Order Optimal Bounds for One-Shot Federated Learning over non-Convex Loss Functions

Arsalan Sharifnassab, Saber Salehkaleybar, S. Jamaloddin Golestani

***Abstract*—We consider the problem of federated learning in a one-shot setting in which there are m machines, each observing n sample functions from an unknown distribution on non-convex loss functions. Let $F : [-1, 1]^d \to \mathbb{R}$ be the expected loss function with respect to this unknown distribution. The goal is to find an estimate of the minimizer of F. Based on its observations, each machine generates a signal of bounded length B and sends it to a server. The server collects signals of all machines and outputs an estimate of the minimizer of F. We show that the expected loss of any algorithm is lower bounded by $\max\big(1/(\sqrt{n}(mB)^{1/d}), 1/\sqrt{mn}\big)$, up to a logarithmic factor. We then prove that this lower bound is order optimal in m and n by presenting a distributed learning algorithm, called Multi-Resolution Estimator for Non-Convex loss function (MRE-NC), whose expected loss matches the lower bound for large mn up to polylogarithmic factors.**

***Index Terms*—Federated learning, Distributed learning, Communication efficiency, non-Convex Optimization.**

I. Introduction

A. General Background

CONSIDER a set of m machines where each machine has access to n samples drawn from an unknown distribution P. Based on its observed samples, each machine sends a single message of bounded length B to a server. The server then collects messages from all machines and estimates values for model's parameters that minimize an expected loss function with respect to distribution P.

The above one-shot setting, in which there is a single message transmission between machines and the server, is one of the scenarios in a machine learning paradigm known as "Federated Learning". With the advances in smart phones, these devices can collect unprecedented amount of data from interactions between users and mobile applications. This huge amount of data can be exploited to improve the performance of learned models running in smart devices. Due to the sensitive nature of the data and privacy concerns, federated learning paradigm suggests to keep users' data in the devices and train the parameters of the models by passing messages between the devices and a central server. Since mobile phones are often off-line or their connection speeds in uplink direction might be slow, it is desirable to train models with minimum number of message transmissions.

A. Sharifnassab is with the Computing Science Department, University of Alberta, Edmonton, Canada (e-mail: sharifna@ualberta.ca).

S. Salehkaleybar is with the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, Netherlands (email:s.salehkaleybar@liacs.leidenuniv.nl).

S. J. Golestani is with the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran (e-mails:golestani@sharif.edu).

Several works have studied the problem of minimizing a convex loss function in the context of one shot distributed learning, and order optimal lower bounds and algorithms are available [1]. However, error lower bounds in the more practical case where the loss function is non-convex, have not been well-studied.

B. Our Contributions

In this paper, we first focus on a regime where mB is large and propose a lower bound on the performance of all one-shot federated learning algorithms. We show that for sufficiently large number of machines m and for any estimator $\hat{\theta}$, there exists a distribution P and the corresponding loss function F such that $\mathbb{E}\big[F(\hat{\theta}) - F(\theta^*)\big] \geq \max\big(1/(\sqrt{n}(mB)^{1/d}\ln mB), 1/\sqrt{mn}\big)$, where n is number of samples per machine, d is dimension of model's parameters, B is signal length in bits, and θ^* is the global minimizer of F. Furthermore, we show that this lower bound is order optimal in terms of n, m, and B. In particular, we propose an estimator, called Multi-Resolution Estimator for Non-Convex loss function (MRE-NC), and show that for large values of mn, the output $\hat{\theta}$ of the MRE-NC algorithm satisfies $\mathbb{E}\big[F(\hat{\theta}) - F(\theta^*)\big] \simeq \sqrt{d}\max\big(1/(\sqrt{n}(mB)^{1/d}), 1/\sqrt{mn}\big)$. We also study error-bounds under tiny communication budget, and show that if B is a constant and $n = 1$, the minimax error[1] does not go to zero even if m approaches infinity and even if $d = 1$.

We adopt an information-theoretic approach, focusing primarily on sample complexity rather than computational complexity, while assuming the availability of unlimited computational resources. Our results reveal a fundamental limitation of federated learning in the presence of a restricted communication budget. The lower bound in Theorem 1 demonstrates a "curse of dimensionality" for scenarios in which mB is sub-exponential in d. Specifically, in a centralized setting where all nm data functions are accessible on the server, a computationally exhaustive algorithm can achieve a minimax error of $1/\sqrt{mn}$, significantly lower than our federated learning minimax lower bound $(1/\sqrt{n})\max(1/\sqrt{m}, 1/(mB)^{1/d})$ for a sub-exponential communication budget B with respect to d. Interestingly, the bound primarily depends on the total number of bits mB received at the server. As a consequence, one can trade off the number of machines m and the communication budget B to maintain a fixed minimax error bound.

[1]The minimax error is defined as the smallest achievable error by an optimal estimator, considering the worst-case scenario across all possible distributions P.

Another contribution of this paper is the development of novel machinery in the proof of Theorem 1. The machinery involves an information-theoretic lower bound for an abstract coin-flipping system (see Section VI-B), which can be of broader interest for the analysis of other federated learning settings, and distributed systems in general.

C. Related Works

McMahan et al. [2] considered a decentralized setting in which each machine has access to a local training data and a global model is trained by aggregating local updates. They termed this setting, "Federated Learning" and mentioned some of its key properties such as severe communication constraints and massively distributed data with non-i.i.d distribution. To address some of these challenges, they proposed "FedAvg" algorithm, which executes in several synchronous rounds. In each round, the server randomly selects a fraction of machines and sends them the current model. Each machine performs a pre-determined number of training phases over its own data. Finally, the updated model at the server is obtained by averaging received models from the machines. The authors trained deep neural networks for tasks of image classification and next word prediction in a text and experimental results showed that the proposed approach can reduce the communication rounds by $10-100$ times compared with the stochastic gradient descent (SGD) algorithm. FedAvg is generally guaranteed to converge to a first-order stationary point [3], [4], which differs from the notion of convergence to global minimum considered in this work. Moreover, the FedAvg setting is not a one-shot scenario and involves two-way communication between the server and the machines, resulting in sub-optimal performance in one-shot setting.

After introducing the setting of federated learning by McMahan et al. [2], several research work addressed its challenges such as communication constraints, system heterogeneity (different computational and communication capabilities of the machines), statistical heterogeneity (data is generated in non-identically distributed manner), privacy concerns, and malicious activities. For instance, different approaches have been proposed in order to reduce the size of messages by performing quantization techniques [5], [6], updating the model from a restricted space [5], or utilizing compression schemes [7]–[10]. To resolve system heterogeneity issues such as stragglers, asynchronous communication schemes with the assumption of bounded delay between the server and the machines have been devised [11]–[15]. There are also several works providing convergence guarantees for the case of non-i.i.d. samples distributed among the machines [16]–[25]. Moreover, some notions of privacy can be preserved by utilizing differential privacy techniques [26]–[35] or secure multi-party computation [36]–[39].

A similar setting to federated learning has been studied extensively in the literature of distributed statistical optimization/estimation with the main focus on minimizing convex loss functions with communication constraints. In this setting, machines mainly reside in a data center, they are much more reliable than mobile devices, and straggle nodes are less problematic. If there is no limit on the number of bits that can be sent by the machines, then each machine can send its whole data to the server. In this case, we can achieve the estimation error of a centralized solution that has access to entire data. The problem becomes non-trivial if each machine can only send a limited number of bits to the server. In the one-shot setting, where there is only a single one-directional message transmission from each machine to the server, Zhang et al. [40] proposed a simple averaging method, in which each machine computes an estimate of optimal parameters that minimizes the empirical loss function over its own data and sends them to the server. The output of the server is the averege over the received values. For the convex functions with some additional assumptions, they showed that this method has expected error $O(1/\sqrt{mn}+1/n)$. It can be shown that this bound can be improved to $O(1/\sqrt{mn}+1/n^{1.5})$ via boot-strapping [40] or $O(1/\sqrt{mn}+1/n^{9/4})$ by optimizing a surrogate loss function using Taylor series expansion [41].

Recently, for the convex loss functions in the setting of one-shot federated learning, Salehkaleybar et al. [1] proposed a lower bound on the estimation error achievable by any algorithm. They also proposed an order-optimal estimator whose expected error meets the mentioned lower bound up to a polylogarithmic factor. Our bounds have three main differences with respect to [1]:

- Here we consider general non-convex loss functions as opposed to the convex loss assumption in [1];
- We bound $F(\theta)-F(\theta^*)$, whereas the bound in [1] is on $\|\theta-\theta^*\|$ and translates into a much weaker bound on $F(\theta)-F(\theta^*)$ compared to the results of this paper[2]; and
- The proof of the present bound requires a whole new machinery that is completely different from the proof techniques used in [1].

Zhou et al. [42] proposed a one-shot distillation method where each machine distills its own its data and sends the synthetic data to the server, which then trains the model over whole collected data. Moreover, they evaluated the proposed method experimentally on some real data, showing remarkable reduction in the communication costs. Later, Armacki et al. [43] considered clustered federated learning [44] with one round of communication between machines and the server. They showed that for the strongly convex case, local computations at the machines and a convex clustering based aggregation step at the server can provide an order-optimal mean-square error rate in terms of sample complexity.

For the case of multi-shot setting, a popular approach is based on stochastic gradient descent (SGD) in which the server queries the gradient of empirical loss function at a certain point in each iteration and the gradient vectors are aggregated by averaging to update the model's parameters [2], [45], [46].

[2]Note that for a constant $\delta>0$ and a twice differentiable function F, a bound of size δ on $\|\theta-\theta^*\|$ translates into a bound of size δ^2 on $F(\theta)-F(\theta^*)$, whereas in this paper we prove a much stronger (i.e., larger) lower bound of size δ on $F(\theta)-F(\theta^*)$. More concretely, letting $\delta=1/n^{1/2}(mB)^{1/d}$, for a convex function F with bounded second derivative (as assumed in [1]), a lower bound $\|\theta-\theta^*\|=\Omega(\delta)$ can only imply $F(\theta)-F(\theta^*)=\Omega(\delta^2)$. In the present work, we prove a much stronger lower bound $F(\theta)-F(\theta^*)=\Omega(\delta)$. As such, the bounds in this work and [1], despite their similarities, do not imply each other.

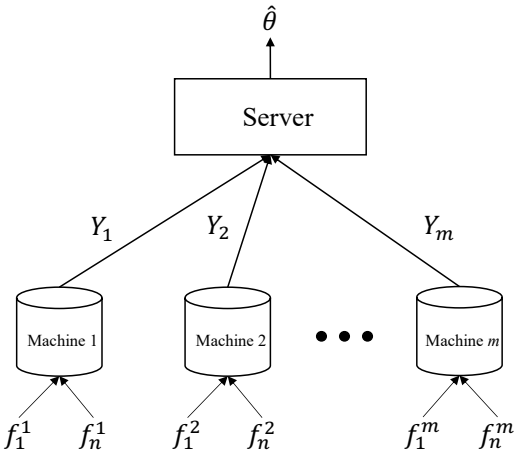


Fig. 1: The considered distributed system consists of m identical machines, each observing n independent sample functions from an unknown distribution P. Each machine i sends signal Y_i of length B bits to a server. The sever collects all the signals and returns an estimate $\hat{\theta}$ for the optimization problem in (3).

In fact, FedAvg algorithm [2] can be seen as an extension of SGD algorithm where each machine perform a number of training phases over its own data in each round. Although these solutions can be applied to non-convex loss functions, there is no theoretical guarantee on the quality of the output. Moreover, in the one-shot setting, the problem becomes more challenging since these gradient descent based methods cannot be adopted easily to this setting.

D. Outline

The paper is organized as follows. We begin with a detailed model and problem definition in Section II. In Section III, we provide a lower bound on the performance of any algorithm. In Section IV, we present the MRE-NC algorithm and an upper bound on its expected error that matches the previous lower bound up to polylogarithmic factors in m and n. We propose a constant lower bound on achievable error under tiny (constant) communication budget in Section V. The proofs of our main results are presented in Sections VI and VII, with the details relegated to appendices for improved readability. Afterwards, we report some numerical experiments on small size problems in Section VIII. Finally, in Section IX, we conclude with some remarks and open problems.

II. Problem Definition

Consider a positive integer d and let $\mathcal{F}$ be the collection of Lipschitz continuous functions over $[-1,1]^d$. More concretely, for any $f \in \mathcal{F}$ and any $\theta, \theta' \in [-1,1]^d$, we have

$$|f(\theta) - f(\theta')| \leq \|\theta - \theta'\|. \tag{1}$$

Let P be an unknown probability distribution over the functions in $\mathcal{F}$. We define the expected loss function as follows:

$$F(\theta) = \mathbb{E}_{f\sim P}\big[f(\theta)\big], \qquad \theta \in [-1,1]^d. \tag{2}$$

Our goal is to estimate a parameter θ^* that minimizes F:

$$\theta^* = \operatorname*{argmin}_{\theta\in[-1,1]^d} F(\theta). \tag{3}$$

We assume that θ^* lies in the interior of the cube $[-1,1]^d$.

The objective function is to be minimized in a distributed manner, as follows. The distributed systems consists of m machines and a server. Each machine i observes n independently and identically distributed samples $\{f_1^i, \cdots, f_n^i\}$ drawn from the probability distribution P. Based on its observed samples, machine i sends a signal Y_i of length B bits to the server.[3] The server collects the signals from all machines and returns an estimation of θ^*, which we denote by $\hat{\theta}$. Note that in this model we consider one-way one-shot communication between machines and the server, in the sense that each machine sends a single message to the server, while receiving no message from the server. We also assume that all machines are identical and are not enumerated in advance. Please refer to Fig. 1 for an illustration of the distributed system.[4]

III. The Lower Bound

In this section, we propose our main result, that is a lower bound on the estimation error of any algorithm. We consider a regime where mB is large. In particular, for any constant $\mathcal{C} \geq 1$, given B and n, we let $M_{\mathcal{C}}$ be the smallest number m that satisfies all of the following equations:

$$\mathcal{C}\sqrt{\ln(mB)} \geq 15, \tag{4}$$

$$mB \geq 10240, \tag{5}$$

$$\frac{23}{\mathcal{C}\sqrt{mB}} + \frac{1}{mB} \leq \frac{1}{7}, \tag{6}$$

$$\frac{1}{B\log_2 mB}\left[\left(\frac{313}{\mathcal{C}}\right)^2 + \frac{94^2}{\mathcal{C}\sqrt{mB}} + \frac{192}{mB} + \frac{15}{(mB)^{1.5}} + \frac{49+6B}{(mB)^2}\right] \leq \frac{1}{10}, \tag{7}$$

$$mn \geq 350000. \tag{8}$$

As an example, these conditioned are satisfied for for $\mathcal{C} = 25$, $n = 1$, $B = 64$, and $M_{\mathcal{C}} = 4 \times 10^5$. The following theorem presents our main lower bound.

Theorem 1. *For any $\mathcal{C} \geq 1$, any $m \geq M_{\mathcal{C}}$, and any estimator with output denoted by $\hat{\theta}$, there exists a distribution P and corresponding function F defined in* (2)*, for which with probability at least* $1/2$*,*

$$F\big(\hat{\theta}\big) - F\left(\theta^*\right) \geq \max\left(\frac{1}{\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB}, \frac{1}{4\sqrt{mn}}\right).$$

The proof is given in Section VI, and involves reducing the problem to the problem of identifying an unfair coin among several fair coins in a specific *coin-flipping system*. We then

[3] In this context, the letter B represents the number of bits in each message and should not be confused with communication bandwidth, which is also commonly denoted by B in the communication literature.

[4] The model of the distributed system here is similar to the one in [47].

rely on tools from information theory to derive a lower bound on the error probability of the latter problem.

As an immediate corollary of Theorem 1, we have

Corollary 1. *For $m \geq M_{\mathcal{C}}$, the expected error of any estimator with output $\hat{\theta}$ is lower bounded by*

$$\mathbb{E}\Big[F\big(\hat{\theta}\big)-F(\theta^*)\Big] \geq \max\left(\frac{1}{2\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB}, \frac{1}{8\sqrt{mn}}\right).$$

For $d \geq 10$, in the previous example where $\mathcal{C} = 25$, $B = 64$, and $m \geq 4 \times 10^5$, the lower bound in Corollary 1 would be $1/(50\sqrt{n}(mB)^{1/d}\ln mB)$.

IV. Order Optimality of the Lower Bound and the MRE-NC Algorithm

Here, we show that the lower bound in Theorem 1 is order optimal. We do this by proposing the MRE-NC estimator and showing that its error upper bound matches the lower bound up to polylogarithmc factors in mn. We should however note that despite its guaranteed order optimal worst-case error bound, benefits of applying the MRE-NC algorithm to real world problems are fairly limited. We refer the interested reader to Section IX for discussions on the shortcomings and scope of the MRE-NC algorithm. We consider general communication budget $B \geq d\log_2 mn$.

The main idea of the MRE-NC algorithm is to find an approximation of F over the domain and then let $\hat{\theta}$ be the minimizer of this approximation. In order to approximate the function efficiently, transmitted signals are constructed such that the server can obtain a multi-resolution view of function $F(\cdot)$ in a grid. Thus, we call the proposed algorithm "Multi-Resolution Estimator for Non-Convex loss (MRE-NC)". The description of MRE-NC is as follows:

Each machine i has access to n functions and sends a signal Y^i comprising $\lfloor B/(d\log_2 mn)\rfloor$ sub-signals of length $\lfloor d\log_2 mn\rfloor$. Each sub-signal has four parts of the form $(p, \Delta, \theta^p, \eta)$. The four parts p, Δ, θ^p, η are as follows:

- Part p: Let

$$\delta \triangleq \ln(mn)\max\left(\frac{\ln mn}{(mB)^{1/d}}, \frac{1}{m^{1/2}}\right). \tag{9}$$

  Let $t = \log_2(1/\delta)$. Without loss of generality, assume that t is a non-negative integer.[5] Consider a sequence of $t+1$ grids on $[-1,1]^d$ as follows. For $l = 0, \ldots, t$, we partition the cube $[-1,1]^d$ into 2^{ld} smaller equal sub-cubes with edge size 2^{-l}. The lth grid G^l contains the centers of these smaller cubes. Thus, each G^l has 2^{ld} grid points. For any point p' in G^l, we say that p' is the parent of all 2^d points in G^{l+1} that are in the 2^{-l}-cube centered at p' (see Fig. 2). Therefore, each point G^l $(l < t)$ has 2^d children.

