DISTRIBUTED GAUSSIAN MEAN ESTIMATION UNDER COMMUNICATION CONSTRAINTS: OPTIMAL RATES AND COMMUNICATION-EFFICIENT ALGORITHMS†

By T. Tony Cai and Hongji Wei

University of Pennsylvania

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

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

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

In the age of big data, communication/computational concerns associated with a statistical procedure are becoming increasingly important in contemporary applications. One of the difficulties for analyzing large datasets is

†The research was supported in part by NSF Grant DMS-1712735 and NIH grants R01-GM129781 and R01-GM123056.

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

Keywords and phrases: Communication constraints, distributed learning, Gaussian mean, minimax lower bound, optimal rate of convergence
that data are distributed, instead of in a single centralized location. This setting arises naturally in many statistical practices.

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

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

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

For large-scale data analysis, communications between machines can be slow and expensive and limitation on bandwidth and communication sometimes becomes the main bottleneck on statistical efficiency. It is therefore necessary to take communication constraints into consideration when constructing statistical procedures. When the communication budget is limited, the algorithm must carefully “compress” the information contained in the data as efficiently as possible, leading to a trade-off between communication costs and statistical accuracy. The precisely quantification of this trade-off is an important and challenging problem.
Estimation of a Gaussian mean occupies a central position in parametric statistical inference. In the present paper we consider distributed Gaussian mean estimation under the communication constraints in both the univariate and multivariate settings. Although optimal estimation of a Gaussian mean is a relatively simple problem in the conventional setting, this problem is quite involved under the communication constraints, both in terms of the construction of the rate optimal distributed estimator and the lower bound argument. Optimal distributed estimation of a Gaussian mean also serves as a starting point for investigating other more complicated statistical problems in distributed learning including distributed nonparametric function estimation, distributed high-dimensional linear regression, and distributed large-scale multiple testing.

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

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

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

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

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

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

In addition to the independent protocol, there are other more general and interactive distributed protocols including the sequential protocol and
blackboard protocol, which are two popular communication protocols considered in the literature (Zhang et al., 2013a; Barnes et al., 2019). We shall first focus on the independent protocol, then introduce the sequential and blackboard protocols and establish optimality results for these two types of distributed protocols in Section 4.

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

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

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

As usual, the estimation accuracy of a distributed estimator $\hat{\theta}$ is measured by the mean squared error (MSE), $\mathbb{E}_{P_{\theta}}\|\hat{\theta} - \theta\|^2_2$, where the expectation is taken over the randomness in both the data and construction of the transcripts and estimator. As in the conventional decision theoretical framework,
a quantity of particular interest in distributed learning is the minimax risk for the distributed protocols
\[
\inf_{\hat{\theta} \in A_{\text{ind}}(b_{1:m})} \sup_{P_0 \in \mathcal{P}} \mathbb{E}_{P_0} \| \hat{\theta} - \theta \|^2,
\]
which characterizes the difficulty of the distributed learning problem under the communication constraints \(b_{1:m}\). As mentioned earlier, in a rigorous decision theoretical formulation of distributed learning, the communication constraints are essential. Without the constraints, one can always output the original data from the local machines to the central machine and the problem is then reduced to the usual centralized setting.

1.2. Distributed estimation of a univariate Gaussian mean. We first consider distributed estimation of a univariate Gaussian mean under the communication constraints \(b_{1:m}\), where \(P_0 = N(\theta, \sigma^2)\) with \(\theta \in [0,1]\) and the variance \(\sigma^2\) known. Set \(\sigma_n = \sigma/\sqrt{n}\). Note that by a sufficiency argument, one can estimate \(\theta\) based on the sample means \(X_i \triangleq \frac{1}{n} \sum_{j=1}^{n} X_{i,j}\) on the local machines, and the problem is the same as if one only observes \(X_i \sim N(\theta, \sigma_n^2)\) on the \(i\)-th machine, for \(i = 1, \ldots, m\).

Our analysis in Section 2 establishes the following minimax rate of convergence for distributed univariate Gaussian mean estimation under the communication constraints \(b_{1:m}\),

\[
\inf_{\hat{\theta} \in A_{\text{ind}}(b_{1:m})} \sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \asymp \begin{cases} 2^{-2B} & \text{if } B < \log_2 \frac{1}{\sigma_n} + 2 \\ \frac{\sigma_n^2}{(B - \log_2 \frac{1}{\sigma_n})} & \text{if } \log_2 \frac{1}{\sigma_n} + 2 \leq B < \log_2 \frac{1}{\sigma_n} + m \\ \min \left\{ \frac{\sigma_n^2}{m}, 1 \right\} & \text{if } B \geq \log_2 \frac{1}{\sigma_n} + m \end{cases},
\]

where \(B = \sum_{i=1}^{m} b_i\) is the total communication budget, and \(a \asymp b\) denotes \(cb \leq a \leq Cb\) for some constants \(c, C > 0\). The same optimal rate of convergence holds for the class of sequential protocols and blackboard protocols.

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

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

1. Localization phase. When \(B < \log_2 \frac{1}{\sigma_n} + 2\), as a function of \(B\), the minimax risk decreases fast at an exponential rate. In this phase, having
more communication budget is very beneficial in terms of improving the estimation accuracy.

2. Refinement phase. When \( \log_2 \frac{1}{\sigma_n} + 2 \leq B < \log_2 \frac{1}{\sigma_n} + m \), as a function of \( B \), the minimax risk decreases relatively slowly and is inverse-proportional to the total communication budget \( B \).

3. Optimal-rate phase. When \( B \geq \log_2 \frac{1}{\sigma_n} + m \), the minimax rate does not depend on \( B \), and is the same as in the centralized setting where all the data are combined (Bickel, 1981).

An essential technique for solving this problem is the decomposition of the minimax estimation problem into two steps, localization and refinement. This critical decomposition provides a framework for both the lower bound analysis and optimal procedure design. In the lower bound analysis, the statistical error is decomposed into “localization error” and “refinement error”. It is shown that one of these two terms is inevitably large under the communication constraints. In our optimal procedure called MODGAME, bits of the transcripts are divided into three types: crude localization bits, finer localization bits, and refinement bits. They compress the local data in a way that both the localization and refinement errors can be optimally reduced. Further technical details and discussion are presented in Section 2. Furthermore, it will be shown that MODGAME is also robust against departures from Gaussianity. See Section 5 for a detailed discussion.
1.3. Distributed estimation of a multivariate Gaussian mean. We then consider the multivariate case under the communication constraints \( b_1:m \), where \( P_\theta = N_d(\theta, \sigma^2 I_d) \) with \( \theta \in [0,1]^d \) and the noise level \( \sigma \) is known. As in the univariate case, by a sufficiency argument, it is equivalent to consider distributed estimation where each local machine only observes a local sample mean vector \( X_i \sim N_d(\theta, \sigma^2_n I_d) \), with \( \sigma_n = \sigma/\sqrt{n} \). The goal is to optimally estimate the mean vector \( \theta \) under the squared error loss.

