Submitted to the Annals of Statistics
arXiv: 0000.0000

DISTRIBUTED ADAPTIVE GAUSSIAN MEAN ESTIMATION WITH UNKNOWN VARIANCE: INTERACTIVE PROTOCOL HELPS ADAPTATION[†]

By T. Tony Cai and Hongji Wei

University of Pennsylvania

Distributed estimation of a Gaussian mean with unknown variance under communication constraints is studied. Necessary and sufficient communication costs under different types of distributed protocols are derived for any estimator that is adaptively rate-optimal over a range of possible values for the variance. Communication-efficient and statistically optimal procedures are developed.

The analysis reveals an interesting and important distinction among different types of distributed protocols: compared to the independent protocols, interactive protocols such as the sequential and blackboard protocols require less communication costs for rate-optimal adaptive Gaussian mean estimation. The lower bound techniques developed in the present paper are novel and can be of independent interest.

1. Introduction. Distributed statistical analysis is becoming increasingly important and challenging, as distributed data sets naturally arise in a range of applications due to size constraints, security concerns, or privacy considerations. For large-scale data analysis, communication costs can be expensive and become the main bottleneck in the learning process. When communication resources are limited, it is important to understand the interplay between the communication constraints and statistical accuracy in order to construct optimal estimation and inference procedures under the communication constraints.

Significant recent effort has been made to gain fundamental understanding of distributed estimation. For example, Zhang et al. (2013); Garg et al. (2014); Braverman et al. (2016); Han et al. (2018); Barnes et al. (2019b) developed lower bound techniques for distributed parametric estimation. Zhu and Lafferty (2018); Szabó and van Zanten (2018); Cai and Wei (2020); Szabo and van Zanten (2020); Cai and Wei (2021); Szabó et al. (2020) considered information-theoretical limits under communication constraints for

[†]The research was supported in part by NSF Grant DMS-2015259 and NIH grants R01-GM129781 and R01-GM123056.

MSC 2010 subject classifications: Primary 62F30; secondary 62B10, 62F12

Keywords and phrases: Adaptive estimation, communication constraints, distributed learning, Gaussian mean, minimax lower bound, optimal rate of convergence

various distributed problems, such as Gaussian mean estimation, linear regression, nonparametric regression and testing. Optimality results have been established under different communication constraints. Besides theoretical analysis, progress has also been made on developing practical methodologies for distributed estimation. See, for example, Kleiner et al. (2014); Deisenroth and Ng (2015); Lee et al. (2017); Diakonikolas et al. (2017); Jordan et al. (2019); Battey et al. (2018); Fan et al. (2019).

In the present paper we study distributed adaptive Gaussian mean estimation with unknown variance in a decision-theoretical framework. This is a basic yet fundamental distributed estimation problem. Gaussian mean estimation with known variance has been intensively studied in the distributed setting. See, for example, Garg et al. (2014); Braverman et al. (2016); Barnes et al. (2019b); Cai and Wei (2020). The optimality results in these papers were established in the non-adaptive setting where the variance of Gaussian observations is known a priori, and the estimation procedures and statistical lower bound arguments critically depend on the knowledge of variance. In a wide range of statistical applications, the variance of the observations is unknown and the procedures and results developed in the aforementioned papers are no longer applicable. Adaptive Gaussian mean estimation with unknown variance is technically challenging, and differs significantly from the non-adaptive setting. Understanding distributed adaptive Gaussian mean estimation with unknown variance also provides insight into other related statistical problems including distributed density estimation and distributed nonparametric regression with random design.

The primary goal of the present paper is to precisely characterize the minimal communication costs for adaptive Gaussian mean estimation without prior knowledge of variance under different types of distributed protocols, and construct communication-efficient estimators. Our analysis shows that the case of unknown variance differs significantly from the case when σ^2 is known. In particular, in sharp contrast to the known variance case, the behaviors of adaptive Gaussian mean estimation with unknown variance are very different under the independent and interactive protocols.

1.1. *Distributed estimation framework and distributed protocols.* We begin by introducing a general framework for distributed estimation by giving a formal definition of transcript, distributed estimator, and distributed protocols. Let $\mathcal{P} = \{P_{\theta,\xi} : \theta \in \Theta, \xi \in \Xi\}$ be a parametric family of distributions supported on space $\mathcal{X}$, where $\theta \in \Theta$ is the parameter of interest and $\xi \in \Xi$ are nuisance parameters. Suppose there are m local machines and a central machine, where the local machines contain the observations and each local

machine has access only to data in that machine, 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 $P_{\theta,\xi} \in \mathcal{P}$:

$$X_i \overset{\text{i.i.d.}}{\sim} P_{\theta,\xi}, \quad i = 1, \ldots, m,$$

where the i-th local machine has access to X_i only.

On each machine, because of limited communication budget, the observation X_i on the i-th local machine needs to be processed to a uniquely decodable binary string Z_i. The resulting string Z_i, which is called the **transcript** from the i-th machine, is transmitted to the central machine. Finally, after all transcripts $Z_1, ..., Z_m$ are generated, a **distributed estimator** $\hat{\theta}$ is constructed on the central machine based on the transcripts $Z_1, ..., Z_m$,

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

The rules and constraints related to how transcripts can be constructed, which is called **distributed protocol**, has a lot of different variety. we are primarily interested in three different types of distributed protocols: independent protocol, sequential protocol, and blackboard protocols:

- **Independent protocol.** The local machines simultaneously generate transcripts and then send them to the central machine. The i-th transcript only depends on the observation X_i on the i-th machine, so it can be expressed by $Z_i = \Pi_i(X_i)$ with some (possibly random) function Π_i. Let $|Z_i|_l$ denote the length of transcript Z_i. The class of independent protocols with total communication cost B is defined as

  $$\mathcal{A}_{ind}(B) = \{\hat{\theta} : \hat{\theta} = \hat{\theta}(Z_1, ..., Z_m), Z_i = \Pi_i(X_i), i = 1, ..., m, \sum_{i=1}^{m} |Z_i|_l \leq B\}.$$

- **Sequential protocol.** The local machines sequentially send transcripts to the next local machine, and finally the central machine collects all the transcripts. The transcript Z_i sent by the i-th local machine depends on local observation X_i and the previous transcripts $Z_1, ..., Z_{i-1}$, which can be written as

  $$Z_i = \Pi_i(X_i, Z_1, ..., Z_{i-1})$$

  where Π_i is a (possibly random) function. The class of sequential protocols with total communication cost B is defined as

  $$\mathcal{A}_{seq}(B) = \{\hat{\theta} : \hat{\theta} = \hat{\theta}(Z_1, ..., Z_m), Z_i = \Pi_i(X_i, Z_1, ..., Z_{i-1}), i = 1, ..., m, \sum_{i=1}^{m} |Z_i|_l \leq B\}.$$

- **Blackboard protocol.** The local machines communicate via a publicly shown blackboard. When a local machine writes a message on the blackboard, all other local machines can see the content. Finally, the central machine collects all the information and outputs the final estimate. The total length of the messages written by all local machines is at most B bits. Similarly, we denote the class of blackboard protocols with total communication cost B as $\mathcal{A}_{bb}(B)$, where the estimator is obtained by a blackboard protocol with total communication cost $\sum_{i=1}^{m} |Z_i|_l \leq B$. It is clear by definitions that the sequential protocols can be considered as a special kind of blackboard protocols.

Independent protocols are considered as **non-interactive** whereas sequential and blackboard protocols are considered as **interactive protocols**. See Kushilevitz (1997); Barnes et al. (2019a) for further discussion on these communication protocols.

1.2. *Main results and our contribution.* If a distributed Gaussian mean estimator achieves the same mean squared error as the optimal centralized estimator (up to a constant factor) over a range of possible value of the variance, we call it rate-optimal adaptive Gaussian mean estimator. The present paper first establishes the lower bounds for the communication costs of rate-optimal adaptive Gaussian mean estimators under the independent, sequential or blackboard protocols respectively. The lower bounds serve as a benchmark for the communication-efficiency of any rate-optimal adaptive Gaussian mean estimator. We then develop estimation algorithms that use the minimal communication cost to achieve the statistical optimal rate of convergence. With the matching upper and lower bounds, we derive the necessary and sufficient communication costs for rate-optimal adaptive Gaussian mean estimators under the independent, sequential or blackboard protocols respectively.

The results exhibit interesting new phenomena. First, the behavior of adaptive Gaussian mean estimation with unknown variance differs significantly from the distributed estimation problem with known variance. Compared to the non-adaptive minimax rate in the case of known variance established in Cai and Wei (2020), there is a cost of adaptation in communication budget for Gaussian mean estimation under the independent protocols, whereas no additional communication budget is needed for adaptation under the interactive protocols. Moreover, it is somewhat surprising that the minimal communication cost for distributed adaptive Gaussian mean estimation under the non-interactive and interactive protocols are different. To the best of our knowledge, this is the first example in statistical distributed

estimation showing that interactions could help with estimation.

The technical tools developed in the present paper to prove the main theorems are novel and can be of independent interest. Most of the existing lower bound techniques are universal for all types of distributed protocols, and also lack the ability to study adaptation over nuisance parameters. The proof of the lower bound under the independent protocols (Theorem 1) are dedicated for adaptive estimation under the independent protocols with a non-information theoretic approach.

1.3. *Related Literature.* As mentioned earlier, distributed Gaussian mean estimation has been intensively studied in the setting of known variance. Zhang et al. (2013); Garg et al. (2014) analyze the distributed estimation problems under the independent protocols. Braverman et al. (2016) applied a strong data processing inequality to obtain lower bounds under the blackboard protocols. Kipnis and Duchi (2017) considers distributed estimation with one-bit measurements under the independent and sequential protocols. Han et al. (2018); Barnes et al. (2019b) proposed non-information theoretic approaches to obtain lower bounds for distributed estimation. Cai and Wei (2020) established a sharp minimax rate of convergence for distributed Gaussian mean estimation with known variance under the independent, sequential, and blackboard protocols. In particular, the results show that the optimal rates are the same under the three protocols when σ^2 is known.

The behavior of estimation problems under various types of distributed protocols has been studied in two different settings. One common setting is that i.i.d. data are distributed over different machines. For example, Braverman et al. (2016); Barnes et al. (2019b) developed unified approach to establish lower bounds for distributed estimation in this setting under independent, sequential, and blackboard protocols. More recently, Acharya et al. (2020) proposed private-coin protocol and public-coin protocol and show that they have different behavior in a distributed Gaussian signal detection problem. Another setting is that data are drawn from different distributions on different local machines. Various two-sample estimation and testing problems have been considered in this setting. Xiang and Kim (2013); Liu (2021) showed that in independence testing problem and two-sample joint density estimation problem, interactions between local machines improve statistical accuracy and communication-efficiency, compared to the classical one-shot communication approaches.

An emerging topic in distributed estimation is the interplay between communication constraints and adaptation. The focus so far has been mainly on adaptive nonparametric function estimation with unknown smoothness in

the distributed setting. Szabo and van Zanten (2020); Cai and Wei (2021) showed that additional communication budget is required in order to achieve adaptation in distributed nonparametric function estimation under the independent protocols. This is in sharp contrast to the classical centralized setting where global adaptation can be achieved for free over a wide range of smoothness classes (Donoho and Johnstone, 1995; Johnstone, 2017).

1.4. *Organization of the paper.* We finish this section with notation and definitions. We first formulate the problem in Section 2. Then we derive the minimal communication cost for rate-optimal adaptive Gaussian mean estimation under the independent protocols in Section 3 and establish the minimal communication cost for rate-optimal adaptive Gaussian mean estimation under the sequential and blackboard protocols in Section 4. The numerical performance of the proposed distributed estimators is investigated in Section 5. Further research directions are discussed in Section 6 and the proofs of main theorems and lemmas are provided in Section 7.

1.5. *Notation and definitions.* For any $a \in \mathbb{R}$, let $\lfloor a \rfloor$ denote the floor function (the largest integer not larger than a), and $\lceil a \rceil$ denote the ceiling function (the smallest integer not smaller than a). Unless otherwise stated, we shorthand $\log a$ as the logarithm to the base 2 of a. For any $a, b \in \mathbb{R}$, let $a \wedge b \triangleq \min\{a, b\}$ and $a \vee b \triangleq \max\{a, b\}$. We use $a = O(b)$ or equivalently $b = \Omega(a)$ to denote there exist a constant $C > 0$ such that $a \le Cb$, and we use $a \asymp b$ to denote $a = O(b)$ while $b = O(a)$. We use $\tau_{[a,b]}(x)$ to denote the truncation function, which is the projection of x onto $[a, b]$. Define the density of a Gaussian distribution with mean 0 and standard deviation σ as

$$\phi_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}.$$

and the tail probability of a standard Gaussian distribution with mean 0 and standard deviation 1 as

$$\Phi(x) = \mathbb{P}(N(0,1) > x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \mathrm{d}y.$$

2. Problem Formulation. In this section, we formulate the statistical problem of distributed Gaussian mean estimation with unknown variance σ^2. Suppose there are m local machines, on the i-th machine there is an i.i.d. normal observation:

$$X_i \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2).$$

The goal is to optimally estimate $\theta \in [0,1]$ without knowing σ^2 under a certain distributed protocol with total communication constraint B. In other words, the estimator needs to be adaptive to the unknown variance σ^2.

In the conventional centralized settings, the minimax risk of restricted Gaussian mean estimation is given in Bickel (1981):

$$\inf_{\hat{\theta}} \sup_{\theta\in[0,1]} \mathbb{E}(\hat{\theta}-\theta)^2 = \frac{\sigma^2}{m} - 4\pi^2\frac{\sigma^4}{m^2} + o(\sigma^2) \asymp \frac{\sigma^2}{m} \wedge 1.$$

The above quantity serves as a benchmark for the Gaussian mean estimation problem. For a given $\sigma_0 > 0$, we call distributed estimator $\hat{\theta}$ a *rate-optimal adaptive estimator* if there exists a constant $C > 0$, not depending on σ, σ_0 or m, such that for any $\sigma \geq \sigma_0$, we have

$$\sup_{\theta\in[0,1]} \mathbb{E}(\hat{\theta}-\theta)^2 \leq C\left(\frac{\sigma^2}{m} \wedge 1\right).$$

Such distributed estimators are consider as statistically optimal and adaptive estimators as it achieves the centralized-setting-optimal rate of convergence over a wide range of σ.

At a high level, let $\mathcal{P}_{\sigma_0} = \{P_{\theta,\sigma} = N(\theta,\sigma^2) : \theta \in [0,1], \sigma \in [\sigma_0,\infty)\}$ be the Gaussian location family with unknown variance. The distributed estimation problem of θ is considered with nuisance parameter σ.

Setting a lower bound $\sigma \geq \sigma_0$ is necessary. This is due to the fact that no distributed estimator with a finite total communication cost B is able to achieve the optimal rate of convergence over all $\sigma > 0$. With total communication cost B, the mean squared error of any distributed estimator is at least of order 2^{-2B} due to discretization error, however, the optimal rate of convergence for Gaussian mean estimation is of order $\frac{\sigma^2}{m}$. As a result, when σ is extremely small, any distributed estimator cannot attain optimal rate of convergence. Therefore, there is no distributed estimator with finite communication cost that can be rate-optimal adaptive with all possible positive real number σ. A lower bound on σ is needed here to make the problem well-formulated. With smaller lower bound σ_0, the distributed estimator needs more communication cost in order to be adaptive over the range $\sigma \geq \sigma_0$, and the estimating procedure would be also different. In the real data application, people needs to choose σ_0 as a priori, depending on prior knowledge on the dataset or the communication budget. See also Remark 5 for further discussion on σ_0.

Throughout this paper, we assume $0 < \sigma_0 \leq \frac{1}{2}$. When $\sigma_0 > \frac{1}{2}$, the solution to the problem is essentially identical to the case $\sigma_0 = \frac{1}{2}$. See Remark 3 for further explanation.