  In each sub-signal, to choose p, we randomly select an l from $1, \ldots, t$ with probability

$$\Pr(l) = \frac{2^{(d-2)l}}{\sum_{j=1}^{t} 2^{(d-2)j}}. \tag{10}$$

[5]If $\delta > 1$, we reset the value of δ to $\delta = 1$. It is not difficult to check that the rest of the proof would not be upset in this spacial case.

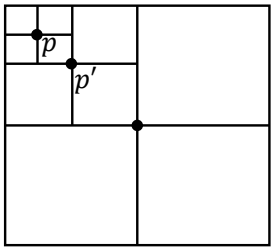


Fig. 2: An illustration of a p-point in $[-1,1]^d$ for $d = 2$. The point p belongs to G^2 and p' is the parent of p.

  We then let p be a uniformly chosen random grid point in G^l. Please note that the level l and point p selected in different sub-signals of a machine are independent and have the same distribution.
- Part Δ: We let

$$F^i(\theta) \triangleq \frac{2}{n}\sum_{j=1}^{n/2} f_j^i(\theta), \quad \text{for } \theta \in [-1,1]^d, \tag{11}$$

  and refer to it as the empirical function of the ith machine. For each sub-signal, based on its selected p part we let

$$\Delta \triangleq F^i(p) - F^i(p'), \tag{12}$$

  where $p' \in G^{l-1}$ is the parent of p.
- Part θ^p, η: In the ith machine, if the p-part of a sub-signal lies in G^t, the machine also appends two extra pieces of information θ^p, η to its sub-signal (otherwise, it sends dummy messages for these parts). We let θ^p be a minimizer of F^i in the G^t-cube containing the point p, where F^i is defined in (11). We then set $\eta = F^i(\theta^p) - F^i(p)$.

At the server, we obtain an approximation $\hat{F}$ of the loss function F over $[-1,1]^d$ as follows. We first eliminate redundant sub-signals so that no two surviving sub-signals from a same machine have the same p-parts. Hence, for each machine, the surviving sub-signals are distinct. We call this process "redundancy elimination". We set $\hat{F}(0) = 0$, and for any $l \geq 1$ and any $p \in G^l$, we let

$$\hat{F}(p) = \hat{F}(p') + \frac{1}{N_p}\sum_{\substack{\text{Subsignals of the form } (p,\Delta,\theta^p,\eta) \\ \text{after redundancy elimination}}} \Delta, \tag{13}$$

where N_p is the number of signals having point p in their first argument after redundancy elimination (with the convention that $0/0 = 0$). After that, for each cube corresponding to a point p in G^t, we choose a single arbitrary sub-signal of the form $(p, \Delta, \theta^p, \eta)$, from some machine i, and let

$$\hat{F}(\theta^p) = \hat{F}(p) + \eta = F^i(\theta^p) + \hat{F}(p) - F^i(p). \tag{14}$$

Finally, the server outputs θ^p with minimum $\hat{F}(\theta^p)$.

Algorithm 1: MRE-NC algorithm

```
// Constructing each sub-signal at machine i
```
$l \leftarrow$ choose randomly from $\{1, \cdots, t\}$ according to (10).
$p \leftarrow$ choose a point from grid G^l uniformly at random.
compute Δ in (12) for the point p.
$\theta^p \leftarrow$ a minimizer of F^i in the G^t-cube centered at p, where F^i is defined in (11).
$\eta \leftarrow F^i(\theta^p) - F^i(p)$.
prepare sub-signal (s, p, θ^p, η) for transmission.
```
// At the server
```
perform the process of "redundancy elimination".
$\hat{F}(0) \leftarrow 0$.
for $l = 1, \ldots, t$ **do**
  for $p \in G^l$ **do**
    compute $\hat{F}(p)$ according to (13).
for each $p \in G^t$, choose an arbitrary sub-signal of the form $(p, \Delta, \theta^p, \eta)$ and compute $\hat{F}(\theta^p)$ according to (14).
return a θ^p with minimum $\hat{F}(\theta^p)$.

The following theorem provides an upper bound on the estimation error of MRE-NC algorithm, for a large-mn and B regime where

$$\begin{aligned} &m \geq \ln^2 mn \\ &\ln mn \geq 8\sqrt{d} \\ &B \geq d \log(mn). \end{aligned} \tag{15}$$

Theorem 2. *Consider a $d \geq 2$ and suppose that* (15) *holds. Let $\hat{\theta}$ be the output of the MRE-NC algorithm. Then, with probability at least $1 - \exp\big(-\Omega(\ln^2 mn)\big)$,*

$$F(\hat{\theta}) - F(\theta^*) \leq 4\sqrt{d}\, \ln^2(mn) \max\left(\frac{\ln mn}{\sqrt{n}\,(mB)^{1/d}},\, \frac{1}{\sqrt{mn}} \right). \tag{16}$$

The proof is given in Section VII and goes by showing that for any $l \leq t$ and any $p \in G^l$, the number of received signals corresponding to p is large enough so that the server obtains a good approximation of F at p. Once we have a good approximation of F over G^t, we can find an approximate minimizer of F over all G^t-cubes. The following is a corollary of Theorem 2.

Corollary 2. *Let $d \geq 2$ and assume* (15)*. Then, for any $k \geq 1$,*

$$\mathbb{E}\Big[\big|F(\hat{\theta}) - F(\theta^*)\big|^k\Big] \leq \max\left(\frac{4\sqrt{d}\, \ln^3 mn}{\sqrt{n}\,(mB)^{1/d}},\, \frac{4\sqrt{d}\, \ln^2 mn}{\sqrt{mn}} \right)^k + \exp\Big(-\Omega\big(\ln^2 mn\big)\Big).$$

The above upper bound matches the lower bound of Corollary 1 up to logarithmic factors with respect to n and m, and is therefore order optimal. This implies the order optimality of the MRE-NC algorithm with respect to n and m for the large-mn regime (15).

Remark 1. Here, we carry out computations for the length of each subsignal. For the p part, we need to represent the level l and the point $p \in G^l$ in that level, which can be done by $\log_2 t + \log_2 2^{dt} < d \log_2 \sqrt{m}$. The Δ and η are scalars in $(-\sqrt{d}/2, \sqrt{d}/2)$ that we need to represent with precision $\epsilon/4t$, where ϵ is the expression in the right hand side of (16). Therefore, $\log_2(4t\sqrt{d}/\epsilon) < \log_2 \sqrt{mn}$ bits suffice to represent each of Δ and η. Finally, θ^p is a point in a 2δ-cube, with a desired entry-wise precision of $\epsilon/4\sqrt{d}$. Therefore, θ^p can be represented by $d \log_2(8\delta\sqrt{d}/\epsilon) < d \log_2 \sqrt{n}$ bits. Combining the above bounds, we obtain the following upper bound on the length of each subsignal: $d \log_2 \sqrt{m} + 2 \log_2 \sqrt{mn} + d \log_2 \sqrt{n} = (d/2 + 1) \log_2 mn \leq d \log_2 mn$.

V. Lower Bound under Tiny Communication Budget

The upper bound in Theorem 2 necessitates $B \geq d \log(mn)$. In this section, we demonstrate that to make the error bound vanish for large m, similar to Theorem 2, we need B to approach infinity as m tends to infinity. Specifically, we examine a low-communication regime where the communication budget B is constrained by a constant independent of m. For this regime and assuming $n = 1$, we prove in the following proposition that the minimax error is lower bounded by a constant, even as m approaches infinity.

Proposition 1. *Let $n = 1$ and suppose that the signal length B is bounded by a constant independent of m. Then, for any estimator $\hat{\theta}$, there is a distribution P over $\mathcal{F}$ such that $F(\hat{\theta}) - F(\theta^*) \geq \epsilon_B$, for all $m \geq 1$, where $\epsilon_B > 0$ is a constant that depends only on B and is independent of m and d. The above constant lower bound holds even when $d = 1$.*

Here, we present a short proof based on Theorem 7 of [1]. Theorem 7 of [1] establishes existence of a distribution P over strongly convex loss function with second derivatives larger than 1 and Lipschitz constant 3, for which an analogous constant lower bound $\|\hat{\theta} - \theta^*\| \geq \epsilon'_B$ holds, where ϵ'_B is a constant independent of m. Given the strong convexity of these loss functions, it follows that $F(\hat{\theta}) - F(\theta^*) \geq (\epsilon'_B)^2$. A normalization by the Lipschitz constant 3 then implies Proposition 1 for $\epsilon_B = (\epsilon'_B)^2/3$.

Proposition 1 shows that the minimax error is lower bounded by a constant regardless of m, when $n = 1$ and B is a constant. The constant ϵ_B in Proposition 1 is exponentially small in B. Note however that this is inevitable, because in view of Theorem 2, when $B = \log m$ and $n = d = 1$, the error of the MRE-NC algorithm is bounded by $\tilde{O}\big(m^{-1/d}\big)$, which is exponentially small in B. Note also that the Proposition 1 relies on the one-shot communication and may not hold in other settings for example when relaxing the assumption of symmetry between machines (i.e. the assumption that the machines run identical algorithms).

VI. Proof of Theorem 1

The desired lower bound is the maximum of two terms,

$$F(\hat{\theta}) - F(\theta^*) \geq 1/\big(\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB\big) \qquad (17)$$

and

$$F(\hat{\theta}) - F(\theta^*) \geq 1/4\sqrt{mn}. \qquad (18)$$

For the more difficult bound $1/\big(\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB\big)$, we first introduce a subclass $\hat{\mathcal{F}}$ of functions in $\mathcal{F}$ and a class of probability distributions over $\hat{\mathcal{F}}$. Under these distributions, each function is generated via a process that involves flipping mB coins, one of which is biased and the rest are fair. For this class, we show that the following property holds. Once we obtain a $1/(2\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB)$-approximate minimizer of the expected loss function F, we can identify the underlying biased coin. We then rely on this observation to reduce the abstract problem of identifying a biased coin among several fair coins via a certain coin flipping process to the problem of loss function minimization. We then use tools from information theory to derive a lower bound on the error probability in the former problem and conclude that the same lower bound applies to the latter problem as well. The second term in the lower bound, i.e. the $1/4\sqrt{mn}$ barrier, is actually well-known to hold in several centralized scenarios. Here, we present a proof based on hypothesis testing. In the rest of this section, we first establish the more difficult bound $F(\hat{\theta}) - F(\theta^*) \geq 1/\big(\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB\big)$ and introduce in Subsection VI-A the function class $\hat{\mathcal{F}}$ and the reduction to the problem of identifying the biased coin. We then describe the coin flipping system in more details in Subsection VI-B and present the lower bound on the error probability in that system. Then, in Subsection VI-C, we provide an information theoretic proof outline for this lower bound, while leaving the details until the appendices for improved readability. Finally, we establish the centralized bound $\Pr\big(F(\hat{\theta}) - F(\theta^*) \geq 1/4\sqrt{mn}\big) \geq 1/2$ in Subsection VI-D.

A. A class of distributions

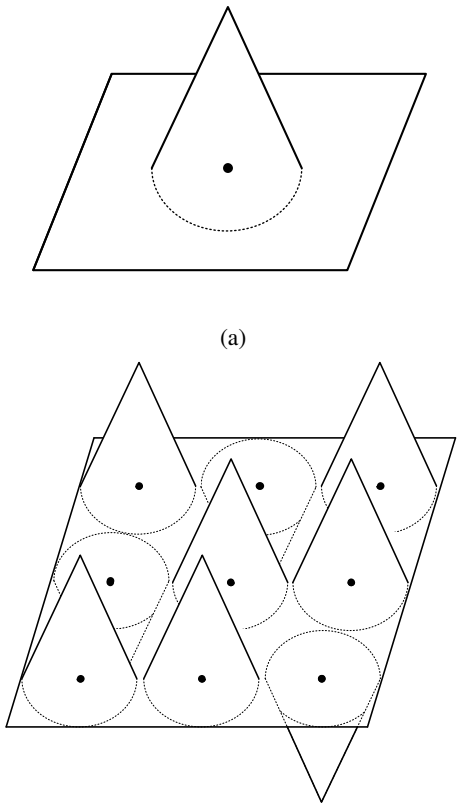
Here, we show that $F(\hat{\theta}) - F(\theta^*) \geq 1/\big(\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB\big)$ with probability at least $1/2$. For simplicity, we assume that $(mB)^{1/d}$ is an integer. Consider a function $h : \mathbb{R}^d \to \mathbb{R}$ as follows. For any $\theta \in \mathbb{R}^n$,

$$h(\theta) \;=\; \begin{cases} (mB)^{-1/d} - \|\theta\| & \text{if } \|\theta\| \leq (mB)^{-1/d}, \\ 0 & \text{otherwise.} \end{cases} \qquad (19)$$

An illustration of $h(\cdot)$ is shown in Fig.3 (a). It is easy to see that $h(\cdot)$ is Lipschitz continuous with Lipschitz constant 1. Consider a regular grid $\mathcal{G}$ with edge size $2/(mB)^{1/d}$ on the cube $[-1,1]^n$. We denote by $\{-1,1\}^{\mathcal{G}}$ the set of all functions from $\mathcal{G}$ to $\{-1,1\}$. To any $\sigma \in \{-1,1\}^{\mathcal{G}}$, we associate a function $f_\sigma : \mathbb{R}^n \to \mathbb{R}$ as follows

$$f_\sigma(\theta) \;\triangleq\; \sum_{p\in\mathcal{G}} \sigma(p)\, h(\theta - p), \qquad \forall \theta \in \mathbb{R}^n. \qquad (20)$$

Fig. 3 (b) illustrates an example of the shape of f_σ. Let $\hat{\mathcal{F}}$ be the set of all functions f_σ, for all $\sigma \in \{-1,1\}^{\mathcal{G}}$. It is easy to see that since $h(\cdot)$ is Lipschitz continuous with Lipschitz constant 1, each function $f_\sigma \in \hat{\mathcal{F}}$ is also Lipschitz continuous with Lipschitz constant 1.



(a)

(b)

Fig. 3: Illustrations of functions h and f_σ for $d = 2$. (a) shows the surface of $h(\cdot)$ defined in (19) and (b) is an example of $f_\sigma(\cdot)$ defined in (20).

For any $p \in \mathcal{G}$, we define a probability distribution P_p over $\hat{\mathcal{F}}$ as follows. For any $\sigma \in \{-1,1\}^{\mathcal{G}}$,

$$P_p(f_\sigma) \;= 2^{-mB}\left(1 - \frac{\sigma(p)}{\mathcal{C}\sqrt{n}\ln mB}\right), \qquad (21)$$

where $\mathcal{C}$ is the constant in the theorem statement. Then, $\sum_{\sigma\in\{-1,1\}^{\mathcal{G}}} P_p(f_\sigma) \;=\; 1$, and as a result, each P_p is a probability distribution. Intuitively, when a function f_σ is sampled from P_p, it is as if for every $q \in \mathcal{G}$ with $q \neq p$, we have $\Pr\big(\sigma(q) = 1\big) = \Pr\big(\sigma(q) = -1\big) = 1/2$, and for $q = p$ we have $\Pr\big(\sigma(p) = 1\big) = 1/2 - 1/\big(2\mathcal{C}\sqrt{n}\ln mB\big)$. This is like, the values of $\sigma(q)$ for $q \neq p$ are chosen independently at random according to the outcome of a fair coin flip, while the value of $\sigma(p)$ is the outcome of an unfair coin flip with bias $-1/\big(2\mathcal{C}\sqrt{n}\ln mB\big)$, i.e., for $q \in \mathcal{G}$,

$$\mathbb{E}_{f_\sigma\sim P_p}\big[\sigma(q)\big] \;=\; \begin{cases} \frac{-1}{\mathcal{C}\sqrt{n}\ln mB} & q = p, \\ 0 & q \neq p. \end{cases} \qquad (22)$$

Therefore, for any $p \in \mathcal{G}$ and any $\theta \in [-1,1]^n$, we have

$$\begin{aligned} F(\theta) &= \mathbb{E}_{f\sim P_p}\big(f(\theta)\big) \\ &= \sum_{\sigma\in\{-1,1\}^{\mathcal{G}}} P_p(f_\sigma) \sum_{q\in\mathcal{G}} \sigma(q) h\big(\theta - q\big) \\ &= \sum_{q\in\mathcal{G}} h\big(\theta-q\big) \sum_{\sigma\in\{-1,1\}^{\mathcal{G}}} P_p(f_\sigma)\sigma(q) \\ &= \sum_{q\in\mathcal{G}} h\big(\theta-q\big)\mathbb{E}_{f_\sigma\sim P_p}\big[\sigma(q)\big] \\ &= \frac{-1}{2\mathcal{C}\sqrt{n}\ln mB}\, h\big(\theta-p\big), \end{aligned} \tag{23}$$

where the last equality is due to (22). Therefore, under probability distribution P_p, $\theta^* = p$ is the global minimizer of $F(\cdot)$, and

$$F(p) \,=\, \frac{-h(0)}{\mathcal{C}\sqrt{n}\ln mB} \,=\, \frac{-1}{\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB}.$$

Moreover, for any $\theta \in \mathbb{R}^n$ with $\|\theta - p\| \geq (mB)^{1/d}$, we have

$$F(\theta) \,=\, 0 \,\geq\, F(\theta^*) + \frac{1}{\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB}, \tag{24}$$

where $\theta^* = p$.

We prove (17) by contradiction. Suppose that there exists an estimator $\mathcal{E}$, such that for any $p \in \mathcal{G}$, when the functions are sampled from distribution P_p, the estimator $\mathcal{E}$ returns an output $\hat{\theta}$, for which with probability at least $1/2$,

$$F(\hat{\theta}) \,<\, F(\theta^*) + \frac{1}{\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB}. \tag{25}$$

Then, it follows from (24) that $\|\hat{\theta} - p\| < (mB)^{1/d}$. In this case, p is the closest grid-point of $\mathcal{G}$ to $\hat{\theta}$. As a result, we can recover p from $\hat{\theta}$, with probability at least $1/2$. More concretely, given estimator $\mathcal{E}$, we can devise an estimator $\mathcal{E}'$ such that for any $p \in \mathcal{G}$ and under distribution P_p, $\mathcal{E}'$ outputs the true p with probability at least $1/2$. This provides a solution for the problem of identifying a biased coin among $mB-1$ unbiased coins, in a coin flipping system that we describe next.

