triangleq \inf_{x \in I_2} \Phi_h(x) \quad \text{and} \quad R_I \triangleq \sup_{x \in I_2} \Phi_h(x).$$

By monotonicity, Φ_h is invertible on I_2. Let $\Phi_h^{-1} : [L_I, R_I] \to I_2$ be the inverse of Φ_h, the distributed estimator $\hat{\theta}_D$ is given by

$$\hat{\theta}_D = \Phi_h^{-1} \left(\tau_{[L_I, R_I]} \left(\frac{1}{u} \sum_{j=1}^{u} V_j \right) \right) \tag{11}$$

where $\tau_{[L_I,R_I]}$ is the truncation function defined in (3).

(b) Otherwise, we have $I_2 \subseteq [(2j+\frac{1}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, (2j+\frac{7}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}]$ for some $j \in \mathbb{Z}$. In this case $\Phi_{\bar{h}}(x)$ is a strictly monotone function on I_2 (proved in Lemma 22 in the Appendix. Denote

$$\bar{L}_I \triangleq \inf_{x \in I_2} \Phi_{\bar{h}}(x) \quad \text{and} \quad \bar{R}_I \triangleq \sup_{x \in I_2} \Phi_{\bar{h}}(x).$$

By monotonicity, $\Phi_{\bar{h}}$ is invertible on I_2. Let $\Phi_{\bar{h}}^{-1} : [\bar{L}_I, \bar{R}_I] \to I_2$ be the inverse of $\Phi_{\bar{h}}$, the distributed estimator $\hat{\theta}_D$ is given by

$$\hat{\theta}_D = \Phi_{\bar{h}}^{-1}\left(\tau_{[\bar{L}_I,\bar{R}_I]}\left(\frac{1}{u}\sum_{j=1}^{u}\bar{V}_j\right)\right) \tag{12}$$

where $\tau_{[\bar{L}_I,\bar{R}_I]}$ is the truncation function defined in (3).

Case 3: $B > \log\frac{1}{\sigma_n} + m$**.** We only need to use part of the total communication budget B as if we deal with the case $B = \lfloor \log\frac{1}{\sigma_n} \rfloor + m$. To be precise, we can always easily find $b_1', b_2', ..., b_m'$ so that $1 \le b_i' \le b_i$ for $i = 1, 2, ..., m$ and

$$\sum_{i=1}^{m} b_i' = \lfloor \log \frac{1}{\sigma_n} \rfloor + m.$$

Then we can implement the procedure introduced in Case 2 where we let the i-th machine only output a transcript of length b_i'.

2.2 Theoretical properties of the MODGAME procedure

Section 2.1 gives a detailed construction of the MODGAME procedure, which clearly satisfies the communication constraints by construction. The following result provides a theoretical guarantee for the statistical performance of MODGAME.

Theorem 1. *Suppose $\sigma_n < 1$. For given communication budgets $b_{1:m}$ with $b_i \ge 1$ for $i = 1, \ldots, m$, let $B = \sum_{i=1}^{m} b_i$ and let $\hat{\theta}_D$ be the MODGAME estimate. Then there exists a constant $C > 0$ such that*

$$\sup_{\theta\in[0,1]} \mathbb{E}(\hat{\theta}_D - \theta)^2 \le \begin{cases} C \cdot 2^{-2B} & \text{if } B < \log\frac{1}{\sigma_n} + 2 \\ C \cdot \frac{\sigma_n^2}{(B - \log\frac{1}{\sigma_n})} & \text{if } \log\frac{1}{\sigma_n} + 2 \le B < \log\frac{1}{\sigma_n} + m \\ C \cdot \frac{\sigma_n^2}{m} & \text{if } B \ge \log\frac{1}{\sigma_n} + m \end{cases}. \tag{13}$$

An interesting and somewhat surprising feature of the upper bound is that it depends on the communication constraints $b_{1:m}$ only through the total budget $B = \sum_{i=1}^{m} b_i$, not the specific value of $b_{1:m}$, so long as each machine can transmit at least one bit. This surprising phenomenon is possibly due to the symmetry among the local machines since samples on different machines are independent and identically distributed. The proof of Theorem 1 is provided in the Appendix.

2.3 Lower bound analysis and discussions

Section 2.1 gives a detailed construction of the MODGAME procedure and Theorem 1 provides a theoretical guarantee for the estimator. We shall now prove that MODGAME is indeed rate optimal among all estimators satisfying the communication constraints by showing that the upper bound in Equation (13) cannot be improved. More specifically, the following lower bound provides a fundamental limit on the estimation accuracy under the communication constraints.

Theorem 2. *Suppose $b_i \geq 1$ for all $i = 1, 2, ..., m$. Let $B = \sum_{i=1}^m b_i$. Then there exists a constant $c > 0$ such that*

$$R_1(b_{1:m}) \geq \begin{cases} c \cdot 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\ c \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\ c \cdot \frac{\sigma_n^2}{m} & \text{if } B \geq \log \frac{1}{\sigma_n} + m. \end{cases}$$

The key novelty in the lower bound analysis is the decomposition of the statistical risk into *localization* error and *refinement* error based on a delicate construction of the following candidate set G_δ:

$$G_\delta \triangleq \left\{ \theta_{u,v} = \sigma_n u + \delta v : u = 0, 1, 2, ..., \left(\lfloor \frac{1}{\sigma_n} \rfloor - 1 \right), v = 0, 1 \right\},$$

where δ is a precision parameter that will be specified later. By assigning θ a uniform prior on the candidate set G_δ, estimation of θ can be decomposed into estimation of u and v. One can view estimation of u as the *localization* step and estimation of v as the *refinement* step. The following lemma is a key technical tool.

Lemma 3. *Let $0 < \sigma_n < 1$ and let u be uniformly distributed on $\{0, 1, ..., \lfloor \frac{1}{\sigma_n} \rfloor - 1\}$ and v be uniformly distributed on $\{0, 1\}$. Let u and v be independent and let $\theta = \theta_{u,v} = \sigma_n u + \delta v$ where $0 < \delta < \frac{\sigma_n}{8}$. Then for all $\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})$,*

$$I(\hat{\theta}; u) + \frac{\sigma_n^2}{64\delta^2} I(\hat{\theta}; v) \leq B. \tag{14}$$

Remark 4. The proof of Lemma 3 mainly relies on the strong data processing inequality (Lemma 14 in Appendix). The strong data processing inequality was originally developed in information theory, for reference see Raginsky (2016). Zhang et al. (2013a) and Braverman et al. (2016) applied this technical tool to obtain lower bounds for distributed mean estimation. However, their lower bounds are not sharp when σ_n is very small, due to the fact that the focus was on bounding the refinement error using the strong data processing inequality, but failed to bound the localization error.

Lemma 3 suggests that under the communication constraints $b_{1:m}$, there is an unavoidable trade-off between the mutual information $I(\hat{\theta}; u)$ and $I(\hat{\theta}; v)$. So one of the above two quantities must be "small". When $I(\hat{\theta}; u)$ (or $I(\hat{\theta}; v)$) is smaller than a certain threshold, it can be shown that the estimator $\hat{\theta}$ cannot accurately estimate u (or v), which means the *localization* error (or the *refinement* error) is large. Given that one of *localization* error and *refinement* error must be larger than a certain value, the desired lower bound follows. A detailed proof of Theorem 2 is given in Section 7.

Minimax rate of convergence. Theorems 1 and 2 together yield a sharp minimax rate for distributed univariate Gaussian mean estimation with independent protocols when $\sigma_n < 1$:

$$\inf_{\hat{\theta}\in\mathcal{A}_{ind}(b_{1:m})} \sup_{\theta\in[0,1]} \mathbb{E}(\hat{\theta}-\theta)^2 \asymp \begin{cases} 2^{-2B} & \text{if } B < \log\frac{1}{\sigma_n}+2 \\ \frac{\sigma_n^2}{(B-\log\frac{1}{\sigma_n})} & \text{if } \log\frac{1}{\sigma_n}+2 \le B < \log\frac{1}{\sigma_n}+m \\ \frac{\sigma_n^2}{m} & \text{if } B \ge \log\frac{1}{\sigma_n}+m \end{cases}. \tag{15}$$

The results also show that MODGAME is rate optimal.

The minimax rate only depends on the total communication budgets $B = \sum_{i=1}^m b_i$. As long as each transcript contains at least one bit, how these communication budgets are allocated to local machines does not affect the minimax rate. This surprising phenomenon is due to the symmetry among the local machines since samples on different machines are independent and identically distributed.

Remark 5. Figure 2 gives an illustration for the minimax rate (15), which is divided into three phases: localization, refinement, and optimal-rate. The minimax risk decreases quickly in the localization phase, when the communication constraints are extremely severe; then it decreases slower in the refinement phase, when there are more communication budgets; finally the minimax rate coincides with the centralized optimal rate (Bickel, 1981) and stays the same, when there are sufficient communication budgets. The value for each additional bit decreases as more bits are allowed.

In the localization phase, the risk is reduced to as small as $O(\sigma_n^2)$, which can be achieved by using the sample on only ONE machine and there is no need to "communicate" with multiple machines. In the refinement phase, the risk is further reduced to $O(\sigma_n^2/m)$. However, one must aggregate information from all machines in order to achieve this rate.

Remark 6. If the central machine itself also has an observation, or equivalently if one of the local machines serves as the central machine, then the communication constraints can be viewed as one of b_i is equal to infinity. This setting is considered in some related literature, for instance, see Jordan et al. (2019). Then according to Theorem 1, MODGAME always achieves the centralized rate $\frac{\sigma_n^2}{m}$, as long as at least one bit is allowed to communicate with each local machine.

Remark 7. Our analysis on the minimax rate can be generalized to the l_r loss for any $r \ge 1$, with suitable modifications on both the lower bound analysis and optimal procedure.

2.4 Optimal procedure when $\sigma_n \ge 1$

For completeness, we briefly discuss the case $\sigma_n \ge 1$. When $\sigma_n \ge 1$, each machine only need to output a one-bit measurement to achieve the global optimal rate as if there are no communication constraints. Therefore, the statistical upper bound automatically matches the trivial lower bound. Some related results are available in Kipnis and Duchi (2019). The following procedure is based on the setting when $b_i = 1$ for all $i = 1, ..., m$. If $b_i > 1$ for some i, then one can simply discard all remaining bits so that only one bit is sent by each machine.

Here is the procedure when $\sigma_n \geq 1$:

Step 1. The i-th machine outputs

$$Z_i = \begin{cases} 0 & \text{if } X_i < 0 \\ 1 & \text{if } X_i \geq 0 \end{cases}.$$

Step 2. The central machine collects $Z_1, Z_2, ..., Z_m$ and estimates θ by

$$\hat{\theta}_D = \tau_{[0,1]}\left(\sigma_n \Phi^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i\right)\right)$$

where τ is the truncation function defined in (3) and Φ is the cumulative distribution function of a standard normal, $\Phi(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{t^2/2}dt$. Here Φ^{-1} is the inverse of Φ and we extend it by defining $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$.

3. Distributed Multivariate Gaussian Mean Estimation

We turn in this section to distributed estimation of a multivariate Gaussian mean under the communication constraints. Similar to the univariate case, suppose we observe on m local machines i.i.d. random samples:

$$X_i \overset{\text{iid}}{\sim} N_d(\theta, \sigma_n^2 I_d), \quad i = 1, \ldots, m,$$

where the i-th machine has access to X_i only. Here we consider the multivariate Gaussian location family

$$\mathcal{P}_{\sigma_n}^d = \left\{N_d(\theta, \sigma_n^2 I_d) : \theta \in [0,1]^d\right\},$$

where $\theta \in [0,1]^d$ is the mean vector of interest and the noise level σ_n is known. Under the constraints on the communication budgets $b_{1:m}$ with $b_i \geq 1$ for $i = 1, \ldots, m$, the goal is to optimally estimate the mean vector θ under the squared error loss. We are interested in the minimax risk for the distributed protocol $\mathcal{A}_{ind}(b_{1:m})$:

$$R_d(b_{1:m}) = \inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_{1:m})} \sup_{\theta \in [0,1]^d} \mathbb{E}\|\hat{\theta} - \theta\|^2.$$

Another goal is to develop a rate-optimal estimator that satisfies the communication constraints. The multivariate case is similar to the univariate setting, but it also has some distinct features, both in terms of the estimation procedure and the lower bound argument. In this section, we still assume $\sigma_n < 1$.

3.1 Lower bound analysis

We first obtain the minimax lower bound which is instrumental in establishing the optimal rate of convergence. The following lower bound on the minimax risk shows a fundamental limit on the estimation accuracy when there are communication constraints. In view of the upper bound to be given in Section 3.2 that is achieved by a generalization of the MODGAME procedure, the lower bound is rate optimal.

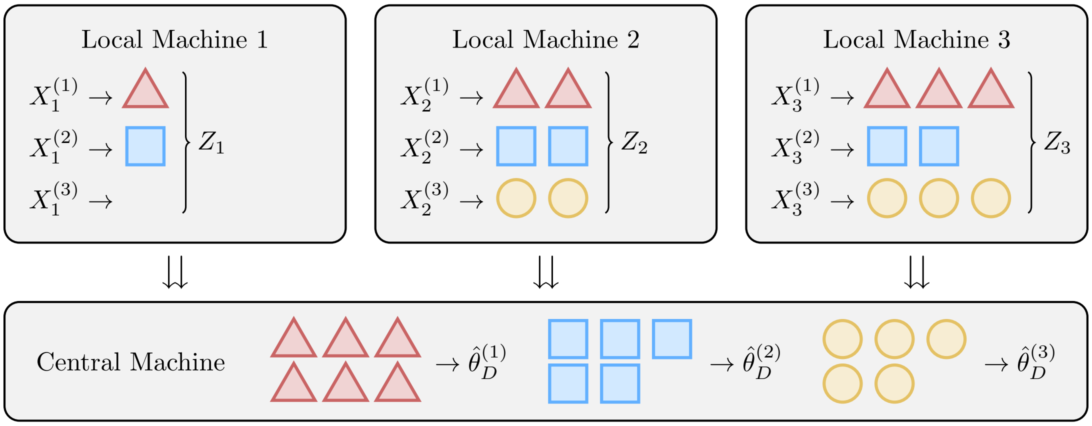


Figure 5: An illustration for multi-MODGAME. Communication budgets are evenly divided into three parts with each part used for estimating a coordinate of θ by the MODGAME procedure.

Theorem 8. *Suppose $b_i \geq 1$ for all $i = 1, 2, ..., m$. Let $B = \sum_{i=1}^m b_i$ and $m' = \frac{1}{d}\sum_{i=1}^m (b_i \wedge d)$, then there exists a constant $c > 0$ such that*

$$R_d(b_{1:m}) \geq \begin{cases} c \cdot 2^{-2B/d} d & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\ c \cdot \frac{d\sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \\ c \cdot \frac{d\sigma_n^2}{m'} & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2) \end{cases} .$$

A detailed proof of Theorem 8 is given in the Appendix.

Remark 9. In the earlier work including Garg et al. (2014); Barnes et al. (2019), a lower bound for distributed Gaussian mean estimation has been established as $\Omega(\frac{\sigma_n^2 d^2}{B})$, where B is the total communication cost. This lower bound is sharp for $\sigma_n \geq 1$. However, when $\sigma_n < 1$, by showing that the additional and exact $\log(1/\sigma_n)$ localization bits are necessary for estimating a Gaussian mean, the lower bound can be improved to $\Omega(\min\{\frac{\sigma_n^2 d^2}{B - d\log(1/\sigma_n)}, \sigma_n^2 d\})$. The improvement is significant when $\log(1/\sigma_n)/m$ is bounded away from 0.

3.2 Optimal procedure

We now construct an estimator of the mean vector under the communication constraints. Roughly speaking, the procedure, called multi-MODGAME, first divides the communication budgets evenly into d parts and then each part of communication budgets will be used to estimate one coordinate of θ. Our analysis shows that multi-MODGAME achieves the minimax optimal rate under the communication constraints. The construction of the distributed estimator $\hat{\theta}_D$ is divided into three steps.

Step 1: *Assign communication budgets.* In this step we will calculate $b_i^{(k)}$ $(i = 1, 2, ..., m; k = 1, 2, ..., d)$ so that

$$b_i = b_i^{(1)} + b_i^{(2)} + ... + b_i^{(d)} \quad \text{for all } i = 1, 2, ...m.$$

where $b_i^{(k)}$ is the number of bits within the transcript Z_i which is associated with estimation of $\hat{\theta}^{(k)}$.

Without loss of generality we assume $b_1 \leq b_2 \leq ... \leq b_m$, which can always be achieved by permuting the indices of the machines. Write $1, 2, 3, ..., d$ repeatedly to form a sequence:

$$Q \triangleq 1, 2, 3, ..., d, 1, 2, 3, ..., d, 1, 2, 3, ...$$

The sequence Q is then divided into subsequences of lengths $b_1, b_2, ..., b_m$. Let Q_1 be the subsequence of Q from index 1 to index b_1; let Q_2 be the next subsequence from index $b_1 + 1$ to $b_1 + b_2$; ... let Q_m be the subsequence from index $\sum_{i=1}^{m-1} b_i + 1$ to $\sum_{i=1}^{m} b_i$. For each $1 \leq k \leq d$, let $b_i^{(k)}$ be the number of occurrence of k within Q_i. To be more precise, $b_i^{(k)}$ can be calculated by

$$b_i^{(k)} = \left\lfloor \frac{\sum_{j=1}^{i} b_j - k}{d} \right\rfloor - \left\lfloor \frac{\sum_{j=1}^{i-1} b_j - k}{d} \right\rfloor .$$

Step 2: *Generate transcripts on local machines.* On the i-th machine, the transcript Z_i is concatenated by short transcripts $Z_i^{(1)}, Z_i^{(2)}, ..., Z_i^{(d)}$, where the length of $Z_i^{(k)}$ is $b_i^{(k)}$ for $k = 1, 2, ..., d$. Note that the k-th coordinate of the observations on each machine, $X_1^{(k)}, X_2^{(k)}, ..., X_m^{(k)}$, can be viewed as i.i.d univariate Gaussian variables with mean $\theta^{(k)}$ and variance σ_n^2. For $1 \leq k \leq d$, the transcripts $Z_1^{(k)}, Z_2^{(k)}, ..., Z_m^{(k)}$ can be generated the same way as if we implement MODGAME to estimate $\theta^{(k)}$ from observations $X_1^{(k)}, X_2^{(k)}, ..., X_m^{(k)}$, within the communication budgets $b_1^{(k)}, b_2^{(k)}, ..., b_m^{(k)}$. Some machines may be assigned zero communication budget, if that happens those machines are ignored and the procedure is implemented as if there are fewer machines.

Step 3: *Construct distributed estimator $\hat{\theta}_D$.* We have collected $Z_i^{(k)}$ $(i = 1, 2, ..., m; k = 1, 2, ..., d)$ from the m local machines. For $1 \leq k \leq d$, as in MODGAME, one can use $Z_1^{(k)}, Z_2^{(k)}, ..., Z_m^{(k)}$ to obtain an estimate for $\hat{\theta}^{(k)}$:

$$\hat{\theta}_D^{(k)} = \hat{\theta}_D^{(k)} \left(Z_1^{(k)}, Z_2^{(k)}, ..., Z_m^{(k)} \right) .$$

The final multi-MODGAME estimator $\hat{\theta}_D$ of the mean vector θ is just the vector consisting of the estimates for the d coordinates:

$$\hat{\theta}_D \triangleq \left(\hat{\theta}_D^{(1)}, \hat{\theta}_D^{(2)}, ..., \hat{\theta}_D^{(d)} \right) .$$

The following result provides a theoretical guarantee for multi-MODGAME.