3. Optimal Adaptive Estimation under the independent protocols. We consider in this section adaptive distributed estimation under the independent protocols. We begin by establishing a lower bound for the minimax relative efficiency under the independent protocols with a given communication budget. A rate-optimal adaptive distributed estimator is then constructed. It is shown that the proposed estimator achieves the minimum communication cost among all rate-optimal adaptive estimators, as is shown by the matching lower bound.

3.1. *Lower bound analysis.* It is difficult to directly derive the minimal communication cost for rate-optimal adaptive estimators. In our analysis, we first analyze the statistical performance of the estimators in the class $\mathcal{A}_{ind}(B)$. Then we argue that only when the communication budget B is larger than a certain value, a distributed estimator in $\mathcal{A}_{ind}(B)$ can possibly be a rate-optimal adaptive estimator. This leads to a lower bound for the communication cost among the rate-optimal adaptive estimators.

We use the relative efficiency as a measure for the statistical performance when we derive the lower bound. The relative efficiency for an estimator $\hat{\theta}$ is defined as

$$r(\hat{\theta}, \theta, \sigma) = \left(\frac{\sigma^2}{m} \wedge 1\right)^{-1} \mathbb{E}(\hat{\theta} - \theta)^2$$

which indicates the gap between the mean squared error of the estimator $\hat{\theta}$ and the optimal rate of convergence when data are drawn from $N(\theta, \sigma^2)$.

We consider the minimax relative efficiency under the total communication constraint B:

$$R_{ind}(\sigma_0, B) = \inf_{\hat{\theta} \in \mathcal{A}_{ind}(B)} \sup_{\theta \in [0,1], \sigma \geq \sigma_0} r(\hat{\theta}, \theta, \sigma).$$

The quantity $R_{ind}(\sigma_0, B)$ is a benchmark for the limit of estimation accuracy under the independent protocols with the total communication constraint B, when σ^2 is unknown.

The relative efficiency is closely related to rate-optimal adaptive estimators. According to the definition, $\hat{\theta}$ is a rate-optimal adaptive estimator over $\sigma \geq \sigma_0$, if and only if the maximum relative efficiency for the estimator $\hat{\theta}$ is bounded by some constant C, i.e.

$$\sup_{\theta \in [0,1], \sigma \geq \sigma_0} \left(\frac{\sigma^2}{m} \wedge 1\right)^{-1} \mathbb{E}(\hat{\theta} - \theta)^2 \leq C.$$

Remark 1. As a contrast, the conventional distributed minimax risk

$$\inf_{\hat{\theta} \in \mathcal{A}_{ind}(B)} \sup_{\theta \in [0,1], \sigma \geq \sigma_0} \mathbb{E}(\hat{\theta} - \theta)^2$$

is not a good proxy to study because the estimation problem becomes more difficult when σ^2 is large. When σ is sufficiently large, say, $\sigma > \sqrt{m}$ this minimax mean squared risk is bounded away from zero according to centralized minimax rate given in Bickel (1981).

The following theorem provides a lower bound on the minimax relative efficiency for estimators in $\mathcal{A}_{ind}(B)$.

THEOREM 1. *If $B > \frac{\frac{1}{\sigma_0}}{m}$, there exists a constant $c > 0$, not depending on σ_0, σ, θ or m, such that*

$$R_{ind}(\sigma_0, B) \geq c\sqrt{\frac{m \log \frac{1}{\sigma_0}}{B}} \wedge 1.$$

The techniques used to prove Theorem 1 are novel and can be of independent interest. Roughly speaking, our goal is to prove that there must exist $\sigma \geq \sigma_0$ and θ, δ such that the central machine cannot tell whether the data are drawn from $N(\theta - \delta, \sigma^2)$ or $N(\theta + \delta, \sigma^2)$ by only looking at these transcripts. To accomplish this goal, we give an upper bound on the integrated squared Hellinger distances over different choices of θ and σ:

$$I = \sum_{i=1}^{m} \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} H^2\left(Z_i | X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2); Z_i | X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2)\right) \mathrm{d}\theta$$

where $\sigma_1, \sigma_2, ..., \sigma_J$ are carefully chosen different levels of σ, λ is a tuning constant factor. $H^2(Z_i | X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2); Z_i | X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2))$ denotes the squared Hellinger distances between distribution of Z_i if $X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2)$ and distribution of Z_i if $X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2)$. If I is proved to be small, then there must exist some θ and σ_j such that

$$\sum_{i=1}^{m} H^2\left(Z_i | X_i \sim N(\theta - \lambda\sigma_j, \sigma_j^2); Z_i | X_i \sim N(\theta + \lambda\sigma_j, \sigma_j^2)\right)$$

is small, and then we can conclude that the central machine does not have enough information to distinguish whether the data are drawn from $N(\theta - \lambda\sigma_j, \sigma_j^2)$ or $N(\theta + \lambda\sigma_j, \sigma_j^2)$. This will give a lower bound on the relative efficiency $R_{ind}(\sigma_0, B)$. The above technique can be summarized into the following lemma:

LEMMA 1. *Let $J > 0$ be an integer. Let $\lambda > 0$, $0 < \sigma_0 < \sigma_1 < ... < \sigma_{J-1}$ satisfy $\lambda\sigma_{J-1} < \frac{1}{6}$. If for any distributed estimator $\hat{\theta} \in \mathcal{A}_{ind}(B)$, we have*

$$I = \sum_{i=1}^{m}\sum_{j=0}^{J-1}\int_{\lambda\sigma_j}^{1-\lambda\sigma_j} H^2\left(Z_i|X_i \sim N(\theta-\lambda\sigma_j,\sigma_j^2); Z_i|X_i \sim N(\theta+\lambda\sigma_j,\sigma_j^2)\right)\mathrm{d}\theta \le \frac{J}{2},$$

then there exists a constant $c > 0$ such that

$$R_{ind}(\sigma_0, B) \ge c\lambda^2 m.$$

Theorem 1 gives a lower bound on the relative efficiency for all distributed estimators from $\mathcal{A}_{ind}(B)$. Note that a rate-optimal adaptive estimator should have bounded relative efficiency, the following Corollary 3.1 can be directly derived from Theorem 1.

COROLLARY 3.1. *If an estimator $\hat{\theta} \in \mathcal{A}_{ind}(B)$ is a rate-optimal adaptive estimator, that is, there exists a constant $C > 0$ such that*

$$\mathbb{E}(\hat{\theta}-\theta)^2 \le C\left(\frac{\sigma^2}{m} \wedge 1\right) \quad \text{for all } \sigma \ge \sigma_0.$$

Then there exists a constant $c > 0$ (which only depends on C) such that

$$B \ge cm\log\frac{1}{\sigma_0}.$$

The above corollary states that the minimum communication cost needed for a rate-optimal adaptive estimator is of order $m\log\frac{1}{\sigma_0}$.

3.2. *Optimal estimator under the independent protocols - $\hat{\theta}_q$.* We now construct a communication efficient rate-optimal adaptive estimator under the independent protocol. The optimal estimator $\hat{\theta}_q$ makes use of $m\log\frac{3}{\sigma_0}$ total communication budget to achieve the centralized optimal rate of convergence for all $\sigma \ge \sigma_0$.

The estimator $\hat{\theta}_q$ can be constructed by the following steps.

Step 1: Generating transcripts. Let $d = 2^{\lfloor\log_2\sigma_0\rfloor}$. Let S_d denote the following grid of interval d between $-1-d$ and 2:

$$S_d = \{-1-d, -1, -1+d, -1+2d, ..., 2-d, 2\}.$$

Let Z_i be the quantized version of X_i and then truncate in $[-1, 2]$. That is,

$$Z_i = \begin{cases} -1-d & \text{if } X_i \le -1 \\ 2 & \text{if } X_i \ge 2 \\ \max\{z \in S_d : z \le X_i\} & \text{if } -1 < X_i < 2 \end{cases}$$

In the third case when $-1 < X_i < 2$, Z_i is the maximum number in S_d that is less than or equal to X_i. Since Z_i has only $3/d + 2$ possible values, it can be encoded using at most $\log\left(\frac{3}{d} + 2\right) \leq \log\left(\frac{6}{\sigma_0} + 2\right)$ bits.

Step 2: Estimation. The central machine receives the transcripts $Z_1, ..., Z_m$ from the local machines. Let $Z_{(1)} \leq ... \leq Z_{(m)}$ be the order statistics of $Z_1, ..., Z_m$. First, we calculate $\hat{\sigma}$ by

$$\hat{\sigma} = \begin{cases} \sigma_0 & \text{if } Z_{(\lceil 0.84m \rceil)} - Z_{(\lfloor 0.16m \rfloor)} < \sigma_0 \\ Z_{(\lceil 0.84m \rceil)} - Z_{(\lfloor 0.16m \rfloor)} & \text{if } \sigma_0 \leq Z_{(\lceil 0.84m \rceil)} - Z_{(\lfloor 0.16m \rfloor)} \leq 1 \\ 1 & \text{if } Z_{(\lceil 0.84m \rceil)} - Z_{(\lfloor 0.16m \rfloor)} > 1 \end{cases}$$

Then, let $\tilde{\sigma} = \min\{2^{-k} : 1 \geq 2^{-k} \geq \hat{\sigma}, k \text{ is an integer}\}$, i.e. the minimum number that is power of 2 and also larger than $\hat{\sigma}$. Let $L = \max\{k\tilde{\sigma} : k\tilde{\sigma} \leq Z_{(\lfloor 0.16m \rfloor)}, k \text{ is an integer}\}$, i.e. the largest multiple of $\tilde{\sigma}$ that is less than or equal to $Z_{(\lfloor 0.16m \rfloor)}$. Similarly we define $R = \max\{k\tilde{\sigma} : k\tilde{\sigma} \geq Z_{(\lceil 0.84m \rceil)}, k \text{ is an integer}\}$, i.e. the smallest multiple of $\tilde{\sigma}$ that is larger than or equal to $Z_{(\lceil 0.84m \rceil)}$. Let $\hat{p}_L = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{\{Z_i < L\}}$ be the proportion of transcripts that is less than L, and $\hat{p}_R = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{\{Z_i \geq R\}}$ be the proportion of transcripts that is larger than or equal to R,

Finally, recall that $\Phi(\cdot)$ denotes the tail probability of a standard Gaussian variable, let $(\hat{\theta}_q, \hat{\sigma}_q)$ be the solution to the equations:

$$\Phi\left(\frac{\hat{\theta}_q - L}{\hat{\sigma}_q}\right) = \hat{p}_L \vee \frac{1}{m},$$

$$\Phi\left(\frac{R - \hat{\theta}_q}{\hat{\sigma}_q}\right) = \hat{p}_R \vee \frac{1}{m}.$$

The above equation always has one unique solution where $\hat{\theta}_q \in [L, R]$, we take this $\hat{\theta}_q$ as the final estimate.

It is easy to verify that the above estimator $\hat{\theta}_q \in \mathcal{A}_{ind}\left(m \log\left(\frac{1}{\sigma_0} + 5\right)\right)$. The next theorem establishes an upper bound on its mean squared error, showing that the estimator is rate-optimal adaptive over $\sigma \geq \sigma_0$.

THEOREM 2. *There exists a constant $C > 0$, not depending on σ_0, σ, θ or m, such that*

$$\sup_{\theta \in [0,1], \sigma \geq \sigma_0} \mathbb{E}(\hat{\theta}_q - \theta)^2 \leq C\left(\frac{\sigma^2}{m} \wedge 1\right).$$

REMARK 2. The construction of the estimator $\hat{\theta}_q$ is involved. A more straightforward and simpler estimator is the quantization-then-average estimator proposed in Zhang et al. (2013). However, it can be shown that the quantization-then-average estimator is not even consistent when each local machine has only limited communication budget, because the quantization bias $(\mathbb{E}Z_i - \theta)$ is not exactly zero if one just rounds the observations to a certain precision on the local machines. As a result, when the number of machines $m \to \infty$, the estimation error will not converge to zero. Therefore, a more sophisticated procedure such as $\hat{\theta}_q$ is necessary to achieve the optimal rate of convergence with the communication constraint.

REMARK 3. The above estimator $\hat{\theta}_q$ is designed under the assumption that $0 < \sigma_0 \leq \frac{1}{2}$. When $\sigma_0 > \frac{1}{2}$, we can use the estimator for the case $\sigma_0 = \frac{1}{2}$, which is rate-optimal adaptive estimator over $\sigma \geq \sigma_0$. The total communication cost is of order m, which cannot be further reduced because each machine needs to transmit at least one bit in order to involve its observation into the estimation procedure. The choice of $\frac{1}{2}$ is for convenience; it can be changed to any positive number and all the results hold with minor modifications.

REMARK 4. Corollary 3.1 and Theorem 2 together show that the necessary and sufficient communication cost for a rate-optimal adaptive estimator is of order $m \log \frac{1}{\sigma_0}$ bits. The order of communication cost of the estimator $\hat{\theta}_q$ cannot be further reduced. Compared to the minimax rate of convergence for non-adaptive Gaussian mean estimation established in the previous complementary work Cai and Wei (2020), the communication cost for adaptive Gaussian mean estimation is larger, so there is a cost of adaptation under the independent protocols.

REMARK 5. The construction of adaptive estimator $\hat{\theta}_q$ requires knowledge of the lower bound σ_0 for unknown σ, which seems unnatural. However, as Theorem 1 suggests, if one lets $\sigma_0 \to 0$, the required communication cost for a distributed estimator to achieve the optimal rate of convergence will go to infinity. Therefore, there is no rate-optimal adaptive estimator for all $\sigma > 0$ without a lower bound on σ. A similar phenomenon also appears in the construction of adaptive confidence ball in nonparametric regression. If one assumes the smoothness $\beta \geq \beta_0$, then it is possible to be adaptive from β_0 to $2\beta_0$. If one does not assume any lower bound for the smoothness, then no adaptation is possible. See Theorem 4 and the discussion thereafter in Cai and Low (2006).

4. Optimal Adaptive Estimate under Interactive Protocols. In the previous section we show that an order of $m \log \frac{1}{\sigma_0}$ bits are necessary and sufficient for an adaptive estimator to achieve its optimal statistical performance under the independent protocols with $\sigma \geq \sigma_0$. However, under the sequential protocols or blackboard protocols, it may require less communication cost to achieve the same statistical performance, because the local machines can "communicate" with each other to some extent. This leads to an interesting question: do we still need $m \log \frac{1}{\sigma_0}$ bits to achieve the optimal rate of convergence over all $\sigma \geq \sigma_0$ under the sequential or blackboard protocols?

We consider in this section distributed estimation under two types of interactive protocols, namely the sequential protocols and the blackboard protocols. We first construct a distributed estimator under the sequential protocols that is statistical optimal for all $\sigma \geq \sigma_0$. A matching lower bound is then established to show that the communication cost of the proposed estimator cannot be further improved for all distributed estimators under the blackboard protocols. Recall that the sequential protocols are a subset of the blackboard protocols, we obtain the sufficient and necessary communication cost for the statistical optimal estimators under interactive protocols. The results show an interesting phenomenon. Compared to the independent protocols, under the sequential protocols or the blackboard protocols, it requires less communication cost for the rate-optimal adaptive estimation. So feedback and information sharing are helpful in distributed Gaussian mean estimation with unknown variance.

4.1. *Optimal estimator under the sequential protocols.* In the following procedure we assume $m \geq 12$. The case of $m \leq 11$ is relatively simple. For example, when $m \leq 11$, the problem can be solved by only looking at the first local machine and outputs its best approximation up to σ_0 precision. The estimation process can be divided into three steps:

Step 1: Preliminary estimation of θ and σ. For the first 11 local machines, the i-th machine ($i = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11$) outputs

$$Z_i = \lfloor \tau_{[-1,2]}(X_i + 1)/\sigma_0 \rfloor.$$

There are at most $\lfloor \frac{3}{\sigma_0} \rfloor + 1$ possible outputs for each local machine, so each transcript Z_i ($i = 1, 2, ..., 11$) can be encoded by no larger than $\log \frac{3}{\sigma_0} + 1$ bits.