B. Coin flipping

Here, we describe an abstract system that aims to identify a biased coin among several fair coins, via observing the outputs of coin flips. We then derive a bound on the error probability of any estimator, and show that no estimator can identify the biased coin with probability at least $1/2$.

Consider k coins, one of which is biased and all others are fair. The outcome of the biased coin has the following distribution:

$$P(1) \,=\, \frac{1}{2} \,+\, \frac{1}{2C\sqrt{n}\ln k}, \qquad P(0) \,=\, \frac{1}{2} \,-\, \frac{1}{2C\sqrt{n}\ln k}. \tag{26}$$

We index the coins by $t = 1, \ldots, k$. The index of the biased coin is unknown initially. Let T denote the index of the biased coin. We assume that T is a random variable, uniformly distributed over $1, \ldots, k$. We aim to estimate T by observing outcomes of coin flips as follows.

Our coin flipping system comprises m machines, called the coin flippers, and a server. Each coin flipper flips each of every coin for n times. Therefore, each coin flipper, i, makes a total number of kn coin flips and collects the outcomes into an $n \times k$ matrix W^i with 0 and 1 entries. The ith coin flipper, for $i = 1, \ldots, m$, then generates a B-bit long signal S^i based on W^i, and sends it to the server. We refer to the (possibly randomized) mapping (or coding) from W^i to S^i by Q^i. The server then collects the signals of all coin flippers and generates an estimate $\hat{T}$ of the true index of the biased coin T.

Let $P_e = \Pr\big(\hat{T} \neq T\big)$ be the probability that the server fails to identify the true biased coin index.

Proposition 2. *Let $k = mB$ and suppose that* (4)–(7) *hold. Then, $P_e > 0.5$.*

The proof is given in the next subsection. The proposition asserts that no estimator can identify the biased coin with probability at least $1/2$. This contradicts the statement in last line of the previous subsection. Hence, our initial assumption on the existence of estimator $\mathcal{E}$ that satisfies (25) cannot be the case. Equivalently, there exists no estimator $\mathcal{E}$ for which with probability at least $1/2$ we have

$$F(\hat{\theta}) \leq F(\theta^*) + 1/\big(\mathcal{C}\sqrt{n}(mB)^{1/d}\ln mB\big). \tag{27}$$

C. Proof of Proposition 2

The proof relies on the following proposition.

Proposition 3. *Suppose that $k = mB$ is large enough so that* (4)*,* (5)*, and* (6) *are satisfied. Then, for each coin flipper, i, and under any coding Q^i, we have*

$$\begin{aligned} I\big(T; S^i\big) \,<\, \frac{3B}{k\ln 2} \,+\, \frac{1}{k}\Bigg[\left(\frac{313}{\mathcal{C}}\right)^2 \,+\, \frac{94^2}{\mathcal{C}\sqrt{k}} \,+\, \frac{192}{k} \\ +\, \frac{15}{k^{1.5}} \,+\, \frac{49+6B}{k^2}\Bigg], \end{aligned} \tag{28}$$

where $I(T; S^i)$ is the mutual information between T and S^i (see [48], page 20, for the definition of mutual information).

The proof of Proposition 3 is pretty lengthy, and is given in Appendix B.

Given the index T of the biased coin, the signals $S^1, \ldots, S^m$ will be independent. As a result,

$$H\big(S^1, \ldots, S^m \mid T\big) \,=\, \sum_{i=1}^m H\big(S^i \mid T\big), \tag{29}$$

where $H(\cdot)$ is the entropy function (see [48], page 14). Consequently,

$$\begin{aligned} I\big(T;\, S^1, \ldots, S^m\big) &= H\big(S^1, \ldots, S^m\big) - H\big(S^1, \ldots, S^m \mid T\big) \\ &= H\big(S^1, \ldots, S^m\big) - \sum_{i=1}^m H\big(S^i \mid T\big) \\ &\leq \sum_{i=1}^m H\big(S^i\big) - \sum_{i=1}^m H\big(S^i \mid T\big) \\ &= \sum_{i=1}^m \Big(H\big(S^i\big) - H\big(S^i \mid T\big)\Big) \\ &= \sum_{i=1}^m I\big(T; S^i\big). \end{aligned} \tag{30}$$

Let

$$\epsilon \triangleq \left(\frac{313}{\mathcal{C}}\right)^2 + \frac{94^2}{\mathcal{C}\sqrt{k}} + \frac{192}{k} + \frac{15}{k^{1.5}} + \frac{49+6B}{k^2}$$

be the expression which is a part of the right hand side of (28). Then, it follows from (7) and $k = mB$ that

$$\frac{\epsilon}{B\log_2 k} \leq \frac{1}{10}. \tag{31}$$

We employ Fano's inequality (see [48], page 37), and write

$$\begin{aligned}
P_e &\geq \frac{H\big(T \mid S^1,\ldots,S^m\big) - 1}{\log_2 k} \\
&= \frac{H(T) \,-\, I\big(T;\, S^1,\ldots,S^m\big) - 1}{\log_2 k} \\
&= \frac{\log_2 k \,-\, I\big(T;\, S^1,\ldots,S^m\big) - 1}{\log_2 k} \\
&= 1 \,-\, \frac{I\big(T;\, S^1,\ldots,S^m\big)}{\log_2 k} \,-\, \frac{1}{\log_2 k} \\
&\geq 1 \,-\, \frac{\sum_{i=1}^m I\big(T;S^i\big)}{\log_2 k} \,-\, \frac{1}{\log_2 k} \\
&> 1 \,-\, \frac{\sum_{i=1}^m \big(3B/(k\ln 2) \,+\, \epsilon/k\big)}{\log_2 k} \,-\, \frac{1}{\log_2 k} \\
&= 1 \,-\, \frac{3mB}{k\ln k} \,-\, \frac{m\epsilon}{k\log_2 k} \,-\, \frac{1}{\log_2 k} \\
&= 1 \,-\, \frac{3}{\ln k} \,-\, \frac{\epsilon}{B\log_2 k} \,-\, \frac{1}{\log_2 k} \\
&\geq 1 \,-\, \frac{4}{10} \,-\, \frac{\epsilon}{B\log_2 k} \\
&\geq 1 \,-\, \frac{4}{10} \,-\, \frac{1}{10} \\
&= \frac{1}{2},
\end{aligned} \tag{32}$$

where the first inequality is by the Fano's inequality, the first equality follows from the definition of mutual information, the second equality is because the biased coin index T has uniform distribution over $1,\ldots,k$, the second inequality is due to (30), the third inequality follows from Proposition 3, the last equality is because of the assumption $k = mB$ in the Proposition, the fourth inequality is due to the assumption $k = mB \geq 10240$ in (5), and the last inequality is due to (31). Proposition 2 then follows from (32).

D. The centralized lower bound

We now proceed to establish

$$\Pr\big(F(\hat{\theta}) - F(\theta^*) \geq 1/4\sqrt{mn}\big) \geq 1/2.$$

Consider 9 coins, one of which is biased and all others are fair. For the biased coin, suppose that $P(1) = 1/2 + 1/4\sqrt{mn}$. The index, T, of the biased coin is initially unknown. We toss each of every coin for mn times, and estimate an index $\hat{T}$ of the biased coin based on the observed outcomes.

Lemma 1. *Assuming* (8)*, under any estimator $\hat{T}$, we have* $\Pr\big(\hat{T} \neq T\big) \geq 1/2$.

The proof is based on the error probability of the optimal hypothesis test, and is given in Appendix C-A. In the rest of the proof, similar to Subsection VI-A, we consider a collection of functions and a probability distribution over them, such that for the corresponding expected loss function F, finding a $\hat{\theta}$ with $F(\hat{\theta}) < F(\theta^*) + 1/4\sqrt{mn}$ leads to the identification of a biased coin in the setting of Lemma 1 with probability at least $1/2$. This is a contradiction, and establishes the nonexistence of such estimator. Since the argument is very similar to the line of arguments in Subsection VI-A, here we simply state the main result in the form of a lemma and defer the detailed proof until Appendix C-B.

Lemma 2. *Assuming* (8)*, for any estimator $\hat{T}$, there exists a distribution under which* $\Pr\big(F(\hat{\theta}) - F(\theta^*) \geq 1/4\sqrt{mn}\big) \geq 1/2$.

Finally, Theorem 1 follows from (27) and Lemma 2.

VII. Proof of Theorem 2

In this proof, we adopt several ideas from the proof of Theorem 4 in [1]. For simplicity and without loss of generality, throughout this proof, we assume that for any $f \in \mathcal{F}$,

$$f(0) = 0. \tag{33}$$

This is without loss of generality because adding to each function $f \in \mathcal{F}$ a constant $-f(0)$ does not change the estimation $\hat{\theta}$. We first show that for $l = 1,\ldots,t$ and for any $p \in G^l$, the number of sub-signals corresponding to p after redundancy elimination is large enough so that the server obtains a good approximation of F at p. Once we have a good approximation of F at all points of G^t, we can find an approximate minimizer of F. Let

$$\begin{aligned}
\epsilon &\triangleq \frac{4\delta\sqrt{d}\ln(mn)}{\sqrt{n}} \\
&= 4\sqrt{d}\,\ln^2(mn)\,\max\left(\frac{\ln mn}{(mB)^{1/d}\sqrt{n}},\, \frac{1}{\sqrt{mn}}\right).
\end{aligned} \tag{34}$$

For any $p \in \bigcup_{l\leq t} G^l$, let N_p be the number of machines that select point p in at least one of their sub-signals. Equivalently, N_p is the number of sub-signals after redundancy elimination that have point p as their second argument. Let $\mathcal{E}$ be the event that for $l = 1,\ldots,t$ and for any $p \in G^l$, we have

$$N_p \geq \frac{2\ln^4(mn)\,2^{-2l}}{n\epsilon^2}. \tag{35}$$

Then,

Lemma 3. $\Pr\big(\mathcal{E}\big) \geq 1 - m^{d/2}\exp\big(-\ln^2(mn)/32d\big)$.

The proof is based on the concentration inequality in Lemma 7 (b), and is given in Appendix D-A.

Capitalizing on Lemma 3, we now obtain a bound on the estimation error of F over G^l. Let $\mathcal{E}'$ be the event that for $l = 1,\ldots,t$ and any grid point $p \in G^l$, we have

$$\big|\hat{F}(p) - F(p)\big| < \frac{\epsilon}{8}. \tag{36}$$

Lemma 4. $\Pr\big(\mathcal{E}'\big) \geq 1 - m^{d/2}\exp\big(-\ln^2(mn)/32d\big) - 2m^{d/2}\exp\big(-\ln^2(mn)/128d\big)$.

The proof is given in Appendix D-B and relies on Hoeffding's inequality and the lower bound on the number of received signals for each grid point, driven in Lemma 3. For each $p \in G^t$, let $cell_p$ be the small cube with edge size 2δ that is centered at p. Let $\mathcal{E}''$ be the event that for any machine i, any $p \in G^t$, and any $\theta \in cell_p$,

$$\Big|\big(F^i(\theta) - F^i(p)\big) - \big(F(\theta) - F(p)\big)\Big| < \frac{\epsilon}{8}. \tag{37}$$

Lemma 5. $\Pr\big(\mathcal{E}''\big) \geq 1 - 2n^{d/2}m^{1+d/2}\exp\big(-\ln^2(mn)/64\big)$.

The proof is given in Appendix D-C. Assuming $\mathcal{E}''$, it follows from (14) and (37) that for any $p \in G^t$, and for the subsignal $(p, \Delta, \theta^p, \eta)$ that is used in the computation of $\hat{F}(\theta^p)$ in (14), we have

$$\Big|\big(\hat{F}(\theta^p) - \hat{F}(p)\big) - \big(F(\theta^p) - F(p)\big)\Big| \leq \frac{\epsilon}{8}. \tag{38}$$

The following auxiliary lemma has a straightforward proof.

Lemma 6. *Consider a $\gamma > 0$ and a function g over a domain $\mathcal{W}$. Let $\hat{g}$ be a uniform γ-approximation of g, that is $|\hat{g}(w) - g(w)| \leq \gamma$, for all $w \in \mathcal{W}$. Let w^* be the minimizer of $\hat{g}$ over $\mathcal{W}$. Then, $g(w^*) \leq \inf_{w\in\mathcal{W}} g(w) + 2\gamma$.*

Consider a point $p \in G^t$ and the subsignal $(p, \Delta, \theta^p, \eta)$ that is used in the computation of $\hat{F}(\theta^p)$ in (14). Suppose that this subsignal has been generated in the ith machine. Let $\hat{g}(\theta) = F^i(\theta) - F^i(p)$, $g(\theta) = F(\theta) - F(p)$, and $\mathcal{W} = cell_p$. Assuming $\mathcal{E}''$, $\hat{g}$ is an $\epsilon/8$-approximation of g, and Lemma 6 implies that $g(\theta^p) \leq g(\theta^*_{cell_p}) + \epsilon/4$, where $\theta^*_{cell_p}$ is the minimizer of F in $cell_p$. Therefore,

$$F(\theta^p) \leq F(\theta^*_{cell_p}) + \frac{\epsilon}{4}. \tag{39}$$

Moreover, assuming $\mathcal{E}'$ and $\mathcal{E}''$, we obtain

$$\begin{aligned}
|\hat{F}(\theta^p) - F(\theta^*_{cell_p})| &= \big|\big(\hat{F}(\theta^p) - \hat{F}(p)\big) - \big(F(\theta^p) - F(p)\big) \\
&\quad + \big(\hat{F}(p) - F(p)\big) + \big(F(\theta^p) - F(\theta^*_{cell_p})\big)\big| \\
&\leq \big|\big(\hat{F}(\theta^p) - \hat{F}(p)\big) - \big(F(\theta^p) - F(p)\big)\big| \\
&\quad + \big|\hat{F}(p) - F(p)\big| + \big|F(\theta^p) - F(\theta^*_{cell_p})\big| \\
&\leq \frac{\epsilon}{8} + \frac{\epsilon}{8} + \frac{\epsilon}{4} \\
&= \frac{\epsilon}{2}
\end{aligned} \tag{40}$$

where the last inequality follows from (38), (36), and (39) By further assuming $\mathcal{E}$, we know that each cell is selected by at least one machine. Then, applying Lemma 6 on (40) with identifications $\mathcal{W} = \{\theta^P : p \in G^t\}$, $\hat{g}(\theta^p) = \hat{F}(\theta^p)$ and $g(\theta^p) = F(\theta^*_{cell_p})$, we obtain

$$F(\hat{\theta}) \leq \min_{p\in G^t} F(\theta^*_{cell_p}) + \epsilon = F(\theta^*) + \epsilon. \tag{41}$$

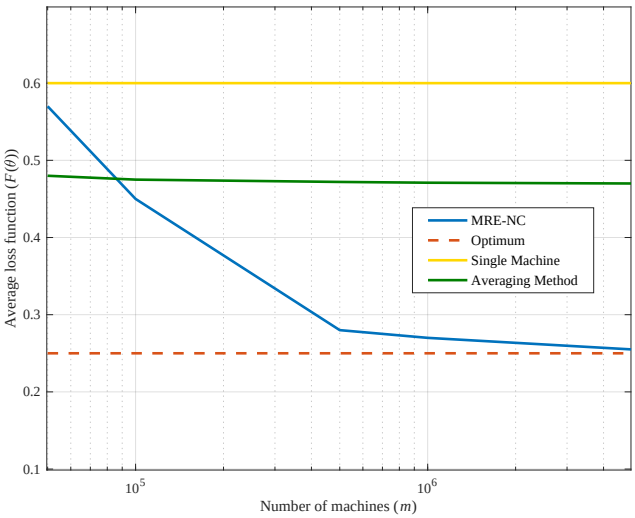


Fig. 4: Comparison of the performance of MRE-NC with two naive approaches. The number of parameters (d) and the number of samples per machine (n) are 6 and 10, respectively.

Substituting the probabilities of events $\mathcal{E}$, $\mathcal{E}'$, and $\mathcal{E}''$ from lemmas 3, 4, and 5, respectively, it follows that $F(\hat{\theta}) \leq F(\theta^*) + \epsilon$ with probability at least

$$\begin{aligned}
&1 - \big(1 - \Pr(\mathcal{E})\big) - \big(1 - \Pr(\mathcal{E}')\big) - \big(1 - \Pr(\mathcal{E}'')\big) \\
&\geq 1 - 2m^{d/2}\left[\exp\left(\frac{-\ln^2 mn}{32d}\right) + \exp\left(\frac{-\ln^2 mn}{128d}\right)\right. \\
&\qquad\qquad \left. + n^{d/2}m\exp\left(\frac{-\ln^2 mn}{64}\right)\right] \\
&\geq 1 - m(mn)^{d/2}\exp\left(\frac{-\ln^2 mn}{128d}\right).
\end{aligned}$$

This completes the proof of Theorem 2.

VIII. Experiments

Here we study performance of the MRE-NC algorithm on problems of small sizes. Note that when d is large, the lower bound $1/\sqrt{n}(mB)^{1/d}$ in Theorem 1, scales poorly with respect to mB. This eliminates the hope for efficient and guaranteed loss minimization in large problems, and limits the applicability of the MRE-NC algorithm to problems with large dimensions. In this view, in this section we focus on small size problems and demonstrate performance of MRE-NC on small toy examples.

A. Synthetic Data

We evaluated the performance of MRE-NC and compared it with two naive approaches: 1- the averaging method from [40]: each machine obtains empirical loss minimizer on its own data and sends to the server. The output would be the average of received signals at the server side. 2- Single machine method: similar to the previous method, each machine sends the empirical loss minimizer to the server. At the server, one of the received signals is picked randomly and returned as the output.

In our experiment, each sample (x, y), $x \in \mathbb{R}^2$, and $y \in \mathbb{R}$ is generated according to $y = \theta_2^T ReLU(\theta_1 x) + N$ where $ReLU(x) = \max(0, x)$ is the rectified linear unit, and the entries $[\theta_1]_{2\times 2}$ are drawn from a uniform distribution in the range $[-2, 2]$ and $[\theta_2]_{2\times 1} = [1, -1]$. Moreover, N is sampled from Gaussian distribution $\mathcal{N}(0, 0.5)$. We considered the mean square error as the loss function.

In Fig. 4, the value of $F(\theta)$ is depicted versus number of machines for MRE-NC and two naive approaches. In this experiment, we assumed that each machine has access to $n = 10$ samples. As can be seen, the MRE-NC algorithm outperforms the two naive methods, its performance improves as the number of machines increases, and approaches to the optimal value.