The construction and the analysis given in Section 3 show that the minimax rate of convergence in this case is given by

\[
\inf_{\hat{\theta}} \sup_{\theta \in A_{\text{ind}}(b_1:m)} \mathbb{E}||\hat{\theta} - \theta||_2^2 = \begin{cases} 
2^{-2B/d} & \text{if } B/d < \log_2 \frac{1}{\sigma_n} + 2 \\
\frac{d \min \left\{ \frac{\sigma_n^2}{m'}, 1 \right\}}{(B/d - \log_2 \frac{1}{\sigma_n})} & \text{if } \log_2 \frac{1}{\sigma_n} + 2 \leq B/d < \log_2 \frac{1}{\sigma_n} + \max\{m', 2\} \\
\min \left\{ \frac{\sigma_n^2}{m'}, 1 \right\} & \text{if } B/d \geq \log_2 \frac{1}{\sigma_n} + \max\{m', 2\}
\end{cases}
\]

where \( B = \sum_{i=1}^m b_i \) is the total communication budgets and \( m' = \sum_{i=1}^m \min \left\{ b_i \frac{1}{d}, 1 \right\} \) is the “effective sample size”. The same optimal rate of convergence holds true for the class of sequential protocols or blackboard protocols.

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

There is a gap between the distributed optimal rate and centralized optimal rate if \( m' \ll m \). See Section 3 for further technical details and discussion.

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

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

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

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

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

### 1.5. Our contribution

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

In particular, the unified minimax rate applies to the case $\sigma_n < 1$. In comparison, when $\sigma_n < 1$, the previous results are not sharp even in the high communication budget regime (i.e. refinement phase and optimal-rate phase). See Remarks 5 and 6 for detailed comparison with previous results. This is an important case that arises in many statistical applications including distributed nonparametric regression and sparse signal recovery. Establishing a sharp and complete minimax rate is not only important for distributed Gaussian mean estimation itself, but also fundamental for solving these related problems.

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

1.6. Organization of the paper. We finish this section with notation and definitions that will be used in the rest of the paper. Section 2 studies distributed estimation of a univariate Gaussian mean under communication constraints with independent protocols and Section 3 considers the multivariate case. Section 4 introduces sequential and blackboard protocols and extends the optimality results to these two types of communication protocols. Section 5 considers the robustness of the proposed procedures against departures from Gaussianity. The numerical performance of the proposed distributed estimators is investigated in Section 6 and further research directions are discussed in Section 7. For reasons of space, we prove the lower bound for the univariate case in Section 8 and defer the proofs of the other main results and the technical lemmas to the Supplementary Material (Cai and Wei, 2020).

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

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

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x),$$

$$H(X|Y) \triangleq -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) \log \mathbb{P}(X = x|Y = y),$$

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

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

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

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

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

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

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

which characterizes the difficulty of the estimation problem with independent protocols under the communication constraints. We first focus on the independent protocols. Same results for sequential and blackboard protocols will be established in Section 4.
We first introduce an estimation procedure and provide an upper bound for its performance and then establish a matching lower bound on the minimax risk. The upper and lower bounds together establish the minimax rate of convergence and the optimality of the proposed estimator.

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

2.1.1. MODGAME procedure when \( \sigma_n < 1 \). When \( \sigma_n < 1 \), MODGAME consists of two steps: localization and refinement. Roughly speaking, the first step utilizes \( \log \frac{1}{\sigma_n} + o(B - \log \frac{1}{\sigma_n}) \) bits, out of the total budget \( B = \sum_{i=1}^{m} b_i \) bits, for localization to roughly locate where \( \theta \) is, up to \( O(\sigma_n) \) error. Building on the location information, the remaining \( B - \log \frac{1}{\sigma_n} \) bits are used for refinement to further increase the accuracy of the estimator. Detailed theoretical analysis will show that the optimality of the final estimator.

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

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

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

\[
g_k(x) = \begin{cases} 
0 & \text{if } [2^k \tau_{[0,1]}(x)] \mod 4 = 0 \text{ or } 3 \\
1 & \text{if } [2^k \tau_{[0,1]}(x)] \mod 4 = 1 \text{ or } 2. 
\end{cases}
\]

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

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

To unify the notation we set \( g_k(x) \equiv \bar{g}_k(x) \equiv 0 \) if \( k < 0 \).
It is worth mentioning that these Gray functions mimic the behavior of the Gray codes (for reference see Savage (1997)). Fix $K \geq 1$, if we treat $(g_1(x), g_2(x), \ldots, g_K(x))$ as a string of code for any source $x \in [0, 1]$, then those $x$ within the interval $[2^{-K}(s - 1), 2^{-K}s)$ where $s$ is an integer will match the same code. Moreover, the code for adjacent intervals only differs by one bit, which is also a key feature for the Gray codes. This key feature guarantees the robustness of the Gray codes. Such robustness makes the Gray functions very useful for distributed estimation. An example for $K = 3$ is shown in Figure 3 to better illustrate the behavior of the Gray functions.

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

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

\[
(4) \quad h(x) \triangleq \left[2^{\left\lfloor \log \frac{1}{\sigma_n} \right\rfloor - 7}x \right] \mod 2 \quad \text{and} \quad \bar{h}(x) \triangleq \left[2^{\left\lfloor \log \frac{1}{\sigma_n} \right\rfloor - 7}x - \frac{1}{2} \right] \mod 2.
\]

For any function $f$, define the convolution function

\[
\Phi_f(x) \triangleq \mathbb{E}_{X \sim N(x, \sigma_n^2)} f(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi \sigma_n}} e^{-\frac{(y-x)^2}{2\sigma_n^2}} f(y) dy.
\]
The above refinement functions and convolution function are used to accurately estimate the mean of the Gaussian observations. In the MODGAME procedure, the central machine collects one-bit measurements of some observations, say \( h(X_1), h(X_2), \ldots, h(X_u) \). By definition, the mean of those one-bit measurements is exactly \( \Phi_h(\theta) \). Note that \( \Phi_h(x) \) is a periodic wave-shape function, therefore after locating \( \theta \) to a short interval of length \( O(\sigma_n) \) during the preliminary steps, the central machine obtains a good estimate for \( \theta \) by solving estimating equation \( \Phi_h(\theta) = u^{-1} \sum_{i=1}^u h(X_i) \). A similar communication strategy is also adopted in Braverman et al. (2016).