On the 12-th and later local machines, based on $Z_1, Z_2, ..., Z_{11}$, each machine can calculate a preliminary estimate of θ and σ by

$$\hat{\theta}_{11} = \sigma_0 Z_{11},$$

$$\hat{\sigma} = \sigma_0 \max\left\{1, \left(\frac{1}{10}\sum_{i=1}^{10}\left(Z_i - \frac{1}{10}\sum_{i=1}^{10} Z_i\right)^2\right)^{1/2}\right\}.$$

Step 2: One-bit passing. Starting at the 12-th local machine, on the i-th local machine, we output

$$Z_i = \text{sign}(X_i - \hat{\theta}_{i-1}),$$

and update the current state $\hat{\theta}$ by

$$\hat{\theta}_i = \hat{\theta}_{i-1} + \hat{\sigma}\gamma_i Z_i$$

where $\gamma_i = i^{-2/3}$.

Step 3: Final estimation of θ. On the central machine, because we have access to $Z_1, Z_2, ..., Z_m$, thus we can calculate $\hat{\theta}_i$ accordingly for all $i = 11, ..., m$. The final estimator of the mean θ is given by

$$\hat{\theta}_{sq} = \frac{1}{m-10}\sum_{i=11}^{m}\hat{\theta}_i.$$

Since each of first 11 local machines outputs at most $\log\frac{3}{\sigma_0} + 1$ bits, and the later local machines only output 1 bit per machine, it is easy to verify that the above proposed estimator $\hat{\theta}_{sq} \in \mathcal{A}_{sq}(11\log\frac{3}{\sigma_0} + m)$. The following theorem gives an upper bound on its mean squared error for all $\sigma \geq \sigma_0$

THEOREM 3. *The estimator $\hat{\theta}_{sq} \in \mathcal{A}_{sq}(11\log\frac{3}{\sigma_0} + m)$ and satisfies*

$$\mathbb{E}(\hat{\theta}_{sq} - \theta)^2 \leq C\left(\frac{\sigma^2}{m} \wedge 1\right),$$

where C is a universal constant not depending on σ_0, σ, θ or m.

That is, the proposed sequential protocol estimator $\hat{\theta}_{sq}$ is rate-optimal for all $\sigma \geq \sigma_0$, whose total communication cost is $\log\frac{3}{\sigma_0} + m$ bits.

REMARK 6. The one-bit passing step of the above estimator $\hat{\theta}_{sq}$ is established in light of the previous work Kipnis and Duchi (2017), where the goal is to construct an estimator using one-bit measurements from local machines. Their proposed estimator was shown to be asymptotically normal. However, the finite sample mean squared error of their estimator was not

guaranteed, as the finite sample performance is significantly influenced by the initial position $\hat{\theta}_{11}$ and the initial step size $\hat{\sigma}$.

In this paper, we introduce the preliminary estimates $\hat{\theta}_{11}$ and $\hat{\sigma}$, which can be obtained with only a small amount of communication cost, as an approximation for the optimal initial position and initial step size. This warm start initialization is the key to obtain finite sample bound in Theorem 3.

The hardcode number "11" in the procedure can be set to any larger constants, but not smaller ones. Because for a technical reason we require the preliminary estimate $\hat{\sigma}$ to have bounded -5 order moment, i.e. $\mathbb{E}\hat{\sigma}^{-5} < \infty$.

REMARK 7. The proof of Theorem 3 extends the techniques developed in the previous seminal work Polyak (1990) on stochastic approximation. Polyak (1990) developed upper bounds for stochastic approximation with averaging. The additional difficulty to prove Theorem 3, compared to the previous work, is to control the uncertainty brought to the estimator $\hat{\theta}_{sq}$ from the random initialization $\hat{\theta}_{11}$ and $\hat{\sigma}$. Much more careful calculation is needed here.

4.2. *Lower bound under interactive protocols.* The above proposed estimator $\hat{\theta}_{sq}$ achieves the optimal rate of convergence for all $\sigma \geq \sigma_0$ with communication cost $(11 \log \frac{3}{\sigma_0} + m)$ bits. The next theorem is a direct corollary derived from Theorem 5 in Cai and Wei (2020). The lower bound argument shows that the communication cost for $\hat{\theta}_{sq}$ cannot be improved.

THEOREM 4. *For any $\hat{\theta} \in \mathcal{A}_{bb}(B)$, if $\hat{\theta}$ is rate-optimal when $\sigma = \sigma_0$, i.e. there is a constant $C > 0$ such that*

$$\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \leq C \left(\frac{\sigma_0^2}{m} \wedge 1 \right).$$

Then there exists a constant $c > 0$ (depends on C) such that

$$B \geq c \left(\log \frac{1}{\sigma_0} + m \right).$$

The above theorem establishes a lower bound on the communication cost for any distributed estimator under the blackboard protocols that achieves optimal rate of convergence when $\sigma = \sigma_0$. The same lower bound also holds for any estimator that achieves the optimal rate of convergence for $\sigma \geq \sigma_0$, as those estimators are under more strict conditions. Recall that the sequential protocols are a subset of the blackboard protocols. Therefore, the lower

bound in Corollary 4, together with the proposed adaptive estimator $\hat{\theta}_{sq}$, shows that order $\log \frac{1}{\sigma_0}+m$ communication cost is necessary and sufficient for rate-optimal adaptive estimation under the interactive protocols, including the sequential and blackboard protocols.

REMARK 8. Recall that for any rate-optimal adaptive estimator under the independent protocols, the minimal communication cost is of order $m \log \frac{1}{\sigma_0}$, which is larger than that for a rate-optimal adaptive estimator under the interactive protocols. Feed-backs and information sharing are necessary to improve communication-efficiency in adaptive Gaussian mean estimation.

REMARK 9. The lower bound on communication cost in Corollary 4 holds for the non-adaptive case when $\sigma = \sigma_0$ is known in advance. Since the adaptive estimator $\hat{\theta}_{sq}$ is constructed with no more communication cost than the non-adaptive case, there is no cost of adaptation for Gaussian mean estimation under the two types of the interactive protocols. In contrast, under the independent protocols, as more communication cost is needed to establish a rate-optimal adaptive estimator, there is a cost of adaptation for Gaussian mean estimation.

TABLE 1
Optimal communication cost for different distributed protocols under adaptive and non-adaptive settings. Adaptive setting: minimal communication cost for rate-optimal adaptive estimator over $\sigma \geq \sigma_0$. Non-adaptive setting: minimal communication cost for rate-optimal estimator with known $\sigma = \sigma_0$.

Protocol	adaptive estimator	non-adaptive estimator
Independent	$O(m \log \frac{1}{\sigma_0})$	$O(m + \log \frac{1}{\sigma_0})$
Sequential	$O(m + \log \frac{1}{\sigma_0})$	$O(m + \log \frac{1}{\sigma_0})$
Blackboard	$O(m + \log \frac{1}{\sigma_0})$	$O(m + \log \frac{1}{\sigma_0})$

5. Numerical Results. The proposed adaptive estimators under independent protocol and under interactive protocols are easy to implement. In this section, we conduct simulation studies to investigate the numerical performance of these two estimators. The numerical results show that the proposed estimators are practically useful, having high statistical accuracy while only requiring a small amount of communication cost.

During the simulation study, we consider a setting where $\sigma_0 = 2^{-12}$, i.e. we know as a prior knowledge that $\sigma \geq \sigma_0 = 2^{-12}$. Assume there are $m = 100$ machines, where each machine has access to a univariate normal variable $X \sim N(\theta, \sigma^2)$, with $\theta = 0.3$ and choices of $\sigma \in \{2^{-2}, 2^{-4}, 2^{-6}, 2^{-8}, 2^{-10}, 2^{12}\}$.

We compare the following three estimators: classical sample-mean estimator (under centralized setting), the adaptive estimator under independent protocol, and the adaptive estimator under sequential protocol. Over 100 random simulations, the average mean squared error of the three different estimators, and the communication cost of the two distributed estimators are given in Table 2.

TABLE 2
Mean squared error and communication cost of the three methods. $\sigma_0 = 2^{-12}$, $m = 100$, $\theta = 0.3$. For distributed estimators, total communication costs (in bits) are given in the parentheses.

σ	Sample-mean	Independent Protocol	Sequential Protocol
2^{-2}	6.17×10^{-4}	4.14×10^{-3}(1500)	2.12×10^{-3}(266)
2^{-4}	4.04×10^{-5}	1.45×10^{-4}(1500)	1.28×10^{-4}(266)
2^{-6}	2.14×10^{-6}	9.02×10^{-6}(1500)	8.31×10^{-6}(266)
2^{-8}	1.46×10^{-7}	5.23×10^{-7}(1500)	4.85×10^{-7}(266)
2^{-10}	8.59×10^{-9}	5.00×10^{-8}(1500)	2.66×10^{-8}(266)
2^{-12}	5.68×10^{-10}	5.10×10^{-9}(1500)	2.47×10^{-9}(266)

As is shown in Table 2, the mean squared errors of adaptive estimators are within 5 to 10 times of the optimal centralized sample-mean estimators. Despite that during the theoretical analysis, the statistical accuracy is proved up to constants with respect to the centralized optimality, the simulation results show that the actual constant gaps are relatively small. Considering their low communication cost, we find the proposed adaptive estimators could be practically useful in real distributed estimation applications.

6. Discussion. We studied in the present paper the problem of distributed adaptive Gaussian mean estimation with unknown σ. In the conventional centralized setting, Gaussian mean estimation with unknown σ is arguably one of the most basic and fundamental problems in classical statistics. As seen in this paper, the theoretical analysis is rich and difficult in the distributed setting.

The insights gained from the analysis can be used to solve other related problems where the variance is unknown. One such problem is nonparametric regression with random design. As pointed out in Cai and Wei (2021), despite being asymptotically equivalent in the centralized setting, the problem of distributed nonparametric regression with random design is significantly different from that with fixed design. For example, when wavelet methods are used, the empirical wavelet coefficients in this case have unknown variance due to the unknown design distribution and the techniques developed in this paper can potentially be used to construct a wavelet estimator in that problem. More discussion on the connections and differences among various

distributed nonparametric function estimation problems can be found in Cai and Wei (2021).

In the present paper, the focus is on the optimal estimation of the mean θ. A closely related problem is statistical inference for the mean including the construction of optimal confidence intervals for θ. This involves optimal estimation of the variance σ^2 in the same setting, which is a challenging problem by itself. We leave the inference problem for future work.

The results in this paper reveal an interesting phenomenon: the communication costs required under different types of distributed protocols can be substantially different. This is in sharp contrast to Gaussian mean estimation with known variance. It is interesting to investigate further the differences among various types of distributed protocols for other distributed statistical problems. It is technically challenging to develop a general optimality theory under different types of communication constraints. More generally, it is of significant interest to understand the interplay between communication cost, statistical accuracy, adaptation, and different types of distributed protocols for a wide range of problems. This is an important topic in data science that is wide open and merits further study.

7. Proofs. We prove the main results in this section. Throughout this section, L_x^1 denotes the L^1 function space with respect to the x variable and $\mathbb{I}_{\{\}}$ denotes the indicator function taking values in $\{0, 1\}$. We use shorthand $a \lesssim b$ to denote there exists a universal constant $C > 0$ such that $a \leq Cb$. With slight abuse of notation, we define ϕ be the standard Gaussian density, ϕ_σ be the density of $N(0, \sigma^2)$, and $\phi_{\theta,\sigma}$ be the density of $N(\theta, \sigma^2)$.

7.1. *Proof of Theorem 1.* We first define several quantities. They play important roles to establish the proof.

Let P, Q be two distributions that are absolutely continuous with respect to a Lebesgue measure on the measurable space $\mathcal{Z}$. p, q are the density functions of P, Q respectively. Define squared Hellinger distance $H^2(P, Q)$ as

$$H^2(P, Q) \triangleq \frac{1}{2} \int_{\mathcal{Z}} (\sqrt{p} - \sqrt{q})^2 \mathrm{d}x.$$

Define total variation distance $TV(P, Q)$ as

$$TV(P, Q) \triangleq \frac{1}{2} \int_{\mathcal{Z}} |p - q| \mathrm{d}x.$$

Let $\mathcal{Z}$ be a finite set, $h : \mathbb{R} \to \mathcal{Z}$ a random function, and $f, g \in L^1(\mathbb{R})$ are non-negative functions. Define "generalized squared Hellinger distance" for

Z be

$$H^2(h;f,g) \triangleq \frac{1}{2}\sum_{z\in\mathcal{Z}}\left(\sqrt{\int_{-\infty}^{\infty} f(x)\mathbb{P}(h(x)=z)dx} - \sqrt{\int_{-\infty}^{\infty} g(x)\mathbb{P}(h(x)=z)\mathrm{d}x}\right)^2.$$

Note that when f, g are densities, $H^2(h;f,g)$ is exactly the squared Hellinger distance between distribution of $h(X)$ when $X\sim f$, and distribution of $h(X)$ when $X\sim g$. This is why we call this quantity generalized squared Hellinger distance.

Similarly, we define "generalized total variation distance" as

$$TV(h;f,g) \triangleq \frac{1}{2}\sum_{z\in\mathcal{Z}}\left|\int_{-\infty}^{\infty} f(x)\mathbb{P}(h(x)=z)\mathrm{d}x - \int_{-\infty}^{\infty} g(x)\mathbb{P}(h(x)=z)\mathrm{d}x\right|.$$

Also when f, g are densities, $TV(h;f,g)$ is exactly the total variation distance between distribution of $h(X)$ when $X\sim f$, and distribution of $h(X)$ when $X\sim g$.

The following lemma provides two basic but useful inequalities for $H^2(h;f,g)$ and $TV(h;f,g)$.

LEMMA 2. *For any random function $h:\mathbb{R}\to\mathcal{Z}$, the following two inequalities hold:*

(a) Sub-additivity of $H^2(h;f,g)$: if $f(x,s), g(x,s)\in L_x^1(\mathbb{R})$ are non-negative functions for each $s\in(s_l,s_r)$, and $\int_{s_l}^{s_r} f(x,s)ds, \int_{s_l}^{s_r} g(x,s)\mathrm{d}s \in L_x^1(\mathbb{R})$. Then we have

$$H^2(h;\int_{s_l}^{s_r} f(\cdot,s)ds, \int_{s_l}^{s_r} g(\cdot,s)ds) \le \int_{s_l}^{s_r} H^2(h;f(\cdot,s),g(\cdot,s))\mathrm{d}s. \tag{1}$$

(b) Bound between $TV(h;f,g)$ and $H^2(h;f,g)$: if f and g have the same support (i.e. $\{x:f(x)>0\}=\{x:g(x)>0\}$) and there exist $M\ge 1$ such that $1/M\le f(x)/g(x)\le M$ for all $x\in\{x:g(x)>0\}$. Then we have

$$H^2(h;f,g) \le \frac{\sqrt{M}-1}{\sqrt{M}+1} TV(h;f,g). \tag{2}$$

Besides, we define $\phi_{\theta,\sigma}$ as the density function of $N(\theta,\sigma^2)$, i.e.

$$\phi_{\theta,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}.$$

Now we move back to the main proof. Let $\lambda = c_\lambda \left(\frac{\log \frac{3}{\sigma_0}}{mB}\right)^{1/4}$ where c_λ is a positive constant that will be specified later. Let J be the maximum integer such that $2^{-J} \geq \sigma_0$. Let $\sigma_j = 2^j \sigma_0, j = 1, 2, ..., J-1$.

We are interested in the following *integrated squared Hellinger distances*:

$$I \triangleq \sum_{i=1}^{m} \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} H^2(\Pi_i; \phi_{(\theta-\lambda\sigma_j),\sigma_j}, \phi_{(\theta+\lambda\sigma_j),\sigma_j}) \mathrm{d}\theta. \tag{3}$$

The following subsection is dedicated to show that under proper choice of the constant c_λ, we have $I \leq \frac{1}{2}m$.