B. Real Data

In this part, we apply the MRE-NC algorithm to the task of classifying images of digits in the MNIST dataset [49]. We employed an ensemble learning technique [50] to build a model at the server side. In ensemble learning, we train a set of models, commonly called weak learners, that perform slightly better than random guess. Afterwards, a strong model can be built based on the models through different techniques such as boosting, stacking, or even picking the weak learner with the best performance. In this experiment, we obtained a collection of weak learners by running multiple instances of MRE-NC algorithm in parallel and then selected the one which has the lowest estimated empirical loss. More specifically, we assumed that each machine has access to $n = 10$ random samples from MNIST dataset. Furthermore, for each image $X \in \mathbb{R}^{28\times 28}$, residing in each machine, that machine splits X horizontally or vertically at pixel $p \in \{7, 14, 21\}$ into two parts, computes the average values of pixels in each part, and finally scales these average values into the range $[0, 100]$. Let $(Z_1^{h,p}, Z_2^{h,p})$ and $(Z_1^{v,p}, Z_2^{v,p})$ be the resulted values for the horizontal or vertical split at pixel p, respectively. We considered the model $h(Z) = sigmoid(\theta_1^T Z + \theta_2)$, where $sigmoid(x) = 1/(1 + \exp(-x))$, $\theta_1 \in \mathbb{R}^2$, $\theta_2 \in \mathbb{R}$, and Z is the sample obtained after pre-processing at the machine as described above for any horizontal or vertical split. We considered the cross-entropy loss function (see page 72 in [51]) and trained six models by executing MRE-NC algorithm on the data obtained from each horizontal/vertical split at pixel $p \in \{7, 14, 21\}$. At the server side, the model with the minimum $\hat{F}$ was selected. In our experiments, we considered images of only two digits 3 and 4 and tried to classify them[6]. Fig. 5, depicts the true $F(\hat{\theta})$ and the error in classification averaged over 10 instances of the problem. As can be seen, both metrics decrease as the number of machines increases. Moreover, these metrics approach the optimal values corresponding to the centralized solution in which the server has access to the entire data.

[6]We considered a binary classification problem in our experiment, and any pair of digits with different shapes can be chosen for the considered task. Herein, we picked the two digits 3 and 4 that are different in shape, and the six weak learners have a wide range of performance in terms of accuracy on these two digits.

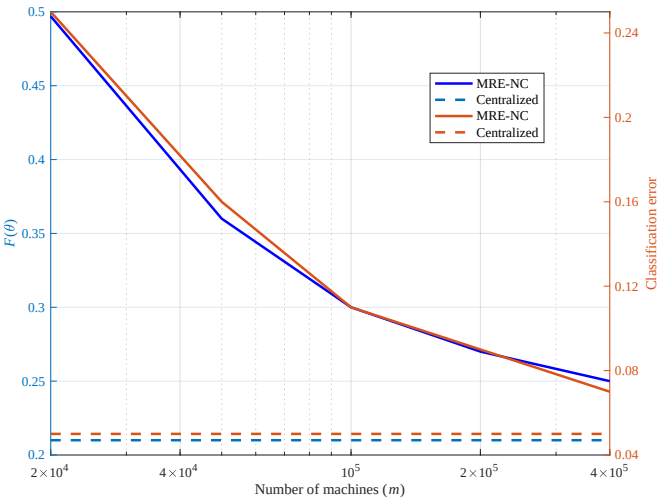


Fig. 5: The performance of MRE-NC (loss and classification error) on classifying digits in MNIST dataset against the number of machines. The left and right y-axes correspond to the true loss function and the classification error of the obtained model, respectively. The number of parameters per weak learner and the number of samples per machine (n) are 3 and 10, respectively.

IX. Discussions

In this paper, we studied the problem of federated learning in a one-shot setting and under limited communication budget B. We presented a general lower bound and showed that, ignoring logarithmic factors, the expected loss $\mathbb{E}\big[F(\hat{\theta}) - F(\theta^*)\big]$ of any estimator is lower bounded by $\max\left(n^{-1/2}(mB)^{-1/d}, (mn)^{-1/2}\right)$. We then proposed an estimator called MRE-NC, whose expected loss matches the above lower bound, and is therefore optimal. We also established a constant lower bound on minimax error when the communication budget is constrained by a constant. The class of functions we considered in this paper is pretty general. We do not assume differentiability and our class includes all Lipschitz continuous functions over $[-1, 1]^d$. This makes the model suitable for use in modern machine learning settings such as neural networks.

The MRE-NC algorithm works by finding an $O(1/\sqrt{n}(mB)^{1/d})$-approximation of the value of the expected loss function F over a fine grid of size mB. To do this, the algorithm adopts a multi-resolution idea from the MRE-C algorithm [1] which was previously proposed for the case of convex loss functions. The overall structure and the details of the MRE-NC algorithm are however different from those in [1]. While our upper bound proof incorporates several ideas from the upper bound proof in [1], the proof of our lower bound is novel and relies on reductions from the problem of identifying an unbiased coin in a certain coin flipping system. The proof involves information theory, and despite the simple appearance of the coin flipping problem, it has not been studied previously, to the best of our knowledge.

Our lower bound implies that the worst case expected error of no estimator can decrease faster than roughly $1/\sqrt{n}(mB)^{1/d}$.

When d is large, the error bound scales poorly with respect to mB. This eliminates the hope for efficient and guaranteed loss minimization in large problems, and limits the applicability of the MRE-NC algorithm to the problems with large dimensions. On the positive side, as we demonstrated in the numerical experiments, the MRE-NC algorithm can be effectively employed to solve small size problems. Moreover, for large dimensional problems, when incorporated into an ensemble learning system, it proves effective for training weak learners (refer to Section VIII for further discussions).

A drawback of the MRE-NC algorithms is that each machine requires to know m in order to set the number of levels for the grids. This can be circumvented by considering infinite number of levels, and letting the probability that p is chosen from level l decrease exponentially with l. As another drawback of the MRE-NC algorithms, note that each machine i needs to compute the minimizer θ^p of its local function F^i in a small cube around the corresponding point p. Since F^i is a non-convex function, finding θ^p is in general computationally exhaustive. Although this will not affect our theoretical bounds, it would further limit the applicability of MRE-NC algorithm in practice. Moreover, it is good to point a possible trade off between the coefficients in the precision and probability exponent of our bounds. More specifically, if we multiply the upper bound in Theorem 2 by a constant, then the corresponding probability exponent will be multiplied by the square of the same constant. In this way, one can obtain smaller upper bounds for larger values of mn.

For future works, given the poor scaling of the lower bound in terms of m and d, it would be important to devise scalable heuristics that are practically efficient in one shot learning system classes of interest, like neural networks. Moreover, efficient accurate solutions might be possible under further assumptions on the class of functions and distributions. On the theory side, the bounds in this paper are minimax bounds. From a practical perspective, it is important to develop average case bounds under reasonable assumptions. Another interesting direction is to relax the assumption of fixed n number of samples per machine, and to prove lower and upper bounds if the ith machine receives n_i samples.

REFERENCES

[1] S. Salehkaleybar, A. Sharifnassab, and S. J. Golestani, "One-shot federated learning: theoretical limits and algorithms to achieve them," *Journal of Machine Learning Research*, vol. 22, pp. 1–47, 2021.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[3] F. Zhou and G. Cong, "On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018, pp. 3219–3227.

[4] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 9709–9758, 2021.

[5] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[6] P. Jiang and G. Agrawal, "A linear speedup analysis of distributed deep learning with sparse and quantized communication," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 2530–2541.

[7] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," *arXiv preprint arXiv:1812.07210*, 2018.

[8] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[9] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," *Nature Communications*, vol. 13, no. 1, pp. 1–8, 2022.

[10] H. Tang, S. Gan, A. A. Awan, S. Rajbhandari, C. Li, X. Lian, J. Liu, C. Zhang, and Y. He, "1-bit adam: Communication efficient large-scale training with adam's convergence speed," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 118–10 129.

[11] M. Zinkevich, M. Weimer, L. Li, and A. Smola, "Parallelized stochastic gradient descent," *Advances in Neural Information Processing Systems*, vol. 23, pp. 2595–2603, 2010.

[12] Q. Ho, J. Cipar, H. Cui, S. Lee, J. K. Kim, P. B. Gibbons, G. A. Gibson, G. Ganger, and E. P. Xing, "More effective distributed ml via a stale synchronous parallel parameter server," in *Advances in Neural Information Processing Systems*, 2013, pp. 1223–1231.

[13] W. Dai, A. Kumar, J. Wei, Q. Ho, G. Gibson, and E. Xing, "High-performance distributed ml at scale through parameter server consistency models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[14] J. Nguyen, K. Malik, H. Zhan, A. Yousefpour, M. Rabbat, M. Malek, and D. Huba, "Federated learning with buffered asynchronous aggregation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 3581–3607.

[15] D. Huba, J. Nguyen, K. Malik, R. Zhu, M. Rabbat, A. Yousefpour, C.-J. Wu, H. Zhan, P. Ustinov, H. Srinivas *et al.*, "Papaya: Practical, private, and scalable federated learning," *Proceedings of Machine Learning and Systems*, vol. 4, pp. 814–832, 2022.

[16] H. Yu, S. Yang, and S. Zhu, "Parallel restarted sgd for non-convex optimization with faster convergence and less communication," *arXiv preprint arXiv:1807.06629*, vol. 2, no. 4, p. 7, 2018.

[17] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7184–7193.

[18] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[19] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[20] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 2022, pp. 965–978.

[21] W. Huang, M. Ye, and B. Du, "Learn from others and be yourself in heterogeneous federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 143–10 153.

[22] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 713–10 722.

[23] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7611–7623, 2020.

[24] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-iid data with reinforcement learning," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1698–1707.

[25] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 878–12 889.

[26] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[27] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*. IEEE, 2014, pp. 464–473.

[28] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.
[29] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable private learning with pate," *arXiv preprint arXiv:1802.08908*, 2018.
[30] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Differentially private asynchronous federated learning for mobile edge computing in urban informatics," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2134–2143, 2019.
[31] F. Mo, H. Haddadi, K. Katevas, E. Marin, D. Perino, and N. Kourtellis, "Ppfl: privacy-preserving federated learning with trusted execution environments," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 94–108.
[32] A. Girgis, D. Data, S. Diggavi, P. Kairouz, and A. T. Suresh, "Shuffled model of differential privacy in federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2521–2529.
[33] X. Gong, A. Sharma, S. Karanam, Z. Wu, T. Chen, D. Doermann, and A. Innanje, "Ensemble attention distillation for privacy-preserving federated learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 076–15 086.
[34] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and S. Y. Philip, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[35] X. Liu, H. Li, G. Xu, Z. Chen, X. Huang, and R. Lu, "Privacy-enhanced federated learning against poisoning adversaries," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4574–4588, 2021.
[36] M. S. Riazi, C. Weinert, O. Tkachenko, E. M. Songhori, T. Schneider, and F. Koushanfar, "Chameleon: A hybrid secure computation framework for machine learning applications," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018, pp. 707–721.
[37] P. Mohassel and P. Rindal, "Aby3: A mixed protocol framework for machine learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 35–52.
[38] B. D. Rouhani, M. S. Riazi, and F. Koushanfar, "Deepsecure: Scalable provably-secure deep learning," in *Proceedings of the 55th Annual Design Automation Conference*, 2018, pp. 1–6.
[39] V. Chen, V. Pastro, and M. Raykova, "Secure computation for machine learning with spdz," *arXiv preprint arXiv:1901.00329*, 2019.
[40] Y. Zhang, M. J. Wainwright, and J. C. Duchi, "Communication-efficient algorithms for statistical optimization," in *Advances in Neural Information Processing Systems*, 2012, pp. 1502–1510.
[41] M. I. Jordan, J. D. Lee, and Y. Yang, "Communication-efficient distributed statistical inference," *Journal of the American Statistical Association*, pp. 1–14, 2018.
[42] Y. Zhou, G. Pu, X. Ma, X. Li, and D. Wu, "Distilled one-shot federated learning," *arXiv preprint arXiv:2009.07999*, 2020.
[43] A. Armacki, D. Bajovic, D. Jakovetic, and S. Kar, "One-shot federated learning for model clustering and learning in heterogeneous environments," *arXiv preprint arXiv:2209.10866*, 2022.
[44] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *IEEE Transactions on Information Theory*, vol. 68, no. 12, pp. 8076–8091, 2022.
[45] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT*. Springer, 2010, pp. 177–186.
[46] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Advances in Neural Information Processing Systems*, 2015, pp. 2737–2745.
[47] S. Salehkaleybar, A. Sharifnassab, and S. J. Golestani, "One-shot distributed learning: theoretical limits and algorithms to achieve them," *arXiv preprint arXiv:1905.04634v1*, 2019.
[48] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.
[49] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.
[50] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
[51] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[52] R. Motwani and P. Raghavan, *Randomized algorithms*. Cambridge University Press, 1995.
[53] R. Ash, *Information Theory*, ser. Dover books on advanced mathematics. Dover Publications, 1990.
[54] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.
[55] R. J. Serfling, *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009, vol. 162.

APPENDIX A
CONCENTRATION INEQUALITIES

Here, we collect two well-known concentration inequalities that will be used in the proofs of our main results.

Lemma 7. *(Concentration inequalities)*
(a)(Hoeffding's inequality) Let $X_1, \cdots, X_n$ be independent random variables ranging over the interval $[a, a+\gamma]$. Let $\bar{X} = \sum_{i=1}^n X_i/n$ and $\mu = \mathbb{E}[\bar{X}]$. Then, for any $\alpha > 0$,

$$\Pr\left(|\bar{X} - \mu| > \alpha\right) \leq 2\exp\left(\frac{-2n\alpha^2}{\gamma^2}\right).$$

(b)(Theorem 4.2 in [52]) Let $X_1, \cdots, X_n$ be independent Bernoulli random variables, $X = \sum_{i=1}^n X_i$, and $\mu = \mathbb{E}[X]$. Then, for any $\alpha \in (0, 1]$,

$$\Pr\left(X < (1-\alpha)\mu\right) \leq \exp\left(-\frac{\mu\alpha^2}{2}\right).$$

APPENDIX B
PROOF OF PROPOSITION 3

The proof comprises a series of lemmas whose proofs are given in the form of separate subsections at the end of this appendix, for improved readability. For simplicity of the notation, we drop the coin-flipper's index from all equations, and will write S, W, and Q in places of S^i, W^i, and Q^i, respectively. Recall that for a coin flipper, W is the $n \times k$ binary matrix of its coin flip outcomes, so that the jth column of W corresponds to the outcomes of the jth coin, for $j = 1, \ldots, k$. We refer to the (possibly randomized) mapping (or coding) from W to the B-bit signal S by Q. More concretely, $Q\big(S \mid W\big)$ denotes the probability that a machine outputs signal S given the coin flipping outcomes W, for all $W \in \mathcal{W}$ and all $S \in \mathcal{S}$, where $\mathcal{W}$ is the set of all $n \times k$ matrices with 0 and 1 entries, and $\mathcal{S}$ is the set of all B-bit signals. We begin by showing that the mutual information $I(T;S)$ is maximized via a coding Q that is deterministic. We call a coding Q deterministic if $Q(S \mid W)$ is either 0 or 1, for all $W \in \mathcal{W}$ and $S \in \mathcal{S}$.

Lemma 8. *Among all randomized codings Q, there exists a deterministic coding that maximizes $I(T;S)$.*

The proof relies on a well-known result on the convexity of mutual information with respect to $\Pr(S \mid T)$, and is given in Appendix B-A.

In light of Lemma 8, for the rest of the proof without loss of generality we assume that the coding Q is deterministic. Equivalently, corresponding to each $s \in \mathcal{S}$, we associate a subset of $\mathcal{W}$ whose elements are mapped to s. With an abuse of notation, we denote this subset of $\mathcal{W}$ by s. In other words, to any $s \in \mathcal{S}$, is associated a subset $s \subseteq \mathcal{W}$ containing all $w \in \mathcal{W}$ that are mapped to s via the deterministic coding. For any $w \in \mathcal{W}$ and for $t = 1, \ldots, k$, we denote the tth column of w by w_t. Given $w \in \mathcal{W}$, we let $P(w)$ be the probability that w is the outcome matrix of coin-flips when T is chosen uniformly at random from $1, \ldots, k$. Moreover, given $t \leq k$, we let $P(w \mid T = t)$ be the probability that w is the outcome matrix of coin-flips when $T = t$.

Lemma 9. *There exists a subset $\bar{\mathcal{W}} \subseteq \mathcal{W}$ with $\Pr(\bar{\mathcal{W}}) \geq 1 - 6k^{-3}$, such that for any $w \in \bar{\mathcal{W}}$,*

$$\Pr\left(W_t = w_t \mid T = t\right) \leq \frac{5 \times 2^{-n}}{3}, \qquad \textit{for } t = 1, \ldots, k, \tag{42}$$

and

$$2^{-kn}\,(1-\delta) \;\leq\; P(w) \;\leq\; 2^{-kn}\,(1+\delta), \tag{43}$$

where

$$\delta \;\triangleq\; \frac{23}{\mathcal{C}\sqrt{k}} + \frac{1}{k}. \tag{44}$$

The proof relies on concentration inequalities, and is presented in Appendix B-B. For the rest of this appendix, we fix the constant δ and the set $\bar{\mathcal{W}}$ as defined in Lemma 9.

Recall the convention that for any $s \in \mathcal{S}$, we denote the subset of $\mathcal{W}$ that is mapped to s, also by s. For the simplicity of notation, for the rest of the proof, for any $s \in \mathcal{S}$, we let $\bar{s} \triangleq s \bigcap \bar{\mathcal{W}}$ and $P(\bar{s} \mid T = t) \triangleq P(w \in \bar{s} \mid T = t)$. We make the convention that $0/0 = 1$.

Lemma 10. *a) The entropy $H(S)$ of signal S satisfies*

$$H(S) \;\geq\; \left(\sum_{s \in \mathcal{S}} P(\bar{s}) \log_2 \frac{1}{P(\bar{s})}\right) \;-\; \frac{9}{k^3}. \tag{45}$$

b) The mutual information $I(T;S)$ satisfies

$$I(T;S) \;\le\; \frac{1}{k\ln 2}\sum_{s\in\mathcal{S}} P(\bar{s})\sum_{t=1}^{k}\left(\frac{P(\bar{s}\mid T=t)}{P(\bar{s})}\;-1\right)^2 \;\;+\;\frac{9+6B+15\sqrt{k}}{k^3}. \tag{46}$$

The proof is given in Appendix B-C. Our next lemma provides a bound on the weighted sum of a probability mass function in terms of its entropy.