Theorem 10. *Let $B = \sum_{i=1}^{m} b_i$ and $m' = \frac{1}{d} \sum_{i=1}^{m} (b_i \wedge d)$. Then there exists a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]^d} \mathbb{E}\|\hat{\theta}_D - \theta\|^2 \leq \begin{cases} C \cdot 2^{-2B/d} d & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\ C \cdot \frac{d\sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \\ C \cdot \frac{d\sigma_n^2}{m'} & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2). \end{cases} \quad (16)$$

Remark 11. Compared to the state-of-art results in the literature including Braverman et al. (2016), the multi-MODGAME procedure is more communication-efficient and more flexible in communication budget allocation. To be specific, the algorithm proposed in Braverman et al. (2016) achieves the mean squared error $O(\frac{\sigma_n^2 d}{\alpha m})$ with the total communication cost of order $\alpha m d + d\log^2(\alpha m d/\sigma_n))$. In comparison, to achieve the same statistical performance, MODGAME only needs $\alpha m d + d\log(1/\sigma_n)$ bits. The difference could be significant when $\sigma_n \ll 1$.

Moreover, multi-MODGAME achieves the optimal statistical performance in the distributed setting with any pre-specified communication budget allocation $(b_1, b_2, ..., b_m)$. That is, the constraint is imposed on each individual local machine. In comparison, the protocol in Braverman et al. (2016) assigns the total communication budget by the algorithm thus in a way solves a simpler "total communication constrained" problem.

The lower and upper bounds given Theorems 8 and 10 together establish the minimax rate for distributed multivariate Gaussian mean estimation:

$$\inf_{\hat\theta\in\mathcal{A}_{ind}(b_{1:m})} \sup_{\theta\in[0,1]^d} \mathbb{E}\|\hat\theta-\theta\|^2 \asymp \begin{cases} 2^{-2B/d}d & \text{if } B/d < \log\frac{1}{\sigma_n}+2 \\ \frac{d\sigma_n^2}{(B/d-\log\frac{1}{\sigma_n})} & \text{if } \log\frac{1}{\sigma_n}+2 \le B/d < \log\frac{1}{\sigma_n} + (m'\vee 2) \\ \frac{d\sigma_n^2}{m'} & \text{if } B/d \ge \log\frac{1}{\sigma_n} + (m'\vee 2) \end{cases} \tag{17}$$

where $B = \sum_{i=1}^m b_i$ is the total communication budget and $m' = \frac{1}{d}\sum_{i=1}^m (b_i \wedge d)$ is the "effective sample size". In particular, the minimax rate (15) for the univariate case is an special case for the above minimax rate (17) with $d = 1$.

Remark 12. Different from the univariate case, in the multivariate case the minimax rate depends on not only the total communication budget B, but also the effective sample size m'. How the communication budgets assigned to individual local machines affects the difficulty of the estimation problem. If the communication budgets are tight on some machines, then one may have $m' \ll m$, which means the centralized minimax rate cannot be achieved even if the total communication budget B is sufficiently large.

Remark 13. The present paper focuses on the unit hypercube $[0, 1]^d$ as the parameter space. A similar analysis can be applied to other "regular" shape constraints, such as a ball or a simplex, and the minimax rate depends on the constraint.

4. Robustness Against Departures from Gaussianity

We have so far focused exclusively on the Gaussian location families. Both the optimal distributed procedures and lower bound arguments are established under the assumption of Gaussian observations. We consider in this section robustness of the proposed MODGAME and multi-MODGAME procedures against departures from Gaussianity.

Even if the i.i.d observations $X_{i,j}, i = 1, 2, .., m, j = 1, 2, ..., n$ are drawn from a non-Gaussian distribution, after taking the sample mean on each local machine, according to the central limit theorem, the distribution of these sample means is close to a Gaussian distribution when n is large. Thus intuitively the proposed procedures should still work even when the original observations are nongaussian.

For simplicity we focus on the one-dimensional estimation problem. The multivariate case can be considered as a direct generalization to the univariate case. Let P_θ be a location family where θ is the mean, and its variance is σ^2. Denote $\bar{P}_\theta^n$ as the distribution of the mean of n *i.i.d* copies drawn from P_θ. If on each local machine we can access to n i.i.d observations $X_{i,1}, X_{i,2}, ..., X_{i,n} \sim P_\theta$, then each machine can take the local sample mean $X_i \triangleq \sum_{j=1}^n X_{i,j} \sim \bar{P}_\theta^n$. Even though $\bar{P}_\theta^n$ is a non-Gaussian distribution, the MODGAME procedure can take X_i as inputs to generate a final estimate.

Recall that MODGAME is divided into three steps: crude localization step, finer localization step, and refinement step. During the first two steps, in order to obtain the desired statistical guarantee for the "confidence interval" I_2, we only need sub-Gaussian tail condition for X_i. During the refinement step, the key is to use Φ_h or $\Phi_{\bar{h}}$ to generate estimates from the one-bit measurements. If X_i is not drawn from a Gaussian distribution, there is additional bias that could be controlled under certain conditions.

Let $TV(\cdot, \cdot)$ denote the total variation distance between two probability distributions. A random variable X (or a distribution P where $X \sim P$) is called v-subgaussian if $\mathbb{E}\exp(s(X - \mathbb{E}X)) \le \exp(\frac{v^2 s^2}{2}),\ \forall s \in \mathbb{R}$. The following theorem shows that when the total variation distance between the distribution $\bar{P}_\theta^n$ of the local sample mean and the Gaussian distribution $N(\theta, \sigma_n^2)$ is sufficiently small, MODGAME has the same theoretical guarantee as in the Gaussian case. This implies that MODGAME is robust against departures from the Gaussian distribution.

Theorem 14. *Suppose $\sigma_n < 1$. If $\bar{P}_\theta^n$ is a $D\sigma_n$-subgaussian distribution and $TV(\bar{P}_\theta^n, N(\theta, \sigma_n^2)) \le \frac{D}{\sqrt{m}}$ for some $D > 0$. Then there exists a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \le C \cdot \begin{cases} 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\ \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \le B < \log \frac{1}{\sigma_n} + m \\ \frac{\sigma_n^2}{m} & \text{if } B \ge \log \frac{1}{\sigma_n} + m \end{cases}. \tag{18}$$

where $\hat{\theta}$ is the output of the MODGAME procedure and $B = \sum_{i=1}^m b_i$ is the total communication cost.

A sketch of the proof is given in the Appendix. Note that $X_i \sim \bar{P}_\theta^n$ is the mean of i.i.d observations in the ith local machine. The L_1 Berry-Esseen bound (e.g. (Chen et al., 2010, Corollary 4.2)) suggests $TV(\bar{P}_\theta^n, N(\theta, \sigma_n^2)) \le \frac{\mathbb{E}(|X_1 - \theta|/\sigma)^3}{2\sqrt{n}}$. If X_1 is a $D\sigma$-subgaussian distribution, then $\mathbb{E}(|X_1 - \theta|/\sigma)^3$ is bounded by a constant (depending on D). Hence the following corollary holds.

Corollary 15. *Suppose $\sigma_n < 1$. If P_θ is a $D\sigma$-subgaussian distribution, and $m \le Dn$ for some $D > 0$. Then there exist a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \le C \cdot \begin{cases} 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\ \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \le B < \log \frac{1}{\sigma_n} + m \\ \frac{\sigma_n^2}{m} & \text{if } B \ge \log \frac{1}{\sigma_n} + m \end{cases}. \tag{19}$$

where $\hat{\theta}$ is the output of the MODGAME procedure. $B = \sum_{i=1}^m b_i$ is the total communication cost.

Corollary 15 shows that, if n/m is asymptotically bounded away from 0, then MODGMAE achieves the same statistical performance as in the Gaussian case as long as the observations are drawn from a subgaussian distribution.

Theorem and Corollary 15 are stated in univariate Gaussian mean estimation setting. It is not difficult to extend the current result to multivariate setting since multi-MODGAME procedure can be viewed as an aggregation of MODGAME procedures in each dimension.

5. Simulation Studies

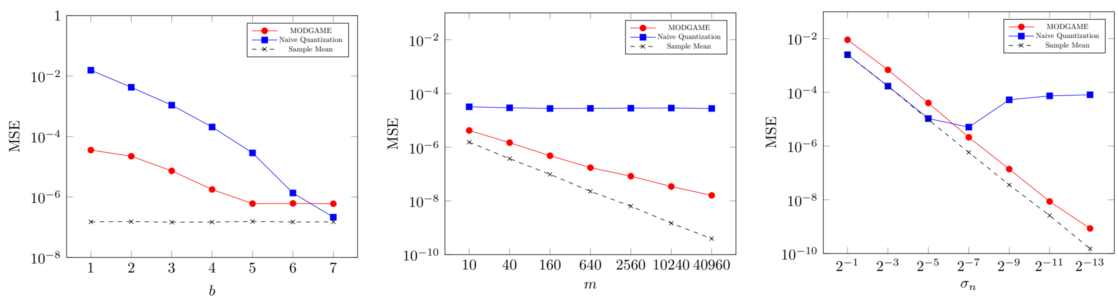
It is clear by construction that MODGAME and multi-MODGAME satisfy the communication constraints and are easy to implement. We investigate in this section their numerical performance through simulation studies. Comparisons with the existing methods are given and the results are consistent with the theory. In this section, we implement a slightly modified version of MODGAME procedure, where each local machine output three refinement bits instead of one. This slightly modified MODGAME procedure has better numerical performance and also has the same theoretical guarantee as what is stated in Section 2.

We first consider MODGAME for estimating a univariate Gaussian mean. In this case, we set $d = 1$ and $b_1 = b_2 = ... = b_m = b$, i.e. the communication budgets for all machines are equal, and compare the empirical MSEs of MODGAME, naive quantization (see e.g. Zhang et al. (2013a)), and sample mean. For naive quantization, each machine projects its observation to $[0, 1]$ and quantizes it to precision 2^{-b}. The quantized observation is sent to the central machine and the central machine uses their average as the final estimate. The sample mean is the efficient estimate when there are no communication constraints, which can be viewed as a benchmark for any distributed Gaussian mean estimation procedure.

First, we fix $m = 100$, $\sigma_n = 2^{-8}$ and assign the communication budget for each machine b from 1 to 7. The MSEs of the three estimators are shown in Figure 6a, which shows that MODGAME makes better use of the communication resources in comparison to naive quantization. It can be seen from the figure, MODGAME outperforms naive quantization when the communication constraints are extremely severe. As the communication budgets increases, naive quantization can nearly achieve the optimal MSE, meanwhile MODGAME still performs very well.

In the second setting, we fix $\sigma_n = 2^{-8}$, $b = 5$ and vary the number of machines m from 10 to 40960. Figure 6b plots the MSEs of the three methods. The MSE of MODGAME decreases as number of machine increases and outperforms naive quantization; the MSE of naive quantization remains constant as the quantization error plays a dominant role in the MSE.

Finally, we fix $b = 5$, $m = 100$ and vary the standard deviation σ_n from 2^{-1} to 2^{-13}. Figure 6c shows the MSEs of the three estimators. It can be seen that MODGAME is robust for all choices of σ_n. The difference between the MSE of MODGAME and the optimal MSE for non-distributed sample mean is small. For naive quantization, it is as good as the optimal non-distributed sample mean when σ_n is large. However, as seen in the previous experiment, when σ_n is small, the MSE of naive quantization is dominated by the quantization error and is much larger than the MSE of MODGAME. In all three settings, it can be seen clearly that the MSE of MODGAME decreases as the communication



(a) Fixed m and σ_n (b) Fixed b and σ_n (c) Fixed b and m

Figure 6: Comparisons of the MSEs of MODGAME (red), naive quantization (blue) and sample mean (black). MSEs are plotted on log-scale. In 6b and 6c, m and σ_n are plotted on log-scale.

budgets increases. This is consistent with the theoretical results established in Section 2 and demonstrates the tradeoff between the communication costs and statistical accuracy.

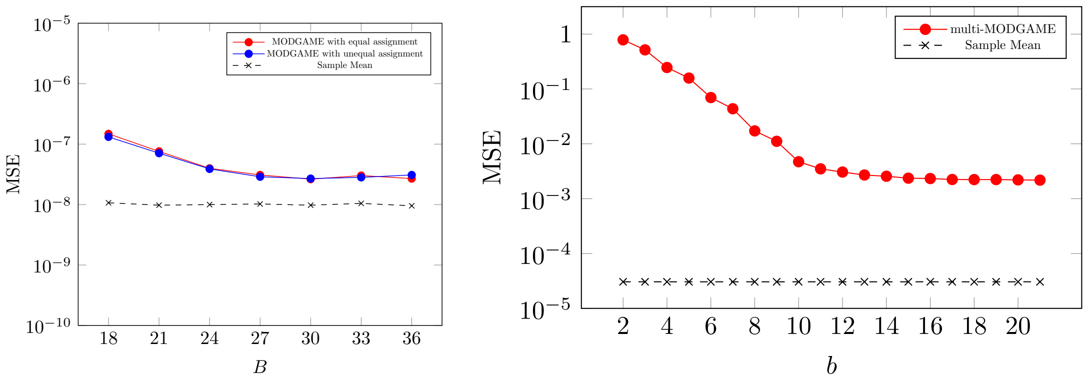
Besides, to demonstrate that the performance of the MODGAME procedure only depends on total communication budget B, we implement another simulation. We fix $m = 6$, $\sigma_n = 2^{-12}$ and assign the total communication budgets B from 18 to 36. We compare the performance of the MODGAME procedure with different communication allocation. That is, in one simulation we assign $b_i = 3$ bits to each local machine except one, and that one machine are assigned $B - 3(m-1)$ bits. In another simulation we assign equal communication budget $b_i = B/m$ to each machine. As a benchmark, we also implement non-distributed sample mean estimator. Figure 7a shows the MSEs of the above three methods. It is shown clearly that how communication budgets are assigned to local machines doesn't affect the performance of the MODGAME procedure, which is consistent with our theory.

We now turn to multi-MODGAME. Different values of the dimension d yield similar phenomena. We use $d = 50$ here for illustration. When d is larger than the number of bits that is allowed to communicate on each machine, naive quantization is not valid as it is unclear how to quantize the d coordinates of the observed vector. As a comparison, it can be seen in the following experiments that multi-MODGAME still performs well even if d is large and the communication budgets are tight.

Same as before, we set $b_1 = b_2 = ... = b_m = b$, i.e. the communication budgets for all machines are equal. We set $d = 50$, $\sigma_n = 2^{-8}, m = 25$ and assign the communication budgets b for each machine from 2 to 21. The MSEs of different methods are shown in Figure 7b. A phase transition at $b = 10$ can be clearly seen. When $b \le 10$, the MSE decreases quickly at an exponential rate. When $b > 10$, the decrease becomes relatively slow. This phenomenon is consistent with the theoretical prediction that different phases appear in the convergence rate for multi-MODGAME (Theorem 10).

6. Discussion

We established in the present paper the sharp minimax optimal rate that holds for all values of the parameters d, m, n, and σ in all communication budget regimes under the



(a) Different allocation

(b) Multi-MODGAME

Figure 7: Left panel: Comparisons of the MSEs of MODGAME with equal assignment (red), MODGAME with unequal assignment (blue) and sample mean (black). Right panel: Comparisons of the MSEs of multi-MODGAME (red) and sample mean (black). MSEs are plotted on log-scale.

independent protocol. A key technique is the decomposition of the minimax estimation problem into two steps, *localization* and *refinement*, which appears in both the lower bound analysis and optimal procedure design. The optimality results and techniques developed can be useful for solving other problems such as distributed nonparametric function estimation and distributed sparse signal recovery.

In spite of these optimality results, there are still several open problems on distributed Gaussian mean estimation. For example, an interesting problem is the optimal estimation of the mean θ when the variance σ^2 is unknown. The lack of knowledge of σ^2 requires additional communication efforts for optimally estimating θ. When there are more than one sample available on each local machine, a natural approach is to estimate σ^2 on each local machine and then use MODGAME to estimate θ. It would be interesting to investigate the performance of such an estimator. Other than estimating the mean θ, distributed estimation of the variance σ^2 is also an interesting and important problem. When there are multiple samples on each local machine, the local estimate of σ^2 can be viewed as an observation drawn from a scaled χ^2 distribution. The problem then becomes a distributed χ^2 estimation problem and it might be solved by using a similar approach to the one used in the present paper. We leave these for future work.

Optimal estimation of the mean of a multivariate Gaussian distribution with a general (known) covariance matrix is another interesting problem. A naive approach is to ignore the dependency and apply MODGAME to estimate the coordinates individually, this is arguably not communication efficient in general. For instance, if the correlation between certain coordinates is large, it may be possible to save a significant amount of communication budget by utilizing the information from one coordinate to help estimate the other. Another approach is to use multi-MODGAME after orthogonalization. More specifically, consider the Gaussian location family with a general non-singular covariance matrix Σ. Let $\lambda_{\min} > 0$ be the smallest eigenvalue of Σ. For $X \sim N_d(\theta, \Sigma)$, $\lambda_{\min}^{1/2}(d\Sigma)^{-1/2}X \sim N_d\left(\lambda_{\min}^{1/2}(d\Sigma)^{-1/2}\theta, \frac{\lambda_{\min}}{d} I_d\right)$.

Note that $\lambda_{\min}^{1/2}(d\Sigma)^{-1/2}\theta \in [0,1]^d$ for any $\theta \in [0,1]^d$, therefore one can apply multi-MODGAME to estimate $\lambda_{\min}^{1/2}(d\Sigma)^{-1/2}\theta$, then transform it back to get an estimate for θ. However, this is generally not rate-optimal. A systematic study is needed for this problem. Another related and more challenging problem is optimal distributed estimation of the covariance matrix Σ.

This paper arguably considered one of the simplest settings for optimal distributed estimation under the communication constraints, but as can be seen in the paper, both the construction of the rate optimal estimators and the theoretical analysis are already quite involved for such a seemingly simple problem. As we deepen our understanding on distributed learning under the communication constraints, we hope to extent this line of work to investigate other statistical problems in distributed settings, including nonparametric function estimation, high-dimensional linear regression, and large-scale multiple testing.

We hope that the results and techniques developed in this paper serve as a starting point for developing optimality theories for other distributed learning problems.

7. Proofs

In this section we prove Theorem 2 for the univariate case. For reasons of space, Theorems 1, 8, 10, 4 and the technical lemmas are proved in the Appendix.

We prove separately the three cases in Theorem 2: $B < \log\frac{1}{\sigma_n} + 2$, $\log\frac{1}{\sigma_n} + 2 \le B < \log\frac{1}{\sigma_n} + m$, and $B \ge \log\frac{1}{\sigma_n} + m$. We first focus on the most important case $\log\frac{1}{\sigma_n} + 2 \le B < \log\frac{1}{\sigma_n} + m$. New technical tools are developed in the proof. The other two cases are relatively easy.

Case 1: $\log\frac{1}{\sigma_n} + 2 \le B < \log\frac{1}{\sigma_n} + m$. Note that $b_i \ge 1$ for all $i = 1, 2, ..., m$ implies that $B = \sum_{i=1}^m b_i \ge m$. Therefore in this case we must have $\sigma_n < 1$.

Let $0 < \delta < \frac{1}{8}\sigma_n$ be a parameter to be specified later. Define a grid of candidate values of θ as

$$G_\delta \triangleq \left\{\theta_{u,v} = \sigma_n u + \delta v : u = 0, 1, 2, ..., \left(\lfloor\frac{1}{\sigma_n}\rfloor - 1\right), v = 0, 1\right\}. \tag{20}$$

Let $\mathbb{U}(G_\delta)$ be a uniform prior of θ on G_δ. Note that $G_\delta \subset [0,1]$, so the minimax risk is lower bounded by the Bayesian risk:

$$\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})} \sup_{\theta\in[0,1]} (\hat{\theta} - \theta)^2 \ge \inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})} \mathbb{E}_{\theta\sim\mathbb{U}(G_\delta)} (\hat{\theta} - \theta)^2. \tag{21}$$

For any estimator $\hat{\theta} \in \mathcal{A}(b_{1:m})$, the rounded estimator $\hat{\theta}' \triangleq \mathrm{argmin}_{\tilde{\theta}\in G_\delta} |\tilde{\theta} - \hat{\theta}|$ always satisfy $(\hat{\theta} - \theta)^2 \ge \frac{1}{4}(\hat{\theta}' - \theta)^2$ for all $\theta \in G_\delta$. Note that $\hat{\theta}'$ also belongs to the protocol class $\mathcal{A}(b_{1:m})$, and only takes value in G_δ, this implies

$$\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})} \mathbb{E}_{\theta\sim\mathbb{U}(G_\delta)} (\hat{\theta} - \theta)^2 \ge \frac{1}{4} \inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})\cap G_\delta} \mathbb{E}_{\theta\sim\mathbb{U}(G_\delta)} (\hat{\theta} - \theta)^2, \tag{22}$$

where $\mathcal{A}(b_{1:m}) \cap G_\delta$ is a shorthand for $\mathcal{A}(b_{1:m}) \cap \{\hat{\theta} : \hat{\theta} \text{ only takes value in } G_\delta\}$.