7.1.1. *Bound integrated integrated squared Hellinger distances I.* We first "slice" $\phi_{(\theta-\lambda\sigma_j),\sigma_j}$ and $\phi_{(\theta+\lambda\sigma_j),\sigma_j}$ in (3) into pieces so that we can apply Lemma 2(a) to give an upper bound for I. Let

$$s^* = \sup_{x\in\mathbb{R}} |\phi_{-\lambda,1}(x) - \phi_{\lambda,1}(x)|,$$

$$A(s) = \{x : |\phi_{-\lambda,1}(x) - \phi_{\lambda,1}(x)| \geq s\}, \quad 0 < s < s^*,$$

$$x_s = \sup A(s),$$

$$f(x,s) = \mathbb{I}_{\{x\in A(s)\}} \frac{\phi_{-\lambda,1}(x)}{|\phi_{-\lambda,1}(x) - \phi_{\lambda,1}(x)|}, \quad x \neq 0, 0 < s < s^*,$$

$$g(x,s) = \mathbb{I}_{\{x\in A(s)\}} \frac{\phi_{\lambda,1}(x)}{|\phi_{-\lambda,1}(x) - \phi_{\lambda,1}(x)|}, \quad x \neq 0, 0 < s < s^*.$$

When $x = 0$ we set $f(0,s) = g(0,s) = \phi_{\lambda,1}(0)/s^*$. It is easy to verify that

$$\phi_{-\lambda,1}(x) = \int_0^{s^*} f(x,s)ds \quad \text{and} \quad \phi_{\lambda,1}(x) = \int_0^{s^*} g(x,s)\mathrm{d}s.$$

The reason why we design the function f and g is for a good property: $g(x,s) - f(x,s) = \mathbb{I}_{\{x\in A(s)\}} \operatorname{sign}(x)$, which is a compact supported piecewise function only taking values in $\{-1, 0, 1\}$. By this way we "discretize" the problem and is able to adopt combinatoric techniques (in Lemma 9).

Note that $\phi_{(\theta-\lambda\sigma_j),\sigma_j}(x) = \frac{1}{\sigma_j}\phi_{-\lambda,1}((x-\theta)/\sigma_j)$ and $\phi_{(\theta+\lambda\sigma_j),\sigma_j}(x) = \frac{1}{\sigma_j}\phi_{\lambda,1}((x-\theta)/\sigma_j)$, so we have

$$\phi_{(\theta-\lambda\sigma_j),\sigma_j}(x) = \int_0^{s^*} \frac{1}{\sigma_j} f((x-\theta)/\sigma_j, s)\mathrm{d}s,$$

$$\phi_{(\theta+\lambda\sigma_j),\sigma_j}(x) = \int_0^{s^*} \frac{1}{\sigma_j} g((x-\theta)/\sigma_j, s)\mathrm{d}s.$$

The above equations and Lemma 2(a) implies

(4)
$$I = \sum_{i=1}^{m}\sum_{j=0}^{J-1}\int_{\lambda\sigma_j}^{1-\lambda\sigma_j} H^2(\Pi_i; \phi_{(\theta-\lambda\sigma_j),\sigma_j}, \phi_{(\theta+\lambda\sigma_j),\sigma_j})\mathrm{d}\theta$$

(5)
$$\leq \sum_{i=1}^{m}\sum_{j=0}^{J-1}\int_{\lambda\sigma_j}^{1-\lambda\sigma_j} \mathrm{d}\theta \int_0^{s^*} H^2\left(\Pi_i(x); \frac{1}{\sigma_j} f((x-\theta)/\sigma_j, s), \frac{1}{\sigma_j} g((x-\theta)/\sigma_j, s)\right)\mathrm{d}s$$

(6)
$$= \sum_{i=1}^{m}\sum_{j=0}^{J-1}\frac{1}{\sigma_j}\int_{\lambda\sigma_j}^{1-\lambda\sigma_j} d\theta \int_0^{s^*} H^2\left(\Pi_i(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right)\mathrm{d}s.$$

Note that $f(x,s)$ and $g(x,s)$ both are supported on $A(s)$ and when $x \in A(s)$,

$$f(x,s)/g(x,s) = \phi_{-\lambda,1}(x)/\phi_{\lambda,1}(x) = e^{2\lambda x} \in [e^{-2\lambda x_s}, e^{2\lambda x_s}].$$

Apply Lemma 2(b), we have

$$H^2\left(\Pi_i(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right) \leq \frac{e^{\lambda x_s}-1}{e^{\lambda x_s}+1} TV\left(\Pi_i(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right).$$

Substitute into (4) and apply Fubini's theorem, we get

(7)
$$I \leq \int_0^{s^*} ds \frac{e^{\lambda x_s}-1}{e^{\lambda x_s}+1} \sum_{i=1}^{m}\sum_{j=0}^{J-1}\frac{1}{\sigma_j}\int_{\lambda\sigma_j}^{1-\lambda\sigma_j} TV\left(\Pi_i(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right)\mathrm{d}\theta.$$

The following lemma bridges the partial total variation distances and communication costs, which is crucial to our proof.

LEMMA 3. *If $\Pi : \mathbb{R} \to \{0,1\}^b$ takes value in $\{0,1\}^b$, then there exist a constant $C_1 > 0$ such that*

$$\sum_{j=0}^{J-1}\frac{1}{\sigma_j}\int_{\lambda\sigma_j}^{1-\lambda\sigma_j} TV\left(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right)\mathrm{d}\theta \leq C_1 x_s(1+x_s)\sqrt{J(b\wedge J)}.$$

Another Lemma gives an upper bound on the integral by analysis.

LEMMA 4. *If $\lambda \le 1/6$ then there exists a constant $C_2 > 0$ such that*

$$\int_0^{s^*} \frac{e^{\lambda x_s} - 1}{e^{\lambda x_s} + 1} x_s(1 + x_s)\mathrm{d}s \le C_2\lambda^2.$$

Apply Lemma 3 and 4 on (7), we have

$$I \le C_1C_2\sqrt{J}\lambda^2 \sum_{i=1}^{n} \sqrt{(b_i \wedge J)}.$$

Jensen's inequality implies that $\sum_{i=1}^m \sqrt{b_i} \le m\sqrt{1/m\sum_{i=1}^n b_i} = \sqrt{mB}$, therefore we have

$$I \le C_1C_2\sqrt{mJB}\lambda^2.$$

Recall that $\lambda = c_\lambda \left(\frac{\log \frac{3}{\sigma_0}}{mB}\right)^{1/4}$. Note that $\log \frac{3}{\sigma_0} \le 2J$, so when c_λ is a sufficiently small constant such that $0 < c_\lambda < \frac{1}{\sqrt{8C_1C_2}}$, we have

$$I \le \frac{J}{2}.$$

Recall the definition of I in (3):

$$I = \sum_{j=0}^{J-1} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} \sum_{i=1}^{m} H^2(\Pi_i; \phi_{(\theta-\lambda\sigma_j),\sigma_j}, \phi_{(\theta+\lambda\sigma_j),\sigma_j})\mathrm{d}\theta.$$

The above upper bound $I \le J/2$ holds for any distributed estimator $\hat{\theta}$. Note that we have $B > \frac{\frac{1}{\sigma_0}}{m}$ thus $\lambda\sigma_{J-1} < 1/6$ if we set $c_\lambda < 1/6$. Apply Lemma 1, we can conclude the desired lower bound:

$$R_{ind}(\sigma_0, B) \ge c\lambda^2 m \ge c_1c_\lambda^2\sqrt{\frac{m\log\frac{3}{\sigma_0}}{B}}. \quad \square$$

7.2. *Proof of Theorem 2.* For simplicity of notations we define $Z_{(-)} = Z_{(\lfloor 0.16m \rfloor)}$ and $Z_{(+)} = Z_{(\lceil 0.84m \rceil)}$. Before we proceed to the proof, we give a lemma showing large deviation bounds on $Z_{(-)}$ and $Z_{(+)}$. These bounds can be directly derived using Gaussian tail bounds so we omit the proof.

LEMMA 5. *There exists universal constants $C, c > 0$ such that for any $k \ge 2$, we have*

$$\mathbb{P}(Z_{(-)} < \theta - k\sigma) \le C\exp(-ck^2m),$$

$$\mathbb{P}(Z_{(-)} > \theta - \sigma/2) \le C\exp(-cm),$$
$$\mathbb{P}(Z_{(+)} > \theta + k\sigma) \le C\exp(-ck^2 m),$$
$$\mathbb{P}(Z_{(+)} < \theta + \sigma/2) \le C\exp(-cm).$$

We first define several events:

$$E_1 = \{\theta \notin [Z_{(-)}, Z_{(+)}]\},$$
$$E_2 = \{\theta \in [Z_{(-)}, Z_{(+)}], \hat{\sigma} \notin [\min\{1, \frac{1}{2}\sigma\}, 4\sigma]\},$$
$$E_3 = (E_1 \bigcup E_2)^c = \{\theta \in [Z_{(-)}, Z_{(+)}], \min\{1, \frac{1}{2}\sigma\} \le \hat{\sigma} \le 4\sigma\}.$$

Note that we have

$$\mathbb{E}(\hat{\theta}_q - \theta)^2 = \sum_{k=1}^{3} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_k\}}.$$

Therefore, the proof can be divided into showing $\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_k\}} \le C_k \frac{\sigma^2}{m}$ with some universal constant C_k respectively for $k = 1, 2, 3$.

1. Bound on $\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_1\}}$.

Under E_1, we have either $E_{11} = \{Z_{(-)} > \theta\}$ or $E_{12} = \{Z_{(+)} < \theta\}$ happens.

Define $E_{11,k} = \{Z_{(-)} > \theta, \theta + k\sigma < Z_{(+)} \le \theta + (k+1)\sigma\}$. Under $E_{11,k}$, note that we have $Z_{(+)} - Z_{(-)} \le (k+1)\sigma$, this implies $\hat{\sigma} \le (k+1)\sigma$, then $\tilde{\sigma} \le 2(k+1)\sigma$, thus $R \le \theta + 3(k+1)\sigma$. Note that the final estimate $\hat{\theta}_q$ must lie in the interval $[L, R]$, So we have $|\hat{\theta}_q - \theta| \le 3(k+1)\sigma$ under event $E_{11,k}$.

Apply Lemma 5, when $k = 0, 1$ we have $\mathbb{P}(E_{11,k}) \le \mathbb{P}(E_{11}) \le C\exp(-cm)$. When $k \ge 2$ we have $\mathbb{P}(E_{11,k}) \le \mathbb{P}(Z_{(+)} > \theta + k\sigma) \le C\exp(-ck^2 m)$. Therefore, note that $E_{11} = \bigcup_{k=0}^{\infty} E_{11,k}$, we have

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{11}\}} &\le \sum_{k=0}^{\infty} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_{11,k}\}} \\
&\le (6\sigma)^2 \cdot 2C\exp(-cm) + \sum_{k=2}^{\infty} (3(k+1)\sigma)^2 \cdot C\exp(-ck^2 m) \\
&\le C' \frac{\sigma^2}{m}.
\end{aligned}$$

with some universal constant $C' > 0$.

By a symmetric argument we can also prove that $\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_{12}\}} \le C'\frac{\sigma^2}{m}$. Therefore, we conclude that

$$\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_1\}} \le \mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_{11}\}} + \mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_{12}\}} \le 2C'\frac{\sigma^2}{m}.$$

2. Bound on $\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_2\}}$.

Let $E_{21} = \{\theta \in [Z_{(-)}, Z_{(+)}], \hat{\sigma} < \min\{1, \frac{1}{2}\sigma\}\}$ and $E_{22,k} = \{\theta \in [Z_{(-)}, Z_{(+)}], k\sigma < \hat{\sigma} \le (k+1)\sigma\}$ $(k \ge 4)$.

Under E_{21} we have $|\hat{\theta}-\theta| < \frac{3}{2}\sigma$, under $E_{22,k}$ we have $|\hat{\theta}-\theta| < 3(k+1)\sigma$. Moreover, we have the probability bounds

$$\mathbb{P}(E_{21}) \le \mathbb{P}(Z_{(+)} < \theta + \sigma/2) \le C\exp(-ck^2m),$$

$$\mathbb{P}(E_{22,k}) \le \mathbb{P}(Z_{(+)} > \theta + \frac{k}{2}\sigma) + \mathbb{P}(Z_{(-)} < \theta - \frac{k}{2}\sigma) \le 2C\exp(-ck^2m/4).$$

Note that $E_2 = E_{21}\bigcup\bigcup_{k=4}^{\infty} E_{22,k}$, we have

$$\begin{aligned}\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_2\}} &\le \mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_21\}} + \sum_{k=4}^{\infty}\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_22,k\}}\\ &\le (\frac{3}{2}\sigma)^2\cdot C\exp(-ck^2m) + \sum_{k=4}^{\infty}(3(k+1)\sigma)^2\cdot 2C\exp(-ck^2m/4)\\ &\le C''\frac{\sigma^2}{m}\end{aligned}$$

with some universal constant $C'' > 0$.

3. Bound on $\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_3\}}$.

Under event E_3, because we have $\min\{1, \frac{1}{2}\sigma\} \le \hat{\sigma} \le 4\sigma$, also note that $\hat{\sigma} \le 1$ almost surely, so there are at most 5 possible values of $\tilde{\sigma}$, whose range is between $\min\{1, \frac{1}{2}\sigma\}$ to $\min\{1, 8\sigma\}$ (recall that $\tilde{\sigma}$ is chosen only from powers of 2).

For each possible value of $\tilde{\sigma}$, the length of the interval $[L, R]$ is either $\tilde{\sigma}$ or $2\tilde{\sigma}$. Recall we have requirements that L, R are multiples of $\tilde{\sigma}$ and event E_3 suggest $\theta \in [L, R]$, so there are at most 5 possible values of (L, R) pairs for each possible value of $\tilde{\sigma}$. Putting together, we can conclude that under event E_3, the possible values of the pair (L, R) is at most 25. We use $(L_1, R_1), (L_2, R_2), ..., (L_{25}, R_{25})$ to denote these 25 possible values of the (L,R) pair. Thus we have the following decomposition:

(8)

$$\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_3\}} = \sum_{k=1}^{25}\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{E_3,(L,R)=(L_k,R_k)\}} \le \sum_{k=1}^{25}\mathbb{E}(\hat{\theta}_q-\theta)^2\mathbb{I}_{\{(L,R)=(L_k,R_k)\}}.$$

For each possible pair (L_k, R_k) $(k = 1, 2, ..., 25)$, we have $R_k - L_k \leq 2\tilde{\sigma} \leq 16\sigma$. Define the function $F_k : (-\infty, \infty) \times (0, \infty) \to (0, 1) \times (0, 1)$ as

$$F_k(t, s) = \begin{pmatrix} \Phi\left(\frac{t-L_k}{s}\right) \\ \Phi\left(\frac{R_k-t}{s}\right) \end{pmatrix}.$$

When $(L, R) = (L_k, R_k)$, we have

(9)

$$\mathbb{E}\|F_k(\hat{\theta}_{sq}, \hat{\sigma}_{sq}) - F_k(\theta, \sigma)\|^2 = \mathbb{E}(\max\{\hat{p}_L, \frac{1}{m}\} - \mathbb{P}(X_1 < L))^2 + \mathbb{E}(\max\{\hat{p}_R, \frac{1}{m}\} - \mathbb{P}(X_1 > R))^2 \leq \frac{4}{m}$$

where the last inequality is due to mp_L is binomial distributed with mean $m\mathbb{P}(X_1 < L)$, and mp_R is binomial distributed with mean $m\mathbb{P}(X_1 > R)$.