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

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

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

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

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

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

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

\[
Z_i = (U_{s_{i-1}+1}, U_{s_{i-1}+2}, \ldots, U_{s_{i-1}+b_i}),
\]

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

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

\[
\hat{\theta}_D = \min \{ x : x \in \text{Dec}_B(U_1, U_2, \ldots, U_B) \}.
\]
Case 2: \( \log \frac{1}{\sigma_n} + 2 \leq B \leq \log \frac{1}{\sigma_n} + m \). Let

\[
(5) \quad u \triangleq \max \left\{ s \in \mathbb{N} : \left( \log s \right)^2 + 2s \leq B - \left\lfloor \log \frac{1}{\sigma_n} \right\rfloor \right\},
\]

and define finer localization functions:

\[
(6) \quad f_1(x) \triangleq g_{\left\lfloor \log \frac{1}{\sigma_n} - \left\lfloor \log u \right\rfloor - 2 \right\rfloor}(x),
\]

\[
f_2(x) \triangleq g_{\left\lfloor \log \frac{1}{\sigma_n} - \left\lfloor \log u \right\rfloor - 2 \right\rfloor}(x),
\]

\[
f_k(x) \triangleq g_{\left\lfloor \log \frac{1}{\sigma_n} - \left\lfloor \log u \right\rfloor - 4 + k \right\rfloor}(x) \text{ for } k \geq 3.
\]

In this case the total communication budget is divided into 3 parts: crude localization bits (roughly \( b \log \frac{1}{\sigma_n} \) bits), finer localization bits (\( \left\lfloor \log u \right\rfloor^2 \) bits), and refinement bits (2\( u \) bits). The crude localization bits are the values of the functions \( g_1(\cdot), g_2(\cdot), \ldots, g_{\left\lfloor \log \frac{1}{\sigma_n} \right\rfloor}(\cdot) \), each evaluated on a local sample. We denote those resulting binary bits by \( U_1, U_2, \ldots, U_{\left\lfloor \log \frac{1}{\sigma_n} \right\rfloor} \). The finer localization bits are the values of the functions \( f_1(\cdot), f_2(\cdot), \ldots, f_{\left\lfloor \log u \right\rfloor}(\cdot) \), each function is evaluated on \( b \log \frac{1}{\sigma_n} \) different local samples. The function values of \( f_k(\cdot) \) are denoted by \( W_{k,1}, W_{k,2}, \ldots, W_{k,\left\lfloor \log u \right\rfloor} \). The refinement bits are the values of the function \( h(\cdot) \), evaluated on \( u \) local samples; and the values of the function \( \bar{h}(\cdot) \), evaluated on \( u \) different local samples. The resulting binary bits are denoted by \( V_1, V_2, \ldots, V_n \) and \( \bar{V}_1, \bar{V}_2, \ldots, \bar{V}_n \) respectively.

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

\[
\left\lfloor \log u \right\rfloor^2 + 2u \leq B - \left\lfloor \log \frac{1}{\sigma_n} \right\rfloor \leq m.
\]

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

The procedure can be summarized as follows:

**Step 1:** Generate transcripts on local machines. Define \( s_i = \sum_{j=1}^{i} (b_j - \mathbb{I}_{\left\{ j \leq \left\lfloor \log u \right\rfloor^2 + 2u \right\}}) \) and \( s_0 = 0 \). On the \( i \)-th machine:
Fig 4: An illustration of MODGAME. The bits in the transcripts are transmitted to the central machine and divided into three types: crude localization bits, finer localization bits, and refinement bits. One then constructs on the central machine a crude interval $I_1$, a finer interval $I_2$, and the final estimate $\hat{\theta}_D$ step by step.

- If $(j-1)[\log u] + 1 \leq i \leq j[\log u]$ for some integer $1 \leq j \leq [\log u]$, output
  \[ Z_i = (U_{s_1+1}, U_{s_1+2}, \ldots, U_{s_i-1+b_i-1}, W_{j,i-(j-1)[\log u]}); \]
  (If $b_i = 1$, just output $Z_i = W_{j, i-(j-1)[\log u]}$.)

- If $[\log u]^2 + 1 \leq i \leq [\log u]^2 + u$, output
  \[ Z_i = (U_{s_1+1}, U_{s_1+2}, \ldots, U_{s_i-1+b_i-1}, V_{i-[\log u]^2}); \]
  (If $b_i = 1$, just output $Z_i = V_{i-[\log u]^2}$.)

- If $[\log u]^2 + u + 1 \leq i \leq [\log u]^2 + 2u$, output
  \[ Z_i = (U_{s_1+1}, U_{s_1+2}, \ldots, U_{s_i-1+b_i-1}, \bar{V}_{i-([\log u]^2+u)}); \]
  (If $b_i = 1$, just output $Z_i = \bar{V}_{i-([\log u]^2+u)}$.)

- If $i \geq [\log u]^2 + 2u + 1$, output
  \[ Z_i = (U_{s_1+1}, U_{s_1+2}, \ldots, U_{s_i-1+b_i}). \]
where the above binary bits are calculated by

\[ U_{s_{i-1}+k} \triangleq g_{s_{i-1}+k}(X_i) \quad \text{for} \quad i = 1, 2, ..., m; \quad k = 1, 2, ..., b_i. \]

\[ W_{j,i-\lfloor \log u \rfloor} \triangleq f_j(X_i) \quad \text{for} \quad j = 1, 2, ..., \lfloor \log u \rfloor - 1; \]

\[ \bar{V}_i-\lfloor \log u \rfloor^2 \triangleq h(X_i) \quad \text{for} \quad i = \lfloor \log u \rfloor^2 + 1, \lfloor \log u \rfloor^2 + 2, ..., \lfloor \log u \rfloor^2 + u. \]

\[ \bar{W}_i-(\lfloor \log u \rfloor^2 + u) \triangleq \bar{h}(X_i) \quad \text{for} \quad i = \lfloor \log u \rfloor^2 + u + 1, ..., \lfloor \log u \rfloor^2 + 2u. \]

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

**Step 2.1:** First, we use crude localization bits \( U_1, U_2, ... U_{\lfloor \log \frac{1}{\sigma_n} \rfloor} \) to roughly locate \( \theta \). The “crude interval” \( I_1 \) will be obtained in this step.