Now we have $\hat{\theta}, \theta \in G_\delta$ thus they can be reparametrized by $\hat{\theta} = \theta_{\hat{u},\hat{v}}$ and $\theta = \theta_{u,v}$. It is easy to verify the inequality

$$(\hat{\theta}_{\hat{u},\hat{v}} - \theta_{u,v})^2 \geq \max\left\{\frac{\sigma_n^2}{4}(\hat{u}-u)^2, \delta^2 \mathbb{I}_{\{\hat{v}\neq v\}}\right\}.$$

Hence

$$\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})\cap G_\delta} \mathbb{E}_{\theta\sim\mathbb{U}(G_\delta)}(\hat{\theta}-\theta)^2 \geq \inf_{\theta_{\hat{u},\hat{v}}\in\mathcal{A}(b_{1:m})\cap G_\delta} \mathbb{E}_{\theta_{u,v}\sim\mathbb{U}(G_\delta)} \max\left\{\frac{\sigma_n^2}{4}(\hat{u}-u)^2, \delta^2 \mathbb{I}_{\{\hat{v}\neq v\}}\right\}. \tag{23}$$

Putting together (21), (22), and (23), we have

$$\begin{aligned}\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})} \sup_{\theta\in[0,1]} (\hat{\theta}-\theta)^2 &\geq \frac{1}{4} \inf_{\theta_{\hat{u},\hat{v}}\in\mathcal{A}(b_{1:m})\cap G_\delta} \mathbb{E}_{\theta_{u,v}\sim\mathbb{U}(G_\delta)} \max\left\{\frac{\sigma_n^2}{4}(\hat{u}-u)^2, \delta^2 \mathbb{I}_{\{\hat{v}\neq v\}}\right\} \\ &\geq \inf_{\theta_{\hat{u},\hat{v}}\in\mathcal{A}(b_{1:m})\cap G_\delta} \max\left\{\frac{\sigma_n^2}{16}\mathbb{E}_{\theta_{u,v}\sim\mathbb{U}(G_\delta)}(\hat{u}-u)^2, \frac{\delta^2}{4}\mathbb{P}_{\theta_{u,v}\sim\mathbb{U}(G_\delta)}(\hat{v}\neq v)\right\}.\end{aligned} \tag{24}$$

Therefore, by assigning a prior $\theta \sim \mathbb{U}(G_\delta)$, we have successfully decomposed the estimation problem of θ into estimation problems of u and v. We can view estimation of u as "localization" step and estimation of v as "refinement" step, so (24) essentially has decomposed the statistical risk into localization error and refinement error. To lower bound the right hand side of (24), we show that under communication constraints, one cannot simultaneously estimate both u and v accurately, i.e. the localization and refinement errors cannot be both too small. Lemma 3, which shows that for any distributed estimator $\hat{\theta}$, there is unavoidable trade-off between the mutual information $I(\hat{\theta}; u)$ and $I(\hat{\theta}; v)$, is a key step.

We set $\delta = \frac{\sigma_n}{\sqrt{256(B+1-\log(\lfloor\frac{1}{\sigma_n}\rfloor))}}$, and assign the uniform prior $\mathbb{U}(G_\delta)$ to the parameter $\theta = \theta_{u,v}$. One can easily verify $\delta < \frac{1}{8}\sigma_n$, and u, v are independent random variables where u is uniform distributed on $\{0, 1, ..., \lfloor\frac{1}{\sigma_n}\rfloor - 1\}$, and v is uniform distributed on $\{0, 1\}$. Therefore, we can apply Lemma 3 to get inequality (14). From the inequality (14) we can further get, for any $\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap G_\delta$, one of the following two inequalities

$$I(\hat{\theta}; u) \leq \log(\lfloor\frac{1}{\sigma_n}\rfloor) - 1 \quad \text{or} \quad I(\hat{\theta}; v) \leq \frac{64\delta^2}{\sigma_n^2}\left(B + 1 - \log(\lfloor\frac{1}{\sigma_n}\rfloor)\right)$$

must hold. We show that either of the above bounds on the mutual information will result in a large statistical risk.

Case 1.1: $I(\hat{\theta}; u) \leq \log(\lfloor\frac{1}{\sigma_n}\rfloor) - 1$. Note that $\hat{u}$ is a function on $\hat{\theta}$, thus by data processing inequality, $I(\hat{u}; u) \leq I(\hat{\theta}; u) \leq \log(\lfloor\frac{1}{\sigma_n}\rfloor) - 1$. Note that u is uniform distributed on $\{0, 1, ..., \lfloor\frac{1}{\sigma_n}\rfloor - 1\}$, thus $H(u) = \log(\lfloor\frac{1}{\sigma_n}\rfloor)$. We have

$$H(u|\hat{u}) = H(u) - I(\hat{u}; u) \geq 1. \tag{25}$$

The following lemma shows that large conditional entropy will result in large L_2 distance between two integer-valued random variables.

Lemma 16. *Suppose A, D are two integer-valued random variables. If $H(A|D) \geq \frac{1}{2}$, then there exist a constant $c_2 > 0$ such that*

$$\mathbb{E}(A - D)^2 \geq c_2.$$

Given (25) and the fact that $\hat{u}$, u are integer valued, Lemma 16 yields

$$\mathbb{E}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{u} - u)^2 \geq c_2. \tag{26}$$

Case 1.2: $I(\hat{\theta}; v) \leq \frac{\delta^2}{c_1 \sigma_n^2}(B + 1 - \log(\lfloor \frac{1}{\sigma_n} \rfloor))$. By the strong data processing inequality, plug in $\delta = \frac{\sigma_n}{\sqrt{256(B+1-\log(\lfloor \frac{1}{\sigma_n} \rfloor))}}$ we have $I(\hat{v}; v) \leq I(\hat{\theta}; v) \leq \frac{1}{4}$, so $H(v|\hat{v}) = H(v) - I(\hat{v}; v) \geq \frac{3}{4}$. It follows from Lemma 16 that

$$\mathbb{P}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{v} \neq v) = \mathbb{E}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{v} - v)^2 \geq c_2. \tag{27}$$

Combine (26) for Case 1.1 and (27) for Case 1.2 together, we have for any $\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap G_\delta$,

$$\begin{aligned} &\max\left\{\frac{\sigma_n^2}{16}\mathbb{E}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{u} - u)^2, \frac{\delta^2}{4}\mathbb{P}_{\theta_{u,v} \sim \mathbb{U}(G_\delta)}(\hat{v} \neq v)\right\} \\ \geq & c_2 \min\left\{\frac{\sigma_n^2}{16}, \frac{\delta^2}{4}\right\} = \frac{c_2 \sigma_n^2}{1024(B + 1 - \log(\lfloor \frac{1}{\sigma_n} \rfloor))} \geq \frac{c_2}{2048} \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} . \end{aligned} \tag{28}$$

The minimax lower bound follows by combining (24) and (28),

$$\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]} (\hat{\theta} - \theta)^2 \geq \frac{c_2}{2048} \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} .$$

Case 2: $B < \log \frac{1}{\sigma_n} + 2$. Let $S = 2^{B+1}$ and $K_S \triangleq \{\frac{i}{S} : i = 0, 1, ..., S-1\}$. Denote by $\mathbb{U}(K_S)$ the uniform distribution on K_S. For the same reason as in (21) and (22) we have

$$\begin{aligned} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]} (\hat{\theta} - \theta)^2 &\geq \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \mathbb{E}_{\theta \sim \mathbb{U}(K_S)}(\hat{\theta} - \theta)^2 \geq \frac{1}{4} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap K_S} \mathbb{E}_{\theta \sim \mathbb{U}(K_S)}(\hat{\theta} - \theta)^2 \\ &= \frac{1}{4S^2} \inf_{\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap K_S} \mathbb{E}_{\theta \sim \mathbb{U}(K_S)}(S\hat{\theta} - S\theta)^2. \end{aligned} \tag{29}$$

The parameter θ can be treated as a random variable drawn from $\mathbb{U}(K_S)$. Note that by the data processing inequality, for any $\hat{\theta} \in \mathcal{A}(b_{1:m})$,

$$I(\hat{\theta}; \theta) = I(\hat{\theta}(Z_1, Z_2, ..., Z_m); \theta) \leq I(Z_1, Z_2, ..., Z_m; \theta) \leq \sum_{i=1}^m H(Z_i) \leq B.$$

By $\theta \sim \mathbb{U}(K_S)$ we have $H(\theta|\hat{\theta}) = H(\theta) - I(\hat{\theta}; \theta) \geq \log S - B \geq 1$. Note that when $\theta \sim \mathbb{U}(K_S)$, for any $\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap K_S$, $S\hat{\theta}$ and $S\theta$ both take value in $\{0, 1, 2, ..., S-1\}$. Also we have $H(S\theta|S\hat{\theta}) = H(\theta|\hat{\theta}) \geq 1$. Therefore, Lemma 16 yields that $\mathbb{E}_{\theta\sim\mathbb{U}(K_S)}(S\hat{\theta} - S\theta)^2 \geq c_2$. We thus conclude that

$$\frac{1}{4S^2} \inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})\cap K_S} \mathbb{E}_{\theta\sim\mathbb{U}(K_S)}(S\hat{\theta} - S\theta)^2 \geq \frac{c_2}{4\cdot 2^{2(B+1)}} = \frac{c_2}{16}\cdot 2^{-2B}.$$

The desired lower bound follows by plugging into (29).

Case 3: $B \geq \log\frac{1}{\sigma_n} + m$. The minimax risk for distributed protocols is always lower bounded by the minimax risk with no communication constraints:

$$\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})} \sup_{\theta\in[0,1]} (\hat{\theta} - \theta)^2 \geq \inf_{\hat{\theta}} \sup_{\theta\in[0,1]} (\hat{\theta} - \theta)^2 \asymp \frac{\sigma_n^2}{m}.$$

which is given in Bickel (1981). □

Acknowledgments

We would like to acknowledge support for this project from the NSF Grant DMS-1712735 and NIH grants R01-GM129781 and R01-GM123056.

Appendix A. Notation and definitions

There will be several new notation and definitions involved in the following proofs. For any distributions P and Q with common support $\mathcal{X}$, denote the Kullback-Leibler divergence between P and Q as

$$D_{KL}(P\|Q) \triangleq \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx,$$

where p, q are densities of P, Q respectively.

For continuous random variables X, Y supported on $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ with a joint probability density function $f(x, y)$, the differential entropy $H(X)$, conditional differential entropy $H(X|Y)$ and mutual information $I(X;Y)$ are defined as

$$\begin{aligned} H(X) &\triangleq -\int_{\mathcal{X}} f(x)\log f(x)dx, \\ H(X|Y) &\triangleq -\int_{\mathcal{X},\mathcal{Y}} f(x,y)\log f(x|y)dxdy, \\ I(X;Y) &\triangleq \int_{\mathcal{X},\mathcal{Y}} f(x,y)\log\frac{f(x|y)}{f(x)}dxdy. \end{aligned}$$

where $f(x)$ is the marginal density function of X and $f(x|y)$ is the conditional density function of X given Y.

Appendix B. Proof of Theorem 1

We first define the "change points sets" for the Gray functions $g_k(x)$ and conjugate Gray functions $\bar{g}_k(x)$. For any $k \geq 1$, let G_k be the change points set for g_k, which is defined as

$$G_k \triangleq \{(2j-1)\cdot 2^{-k} : 1 \leq j \leq 2^{k-1}\}.$$

Similarly, let $\bar{G}_k$ be the change-points set for $\bar{g}_k$, which is defined as

$$\bar{G}_k \triangleq \{j \cdot 2^{1-k} : 1 \leq j \leq 2^{k-1} - 1\}.$$

As the name suggests, the change points set for a Gray function (or a conjugate Gray function) is the collection of points $x \in [0,1]$ where $g_k(x)$ (or $\bar{g}_k(x)$) changes its value from 0 to 1 or from 1 to 0. More precisely,

$$G_k = \{x : \lim_{y\to x^-} g_k(y) \neq \lim_{y\to x^+} g_k(y)\} \quad \text{and} \quad \bar{G}_k = \{x : \lim_{y\to x^-} \bar{g}_k(y) \neq \lim_{y\to x^+} \bar{g}_k(y)\}.$$

An important property for the change-points sets is that for any $k \geq 1$,

$$\bar{G}_{k+1} = \bigcup_{i=1}^{k} G_k \quad \text{and} \quad G_i \cap G_j = \varnothing \ \forall 1 \leq i < j \leq k. \tag{30}$$

Case 1: $B < \log\frac{1}{\sigma_n} + 2$**.** We first state several technical lemmas in general forms. These lemmas will also be used in Case 2.

Lemma 17. *Let $x \in [0,1]$ and $K \geq 1$ be an integer. Let $g_1, ..., g_K$ be the Gray functions and let $G_1, ..., G_K$ be the corresponding sets of change points. Assume that for any $k \leq K$, either $z_k = g_k(x)$ or $d(x, G_k) \leq 2^{-(K+2)}$. Then we have*

$$d(x, Dec_K(z_1, z_2, ..., z_K)) \leq 2^{-(K+2)}$$

Lemma 18. *If $z_k = g_k(X)$ where $X \sim N(x, \sigma_n^2)$, then*

$$\mathbb{P}(z_k \neq g_k(x)) \leq 2e^{-\frac{d(x,G_k)^2}{2\sigma_n^2}}. \tag{31}$$

Similarly, if $\bar{z}_k = g_k(X)$ where $X \sim N(x, \sigma_n^2)$, then

$$\mathbb{P}(\bar{z}_k \neq \bar{g}_k(x)) \leq 2e^{-\frac{d(x,\bar{G}_k)^2}{2\sigma_n^2}}. \tag{32}$$

Lemma 19. *Fix any $x \in [0,1]$ and integer $1 \leq K \leq \log \frac{1}{\sigma_n} + 2$. For any $1 \leq k \leq K$, let $z_k = g_k(X_k)$ where $X_k \sim N(x, \sigma_n^2)$. ($X_1$,$X_2$,...,$X_K$ can be correlated.) Then there exists a constant $C_1 > 0$ such that, for any $L \leq K$,*

$$\mathbb{P}(d(x, Dec_K(z_1, z_2, ..., z_K)) \geq \frac{5}{4}2^{-L} - 2^{-K}) \leq C_1 e^{-\frac{2^{-2(L+2)}}{2\sigma_n^2}}.$$

Now we prove Case 1. For simplicity denote $A = d(\theta, \mathrm{Dec}_B(U_1, U_2, ..., U_B))$. Note that $A \leq 1$, so we have

$$\begin{aligned}
\mathbb{E}A^2 &\leq \mathbb{P}(A \leq \frac{5}{4}2^{-B}) \cdot (\frac{5}{4}2^{-B})^2 + \sum_{k=0}^{B-1} \mathbb{P}(\frac{5}{4}2^{-B+k} \leq A \leq \frac{5}{4}2^{-B+k+1}) \cdot (\frac{5}{4}2^{-B+k+1})^2 \\
&\leq 1 \cdot (\frac{5}{4}2^{-B})^2 + \sum_{k=0}^{B-1} \mathbb{P}(A \geq \frac{5}{4}2^{-B+k}) \cdot (\frac{5}{4}2^{-B+k+1})^2.
\end{aligned}$$

Note that $B \leq \log \frac{1}{\sigma_n} + 2$, and U_k has the same distribution as $g_k(X)$ where $X \sim N(\theta, \sigma_n^2)$. We can apply Lemma 19 and further get

$$\begin{aligned}
\mathbb{E}A^2 &\leq \frac{25}{16}2^{-2B} + \sum_{k=0}^{B-1} C_1 e^{-\frac{2^{-2(B-k+2)}}{2\sigma_n^2}} \cdot \left(\frac{5}{4}2^{-B+k+1}\right)^2 \\
&\leq \frac{25}{16}2^{-2B} \left(1 + C_1 \sum_{k=0}^{B-1} 2^{-(2k+2)} e^{-\frac{2^{-2(-\log \sigma_n + 2 - k + 2)}}{2\sigma_n^2}}\right) \\
&\leq C_2 \cdot 2^{-2B},
\end{aligned}$$

where $C_2 \triangleq \frac{25}{16}\left(1 + C_1 \sum_{k=0}^{\infty} 2^{-(2k+2)} e^{-2^{(2k-9)}}\right)$ is summable.

Finally, we have $\hat{\theta}_D \in \mathrm{Dec}_B(U_1, U_2, ..., U_B)$ and note that the length of $\mathrm{Dec}_B(U_1, U_2, ..., U_B)$ is 2^{-B}, therefore we conclude that

$$\mathbb{E}(\hat{\theta}_D - \theta)^2 \leq \mathbb{E}(A + 2^{-B})^2 \leq 2\mathbb{E}A^2 + 2^{-2B} \leq (2C_2 + 1)2^{-2B}.$$

The upper bound in (13) for Case 1 is proved.
Case 2: $\log\frac{1}{\sigma_n}+2\le B<\log\frac{1}{\sigma_n}+m$. We define

$$\tilde{I}_1=\{x:d(x,I_1')\le 2^{-(\lfloor\log\frac{1}{\sigma_n}\rfloor-\lfloor\log u\rfloor)}\}\cap[0,1]. \tag{33}$$

which is the interval that stretches out $\frac{1}{4}$ the length of I_1' on both sides.

The proof is divided into three steps with each step summarized as a lemma below. These lemmas also imply the purpose of constructing intervals I_1 and I_2: they are confidence intervals with small risks of θ falling outside.

Lemma 20. *There exists a constant $C_3>0$ such that*

$$\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\notin\tilde{I}_1\}})\le\frac{C_3\sigma_n^2}{u}.$$

Lemma 21. *The set I_2' defined in* (10) *is an interval and there exists a constant $C_4>0$ such that*

$$\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in\tilde{I}_1,\theta\notin I_2\}})\le\frac{C_4\sigma_n^2}{u}.$$

Lemma 22. *(1) One of the following two conditions must hold:*

$$I_2\subseteq\left[(2j-\frac{3}{4})\cdot 2^{-(\lfloor\log\frac{1}{\sigma_n}\rfloor-6)},(2j+\frac{3}{4})\cdot 2^{-(\lfloor\log\frac{1}{\sigma_n}\rfloor-6)}\right]\ \textit{for some } j\in\mathbb{Z}$$

or

$$I_2\subseteq\left[(2j+\frac{1}{4})\cdot 2^{-(\lfloor\log\frac{1}{\sigma_n}\rfloor-6)},(2j+\frac{7}{4})\cdot 2^{-(\lfloor\log\frac{1}{\sigma_n}\rfloor-6)}\right]\ \textit{for some } j\in\mathbb{Z}.$$

(2) There exists a constant $C_5>0$ such that

$$\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in I_2\}})\le\frac{C_5\sigma_n^2}{u}.$$

From the above three lemmas we get

$$\begin{aligned}\mathbb{E}((\hat{\theta}_D-\theta)^2&\le\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\notin\tilde{I}_1\}})+\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in\tilde{I}_1,\theta\notin I_2\}})+\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in I_2\}})\\&\le(C_3+C_4+C_5)\frac{\sigma_n^2}{u}.\end{aligned}$$

By the definition of u in (5), and $u\ge1$, we know

$$B-\lfloor\log\frac{1}{\sigma_n}\rfloor<\lfloor\log(u+1)\rfloor^2+2(u+1)<2u+2(u+1)\le 6u.$$

Hence

$$\mathbb{E}((\hat{\theta}_D-\theta)^2\le 6(C_3+C_4+C_5)\frac{\sigma_n^2}{B-\lfloor\log\frac{1}{\sigma_n}\rfloor}\le 6(C_3+C_4+C_5)\frac{\sigma_n^2}{B-\log\frac{1}{\sigma_n}}.$$

Case 3: $B>\log\frac{1}{\sigma_n}+m$. We can apply the procedure described in Case 2 (or Case 1 if $m=1$) as if we have $B'=\lfloor\log\frac{1}{\sigma_n}\rfloor+m$ total communication budgets. So for some constant $C>0$ we have the guaranteed upper bound

$$\mathbb{E}((\hat{\theta}_D-\theta)^2\le C\frac{\sigma_n^2}{B'-\log\frac{1}{\sigma_n}}\le 2C\cdot\frac{\sigma_n^2}{m}\quad\text{if } m\ge2$$

or $\mathbb{E}((\hat{\theta}_D-\theta)^2\le C\cdot 2^{-2B'}\le C\cdot\frac{\sigma_n^2}{m}$ if $m=1$. ◻

Appendix C. Proof of Theorem 8

Similar as how we proved Theorem 2, we are going to divide our proof into three cases: $B/d < \log\frac{1}{\sigma_n} + 2$, $\log\frac{1}{\sigma_n} + 2 \le B/d < \log\frac{1}{\sigma_n} + m'$, and $B/d \ge \log\frac{1}{\sigma_n} + m'$. We first prove the case $\log\frac{1}{\sigma_n} + 2 \le B < \log\frac{1}{\sigma_n} + m'$, then the case $B/d \ge \log\frac{1}{\sigma_n} + m'$. Some new technical tools are involved in the proof of these two cases. Finally we will prove the relatively easier case $B/d < \log\frac{1}{\sigma_n} + 2$.