Note that $R_k - \theta \leq R_k - L_k \leq 16\sigma$, therefore $\frac{t-L_k}{\sigma} < 16$ and $\frac{R_k-t}{\sigma} < 16$ for $t \in [L_k, R_k]$. Then it is easy to prove that there exists a constant $c' = |\frac{d\Phi(x)}{x}|_{x=16}| > 0$ such that for any $t, \in [L_k, R_k]$,

$$\left|\Phi\left(\frac{t - L_k}{\sigma}\right) - \Phi\left(\frac{\theta - L_k}{\sigma}\right)\right| \geq c' \frac{|t - \theta|}{\sigma};$$

$$\left|\Phi\left(\frac{R_k - t}{\sigma}\right) - \Phi\left(\frac{R_k - \theta}{\sigma}\right)\right| \geq c' \frac{|t - \theta|}{\sigma}.$$

Besides, note that for any $t \in [L, R]$ and $s > 0$, at least one of the following inequalities holds:

$$\left|\Phi\left(\frac{t - L_k}{s}\right) - \Phi\left(\frac{\theta - L_k}{\sigma}\right)\right| > \left|\Phi\left(\frac{t - L_k}{\sigma}\right) - \Phi\left(\frac{\theta - L_k}{\sigma}\right)\right|;$$

$$\left|\Phi\left(\frac{R_k - t}{s}\right) - \Phi\left(\frac{R_k - \theta}{\sigma}\right)\right| \geq \left|\Phi\left(\frac{R_k - t}{\sigma}\right) - \Phi\left(\frac{R_k - \theta}{\sigma}\right)\right|.$$

Combine the two observations above, we have

$$\|F_k(\hat{\theta}_{sq}, \hat{\sigma}_{sq}) - F_k(\theta, \sigma)\|^2 \geq \frac{(c')^2}{\sigma^2}(\hat{\theta}_{sq} - \theta)^2.$$

Substitute above inequality into (9), when $(L, R) = (L_k, R_k)$ we have

$$\mathbb{E}(\hat{\theta}_{sq} - \theta)^2 \leq \frac{4}{(c')^2}\frac{\sigma^2}{m}.$$

Substitute the above inequality into (8) we obtained the desired bound

$$\mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{E_3\}} \leq \sum_{k=1}^{25} \mathbb{E}(\hat{\theta}_q - \theta)^2 \mathbb{I}_{\{(L,R)=(L_k,R_k)\}} \leq \frac{100}{(c')^2}\frac{\sigma^2}{m}. \quad \square$$

7.3. *Proof of Theorem 3.* The proof of Theorem 3 will be carried out by several stages. Throughout the proof, we define $\delta_k = \frac{\hat{\theta}_k - \theta}{\sigma}$, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is standard Gaussian density. $\Phi(x) = \mathbb{P}(X > x)$ where $X \sim N(0,1)$, and $\Lambda(x) = 1 - 2\Phi(x)$. We also define $\mu_k = \frac{1}{k-10}$.

We first give a lemma that will be useful in the proof.

LEMMA 6. *Let $\{A_k\}_{k=0}^{\infty}$ be a positive sequence, and $\{b_k\}_{k=0}^{\infty}, \{d_k\}_{k=0}^{\infty}$ be two decreasing positive sequences that satisfy*

$$A_k \le (1 - \alpha b_k) A_{k-1} + \beta b_k d_k, \quad k = 1, 2, ...$$

where $\alpha, \beta > 0$. If there exists $K > 0$ such that

$$\frac{d_{k-1}}{d_k} \le 1 + \frac{\alpha}{2} b_k \text{ for all } k \ge K.$$

Then we have for all $k \ge 0$,

$$A_k \le \left(\frac{A_0 + \beta \sum_{i=1}^{K} b_i d_i}{d_K} + \frac{2\beta}{\alpha} \right) d_k. \tag{10}$$

Then we provide several claims and show the proof to each claim directly after their statement.

CLAIM 1. *There exists a constant $C_1 > 0$ (doesn't depend on θ, σ or $\hat{\sigma}$) such that for all $k \ge 11$, we have*

$$\mathbb{E}[(\hat{\theta}_k - \theta)^2 | \hat{\sigma}] \le C_1 \left((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2 \right) \sigma^2 \gamma_k. \tag{11}$$

PROOF OF CLAIM 1. Define the Lyapunov function

$$L(x) = \begin{cases} x^2 & \text{if } -2 < x < 2 \\ 4e^{|x|/2 - 1} & \text{if } |x| \ge 2 \end{cases}.$$

Note that $x^2 \lesssim L(x)$, therefore to prove Claim 1, it suffices to show that

$$\mathbb{E}\left[L\left(\frac{\hat{\theta}_k - \theta}{\sigma} \right) \middle| \hat{\sigma} \right] \lesssim \left((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2 \right) \gamma_k. \tag{12}$$

We have the following lemma.

LEMMA 7. *(a) If $\gamma_k \hat{\sigma}/\sigma \ge 2$, we have*

$$\mathbb{E}[L(\delta_k) | \hat{\sigma}] \le 11 e^{\gamma_k \hat{\sigma}/(2\sigma)}. \tag{13}$$

(b) If $\gamma_k\hat{\sigma}/\sigma \le 2$ and $k \ge 12$, we have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \le (1 - 0.25\gamma_k\hat{\sigma}/\sigma)\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}] + (\gamma_k\hat{\sigma}/\sigma)^2. \tag{14}$$

Case 1: When $\gamma_k\hat{\sigma}/\sigma \ge 2$. Consider surrogate function

$$\tilde{L}(x) = \begin{cases} 4 & \text{if } 0 \le y \le 4 \\ 4e^{\sqrt{x}/2-1} & \text{if } x \ge 4 \end{cases}.$$

Note that $\tilde{L}(\delta_k^2) \le L(\delta_k) + 4$ and $\tilde{L}(x)$ is convex, apply Lemma 7(a) and Jensen's inequality we have

$$\tilde{L}(\mathbb{E}[\delta_k^2|\hat{\sigma}]) \le \mathbb{E}[\tilde{L}(\delta_k^2)|\hat{\sigma}] \le \mathbb{E}[L(\delta_k)|\hat{\sigma}] + 4 \le 11e^{\gamma_k\hat{\sigma}/(2\sigma)} + 4$$

which suggests that

$$\mathbb{E}[\delta_k^2|\hat{\sigma}] \lesssim (\gamma_k\hat{\sigma}/\sigma)^2 \le (\hat{\sigma}/\sigma)^2\gamma_k.$$

Case 2: When $\gamma_k\hat{\sigma}/\sigma < 2$. Let k_0 be the largest k such that $\gamma_k\hat{\sigma}/\sigma \ge 2$ (if there is no such k, set $k_0 = 0$). Given Lemma 7(b), we can apply Lemma 6 with $A_i = \mathbb{E}[L(\delta_{k_0+i})|\hat{\sigma}]$, $b_i = d_i = \gamma_{k_0+i}\hat{\sigma}/\sigma$, $\alpha = 0.25$, $\beta = 1$, and $K = \lceil 8^3(\hat{\sigma}/\sigma)^{-3}\rceil - k_0$. This is a valid K value because

$$d_{i-1}/d_i = (1-1/k)^{-2/3} \le 1+1/k \le 1+\frac{k^{-2/3}\hat{\sigma}}{8\sigma} = 1+\frac{\alpha}{2}b_k \text{ when } i \ge 8^3(\hat{\sigma}/\sigma)^{-3}$$

where $k = k_0 + i$.

Also note that we have $\gamma_{k_0}\hat{\sigma}/\sigma \le 4$ due to the definition of k_0, thus $A_0 \le 11e^2$ according to Lemma 7(a). And note that $\sum_{i=1}^K b_i d_i < \sum_{i=1}^\infty b_i d_i = (\hat{\sigma}/\sigma)^2 \sum_{i=1+k_0}^\infty \gamma_i^2 < \infty$. Therefore, apply Lemma 6, we have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \lesssim \left(\frac{11e^2 + (\hat{\sigma}/\sigma)^2\sum_{i=1+k_0}^\infty \gamma_i^2}{(\hat{\sigma}/\sigma)^3} + 8\right) d_k \lesssim ((\hat{\sigma}/\sigma)^{-3} + 1)\hat{\sigma}/\sigma\gamma_k.$$

Combine the two cases above, we prove the desired bound (11).

Claim 2. *There exists a constant $C_2 > 0$ (doesn't depend on θ, σ or $\hat{\sigma}$) such that for all $k \ge 11$, we have*

$$\mathbb{E}[(\hat{\theta}_k - \theta)^4|\hat{\sigma}] \le C_2\left((\hat{\sigma}/\sigma)^{-4} + (\hat{\sigma}/\sigma)^4\right)\sigma^4\gamma_k^2. \tag{15}$$

PROOF OF CLAIM 2. The proof is very similar to Claim 1. We will omit some details in the proof. Re-define the Lyapunov function

$$L(x) = \begin{cases} x^4 & \text{if } -2 < x < 2 \\ 16e^{|x|/2-1} & \text{if } |x| \geq 2 \end{cases}.$$

We have the following lemma.

LEMMA 8. *(a) If $\gamma_k\hat{\sigma}/\sigma \geq 2$, we have*

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \leq 44e^{\gamma_k\hat{\sigma}/(2\sigma)}. \tag{16}$$

(b) If $\gamma_k\hat{\sigma}/\sigma \leq 2$ and $k \geq 12$, we have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \leq (1-0.25\gamma_k\hat{\sigma}/\sigma)\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}, \delta_{k-1}]+(6C_1+1)((\hat{\sigma}/\sigma)^4+1)\gamma_k^3. \tag{17}$$

Case 1: When $\gamma_k\hat{\sigma}/\sigma \geq 2$. Similarly we can conclude that

$$\mathbb{E}[\delta_k^4|\hat{\sigma}] \lesssim (\gamma_k\hat{\sigma}/\sigma)^4 \leq (\hat{\sigma}/\sigma)^4\gamma_k^2.$$

Case 2: When $\gamma_k\hat{\sigma}/\sigma < 2$. Let $i = k - k_0$ where k_0 is defined as in the proof of Claim 1. Given Lemma 8(b), we can apply Lemma 6 with $b_i = \gamma_k\hat{\sigma}/\sigma$, $d_i = \gamma_k^2$, $\alpha = 0.25$, $\beta = (6C_1+1)((\hat{\sigma}/\sigma)^{-1} + (\hat{\sigma}/\sigma)^3)$, and $K = \lceil 16^3(\hat{\sigma}/\sigma)^{-3}\rceil - k_0$, then we have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \lesssim \left(\frac{44e^2 + (\hat{\sigma}/\sigma)^4 + 1}{(\hat{\sigma}/\sigma)^4} + 8((\hat{\sigma}/\sigma)^{-1} + (\hat{\sigma}/\sigma)^3)\right)\gamma_k^2 \lesssim ((\hat{\sigma}/\sigma)^{-4}+(\hat{\sigma}/\sigma)^3)\gamma_k^2.$$

Combine the two cases above, and note that $\mathbb{E}[(\hat{\theta}_k - \theta)^4 \lesssim \mathbb{E}[L(\delta_k)|\hat{\sigma}]$, we can conclude (15).

CLAIM 3. *Let $\bar{\theta}_k = \frac{1}{k}\sum_{i=1}^k \hat{\theta}_i$. There exists a constant $C_3 > 0$ (doesn't depend on θ, σ or $\hat{\sigma}$) such that*

$$\mathbb{E}[(\bar{\theta}_k - \theta)^2|\hat{\sigma}] \leq C_3\left((\hat{\sigma}/\sigma)^{-2} + (\hat{\sigma}/\sigma)^2\right)\sigma^2\gamma_k.$$

PROOF OF CLAIM 3. Let $\mu_k = \frac{1}{k-10}$, note that

$$\bar{\theta}_{k+1} - \theta = (1-\mu_k)(\bar{\theta}_k - \theta) + \mu_k(\hat{\theta}_{k+1} - \theta).$$

This implies

$$\mathbb{E}[(\bar{\theta}_{k+1} - \theta)^2|\hat{\sigma}]^{1/2} \leq (1-\mu_k)\mathbb{E}[(\bar{\theta}_k - \theta)^2|\hat{\sigma}]^{1/2} + \mu_k\mathbb{E}[(\hat{\theta}_i - \theta)^2|\hat{\sigma}]^{1/2}. \tag{18}$$

From the above inequality we can show

$$\mathbb{E}[(\bar{\theta}_k-\theta)^2|\hat{\sigma}]^{1/2} \le 3\left(C_1\left((\hat{\sigma}/\sigma)^{-2}+(\hat{\sigma}/\sigma)^2\right)\sigma^2\gamma_k\right)^{1/2} \tag{19}$$

holds for all $k \ge 1$ by induction, which suggest Claim 3 holds with $C_3 = 9C_1$. The induction is concluded by:

1. From Claim 1 we have

   $$\mathbb{E}[(\bar{\theta}_{11}-\theta)^2|\hat{\sigma}]^{1/2} = E[(\hat{\theta}_{11}-\theta)^2|\hat{\sigma}]^{1/2} \le \left(C_1\left((\hat{\sigma}/\sigma)^{-2}+(\hat{\sigma}/\sigma)^2\right)\sigma^2\gamma_{11}\right)^{1/2},$$

   therefore (19) holds when $k = 11$.
2. If (19) holds for k, from (18) and Claim 1 we have

   $$\mathbb{E}[(\bar{\theta}_{k+1}-\theta)^2|\hat{\sigma}]^{1/2} \le \left(3(1-\mu_k)\sqrt{\frac{\gamma_k}{\gamma_{k+1}}}+\mu_k\right)\left(C_1\left((\hat{\sigma}/\sigma)^{-2}+(\hat{\sigma}/\sigma)^2\right)\sigma^2\gamma_{k+1}\right)^{1/2}.$$

   Note that $\sqrt{\frac{\gamma_k}{\gamma_{k+1}}} < (1-\mu_k)^{-1/3}$ and $3(1-\mu_k)^{2/3}+\mu_k \le 3$ for all k. So we have (19) holds for $k+1$.

Claim 4. *Let $\bar{\theta}_k = \frac{1}{k}\sum_{i=1}^k \hat{\theta}_i$. There exists a constant $C_4 > 0$ (doesn't depend on θ, σ or $\hat{\sigma}$) such that*

$$\mathbb{E}[(\bar{\theta}_k-\theta)(\hat{\theta}_k-\theta)|\hat{\sigma}] \le C_4\left((\hat{\sigma}/\sigma)^{-5}+(\hat{\sigma}/\sigma)^3\right)\sigma^2\mu_k. \tag{20}$$

Proof of Claim 4. We have

$$\mathbb{E}[(\bar{\theta}_{k+1}-\theta)(\hat{\theta}_{k+1}-\theta)|\hat{\sigma},\hat{\theta}_k] = (1-\mu_k)\mathbb{E}[(\bar{\theta}_k-\theta)(\hat{\theta}_{k+1}-\theta)|\hat{\sigma},\hat{\theta}_k]+\mu_k\mathbb{E}[(\hat{\theta}_{k+1}-\theta)^2|\hat{\sigma},\hat{\theta}_k].$$

Take expectation with respect to $\hat{\theta}_k$ we have

$$\begin{aligned}
&\mathbb{E}[(\bar{\theta}_{k+1}-\theta)(\hat{\theta}_{k+1}-\theta)|\hat{\sigma}]\\
=&(1-\mu_k)\mathbb{E}[(\bar{\theta}_k-\theta)(\hat{\theta}_k-\theta-\hat{\sigma}\gamma_k\Lambda(\delta_k))|\hat{\sigma}]+\mu_k\mathbb{E}[(\theta_{k+1}-\theta)^2|\hat{\sigma}]\\
=&(1-\mu_k)(1-2\hat{\sigma}/\sigma\gamma_k\phi(0))\mathbb{E}[(\bar{\theta}_k-\theta)(\hat{\theta}_k-\theta)|\hat{\sigma}]+(1-\mu_k)\hat{\sigma}\gamma_k\mathbb{E}[(\bar{\theta}_k-\theta)(2\phi(0)\delta_k-\Lambda(\delta_k))|\hat{\sigma}]\\
&+\mu_k\mathbb{E}[(\hat{\theta}_{k+1}-\theta)^2|\hat{\sigma}].
\end{aligned}$$

Note that $2\phi(0)\delta_k - \Lambda(\delta_k) \lesssim \delta_k^2$. Cauchy-Schwartz inequality suggests

$$\begin{aligned}
\mathbb{E}[(\bar{\theta}_k-\theta)(2\phi(0)\delta_k-\Lambda(\delta_k))|\hat{\sigma}]^2 &\le \mathbb{E}[(\bar{\theta}_k-\theta)^2|\hat{\sigma}]\mathbb{E}[((2\phi(0)\delta_k-\Lambda(\delta_k))^2|\hat{\sigma}]\\
&\lesssim \mathbb{E}[(\bar{\theta}_k-\theta)^2|\hat{\sigma}]\mathbb{E}[\delta_k^4|\hat{\sigma}]\\
&\lesssim \left((\hat{\sigma}/\sigma)^{-2}+(\hat{\sigma}/\sigma)^2\right)\sigma^2\gamma_k\cdot\left((\hat{\sigma}/\sigma)^{-4}+(\hat{\sigma}/\sigma)^4\right)\gamma_k^2\\
&\lesssim \left((\hat{\sigma}/\sigma)^{-6}+(\hat{\sigma}/\sigma)^6\right)\sigma^2\gamma_k^3.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\mathbb{E}[(\bar{\theta}_{k+1}-\theta)(\hat{\theta}_{k+1}-\theta)|\hat{\sigma}] \\
\leq&(1-\mu_k)(1-2\hat{\sigma}/\sigma\gamma_k\phi(0))\mathbb{E}[(\bar{\theta}_k-\theta)(\hat{\theta}_k-\theta)|\hat{\sigma}] \\
&+C_4'(1-\mu_k)\left((\hat{\sigma}/\sigma)^{-3}+(\hat{\sigma}/\sigma)^3\right)\hat{\sigma}\sigma\gamma_k^{5/2} \\
&+\mu_k C_1\left((\hat{\sigma}/\sigma)^{-2}+(\hat{\sigma}/\sigma)^2\right)\sigma^2\gamma_k \\
\leq&(1-2\hat{\sigma}/\sigma\gamma_k\phi(0))\mathbb{E}[(\bar{\theta}_k-\theta)(\hat{\theta}_k-\theta)|\hat{\sigma}]+(C_4'+C_1)\left((\hat{\sigma}/\sigma)^{-2}+(\hat{\sigma}/\sigma)^4\right)\sigma^2\mu_k\gamma_k.
\end{aligned}$$

with some constant $C_4' > 0$. The last inequality is due to the fact that $\gamma_k^{5/2} \leq \mu_k\gamma_k$.