Lemma 11. *Consider an integer $n \ge 1$ and a set $\{\alpha_u \mid u \in \{0,1\}^n\}$ of real numbers such that $\alpha_u \in [-1,1]$, for all $u \in \{0,1\}^n$, and $\sum_{u\in\{0,1\}^n}\alpha_u = 0$. Let U be a random variable on $\{0,1\}^n$ with probability distribution P, such that for any $u \in \{0,1\}^n$, we have $U = u$ with probability $P(u)$. Then,*

$$\left(\sum_{u\in\{0,1\}^n}\alpha_u P(u)\right)^2 \;\le\; 1.5\big(n - H(U)\big), \tag{47}$$

where $H(U)$ is the entropy of U.

The proof is presented in Appendix B-D. We now have all the required lemmas, and are ready to prove Proposition 3.

For any $t \le k$, any $u \in \{0,1\}^n$, and any $s \in \mathcal{S}$, let $N_t^{\bar{s}}(u)$ be the number of $w \in \bar{s}$ such that $w_t = u$. Also, let $|\bar{s}|$ be the size of the set $\bar{s}$. Then, for any $u \in \{0,1\}^n$,

$$\begin{aligned}\Pr\big(W_t = u \mid W \in \bar{s}\big) \;&=\; \frac{\sum_{\substack{w\in\bar{s}\\ w_t=u}} P(w)}{\sum_{w\in\bar{s}} P(w)}\\ &\le\; \frac{\sum_{\substack{w\in\bar{s}\\ w_t=u}} (1+\delta)2^{-kn}}{\sum_{w\in\bar{s}}(1-\delta)2^{-kn}}\\ &=\; \frac{1+\delta}{1-\delta}\,\frac{N_t^{\bar{s}}(u)}{|\bar{s}|},\end{aligned} \tag{48}$$

where the inequality follows from (43). In the same vein, for any $u \in \{0,1\}^n$,

$$\Pr\big(W_t = u \mid W \in \bar{s}\big) \;\ge\; \frac{1-\delta}{1+\delta}\,\frac{N_t^{\bar{s}}(u)}{|\bar{s}|}. \tag{49}$$

Let

$$\mathcal{U} \;\triangleq\; \left\{u \in \{0,1\}^n \,\middle|\quad \Pr\big(W_1 = u \mid T = 1\big) \le \frac{5\times 2^{-n}}{3}\right\}. \tag{50}$$

It follows from (42) that

$$\text{if}\quad w \in \bar{\mathcal{W}},\quad \text{then}\quad w_t \in \mathcal{U},\quad \text{for}\quad t = 1,\ldots,k. \tag{51}$$

Therefore, for any $u \in \{0,1\}^n$, any $t \le k$, and any $s \in \mathcal{S}$,

$$\text{if}\quad u \notin \mathcal{U},\quad \text{then}\quad N_t^{\bar{s}}(u) = 0. \tag{52}$$

Let

$$\gamma \;\triangleq\; \sum_{u\in\mathcal{U}} \Pr\big(W_1 = u \mid T = 1\big). \tag{53}$$

Then, for the random outcome matrix W of the coin flipping, we have

$$\gamma \;=\; \Pr\big(W_1 \in \mathcal{U} \mid T = 1\big) \;\ge\; \Pr\big(W \in \bar{\mathcal{W}} \mid T = 1\big) \;=\; \Pr\big(W \in \bar{\mathcal{W}}\big) \;=\; P(\bar{\mathcal{W}}) \;\ge\; 1 - 6k^{-3} \;\ge\; \frac{5}{6}, \tag{54}$$

where the first inequality follows from (51), the first equality is due to the symmetry and invariance of the set $\bar{\mathcal{W}}$ with respect to permutation of different columns, the second inequality is due to Lemma 9, and the third inequality is because $6k^{-3} \le 1/6$ (see (5) with identification $k = mB$). For any $u \in \{0,1\}^n$, let

$$\alpha_u \;\triangleq\; \begin{cases}\frac{2^n}{\gamma}\Pr\big(W_1 = u \mid T = 1\big) - 1 & \text{if } u \in \mathcal{U},\\ -1 & \text{if } u \notin \mathcal{U}.\end{cases} \tag{55}$$

It follows from (54) and the definition of $\mathcal{U}$ in (50) that for any $u \in \mathcal{U}$, we have $2^n P(W_1 = u \mid T = 1)/\gamma \le 2^n P(W_1 = u \mid T = 1) \times 6/5 \le 2$. Therefore, $\alpha_u \in [-1,1]$, for all $u \in \{0,1\}^n$. Moreover,

$$\sum_{u\in\{0,1\}^n}\alpha_u \;=\; -2^n + \frac{2^n}{\gamma}\sum_{u\in\mathcal{U}}\Pr\big(W_1 = u \mid T = 1\big) \;=\; -2^n + \frac{2^n}{\gamma}\times\gamma \;=\; 0,$$

where the second equality is from the definition of γ in (53). Hence, the set of numbers α_u, for $u \in \{0,1\}^n$, satisfies all of the conditions in Lemma 11. Therefore, it follows from Lemma 11 that for any $s \in \mathcal{S}$ and any $t \leq k$,

$$\left(\sum_{u \in \{0,1\}^n} \alpha_u \Pr\big(W_t = u \mid W \in \bar{s}\big) \right)^2 \leq 1.5\Big(n - H\big(W_t \mid W \in \bar{s}\big)\Big). \tag{56}$$

In what follows, we try to derive a bound on $P\big(\bar{s} \mid t\big)/P(\bar{s})$ in terms of α_u. We then use (56) and Lemma 10 to obtain the desired bound on $I(T;S)$. We now elaborate on $P\big(\bar{s} \mid t\big)$,

$$\begin{aligned} P\big(\bar{s} \mid T = t\big) &= 2^{-n(k-1)} \sum_{w \in \bar{s}} P\big(w_t \mid T = t\big) \\ &= 2^{-n(k-1)} \sum_{u \in \{0,1\}^n} \Pr\big(W_t = u \mid T = t\big) N_t^{\bar{s}}(u) \\ &= 2^{-nk} \sum_{u \in \{0,1\}^n} \Big(2^n \Pr\big(W_t = u \mid T = t\big)\Big) N_t^{\bar{s}}(u) \\ &= 2^{-nk} \sum_{u \in \{0,1\}^n} \Big((\alpha_u + 1)\gamma\Big) N_t^{\bar{s}}(u), \end{aligned} \tag{57}$$

where the last equality is due to (52) and the definition of α_u in (55). On the other hand, since $P(\bar{s}) = \sum_{w \in \bar{s}} P(w)$, it follows from (43) that

$$\begin{aligned} P(\bar{s}) &\leq \sum_{w \in \bar{s}} (1+\delta)\, 2^{-kn} = 2^{-kn}\,(1+\delta)\,|\bar{s}|, \\ P(\bar{s}) &\geq \sum_{w \in \bar{s}} (1-\delta)\, 2^{-kn} = 2^{-kn}\,(1-\delta)\,|\bar{s}|. \end{aligned} \tag{58}$$

Combining (57) and (58), we obtain

$$\begin{aligned} \frac{P\big(\bar{s} \mid T = t\big)}{P(\bar{s})} &\leq \frac{1}{1-\delta} \sum_{u \in \{0,1\}^n} (1+\alpha_u) \frac{N_t^{\bar{s}}(u)}{|s|}, \\ \frac{P\big(\bar{s} \mid T = t\big)}{P(\bar{s})} &\geq \frac{1}{1+\delta} \sum_{u \in \{0,1\}^n} (1+\alpha_u) \frac{N_t^{\bar{s}}(u)}{|s|}. \end{aligned} \tag{59}$$

It then follows from (59) and (48) that

$$\begin{aligned} \frac{P\big(\bar{s} \mid T = t\big)}{P(\bar{s})} &\geq \frac{1}{1+\delta} \sum_{u \in \{0,1\}^n} (1+\alpha_u) \frac{N_t^{\bar{s}}(u)}{|s|} \\ &\geq \frac{1-\delta}{(1+\delta)^2} \sum_{u \in \{0,1\}^n} \big(1+\alpha_u\big) \Pr\big(W_t = u \mid W \in \bar{s}\big) \\ &\geq \big(1-4\delta\big) \sum_{u \in \{0,1\}^n} \big(1+\alpha_u\big) \Pr\big(W_t = u \mid W \in \bar{s}\big) \\ &\geq \sum_{u \in \{0,1\}^n} \big(1+\alpha_u\big) \Pr\big(W_t = u \mid W \in \bar{s}\big) - 4\delta \sum_{u \in \{0,1\}^n} 2 \Pr\big(W_t = u \mid W \in \bar{s}\big) \\ &= \sum_{u \in \{0,1\}^n} \big(1+\alpha_u\big) \Pr\big(W_t = u \mid W \in \bar{s}\big) - 8\delta \\ &= \sum_{u \in \{0,1\}^n} \alpha_u \Pr\big(W_t = u \mid W \in \bar{s}\big) + 1 - 8\delta, \end{aligned} \tag{60}$$

where the first inequality is from (59), the second inequality follows from (48), the fourth inequality is because $\alpha_u \leq 2$ for all $u \in \{0,1\}^n$, and the third inequality is due to the assumption that $\delta \leq 1/7$ (see (6)) and the following inequality (which is easy to verify with a computer program)

$$\frac{1-x}{(1+x)^2} \geq 1 - 4x \quad \text{and} \quad \frac{1+x}{(1-x)^2} \leq 1 + 4x, \qquad \forall x \in [0, 1/7].$$

Following a similar line of arguments and using (49) instead of (48), we obtain

$$\frac{P\big(\bar{s} \mid T = t\big)}{P(\bar{s})} \leq \sum_{u \in \{0,1\}^n} \alpha_u \Pr\big(W_t = u \mid W \in \bar{s}\big) + 1 + 8\delta. \tag{61}$$

Combining (60) and (61), we obtain

$$\begin{aligned}\left(\frac{P\big(\bar{s} \mid T=t\big)}{P(\bar{s})}-1\right)^2 &\leq \left(\Big|\sum_{u\in\{0,1\}^n}\alpha_u \Pr\big(W_t=u \mid W\in\bar{s}\big)\Big| + 8\delta\right)^2 \\ &\leq 2\left(\sum_{u\in\{0,1\}^n}\alpha_u \Pr\big(W_t=u \mid W\in\bar{s}\big)\right)^2 + 2(8\delta)^2 \\ &\leq 3\Big(n - H\big(W_t \mid W\in\bar{s}\big)\Big) + 128\delta^2,\end{aligned} \tag{62}$$

where the first inequality is due to (60) and (61), the second inequality is because $(a+b)^2 \leq 2a^2+2b^2$, for all $a,b\in\mathbb{R}$, and the last inequality follows from (56).

On the other hand,

$$\begin{aligned}H\big(W \mid W\in\bar{s}\big) &= \sum_{w\in\bar{s}} P\big(w \mid w\in\bar{s}\big)\,\log_2\frac{1}{P\big(w \mid w\in\bar{s}\big)} \\ &= \sum_{w\in\bar{s}}\frac{P(w)}{P(\bar{s})}\,\log_2\frac{P(\bar{s})}{P(w)} \\ &\geq \sum_{w\in\bar{s}}\frac{P(w)}{P(\bar{s})}\,\log_2\frac{P(\bar{s})}{(1+\delta)\,2^{-kn}} \\ &= \log_2\frac{P(\bar{s})}{(1+\delta)\,2^{-kn}} \\ &= kn + \log_2 P(\bar{s}) - \log_2(1+\delta) \\ &\geq kn - 1.5\delta + \log_2 P(\bar{s}),\end{aligned} \tag{63}$$

where the first inequality is due to Lemma 9, and the last inequality is because $\log_2(1+x)\leq 1.5x$, for all $x>-1$. Moreover,

$$H\big(W \mid W\in\bar{s}\big) = H\big(W_1,\ldots,W_k \mid W\in\bar{s}\big) \leq \sum_{t=1}^{k} H\big(W_t \mid W\in\bar{s}\big), \tag{64}$$

where the inequality is from the sub-additive property of the entropy (see [48], page 41). Plugging (63) into (64), we obtain

$$\sum_{t=1}^{k} H\big(W_t \mid W\in\bar{s}\big) \geq H\big(W \mid W\in\bar{s}\big) \geq \log_2 P(\bar{s}) + kn - 1.5\delta. \tag{65}$$

Combining everything together, we finally have

$$\begin{aligned}I(T;S) &\leq \frac{1}{k\ln 2}\sum_{s\in\mathcal{S}}P(\bar{s})\sum_{t=1}^{k}\left(\frac{P(\bar{s}\mid T=t)}{P(\bar{s})}-1\right)^2 + \frac{9+6B+15\sqrt{k}}{k^3} \\ &\leq \frac{1}{k\ln 2}\sum_{s\in\mathcal{S}}P(\bar{s})\sum_{t=1}^{k}\Big(3n - 3H\big(W_t \mid W_t\in\bar{s}\big) + 128\delta^2\Big) + \frac{9+6B+15\sqrt{k}}{k^3} \\ &= \frac{3n}{k\ln 2} - \frac{3}{k\ln 2}\sum_{s\in\mathcal{S}}P(\bar{s})\sum_{t=1}^{k}H\big(W_t \mid W_t\in\bar{s}\big) + \frac{128\delta^2}{\ln 2} + \frac{9+6B+15\sqrt{k}}{k^3} \\ &\leq \frac{3n}{k\ln 2} - \frac{3}{k\ln 2}\sum_{s\in\mathcal{S}}P(\bar{s})\Big(\log_2 P(\bar{s}) + nk - 1.5\delta\Big) + 185\delta^2 + \frac{9+6B+15\sqrt{k}}{k^3} \\ &= \frac{3}{k\ln 2}\sum_{s\in\mathcal{S}}P(\bar{s})\log_2\frac{1}{P(\bar{s})} + \frac{4.5\delta}{k\ln 2} + 185\delta^2 + \frac{9+6B+15\sqrt{k}}{k^3} \\ &\leq \frac{3}{k\ln 2}H(S) + \frac{27}{k^3\ln 2} + \frac{6.5\delta}{k} + 185\delta^2 + \frac{9+6B+15\sqrt{k}}{k^3} \\ &\leq \frac{3B}{k\ln 2} + \frac{40}{k^3} + \frac{6.5\delta}{k} + 185\delta^2 + \frac{9+6B+15\sqrt{k}}{k^3} \\ &< \frac{3B}{k\ln 2} + \frac{1}{k}\left[\left(\frac{313}{\mathcal{C}}\right)^2 + \frac{94^2}{\mathcal{C}\sqrt{k}} + \frac{192}{k} + \frac{15}{k^{1.5}} + \frac{49+6B}{k^2}\right],\end{aligned}$$

where the first inequality follows from Lemma 10 (b), the second inequality is due to (62), the third inequality is from (65), the fourth inequality results from Lemma 10 (a), the fifth inequality is because S is a B-bit signal and as a result, $H(S) \leq B$, and the last inequality is by substituting δ from (44) and simple calculations. This implies (28) and completes the proof of Proposition 3.

A. Proof of Lemma 8

The proof relies on a known property of mutual information (see Theorem 2.7.4 of [48] on page 33), according to which

$$I(S;T) \text{ is a convex function with respect to } P\big(S \mid T\big). \tag{66}$$

Let Q be a randomized coding, under which a machine outputs signal S given the coin-flipping outcome vector W with probability $Q\big(S \mid W\big)$. For any $s \in \mathcal{S}$ and $t = 1, \ldots, k$, let

$$P_Q\big(s \mid t\big) \triangleq \sum_{w \in \mathcal{W}} P\big(w \mid t\big)\, Q\big(s \mid w\big) \tag{67}$$

be the probability of signal s given the biased coin index t. Let $\mathcal{Q}$ be the set of all deterministic mappings (or functions) from $\mathcal{W}$ to $\mathcal{S}$. Corresponding to any $g \in \mathcal{Q}$, we consider a deterministic coding Q_g as follows

$$Q_g\big(s \mid w\big) = \begin{cases} 1 & \text{if } g(w) = s, \\ 0 & \text{otherwise.} \end{cases} \tag{68}$$

We also let

$$P_g\big(s \mid t\big) \triangleq \sum_{w \in \mathcal{W}} P\big(w \mid t\big)\, Q_g\big(s \mid w\big) \tag{69}$$

be the probability of signal s given the biased coin index t, under the coding Q_g. We will show that for any stochastic coding Q, $P_Q(\cdot)$ is a convex combination of $P_g(\cdot)$, for $g \in \mathcal{Q}$, in the sense that there exist non-negative coefficients α_g, for $g \in \mathcal{Q}$, such that $\sum_{g \in \mathcal{Q}} \alpha_g = 1$ and

$$P_Q\big(s \mid t\big) = \sum_{g \in \mathcal{Q}} \alpha_g\, P_g\big(s \mid t\big), \qquad \forall s \in \mathcal{S}, \quad t = 1, \ldots, k. \tag{70}$$

Once we establish (70), it follows from (66) that[7]

$$I\big(S;T\big) \leq \sum_{g \in \mathcal{Q}} \alpha_g I\big(S_g;T\big) \leq \max_{g \in \mathcal{Q}} I\big(S_g;T\big), \tag{71}$$

where S is a random signal generated via coding Q, and for $g \in \mathcal{Q}$, S_g is a random signal generated under coding Q_g. As a result, there exists a $g \in \mathcal{Q}$ such that the mutual information under deterministic coding Q_g is no smaller than the mutual information under the randomized coding Q. This shows that the mutual information is maximized under a deterministic coding, which in turn implies the lemma. In the rest of the proof, we will establish (70).