Case 1: $\log\frac{1}{\sigma_n} + 2 \le B/d < \log\frac{1}{\sigma_n} + (m' \vee 2)$.

If $m' \le 2$ then no choices of B, m' satisfy this case. So we have $m' > 2$ and $\log\frac{1}{\sigma_n} + 2 \le B/d < \log\frac{1}{\sigma_n} + m'$. Note that $B/d \ge m'$ always holds, so this further implies $B/d > 2$ thus we must have $\sigma_n < 1$ in this case.

We define the same grid of candidate values of θ as in (20):

$$G_\delta \triangleq \{\theta_{u,v} = \sigma_n u + \delta v : u = 0, 1, 2, ..., (\lfloor \frac{1}{\sigma_n} \rfloor - 1), v = 0, 1\}$$

where $0 < \delta < \frac{1}{8}\sigma_n$ is a parameter to be specified later. Let $\mathbb{U}(G_\delta^d)$ be a uniform prior on $G_\delta^d \triangleq G_\delta \times G_\delta \times ... \times G_\delta$. This is equivalent to

$$\theta^{(k)} \overset{\text{iid}}{\sim} \mathbb{U}(G_\delta) \quad \text{for } k = 1, 2, ..., d.$$

Follow the same reason as (21) and (22), we have

$$\inf_{\hat\theta \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]^d} \|\hat\theta - \theta\|^2 \ge \frac{1}{4} \inf_{\hat\theta \in \mathcal{A}(b_{1:m}) \cap G_\delta^d} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \|\hat\theta - \theta\|^2 \tag{34}$$

where $\mathcal{A}(b_{1:m}) \cap G_\delta^d$ is a shorthand for $\mathcal{A}(b_{1:m}) \cap \{\hat\theta : \hat\theta \text{ only takes value in } G_\delta^d\}$.

When $\theta, \hat\theta \in G_\delta^d$, for any $k = 1, 2, ..., d$ we have $\theta^{(k)}, \hat\theta^{(k)} \in G_\delta$, thus we can reparametrize $\theta^{(k)}, \hat\theta^{(k)}$ as

$$\theta^{(k)} = \theta_{u^{(k)}, v^{(k)}} = \sigma_n u^{(k)} + \delta v^{(k)} \quad \text{and} \quad \hat\theta^{(k)} = \theta_{\hat u^{(k)}, \hat v^{(k)}} = \sigma_n \hat u^{(k)} + \delta \hat v^{(i)} \tag{35}$$

so that in the following proof we can view $u^{(k)}, v^{(k)}$ as functions of θ and $\hat u^{(k)}, \hat v^{(k)}$ as functions of $\hat\theta$.

Follow the same reason as (23), we have

$$\begin{aligned}
\inf_{\hat\theta \in \mathcal{A}(b_{1:m}) \cap G_\delta^d} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \|\hat\theta - \theta\|^2 &= \inf_{\hat\theta \in \mathcal{A}(b_{1:m}) \cap G_\delta^d} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \sum_{k=1}^d (\hat\theta^{(k)} - \theta^{(k)})^2 \\
&\ge \inf_{\hat\theta \in \mathcal{A}(b_{1:m}) \cap G_\delta^d} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \sum_{k=1}^d \max\left\{ \frac{\sigma_n^2}{4} (\hat u^{(k)} - u^{(k)})^2, \delta^2 \mathbb{I}_{\{\hat v^{(k)} \ne v^{(k)}\}} \right\} \\
&\ge \inf_{\hat\theta \in \mathcal{A}(b_{1:m}) \cap G_\delta^d} \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \max\left\{ \frac{\sigma_n^2}{4} \sum_{k=1}^d (\hat u^{(k)} - u^{(k)})^2, \delta^2 \sum_{k=1}^d \mathbb{I}_{\{\hat v^{(k)} \ne v^{(k)}\}} \right\} \\
&\ge \inf_{\hat\theta \in \mathcal{A}(b_{1:m}) \cap G_\delta^d} \max\left\{ \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \frac{\sigma_n^2}{4} \sum_{k=1}^d (\hat u^{(k)} - u^{(k)})^2, \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \delta^2 \sum_{k=1}^d \mathbb{I}_{\{\hat v^{(k)} \ne v^{(k)}\}} \right\}.
\end{aligned} \tag{36}$$

Combine (34) and (36) we have

$$\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})} \sup_{\theta\in[0,1]^d} \|\hat{\theta}-\theta\|^2 \geq \inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})\cap G_\delta^d} \max\left\{\mathbb{E}_{\theta\sim\mathbb{U}(G_\delta^d)}\frac{\sigma_n^2}{16}\sum_{k=1}^d(\hat{u}^{(k)}-u^{(k)})^2, \mathbb{E}_{\theta\sim\mathbb{U}(G_\delta^d)}\frac{\delta^2}{4}\sum_{k=1}^d \mathbb{I}_{\{\hat{v}^{(k)}\neq v^{(k)}\}}\right\}. \tag{37}$$

Therefore, by assigning a prior $\theta\sim\mathbb{U}(G_\delta^d)$, we have successively decompose the statistical error into multi-dimensional "localization error" and "refinement error", similar as what we do in the proof of Theorem 2.

Next we are going to provide a lemma that shows a trade-off between $\sum_{k=1}^d I(\hat{\theta},u^{(k)})$ and $\sum_{k=1}^d I(\hat{\theta},v^{(k)})$. This lemma is an analog to Lemma 3.

Lemma 23. *If $\sigma_n<1$, $\theta\sim\mathbb{U}(G_\delta^d)$ and $u^{(k)},v^{(k)}$ is defined as in* (35) *for all $k=1,2,...,d$. Then we have*

$$\sum_{k=1}^d I(\hat{\theta};u^{(k)}) + \frac{\sigma_n^2}{64\delta^2}\sum_{k=1}^d I(\hat{\theta};v^{(k)}) \leq B \tag{38}$$

.

Another lemma is given as an extension to Lemma 16.

Lemma 24. *Suppose $A_1,A_2,...,A_d$ and $D_1,D_2,...,D_d$ are integer-value random variables. If*

$$\frac{1}{d}\sum_{k=1}^d H(A_k|D_k)\geq\frac{1}{2},$$

then there exist a constant $c_3>0$ such that

$$\frac{1}{d}\sum_{k=1}^d \mathbb{E}(A_k-D_k)^2\geq c_3.$$

From now on we set $\delta=\frac{\sigma_n}{\sqrt{256(B/d+1-\log(\lfloor\frac{1}{\sigma_n}\rfloor))}}$, and threat θ is a random variable drawn from $\mathbb{U}(G_\delta^d)$. It is easy to verify that $\delta<\frac{\sigma_n}{8}$. So from Lemma 23 we know that, for any $\hat{\theta}\in\mathcal{A}(b_{1:m})\cap G_\delta^d$, one of the following two inequalities must hold:

$$\sum_{k=1}^d I(\hat{\theta};u^{(k)})\leq d\left(\log(\lfloor\frac{1}{\sigma_n}\rfloor)-1\right) \quad \text{or} \quad \sum_{k=1}^d I(\hat{\theta};v^{(k)})\leq\frac{64d\delta^2}{\sigma_n^2}\left(B/d+1-\log(\lfloor\frac{1}{\sigma_n}\rfloor)\right).$$

Case 1.1: $\sum_{k=1}^d I(\hat{\theta};u^{(k)})\leq d\left(\log(\lfloor\frac{1}{\sigma_n}\rfloor)-1\right)$.

Note that for any $k=1,2,...d$, $u^{(k)}$ is uniformly distributed on $\{1,2,...,\lfloor\frac{1}{\sigma_n}\rfloor\}$. This implies $H(u^{(k)})=\log\lfloor\frac{1}{\sigma_n}\rfloor$ for all k.

Note that $\hat{u}^{(k)}$ is a function of $\hat{\theta}$ for all k, by data processing inequalities we have $I(\hat{u}^{(k)},u^{(k)})\leq I(\theta;u^{(k)})$ for all k. Therefore, we can get

$$\sum_{k=1}^d H(u^{(k)}|\hat{u}^{(k)})=\sum_{k=1}^d\left(H(u^{(k)})-I(\hat{u}^{(k)};u^{(k)})\right)$$

$$\geq \sum_{k=1}^{d} H(u^{(k)}) - \sum_{k=1}^{d} I(\hat{\theta}, u^{(k)})$$
$$\geq d \log\lfloor \frac{1}{\sigma_n} \rfloor - d\left(\log(\lfloor \frac{1}{\sigma_n} \rfloor) - 1\right) = d.$$

$\hat{u}^{(k)}$ and $u^{(k)}$ only take integer values. So we can apply Lemma 24 to get

$$\sum_{k=1}^{d} (\hat{u}^{(k)} - u^{(k)})^2 \geq c_3 d. \tag{39}$$

Case 1.2: $\sum_{k=1}^{d} I(\hat{\theta}; v^{(k)}) \leq \frac{64 d \delta^2}{\sigma_n^2}\left(B/d + 1 - \log(\lfloor \frac{1}{\sigma_n} \rfloor)\right)$.
By data-processing inequality, plug in $\delta = \frac{\sigma_n}{\sqrt{256(B/d+1-\log(\lfloor \frac{1}{\sigma_n} \rfloor))}}$, we have

$$\sum_{k=1}^{d} I(\hat{v}^{(k)}; v^{(k)}) \leq \sum_{k=1}^{d} I(\hat{\theta}; v^{(k)}) \leq \frac{d}{4}.$$

For all k, $v^{(k)}$ is a Bernoulli random variable with mean $\frac{1}{2}$, so $H(v^{(k)}) = 1$. We can get

$$\sum_{k=1}^{d} H(v^{(k)} | \hat{v}^{(k)}) = \sum_{k=1}^{d} \left(H(v^{(k)}) - I(\hat{v}^{(k)}; v^{(k)})\right) \geq \frac{3}{4} d.$$

Apply Lemma 24, and note that $\hat{v}^{(k)}, v^{(k)}$ only take values in $\{0, 1\}$, we have

$$\sum_{k=1}^{d} \mathbb{I}_{\{\hat{v}^{(k)} \neq v^{(k)}\}} = \sum_{k=1}^{d} (\hat{v}^{(k)} - v^{(k)})^2 \geq c_3 d. \tag{40}$$

Combine (39) for Case 1.1 and (40) for Case 1.2 together, we have for any $\hat{\theta} \in \mathcal{A}(b_{1:m}) \cap G_\delta^d$,

$$\begin{aligned}
\max\left\{\mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \frac{\sigma_n^2}{16} \sum_{k=1}^{d} (\hat{u}^{(k)} - u^{(k)})^2, \mathbb{E}_{\theta \sim \mathbb{U}(G_\delta^d)} \frac{\delta^2}{4} \sum_{k=1}^{d} \mathbb{I}_{\{\hat{v}^{(k)} \neq v^{(k)}\}}\right\} &\geq c_3 d \min\{\frac{\sigma_n^2}{16}, \frac{\delta^2}{4}\} \\
&= \frac{c_3 d \sigma_n^2}{1024(B/d + 1 - \log(\lfloor \frac{1}{\sigma_n} \rfloor))} \\
&\geq \frac{c_3}{2048} \cdot \frac{d \sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})}.
\end{aligned} \tag{41}$$

Combine (37) and (41) we get the minimax lower bound

$$\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]^d} \|\hat{\theta} - \theta\|^2 \geq \frac{c_3}{2048} \cdot \frac{d \sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})}.$$

Case 2: $B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2)$

In this case we assign a two-point prior $\theta \sim \mathbb{U}(\{0,\delta\}^d)$, where $\delta \le 1$ is a parameter that will be defined later.

Then we have

$$\begin{aligned}\inf_{\hat\theta\in\mathcal{A}(b_{1:m})}\sup_{\theta\in[0,1]^d}\|\hat\theta-\theta\|^2 &\ge \frac{1}{4}\inf_{\hat\theta\in\mathcal{A}(b_{1:m})\cap\{0,\delta\}^d}\mathbb{E}_{\theta\sim\mathbb{U}(\{0,\delta\}^d)}\|\hat\theta-\theta\|^2 \\ &= \frac{\delta^2}{4}\inf_{\hat\theta\in\mathcal{A}(b_{1:m})\cap\{0,\delta\}^d}\mathbb{E}_{\theta\sim\mathbb{U}(\{0,\delta\}^d)}\sum_{k=1}^d(\frac{\hat\theta^{(k)}}{\delta}-\frac{\theta^{(k)}}{\delta})^2.\end{aligned}\tag{42}$$

From now on we treat θ as a random variable drawn from $\mathbb{U}(\{0,\delta\}^d)$, and assume $\hat\theta$ is an arbitrary distributed estimator from protocol $\mathcal{A}(b_{1:m})\cap\{0,\delta\}^d$. Now we provide a lemma which can be viewed as "multidimensional strong data processing inequality":

Lemma 25 (Multidimensional strong data processing inequality)**.** *Suppose $T=(T^{(1)},T^{(2)},...,T^{(d)})$ where each coordinate is an i.i.d Bernoulli random variable with mean $\frac{1}{2}$. Let μ_0 be a d-dimensional vector and $\Delta>0$ be a positive real number. Let X be a d-dimensional Gaussian random variable where $X^{(1)},X^{(2)},...,X^{(d)}$ are independent with distribution*

$$X^{(k)}\sim N(\mu_0^{(k)}+T^{(k)}\Delta,\sigma_n^2).$$

Let Z be a discrete random variable such that $T\to X\to Z$ is a Markov chain, i.e. $Z\perp T|X$. Then the following multidimensional strong data processing inequality holds:

$$I(T;Z)\le 64\left(\frac{\Delta}{\sigma_n}\right)^2 I(X;Z).\tag{43}$$

Let T be a vector where $T^{(k)}\triangleq\theta^{(k)}/\delta$. Recall that $\theta\in\{0,\delta\}^d$, thus for any $1\le i\le m$, on the $i-th$ machine, $T\to X_i\to Z_i$ forms a Markov chain satisfying conditions in Lemma 25 with $\Delta=\delta$. Therefore, we can apply Lemma 25 to get

$$I(\theta;Z_i)\le 64(\frac{\delta}{\sigma_n})^2 I(X_i;Z_i)\le 64(\frac{\delta}{\sigma_n})^2 b_i.\tag{44}$$

where the last inequality is due to $I(X_i;Z_i)\le H(Z_i)\le b_i$.

Also by data processing inequality, we have for any $1\le i\le m$,

$$I(\theta;Z_i)\le I(\theta;X_i)=d\cdot I(\theta^{(1)},X_i^{(1)})\le d\cdot\frac{\delta^2}{4\sigma_n^2}.\tag{45}$$

where the last inequality is due to the facts $I(\theta^{(1)},X_i^{(1)})=H(X^{(1)})-H(X^{(1)}|\theta^{(1)})$ and an upper bound of differential entropy $H(X^{(1)})$ proved in Michalowicz et al. (2008):

$$H(X^{(1)})\le\frac{1}{2}\ln(2\pi e\sigma_n^2)+\frac{\delta^2}{4\sigma_n^2}=H(X^{(1)}|\theta^{(1)})+\frac{\delta^2}{4\sigma_n^2}.$$

Apply bound (44) for machines with $b_i<d$ and apply bound (45) for those machines with $b_i\ge d$, after taking the summation we can get

$$\sum_{i=1}^d I(\theta;Z_i)\le\sum_{i=1}^d\left(64\frac{\delta^2}{\sigma_n^2}b_i\mathbb{I}_{\{b_i<d\}}+\frac{d\delta^2}{4\sigma_n^2}\mathbb{I}_{\{b_i\ge d\}}\right)$$

$$\leq 64\frac{\delta^2}{\sigma_n^2}\sum_{i=1}^{d}\min\{b_i, d\}$$
$$= 64\frac{d\delta^2}{\sigma_n^2}m'.$$

Then we provide a technical lemma which is a decomposition on the mutual information between multiple random variables.

Lemma 26. *Suppose A and D are two random variables. $Y_1, Y_2, ..., Y_m$ are mutually independent random variables conditional on A and D. Then we have*

$$I(Y_1, Y_2, ..., Y_m; A|D) \leq \sum_{i=1}^{m} I(Y_i; A|D).$$

Particularly, if $D = \varnothing$, i.e. $Y_1, Y_2, ..., Y_m$ are mutually independent random variables conditional on A. Then we have

$$I(Y_1, Y_2, ..., Y_m; A) \leq \sum_{i=1}^{m} I(Y_i; A).$$

Note that $X_1, X_2, ..., X_m$ are independent conditional on θ, thus $Z_1, Z_2, ..., Z_m$ are also independent conditional on θ, apply Lemma 26 and data processing inequality, we have

$$I(\theta; \hat{\theta}) \leq I(\theta; Z_{1:m}) \leq \sum_{i=1}^{m} I(\theta; Z_i).$$

So now we have

$$I(\theta; \hat{\theta}) \leq 64\frac{d\delta^2}{\sigma_n^2}m'. \tag{46}$$

Set $\delta = \frac{1}{16}\left(\sqrt{\frac{\sigma_n^2}{m'}} \wedge 1\right)$. It is easy to verify that $\delta < 1$ is a feasible value. From (46), we have $I(\theta; \hat{\theta}) \leq \frac{d}{4}$. Note that $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(k)}$ are independent Bernoulli variables with mean $\frac{1}{2}$, we have

$$\begin{aligned}\sum_{k=1}^{d} H(\theta^{(k)}|\hat{\theta}^{(k)}) &= \sum_{k=1}^{d}[H(\theta^{(k)}) - I(\theta^{(k)}; \hat{\theta}^{(k)})] \\ &\geq \sum_{k=1}^{d}[1 - I(\theta^{(k)}; \hat{\theta})] \\ &\geq d - I(\theta; \hat{\theta}) \geq \frac{3}{4}d\end{aligned} \tag{47}$$

where the first inequality is due to $H(\theta^{(k)}) = 1$ and data processing inequality; the second inequality is due to the following Lemma 27.