Now we apply Lemma 6 with $b_k = \hat{\sigma}/\sigma\gamma_k$, $d_k = \mu_k$, $\alpha = 2\phi(0)$, $\beta = (C_4'+C_1)\left((\hat{\sigma}/\sigma)^{-3}+(\hat{\sigma}/\sigma)^3\right)\sigma^2$, and $K = (\phi(0)\hat{\sigma}/\sigma)^{-3}$. We have

$$\mathbb{E}[(\bar{\theta}_k-\theta)(\hat{\theta}_k-\theta)|\hat{\sigma}] \lesssim \left(\frac{\sigma^2+\left((\hat{\sigma}/\sigma)^{-2}+(\hat{\sigma}/\sigma)^4\right)\sigma^2}{(\hat{\sigma}/\sigma)^3}+\left((\hat{\sigma}/\sigma)^{-3}+(\hat{\sigma}/\sigma)^3\right)\sigma^2\right)\mu_k.$$

Then we can conclude (20).

CLAIM 5. *There exists a constant $C_5 > 0$ such that*

$$\mathbb{E}[(\bar{\theta}_k-\theta)^2|\hat{\sigma}] \leq C_5\left((\hat{\sigma}/\sigma)^{-5}+(\hat{\sigma}/\sigma)^3\right)\sigma^2\mu_k. \tag{21}$$

PROOF OF CLAIM 5. Note that we have

$$\begin{aligned}
&E[(\bar{\theta}_{k+1}-\theta)^2|\hat{\sigma}] \\
=&(1-\mu_k)^2E[(\bar{\theta}_k-\theta)^2|\hat{\sigma}]+2\mu_k(1-\mu_k)E[(\bar{\theta}_k-\theta)(\hat{\theta}_{k+1}-\theta)|\hat{\sigma}]+\mu_k^2E[(\hat{\theta}_{k+1}-\theta)^2|\hat{\sigma}] \\
\leq&(1-2\mu_k+\mu_k^2)E[(\bar{\theta}_k-\theta)^2|\hat{\sigma}] \\
&+2C_4\left((\hat{\sigma}/\sigma)^{-5}+(\hat{\sigma}/\sigma)^3\right)\sigma^2\mu_k^2 \\
&+C_3\left((\hat{\sigma}/\sigma)^{-2}+(\hat{\sigma}/\sigma)^2\right)\sigma^2\gamma_k\mu_k^2 \\
\leq&(1-2\mu_k+\mu_k^2)E[(\bar{\theta}_k-\theta)^2|\hat{\sigma}]+(2C_4+C_3)\left((\hat{\sigma}/\sigma)^{-5}+(\hat{\sigma}/\sigma)^3\right)\sigma^2\mu_k^2
\end{aligned}$$

which implies there exists a constant $C_5 > 0$ such that

$$E[(\bar{\theta}_{k+1}-\theta)^2|\hat{\sigma}] \leq C_5\left((\hat{\sigma}/\sigma)^{-5}+(\hat{\sigma}/\sigma)^3\right)\sigma^2\mu_k.$$

PROOF OF THE THEOREM. Now we are ready to prove the theorem. Take expectation on (21) with respect to $\hat{\sigma}$, we have

$$\mathbb{E}[(\bar{\theta}_k-\theta)^2] \leq C_5\mathbb{E}[\left((\hat{\sigma}/\sigma)^{-5}+(\hat{\sigma}/\sigma)^3\right)]\sigma^2\mu_k.$$

Note that $\hat{\sigma}$ is the empirical estimate of σ over 10 observations. Thus we have $\mathbb{E}((\hat{\sigma}/\sigma)^{-5}) < \infty$ and $\mathbb{E}(\hat{\sigma}/\sigma)^3 < \infty$, therefore there exists a constant $C_6 > 0$ such that

$$\mathbb{E}[(\bar{\theta}_k - \theta)^2] \leq C_6 \sigma^2 \mu_k.$$

Substitute $k = m$ into the above equation, also note that $(\hat{\theta}_{sq} - \theta)^2 \leq (\bar{\theta}_k - \theta)^2$ and $(\hat{\theta}_{sq} - \theta)^2 \leq 1$, we conclude the theorem. □

7.4. *Proof of Lemma 1.* Note that the length of each integral interval $(\lambda\sigma_j, 1 - \lambda\sigma_j)$ is at least $2/3$. Therefore, for any distributed estimator $\hat{\theta}$, there exists $0 \leq j^* \leq J - 1$ and $\theta^* \in [0, 1]$ such that

$$\sum_{i=1}^{m} H^2(Z_i|X_i \sim N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2), Z_i|X_i \sim N(\theta^* + \lambda\sigma_{j^*}, \sigma_{j^*}^2)) \leq \frac{3}{4}$$

where $Z_i|X_i \sim P$ denotes the distribution of $\Pi_i(X_i)$ when $X_i \sim P$, H^2 denotes the squared Hellinger distances.

Note that $Z_i, i = 1, 2, ..., m$ are independent, by sub-additivity of squared Hellinger distances for product measures, we have

$$H^2((Z_i)_{i=1}^m|_{X_i \sim N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2), \text{ for } i=1,2,...,m}, (Z_i)_{i=1}^m|_{X_i \sim N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2), \text{ for } i=1,2,...,m}) \leq \frac{3}{4}$$

where $(Z_i)_{i=1}^m$ is a shorthand for $(Z_1, Z_2, ..., Z_m)$

Note that the distributed estimator $\hat{\theta}$ is a (possibly random) function of $(Z_i)_{i=1}^m$. Given that the squared Hellinger distance between the distribution of $(Z_i)_{i=1}^m$ under $N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2)$ and $N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2)$ are bounded by $3/4$, which means we cannot "distinguish" whether the data are drawn from $N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2)$ or $N(\theta^* - \lambda\sigma_{j^*}, \sigma_{j^*}^2)$ by looking at those transcripts, we can apply Le Cam's method to conclude a minimax lower bound: when $\sigma = \sigma_{j^*}$, there exists a constant $c_1 > 0$ such that

$$\sup_{\theta \in \{\theta^* - \lambda\sigma_{j^*}, \theta^* + \lambda\sigma_{j^*}\}} \mathbb{E}(\hat{\theta} - \theta)^2 \geq c_1 \lambda^2 \sigma^{*2},$$

which is equivalent to

$$\sup_{\theta \in \{\theta^* - \lambda\sigma_{j^*}, \theta^* + \lambda\sigma_{j^*}\}} \left(\frac{\sigma^{*2}}{m}\right)^{-1} \mathbb{E}(\hat{\theta} - \theta)^2 \geq c_1 \lambda^2 m.$$

Thus we can conclude that

$$R_{ind}(\sigma_0, B) \geq c\lambda^2 m. \quad \square$$

7.5. *Proof of Lemma 2.* PROOF OF (1). For any $z \in \mathcal{Z}$, define

$$F_z(s) = \int_{-\infty}^{\infty} f(x,s)\mathbb{P}(h(x)=z)\mathrm{d}x,$$

$$G_z(s) = \int_{-\infty}^{\infty} g(x,s)\mathbb{P}(h(x)=z)\mathrm{d}x.$$

By definition, we have

$$\begin{aligned} H^2(h;\int_{s_l}^{s_r} f(\cdot,s)\mathrm{d}s, \int_{s_l}^{s_r} g(\cdot,s)\mathrm{d}s) &= \frac{1}{2}\sum_{z\in\mathcal{Z}}\left(\sqrt{\int_{s_l}^{s_r} F_z(s)\mathrm{d}s} - \sqrt{\int_{s_l}^{s_r} G_z(s)\mathrm{d}s}\right)^2 \\ &= \frac{1}{2}\sum_{z\in\mathcal{Z}}\left(\int_{s_l}^{s_r} F_z(s)\mathrm{d}s + \int_{s_l}^{s_r} G_z(s)\mathrm{d}s - 2\sqrt{\int_{s_l}^{s_r} F_z(s)\mathrm{d}s \int_{s_l}^{s_r} G_z(s)\mathrm{d}s}\right), \end{aligned}$$

$$\begin{aligned} \int_{s_l}^{s_r} H^2(h;f(\cdot,s),g(\cdot,s))\mathrm{d}s &= \frac{1}{2}\sum_{z\in\mathcal{Z}}\int_{s_l}^{s_r}\left(\sqrt{F_z(s)} - \sqrt{G_z(s)}\right)^2\mathrm{d}s \\ &= \frac{1}{2}\sum_{z\in\mathcal{Z}}\left(\int_{s_l}^{s_r} F_z(s)\mathrm{d}s + \int_{s_l}^{s_r} G_z(s)\mathrm{d}s - 2\int_{s_l}^{s_r}\sqrt{F_z(s)G_z(s)\mathrm{d}s}\right). \end{aligned}$$

Therefore, from Cauchy-Schwartz inequality

$$\sqrt{\int_{s_l}^{s_r} F_z(s)\mathrm{d}s \int_{s_l}^{s_r} G_z(s)\mathrm{d}s} \geq \int_{s_l}^{s_r}\sqrt{F_z(s)G_z(s)\mathrm{d}s},$$

we can conclude (1).

PROOF OF (2). Define

$$F_z = \int_{-\infty}^{\infty} f(x)\mathbb{P}(h(x)=z)\mathrm{d}x,$$

$$G_z = \int_{-\infty}^{\infty} g(x)\mathbb{P}(h(x)=z)\mathrm{d}x.$$

$1/M \leq f(x)/g(x) \leq M$ for all $x \in \{x : g(x) > 0\}$ implies that $1/M \leq F_z/G_z \leq M$ when $F(z) > 0$ or $G(z) > 0$. This suggests that

$$\left|\frac{\sqrt{F_z}-\sqrt{G_z}}{\sqrt{F_z}+\sqrt{G_z}}\right| \leq \frac{\sqrt{M}-1}{\sqrt{M}+1}.$$

By definition we have

$$H^2(h;f,g) = \frac{1}{2}\sum_{z\in\mathcal{Z}}(\sqrt{F_z}-\sqrt{G_z})^2 = \frac{1}{2}\sum_{z\in\mathcal{Z}}\left|\frac{\sqrt{F_z}-\sqrt{G_z}}{\sqrt{F_z}+\sqrt{G_z}}\right||\sqrt{F_z}-\sqrt{G_z}| \leq \frac{\sqrt{M}-1}{\sqrt{M}+1}TV(h;f,g). \quad \square$$

7.6. *Proof of Lemma 3.* First, note that by definition we have

$$TV\left(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right) \leq x_s \sigma_j.$$

So we have

$$\sum_{j=0}^{J-1} \frac{1}{\sigma_j} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} TV\left(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right) \mathrm{d}\theta \leq x_s J.$$

Therefore, it only remains to prove

$$\sum_{j=0}^{J-1} \frac{1}{\sigma_j} \int_{\lambda\sigma_j}^{1-\lambda\sigma_j} TV\left(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right) \mathrm{d}\theta \leq C_1 x_s (1+x_s)\sqrt{Jb}.$$

The next technical lemma is the key to prove Lemma 3.

LEMMA 9. *Let $\{a_k\}, k = 1, 2, ..., 2^J$ be a non-negative sequence such that*

$$0 \leq a_k \leq 1, k = 1, 2, ..., 2^J.$$

Then there exists a constant $C_3 > 0$ such that

$$\sum_{j=1}^{J} \sum_{l=1}^{2^{J-j}} \left| \sum_{k=(l-1)2^j+1}^{(l-1)2^j+2^{j-1}} a_k - \sum_{k=(l-1)2^j+2^{j-1}+1}^{l\cdot 2^j} a_k \right| \leq C_3 2^J \sqrt{J} \int_0^w \sqrt{-\log t}\mathrm{d}t$$

where $w = 2^{-J}\sum_{k=1}^{2^J} a_k$ is the mean of the sequence.