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

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

\[ I_1' \triangleq \text{Dec}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3}(U_1, U_2, ..., U_{\lfloor \log \frac{1}{\sigma_n} \rfloor - \lfloor \log u \rfloor - 3}). \]

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

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

**Step 2.2:** Then, we use finer localization bits to locate \( \theta \) to a smaller interval of length roughly \( O(\sigma_n) \). A ”finer interval” \( I_2 \) will be generated in this step. For any \( 1 \leq k \leq \lfloor \log u \rfloor \), let

\[ W_k = \mathbb{I}_{\left\{ \sum_{j=1}^{\lfloor \log u \rfloor} W_{k,j} \geq \frac{1}{2} \lfloor \log u \rfloor \right\}} \]

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

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

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

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

\[ I_2' \triangleq \text{Dec}_{\lfloor \log \frac{1}{\sigma_n} \rfloor - 4}(W_{\lfloor \log u \rfloor - \lfloor \log \frac{1}{\sigma_n} \rfloor + 5}, W_{\lfloor \log u \rfloor - \lfloor \log \frac{1}{\sigma_n} \rfloor + 6}, ..., W_{\lfloor \log u \rfloor}). \]
Then we further stretch \( I_2' \) to a larger interval \( I_2 \) so that \( I_2 \) will double the length of \( I_2' \):

\[
I_2 \triangleq \left\{ x : d(x, I_2') \leq 2^{-\left(\log \frac{1}{\sigma_n} \right)^{-3}} \right\}.
\]

(c) If \( \left[ \log \frac{1}{\sigma_n} \right] - \left[ \log u \right] \geq 4 \), let

\[
I_2' \triangleq \{ x \in I_1 : f_k(x) = W_k \text{ for all } k = 1, 2, ..., \left[ \log u \right] \}.
\]

Lemma 7 in the Supplementary Material Cai and Wei (2020) shows \( I_2' \) is an interval. Then we further stretch \( I_2' \) to a larger interval \( I_2 \) so that \( I_2 \) will double the length of \( I_2' \):

\[
I_2 \triangleq \left\{ x : d(x, I_2') \leq 2^{-\left(\log \frac{1}{\sigma_n} \right)^{-3}} \right\}.
\]

**Step 2.3:** Finally, we use refinement bits \( V_1, V_2, ..., V_u \) and \( \bar{V}_1, \bar{V}_2, ..., \bar{V}_u \) to get an accurate estimate \( \hat{\theta}_D \). Lemma 8 in the Supplementary Material Cai and Wei (2020) shows that one of the following two conditions must hold:

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

or

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

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

(a) If \( I_2 \subseteq \left[ (2j - \frac{3}{4}) \cdot 2^{-\left(\log \frac{1}{\sigma_n} \right)^{-6}}, (2j + \frac{3}{4}) \cdot 2^{-\left(\log \frac{1}{\sigma_n} \right)^{-6}} \right] \) for some \( j \in \mathbb{Z} \). Then \( \Phi_h(x) \) is a strictly monotone function on \( I_2 \) (proved in Lemma 8 in the Supplementary Material Cai and Wei (2020)). Denote

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

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

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

where \( \tau_{[L_I, R_I]} \) is the truncation function defined in (3).
(b) Otherwise, we have
\[ I_2 \subseteq [(2j + \frac{1}{4}) \cdot 2^{-(\log \frac{1}{\sigma_n})^{-6}}, (2j + \frac{7}{4}) \cdot 2^{-(\log \frac{1}{\sigma_n})^{-6}}] \]
for some \( j \in \mathbb{Z} \). In this case \( \Phi_h(x) \) is a strictly monotone function on \( I_2 \) (proved in Lemma 8 in the Supplementary Material Cai and Wei (2020)). Denote
\[ \bar{L} \triangleq \inf_{x \in I_2} \Phi_h(x) \quad \text{and} \quad \bar{R} \triangleq \sup_{x \in I_2} \Phi_h(x). \]
By monotonicity, \( \Phi_h \) is invertible on \( I_2 \). Let \( \bar{L}, \bar{R} \) be the inverse of \( \Phi_h \), the distributed estimator \( \hat{\theta} \) is given by
\[ \hat{\theta} = \Phi_h^{-1} \left( \tau_{[\bar{L}, \bar{R}]} \left( \frac{1}{m} \sum_{j=1}^{u} \tilde{v}_j \right) \right) \]
where \( \tau_{[\bar{L}, \bar{R}]} \) is the truncation function defined in (3).

**Case 3:** \( B > \log \frac{1}{\sigma_n} + m \). We only need to use part of the total communication budget \( B \) as if we deal with the case \( B = \log \frac{1}{\sigma_n} + m \). To be precise, we can always easily find \( b'_1, b'_2, \ldots, b'_m \) so that 1 \( \leq b'_i \leq b_i \) for \( i = 1, 2, \ldots, m \) and
\[ \sum_{i=1}^{m} b'_i = \log \frac{1}{\sigma_n} + m. \]
Then we can implement the procedure introduced in Case 2 where we let the \( i \)-th machine only output a transcript of length \( b'_i \).

2.1.2. **MODGAME procedure when \( \sigma_n \geq 1 \).** When \( \sigma_n \geq 1 \), each machine only need to output a one-bit measurement to achieve the global optimal rate as if there are no communication constraints. Some related results are available in Kipnis and Duchi (2019). The following procedure is based on the setting when \( b_i = 1 \) for all \( i = 1, \ldots, m \). If \( b_i > 1 \) for some \( i \), then one can simply discard all remain bits so that only one bit is sent by each machine.

Here is the MODGAME procedure when \( \sigma_n \geq 1 \):

**Step 1.** The \( i \)-th machine outputs
\[ Z_i = \begin{cases} 0 & \text{if } X_i < 0 \\ 1 & \text{if } X_i \geq 0 \end{cases}. \]

**Step 2.** The central machine collects \( Z_1, Z_2, \ldots, Z_m \) and estimates \( \theta \) by
\[ \hat{\theta} = \tau_{[0,1]} \left( \sigma_n \Phi^{-1} \left( \frac{1}{m} \sum_{i=1}^{m} Z_i \right) \right) \]
where \( \tau \) is the truncation function defined in (3) and \( \Phi \) is the cumulative distribution function of a standard normal, \( \Phi(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{t^2/2} dt \).

Here \( \Phi^{-1} \) is the inverse of \( \Phi \) and we extend it by defining \( \Phi^{-1}(0) = -\infty \) and \( \Phi^{-1}(1) = \infty \).

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

**Theorem 1.** For given communication budgets \( b_1: m \) with \( b_i \geq 1 \) for \( i = 1, \ldots, m \), let \( B = \sum_{i=1}^{m} b_i \) and let \( \hat{\theta}_D \) be the MODGAME estimate. Then there exists a constant \( C > 0 \) such that

\[
\sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta}_D - \theta)^2 \leq \begin{cases} 
C \cdot 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\
C \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\
C \cdot \left( \frac{\sigma_n^2}{m} \wedge 1 \right) & \text{if } B \geq \log \frac{1}{\sigma_n} + m 
\end{cases}
\]

An interesting and somewhat surprising feature of the upper bound is that it depends on the communication constraints \( b_1:m \) only through the total budget \( B = \sum_{i=1}^{m} b_i \), not the specific value of \( b_1:m \), so long as each machine can transmit at least one bit. The proof of Theorem 1 is provided in the Supplementary Material (Cai and Wei, 2020).