Lemma 27. *If A is a random variable and $Y_1, Y_2, ..., Y_d$ are independent random variables, then*

$$I(A; (Y_1, Y_2, ..., Y_d)) \geq \sum_{k=1}^{d} I(A; Y_k).$$

Finally, note that $\frac{\theta^{(k)}}{\delta}$ and $\frac{\hat{\theta}^{(k)}}{\delta}$ only take integer values for all $1 \le k \le d$. Thus we can apply Lemma 24 to get

$$\mathbb{E}\sum_{k=1}^{d}(\frac{\hat{\theta}^{(k)}}{\delta} - \frac{\theta^{(k)}}{\delta})^2 \ge c_3 d.$$

Substitute the above inequality into (42). we prove the minimax lower bound

$$\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})} \sup_{\theta\in[0,1]^d} \|\hat{\theta}-\theta\|^2 \ge \frac{c_3\delta^2 d}{4} = \frac{c_3}{1024}\cdot d\left(\frac{\sigma_n^2}{m'} \wedge 1\right).$$

Case 3: $B/d < \log\frac{1}{\sigma_n} + 2$
Let $S = 2^{\lfloor B/d \rfloor + 2}$. Define

$$K_S \triangleq \{\frac{i}{S} : i = 0, 1, 2, ..., S-1\}.$$

We are going to assign a uniform prior on $K_S^d \triangleq K_S \times K_S \times ... \times K_S$ to θ. Let $\mathbb{U}(K_S^d)$ denote the uniform distribution on K_S^d. Similar as (29) we have

$$\begin{aligned}\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})} \sup_{\theta\in[0,1]^d} \|\hat{\theta}-\theta\|^2 &\ge \frac{1}{4}\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})\cap K_S^d} \mathbb{E}_{\theta\sim\mathbb{U}(K_S^d)}\|\hat{\theta}-\theta\|^2 \\ &= \frac{1}{4S^2}\inf_{\hat{\theta}\in\mathcal{A}(b_{1:m})\cap K_S^d} \mathbb{E}_{\theta\sim\mathbb{U}(K_S^d)}\sum_{k=1}^{d}(S\hat{\theta}^{(k)} - S\theta^{(k)})^2.\end{aligned} \tag{48}$$

From now on we treat the parameter θ as a random variable drawn from the prior distribution $\mathbb{U}(K_S^d)$, and $\hat{\theta}$ is an arbitrary distributed estimator from protocol $\mathcal{A}(b_{1:m})\cap K_S^d$. Note that by data processing inequality, for any $\hat{\theta} \in \mathcal{A}(b_{1:m})$,

$$\begin{aligned} I(\hat{\theta};\theta) &= I(\hat{\theta}(Z_1, Z_2, ..., Z_m);\theta) \\ &\le I(Z_1, Z_2, ..., Z_m;\theta) \\ &\le H(Z_1, Z_2, ..., Z_m) \\ &\le \sum_{i=1}^{m} H(Z_i) \le \sum_{i=1}^{m} b_i = B. \end{aligned}$$

By $\theta \sim \mathbb{U}(K_S^d)$ we have $H(\theta^{(k)}) = \log S$ for all k. Similar as (47), we can apply Lemma 27 to get

$$\begin{aligned} \sum_{k=1}^{d} H(\theta^{(k)}|\hat{\theta}^{(k)}) &= \sum_{k=1}^{d}[H(\theta^{(k)}) - I(\theta^{(k)};\hat{\theta}^{(k)})] \\ &\ge \sum_{k=1}^{d}[\log S - I(\theta^{(k)};\hat{\theta})] \\ &\ge B + d - I(\theta;\hat{\theta}) \ge d. \end{aligned}$$

Note that $S\hat{\theta}$ and $S\theta$ both take integer values in $\{0, 1, 2, ..., S-1\}$. Therefore, we can apply Lemma 24 to have

$$\mathbb{E}(S\hat{\theta} - S\theta)^2 \geq c_3 d.$$

Plug into (48) we can obtain the desired lower bound

$$\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]^d} \|\hat{\theta} - \theta\|^2 \geq \frac{c_3 d}{4} 2^{-2(\lfloor B/d \rfloor + 2)} \geq \frac{c_3}{64} 2^{-2B/d} \cdot d.$$

$\square$

Appendix D. Proof of Theorem 10

Case 1: $B < d$. We first consider the case when $B < d$. When $B < d$ we have $m' \leq B/d < 1$. This implies we have either $B/d < \log \frac{1}{\sigma_n} + 2$ or $B/d > \log \frac{1}{\sigma_n} + m' \vee 2$.

If $B/d < \log \frac{1}{\sigma_n} + 2$, the trivial bound

$$\sup_{\theta \in [0,1]^d} \mathbb{E}\|\hat{\theta}_D - \theta\|^2 \leq d \leq C \cdot 2^{-2B/d} d$$

holds for any $C > 4$.

If $B/d > \log \frac{1}{\sigma_n} + m' \vee 2$, because $m' \leq B/d < 1$ we have $\sigma_n > 1$. So $\frac{\sigma_n^2}{m'} > 1$, thus we have

$$\sup_{\theta \in [0,1]^d} \mathbb{E}\|\hat{\theta}_D - \theta\|^2 \leq d \leq C \cdot d \left(\frac{\sigma_n^2}{m'} \wedge 1 \right)$$

holds as long as $C \geq 1$.

Case 2: $B \geq d$. In this case we assume $B \geq d$ in the following discussion. First, we are going to show for any $k = 1, 2, ..., m$:

- There are at least $\lfloor m' \rfloor$ machines having positive $b_i^{(k)}$, i.e.

$$\#\{1 \leq i \leq m : b_i^{(k)} \geq 1\} \geq \lfloor m' \rfloor. \tag{49}$$

- The total communication budgets for estimating k-th coordinate is at least $\lfloor B/d \rfloor$ bits, i.e.

$$\sum_{i=1}^{m} b_i^{(k)} \geq \lfloor B/d \rfloor. \tag{50}$$

Since $b_1 \leq b_2 \leq ... \leq b_m$, let $L \geq 0$ be the index that

$$b_i \leq d - 1 \text{ for all } i \leq L; \quad b_i \geq d \text{ for all } i \geq L + 1.$$

Fix any $1 \leq k \leq d$, for all $i \leq L$ we have $b_i^{(k)} \leq 1$. And we have

$$\sum_{i=1}^{L} b_i^{(k)} = \left\lfloor \frac{\sum_{j=1}^{L} b_j - k}{d} \right\rfloor - \left\lfloor \frac{-k}{d} \right\rfloor \geq \left\lfloor \frac{\sum_{j=1}^{L} b_j}{d} \right\rfloor.$$

Thus

$$\#\{1 \leq i \leq L : b_i^{(k)} \geq 1\} \geq \left\lfloor \frac{\sum_{j=1}^{L} b_j}{d} \right\rfloor.$$

On the contrary, for all $i \geq L+1$ we have $b_i^{(k)} \geq 1$, thus we have

$$\#\{L+1 \leq i \leq m : b_i^{(k)} \geq 1\} = m - L.$$

Combine the two inequalities above we get

$$\#\{1 \leq i \leq m : b_i^{(k)} \geq 1\} \geq \left\lfloor \frac{\sum_{j=1}^{L} b_j}{d} \right\rfloor + (m-L) = \left\lfloor \frac{\sum_{j=1}^{m} b_j \wedge d}{d} \right\rfloor = \lfloor m' \rfloor.$$

So we have proved (49)

Moreover, (50) can be proved by

$$\sum_{i=1}^{m} b_i^{(k)} = \left\lfloor \frac{\sum_{j=1}^{m} b_j - k}{d} \right\rfloor - \left\lfloor \frac{-k}{d} \right\rfloor \geq \left\lfloor \frac{B}{d} \right\rfloor.$$

Note that $\hat{\theta}_D^{(k)}$ is obtained by the MODGAME procedure with communication budgets $b_1^{(k)}, b_2^{(k)}, ..., b_m^{(k)}$. $B \geq d$ implies $B/d \geq 1$ and $m' \geq 1$. By (49), (50) and Theorem 1, we have

$$\sup_{\theta^{(k)} \in [0,1]} \mathbb{E}(\hat{\theta}_D^{(k)} - \theta^{(k)})^2 \leq \begin{cases} C \cdot 2^{-2\lfloor B/d \rfloor} & \text{if } \lfloor B/d \rfloor < \log \frac{1}{\sigma_n} + 2 \\ C \cdot \frac{\sigma_n^2}{(\lfloor B/d \rfloor - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq \lfloor B/d \rfloor < \log \frac{1}{\sigma_n} + \lfloor m' \rfloor \\ C \cdot \frac{\sigma_n^2}{\lfloor m' \rfloor} & \text{if } \lfloor B/d \rfloor \geq \log \frac{1}{\sigma_n} + \lfloor m' \rfloor \end{cases}$$

So finally we can conclude

$$\begin{aligned} \sup_{\theta \in [0,1]^d} \mathbb{E}\|\hat{\theta}_D - \theta\|^2 &= \sup_{\theta \in [0,1]^d} \sum_{k=1}^{m} \mathbb{E}(\hat{\theta}_D^{(k)} - \theta^{(k)})^2 \\ &\leq \sum_{k=1}^{m} \sup_{\theta \in [0,1]^d} \mathbb{E}(\hat{\theta}_D^{(k)} - \theta^{(k)})^2 \\ &\leq \begin{cases} C \cdot 2^{-2\lfloor B/d \rfloor} d & \text{if } \lfloor B/d \rfloor < \log \frac{1}{\sigma_n} + 2 \\ C \cdot \frac{d\sigma_n^2}{(\lfloor B/d \rfloor - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq \lfloor B/d \rfloor < \log \frac{1}{\sigma_n} + \lfloor m' \rfloor \\ C \cdot d \frac{\sigma_n^2}{\lfloor m' \rfloor} & \text{if } \lfloor B/d \rfloor \geq \log \frac{1}{\sigma_n} + \lfloor m' \rfloor \end{cases} \end{aligned}$$

The above inequality is equivalent to (16) by properly scaling the constant C. □

Appendix E. Proof of Theorem 4

The proof is almost the same as the proof of Theorem 1. Here we only focus on critical points that make the proof go through.

Based on the subgaussian tail bound of X_i, we can finish the proof for the case $B < \log\frac{1}{\sigma_n}+2$, and show that Lemma 20 and 21 for the case $\log\frac{1}{\sigma_n}+2 \le B < \log\frac{1}{\sigma_n}+m$ still hold with the same proof as well.

So now it remains to show the Lemma 22 in order to complete the proof. The only difference is to show

$$\mathbb{E}\left(\frac{1}{u}\sum_{i=1}^{u} V_i - \Phi_h(\theta)\right)^2 \le \frac{C}{u}. \tag{51}$$

still holds true for some $C > 0$ under the condition of Theorem 4.

Define

$$\tilde{\Phi}_h(\theta) \triangleq \mathbb{E}_{X\sim\bar{P}_\theta^n} h(X) = \int_{-\infty}^{\infty} h(y)d\bar{P}_\theta^n(y).$$

Note that $\mathbb{E}V_i = \tilde{\Phi}_h(\theta)$ thus we have

$$\mathbb{E}\left(\frac{1}{u}\sum_{i=1}^{u} V_i - \tilde{\Phi}_h(\theta)\right)^2 \le \frac{1}{4u}$$

Also note that $u \le m$, hence in order to show (51) it suffices to show

$$|\tilde{\Phi}_h(\theta) - \Phi_h(\theta)| \le \frac{C}{\sqrt{m}} \tag{52}$$

holds for all θ with some $C > 0$. The above inequality can be proved by Hölder's inquality:

$$|\tilde{\Phi}_h(\theta) - \Phi_h(\theta)| = |\mathbb{E}_{X\sim\bar{P}_\theta^n} h(X) - \mathbb{E}_{X\sim N(\theta,\sigma_n^2)} h(X)| \le TV(\bar{P}_\theta^n, N(\theta,\sigma_n^2)) \cdot \|h\|_\infty \le \frac{D}{\sqrt{m}}.$$

Appendix F. Proof of Lemma 3

First we state strong data processing inequality on Gaussian channels as a lemma:

Lemma 28 (Strong data processing inequality)**.** *Suppose T is a Bernoulli random variable taking values in $\{0,1\}$ with probability $\frac{1}{2}$ each. $\mu_0 < \mu_1 \in \mathbb{R}$. X is a normal variable with mean μ_T and variance σ_n^2*

$$X \sim N(\mu_T, \sigma_n^2).$$

Let Z be a finite, discrete random variable such that $T \to X \to Z$ is a Markov chain, i.e. $Z \perp T|X$. Then the strong data processing inequality holds:

$$I(T;Z) \le 64\left(\frac{\mu_1-\mu_0}{\sigma_n}\right)^2 I(X;Z). \tag{53}$$

Fix any $i \in \{0,1,...,m\}$. Conditional on $u = j$ for any $j \in \{0,1,...,\lfloor\frac{1}{\sigma_n}\rfloor - 1\}$, we have

$$\mathbb{P}(v=0|u=j) = \mathbb{P}(v=1|u=j) = \frac{1}{2} \quad \text{and} \quad X_i|_{u=j} \sim N(j\sigma_n + v\delta, \sigma_n^2).$$

The fact $Z_i = \Pi_i(X_i)$ implies $v \to X_i \to Z_i$ forms a Markov chain. Therefore, apply Lemma 28 we have

$$I(v; Z_i|u=j) \le \frac{64\delta^2}{\sigma_n^2} I(X_i; Z_i|u=j).$$

Further we can get

$$
\begin{aligned}
I(v;Z_i|u) &= \sum_{j\in\{0,1,\dots,\lfloor\frac{1}{\sigma_n}\rfloor-1\}} \mathbb{P}(u=j)I(v;Z_i|u=j) \\
&\leq \sum_{j\in\{0,1,\dots,\lfloor\frac{1}{\sigma_n}\rfloor-1\}} \frac{64\delta^2}{\sigma_n^2}\mathbb{P}(u=j)I(X_i;Z_i|u=j) \\
&= \frac{64\delta^2}{\sigma_n^2}I(X_i;Z_i|u).
\end{aligned} \tag{54}
$$

From the above inequality we have

$$
\begin{aligned}
H(Z_i) &= I(Z_i;u) + H(Z_i|u) \\
&\geq I(Z_i;u) + I(Z_i;X_i|u) \\
&\geq I(Z_i;u) + \frac{\sigma_n^2}{64\delta^2}I(Z_i;v|u) \\
&= I(Z_i;u,v) + (\frac{\sigma_n^2}{64\delta^2}-1)I(Z_i;v|u).
\end{aligned} \tag{55}
$$

Denote $Z_{1:m}$ be the tuple $(Z_1, Z_2, ..., Z_m)$. Note that $X_1, X_2, ..., X_m$ are independent conditional on u and v, thus $Z_1, Z_2, ..., Z_m$ are also independent conditional on u and v. Therefore by Lemma 26 we have

$$
I(Z_{1:m};u,v) \leq \sum_{i=1}^m I(Z_i;u,v). \tag{56}
$$

and

$$
I(Z_{1:m};v|u) \leq \sum_{i=1}^m I(Z_i;v|u). \tag{57}
$$

Now take summation over (55). Note that $\delta < \frac{\sigma_n}{8}$ suggest that $\frac{\sigma_n^2}{64\delta^2} - 1 > 0$, applying (56) and (57) we have

$$
\begin{aligned}
\sum_{i=1}^m H(Z_i) &\geq \sum_{i=1}^m I(Z_i;u,v) + (\frac{\sigma_n^2}{64\delta^2}-1)\sum_{i=1}^m I(Z_i;v|u) \\
&\geq I(Z_{1:m};u,v) + (\frac{\sigma_n^2}{64\delta^2}-1)I(Z_{1:m};v|u) \\
&= I(Z_{1:m};u) + \frac{\sigma_n^2}{64\delta^2}I(Z_{1:m};v|u).
\end{aligned}
$$

$u \perp v$ suggests $H(v|u) = H(v)$, thus we can get

$$
I(Z_{1:m};v|u) = H(v|u) - H(v|Z_{1:m},u) \geq H(v) - H(v|Z_{1:m}) = I(Z_{1:m};v).
$$

$\hat{\theta}$ is a function on $Z_{1:m}$, by data processing inequality we have

$$
I(\hat{\theta};u) \leq I(Z_{1:m};u) \quad \text{and} \quad I(\hat{\theta};v) \leq I(Z_{1:m};v).
$$

Combine the above three inequalities, we have

$$I(\hat{\theta}; u) + \frac{\sigma_n^2}{64\delta^2} I(\hat{\theta}; v) \leq \sum_{i=1}^{m} H(Z_i).$$

The proof is completed by the bound $\sum_{i=1}^{m} H(Z_i) \leq \sum_{i=1}^{m} b_i = B$. $\square$

Appendix G. Proof of Lemma 17

From (30) we know that $\bigcup_{k=1}^{K} G_k$ is a lattice on $[0,1]$ of interval length 2^{-K}, and $G_1, G_2, ..., G_K$ are mutually disjoint. So there is at most one $k \in \{0, 1, ..., K\}$ satisfying $d(x, G_k) \leq 2^{-(K+2)}$.

If there is no $k \in \{0, 1, ..., K\}$ satisfying $d(x, G_k) \leq 2^{-(K+2)}$, then we know that $z_k = g_k(x)$ for all $k \in \{0, 1, ..., K\}$. This apparently implies $x \in Dec_K(z_1, z_2, ..., z_K)$ thus

$$d(x, \mathrm{Dec}_K(z_1, z_2, ..., z_K)) = 0.$$

If there is one $k \in \{0, 1, ..., K\}$ satisfying $d(x, G_k) \leq 2^{-(K+2)}$, denote this number as k', and denote the nearest point in $G_{k'}$ to x as x'. Now we know that $z_k = g_k(x)$ for all $k \neq k'$. If $z_{k'} = g_{k'}(x)$, then similar as above we have $x \in Dec_K(z_1, z_2, ..., z_K)$; If $z_{k'} \neq g_{k'}(x)$, this implies

$$x \in \mathrm{Dec}_K(z_1, z_2, ..., z_{k'-1}, 1 - z_{k'}, z_{k'+1}, ..., z_K).$$

Note that $x' \in G_{k'}$ suggests x' is a change-point for $g_{k'}$, and also note that x' is a endpoint of the interval $\mathrm{Dec}_K(z_1, z_2, ..., z_{k'-1}, 1 - z_{k'}, z_{k'+1}, ..., z_K)$ (because x' is the nearest point in the lattice $\bigcup_{k=1}^{K} G_k$ to x), so it is not difficult to conclude that $\mathrm{Dec}_K(z_1, z_2, ..., z_{k'-1}, 1 - z_{k'}, z_{k'+1}, ..., z_K)$ and $\mathrm{Dec}_K(z_1, z_2, ..., z_{k'}, ..., z_K)$ are adjacent intervals joint at x'. So in this case we have

$$d(x, \mathrm{Dec}_K(z_1, z_2, ..., z_K)) = d(x, x') \leq 2^{-(K+2)}.$$

$\square$

Appendix H. Proof of Lemma 18

We only prove (31) here. The proof for (32) is similar.

When $|X - x| < d(x, G_k)$, there is no change-point in G_k between X and x thus $g_k(x) = g_k(X) = z_k$. So we have

$$\mathbb{P}(z_k \neq g_k(x)) \leq \mathbb{P}(|X - x| \geq d(x, G_k)) \leq 2e^{-\frac{d(x,G_k)^2}{2\sigma_n^2}}.$$

where the second inequality comes from the Gaussian tail bound for $X \sim N(x, \sigma_n^2)$. $\square$

Appendix I. Proof of Lemma 19

Note that $\mathrm{Dec}_K(z_1, z_2, ..., z_K) \subseteq \mathrm{Dec}_L(z_1, z_2, ..., z_L)$ and the length of $\mathrm{Dec}_L(z_1, z_2, ..., z_L)$ is 2^{-L}, the length of $\mathrm{Dec}_K(z_1, z_2, ..., z_L)$ is 2^{-K}, so we have

$$d(x, \mathrm{Dec}_K(z_1, z_2, ..., z_K)) \leq d(x, \mathrm{Dec}_L(z_1, z_2, ..., z_L)) + (2^{-L} - 2^{-K}).$$

This implies

$$\mathbb{P}(d(x, \mathrm{Dec}_K(z_1, z_2, ..., z_K)) \geq \frac{5}{4}2^{-L} - 2^{-K}) \leq \mathbb{P}(d(x, \mathrm{Dec}_L(z_1, z_2, ..., z_L)) \geq 2^{-(L+2)}).$$

Apply Lemma 17 and the union bound, then apply lemma 18, we have

$$\begin{aligned}
\mathbb{P}(d(x, \mathrm{Dec}_L(z_1, z_2, ..., z_L)) \geq 2^{-(L+2)}) &\leq \mathbb{P}(\exists\ k \leq L : d(x, G_k) > 2^{-(L+2)}, z_k \neq g_k(x)) \\
&\leq \sum_{k \leq L : d(x,G_k) > 2^{-(L+2)}} \mathbb{P}(z_k \neq g_k(x)) \\
&\leq \sum_{k=1}^{L} 2e^{-\frac{d(x,G_k)^2}{2\sigma_n^2}} \mathbb{I}_{\{d(x,G_k) > 2^{-(L+2)}\}} \\
&\leq \sum_{k=1}^{L} \sum_{y \in G_k} 2e^{-\frac{(x-y)^2}{2\sigma_n^2}} \mathbb{I}_{\{|x-y| > 2^{-(L+2)}\}} \\
&= \sum_{y \in \bigcup_{k=1}^{L} G_k} 2e^{-\frac{(x-y)^2}{2\sigma_n^2}} \mathbb{I}_{\{|x-y| > 2^{-(L+2)}\}}.
\end{aligned}$$

where the last equality is due to $G_1, G_2, ..., G_L$ are mutually disjoint.