Let $x'_s = \inf_{x\in A(s)} |x|$. For any real number θ, $z \in \{0,1\}^b$ and $k \in [2^J]$, let $a_k(\theta, z) = \int_{\theta+(k-1)x_s\sigma_0}^{\theta+kx_s\sigma_0} \mathbb{P}(\Pi(x) = z)\mathrm{d}x$ and $a'_k(\theta, z) = \int_{\theta+(k-1)x'_s\sigma_0}^{\theta+kx'_s\sigma_0} \mathbb{P}(\Pi(x) =$

$z)\mathrm{d}x$. Note that it is easy to check $A(s) = [-x_s, -x'_s] \cup [x'_s, x_s]$, so we have

$$
\begin{aligned}
&TV\left(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right)\\
=&\frac{1}{2}\sum_{z\in\{0,1\}^b}\left|\int_{\theta-\sigma_j x_s}^{\theta-\sigma_j x'_s}\mathbb{P}(\Pi(x)=z)\mathrm{d}x - \int_{\theta+\sigma_j x'_s}^{\theta+\sigma_j x_s}\mathbb{P}(\Pi(x)=z)\mathrm{d}x\right|\\
\le&\frac{1}{2}\sum_{z\in\{0,1\}^b}\left|\int_{\theta-\sigma_j x_s}^{\theta}\mathbb{P}(\Pi(x)=z)\mathrm{d}x - \int_{\theta}^{\theta+\sigma_j x_s}\mathbb{P}(\Pi(x)=z)\mathrm{d}x\right|\\
&+\frac{1}{2}\sum_{z\in\{0,1\}^b}\left|\int_{\theta-\sigma_j x'_s}^{\theta}\mathbb{P}(\Pi(x)=z)\mathrm{d}x - \int_{\theta}^{\theta+\sigma_j x'_s}\mathbb{P}(\Pi(x)=z)\mathrm{d}x\right|\\
=&\frac{1}{2^{J-j}}\sum_{z\in\{0,1\}^b}\sum_{r=1}^{2^{J-j-1}}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a_k(\theta-\sigma_j x_s - 2^{j+1}(r-1)x_s\sigma_0, z)\right.\\
&\qquad\left.-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a_k(\theta-\sigma_j x_s - 2^{j+1}(r-1)x_s\sigma_0, z)\right|\\
&+\frac{1}{2^{J-j}}\sum_{z\in\{0,1\}^b}\sum_{r=1}^{2^{J-j-1}}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j} a'_k(\theta-\sigma_j x'_s - 2^{j+1}(r-1)x'_s\sigma_0, z)\right.\\
&\qquad\left.-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r} a'_k(\theta-\sigma_j x'_s - 2^{j+1}(r-1)x'_s\sigma_0, z)\right|.
\end{aligned}
$$

Substitute the above inequality and rewrite the integral variable, also

recall that $\sigma_j = 2^j\sigma_0$, we have

(22)

$$\begin{aligned}
&\sum_{j=0}^{J-1}\frac{1}{\sigma_j}\int_{\lambda\sigma_j}^{1-\lambda\sigma_j} TV\left(\Pi(x); f((x-\theta)/\sigma_j, s), g((x-\theta)/\sigma_j, s)\right)\mathrm{d}\theta\\
\leq&\tfrac{1}{2^J\sigma_0}\textstyle\sum_{j=0}^{J-1}\sum_{z\in\{0,1\}^b}\sum_{r=1}^{2^{J-j-1}}\int_{\lambda\sigma_j-\sigma_j x_s-2^{j+1}(r-1)x_s\sigma_0}^{1-\lambda\sigma_j-\sigma_j x_s-2^{j+1}(r-1)x_s\sigma_0}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a_k(\theta,z)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a_k(\theta,z)\right|\mathrm{d}\theta\\
&+\tfrac{1}{2^J\sigma_0}\textstyle\sum_{j=0}^{J-1}\sum_{z\in\{0,1\}^b}\sum_{r=1}^{2^{J-j-1}}\int_{\lambda\sigma_j-\sigma_j x'_s-2^{j+1}(r-1)x'_s\sigma_0}^{1-\lambda\sigma_j-\sigma_j x'_s-2^{j+1}(r-1)x'_s\sigma_0}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a'_k(\theta,z)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a'_k(\theta,z)\right|\mathrm{d}\theta\\
\leq&\frac{1}{2^J\sigma_0}\sum_{j=0}^{J-1}\sum_{z\in\{0,1\}^b}\sum_{r=1}^{2^{J-j-1}}\int_{-x_s}^{1}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a_k(\theta,z)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a_k(\theta,z)\right|\mathrm{d}\theta\\
&+\frac{1}{2^J\sigma_0}\sum_{j=0}^{J-1}\sum_{z\in\{0,1\}^b}\sum_{r=1}^{2^{J-j-1}}\int_{-x'_s}^{1}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a'_k(\theta,z)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a'_k(\theta,z)\right|\mathrm{d}\theta\\
=&\frac{x_s\sigma_0}{2^J\sigma_0}\int_{-x_s}^{1}\mathrm{d}\theta\sum_{z\in\{0,1\}^b}\sum_{j=0}^{J-1}\sum_{r=1}^{2^{J-j-1}}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a_k(\theta,z)/(x_s\sigma_0)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a_k(\theta,z)/(x_s\sigma_0)\right|\\
&+\frac{x'_s\sigma_0}{2^J\sigma_0}\int_{-x'_s}^{1}\mathrm{d}\theta\sum_{z\in\{0,1\}^b}\sum_{j=0}^{J-1}\sum_{r=1}^{2^{J-j-1}}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a'_k(\theta,z)/(x'_s\sigma_0)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a'_k(\theta,z)/(x'_s\sigma_0)\right|.
\end{aligned}$$

Define $w(\theta,z)\triangleq 2^{-J}\sum_{k=1}^{2^J}a_k(\theta,z)/(x_s\sigma_0)=\frac{1}{2^J x_s\sigma_0}\int_\theta^{\theta+2^J x_s\sigma_0}\mathbb{P}(\Pi(x)=z)$. Note that $a_k(\theta,z)/(x_s\sigma_0)\in[0,1]$, apply Lemma 9 gives

$$\begin{aligned}
&\sum_{z\in\{0,1\}^b}\sum_{j=0}^{J-1}\sum_{r=1}^{2^{J-j-1}}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a_k(\theta,z)/(x_s\sigma_0)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a_k(\theta,z)/(x_s\sigma_0)\right|\\
\leq&C_3 2^J\sqrt{J}\sum_{z\in\{0,1\}^b}\int_0^{w(\theta,z)}\sqrt{-\log t}\mathrm{d}t.
\end{aligned}$$

Then note that $\int_0^w\sqrt{-\log t}\mathrm{d}t$ is a concave function of w and $\sum_{z\in\{0,1\}^b}w(\theta,z)=1$, we can apply Jensen's inequality to get

$$\sum_{z\in\{0,1\}^b}\int_0^{w(\theta,z)}\sqrt{-\log t}\mathrm{d}t\leq 2^b\int_0^{2^{-b}}\sqrt{-\log t}\mathrm{d}t.$$

It is not difficult to prove that there exists a constant $C_{1,1}$ such that

$$\int_0^{2^{-b}}\sqrt{-\log t}\mathrm{d}t\leq C_{1,1}2^{-b}\sqrt{b}.$$

Combine the three inequalities above we can conclude

$$\frac{x_s\sigma_0}{2^J\sigma_0}\int_{-x_s}^{1}\mathrm{d}\theta\sum_{z\in\{0,1\}^b}\sum_{j=0}^{J-1}\sum_{r=1}^{2^{J-j-1}}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a_k(\theta,z)/(x_s\sigma_0)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a_k(\theta,z)/(x_s\sigma_0)\right|$$
$$\leq C_3C_{1,1}x_s\int_{-x_s}^{1}\sqrt{Jb}\mathrm{d}\theta=C_3C_{1,1}x_s(1+x_s)\sqrt{Jb}.$$

By a similar argument we also have

$$\frac{x'_s\sigma_0}{2^J\sigma_0}\int_{-x'_s}^{1}\mathrm{d}\theta\sum_{z\in\{0,1\}^b}\sum_{j=0}^{J-1}\sum_{r=1}^{2^{J-j-1}}\left|\sum_{k=2^{j+1}(r-1)+1}^{2^{j+1}(r-1)+2^j}a_k(\theta,z)/(x'_s\sigma_0)-\sum_{k=2^{j+1}(r-1)+2^j+1}^{2^{j+1}r}a_k(\theta,z)/(x'_s\sigma_0)\right|$$
$$\leq C_3C_{1,1}x'_s(1+x'_s)\sqrt{Jb}.$$

Substitute the above two inequalities into (22) and note that $x'_s\leq x_s$, we conclude Lemma 3. □

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to "Distributed Adaptive Gaussian Mean Estimation with Unknown Variance: Interactive Protocol Helps Adaptation"
(doi: url to be specified). In this supplementary material, we present proofs for several technical lemmas, namely Lemmas 4,6,7,8, and 9.

References.

Acharya, J., Canonne, C. L., and Tyagi, H. (2020). Distributed signal detection under communication constraints. In *Conference on Learning Theory*, pages 41–63. PMLR.

Barnes, L. P., Han, Y., and Özgür, A. (2019a). Fisher information for distributed estimation under a blackboard communication protocol. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2704–2708. IEEE.

Barnes, L. P., Han, Y., and Özgür, A. (2019b). Learning distributions from their samples under communication constraints. *CoRR*, abs/1902.02890.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352.

Bickel, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.*, 9(6):1301–1309.

Braverman, M., Garg, A., Ma, T., Nguyen, H. L., and Woodruff, D. P. (2016). 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.

Cai, T. T. and Low, M. G. (2006). Adaptive confidence balls. *The Annals of Statistics*, 34:202–228.

Cai, T. T. and Wei, H. (2020). Distributed Gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *arXiv preprint arXiv:2001.08877*.

Cai, T. T. and Wei, H. (2021). Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation.

Deisenroth, M. and Ng, J. W. (2015). Distributed Gaussian processes. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1481–1490.

Diakonikolas, I., Grigorescu, E., Li, J., Natarajan, A., Onak, K., and Schmidt, L. (2017). Communication-efficient distributed learning of discrete distributions. In *Advances in Neural Information Processing Systems*, pages 6391–6401.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

Fan, J., Wang, D., Wang, K., and Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031.

Garg, A., Ma, T., and Nguyen, H. (2014). On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734.

Han, Y., Özgür, A., and Weissman, T. (2018). Geometric lower bounds for distributed parameter estimation under communication constraints. *arXiv preprint arXiv:1802.08417*.

Johnstone, I. M. (2017). *Gaussian Estimation: Sequence and Wavelet Models*. Manuscript.

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

Kipnis, A. and Duchi, J. C. (2017). Mean estimation from adaptive one-bit measurements. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1000–1007.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816.

Kushilevitz, E. (1997). Communication complexity. In *Advances in Computers*, volume 44, pages 331–360. Elsevier.

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

Liu, J. (2021). A few interactions improve distributed nonparametric estimation, optimally.

Polyak, B. (1990). New stochastic approximation type procedures. *Avtomatica i Telemekhanika*, 7:98–107.

Szabó, B. and van Zanten, H. (2018). Adaptive distributed methods under communication constraints. *arXiv preprint arXiv:1804.00864*.

Szabo, B. and van Zanten, H. (2020). Distributed function estimation: adaptation using minimal communication. *arXiv preprint arXiv:2003.12838*.

Szabó, B., Vuursteen, L., and van Zanten, H. (2020). Optimal distributed testing in high-dimensional gaussian models. *CoRR*, abs/2012.04957.

Xiang, Y. and Kim, Y.-H. (2013). Interactive hypothesis testing against independence. In *2013 IEEE International Symposium on Information Theory*, pages 2840–2844. IEEE.

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

Zhu, Y. and Lafferty, J. (2018). Distributed nonparametric regression under communication constraints. *arXiv preprint arXiv:1803.01302*.

Submitted to the Annals of Statistics
arXiv: 0000.0000

SUPPLEMENT TO "DISTRIBUTED ADAPTIVE GAUSSIAN MEAN ESTIMATION WITH UNKNOWN VARIANCE: INTERACTIVE PROTOCOL HELPS ADAPTATION" †

By T. Tony Cai, and Hongji Wei

University of Pennsylvania

We present in this supplement the detailed proofs of Lemmas 4,6,7,8,9 in the paper "Distributed Adaptive Gaussian Mean Estimation with Unknown Variance: Interactive Protocol Helps Adaptation".

1. Proof of Lemma 4. Note that when $0 < x < 1/\lambda$ we have $e^{\lambda x} - e^{-\lambda x} < 3\lambda x$, thus

$$\phi_{\lambda,1}(x) - \phi_{-\lambda,1}(x) = e^{-\frac{x^2+\lambda^2}{2}}(e^{\lambda x} - e^{-\lambda x}) < 3\lambda x e^{-x^2/2}.$$

When $x \geq 1/\lambda$, note that $\lambda < 1/6$, we have

$$\phi_{\lambda,1}(x) - \phi_{-\lambda,1}(x) < \phi_{\lambda,1}(x) = e^{-\frac{(x-\lambda)^2}{2}} \leq e^{-\frac{(1/\lambda-\lambda)^2}{2}} < \lambda.$$

Therefore, by definition of s^* and the above two inequalities, we have

$$s^* = \sup_x |\phi_{\lambda,1}(x) - \phi_{-\lambda,1}(x)| < 3(\sup_x x e^{-x^2/2})\lambda = 3\lambda/e.$$

When $e^{-\frac{(1/\lambda-\lambda)^2}{2}} < s \leq s^*$, it is easy to see that $x_s > 1/\lambda$, thus we have

$$s = \phi_{\lambda,1}(x_s) - \phi_{-\lambda,1}(x_s) < 3\lambda x_s e^{-x_s^2/2}.$$

Let $y = y(t) \geq 1$ be the root to the equation: $ye^{-y^2/2} = t$, then we can conclude

$$x_s \leq y\left(\frac{s}{3\lambda}\right) \text{ when } e^{-\frac{(1/\lambda-\lambda)^2}{2}} < s \leq s^*. \tag{23}$$

When $s \leq e^{-\frac{(1/\lambda-\lambda)^2}{2}}$, we have

$$s = \phi_{\lambda,1}(x_s) - \phi_{-\lambda,1}(x_s) < e^{-\frac{(x_s-\lambda)^2}{2}}$$

†The research was supported in part by NSF Grant DMS-2015259 and NIH grants R01-GM129781 and R01-GM123056.

then we can conclude

$$x_s \le \sqrt{-2\log s} + \lambda \text{ when } s \le e^{-\frac{(1/\lambda-\lambda)^2}{2}}. \tag{24}$$

Now we are ready to prove the original equality in Lemma 4. Note that

$$\int_0^{s^*} \frac{e^{\lambda x_s}-1}{e^{\lambda x_s}+1} x_s(1+x_s)\mathrm{d}s \lesssim \int_0^{s^*} \lambda(x_s^2+x_s^3)\mathrm{d}s = \lambda \int_0^{e^{-\frac{(1/\lambda-\lambda)^2}{2}}} (x_s^2+x_s^3)\mathrm{d}s + \lambda \int_{e^{-\frac{(1/\lambda-\lambda)^2}{2}}}^{s^*} (x_s^2+x_s^3)\mathrm{d}s.$$

For the first part, apply (24), we have

$$\lambda \int_0^{e^{-\frac{(1/\lambda-\lambda)^2}{2}}} (x_s^2+x_s^3)\mathrm{d}s \lesssim \lambda \int_0^{e^{-\frac{(1/\lambda-\lambda)^2}{2}}} (\sqrt{-2\log s}+\lambda)^3\mathrm{d}s \lesssim \lambda \frac{1}{\lambda^3} e^{-\frac{(1/\lambda-\lambda)^2}{2}} \lesssim \lambda^2.$$

For the second part, apply (23), we have

$$\begin{aligned}\lambda \int_{e^{-\frac{(1/\lambda-\lambda)^2}{2}}}^{x_s} (x_s^2+x_s^3)\mathrm{d}s &< \lambda \int_0^{s^*} (y(\frac{s}{3\lambda})^2 + y\left(\frac{s}{3\lambda}\right)^3)\mathrm{d}s = 3\lambda^2 \int_0^{3\lambda s^*} (y(t)^3+y(t)^2)\mathrm{d}t \\ &\le 3\lambda^2 \int_0^{1/e} (y(t)^3+y(t)^2)\mathrm{d}t.\end{aligned}$$

Recall that $y(t)$ is the root to $ye^{-y^2/2} = t$ thus the integral $\int_0^{1/e}(y(t)^3 + y(t)^2)\mathrm{d}t < \infty$. Therefore the second part is also $O(\lambda^2)$.

Combine the two parts together we have proved that

$$\int_0^{s^*} \frac{e^{\lambda x_s}-1}{e^{\lambda x_s}+1} x_s(1+x_s)\mathrm{d}s \lesssim \lambda^2. \quad \square$$

2. Proof of Lemma 6. When $k \le K$, note that we have $A_k \le A_{k-1} + \beta b_k d_k$, thus

$$A_k \le A_0 + \beta \sum_{i=1}^{k} b_i d_i \le \frac{A_0 + \beta \sum_{i=1}^{K} b_i d_i}{d_K} d_k.$$

So the inequality (10) is proved. When $k \ge K+1$, we prove by induction. Suppose (10) holds for $k-1$, that is,

$$A_{k-1} \le C_K d_{k-1}$$

where $C_K = \frac{A_0+\beta\sum_{i=1}^{K} b_i d_i}{d_K} + \frac{2\beta}{\alpha}$. Then we have

$$\begin{aligned}A_k &\le (1-\alpha b_k)A_{k-1} + \beta b_k d_k \le (1-\alpha b_k) C_K d_k (1+\frac{\alpha}{2}b_k) + \beta b_k d_k \\ &< \left(C_K(1-\frac{\alpha}{2}b_k) + \beta b_k\right) d_k \le C_k d_k\end{aligned}$$

where the last inequality is due to the fact that $\frac{\alpha}{2}C_K \geq b_k$.