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

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

\[
R_1(b_1:m) \geq \begin{cases} 
c \cdot 2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\
c \cdot \frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\
c \cdot \left( \frac{\sigma_n^2}{m} \wedge 1 \right) & \text{if } B \geq \log \frac{1}{\sigma_n} + m. 
\end{cases}
\]
The key novelty in the lower bound analysis is the decomposition of the statistical risk into localization error and refinement error based on a delicate construction of the following candidate set $G_\delta$:

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

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

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

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

**Remark 1.** The proof of Lemma 1 mainly relies on the strong data processing inequality (Lemma 14 in Cai and Wei (2020)). The strong data processing inequality was originally developed in information theory, for reference see Raginsky (2016). Zhang et al. (2013a) and Braverman et al. (2016) applied this technical tool to obtain lower bounds for distributed mean estimation. However, their lower bounds are not sharp when $\sigma_n$ is very small, due to the fact that the focus was on bounding the refinement error using the strong data processing inequality, but failed to bound the localization error.

Lemma 1 suggests that under the communication constraints $b_{1,m}$, there is an unavoidable trade-off between the mutual information $I(\hat{\theta}; u)$ and $I(\hat{\theta}; v)$. So one of the above two quantities must be “small”. When $I(\hat{\theta}; u)$ (or $I(\hat{\theta}; v)$) is smaller than a certain threshold, it can be shown that the estimator $\hat{\theta}$ cannot accurately estimate $u$ (or $v$), which means the localization error (or the refinement error) is large. Given that one of localization error and refinement error must be larger than a certain value, the desired lower bound follows. A detailed proof of Theorem 2 is given in Section 8.

**Minimax rate of convergence.** Theorems 1 and 2 together yield a sharp minimax rate for distributed univariate Gaussian mean estimation with inde-
pendent protocols:

\[(15)\]

\[
\inf_{\hat{\theta} \in A_{\text{ind}}(b_1:m)} \sup_{\theta \in [0,1]} \mathbb{E}(\hat{\theta} - \theta)^2 \preceq \begin{cases} 
2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\
\frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\
\frac{\sigma_n^2}{m} \land 1 & \text{if } B \geq \log \frac{1}{\sigma_n} + m
\end{cases}
\]

The results also show that MODGAME is rate optimal.

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

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

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

**Remark 3.** If the central machine itself also has an observation, or equivalently if one of the local machines serves as the central machine, then the communication constraints can be viewed as one of \(b_i\) is equal to infinity. This setting is considered in some related literature, for instance, see Jordan et al. (2019). Then according to Theorem 1, MODGAME always achieves the centralized rate \(\frac{\sigma_n^2}{m} \land 1\), as long as at least one bit is allowed to communicate with each local machine.

**Remark 4.** Our analysis on the minimax rate can be generalized to the \(l_r\) loss for any \(r \geq 1\), with suitable modifications on both the lower bound analysis and optimal procedure.
3. Distributed Multivariate Gaussian Mean Estimation. We turn in this section to distributed estimation of a multivariate Gaussian mean under the communication constraints. Similar to the univariate case, suppose we observe on \( m \) local machines i.i.d. random samples:

\[
X_i \sim \mathcal{N}_d(\theta, \sigma_n^2 I_d), \quad i = 1, \ldots, m,
\]

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

\[
\mathcal{P}^d_{\sigma_n} = \left\{ \mathcal{N}_d(\theta, \sigma_n^2 I_d) : \theta \in [0,1]^d \right\},
\]

where \( \theta \in [0,1]^d \) is the mean vector of interest and the noise level \( \sigma_n \) is known. Under the constraints on the communication budgets \( b_1: m \) with \( b_i \geq 1 \) for \( i = 1, \ldots, m \), the goal is to optimally estimate the mean vector \( \theta \) under the squared error loss. We are interested in the minimax risk for the distributed protocol \( A_{\text{ind}}(b_1:m) \):

\[
R_d(b_1:m) = \inf_{\hat{\theta} \in A_{\text{ind}}(b_1:m)} \sup_{\theta \in [0,1]^d} \mathbb{E}\|\hat{\theta} - \theta\|^2.
\]

Another goal is to develop a rate-optimal estimator that satisfies the communication constraints. The multivariate case is similar to the univariate setting, but it also has some distinct features, both in terms of the estimation procedure and the lower bound argument.

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

**Theorem 3.** Suppose \( b_i \geq 1 \) for all \( i = 1, 2, \ldots, m \). Let \( B = \sum_{i=1}^m b_i \) and \( m' = \frac{1}{d} \sum_{i=1}^m (b_i \wedge d) \), then there exists a constant \( c > 0 \) such that

\[
R_d(b_1:m) \geq \begin{cases} 
  c \cdot 2^{-2B/d} & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\
  c \cdot \frac{\sigma_n^2}{B/d - \log \frac{1}{\sigma_n}} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \\
  c \cdot d \left( \frac{\sigma_n^2}{m'} \wedge 1 \right) & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2)
\end{cases}
\]

A detailed proof of Theorem 3 is given in the Supplementary Material (Cai and Wei, 2020).
Remark 5. In the earlier work including Garg et al. (2014); Barnes et al. (2019), a lower bound for distributed Gaussian mean estimation has been established as \( \Omega(\frac{\sigma_n^2 d^2}{B}) \), where \( B \) is the total communication cost. This lower bound is sharp for \( \sigma_n \geq 1 \). However, when \( \sigma_n < 1 \), by showing that the additional and exact \( \log(1/\sigma_n) \) localization bits are necessary for estimating a Gaussian mean, the lower bound can be improved to \( \Omega(\min\{\frac{\sigma_n^2 d^2}{B - d \log(1/\sigma_n)}, \sigma_n^2 d\}) \). The improvement is significant when \( \log(1/\sigma_n)/m \) is bounded away from 0.

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

![Diagram of multi-MODGAME procedure](image)

**Fig 5:** An illustration for multi-MODGAME. Communication budgets are evenly divided into three parts with each part used for estimating a coordinate of \( \theta \) by the MODGAME procedure.

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

\[
\begin{align*}
    b_i &= b_i^{(1)} + b_i^{(2)} + ... + b_i^{(d)} \\
    \text{for all } &i = 1, 2, ...m.
\end{align*}
\]

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

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

\[ Q \triangleq 1, 2, 3, \ldots, d, 1, 2, 3, \ldots, d, 1, 2, 3, \ldots \]

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

\[
\sum_{j=1}^{i} b_j - k \quad \frac{\sum_{j=1}^{i-1} b_j - k}{d}.
\]

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

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

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

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

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

The following result provides a theoretical guarantee for multi-MODGAME.
Theorem 4. Let $B = \sum_{i=1}^{m} b_i$ and $m' = \frac{1}{d} \sum_{i=1}^{m} (b_i \wedge d)$. Then there exists a constant $C > 0$ such that

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

Remark 6. Compared to the state-of-art results in the literature including Braverman et al. (2016), the multi-MODGAME procedure is more communication-efficient and more flexible in communication budget allocation. To be specific, the algorithm proposed in Braverman et al. (2016) achieves the mean squared error $O\left( \frac{\sigma_n^2}{m} \right)$ with the total communication cost of order $\alpha md + d \log^2(\alpha md/\sigma_n)$. In comparison, to achieve the same statistical performance, MODGAME only needs $\alpha md + d \log(1/\sigma_n)$ bits. The difference could be significant when $\sigma_n \ll 1$.