From (30) we know $\bigcup_{k=1}^{L} G_k$ is a lattice on $[0, 1]$ of interval length 2^{-L}. Also note that $2^{-L} \leq 2^{-K} \leq \frac{1}{4}\sigma_n$, thus we have

$$\begin{aligned}
\sum_{y \in \bigcup_{k=1}^{L} G_k} 2e^{-\frac{(x-y)^2}{2\sigma_n^2}} \mathbb{I}_{\{|x-y| > 2^{-(L+2)}\}} &\leq 4\sum_{j=0}^{\infty} e^{-\frac{(2^{-(L+2)} + j\cdot 2^{-L})^2}{2\sigma_n^2}} \\
&\leq 4e^{-\frac{2^{-2(L+2)}}{2\sigma_n^2}} \sum_{j=0}^{\infty} e^{-j^2 \frac{2^{-2L}}{2\sigma_n^2}} \\
&\leq 4e^{-\frac{2^{-2(L+2)}}{2\sigma_n^2}} \sum_{j=0}^{\infty} e^{-j^2/32} \\
&= C_1 e^{-\frac{2^{-2(L+2)}}{2\sigma_n^2}}.
\end{aligned}$$

where $C_1 \triangleq 4\sum_{j=0}^{\infty} e^{-j^2/32}$ is summable. □

Appendix J. Proof of Lemma 20

If $I_1 = [0, 1]$, in this case we have $\tilde{I}_1 = I_1' = [0, 1]$ thus the lemma automatically holds. So now we assume I_1 is a proper subset of $[0, 1]$. In this case we have $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \geq 4$. For simplicity of notations we denote $a = \lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3$. Recall the definition of I_1' in (7)

$$I_1' = \mathrm{Dec}_a(U_1, U_2, ..., U_a)$$

and the definition of I_1 in (33)

$$\tilde{I}_1 = \{x : d(x, I_1') \leq \frac{1}{4} \cdot 2^{-a}\}.$$

Note that $d(\theta, I') \leq 1$, so we have the decomposition

$$
\begin{aligned}
\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\notin\tilde{I}_1\}}) =&\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{d(\theta,I_1')>\frac{1}{4}\cdot 2^{-a}\}})\\
\leq&\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\frac{1}{4}\cdot 2^{-a}<d(\theta,I_1')\leq\frac{5}{4}\cdot 2^{-a+1}\}})+\sum_{k=1}^{a-1}\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\frac{5}{4}\cdot 2^{-a+k}<d(\theta,I_1')\leq\frac{5}{4}\cdot 2^{-a+k+1}\}})\\
\leq&\mathbb{P}(d(\theta,I_1')>\frac{1}{4}\cdot 2^{-a})\cdot(\frac{5}{4}\cdot 2^{-a+1}+2^{-a+1})^2\\
&+\sum_{k=1}^{a-1}\mathbb{P}(d(\theta,I_1')>\frac{5}{4}\cdot 2^{-a+k})\cdot(\frac{5}{4}\cdot 2^{-a+k+1}+2^{-a+1})^2\\
\leq&25\mathbb{P}(d(\theta,I_1')>\frac{1}{4}\cdot 2^{-a})\cdot 2^{-2a}+25\sum_{k=1}^{a-1}\mathbb{P}(d(\theta,I_1')>\frac{5}{4}\cdot 2^{-a+k})\cdot 2^{-2a+2k},
\end{aligned}
\tag{58}
$$

where the second inequality is due to the fact that

$$
|\hat{\theta}_D-\theta|\leq d(\theta,I_1)+2^{-a+1}\leq d(\theta,I_1')+2^{-a+1}
$$

because $\hat{\theta}_D\in I_1$ and the length of I_1 is 2^{-a+1}.

Note that $a\leq\log\frac{1}{\sigma_n}-\log u-2\leq\log\frac{1}{\sigma_n}-2$, thus we can apply Lemma 19 to right hand side of (58) then we have

$$
\begin{aligned}
\mathbb{E}((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\notin\tilde{I}_1\}}) &\leq 25C_1\sum_{k=0}^{a-1}e^{-\frac{2^{-2(a-k+2)}}{2\sigma_n^2}}\cdot 2^{-2a+2k}\\
&\leq 25C_1\cdot e^{-\frac{2^{-2(a+2)}}{2\sigma_n^2}}\cdot 2^{-2a}\left(1+\sum_{k=1}^{a-1}e^{-\frac{2^{-2(a-k+2)}-2^{-2(a+2)}}{2\sigma_n^2}}\cdot 2^{2k}\right)\\
&\leq 25C_1\cdot e^{-\frac{2^{-2(\log\frac{1}{\sigma_n}-\log n)}}{2\sigma_n^2}}\cdot 2^{-2(\log\frac{1}{\sigma_n}-\log n-4)}\left(1+\sum_{k=1}^{a-1}e^{-\frac{2^{-2a+2k-5}}{2\sigma_n^2}}\cdot 2^{2k}\right)\\
&\leq 25C_1\cdot e^{-\frac{n^2}{2}}\cdot 256n^2\sigma_n^2\cdot\left(1+\sum_{k=1}^{a-1}e^{-2^{2k-1}}\cdot 2^{2k}\right)\\
&\leq\frac{C_3\sigma_n^2}{n}.
\end{aligned}
$$

where $C_3\triangleq 6400C_1\cdot\left(\sup_{x\geq 1}x^3e^{-\frac{x^2}{2}}\right)\cdot\left(1+\sum_{k=1}^{\infty}e^{-2^{2k-1}}\cdot 2^{2k}\right)$ is a finite positive constant.

Appendix K. Proof of Lemma 21

In the following proof we focus on the case $\lfloor\log\frac{1}{\sigma_n}\rfloor-\lfloor\log u\rfloor\geq 4$, i.e. I_1 is a proper subset of $[0,1]$. The proof of the case $\lfloor\log\frac{1}{\sigma_n}\rfloor-\lfloor\log u\rfloor<4$ is almost the same except some modification on notations and definitions (the constructed of nested intervals will be $\mathrm{Dec}_k(W_{\lfloor\log u\rfloor-\lfloor\log\frac{1}{\sigma_n}\rfloor+5},W_{\lfloor\log u\rfloor-\lfloor\log\frac{1}{\sigma_n}\rfloor+6},\ldots,W_{\lfloor\log u\rfloor-\lfloor\log\frac{1}{\sigma_n}\rfloor+4+k})$ instead).

When $n = 1$ we have $I_1 \subset I_2$ so the lemma automatically holds. Therefore we assume $n \geq 2$ during the following proof.

For simplicity we still denote $a = \lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3$. Define

$$F_1 \triangleq G_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 2}, \quad F_2 \triangleq \bar{G}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 2},$$

$$\text{and} \quad F_k \triangleq G_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 4 + k} \quad \text{for } k \geq 3.$$

By definition of f in (6), we know for all k, F_k is the change points set for f_k. Also, from (30) we have for all $2 \leq K \leq \lfloor \log u \rfloor$,

$$\bigcup_{k=1}^{K} F_k = \bar{G}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3 + K} \quad \text{and} \quad F_1, F_2, ..., F_k \text{ are mutually disjoint.} \tag{59}$$

Now let's define $J_0 \triangleq I_1$, and for any $1 \leq k \leq \lfloor \log u \rfloor$,

$$J_k \triangleq \{x \in I_1 : f_1(x) = W_1, f_2(x) = W_2, ..., f_k(x) = W_k\}.$$

By definition we know

$$I_2' = J_{\lfloor \log u \rfloor}.$$

Next we are going to provide several claims and show the proof directly after each claim.

Claim 1. *For any $0 \leq k \leq \lfloor \log u \rfloor$, J_k is an interval of length $2^{-(a+k-1)}$ and they are nested:*

$$I_2' = J_{\lfloor \log u \rfloor} \subset J_{\lfloor \log u \rfloor - 1} \subset ... \subset J_0 = I_1.$$

Note that $J_0 = I_1$ is an interval of the form $[(2j-1) \cdot 2^{-(a+1)}, (2j+3) \cdot 2^{-(a+1)})$ for some $j \in \mathbb{Z}$, whose length is $2^{-(a-1)}$. It is not difficult to see that $J_1 \subset J_0$ is an interval of the form $[(2j-1) \cdot 2^{-(a+1)}, (2j+1) \cdot 2^{-(a+1)})$ for some $j \in \mathbb{Z}$, whose length is 2^{-a}... Further induction can prove that for all $2 \leq k \leq \lfloor \log u \rfloor$, $J_k \subset J_{k-1}$ is an interval of the form $[j \cdot 2^{-(a+k-1)}, (j+1) \cdot 2^{-(a+k-1)}]$ for some $j \in \mathbb{Z}$, whose length is $2^{-(a+k-1)}$. Thus the claim is proved.

Claim 2. *Fix integer $2 \leq K \leq \lfloor \log u \rfloor$. If $\theta \in \tilde{I}_1$, and*

$$W_k = f_k(\theta) \quad \textit{for all } 1 \leq k \leq K \textit{ satisfying } d(\theta, F_k) \geq 2^{-(a+K)}$$

Then we have

$$d(x, J_K) \leq 2^{-(a+K)}.$$

This claim is an analog to Lemma 17. Let's assume the conditions stated in the claim hold. From equation (59) we know that $\bigcup_{k=1}^{K} F_k$ is a lattice on $[0,1]$ of interval length 2^{a+K-1}. Also note that $F_1, F_2, ..., F_K$ are mutually disjoint. So there is at most one set $F' \in \{F_1, F_2, ..., F_K\}$ satisfying $d(\theta, F') < 2^{-(a+K)}$.

If there is no F' satisfying $d(\theta, F') < 2^{-(a+K)}$, then we know that $W_k = f_k(\theta)$ for all $1 \leq k \leq K$. This apparently implies $\theta \in J_K$ thus

$$d(\theta, J_K) = 0.$$

If there is one F' satisfying $d(\theta, F') < 2^{-(a+K)}$, denote the nearest point in F' to θ as θ'. Note that from $K \geq 2$ and $\theta \in \tilde{I}_1$, we can point out that θ' is NOT one of the two endpoints of the interval I_1 (this is important!). So similar as the proof of Lemma 17, we can show that J_K must be one of the two adjacent intervals joint at θ'. Thus we have

$$\text{either} \quad \theta \in J_K \quad \text{or} \quad d(\theta, J_K) = d(\theta, \theta') \leq 2^{-(a+K)}.$$

Therefore we can conclude that

$$d(\theta, J_K) \leq 2^{-(a+K)}$$

if the conditions stated in the claim hold.

Claim 3. *For all* $1 \leq k \leq \lfloor \log u \rfloor$ *we have*

$$\mathbb{P}(W_k \neq f_k(\theta)) \leq 2^{-\left(\frac{d(\theta, F_k)^2}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor}. \tag{60}$$

This claim is an analog to Lemma 18. Because $W_1, W_2, ..., W_{\lfloor \log u \rfloor}$ are generated by majority voting, the tail bounds are tighter compared to the tail bound in Lemma 18.

For any $1 \leq k \leq \lfloor \log u \rfloor$, recall that W_k is calculated by the majority voting:

$$W_k = \mathbb{I}_{\{\sum_{j=1}^{\lfloor \log u \rfloor} W_{k,j} \geq \frac{1}{2}\lfloor \log u \rfloor\}}.$$

So we have

$$\mathbb{P}(W_k \neq f_k(\theta)) = \mathbb{P}\left(\frac{1}{\lfloor \log u \rfloor}\sum_{j=1}^{\lfloor \log u \rfloor} \mathbb{I}_{\{W_{k,j} \neq f_k(\theta)\}} \geq \frac{1}{2}\right). \tag{61}$$

Note that $W_{k,1}, W_{k,2}, ..., W_{k,\lfloor \log u \rfloor}$ come from different machines, so $\mathbb{I}_{\{W_{k,j} \neq f_k(\theta)\}}$ ($j = 1, 2, ..., \lfloor \log u \rfloor$) are i.i.d Bernoulli variables. From Lemma 18 we have an upper bound on their success probabilities:

$$\mathbb{P}(W_{k,j} \neq f_k(\theta)) \leq 2e^{-\frac{d(\theta, F_k)^2}{2\sigma_n^2}} \quad \text{for all } j = 1, 2, ..., \lfloor \log u \rfloor.$$

When $d(\theta, F_k)^2 \leq 6\sigma_n^2$ the bound (60) is trivial. So in the following proof we assume $d(x, F_k)^2 \geq 6\sigma_n^2$. This implies

$$\mathbb{P}(W_{k,j} \neq f_k(\theta)) \leq 2e^{-3} < \frac{1}{2}.$$

For simplicity denote $p \triangleq 2e^{-\frac{d(\theta, F_k)^2}{2\sigma_n^2}}$. Apply Chernoff-Hoeffding bound to the right hand side of (61), we have

$$\begin{aligned}
\mathbb{P}(W_k \neq f_k(\theta)) &\leq 2^{-\left(\frac{1}{2}\log\frac{1/2}{p} + \frac{1}{2}\log\frac{1/2}{1-p}\right)\lfloor \log u \rfloor} \\
&\leq 2^{-\left(-\frac{1}{2}\log p - 1\right)\lfloor \log u \rfloor} \\
&= 2^{-\left(\frac{1}{2}\log e^{\frac{d(\theta, F_k)^2}{2\sigma_n^2}} - \frac{3}{2}\right)\lfloor \log u \rfloor} \\
&\leq 2^{-\left(\frac{d(\theta, F_k)^2}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor}.
\end{aligned}$$

Claim 4. *There exist a constant $C_4' > 0$ such that for any $2 \leq K \leq \lfloor \log u \rfloor$ we have*

$$\mathbb{P}(\theta \in \tilde{I}_1, d(\theta, I_2') \geq \frac{3}{2} \cdot 2^{-(a+K-1)} - 2^{-(a+\lfloor \log u \rfloor -1)}) \leq C_4' 2^{-\left(\frac{2^{-2(a+K)}}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor}.$$

This claim is an analog to Lemma 19. Note that by Claim 1 we have $I_2' \subseteq J_K$ and the length of J_K is $2^{-(a+K-1)}$, the length of I_2' is $2^{-a+\lfloor \log u \rfloor -1}$, so we have

$$d(\theta, I_2') \leq d(\theta, J_K) + (2^{-(a+K-1)} - 2^{-a+\lfloor \log u \rfloor -1}).$$

This implies

$$\mathbb{P}(\theta \in \tilde{I}_1, d(\theta, I_2') \geq \frac{3}{2} \cdot 2^{-(a+K-1)} - 2^{-a+\lfloor \log u \rfloor -1}) \leq \mathbb{P}(\theta \in \tilde{I}_1, d(\theta, J_K) \geq 2^{-(a+K)}).$$

Apply Claim 2 and the union bound, then apply Claim 3, we have

$$\begin{aligned}
\mathbb{P}(\theta \in \tilde{I}_1, d(\theta, J_K) \geq 2^{-(a+K)}) \leq& \mathbb{P}\left(\exists k \leq K : d(\theta, F_k) \geq 2^{-(a+K)}, W_k \neq f_k(\theta)\right) \\
\leq& \sum_{k=1}^{K} 2^{-\left(\frac{d(\theta, F_k)^2}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor} \mathbb{I}_{\{d(\theta, F_k) \geq 2^{-(a+K)}\}} \\
\leq& \sum_{y \in \bigcup_{k=1}^{K} F_k} 2^{-\left(\frac{(\theta - y)^2}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor} \mathbb{I}_{\{|\theta - y| \geq 2^{-(a+K)}\}}
\end{aligned}$$

where the last inequality is due to $F_1, F_2, ..., F_K$ are mutually disjoint.

From (59) we know $\bigcup_{k=1}^{K} F_k$ is a lattice on [0,1] of interval length $2^{-(a+K-1)}$. Also note that $2^{-(a+K-1)} \geq 2^{-(a+\lfloor \log u \rfloor -1)} = 2^{-\lfloor \log \frac{1}{\sigma_n} \rfloor + 3} \geq 4\sigma_n$, thus we have

$$\begin{aligned}
&\sum_{y \in \bigcup_{k=1}^{K} F_k} 2^{-\left(\frac{(\theta - y)^2}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor} \mathbb{I}_{\{|\theta - y| \geq 2^{-(a+K)}\}} \\
&\leq 2 \sum_{j=0}^{\infty} 2^{-\left(\frac{(2^{-(a+K)} + j \cdot 2^{-(a+K-1)})^2}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor} \\
&\leq 2 \cdot 2^{-\left(\frac{2^{-2(a+K)}}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor} \cdot \left(\sum_{j=0}^{\infty} 2^{-\left(\frac{j^2 \cdot 2^{-2(a+K-1)}}{4\sigma_n^2}\right)\lfloor \log u \rfloor}\right) \\
&\leq 2 \cdot 2^{-\left(\frac{2^{-2(a+K)}}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor} \cdot \left(\sum_{j=0}^{\infty} 2^{-4j^2 \lfloor \log u \rfloor}\right) \\
&\leq C_4' 2^{-\left(\frac{2^{-2(a+K)}}{4\sigma_n^2} - \frac{3}{2}\right)\lfloor \log u \rfloor}
\end{aligned}$$

where $C_4' \triangleq 2\left(\sum_{j=0}^{\infty} 2^{-4j^2}\right)$ is summable.