Lets fix a randomized coding Q. We enumerate the set $\mathcal{W}$ and let $\mathcal{W} = \big\{w^1, \ldots, w^{2^{kn}}\big\}$. For any $g \in \mathcal{Q}$ let

$$\alpha_g \triangleq \prod_{w \in \mathcal{W}} Q\big(g(w) \mid w\big) = \prod_{i=1}^{2^{kn}} Q\big(g(w^i) \mid w^i\big). \tag{72}$$

Then,

$$\begin{aligned} \sum_{g \in \mathcal{Q}} \alpha_g &= \sum_{g \in \mathcal{Q}} \prod_{i=1}^{2^{kn}} Q\big(g(w^i) \mid w^i\big) \\ &= \sum_{s_1 \in \mathcal{S}} \cdots \sum_{s_{2^{kn}} \in \mathcal{S}} \prod_{i=1}^{2^{kn}} Q\big(s_i \mid w^i\big) \\ &= \left(\sum_{s_1 \in \mathcal{S}} Q\big(s_1 \mid w^1\big)\right) \times \cdots \times \left(\sum_{s_{2^{kn}} \in \mathcal{S}} Q\big(s_{2^{kn}} \mid w^{2^{kn}}\big)\right) \\ &= 1 \times \cdots \times 1 \\ &= 1, \end{aligned} \tag{73}$$

[7]Please note that $I(S;T)$ can be seen as a convex function of a vector $\alpha = (\alpha_1, \cdots, \alpha_{|\mathcal{Q}|})$ where $\sum_{g \in \mathcal{Q}} \alpha_g = 1$. Moreover, the value of this function at the standard basis vector e_i, $1 \leq i \leq |\mathcal{Q}|$, would be $I\big(S_g;T\big)$. Thus, the value of $I(S;T)$ is less than the linear combination of values of this function at basis vectors with weights given in α.

where the second equality is because $\mathcal{Q}$ is the set of all deterministic functions from $\mathcal{W}$ to $\mathcal{S}$ and for any $s_1, \ldots, s_{2^{kn}} \in \mathcal{S}$, there exists a $g \in \mathcal{Q}$ such that $g(w^i) = s_i$ for $i = 1, \ldots, 2^{kn}$; and the last inequality is because for any $w \in \mathcal{W}$, $Q\big(\cdot \mid w\big)$ is a probability mass function over $\mathcal{S}$.

On the other hand, for any $s \in \mathcal{S}$,

$$\begin{aligned}
\sum_{g\in\mathcal{Q}} \alpha_g Q_g\big(s \mid w^1\big) \;&=\; \sum_{\substack{g\in\mathcal{Q}\\ g(w^1)=s}} \alpha_g \\
&= \sum_{\substack{g\in\mathcal{Q}\\ g(w^1)=s}} \prod_{i=1}^{2^{kn}} Q\big(g(w^i) \mid w^i\big) \\
&= Q\big(s \mid w^1\big) \sum_{\substack{g\in\mathcal{Q}\\ g(w^1)=s}} \prod_{i=2}^{2^{kn}} Q\big(g(w^i) \mid w^i\big) \\
&= Q\big(s \mid w^1\big) \sum_{s_2\in\mathcal{S}} \cdots \sum_{s_{2^{kn}}\in\mathcal{S}} \prod_{i=2}^{2^{kn}} Q\big(g(w^i) \mid w^i\big) \\
&= Q\big(s \mid w^1\big) \left(\sum_{s_2\in\mathcal{S}} Q\big(s_2 \mid w^2\big)\right) \times \cdots \times \left(\sum_{s_{2^{kn}}\in\mathcal{S}} Q\big(s_{2^{kn}} \mid w^{2^{kn}}\big)\right) \\
&= Q\big(s \mid w^1\big) \times 1 \times \cdots \times 1 \\
&= Q\big(s \mid w^1\big),
\end{aligned} \tag{74}$$

where the first equality follows from the definition of Q_g in (68), the fourth equality is because for any $s_1, \ldots, s_{2^{kn}} \in \mathcal{S}$, there exists a $g \in \mathcal{Q}$ such that $g(w^i) = s_i$ for $i = 1, \ldots, 2^{kn}$, and the sixth equality is because for any $w \in \mathcal{W}$, $Q\big(\cdot \mid w^1\big)$ is a probability mass function over $\mathcal{S}$. In the same vein, for any $w \in \mathcal{W}$ and any $s \in \mathcal{S}$, we have

$$Q\big(s \mid w\big) \;=\; \sum_{g\in\mathcal{Q}} \alpha_g Q_g\big(s \mid w\big). \tag{75}$$

Therefore, for $t = 1, \ldots, k$ and for any $s \in \mathcal{S}$,

$$\begin{aligned}
P_Q\big(s \mid t\big) \;&=\; \sum_{w\in\mathcal{W}} P\big(w \mid t\big)\, Q\big(s \mid w\big) \\
&= \sum_{w\in\mathcal{W}} P\big(w \mid t\big) \sum_{g\in\mathcal{Q}} \alpha_g\, Q_g\big(s \mid w\big) \\
&= \sum_{g\in\mathcal{Q}} \alpha_g \sum_{w\in\mathcal{W}} P\big(w \mid t\big)\, Q_g\big(s \mid w\big) \\
&= \sum_{g\in\mathcal{Q}} \alpha_g P_g\big(s \mid t\big),
\end{aligned} \tag{76}$$

where the first equality is from the definition of $P_Q(\cdot)$ in (67), the second equality follows from (75), and the last equality is due to the definition of $P_g(\cdot)$ in (69). This implies (70). Lemma 8 then follows from the argument following (71).

B. Proof of Lemma 9

Fix a $t_0 \leq k$ and let ψ be a random outcome of the coin flipping matrix generated via distribution $P(W \mid T = t_0)$. For $t = 1, \ldots, k$ let δ_t denote the number of 1s in the tth column of ψ. Therefore,

$$\mathbb{E}\big[\delta_{t_0}\big] \;=\; \frac{n}{2} \;+\; \frac{\sqrt{n}}{2\mathcal{C}\,\ln k}, \tag{77}$$

and for any $t \neq t_0$,

$$\mathbb{E}\big[\delta_t\big] \;=\; \frac{n}{2}. \tag{78}$$

We now capitalizing on the Hoeffding's inequality (see Lemma 7 (a)) to obtain

$$\begin{aligned}\Pr\Big(\big|\delta_{t_0}-\frac{n}{2}\big|\geq 2.5\sqrt{n\ln k}\Big) &\leq \Pr\left(\big|\delta_{t_0}-\Big(\frac{n}{2}+\frac{\sqrt{n}}{2\mathcal{C}\ln k}\Big)\big|\geq 2.5\sqrt{n\ln k}-\frac{\sqrt{n}}{2\mathcal{C}\ln k}\right)\\ &= \Pr\left(\big|\delta_{t_0}-\mathbb{E}[\delta_{t_0}]\big|\geq 2.5\sqrt{n\ln k}-\frac{\sqrt{n}}{2\mathcal{C}\ln k}\right)\\ &\leq \Pr\Big(\big|\delta_{t_0}-\mathbb{E}[\delta_{t_0}]\big|\geq 2\sqrt{n\ln k}\Big)\\ &\leq 2\exp\left(\frac{-8n\ln k}{n}\right)\\ &\leq 2\exp\big(-8\ln k\big)\\ &\leq \frac{2}{k^4},\end{aligned} \tag{79}$$

where the first equality is from (77) and the third inequality is due to the Hoeffding's inequality. In the same vein, for any $t\neq t_0$,

$$\begin{aligned}\Pr\Big(\big|\delta_t-\frac{n}{2}\big|\geq 2.5\sqrt{n\ln k}\Big) &= \Pr\Big(\big|\delta_t-\mathbb{E}[\delta_t]\big|\geq 2.5\sqrt{n\ln k}\Big)\\ &\leq 2\exp\left(\frac{-12.5n\ln k}{n}\right)\\ &\leq \frac{2}{k^4},\end{aligned} \tag{80}$$

where the equality is due to (78) Therefore, for $t=1,\ldots,k$,

$$\Pr\Big(\big|\delta_t-\frac{n}{2}\big|\geq 2.5\sqrt{n\ln k}\Big)\leq \frac{2}{k^4}. \tag{81}$$

It is easy to verify via a simple computer program that $e^x\leq 1+4x/3$, for all $x\in[0,0.5]$. It then follows from (4) with $k=mB$ that

$$\exp\left(\frac{15}{2\mathcal{C}\sqrt{\ln k}}\right)\leq 1+\frac{10}{\mathcal{C}\sqrt{\ln k}}. \tag{82}$$

Let

$$\epsilon\triangleq\frac{1}{\mathcal{C}\sqrt{n}\ln k}. \tag{83}$$

In the same vein, we have $(1+x)/(1-x)\leq e^{3x}$, for all $x\in[0,1/3]$. Therefore, in view of (4), $\epsilon\leq 1/3$, and hence,

$$\frac{1+\epsilon}{1-\epsilon}\leq e^{3\epsilon}. \tag{84}$$

Moreover, for any $x\in[0,0.5]$, we have $1-x\geq e^{-2x}$. Consequently,

$$\big(1-\epsilon^2\big)^{n/2}\geq\big(\exp(-2\epsilon^2)\big)^{n/2}=\exp\big(-n\epsilon^2\big)=\exp\left(\frac{-1}{\mathcal{C}^2\ln^2 k}\right). \tag{85}$$

Once again, we emphasize that ψ is sampled from a distribution in which the t_0th coin is biased. Then, for $t=1,\ldots,k$,

$$\begin{aligned}\Pr\big(W_t=\psi_t\mid T=t\big) &= \left(\frac{1}{2}+\frac{\epsilon}{2}\right)^{\delta_t}\left(\frac{1}{2}-\frac{\epsilon}{2}\right)^{n-\delta_t}\\ &= 2^{-n}\,(1+\epsilon)^{\frac{n}{2}+(\delta_t-\frac{n}{2})}\,(1-\epsilon)^{\frac{n}{2}-(\delta_t-\frac{n}{2})}\\ &= 2^{-n}\left(1-\epsilon^2\right)^{\frac{n}{2}}\left(\frac{1+\epsilon}{1-\epsilon}\right)^{\delta_t-\frac{n}{2}},\end{aligned} \tag{86}$$

where the first equality is due to (26) and the definition of ϵ in (83). Assuming $|\delta_t-n/2|\leq 2.5\sqrt{n\ln k}$, (86) simplifies to

$$\begin{aligned}P\big(W_t=\psi_t\mid T=t\big) &\leq 2^{-n}\left(\frac{1+\epsilon}{1-\epsilon}\right)^{\delta_t-\frac{n}{2}}\\ &\leq 2^{-n}\left(e^{3\epsilon}\right)^{|\delta_t-\frac{n}{2}|}\\ &\leq 2^{-n}\exp\left(7.5\epsilon\sqrt{n\ln k}\right)\\ &= 2^{-n}\exp\left(\frac{15}{2\mathcal{C}\sqrt{\ln k}}\right)\\ &\leq 2^{-n}\left(1+\frac{10}{\mathcal{C}\sqrt{\ln k}}\right),\end{aligned} \tag{87}$$

where the second inequality follows from (84), the third inequality is due to the assumption $|\delta_t - n/2| \leq 2.5\sqrt{n \ln k}$, the equality is by the definition of ϵ in (83), and the last inequality is from (82).

In the same vein, assuming $|\delta_t - n/2| \leq 2.5\sqrt{n \ln k}$, (86) can be simplified as

$$\begin{aligned} P\big(W_t = \psi_t \mid T = t\big) \;&=\; 2^{-n} \left(1 - \epsilon^2\right)^{\frac{n}{2}} \left(\frac{1+\epsilon}{1-\epsilon}\right)^{\delta_t - \frac{n}{2}} \\ &\geq\; 2^{-n} \exp\left(\frac{-1}{\mathcal{C}^2 \ln^2 k}\right) \left(\frac{1+\epsilon}{1-\epsilon}\right)^{-|\delta_t - \frac{n}{2}|} \\ &\geq\; 2^{-n} \exp\left(\frac{-1}{\mathcal{C}^2 \ln^2 k}\right) \exp\left(3\epsilon\right)^{-|\delta_t - \frac{n}{2}|} \\ &\geq\; 2^{-n} \exp\left(\frac{-1}{\mathcal{C}^2 \ln^2 k} \;-\; 7.5\epsilon\sqrt{n \ln k}\right) \\ &=\; 2^{-n} \exp\left(\frac{-1}{\mathcal{C}^2 \ln^2 k} \;-\; \frac{15}{2\mathcal{C}\sqrt{\ln k}}\right) \\ &\geq\; 2^{-n} \exp\left(\frac{-8}{\mathcal{C}\sqrt{\ln k}}\right) \\ &\geq\; 2^{-n} \left(1 - \frac{8}{\mathcal{C}\sqrt{\ln k}}\right), \end{aligned} \tag{88}$$

where the first inequality follows from (85), the second inequality is due to (84), the third inequality is from the assumption that $|\delta_t - n/2| \leq 2.5\sqrt{n \ln k}$, the second equality is from the definition of ϵ in (83), the fourth inequality is because $\mathcal{C} \ln k \geq 2$ (see (4) with identification $k = mB$), and the last inequality is because $e^{-x} \geq 1 - x$, for all $x \in \mathbb{R}$.

Combining (81), (87), and (88), it follows that for $t = 1, \ldots, k$, with probability at least $1 - 2k^{-4}$ we have

$$\Pr\big(W_t = \psi_t \mid T = t\big) \;\in\; 2^{-n} \;\times\; \left(1 - \frac{8}{\mathcal{C}\sqrt{\ln k}},\, 1 + \frac{10}{\mathcal{C}\sqrt{\ln k}}\right). \tag{89}$$

Let $\bar{\mathcal{W}}_1$ be a subset of $\mathcal{W}$ that contains all $w \in \mathcal{W}$ for which $|\delta_t - n/2| \leq 2.5\sqrt{n \ln k}$, for $t = 1, \ldots, k$. Then, from (87) and (4), for any $w \in \bar{\mathcal{W}}_1$, we obtain:

$$\Pr\big(W_t = w_t \mid T = t\big) \;\leq\; 2^{-n} \left(1 + \frac{10}{15}\right) \;=\; \frac{5 \times 2^{-n}}{3}. \tag{90}$$

Moreover, it follows from (81) and the union bound that

$$P\big(\bar{\mathcal{W}}_1\big) \;\geq\; 1 - 2k^{-3}. \tag{91}$$

We now proceed to prove the second part of the lemma, i.e. (43). Again, fix a $t_0 \leq k$ and let ψ be a random matrix of coin-flip outcomes in which the biased coin has index $T = t_0$. In this case, the columns $\psi_1, \ldots, \psi_k$ of ψ are independent random vectors. For $t = 1, \ldots, k$, let

$$y_t \;=\; f(\psi_t) \;\triangleq\; \min\left(\max\left(2^n \Pr\big(W_t = \psi_t \mid T = t\big),\; 1 - \frac{8}{\mathcal{C}\sqrt{\ln k}}\right),\; 1 + \frac{10}{\mathcal{C}\sqrt{\ln k}}\right). \tag{92}$$

Since each y_t is only a function of ψ_t, it follows that $y_1, \ldots, y_k$ are independent random variables. Moreover, every y_t lies in an interval of length $18/\mathcal{C}\sqrt{\ln k}$. On the other hand, it follows from (89) that for $t = 1, \ldots, k$, with probability at least $1 - 2k^{-4}$, we have $y_t = 2^n \Pr\big(W_2 = \psi_t \mid T = t\big)$. The union bound then implies that with probability at least $1 - 2k^{-3}$,

$$y_t = 2^n \Pr\big(W_t = \psi_t \mid T = t\big), \qquad \text{for } t = 1, \ldots, k. \tag{93}$$

Therefore, for the random matrix ψ sampled from a distribution with biased coin index $T = t_0$, we have

$$\begin{aligned} P(\psi) \;&=\; \Pr(W = \psi) \\ &=\; \frac{1}{k} \sum_{t=1}^{k} \Pr\big(W = \psi \mid T = t\big) \\ &=\; \frac{1}{k} \sum_{t=1}^{k} 2^{-(k-1)n} \Pr\big(W_t = \psi_t \mid T = t\big) \\ &=\; \frac{2^{-kn}}{k} \sum_{t=1}^{k} 2^n \Pr\big(W_t = \psi_t \mid T = t\big) \end{aligned} \tag{94}$$

It then follows from (93) that with probability at least $1 - 2k^{-3}$,

$$P(\psi) \;=\; \frac{2^{-kn}}{k} \sum_{t=1}^{k} y_t. \tag{95}$$

Let

$$\beta \triangleq \mathbb{E}\left[\frac{2^{-kn}}{k} \sum_{t=1}^{k} y_t\right]. \tag{96}$$

Claim 1. $|\beta - 2^{-kn}| \leq 2^{-kn}/k$.

Proof. Temporarily, fix a $t \neq t_0$ and let

$$\mathcal{U}^+ \;\triangleq\; \left\{u \in \{0,1\}^n \;\;\middle|\;\; 2^n \Pr\big(\psi_t = u \mid T = t\big) > 1 + \frac{10}{\mathcal{C}\sqrt{\ln k}}\right\}, \tag{97}$$

$$\mathcal{U}^- \;\triangleq\; \left\{u \in \{0,1\}^n \;\;\middle|\;\; 2^n \Pr\big(\psi_t = u \mid T = t\big) < 1 - \frac{8}{\mathcal{C}\sqrt{\ln k}}\right\}. \tag{98}$$

Then, it follows from (89) that

$$\sum_{u \in \mathcal{U}^+} \Pr\big(\psi_t = u \mid T = t\big) \;\leq\; 2k^{-4}. \tag{99}$$

On the other hand, (88) implies that for any $u \in \mathcal{U}^-$

$$|\delta(u) - n/2| \geq 2.5\sqrt{n \ln k}, \tag{100}$$

where $\delta(u)$ is the number of 1s in the binary vector u. Let $z_1, \ldots, z_n$ be i.i.d. binary outcomes of a fair coin flip, and let $Z = z_1 + \cdots + z_n$. Then,

$$\begin{aligned}
\sum_{u \in \mathcal{U}^-} \frac{1}{2^n} \;&\leq \sum_{\substack{u \in \{0,1\}^n \\ |\delta(u) - n/2| \geq 2.5\sqrt{n \ln k}}} \frac{1}{2^n} \\
&= \Pr\left(\Big|Z - \frac{n}{2}\Big| \;\geq\; 2.5\sqrt{n \ln k}\right) \\
&\leq 2 \exp\left(\frac{2 \times \left(2.5\sqrt{n \ln k}\right)^2}{n}\right) \\
&\leq 2 \exp\big(-4 \ln k\big) \\
&= 2k^{-4},
\end{aligned} \tag{101}$$

where the first inequality follows from (100), the first equality is because Z has uniform distribution over $\{0,1\}^n$, and the second inequality is due to the Hoeffding's inequality.