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

The lower and upper bounds given Theorems 3 and 4 together establish the minimax rate for distributed multivariate Gaussian mean estimation:

\[
\inf_{\hat{\theta} \in \mathcal{A}_{ind}(b_1, m)} \sup_{\theta \in [0,1]^d} \mathbb{E}[\|\hat{\theta} - \theta\|^2] \asymp \begin{cases} 
2^{-2B/d} & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\
\frac{d\sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \\
d \left( \frac{\sigma_n^2}{m'} \wedge 1 \right) & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2)
\end{cases}
\]

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

Remark 7. Different from the univariate case, in the multivariate case the minimax rate depends on not only the total communication budget $B$, but also the effective sample size $m'$. How the communication budgets assigned to individual local machines affects the difficulty of the estimation problem. If the communication budgets are tight on some machines, then
one may have $m' \ll m$, which means the centralized minimax rate cannot be achieved even if the total communication budget $B$ is sufficiently large.

Remark 8. The present paper focuses on the unit hypercube $[0, 1]^d$ as the parameter space. A similar analysis can be applied to other “regular” shape constraints, such as a ball or a simplex, and the minimax rate depends on the constraint.

4. Optimal Distributed Estimation with Sequential and Blackboard Protocols. Independent protocols are considered as a “non-interactive” communication strategy, where each machine can only access its own samples. However, feedback could be helpful in the learning process. There are other more general and interactive communication protocols considered in the literature, including the sequential protocols and blackboard protocols (Zhang et al., 2013a; Barnes et al., 2019).

• Sequential protocols. 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, which is at most $b_i$ bits, can depend on the previous transcripts $Z_1, Z_2, ..., Z_{i-1}$.

• Blackboard protocols. 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 the $i$-th local machine is at most $b_i$ bits.

As for distributed estimation with the independent protocols, it is interesting to establish the optimal rates of convergence for the sequential protocols and blackboard protocols. This is also related to a question of both theoretical and practical interest: is feedback useful for distributed Gaussian mean estimation?

Note that any independent protocol can be viewed as a sequential protocol (by ignoring messages provided by the previous machines). Similarly, any sequential protocol can be implemented as a blackboard protocol. Therefore, the upper bounds (13) for the proposed MODGAME procedure and (16) for multi-MODGAME still hold over the class of sequential protocols and blackboard protocols. The question is: Can these bounds be improved by using more sophisticated algorithms?

The answer is no. The following theorem provides a lower bound for $d$-dimensional distributed Gaussian mean estimation with blackboard proto-
clos. We denote by $\mathcal{A}_{\text{seq}}(b_{1:m})$ and $\mathcal{A}_{BB}(b_{1:m})$ the class of sequential protocols and blackboard protocols respectively.

**Theorem 5.** Suppose $b_i \geq 1$ for all $i = 1, 2, ..., m$. Let $B = \sum_{i=1}^{m} b_i$ and $m' = \frac{1}{d} \sum_{i=1}^{m} (b_i \wedge d)$, then there exists a constant $c > 0$ such that

$$\inf_{\hat{\theta} \in \mathcal{A}_{BB}(b_{1:m})} \sup_{\theta \in [0,1]^d} \mathbb{E}\|\hat{\theta} - \theta\|^2 \geq \begin{cases} c \cdot 2^{-2B/d}d & \text{if } B/d < \log \frac{1}{\sigma_n} + 2 \\ c \cdot \frac{d \sigma_n^2}{(B/d - \log \frac{1}{\sigma_n})} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B/d < \log \frac{1}{\sigma_n} + (m' \vee 2) \\ c \cdot d \left( \frac{n^2}{m' + 1} \right) & \text{if } B/d \geq \log \frac{1}{\sigma_n} + (m' \vee 2) \end{cases}$$

The proof of Theorem 5 is also based on the localization-refinement error decomposition. A sketch of the proof is given in the Supplementary material Cai and Wei (2020). Theorem 5 and the upper bound given in (16) together show that the optimal rate of convergence is the same and MODGAME and multi-MODGAME are rate-optimal for the three classes of communication protocols–independent, sequential, and blackboard. To some extent, the results imply that feedback is not necessary to achieve communication-efficiency for distributed Gaussian mean estimation.

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

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

For simplicity we focus on the one-dimensional estimation problem. The multivariate case can be considered as a direct generalization to the univariate case. Let $P_{\theta}$ be a location family where $\theta$ is the mean, and its variance is $\sigma^2$. Denote $P^n_{\theta}$ as the distribution of the mean of $n$ i.i.d copies drawn from $P_{\theta}$. If on each local machine we can access to $n$ i.i.d observations $X_{i,1}, X_{i,2}, ..., X_{i,n} \sim P_{\theta}$, then each machine can take the local sample mean $X_i \triangleq \frac{1}{n} \sum_{j=1}^{n} X_{i,j} \sim P^n_{\theta}$. Even though $P^n_{\theta}$ is a non-Gaussian distribution, the MODGAME procedure can take $X_i$ as inputs to generate a final estimate.
Recall that MODGAME is divided into three steps: crude localization step, finer localization step, and refinement step. During the first two steps, in order to obtain the desired statistical guarantee for the “confidence interval” \( I_2 \), we only need sub-Gaussian tail condition for \( X_i \). During the refinement step, the key is to use \( \Phi_h \) or \( \bar{\Phi}_h \) to generate estimates from the one-bit measurements. If \( X_i \) is not drawn from a Gaussian distribution, there is additional bias that could be controlled under certain conditions.

Let \( TV(\cdot, \cdot) \) denote the total variation distance between two probability distributions. A random variable \( X \) (or a distribution \( P \) where \( X \sim P \)) is called \( v \)-subgaussian if
\[
E \exp(s(X - E X)) \leq \exp\left(\frac{v^2 s^2}{2}\right), \quad \forall s \in \mathbb{R}.
\]

The following theorem shows that when the total variation distance between the distribution \( \bar{P}_n \) of the local sample mean and the Gaussian distribution \( N(\theta, \sigma^2_n) \) is sufficiently small, MODGAME has the same theoretical guarantee as in the Gaussian case. This implies that MODGAME is robust against departures from the Gaussian distribution.