So (10) also holds for k. Note that the starting point $k = K$ is proved in the case $k \leq K$, so we conclude that when $k \geq K+1$ the inequality (10) holds. □

3. Proof of Lemma 7. We prove (13) by induction. Note that $\hat{\theta}_{11}$ is quantized version of $X_{11} \sim N(\theta, \sigma^2)$ with approximation error at most σ_0 (which is less than σ). Also note that $\gamma_{11}\hat{\sigma}/\sigma \geq \gamma_k\hat{\sigma}/\sigma \geq 2$ and $L(x) \leq 4e^{|x/2|-1}$. Thus we have

$$\mathbb{E}[L(\delta_{11})|\hat{\sigma}] \leq \mathbb{E}[4e^{\frac{|X_{11}-\theta|+\sigma}{2\sigma}-1}] < 8e^{-1/4} < 11e^{\gamma_{11}\hat{\sigma}/(2\sigma)},$$

so (13) holds when $k = 11$.

Next let's assume (13) holds for $k-1$, i.e.

$$\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}] \leq 11e^{\gamma_{k-1}\hat{\sigma}/(2\sigma)}.$$

We have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}, \hat{\theta}_{k-1}] = (1-\Phi(\delta_{k-1}))L(\delta_{k-1}-\gamma_k\hat{\sigma}/\sigma) + \Phi(\delta_{k-1})L(\delta_{k-1}+\gamma_k\hat{\sigma}/\sigma).$$

When $|\delta_{k-1}| \leq \gamma_k\hat{\sigma}/\sigma$, note that $L(x) \leq 4e^{|x/2|-1}$, we have

$$\begin{aligned}
\mathbb{E}[L(\delta_k)|\hat{\sigma}, \delta_{k-1}] &\leq \frac{4}{e}(1-\Phi(|\delta_{k-1}|))e^{(\gamma_k\hat{\sigma}/\sigma-|\delta_{k-1}|)/2} + \frac{4}{e}\Phi(|\delta_{k-1}|)e^{(\gamma_k\hat{\sigma}/\sigma+|\delta_{k-1}|)/2} \\
&\leq \frac{4}{e}e^{\gamma_k\hat{\sigma}/(2\sigma)} \sup_{x>0}\Big((1-\Phi(x))e^{-x/2} + \Phi(x)e^{x/2}\Big) \\
&< \frac{4}{e}e^{\gamma_k\hat{\sigma}/(2\sigma)} \cdot 1.10 < 2e^{\gamma_k\hat{\sigma}/(2\sigma)}.
\end{aligned}$$

When $|\delta_{k-1}| > \gamma_k\hat{\sigma}/\sigma$, note that $\Phi(|\delta_{k-1}|) < e^{-\delta_{k-1}^2/2}$, we have

$$\begin{aligned}
\mathbb{E}[L(\delta_k)|\hat{\sigma}, \delta_{k-1}] &\leq \frac{4}{e}(1-\Phi(|\delta_{k-1}|))e^{(|\delta_{k-1}|-\gamma_k\hat{\sigma}/\sigma)/2} + \frac{4}{e}\Phi(|\delta_{k-1}|)e^{(\gamma_k\hat{\sigma}/\sigma+|\delta_{k-1}|)/2} \\
&\leq \frac{4}{e}e^{|\delta_{k-1}|/2-\gamma_k\hat{\sigma}/(2\sigma)} + \frac{4}{e}e^{\gamma_k\hat{\sigma}/(2\sigma)+|\delta_{k-1}|/2-\delta_{k-1}^2/2} \\
&\leq \frac{4}{e}e^{|\delta_{k-1}|/2}(e^{-\gamma_k\hat{\sigma}/(2\sigma)} + e^{\gamma_k\hat{\sigma}/(2\sigma)-(\gamma_k\hat{\sigma}/\sigma)^2/2}) \\
&\leq 2\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}, \delta_{k-1}]e^{-\gamma_k\hat{\sigma}/(2\sigma)}
\end{aligned}$$

where the last inequality is due to $\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}, \delta_{k-1}] = 4e^{|\delta_{k-1}|/2-1}$ and $\gamma_k\hat{\sigma}/(2\sigma) - (\gamma_k\hat{\sigma}/\sigma)^2/2 < -\gamma_k\hat{\sigma}/(2\sigma)$ because $\gamma_k\hat{\sigma}/\sigma \geq 2$.

Combine the two inequalities above and take expectation with respect to δ_{k-1}, apply the induction assumption for $k-1$, also note that $\gamma_{k-1} < 2\gamma_k - 0.9\gamma_k$, we have

$$\begin{aligned}\mathbb{E}[L(\delta_k)|\hat{\sigma}] &\le 2e^{-\gamma_k\hat{\sigma}/(2\sigma)}\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}] + 2e^{\gamma_k\hat{\sigma}/(2\sigma)} \\ &\le 22e^{(\gamma_{k-1}-\gamma_k)\hat{\sigma}/(2\sigma)} + 2e^{\gamma_k\hat{\sigma}/(2\sigma)} \\ &< 22e^{\gamma_k\hat{\sigma}/(2\sigma)} \cdot e^{-0.9\gamma_k\hat{\sigma}/(2\sigma)} + 2e^{\gamma_k\hat{\sigma}/(2\sigma)} \\ &= (22e^{-0.9} + 2)e^{\gamma_k\hat{\sigma}/(2\sigma)} < 11e^{\gamma_k\hat{\sigma}/(2\sigma)}.\end{aligned}$$

Therefore the proof of (13) is completed by induction on k.

To prove (14), we consider conditional expectation on δ_{k-1}.

Case 1: When $|\delta_{k-1}| \le 2$, note that $|\delta_k - \delta_{k-1}| = \gamma_k\hat{\sigma}/\sigma < 2$, this implies $|\delta_k| \le 4$. In this range we have $L(\delta_k) \le \delta_k^2$. So when $|\delta_{k-1}| \le 2$, we have

$$\begin{aligned}\mathbb{E}[L(\delta_k)|\hat{\sigma}, \delta_{k-1}] &\le (1-\Phi(\delta_{k-1}))(\delta_{k-1} - \gamma_k\hat{\sigma}/\sigma)^2 + \Phi(\delta_{k-1})(\delta_{k-1} + \gamma_k\hat{\sigma}/\sigma)^2 \\ &= \delta_{k-1}^2 - 2\gamma_k\hat{\sigma}/\sigma(1-2\Phi(\delta_{k-1}))\delta_{k-1} + (\gamma_k\hat{\sigma}/\sigma)^2 \\ &\le (1-0.95\gamma_k\hat{\sigma}/\sigma)\delta_{k-1}^2 + (\gamma_k\hat{\sigma}/\sigma)^2 \\ &= (1-0.95\gamma_k\hat{\sigma}/\sigma)\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}, \delta_{k-1}] + (\gamma_k\hat{\sigma}/\sigma)^2\end{aligned}$$

where the last inequality is due to the fact that $(1-2\Phi(\delta_{k-1}))\delta_{k-1} > 0.475\delta_{k-1}^2$ when $\delta_{k-1} \le 2$.

Case 2: When $|\delta_{k-1}| > 2$, without loss of generality we assume $\delta_{k-1} > 2$. The opposite case $\delta_{k-1} < -2$ has the same result due to symmetry. Note that $L(x) \le 4e^{|x/2|-1}$, we have

$$\begin{aligned}\mathbb{E}[L(\delta_k)|\hat{\sigma}, \delta_{k-1}] &\le (1-\Phi(\delta_{k-1}))4e^{(\delta_{k-1}-\gamma_k\hat{\sigma}/\sigma)/2-1} + \Phi(\delta_{k-1})4e^{(\delta_{k-1}+\gamma_k\hat{\sigma}/\sigma)/2-1} \\ &= 4e^{\delta_{k-1}/2-1}\left((1-\Phi(\delta_{k-1}))e^{-\gamma_k\hat{\sigma}/(2\sigma)} + \Phi(\delta_{k-1})e^{\gamma_k\hat{\sigma}/(2\sigma)}\right).\end{aligned}$$

Note that we have $\delta_{k-1} > 2$ and $0 < \gamma_k\hat{\sigma}/\sigma \le 2$, this implies

$$(1-\Phi(\delta_{k-1}))e^{-\gamma_k\hat{\sigma}/(2\sigma)} + \Phi(\delta_{k-1})e^{\gamma_k\hat{\sigma}/(2\sigma)} \le 1 - 0.25\gamma_k\hat{\sigma}/\sigma.$$

Substitute into above inequality we have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}, \delta_{k-1}] \le (1-0.25\gamma_k\hat{\sigma}/\sigma)\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}, \delta_{k-1}].$$

Combine the two cases above we can conclude (14). □

4. Proof of Lemma 8. The proof of Lemma 8(a) is identical to the proof of Lemma 7(a) except that all the upper bounds are multiplied by 4. To prove Lemma 8(b), we take conditional expectation on δ_{k-1}, and divide the proof into two cases:

Case 1: When $|\delta_{k-1}| \le 2$, similarly we have $|\delta_k| \le 4$, in this range we have $L(\delta_k) \le \delta_k^4$. So in this case we have

$$\begin{aligned}\mathbb{E}[L(\delta_k)|\hat{\sigma}, \delta_{k-1}] &\le (1-\Phi(\delta_{k-1}))(\delta_{k-1} - \gamma_k\hat{\sigma}/\sigma)^4 + \Phi(\delta_{k-1})(\delta_{k-1} + \gamma_k\hat{\sigma}/\sigma)^4 \\ &\le \delta_{k-1}^4 - 4\gamma_k\hat{\sigma}/\sigma(1-2\Phi(\delta_{k-1}))\delta_{k-1}^3 + 6(\gamma_k\hat{\sigma}/\sigma)^2\delta_{k-1}^2 + (\gamma_k\hat{\sigma}/\sigma)^4 \\ &\le (1-0.95\gamma_k\hat{\sigma}/\sigma)\delta_{k-1}^4 + 6(\gamma_k\hat{\sigma}/\sigma)^2\delta_{k-1}^2 + (\gamma_k\hat{\sigma}/\sigma)^4 \\ &= (1-0.95\gamma_k\hat{\sigma}/\sigma)\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}, \delta_{k-1}] + 6(\gamma_k\hat{\sigma}/\sigma)^2\delta_{k-1}^2 + (\gamma_k\hat{\sigma}/\sigma)^4.\end{aligned}$$

Case 2: When $|\delta_{k-1}| > 2$, identical to proof of 7(b) we have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}, \delta_{k-1}] \le (1-0.25\gamma_k\hat{\sigma}/\sigma)\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}, \delta_{k-1}].$$

Combine the two cases above and take expectation with respect to δ_{k-1}, we have

$$\mathbb{E}[L(\delta_k)|\hat{\sigma}] \le (1-0.25\gamma_k\hat{\sigma}/\sigma)\mathbb{E}[L(\delta_{k-1})|\hat{\sigma}, \delta_{k-1}] + 6(\gamma_k\hat{\sigma}/\sigma)^2\mathbb{E}[\delta_{k-1}^2|\hat{\sigma}] + (\gamma_k\hat{\sigma}/\sigma)^4.$$

and finally we can conclude (17) by substitute the upper bound of $\mathbb{E}[\delta_{k-1}^2|\hat{\sigma}]$ given in Claim 1. □

5. Proof of Lemma 9. We prove this lemma by induction. We intend to prove the lemma by assuming that Lemma 9 holds when the number J in the statement is replaced by $J-1$.

Note that by the assumption of induction we have

$$\begin{aligned}&\sum_{j=1}^{J}\sum_{l=1}^{2^{J-j}}\left|\sum_{k=(l-1)2^j+1}^{(l-1)2^j+2^{j-1}} a_k - \sum_{k=(l-1)2^j+2^{j-1}+1}^{l\cdot 2^j} a_k\right| \\ =&\sum_{j=1}^{J-1}\sum_{l=1}^{2^{J-1-j}}\left|\sum_{k=(l-1)2^j+1}^{(l-1)2^j+2^{j-1}} a_k - \sum_{k=(l-1)2^j+2^{j-1}+1}^{l\cdot 2^j} a_k\right| \\ &+\sum_{j=1}^{J-1}\sum_{l=1}^{2^{J-1-j}}\left|\sum_{k=(l-1)2^j+1}^{(l-1)2^j+2^{j-1}} a_{k+2^{J-1}} - \sum_{k=(l-1)2^j+2^{j-1}+1}^{l\cdot 2^j} a_{k+2^{J-1}}\right| \\ &+\left|\sum_{k=1}^{2^{J-1}} a_k - \sum_{k=2^{J-1}+1}^{2^J} a_k\right| \\ \le& C_3\cdot 2^{J-1}\sqrt{J-1}\left(\int_0^u\sqrt{-\log t}\mathrm{d}t + \int_0^v\sqrt{-\log t}\mathrm{d}t\right) + 2^{J-1}|u-v|.\end{aligned}$$

where $u = 2^{-(J-1)}\sum_{k=1}^{2^{J-1}} a_k$ is the mean of the first half of the sequence and $v = 2^{-(J-1)}\sum_{k=2^{J-1}+1}^{2^J} a_k$ is the mean of the second half. Therefore, in order to prove the lemma it suffices to show that

$$\sup_{u+v=2w, 0\le u,v\le 1} C_3\cdot 2^{J-1}\sqrt{J-1}\left(\int_0^u \sqrt{-\log t}\mathrm{d}t + \int_0^v \sqrt{-\log t}\mathrm{d}t\right) + 2^{J-1}|u-v| \le C_3\cdot 2^J\sqrt{J}\int_0^w \sqrt{-\log t}\mathrm{d}t.$$

Let $\epsilon = w-u$, the above inequality is equivalent to

$$\sup_{0\le\epsilon\le w\wedge(1-w)} G_{J,w}(\epsilon) \le C_3\cdot 2\sqrt{J}\int_0^w \sqrt{-\log t}\mathrm{d}t \tag{25}$$

where

$$G_{J,w}(\epsilon) \triangleq C_3\cdot\sqrt{J-1}\left(\int_0^{(w-\epsilon)} \sqrt{-\log t}\mathrm{d}t + \int_0^{(w+\epsilon)} \sqrt{-\log t}\mathrm{d}t\right) + 2\epsilon, \quad 0\le\epsilon\le w\wedge(1-w).$$

Note that $G_{J,w}(\epsilon)$ is concave as a function of ϵ, thus it takes the maximum value at a unique point $\epsilon^*\in[0, w\wedge(1-w)]$. Also by concavity of $(\sqrt{-\log t})' = -\frac{\log e}{2x\sqrt{-\log t}}$ we have

$$\frac{dG_{J,w}(\epsilon)}{d\epsilon} = C_3\sqrt{J-1}\left(\sqrt{-\log(w+\epsilon)} - \sqrt{-\log(w-\epsilon)}\right) + 2 \le -C_3\sqrt{J-1}\frac{2\epsilon\log e}{2w\sqrt{-\log w}} + 2.$$

The above inequality implies when $\epsilon > \frac{2w\sqrt{-\log w}}{C_3\sqrt{J-1}\log e}$, we have $\frac{dG_{J,w}(\epsilon)}{d\epsilon} < 0$. This in turn implies $\epsilon^* \le \frac{2w\sqrt{-\log w}}{C_3\sqrt{J-1}\log e}$. Therefore, we can conclude

$$\sup_{0\le\epsilon\le w\wedge(1-w)} G_{J,w}(\epsilon) = G_{J,w}(\epsilon^*) \le 2C_3\sqrt{J-1}\int_0^w \sqrt{-\log t}\mathrm{d}t + 2\frac{2w\sqrt{-\log w}}{C_3\sqrt{J-1}\log e} \le \left(2C_3\sqrt{J-1} + \frac{4}{C_3\sqrt{J-1}\log e}\right)\int_0^w \sqrt{-\log t}\mathrm{d}t.$$

By choosing C_3 large enough, the inequality $2C_3\sqrt{J-1} + \frac{4}{C_3\sqrt{J-1}\log e} \le 2C_3\sqrt{J}$ can be made to hold for all $J\ge 2$. Therefore we conclude (25). □

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tcai@wharton.upenn.edu
hongjiw@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~tcai/