We now expand $\mathbb{E}[y_t]$ as follows. From (92), we have

$$\begin{aligned}
\mathbb{E}[y_t] &= \mathbb{E}\big[f(\psi_t)\big] \\
&= \sum_{u\in\{0,1\}^n} \Pr\big(\psi_t = u\big) f(u) \\
&= \frac{1}{2^n} \sum_{u\in\{0,1\}^n} f(u) \\
&= \frac{1}{2^n} \sum_{u\in\{0,1\}^n} \min\left(\max\left(2^n \Pr\big(\psi_t = u \mid T = t\big),\ 1 - \frac{8}{\mathcal{C}\sqrt{\ln k}}\right),\ 1 + \frac{10}{\mathcal{C}\sqrt{\ln k}}\right) \\
&= \frac{1}{2^n}\Bigg(\sum_{u\in\{0,1\}^n} 2^n \Pr\big(\psi_t = u \mid T = t\big) \\
&\quad + \sum_{u\in\mathcal{U}^+}\left[\left(1 + \frac{10}{\mathcal{C}\sqrt{\ln k}}\right) - 2^n \Pr\big(\psi_t = u \mid T = t\big)\right] \\
&\quad + \sum_{u\in\mathcal{U}^-}\left[\left(1 - \frac{8}{\mathcal{C}\sqrt{\ln k}}\right) - 2^n \Pr\big(\psi_t = u \mid T = t\big)\right]\Bigg) \\
&= 1 - \sum_{u\in\mathcal{U}^+}\left[\Pr\big(\psi_t = u \mid T = t\big) - 2^{-n}\left(1 + \frac{10}{\mathcal{C}\sqrt{\ln k}}\right)\right] \\
&\quad + \frac{1}{2^n}\sum_{u\in\mathcal{U}^-}\left[\left(1 - \frac{8}{\mathcal{C}\sqrt{\ln k}}\right) - 2^n \Pr\big(\psi_t = u \mid T = t\big)\right]\Bigg),
\end{aligned} \tag{102}$$

where the third equality is because $t \neq t_0$ and as a result $\Pr\big(\psi_t = u\big) = 2^{-n}$ for all $u \in \{0,1\}^n$, and the fifth equality follows from the definitions of $\mathcal{U}^+$ and $\mathcal{U}^-$ in (97) and (98), respectively. From (102), we have

$$\mathbb{E}[y_t] \geq 1 - \sum_{u\in\mathcal{U}^+} \Pr\big(\psi_t = u \mid T = t\big) \geq 1 - 2k^{-4}, \tag{103}$$

where the second inequality is due to (99). Moreover, it follows from (102) that

$$\mathbb{E}[y_t] \leq 1 + \frac{1}{2^n}\sum_{u\in\mathcal{U}^-}\left(1 - \frac{8}{\mathcal{C}\sqrt{\ln k}}\right) \leq 1 + \sum_{u\in\mathcal{U}^-}\frac{1}{2^n} \leq 1 + 2k^{-4}, \tag{104}$$

where the first inequality is due to (102), and the last inequality is form (101). Combining (103) and (104), it follows that for any $t \neq t_0$,

$$\big|\mathbb{E}[y_t] - 1\big| \leq 2k^{-4}. \tag{105}$$

On the other hand, (4) implies that $\mathcal{C}\sqrt{\ln k} \geq 15$. Therefore, from the definition of y_t, we have $y_{t_0} \in \big(1 - 8/15,\ 1 + 10/15\big)$. Therefore,

$$\big|\mathbb{E}[y_{t_0}] - 1\big| \leq \frac{2}{3}. \tag{106}$$

Combining (105) and (106), we obtain

$$\begin{aligned}
\big|\beta - 2^{-kn}\big| &= 2^{-kn}\left|\frac{1}{k}\sum_{t=1}^{k}\mathbb{E}[y_t] - 1\right| \\
&\leq \frac{2^{-kn}}{k}\sum_{t=1}^{k}\big|\mathbb{E}[y_t] - 1\big| \\
&\leq \frac{2^{-kn}}{k}\left[(k-1)\times 2k^{-4} + \frac{2}{3}\right] \\
&\leq \frac{2^{-kn}}{k},
\end{aligned} \tag{107}$$

where the first equality is from the definition of β in (96), the third inequality follows from (105) and (106), and the last inequality is due to the assumption $k^{-3} \leq 1/6$ (see (5) with identification $k = mB$). This completes the proof of Claim 1. $\square$

We proceed with the proof of the lemma. Since $y_1, \ldots, y_k$ are independent random variables over an interval of length $18/\mathcal{C}\sqrt{\ln k}$, employing the Hoeffding's inequality we have

$$\begin{aligned}
\Pr\left(\left|P(\psi) - 2^{-kn}\right| \geq \frac{23 \times 2^{-nk}}{\mathcal{C}\sqrt{k}} + \frac{2^{-kn}}{k}\right) &\leq \Pr\left(\left|P(\psi) - \beta\right| \geq \frac{23 \times 2^{-nk}}{\mathcal{C}\sqrt{k}}\right) \\
&\leq \Pr\left(\left|P(\psi) - \beta\right| \geq \frac{23 \times 2^{-nk}}{\mathcal{C}\sqrt{k}} \;\middle|\; P(\psi) = \frac{2^{-kn}}{k}\sum_{t=1}^{k} y_t\right) \\
&\quad + \Pr\left(P(\psi) \neq \frac{2^{-kn}}{k}\sum_{t=1}^{k} y_t\right) \\
&\leq \Pr\left(\left|\frac{1}{k}\sum_{t=1}^{k} y_t - 2^{kn}\beta\right| \geq \frac{23}{\mathcal{C}\sqrt{k}}\right) + 2k^{-3} \qquad (108) \\
&\leq 2\exp\left(\frac{-2k\left(23/\mathcal{C}\sqrt{k}\right)^2}{\left(18/\mathcal{C}\sqrt{\ln k}\right)^2}\right) + 2k^{-3} \\
&= 2\exp\left(\frac{-2 \times 23^2}{18^2}\ln k\right) + 2k^{-3} \\
&\leq 2\exp\big(-3\ln k\big) + 2k^{-3} \\
&\leq 4k^{-3},
\end{aligned}$$

where the first inequality follows from Claim 1, the third inequality is due to (95) and the fourth inequality follows from the Hoeffding's inequality (see Lemma 7 (a)) and the definition of β in (96).

Since t_0 was chosen arbitrarily, (108) holds when the biased coin has any index in $1, \ldots, k$, and as a result it also holds when the biased coin is chosen uniformly at random from $1, \ldots, k$. Finally, we define a subset $\bar{\mathcal{W}}_2 \subset \mathcal{W}$ as

$$\bar{\mathcal{W}}_2 \triangleq \left\{w \in \mathcal{W} : \quad \left|P(\psi) - 2^{-kn}\right| \leq \frac{23 \times 2^{-nk}}{\mathcal{C}\sqrt{k}} + \frac{2^{-kn}}{k}\right\}, \qquad (109)$$

and let $\bar{\mathcal{W}} = \bar{\mathcal{W}}_1 \bigcap \bar{\mathcal{W}}_2$. Employing a union bound on (91) and (108), it follows that $P\big(\bar{\mathcal{W}}\big) \geq 1 - 6k^{-3}$. Moreover, Equations (42) and (43) in the lemma statement follow from (90) and (109), respectively. This completes the proof of Lemma 9.

C. Proof of Lemma 10

In this appendix, we present the proof of Lemma 10. For Part (a), let $f(x) = x\log_2(1/x)$. Then $f'(x) = \log_2(1/x) - \log_2 e$, where $e \simeq 2.718$ is the basis of the natural logarithm. As a result, $f'(x) \geq -1.5$, for all $x \in (0, 1]$. Consequently, for any $s \in \mathcal{S}$,

$$f\big(P(s)\big) \;\geq\; f\big(P(\bar{s})\big) - 1.5\big(P(s) - P(\bar{s})\big) \qquad (110)$$

Then,

$$\begin{aligned}
H(S) &= \sum_{s\in\mathcal{S}} P(s)\log_2\frac{1}{P(s)} \\
&= \sum_{s\in\mathcal{S}} f\big(P(s)\big) \\
&\geq \sum_{s\in\mathcal{S}} \Big(f\big(P(\bar{s})\big) - 1.5\big(P(s) - P(\bar{s})\big)\Big) \\
&= \left(\sum_{s\in\mathcal{S}} f\big(P(\bar{s})\big)\right) - 1.5\left(\sum_{s\in\mathcal{S}} P(s) - \sum_{s\in\mathcal{S}} P(\bar{s})\right) \\
&= \left(\sum_{s\in\mathcal{S}} P(\bar{s})\log_2\frac{1}{P(\bar{s})}\right) - 1.5\big(1 - P(\bar{\mathcal{W}})\big) \\
&\geq \left(\sum_{s\in\mathcal{S}} P(\bar{s})\log_2\frac{1}{P(\bar{s})}\right) - 9k^{-3},
\end{aligned}$$

where the first equality is from the definition of entropy (see [48], page 14), and the inequalities are due to (110) and Lemma 9, respectively. This completes the proof of Part (a) of the lemma.

We now proceed to the proof of Part (b). Following similar steps as in the proof of Part (a), it can be shown than for $t = 1, \ldots, k$,

$$\begin{aligned} H\big(S \mid T = t\big) \;&\triangleq\; \sum_{s \in \mathcal{S}} P\big(s \mid T = t\big) \log_2 \frac{1}{P\big(s \mid T = t\big)} \\ &\geq \left(\sum_{s \in \mathcal{S}} P\big(\bar{s} \mid T = t\big) \log_2 \frac{1}{P\big(\bar{s} \mid T = t\big)} \right) \;-\; \frac{9}{k^3}. \end{aligned} \tag{111}$$

Let $\bar{\mathcal{W}}^c$ be the complement of the set $\bar{\mathcal{W}}$, and for any $s \in \mathcal{S}$, let $\tilde{s} \triangleq s \bigcap \bar{\mathcal{W}}^c$. Then, from Lemma 9, we have

$$\sum_{s \in \mathcal{S}} P(\tilde{s}) \;=\; P\big(\bar{\mathcal{W}}^c\big) \;\leq\; 6k^{-3}. \tag{112}$$

It is easy to verify that $x \log_2(1/x) \leq 3.2 x^{5/6}$, for all $x \geq 0$. Then,

$$6k^{-3} \log_2 \frac{k^3}{6} \;\leq\; 3.2 \left(6k^{-3}\right)^{5/6} \;\leq\; 15k^{-2.5} \tag{113}$$

Let $|\mathcal{S}|$ be the number of elements in $\mathcal{S}$. Since $\mathcal{S}$ comprises the set of all B-bit signals, we have $|\mathcal{S}| = 2^B$. It follows from the Jensen's inequality (see [48], page 25) that for fixed $\sum_{s \in \mathcal{S}} P(\tilde{s})$, the value of $\sum_{s \in \mathcal{S}} P(\tilde{s}) \log_2 \big(1/P(\tilde{s})\big)$ is maximized when all $P(\tilde{s})$, for $s \in \mathcal{S}$, have equal probability. Therefore,

$$\begin{aligned} \sum_{s \in \mathcal{S}} P(\tilde{s}) \log_2 \frac{1}{P(\tilde{s})} \;&\leq\; \sum_{s \in \mathcal{S}} \frac{\sum_{s \in \mathcal{S}} P(\tilde{s})}{|S|} \log_2 \frac{|S|}{\sum_{s \in \mathcal{S}} P(\tilde{s})} \\ &=\; P\big(\bar{\mathcal{W}}^c\big) \log_2 \frac{|S|}{P\big(\bar{\mathcal{W}}^c\big)} \\ &=\; P\big(\bar{\mathcal{W}}^c\big) B \;+\; P\big(\bar{\mathcal{W}}^c\big) \log_2 \frac{1}{P\big(\bar{\mathcal{W}}^c\big)} \\ &\leq\; \frac{6B}{k^3} \;+\; P\big(\bar{\mathcal{W}}^c\big) \log_2 \frac{1}{P\big(\bar{\mathcal{W}}^c\big)} \\ &\leq\; \frac{6B}{k^3} \;+\; 6k^{-3} \log_2 \frac{k^3}{6} \\ &\leq\; \frac{6B}{k^3} \;+\; \frac{15}{k^{2.5}} \\ &=\; \frac{6B + 15\sqrt{k}}{k^3}, \end{aligned} \tag{114}$$

where the first equality follows from (112), the second equality is because $|\mathcal{S}| = 2^B$, the second inequality is due to (112), the third inequality is again because of (112) and the fact that $x \log_2(1/x)$ is an increasing function for $x \in [0, 1/e]$, and the last inequality follows from (113). Consequently,

$$\begin{aligned} H(S) \;&=\; \sum_{s \in \mathcal{S}} P(s) \log_2 \frac{1}{P(s)} \\ &=\; \sum_{s \in \mathcal{S}} \big(P(\bar{s}) + P(\tilde{s})\big) \log_2 \frac{1}{P(\bar{s}) + P(\tilde{s})} \\ &\leq\; \sum_{s \in \mathcal{S}} P(\bar{s}) \log_2 \frac{1}{P(\bar{s})} \;+\; \sum_{s \in \mathcal{S}} P(\tilde{s}) \log_2 \frac{1}{P(\tilde{s})} \\ &\leq\; \sum_{s \in \mathcal{S}} P(\bar{s}) \log_2 \frac{1}{P(\bar{s})} \;+\; \frac{6B + 15\sqrt{k}}{k^3}, \end{aligned} \tag{115}$$

where the last inequality is due to (114).

On the other hand, for any $x, y > 0$, we have

$$\begin{aligned} y\log_2 y - x\log_2 x &= (y-x)\log_2 x + y\big(\log_2 y - \log_2 x\big) \\ &= (y-x)\log_2 x + y\log_2\frac{y}{x} \\ &\le (y-x)\log_2 x + \frac{y}{\ln 2}\left(\frac{y}{x}-1\right) \\ &= (y-x)\log_2 x + \frac{1}{\ln 2}\left[\frac{y^2-yx}{x} + \frac{x^2-yx}{x} - \frac{x^2-yx}{x}\right] \\ &= (y-x)\log_2 x + \frac{1}{\ln 2}\left[\frac{x^2+y^2-2yx}{x} - (x-y)\right] \\ &= (y-x)\log_2(xe) + \frac{(x-y)^2}{x\ln 2}, \end{aligned} \tag{116}$$

where the inequality is because $\log_2\alpha \le (\alpha-1)/\ln 2$, for all $\alpha > 0$. Combining (111), (115), and (116), we obtain

$$\begin{aligned} I(T;S) &= H(S) - \sum_{t=1}^{k} P\big(T=t\big)H\big(S \mid T=t\big) \\ &= H(S) - \frac{1}{k}\sum_{t=1}^{k} H\big(S \mid T=t\big) \\ &\le H(S) - \frac{1}{k}\sum_{t=1}^{k}\left(\sum_{s\in\mathcal{S}} P\big(\bar{s} \mid t\big)\log_2\frac{1}{P\big(\bar{s} \mid t\big)} - \frac{9}{k^3}\right) \\ &\le \sum_{s\in\mathcal{S}} P(\bar{s})\log_2\frac{1}{P(\bar{s})} + \frac{6B+15\sqrt{k}}{k^3} \\ &\quad - \frac{1}{k}\sum_{t=1}^{k}\left(\sum_{s\in\mathcal{S}} P\big(\bar{s} \mid t\big)\log_2\frac{1}{P\big(\bar{s} \mid t\big)} - \frac{9}{k^3}\right) \\ &= \frac{1}{k}\sum_{t=1}^{k}\sum_{s\in\mathcal{S}}\Big(P\big(\bar{s} \mid t\big)\log_2 P\big(\bar{s} \mid t\big) - P(\bar{s})\log_2 P(\bar{s})\Big) + \frac{9+6B+15\sqrt{k}}{k^3} \\ &\le \frac{1}{k}\sum_{t=1}^{k}\sum_{s\in\mathcal{S}}\left[\Big(P\big(\bar{s} \mid t\big) - P(\bar{s})\Big)\log_2\big(P(\bar{s})e\big) + \frac{\big(P(\bar{s}) - P(\bar{s} \mid t)\big)^2}{P(\bar{s})\ln 2}\right] \\ &\quad + \frac{9+6B+15\sqrt{k}}{k^3} \\ &= \sum_{s\in\mathcal{S}}\log_2\big(P(\bar{s})e\big)\left[\left(\frac{1}{k}\sum_{t=1}^{k} P\big(\bar{s} \mid t\big)\right) - P(\bar{s})\right] \\ &\quad + \frac{1}{k\ln 2}\sum_{s\in\mathcal{S}} P(\bar{s})\sum_{t=1}^{k}\left(\frac{P(\bar{s}) - P(\bar{s} \mid t)}{P(\bar{s})}\right)^2 + \frac{9+6B+15\sqrt{k}}{k^3} \\ &= \sum_{s\in\mathcal{S}}\log_2\big(P(\bar{s})e\big)\times 0 \\ &\quad + \frac{1}{k\ln 2}\sum_{s\in\mathcal{S}} P(\bar{s})\sum_{t=1}^{k}\left(\frac{P(\bar{s} \mid t)}{P(\bar{s})} - 1\right)^2 + \frac{9+6B+15\sqrt{k}}{k^3} \\ &= \frac{1}{k\ln 2}\sum_{s\in\mathcal{S}} P(\bar{s})\sum_{t=1}^{k}\left(\frac{P(\bar{s} \mid T=t)}{P(\bar{s})} - 1\right)^2 + \frac{9+6B+15\sqrt{k}}{k^3}, \end{aligned}$$

where the first equality is from the definition of mutual information (see [48], page 20), the first inequality is due to (111), the second inequality is from (115), and the third inequality follows from (116). This completes the proof of Lemma 10.