**Theorem 6.** If \( \bar{P}_n^\theta \) is a \( D \sigma_n \)-subgaussian distribution and \( TV(\bar{P}_n^\theta, N(\theta, \sigma^2_n)) \leq \frac{D}{\sqrt{n}} \) for some \( D > 0 \). Then there exists a constant \( C > 0 \) such that

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

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

A sketch of the proof is given in the Supplementary Material Cai and Wei (2020). Note that \( X_i \sim \bar{P}_n^\theta \) is the mean of i.i.d observations in the \( i \)th local machine. The \( L_1 \) Berry-Esseen bound (e.g. (Chen et al., 2010, Corollary 4.2)) suggests \( TV(\bar{P}_n^\theta, N(\theta, \sigma^2_n)) \leq \frac{\mathbb{E}(|X_1 - \theta|/\sigma)^3}{2\sqrt{\pi}} \). If \( X_1 \) is a \( D \sigma \)-subgaussian distribution, then \( \mathbb{E}(|X_1 - \theta|/\sigma)^3 \) is bounded by a constant (depending on \( D \)). Hence the following corollary holds.

**Corollary 7.** If \( P_\theta \) is a \( D \sigma \)-subgaussian distribution, and \( m \leq Dn \) for some \( D > 0 \). Then there exist a constant \( C > 0 \) such that

\[
(19) \sup_{\theta \in [0, 1]} E(\hat{\theta} - \theta)^2 \leq C \cdot \begin{cases} 
2^{-2B} & \text{if } B < \log \frac{1}{\sigma_n} + 2 \\
\frac{\sigma_n^2}{(B - \log \frac{1}{\sigma_n})^2} & \text{if } \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \\
\frac{\sigma_n^2}{m} & \text{if } B \geq \log \frac{1}{\sigma_n} + m
\end{cases}
\]
where $\hat{\theta}$ is the output of the MODGAME procedure. $B = \sum_{i=1}^{m} b_i$ is the total communication cost.

Corollary 7 shows that, if $n/m$ is asymptotically bounded away from 0, then MODGMAE achieves the same statistical performance as in the Gaussian case as long as the observations are drawn from a subgaussian distribution.

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

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

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

In the second setting, we fix $\sigma_n = 2^{-8}$, $b = 5$ and vary the number of machines $m$ from 10 to 40960. Figure 6b plots the MSEs of the three methods. The MSE of MODGAME decreases as number of machine increases and outperforms naive quantization; the MSE of naive quantization remains constant as the quantization error plays a dominant role in the MSE.
Finally, we fix $b = 5$, $m = 100$ and vary the standard deviation $\sigma_n$ from $2^{-1}$ to $2^{-13}$. Figure 6c shows the MSEs of the three estimators. It can be seen that MODGAME is robust for all choices of $\sigma_n$. The difference between the MSE of MODGAME and the optimal MSE for non-distributed sample mean is small. For naive quantization, it is as good as the optimal non-distributed sample mean when $\sigma_n$ is large. However, as seen in the previous experiment, when $\sigma_n$ is small, the MSE of naive quantization is dominated by the quantization error and is much larger than the MSE of MODGAME. In all three settings, it can be seen clearly that the MSE of MODGAME decreases as the communication budgets increases. This is consistent with the theoretical results established in Section 2 and demonstrates the trade-off between the communication costs and statistical accuracy.

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

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

Same as before, we set $b_1 = b_2 = \ldots = b_m = b$, i.e. the communication budgets for all machines are equal. We set $d = 50$, $\sigma_\epsilon = 2^{-8}$, $m = 25$ and assign the communication budgets $b$ for each machine from 2 to 21. The MSEs of different methods are shown in Figure 7b. A phase transition at $b = 10$ can be clearly seen. When $b \leq 10$, the MSE decreases quickly at an exponential rate. When $b > 10$, the decrease becomes relatively slow. This phenomenon is consistent with the theoretical prediction that different phases appear in the convergence rate for multi-MODGAME (Theorem 4).

![Figure 7](image)

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

7. **Discussion.** We established in the present paper a sharp and complete minimax rate that holds for all values of the parameters $d, m, n, \sigma$ in all communication budget regimes under the independent, sequential, and blackboard protocols. A key technique is the decomposition of the minimax estimation problem into two steps, localization and refinement, which appears in both the lower bound analysis and optimal procedure design. The optimality results and techniques developed can be useful for solving other problems such as distributed nonparametric function estimation and distributed sparse signal recovery.

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

Optimal estimation of the mean of a multivariate Gaussian distribution with a general (known) covariance matrix is another interesting problem. A naive approach is to ignore the dependency and apply MODGAME to estimate the coordinates individually, this is arguably not communication efficient in general. For instance, if the correlation between certain coordinates is large, it may be possible to save a significant amount of communication budget by utilizing the information from one coordinate to help estimate the other. Another approach is to use multi-MODGAME after orthogonalization. More specifically, consider the Gaussian location family with a general non-singular covariance matrix $\Sigma$. Let $\lambda_{\text{min}} > 0$ be the smallest eigenvalue of $\Sigma$. For $X \sim N_d(\theta, \Sigma)$, $\lambda_{\text{min}}^{1/2}(d\Sigma)^{-1/2}X \sim N_d\left(\lambda_{\text{min}}^{1/2}(d\Sigma)^{-1/2}\theta, \frac{\lambda_{\text{min}}}{d}I_d\right)$. Note that $\lambda_{\text{min}}^{1/2}(d\Sigma)^{-1/2}\theta \in [0, 1]^d$ for any $\theta \in [0, 1]^d$, therefore one can apply multi-MODGAME to estimate $\lambda_{\text{min}}^{1/2}(d\Sigma)^{-1/2}\theta$, then transform it back to get an estimate for $\theta$. However, this is generally not rate-optimal. A systematic study is needed for this problem. Another related and more challenging problem is optimal distributed estimation of the covariance matrix $\Sigma$.

This paper arguably considered one of the simplest settings for optimal distributed estimation under the communication constraints, but as can be seen in the paper, both the construction of the rate optimal estimators and the theoretical analysis are already quite involved for such a seemingly simple problem. As we deepen our understanding on distributed learning under the communication constraints, we hope to extent this line of work to investigate other statistical problems in distributed settings, including nonparametric function estimation, high-dimensional linear regression, and large-scale multiple testing. For Gaussian mean estimation, as we showed in the present paper, the optimal rates of convergence under the three different communication protocols – independent, sequential, and blackboard – are the same. In some more complicated problems, feedback might be useful.
in improving estimation accuracy and the optimal rates will thus be different under these three classes of communication protocols. It is interesting to understand fully when and to what extend feedback helps in terms of improving statistical accuracy.

8. Proofs. In this section we prove Theorem 2 for the univariate case. For reasons of space, Theorems 1, 3, 4, 5, 6 and the technical lemmas are proved in the Supplementary Material (Cai and Wei, 2020).

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

**Case 1:** \( \log \frac{1}{\sigma_n} + 2 \leq B < \log \frac{1}{\sigma_n} + m \). Note that \( b_i \geq 1 \) for all \( i = 1, 2, \ldots, m \) implies that \( \bar{B} = \sum_{i=1}^{m} b_i \geq m \). Therefore in this case we must have \( \sigma_n < 1 \).