PROOF OF LEMMA 21: Next we are going to prove Lemma 21. From Claim 1 we know that $I'_2 = J_{\lfloor \log u \rfloor}$ is an interval. Note that $\hat{\theta}_D \in I'_2 \subseteq I_1$ and the length of I_1 is $2^{-(a-1)}$. Therefore, when $\theta \in \tilde{I}_1$ we have

$$d(\theta, I'_2) \le 2^{-(a-1)}$$

By chaining and apply Claim 3

$$\begin{aligned}
\mathbb{E}\left((\hat{\theta}_D-\theta)\mathbb{I}_{\{\theta\in\tilde{I}_1,\theta\notin I_2\}}\right) =&\mathbb{E}\left((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in\tilde{I}_1,d(\theta,I'_2)>2^{-(a+\lfloor\log u\rfloor)}\}}\right)\\
\le&\mathbb{E}\left((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in\tilde{I}_1,2^{-(a+\lfloor\log u\rfloor)}<d(\theta,I'_2)\le\frac{3}{2}\cdot 2^{-(a+\lfloor\log u\rfloor-2)}\}}\right)\\
&+\sum_{k=3}^{\lfloor\log u\rfloor-1}\mathbb{E}\left((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in\tilde{I}_1,\frac{3}{2}\cdot2^{-(a+k-1)}<d(\theta,I'_2)\le\frac{3}{2}\cdot2^{-(a+k-2)}\}}\right)\\
&+\mathbb{E}\left((\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in\tilde{I}_1,\frac{3}{2}\cdot2^{-(a+1)}<d(\theta,I'_2)\le2^{-(a-1)}\}}\right)\\
\le&\mathbb{P}\left(\theta\in\tilde{I}_1,d(\theta,I'_2)>2^{-(a+\lfloor\log u\rfloor)}\right)\left(\frac{3}{2}\cdot2^{-(a+\lfloor\log u\rfloor-2)}\right)^2\\
&+\sum_{k=3}^{\lfloor\log u\rfloor-1}\mathbb{P}\left(\theta\in\tilde{I}_1,d(\theta,I'_2)>\frac{3}{2}\cdot2^{-(a+k-1)}\right)\left(\frac{3}{2}\cdot2^{-(a+k-2)}\right)^2\\
&+\mathbb{P}\left(\theta\in\tilde{I}_1,d(\theta,I'_2)>\frac{3}{2}\cdot2^{-(a+1)}\right)\left(2^{-(a-1)}\right)^2\\
\le&C'_4\sum_{k=2}^{\lfloor\log u\rfloor}2^{-\left(\frac{2^{-2(a+k)}}{4\sigma_n^2}-\frac{3}{2}\right)\lfloor\log u\rfloor}\cdot\left(2^{-(a+k-3)}\right)^2\\
=&C'_4\cdot2^{-\left(\frac{2^{-2(a+\lfloor\log u\rfloor)}}{4\sigma_n^2}-\frac{3}{2}\right)\lfloor\log u\rfloor}\cdot\left(2^{-(a+\lfloor\log u\rfloor-3)}\right)^2\\
&\cdot\left(1+\sum_{k=2}^{\lfloor\log u\rfloor-1}2^{-\left(\frac{2^{-2(a+k)}}{4\sigma_n^2}-\frac{2^{-2(a+\lfloor\log u\rfloor)}}{4\sigma_n^2}\right)\lfloor\log u\rfloor}\cdot\left(2^{(\lfloor\log u\rfloor-k)}\right)^2\right).
\end{aligned}$$

Further we can get

$$\begin{aligned}
\sum_{k=2}^{\lfloor\log u\rfloor-1}2^{-\left(\frac{2^{-2(a+k)}}{4\sigma_n^2}-\frac{2^{-2(a+\lfloor\log u\rfloor)}}{4\sigma_n^2}\right)\lfloor\log u\rfloor}\cdot\left(2^{(\lfloor\log u\rfloor-k)}\right)^2 \le&\sum_{k=2}^{\lfloor\log u\rfloor-1}2^{-\left(\frac{2^{-2(a+k)}}{8\sigma_n^2}\right)\lfloor\log u\rfloor}\cdot2^{2(\lfloor\log u\rfloor-k)}\\
=&\sum_{j=1}^{\lfloor\log u\rfloor-2}2^{-\left(\frac{2^{-2(\lfloor\log\frac{1}{\sigma_n}\rfloor-j-3)}}{8\sigma_n^2}\right)\lfloor\log u\rfloor}\cdot2^{2j}\\
\le&\sum_{j=1}^{\lfloor\log u\rfloor-2}2^{-2^{2j+1}\cdot\lfloor\log u\rfloor}\cdot2^{2j}\\
\le&\sum_{j=1}^{\infty}2^{-2^{2j+1}\cdot\lfloor\log u\rfloor}\cdot2^{2j},
\end{aligned}$$

and

$$2^{-\left(\frac{2^{-2(a+\lfloor \log u \rfloor)}}{4\sigma_n^2}-\frac{3}{2}\right)\lfloor \log u \rfloor} \cdot \left(2^{-(a+\lfloor \log u \rfloor-3)}\right)^2 = 2^{-\left(\frac{2^{-2(\lfloor \log \frac{1}{\sigma_n} \rfloor-3)}}{4\sigma_n^2}-\frac{3}{2}\right)\lfloor \log u \rfloor} \cdot \left(2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-6)}\right)^2$$
$$\leq 2^{-\frac{5}{2}\lfloor \log u \rfloor} \cdot 2^{14}\sigma_n^2$$
$$\leq \frac{2^{14}\sigma_n^2}{u}.$$

Therefore we have

$$\mathbb{E}\left((\hat{\theta}_D-\theta)\mathbb{I}_{\{\theta\in\tilde{I}_1,\theta\notin I_2\}}\right) \leq 2^{14}C_4'\left(1+\sum_{j=1}^{\infty} 2^{-2^{2j+1}\cdot\lfloor \log u \rfloor}\cdot 2^{2j}\right)\cdot\frac{\sigma_n^2}{u}$$

where $2^{14}C_4'\left(1+\sum_{j=1}^{\infty} 2^{-2^{2j+1}\cdot\lfloor \log u \rfloor}\cdot 2^{2j}\right)$ is a finite constant. □

Appendix L. Proof of Lemma 22

We are going to provide several claims and show the proof directly after each claim.

Claim 5. *One of the following two formulas must hold:*

$$I_2 \subseteq [(2j-\frac{3}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-6)}, (2j+\frac{3}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-6)}],\ \textit{for some } j\in\mathbb{Z} \tag{62}$$

or

$$I_2 \subseteq [(2j+\frac{1}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-6)}, (2j+\frac{7}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-6)}],\ \textit{for some } j\in\mathbb{Z}. \tag{63}$$

This claim is exactly the first part of the Lemma. We first prove that the length of I_2 is at most $2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-5)}$. We divide the discussion into 4 cases:

If $\lfloor \log \frac{1}{\sigma_n} \rfloor \leq 4$. Then we have $I_2=[0,1]$. The length of I_2 is $1<2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-5)}$.

If $\lfloor \log \frac{1}{\sigma_n} \rfloor \geq 5$ and $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \geq 4$, From Claim 1 in Lemma 21 we know the length of I_2' is $2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-4)}$. Thus the length of I_2 is $2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-5)}$.

If $\lfloor \log \frac{1}{\sigma_n} \rfloor \geq 5$ and $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor = 3$, then we have

$$I_2' = \mathrm{Dec}_{\lfloor \log \frac{1}{\sigma_n} \rfloor-4}(W_1, W_3, ..., W_{\lfloor \log u \rfloor})$$

so the length of I_2' is $2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-4)}$. Thus the length of I_2 is $2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-5)}$.

If $\lfloor \log \frac{1}{\sigma_n} \rfloor \geq 5$ and $\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor \leq 2$, then we have

$$I_2' = \mathrm{Dec}_{\lfloor \log \frac{1}{\sigma_n} \rfloor-4}(W_{\lfloor \log u \rfloor-\lfloor \log \frac{1}{\sigma_n} \rfloor+5}, ..., W_{\lfloor \log u \rfloor})$$

so the length of I_2' is $2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-4)}$. Thus the length of I_2 is $2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-5)}$.

Therefore, we can conclude that the length of I_2 is at most $2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-5)}$. Define j'

$$j' \triangleq \max\left\{j\in\mathbb{Z} : (j+\frac{1}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor-6)} \text{ is at the left to } I_2\right\}.$$

Then it is easy to see that $I_2 \subset [(j'+\frac{1}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}, (j'+\frac{7}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}]$ because the length of I_2 is at most $2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-5)}$. By proper reparametrization, $[j'\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-5)}, (j'+2)\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-5)}]$ can be represented as one of the interval forms stated in (62) or (63). Thus the claim is proved.

Claim 6. *For any $j\in\mathbb{Z}$, Φ_h is a strictly monotone function on $[(2j-\frac{3}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}, (2j+\frac{3}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}]$; $\Phi_{\bar{h}}$ is a strictly monotone function on $[(2j+\frac{1}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}, (2j+\frac{7}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}]$. Further, there exist a constant $c_1>0$ such that*

$$\left|\frac{d\Phi_h(x)}{dx}\right| > \frac{c_1}{\sigma_n} \quad \text{for all } x\in[(2j-\frac{3}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}, (2j+\frac{3}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}];$$

$$\left|\frac{d\Phi_{\bar{h}}(x)}{dx}\right| > \frac{c_1}{\sigma_n} \quad \text{for all } x\in[(2j+\frac{1}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}, (2j+\frac{7}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}].$$

We only prove the properties for Φ_h here. The properties for $\Phi_{\bar{h}}$ can be proved by a similar way.

When $x\in[(2j-\frac{3}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}, (2j+\frac{3}{4})\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}]$ for some $j\in\mathbb{Z}$, let

$$z \triangleq x - 2j\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}. \tag{64}$$

Then we have

$$z\in[-\frac{3}{4}\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}, \frac{3}{4}\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-6)}].$$

Note that

$$\Phi_h(x) = \int_{-\infty}^{\infty}\phi(\frac{1}{\sigma_n}(y-x))h(y)dy = \int_{-\infty}^{\infty}\phi(\frac{1}{\sigma_n}y)h(x-y)dy.$$

A direct calculation gives

$$\frac{d\Phi_h(x)}{dx} = \sum_{k=-\infty}^{\infty}\phi\left(\frac{1}{\sigma_n}(x+(2k-1)\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-7)})\right) - \sum_{k=-\infty}^{\infty}\phi\left(\frac{1}{\sigma_n}(x+2k\cdot 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-7)})\right). \tag{65}$$

For simplicity we denote $\Delta \triangleq 2^{-(\lfloor \log \frac{1}{\sigma_n}\rfloor-7)}$. Substitute (64) into (65) we have

$$\begin{aligned}
\left|\frac{d\Phi_h(x)}{dx}\right| &= \left|\sum_{k=-\infty}^{\infty}\phi\left(\frac{1}{\sigma_n}(z+(2k+j-1)\Delta)\right) - \sum_{k=-\infty}^{\infty}\phi\left(\frac{1}{\sigma_n}(x+(2k+j)\Delta)\right)\right| \\
&= \left|\sum_{k=-\infty}^{\infty}\phi\left(\frac{1}{\sigma_n}(z+2k\Delta)\right) - \sum_{k=-\infty}^{\infty}\phi\left(\frac{1}{\sigma_n}(z+(2k-1)\Delta)\right)\right| \\
&= \left| \left(\phi\left(\frac{z}{\sigma_n}\right) - \phi\left(\frac{1}{\sigma_n}(z+\Delta)\right) - \phi\left(\frac{1}{\sigma_n}(z-\Delta)\right)\right)\right. \\
&\quad + \sum_{k=1}^{\infty}\left(\phi\left(\frac{1}{\sigma_n}(z+2k\Delta)\right) - \phi\left(\frac{1}{\sigma_n}(z+(2k+1)\Delta)\right)\right)
\end{aligned}$$

$$+ \sum_{k=1}^{\infty} \left(\phi\left(\frac{1}{\sigma_n}(z - 2k\Delta) \right) - \phi\left(\frac{1}{\sigma_n}(z - (2k+1)\Delta) \right) \right) \Bigg|.$$

Note that $|z| \le \frac{3}{4} \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)} = \frac{3}{8}\Delta$, so for any $k \ge 1$ we have

$$\phi\left(\frac{1}{\sigma_n}(z + 2k\Delta) \right) - \phi\left(\frac{1}{\sigma_n}(z + (2k+1)\Delta) \right) > 0$$

$$\phi\left(\frac{1}{\sigma_n}(z - 2k\Delta) \right) - \phi\left(\frac{1}{\sigma_n}(z - (2k+1)\Delta) \right) > 0.$$

Moreover, by the bound

$$\frac{2^7}{\sigma_n} \le 2^{-(\log \frac{1}{\sigma_n} - 7)} \le \Delta \le 2^{-(\log \frac{1}{\sigma_n} - 8)} \le \frac{2^8}{\sigma_n},$$

we have

$$\begin{aligned}
\phi\left(\frac{z}{\sigma_n} \right) - \phi\left(\frac{1}{\sigma_n}(z + \Delta) \right) - \phi\left(\frac{1}{\sigma_n}(z - \Delta) \right) &= \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{z^2}{2\sigma_n^2}} \left(1 - e^{-\frac{\Delta(2z+\Delta)}{2\sigma_n^2}} - e^{-\frac{\Delta(-2z+\Delta)}{2\sigma_n^2}} \right) \\
&\ge \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(3\Delta/8)^2}{2\sigma_n^2}} \left(1 - 2e^{-\frac{\Delta^2}{8\sigma_n^2}} \right) \\
&\ge \frac{1}{\sqrt{2\pi}\sigma_n} e^{-9 \cdot 2^9} \left(1 - 2e^{-2^{11}} \right).
\end{aligned}$$

Therefore we can conclude that $\left| \frac{d\Phi_h(x)}{dx} \right|$ is lower bounded by $\frac{c_1}{\sigma_n}$ with a positive constant $c_1 = \frac{1}{\sqrt{2\pi}} e^{-9 \cdot 2^9} \left(1 - 2e^{-2^{11}} \right)$, when $x \in [(2j - \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, (2j + \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}]$ for some $j \in \mathbb{Z}$.

Finally, $\frac{d\Phi_h(x)}{dx}$ is a continuous function of x, so $\left| \frac{d\Phi_h(x)}{dx} \right|$ is lower bounded by a positive constant means that $\frac{d\Phi_h(x)}{dx}$ has the same sign for all $x \in [(2j - \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, (2j + \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}]$. Therefore $\Phi_h(x)$ is a strictly monotone function on this interval.

Proof of Lemma 22: Now we are ready to prove Lemma 22. Here we only prove the case $I_2 \subseteq [(2j - \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, (2j + \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}]$ for some $j \in \mathbb{Z}$ because the other case can be proved by a similar way.

From now on we assume $I_2 \subseteq [(2j - \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}, (2j + \frac{3}{4}) \cdot 2^{-(\lfloor \log \frac{1}{\sigma_n} \rfloor - 6)}]$ for some $j \in \mathbb{Z}$. Now $V_1, V_2, ..., V_n$ are i.i.d Bernoulli variables with mean

$$\mathbb{E}V_i = \mathbb{E}_{X \sim N(\theta, \sigma_n^2)} h(X) = \Phi_h(\theta) \quad i = 1, 2, ..., n.$$

So we have

$$\mathbb{E}\left(\frac{1}{u} \sum_{i=1}^{u} V_i - \Phi_h(\theta) \right)^2 \le \frac{1}{4u}.$$

From Claim 6 we can further know Φ_h is invertible on I_2, and its inverse function $\Phi_h^{-1} : [L_I, R_I] \to I_2$ is a $\frac{\sigma_n}{c_1}$-Lipschitz function. Based on this, it is also easy to show that $\Phi_h^{-1}(\tau_{[L_I, R_I]}(\cdot))$ is a $\frac{\sigma_n}{c_1}$-Lipschitz function on $[0, 1]$. Note that when $\theta \in I_2$ we have

$$\Phi_h^{-1}\left(\tau_{[L_I, R_I]}(\Phi_h(\theta)) \right) = \Phi_h^{-1}(\Phi_h(\theta)) = \theta.$$

Therefore, we have

$$\begin{aligned}\mathbb{E}(\hat{\theta}_D-\theta)^2\mathbb{I}_{\{\theta\in I_2\}} &= \mathbb{E}\left(\Phi_h^{-1}\left(\tau_{[L_I,R_I]}(\frac{1}{u}\sum_{i=1}^{u}V_i)\right)-\Phi_h^{-1}\left(\tau_{[L_I,R_I]}(\Phi_h(\theta))\right)\right)^2\mathbb{I}_{\{\theta\in I_2\}}\\ &\le \mathbb{E}\left(\Phi_h^{-1}\left(\tau_{[L_I,R_I]}(\frac{1}{u}\sum_{i=1}^{u}V_i)\right)-\Phi_h^{-1}\left(\tau_{[L_I,R_I]}(\Phi_h(\theta))\right)\right)^2\\ &\le \frac{\sigma_n^2}{c_1^2}\mathbb{E}\left(\frac{1}{u}\sum_{i=1}^{u}V_i-\Phi_h(\theta)\right)^2\\ &\le \frac{\sigma_n^2}{4c_1^2\cdot u}.\end{aligned}$$

□

Appendix M. Proof of Lemma 16

Denote $s \triangleq \mathbb{E}|A-D|$ and $r(s)\triangleq \frac{\sqrt{1+s^2}-1}{s}$. Let

$$R(s)\triangleq -s\log r(s)+\log(1+r(s))-\log(1-r(s)).$$

We are going to prove a stronger result:

$$H(A|D)\le R(s). \tag{66}$$

If inequality (66) holds, then $H(A|D)\ge\frac{1}{2}$ implies $R(s)\ge\frac{1}{2}$. Because $R(s)$ is an strictly increasing function with $R(0)=0$, this suggests that s is lower bounded by some positive constant $c_2=R^{-1}(\frac{1}{2})$. The prove can be completed by

$$\mathbb{E}(A-D)^2\ge\mathbb{E}|A-D|\ge c_2.$$

So it remains to show the inequality (66). For any $d\in\mathbb{Z}$, denote $s_d\triangleq\mathbb{E}\left(|A-D|\middle|D=d\right)$ and $r_d\triangleq\frac{\sqrt{1+s_d^2}-1}{s_d}$. Let

$$p_k\triangleq\mathbb{P}(A=k|D=d),\quad \alpha\triangleq\log(\frac{1+r_d}{e(1-r_d)}),\quad \beta\triangleq-\log r_d,$$

here we omit their dependence on d for simplicity of notations.

From equations

$$\sum_{k=-\infty}^{\infty}p_k=1$$

and

$$\sum_{k=-\infty}^{\infty}|k-d|p_k=\mathbb{E}\left(|A-D|\middle|D=d\right)=s_d$$

we can get

$$
\begin{aligned}
H(A|D=d) &= \sum_{k=-\infty}^{\infty} -p_k \log p_k \\
&= \sum_{k=-\infty}^{\infty} (-p_k \log p_k - \alpha p_k - \beta|k-d|p_k) + \alpha + \beta s_d \\
&\leq \sum_{k=-\infty}^{\infty} \left(\frac{\log e}{e} 2^{-\alpha-\beta|k-d|} \right) + \alpha + \beta s_d \\
&= \log e + \alpha + \beta s_d = R(s_d)
\end{aligned}
$$

where the inequality is due to $-x\log x - tx \leq \frac{\log e}{e} 2^{-t}$ for all t.