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

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

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

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

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

\[
\inf_{\hat{\theta} \in A(b_{1:m})} \mathbb{E}_{\theta \sim U(G_\delta)} (\hat{\theta} - \theta)^2 \geq \frac{1}{4} \inf_{\hat{\theta} \in A(b_{1:m}) \cap G_\delta} \mathbb{E}_{\theta \sim U(G_\delta)} (\hat{\theta} - \theta)^2,
\]

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

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

\[
(\hat{\theta}_{\hat{u},\hat{v}} - \theta_{u,v})^2 \geq \max \left\{ \frac{\sigma_n^2}{4} (\hat{u} - u)^2, \delta^2 \mathbb{1}_{(\hat{v} \neq v)} \right\}.
\]

Hence

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

\[
\geq \frac{1}{4} \inf_{\theta \in U(G_\delta)} \mathbb{E}_{\theta \sim U(G_\delta)} (\hat{\theta}' - \theta)^2.
\]
Putting together (21), (22), and (23), we have
(24)
\[
\inf_{\hat{\theta} \in \mathcal{A}(b_{1:m})} \sup_{\theta \in [0,1]} (\hat{\theta} - \theta)^2 \geq \frac{1}{4} \inf_{\theta_a, \theta_v \in \mathcal{A}(b_{1:m}) \cap G_\delta} \mathbb{E}_{\theta_a, \theta_v \sim \mathcal{U}(G_\delta)} \max \left\{ \frac{\sigma_n^2}{4} (\hat{u} - u)^2, \delta^2 \| \hat{v} - v \|^2 \right\}
\geq \inf_{\theta_a, \theta_v \in \mathcal{A}(b_{1:m}) \cap G_\delta} \max \left\{ \frac{\sigma_n^2}{16} \mathbb{E}_{\theta_a, \theta_v \sim \mathcal{U}(G_\delta)} (\hat{u} - u)^2, \frac{\delta^2}{4} \mathbb{E}_{\theta_a, \theta_v \sim \mathcal{U}(G_\delta)} (\hat{\theta} - \theta)^2 \right\}.
\]

Therefore, by assigning a prior $\theta \sim \mathcal{U}(G_\delta)$, we have successfully decomposed the estimation problem of $\theta$ into estimation problems of $u$ and $v$. We can view estimation of $u$ as “localization” step and estimation of $v$ as “refinement” step, so (24) essentially has decomposed the statistical risk into localization error and refinement error. To lower bound the right hand side of (24), we show that under communication constraints, one cannot simultaneously estimate both $u$ and $v$ accurately, i.e. the localization and refinement errors cannot be both too small. Lemma 1, which shows that for any distributed estimator $\hat{\theta}$, there is unavoidable trade-off between the mutual information $I(\hat{\theta}; u)$ and $I(\hat{\theta}; v)$, is a key step.

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

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

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

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

\[
H(u|\hat{u}) = H(u) - I(\hat{u}; u) \geq 1.
\]

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

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

\[
\mathbb{E}(A - D)^2 \geq c_2.
\]
Given (25) and the fact that \( \hat{u}, u \) are integer valued, Lemma 2 yields

\[ \mathbb{E}_{\hat{u}, v \sim U(G_\delta)}(\hat{u} - u)^2 \geq c_2. \]  

**Case 1.2:** \( I(\hat{\theta}; v) \leq \frac{\delta^2}{\sigma_n^2}(B + 1 - \log(\frac{1}{\sigma_n})) \). By the strong data processing inequality, plug in \( \delta = \frac{\sigma_n}{\sqrt{256(B + 1 - \log(\frac{1}{\sigma_n}))}} \) we have \( I(\hat{v}; v) \leq I(\hat{\theta}; v) \leq \frac{1}{4} \), so

\[ H(v|\hat{v}) = H(v) - I(\hat{v}; v) \geq \frac{3}{4}. \]  

It follows from Lemma 2 that

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

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

\[ \max \left\{ \frac{\sigma_n^2}{16} \mathbb{E}_{\hat{u}, v \sim U(G_\delta)}(\hat{u} - u)^2, \frac{\delta^2}{4} \mathbb{E}_{\hat{u}, v \sim U(G_\delta)}(\hat{v} - v)^2 \right\} \geq c_2 \min \left\{ \frac{\sigma_n^2}{16}, \frac{\delta^2}{4} \right\} = \frac{c_2 \sigma_n^2}{1024(B + 1 - \log(\frac{1}{\sigma_n}))} \geq \frac{c_2}{2048} \cdot \frac{\sigma_n^2}{B - \log(\frac{1}{\sigma_n})}. \]

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

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

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

\[ \inf_{\theta \in \mathcal{A}(b_{1,m})} \sup_{\theta \in [0, 1]} (\hat{\theta} - \theta)^2 \geq \inf_{\theta \in \mathcal{A}(b_{1,m})} \mathbb{E}_{\theta \sim U(K_S)}(\hat{\theta} - \theta)^2 \geq \frac{1}{4} \inf_{\theta \in \mathcal{A}(b_{1,m}) \cap K_S} \mathbb{E}_{\theta \sim U(K_S)}(\hat{\theta} - \theta)^2 \]

\[ = \frac{1}{4S^2} \inf_{\theta \in \mathcal{A}(b_{1,m}) \cap K_S} \mathbb{E}_{\theta \sim U(K_S)}(S\hat{\theta} - S\theta)^2. \]

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

\[ I(\hat{\theta}; \theta) = I(\hat{\theta}(Z_1, Z_2, \ldots, Z_m); \theta) \leq I(Z_1, Z_2, \ldots, Z_m; \theta) \leq \sum_{i=1}^{m} H(Z_i) \leq B. \]
By $\theta \sim U(K_S)$ we have $H(\theta|\hat{\theta}) = H(\theta) - I(\hat{\theta}; \theta) \geq \log S - B \geq 1$. Note that when $\theta \sim U(K_S)$, for any $\hat{\theta} \in A(b_{1:m}) \cap K_S$, $S\hat{\theta}$ and $S\theta$ both take value in $\{0, 1, 2, ..., S-1\}$. Also we have $H(S\theta|S\hat{\theta}) = H(\theta|\hat{\theta}) \geq 1$. Therefore, Lemma 2 yields that $\mathbb{E}_{\theta \sim U(K_S)}(S\hat{\theta} - S\theta)^2 \geq c_2$. We thus conclude that

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

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

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

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

which is given in Bickel (1981). \hfill \Box

SUPPLEMENTARY MATERIAL

**Supplement A:** Supplement to “Distributed Gaussian Mean Estimation under Communication Constraints: Optimal Rates and Communication-Efficient Algorithms”

(doi: url to be specified). In this supplementary material, we prove Theorems 1, 3, 4, 5, 6 and the technical lemmas.

References.


Cai, T. T. and Wei, H. (2020). Supplement to “Distributed Gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms”.


Distributed Gaussian Mean Estimation