Note that $R''(x) = -\frac{\log e}{s\sqrt{1+s^2}}$ so $R(x)$ is a concave function on $[0,\infty)$. Therefore, by Jensen's inequality, we have

$$
\begin{aligned}
H(A|D) &= \sum_{d=-\infty}^{\infty} \mathbb{P}(D=d) H(A|D=d) \\
&\leq \sum_{d=-\infty}^{\infty} \mathbb{P}(D=d) R(s_d) \\
&\leq R\left(\sum_{d=-\infty}^{\infty} \mathbb{P}(D=d) R(s_d) \right) \\
&= R(s).
\end{aligned}
$$

$\square$

Appendix N. Proof of Lemma 23

Denote $u^{(1:d)}$ be the tuple $(u^{(1)}, u^{(2)}, ..., u^{(d)})$ and $v^{(1:d)}$ be the tuple $(v^{(1)}, v^{(2)}, ..., v^{(d)})$. Fix $1 \leq i \leq m$. For any $\tilde{u} \in \{0, 1, ..., \lfloor \frac{1}{\sigma_n} \rfloor - 1\}^d$, conditional on $u^{(1:d)} = \tilde{u}$, $v^{(1:d)} \to X_i \to Z_i$ is a Markov chain that satisfies conditions of Lemma 25 with $\Delta = \delta$. Therefore, we have

$$
I(v^{(1:d)}; Z_i | u^{(1:d)} = \tilde{u}) \leq 64 \frac{\delta^2}{\sigma_n^2} I(X_i; Z_i | u^{(1:d)} = \tilde{u}).
$$

Mimic the summation trick in (54), the above inequality implies

$$
I(v^{(1:d)}; Z_i | u^{(1:d)}) \leq 64 \frac{\delta^2}{\sigma_n^2} I(X_i; Z_i | u^{(1:d)}).
$$

Now we have

$$
\begin{aligned}
H(Z_i) &= I(u^{(1:d)}; Z_i) + H(Z_i | u^{(1:d)}) \\
&\geq I(u^{(1:d)}; Z_i) + I(X_i; Z_i | u^{(1:d)}) \\
&\geq I(u^{(1:d)}; Z_i) + \frac{\sigma_n^2}{64\delta^2} I(v^{(1:d)}; Z_i | u^{(1:d)}) \\
&= I(u^{(1:d)}, v^{(1:d)}; Z_i) + \left(\frac{\sigma_n^2}{64\delta^2} - 1 \right) I(v^{(1:d)}; Z_i | u^{(1:d)}).
\end{aligned}
\tag{67}
$$

Denote $Z_{1:m}$ be the tuple $(Z_1, Z_2, ..., Z_m)$. Conditional on $u^{(1:d)}$ and $v^{(1:d)}$, $Z_1, Z_2, ..., Z_m$ are independent, thus we can apply Lemma 26 to get

$$I(u^{(1:d)}, v^{(1:d)}; Z_{1:m}) \leq \sum_{i=1}^{m} I(u^{(1:d)}, v^{(1:d)}; Z_i)$$

and

$$I(v^{(1:d)}; Z_{1:m}|u^{(1:d)}) \leq \sum_{i=1}^{m} I(v^{(1:d)}; Z_i|u^{(1:d)}).$$

Note that $\delta < \frac{1}{8}\sigma_n$ implies $\frac{\sigma_n^2}{64\delta^2} - 1 > 0$. Therefore, taking summation over (67) we have

$$\begin{aligned}\sum_{i=1}^{m} H(Z_i) &\geq I(u^{(1:d)}, v^{(1:d)}; Z_{1:m}) + \left(\frac{\sigma_n^2}{64\delta^2} - 1\right) I(v^{(1:d)}; Z_{1:m}|u^{(1:d)}) \\ &= I(u^{(1:d)}; Z_{1:m}) + \frac{\sigma_n^2}{64\delta^2} I(v^{(1:d)}; Z_{1:m}|u^{(1:d)}).\end{aligned}$$

$u^{(1)}, u^{(2)}, ..., u^{(k)}$ are independent, thus by Lemma 27 and data processing inequality we have

$$I(u^{(1:d)}; Z_{1:m}) \geq \sum_{k=1}^{d} I(u^{(k)}; Z_{1:m}) \geq \sum_{k=1}^{d} I(u^{(k)}; \hat{\theta}).$$

Similarly, $v^{(1)}, v^{(2)}, ..., v^{(k)}, u^{(1:d)}$ are independent so we have

$$\begin{aligned}I(v^{(1:d)}; Z_{1:m}|u^{(1:d)}) &= I(v^{(1:d)}, u^{(1:d)}; Z_{1:m}) - I(u^{(1:d)}; Z_{1:m}) \\ &\geq \sum_{k=1}^{d} I(v^{(k)}; Z_{1:m}) + I(u^{(1:d)}; Z_{1:m}) - I(u^{(1:d)}; Z_{1:m}) \\ &\geq \sum_{k=1}^{d} I(v^{(k)}; \hat{\theta}).\end{aligned}$$

So finally we can conclude

$$\sum_{k=1}^{d} I(u^{(k)}; \hat{\theta}) + \frac{\sigma_n^2}{64\delta^2} \sum_{k=1}^{d} I(v^{(k)}; \hat{\theta}) \leq \sum_{i=1}^{m} H(Z_i).$$

The proof is completed by the bound $\sum_{i=1}^{m} H(Z_i) \leq \sum_{i=1}^{m} b_i = B$. $\square$

Appendix O. Proof of Lemma 24

From (66) in the proof of Lemma 16, and the fact $R(x)$ is a concave function on $[0, \infty)$, by Jensen's inequality we can get

$$\frac{1}{d}\sum_{k=1}^{d} H(A_k|D_k) \leq \frac{1}{d}\sum_{k=1}^{d} R\left(\mathbb{E}|A_k - D_k|\right)$$

$$\leq R\left(\frac{1}{d}\sum_{k=1}^{d}\mathbb{E}|A_k - D_k|\right).$$

Therefore, $\frac{1}{d}\sum_{k=1}^{d} H(A_k|D_k) \geq \frac{1}{2}$ implies $R\left(\frac{1}{d}\sum_{k=1}^{d}\mathbb{E}|A_k - D_k|\right) \geq \frac{1}{2}$. Note that $R(x)$ is a strictly increasing function on $[0,\infty)$ with $R(0) = 0$, so $\frac{1}{d}\sum_{k=1}^{d}\mathbb{E}|A_k - D_k|$ is lower-bounded by some positive constant $c_3 \triangleq R^{-1}(\frac{1}{2})$. The proof is completed by the inequality

$$\frac{1}{d}\sum_{k=1}^{d}\mathbb{E}(A_k - D_k)^2 \geq \frac{1}{d}\sum_{k=1}^{d}\mathbb{E}|A_k - D_k| \geq c_3.$$

□

Appendix P. Proof of Lemma 25

Denote $T^{(1:k)}$ be the tuple $\{T^{(1)}, T^{(2)}, ..., T^{(k)}\}$ and denote $X^{(1:k)}$ be the tuple $\{X^{(1)}, X^{(2)}, ..., X^{(k)}\}$. Note that for any $1 \leq k \leq d$, $(T^{(k)}, X^{(k)}) \perp T^{(1:k-1)}$. So conditional on $T^{(1:k-1)}$, $T^{(k)} \to X^{(k)} \to Z$ is a Markov chain where $T^{(k)}$ is always a Bernoulli variable with mean $\frac{1}{2}$, and $X^{(k)}$ is a normal of mean $\mu_0^{(k)} + T^{(k)}\Delta$. Apply Lemma 28 and mimic the summation trick in (54), we have

$$I(T^{(k)}; Z|T^{(1:k-1)}) \leq 64(\frac{\Delta}{\sigma_n})^2 I(X^k; Z|T^{(1:k-1)}).$$

Therefore we can get

$$\begin{aligned} I(T;Z) &= I\left((T^{(1)}, T^{(2)}, ..., T^{(d)}); Z\right) \\ &= \sum_{k=1}^{d} I(T^{(k)}; Z|T^{(1:k-1)}) \leq 64(\frac{\Delta}{\sigma_n})^2 \sum_{k=1}^{d} I(X^k; Z|T^{(1:k-1)}). \end{aligned}$$

We also have

$$\begin{aligned} I(X;Z) &= I\left((X^{(1)}, X^{(2)}, ..., X^{(d)}); Z\right) \\ &= \sum_{k=1}^{d} I(X^{(k)}; Z|X^{(1:k-1)}). \end{aligned}$$

So in order to show the inequality (43), it suffices to show that for all $k = 1, 2, ..., d$,

$$I(X^{(k)}; Z|T^{(1:k-1)}) \leq I(X^{(k)}; Z|X^{(1:k-1)}). \tag{68}$$

When $k = 1$ the above inequality is trivial. When $k \geq 2$, because $X^{(k)} \perp T^{(1:k-1)}$ we have

$$\begin{aligned} I(X^{(k)}; Z|T^{(1:k-1)}) &= I(X^{(k)}; (Z, T^{(1:k-1)})) - I(X^{(k)}, T^{(1:k-1)}) \\ &= I(X^{(k)}; T^{(1:k-1)}|Z) + I(X^{(k)}; Z) - I(X^{(k)}, T^{(1:k-1)}) \\ &= I(X^{(k)}; T^{(1:k-1)}|Z) + I(X^{(k)}; Z). \end{aligned}$$

Because $X^{(k)} \perp X^{(1:k-1)}$ we have

$$
\begin{aligned}
I(X^{(k)}; Z|X^{(1:k-1)}) &= I(X^{(k)}; (Z, X^{(1:k-1)})) - I(X^{(k)}, X^{(1:k-1)}) \\
&= I(X^{(k)}; X^{(1:k-1)}|Z) + I(X^{(k)}; Z) - I(X^{(k)}, X^{(1:k-1)}) \\
&= I(X^{(k)}; X^{(1:k-1)}|Z) + I(X^{(k)}; Z).
\end{aligned}
$$

Note that $X^{(k)} \perp T^{(1:k-1)}|(Z, X^{(1:k-1)})$, this implies the data processing inequality

$$
I(X^{(k)}; T^{(1:k-1)}|Z) \leq I(X^{(k)}; X^{(1:k-1)}|Z).
$$

(68) can be proved by the three formulas given above. □

Appendix Q. Proof of Lemma 26

Denote $Y_{1:i}$ be the tuple $(Y_1, Y_2, ..., Y_i)$ for $1 \leq i \leq m$ and $Y_{1:0} = \varnothing$, The conditional independence implies $Y_i \perp Y_{1:i-1}|A, D$ thus $H(Y_i|A, D, Y_{1:i}) = H(Y_i|A, D)$. Therefore, we have

$$
\begin{aligned}
I(Y_{1:m}; A|D) &= \sum_{i=1}^{m} I(Y_i; A|D, Y_{1:i-1}) \\
&= \sum_{i=1}^{m} [H(Y_i|D, Y_{1:i-1}) - H(Y_i|A, D, Y_{1:i-1})] \\
&\leq \sum_{i=1}^{m} [H(Y_i|D) - H(Y_i|A, D)] \\
&= \sum_{i=1}^{m} I(Y_i; A|D).
\end{aligned}
$$

The inequality is due to the fact that the entropy will not decrease after dropping some conditions.

If $D = \varnothing$, similarly we can prove

$$
I(Y_{1:m}; A) \leq \sum_{i=1}^{m} I(Y_i; A)
$$

by dropping all D in the above derivations.

Appendix R. Proof of Lemma 27

$Y_1, Y_2, ..., Y_d$ are independent so we have

$$
H(Y_1, Y_2, ..., Y_d) = \sum_{k=1}^{d} H(Y_k).
$$

From the basic entropy inequality

$$
H(Y_1, Y_2, ..., Y_d|A) \leq \sum_{k=1}^{d} H(Y_k|A)
$$

we have

$$\begin{aligned} I\left(A; (Y_1, Y_2, ..., Y_d)\right) &= H(Y_1, Y_2, ..., Y_d) - H(Y_1, Y_2, ..., Y_d|A) \\ &\geq \sum_{k=1}^{d} H(Y_k) - \sum_{k=1}^{d} H(Y_k|A) \\ &= \sum_{k=1}^{d} I(A; Y_k). \end{aligned}$$

□

Appendix S. Proof of Lemma 28

The proof of this lemma is based on Theorem 3.7 in Raginsky (2016). We specialize their proof to the Gaussian channels.

When $\mu_1 - \mu_0 > \sigma_n$, data processing inequality implies $I(T; Z) \leq I(X; Z)$ so (53) holds automatically. Thus from now on we assume $\mu_1 - \mu_0 < \sigma_n$.

Let P_T, P_X, P_Z denote the marginal distributions for T, X, Z respectively, and denote $P_{\cdot|\cdot}$ be conditional distributions. Denote the support of Z as $\{z_1, z_2, ...z_n\}$. Define the likelihood ratio

$$a_t(x) \triangleq \frac{dP_{X|T=t}}{dP_X}(x) \quad t = 0, 1$$

$$f_i(x) \triangleq \frac{dP_{X|Z=z_i}}{dP_X}(x) \quad i = 1, 2, 3, ..., n.$$

Step 1: In this step we are going to show that $a_t(X)$ is $32(\frac{\mu_1-\mu_0}{\sigma_n})^2$ - subgaussian for both $t = 0, 1$, i.e.

$$\mathbb{E}\exp[s(a_t(X) - 1)] \leq \exp\left(32\sigma_n^{-2}(\mu_1 - \mu_0)^2 s^2/2\right) \quad \forall s. \tag{69}$$

By symmetry we only need to show $a_0(X)$ is $(\frac{\mu_1-\mu_0}{\sigma_n})^2$ - subgaussian. Note that $\mathbb{E}a_0(X) = 1$, direct calculation yields

$$a_0(X) - \mathbb{E}a_0(X) = \frac{1 - \exp(\sigma_n^{-2}(\mu_1 - \mu_0)(X - \frac{\mu_0+\mu_1}{2}))}{1 + \exp(\sigma_n^{-2}(\mu_1 - \mu_0)(X - \frac{\mu_0+\mu_1}{2}))}.$$

Note that $|a_0(x) - \mathbb{E}a_0(x)| < 1$ for all x. For any $0 < s < 1$ we have

$$\begin{aligned} \mathbb{P}(|a_0(X) - \mathbb{E}a_0(X)| \geq s) &= \mathbb{P}\left(|\sigma_n^{-2}(\mu_1 - \mu_0)(X - \frac{\mu_0 + \mu_1}{2}))| \geq \log(\frac{1+s}{1-s})\right) \\ &\leq \mathbb{P}\left(|\sigma_n^{-2}(\mu_1 - \mu_0)(X - \frac{\mu_0 + \mu_1}{2}))| \geq s\right) \\ &\leq 2\exp\left(-\frac{1}{8\sigma_n^2} \cdot \frac{\sigma_n^4 s^2}{(\mu_1 - \mu_0)^2}\right) = 2\exp\left(-\frac{s^2}{2 \cdot 4\sigma_n^{-2}(\mu_1 - \mu_0)^2}\right). \end{aligned} \tag{70}$$

The second inequality is due to the tail bound on the Gaussian mixture X:

$$\mathbb{P}(|X - \frac{\mu_0 + \mu_1}{2}| \geq s) \leq 2 \exp\left(-\frac{s^2}{8\sigma_n^2}\right).$$

for all $s > 0$ when $\mu_1 - \mu_0 \leq \sigma_n$.

Subgaussian tail bound (70) implies the subgaussian bound on moment generating function:

$$\mathbb{E} \exp[s(a_0(X) - 1)] \leq \exp\left(32\sigma_n^{-2}(\mu_1 - \mu_0)^2 s^2/2\right) \quad \forall s.$$

Step 2: In this step we are going to show that for all $i = 1, 2, ..., n$,

$$D_{KL}(P_{T|Z=z_i} \| P_T) \leq 64(\frac{\mu_1 - \mu_0}{\sigma_n})^2 D_{KL}(P_{X|Z=z_i} \| P_X). \tag{71}$$

Note that $Z \perp T | X$ implies

$$\frac{dP_{T|Z=z_i}}{dP_T}(t) = \int \frac{dP_{X|T=t}}{dP_X}(x) \cdot \frac{dP_{X|Z=z_i}}{dP_X}(x) \cdot dP_X(x) = \mathbb{E}\left(a_t(X) f_i(X)\right).$$

So

$$\begin{aligned} D_{KL}(P_{T|Z=z_i} \| P_T) &= \mathbb{E}\left(\frac{dP_{T|Z=z_i}}{dP_T}(T) \log \frac{dP_{T|Z=z_i}}{dP_T}(T)\right) \\ &= \frac{1}{2}\mathbb{E}\left(a_0(X) f_i(X)\right) \log \mathbb{E}\left(a_0(X) f_i(X)\right) + \frac{1}{2}\mathbb{E}\left(a_1(X) f_i(X)\right) \log \mathbb{E}\left(a_1(X) f_i(X)\right) \\ &\leq \left(\mathbb{E}\left(a_0(X) f_i(X)\right) - 1\right)^2 = \mathrm{Cov}\left[a_0(X), f_i(X)\right]^2. \end{aligned} \tag{72}$$

The inequality is due to the facts $\mathbb{E}\left(a_0(X) f_i(X)\right) + \mathbb{E}\left(a_1(X) f_i(X)\right) = 2$ and the general inequality

$$\frac{1}{2} w \log w + \frac{1}{2}(2 - w) \log(2 - w) \leq (w - 1)^2 \quad \forall 0 \leq w \leq 2.$$

Next, we make use of the fact that

$$\mathbb{E}(U \log U) - (\mathbb{E}U) \log(\mathbb{E}U) \geq \mathbb{E}(UZ) - \mathbb{E}U \log \mathbb{E}e^Z \tag{73}$$

for any random variable Z jointly distributed with U and satisfying $\mathbb{E}e^Z < \infty$. (See Theorem 4.13 in Boucheron et al. (2013) for more reference.) We apply (73) with $U = f_i(X)$ and $Z = s(a_0(X) - 1)$ we have

$$\begin{aligned} s\,\mathrm{Cov}\left[a_0(X), f_i(X)\right] &\leq \log \mathbb{E} \exp[s(a_t(X) - 1)] + \mathbb{E}(f_i(X) \log f_i(X)) \\ &\leq 16(\frac{\mu_1 - \mu_0}{\sigma_n})^2 s^2 + D_{KL}(P_{X|Z=z_i} \| P_X). \end{aligned}$$

The second inequality is due to subgaussian bound (69) and the fact $D_{KL}(P_{X|Z=z_i} \| P_X) = \mathbb{E}(f_i(X) \log f_i(X))$.

Plug in $s = \left(32(\frac{\mu_1 - \mu_0}{\sigma_n})^2\right)^{-1} \mathrm{Cov}\left[a_0(X), f_i(X)\right]$ we have

$$\mathrm{Cov}\left[a_0(X), f_i(X)\right]^2 \leq 64(\frac{\mu_1 - \mu_0}{\sigma_n})^2 D_{KL}(P_{X|Z=z_i} \| P_X).$$

Combined with (72) we can conclude

$$D_{KL}(P_{T|Z=z_i}\|P_T) \leq 64(\frac{\mu_1-\mu_0}{\sigma_n})^2 D_{KL}(P_{X|Z=z_i}\|P_X).$$

Step 3: Finally we are going to conclude the desired strong data processing inequality (53). Because we already have inequality (71), we can directly conclude

$$\begin{aligned}
I(T;Z) &= \sum_{i=1}^{n} P_Z(z_i) D_{KL}(P_{T|Z=z_i}\|P_T) \\
&\leq \sum_{i=1}^{n} P_Z(z_i)\cdot 64(\frac{\mu_1-\mu_0}{\sigma_n})^2 D_{KL}(P_{X|Z=z_i}\|P_X) \\
&= 64(\frac{\mu_1-\mu_0}{\sigma_n})^2 I(X;Z).
\end{aligned}$$

□

References

Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. Learning distributions from their samples under communication constraints. *CoRR*, abs/1902.02890, 2019. URL `http://arxiv.org/abs/1902.02890`.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.

PJ Bickel. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9(6):1301–1309, 1981.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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.

Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *The Journal of Machine Learning Research*, 18(1):1493–1514, 2017.

Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein?s method*. Springer Science & Business Media, 2010.

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 6391–6401, 2017.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *arXiv preprint arXiv:1810.00412*, 2018.

Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031, 2019.

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.

Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

Uri Hadar and Ofer Shayevitz. Distributed estimation of gaussian correlations. *IEEE Transactions on Information Theory*, 65(9):5323–5338, 2019.

Y. Han, P. Mukherjee, A. Ozgur, and T. Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 506–510, 2018.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. *arXiv preprint arXiv:1802.08417*, 2018.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Alon Kipnis and John C Duchi. Mean estimation from one-bit measurements. *arXiv preprint arXiv:1901.03403*, 2019.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, 3, 2017.

Joseph Michalowicz, Jonathan Nichols, and Frank Bucholtz. Calculation of differential entropy for a mixed gaussian distribution. *Entropy*, 10(3):200–206, 2008.

Nicole Mücke and Gilles Blanchard. Parallelizing spectrally regularized kernel algorithms. *The Journal of Machine Learning Research*, 19(1):1069–1097, 2018.

Nicole Mücke, Enrico Reiss, Jonas Rungenhagen, and Markus Klein. Data-splitting improves statistical performance in overparameterized regimes. In *International Conference on Artificial Intelligence and Statistics*, pages 10322–10350. PMLR, 2022.

Maxim Raginsky. Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016.

Carla Savage. A survey of combinatorial gray codes. *SIAM review*, 39(4):605–629, 1997.

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.

Botond Szabó and Harry van Zanten. Adaptive distributed methods under communication constraints. *arXiv preprint arXiv:1804.00864*, 2018.

Botond Szabó and Harry van Zanten. An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research*, 20(87):1–30, 2019.

Botond Szabó and Harry van Zanten. Distributed function estimation: Adaptation using minimal communication. *arXiv preprint arXiv:2003.12838*, 2020.

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, 2013a.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013b.

Zhen Zhang and Toby Berger. Estimation via compressed information. *IEEE transactions on Information theory*, 34(2):198–211, 1988.

Yuancheng Zhu and John Lafferty. Distributed nonparametric regression under communication constraints. *arXiv preprint arXiv:1803.01302*, 2018.