D. Proof of Lemma 11

Let $\mathcal{U}^+$ be a subset of $\{0,1\}^n$ that contains the 2^{n-1} elements $u \in \{0,1\}^n$ with largest values of $P(u)$. Also let $\mathcal{U}^-$ be a subset of $\{0,1\}^n$ that contains the 2^{n-1} elements $u \in \{0,1\}^n$ with smallest values of $P(u)$. Then $\mathcal{U}^+$ and $\mathcal{U}^-$ are disjoint sets

with $\mathcal{U}^+ \bigcup \mathcal{U}^- = \{0,1\}^n$. Moreover, for any $u \in \mathcal{U}^+$ and any $v \in \mathcal{U}^-$, we have $P(u) \geq P(v)$. Let $\theta \triangleq P(\mathcal{U}^+) = \sum_{u\in\mathcal{U}^+} P(u)$. Then, $P(\mathcal{U}^-) = 1-\theta$. Since $\sum_{u\in\{0,1\}^n} \alpha_u = 0$ and $\alpha_u \in [-1,1]$, for all $u \in \{0,1\}^n$, it is easy to see that $\sum_{u\in\{0,1\}^n} \alpha_u P(u)$ is maximized for the following choice of alpha:

$$\alpha_u = \begin{cases} 1, & u \in \mathcal{U}^+ \\ -1, & u \in \mathcal{U}^-. \end{cases} \tag{117}$$

Therefore, for any choice of α_u, $u \in \{0,1\}^n$, that satisfy the conditions in the lemma statement, we have

$$\begin{aligned} \left(\sum_{u\in\{0,1\}^n} \alpha_u P(u)\right)^2 &\leq \left(\sum_{u\in\mathcal{U}^+} P(u) - \sum_{u\in\mathcal{U}^-} P(u)\right)^2 \\ &= \big(P(\mathcal{U}^+) - P(\mathcal{U}^-)\big)^2 \\ &= \big(2\theta - 1\big)^2. \end{aligned} \tag{118}$$

Since each of $\mathcal{U}^+$ and $\mathcal{U}^-$ has 2^{n-1} elements, it follows that

$$H\big(U \mid U \in \mathcal{U}^+\big) \leq n-1 \quad \text{and} \quad H\big(U \mid U \in \mathcal{U}^-\big) \leq n-1. \tag{119}$$

It then follows from the *grouping axiom* (see [53], page 8) that

$$\begin{aligned} H(U) &= h(\theta) + \theta H\big(U \mid U \in \mathcal{U}^+\big) + (1-\theta) H\big(U \mid U \in \mathcal{U}^-\big) \\ &\leq h(\theta) + \theta(n-1) + (1-\theta)(n-1) \\ &= h(\theta) + n - 1, \end{aligned} \tag{120}$$

where $h(\theta) = \theta \log_2(1/\theta) + (1-\theta)\log_2\big(1/(1-\theta)\big)$ is the entropy of a binary random variable that equals 1 if $U \in \mathcal{U}^+$ and equals 0 otherwise.

Consider the function $f(x) = (2x-1)^2 + 1.5h(x)$, defined for $x \in [0,1]$. Then, for any $x \in (0,1)$,

$$f''(x) = 8 - \frac{1.5}{x \ln 2} - \frac{1.5}{(1-x)\ln 2} \leq 8 - 2\left(\frac{1}{x} + \frac{1}{1-x}\right) \leq 0. \tag{121}$$

Hence, f is a concave function and is symmetric over $[0,1]$. Therefore, $f(x)$ takes its maximum at $x = 1/2$. As a result, for any $x \in [0,1]$,

$$\big(2x-1\big)^2 + 1.5h(x) = f(x) \leq f(1/2) = 1.5. \tag{122}$$

Combining (118), (120), and (122), we obtain

$$\begin{aligned} \left(\sum_{u\in\{0,1\}^n} \alpha_u P(u)\right)^2 + 1.5H(U) &\leq \big(2\theta-1\big)^2 + 1.5H(U) \\ &\leq \big(2\theta-1\big)^2 + 1.5h(\theta) + 1.5(n-1) \\ &\leq 1.5 + 1.5(n-1) \\ &= 1.5n, \end{aligned} \tag{123}$$

where the inequalities are respectively due to (118), (120), and (122). This implies (47) and completes the proof of Lemma 11.

APPENDIX C
PROOFS OF LEMMAS FOR THE CENTRALIZED LOWER BOUND PROOF IN SECTION VI-D

A. Proof of Lemma 1

For $i = 1,\ldots,9$ and $j = 1,\ldots,mn$, let $x_j^i \in \{-1,1\}$ be the outcome of jth flip of the ith coin. For $i = 1,\ldots,9$, let $N^i = (x_1^i+1)/2 + \cdots + (x_{mn}^i+1)/2$ be the total number of observed 1s for the ith coin. We assume that the index of the biased coin is unknown and has a uniform prior. According to the Neyman-Pearson lemma (see page 59 in [54]), the *most powerful* test is the the likelihood ratio test that outputs a coin index $\hat{T} = i$ with the maximum value of N^i. Below, we derive a lower bound on the error probability of the above test, i.e., $\Pr\big(\hat{T} \neq T\big)$.

Without loss of generality assume that $T = 1$. Then, $\mathbb{E}[x_1^1] = 1/2\sqrt{mn}$, $\mathrm{var}(x_1^1) = 1 - 1/4mn$, and $\mathbb{E}\big[x_1^1 - \mathbb{E}[x_1^1]\big]^3 = 1 - 1/16m^2n^2$. Let

$$Y^1 = \frac{\sum_{j=1}^{mn}\big(x_j^1 - \mathbb{E}[x_j^1]\big)}{\sqrt{mn\ \mathrm{var}(x_1^1)}}, \tag{124}$$

and for $i = 2, \ldots, 9$ let $Y^i = x_1^i + \ldots + x_{mn}^i$. Then,

$$N^1 \;=\; \frac{mn}{2} \;+\; \frac{\sqrt{mn}\sqrt{1-1/4mn}\,Y^1 + \sqrt{mn}/2}{2}, \tag{125}$$

and for $i = 2, \ldots, 9$,

$$N^i \;=\; \frac{mn}{2} \;+\; \frac{\sqrt{mn}\,Y^i}{2}. \tag{126}$$

It then follows from the Berry-Esseen theorem (see [55], page 33) that for any $i \leq 9$ and any $t \in \mathbb{R}$,

$$\left|\Pr\left(Y^i > t\right) - Q(t)\right| \;\leq\; \frac{33}{4}\,\frac{\mathbb{E}\big[x_1^i - \mathbb{E}[x_1^i]\big]^3}{\mathrm{var}(x_1^i)^{1.5}\sqrt{mn}}, \tag{127}$$

where $Q(\cdot)$ is the Q-function of the standard normal distribution. Therefore, ,

$$\begin{aligned}
\Pr\left(N^1 > \frac{mn}{2} + 0.4\sqrt{mn}\right) \;&=\; \Pr\left(\frac{\sqrt{mn}\sqrt{1-1/4mn}\,Y^1 + \sqrt{mn}/2}{2} > 0.4\sqrt{mn}\right) \\
&=\; \Pr\left(Y^1 > \frac{0.3}{\sqrt{1-1/4mn}}\right) \\
&\leq\; Q\left(\frac{0.3}{\sqrt{1-1/4mn}}\right) + \frac{33}{4}\,\frac{1+1/4mn}{(1-1/4mn)^{1.5}\sqrt{mn}} \\
&\leq\; 0.3961,
\end{aligned} \tag{128}$$

where the first equality is due to (125), the first inequality follows from (127), and the last inequality is from the assumption $mn \geq 350000$ in (8). In the same vein, for $i = 2, \ldots, 9$,

$$\begin{aligned}
\Pr\left(N^i \leq \frac{mn}{2} + 0.4\sqrt{mn}\right) \;&=\; \Pr\left(\frac{\sqrt{mn}\,Y^i}{2} \leq 0.4\sqrt{mn}\right) \\
&=\; \Pr\left(Y^i \leq 0.8\right) \\
&\leq\; 1 - Q\left(0.8\right) + \frac{33}{4\sqrt{mn}} \\
&\leq\; 0.8021,
\end{aligned} \tag{129}$$

where the first equality is due to (126), the first inequality follows from (127), and the last inequality is from the assumption $mn \geq 350000$ in (8). Consequently,

$$\Pr\left(\max\left(N^2, \ldots, N^9\right) > \frac{mn}{2} + 0.4\sqrt{mn}\right) \;=\; 1 - \Pr\left(N^2 \leq \frac{mn}{2} + 0.4\sqrt{mn}\right)^8 \;\geq\; 1 - 0.8021^8 \;>\; 0.8286. \tag{130}$$

Finally, for the error probability of the aforementioned maximum likelihood test, we have

$$\begin{aligned}
\Pr\left(\hat{T} \neq T\right) \;&=\; \Pr\Big(\max\left(N^2, \ldots, N^9\right) > N^1\Big) \\
&\geq\; \Pr\left(\max\left(N^2, \ldots, N^9\right) > \frac{mn}{2} + 0.4\sqrt{mn} \quad \text{and} \quad N^1 \leq \frac{mn}{2} + 0.4\sqrt{mn}\right) \\
&=\; \Pr\left(\max\left(N^2, \ldots, N^9\right) > \frac{mn}{2} + 0.4\sqrt{mn}\right) \;\times\; \Pr\left(N^1 \leq \frac{mn}{2} + 0.4\sqrt{mn}\right) \\
&\geq\; 0.8286 \;\times\; \Pr\left(N^1 \leq \frac{mn}{2} + 0.4\sqrt{mn}\right) \\
&\geq\; 0.8286 \;\times\; (1 - 0.3961) \\
&>\; \frac{1}{2},
\end{aligned} \tag{131}$$

where the second equality is due to the independence of different coins, the second inequality follows from (130), and the third inequality is from (128). This completes the proof of Lemma 1.

B. Proof of Lemma 2

Consider a function $\tilde{h} : \mathbb{R}^d \to \mathbb{R}$ as follows. For any $\theta \in \mathbb{R}^n$,

$$h(\theta) \;=\; \begin{cases} 1/2 - \|\theta\| & \text{if } \|\theta\| \leq 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\tilde{\mathcal{G}} = \{-1, 0, 1\}^2$ be the integer grid with 9 points inside $[-1, 1]^2$. To any function $\sigma : \tilde{\mathcal{G}} \to \{-1, 1\}$, we associate a function $\tilde{f}_\sigma(\theta) \triangleq \sum_{p \in \tilde{\mathcal{G}}} \sigma(p)\, \tilde{h}(\theta - p)$ for all $\theta \in \mathbb{R}^n$.

For any $p \in \tilde{\mathcal{G}}$, we define a probability distribution $\tilde{P}_p$ over functions $\tilde{f}_\sigma$ as follows. For any $\sigma : \tilde{\mathcal{G}} \to \{-1, 1\}$,

$$\tilde{P}_p(\tilde{f}_\sigma) = 2^{-9} \left(1 - \frac{\sigma(p)}{2\sqrt{mn}}\right).$$

Intuitively, when a function $\tilde{f}_\sigma$ is sampled from $\tilde{P}_p$, it is as if for every $q \in \tilde{\mathcal{G}}$ with $q \neq p$, we have $\Pr\big(\sigma(q) = 1\big) = \Pr\big(\sigma(q) = -1\big) = 1/2$, and for $q = p$ we have $\Pr\big(\sigma(p) = 1\big) = 1/2 - 1/\big(4\sqrt{mn}\big)$. This is like, the values of $\sigma(q)$ for $q \neq p$ are chosen independently at random according to the outcome of a fair coin flip, while the value of $\sigma(p)$ is the outcome of an unfair coin flip with bias $-1/\big(4\sqrt{mn}\big)$. Similar to (23), it is easy to show that $F(\theta) = h\big(\theta - p\big)/2\sqrt{mn}$. Therefore, under probability distribution P_p, $\theta^* = p$ is the global minimizer of $F(\cdot)$, and for any $\theta \in \mathbb{R}^n$ with $\|\theta - p\| \geq 1/2$, we have $F(\theta) \geq F(\theta^*) + 1/4\sqrt{mn}$. Therefore, if there exists an estimator under which $F(\hat{\theta}) < F(\theta^*) + 1/4\sqrt{mn}$, with probability at least $1/2$, then we have $\|\hat{\theta} - p\| < 1/2$ with probability at least $1/2$. In this case, p is the closest grid-point of $\mathcal{G}$ to $\hat{\theta}$, and we can recover p from $\hat{\theta}$, with probability at least $1/2$. This contradict Lemma 1. Consequently, under any estimator, we have $F(\hat{\theta}) \geq F(\theta^*) + 1/4\sqrt{mn}$, with probability at least $1/2$. This completes the proof of Lemma 2.

APPENDIX D
PROOF OF LEMMAS FOR THE UPPER BOUND PROOF IN SECTION VII

A. Proof of Lemma 3

We begin with a simple inequality: for any $x \in [0, 1]$ and any $k > 0$,

$$1 - (1 - x)^k \geq 1 - e^{-kx} \geq \frac{1}{2} \min\big(kx, 1\big). \tag{132}$$

Let Q_p be the probability that p appears in the p-component of at least one of the sub-signals of machine i. Then, for $p \in G^l$,

$$\begin{aligned} Q_p &= 1 - \left(1 - 2^{-dl} \times \frac{2^{(d-2)l}}{\sum_{j=1}^{t} 2^{(d-2)j}}\right)^{\lfloor B/d \log_2 mn \rfloor} \\ &\geq \frac{1}{2} \min\left(\frac{2^{-2l} \lfloor B/(d \log_2 mn) \rfloor}{\sum_{j=1}^{t} 2^{(d-2)j}}, 1\right) \\ &\geq \frac{1}{2} \min\left(\frac{2^{-2l} B}{2d \ln(mn) \sum_{j=1}^{t} 2^{(d-2)j}}, 1\right), \end{aligned}$$

where the equality is due to the probability of a point p in G^l (see (10)) and the number $\lfloor B/(d \log_2 mn) \rfloor$ of sub-signals per machine, and the first inequality is due to (132). Then,

$$\mathbb{E}\big[N_p\big] = Q_p m \geq \min\left(\frac{2^{-2l} mB}{4d \ln(mn) \sum_{j=1}^{t} 2^{(d-2)j}}, \frac{m}{2}\right). \tag{133}$$

We now bound the two terms on the right hand side of (133). For the second term on the right hand side of (133), we have

$$\begin{aligned} \frac{m}{2} &= \frac{m\epsilon^2}{2\epsilon^2} \\ &\geq \frac{16md \ln^4 mn}{2mn\epsilon^2} \\ &= \frac{8d \ln^4 mn}{n\epsilon^2}, \end{aligned} \tag{134}$$

where the first inequality is from the definition of ϵ in (34). For the first term at the right hand side of (133), note that

$$t = \log_2(1/\delta) \leq \log_2\left(\frac{\sqrt{m}}{\ln mn}\right) < \ln m. \tag{135}$$

It follows that for any $d \geq 1$,

$$\begin{aligned}
\sum_{j=1}^{t} 2^{(d-2)j} &\leq t2^{t(d-2)} \\
&\leq \ln(mn)\, 2^{t(d-2)} \\
&= \ln(mn)\, \left(\frac{1}{\delta}\right)^{(d-2)} \\
&= \ln(mn)\, \delta^2\, \left(\frac{1}{\delta}\right)^{d} \\
&\leq \ln(mn)\, \delta^2\, \frac{mB}{\ln^{2d} mn} \\
&= \ln(mn) \times \frac{n\epsilon^2}{16d \ln^2 mn} \times \frac{mB}{\ln^{2d} mn} \\
&\leq \frac{nmB\epsilon^2}{16d \ln^5 mn},
\end{aligned}$$

where the second inequality is due to (135), the third inequality follows from the definition of δ, the third equality is from the definition of ϵ in (34), and the last inequality is because of the assumption $d \geq 2$. Then,

$$\begin{aligned}
\frac{2^{-2l} mB}{4d \ln(mn) \sum_{j=1}^{t} 2^{(d-2)j}} &\geq \frac{2^{-2l} mB}{4d \ln(mn)} \times \frac{16d \ln^5 mn}{nmB\epsilon^2} \\
&= \frac{4 \ln^4(mn)\, 2^{-2l}}{n\epsilon^2}.
\end{aligned} \tag{136}$$

Consequently,

$$\begin{aligned}
\frac{2^{-2l} mB}{4d \ln(mn) \sum_{j=1}^{t} 2^{(d-2)j}} &\geq \frac{4 \ln^4(mn)\, 2^{-2l}}{n\epsilon^2} \\
&\geq \frac{4 \ln^4(mn)\, 2^{-2t}}{n\epsilon^2} \\
&= \frac{4 \ln^4(mn)\, \delta^2}{n\epsilon^2} \\
&= \frac{4 \ln^4(mn)\, \delta^2}{16d\delta^2\, \ln^2 mn} \\
&= \frac{\ln^2(mn)}{4d},
\end{aligned} \tag{137}$$

where the first equality is due to the definition of $t = \ln_2(1/\delta)$, and the second equality is from the definition of ϵ. Plugging (134) and (136) into (133), it follows that for $l = 1, \ldots, t$ and for any $p \in G^l$,

$$\mathbb{E}\big[N_p\big] \;\geq\; \frac{4 \ln^4(mn)\, 2^{-2l}}{n\epsilon^2}. \tag{138}$$

Moreover, plugging (137) into (133), we obtain

$$\begin{aligned}
\frac{1}{8}\mathbb{E}\big[N_p\big] &\geq \frac{1}{8}\, \min\left(\frac{\ln^2(mn)}{4d},\, \frac{m}{2}\right) \\
&\geq \frac{1}{8}\, \min\left(\frac{\ln^2(mn)}{4d},\, \frac{\ln^2 mn}{2}\right) \\
&\geq \frac{\ln^2(mn)}{32d},
\end{aligned} \tag{139}$$

where the second inequality is because of the assumption $m \geq \ln^2 mn$ in (15). Then, for $l \in 1, \ldots, t$ and any $p \in \tilde{G}^l_{s^*}$,

$$\begin{aligned}
\Pr\left(N_p \leq \frac{2 \ln^4(mn)\, 2^{-2l}}{n\epsilon^2}\right) &\leq \Pr\left(N_p \leq \frac{\mathbb{E}[N_p]}{2}\right) \\
&\leq \exp\big(-(1/2)^2 \mathbb{E}[N_p]/2\big) \\
&\leq \exp\big(-\ln^2(mn)/32d\big),
\end{aligned} \tag{140}$$

where the inequalities are due to (138), Lemma 7 (b), and (139), respectively. Then,

$$\begin{aligned}
\Pr\big(\mathcal{E}\big) \;&=\; \Pr\left(N_p \geq \frac{2\ln^4(mn)\,2^{-2l}}{n\epsilon^2},\quad \forall p\in G^l \text{ and for } l=1,\ldots,t\right)\\
&\geq\; 1-\sum_{l=1}^{t}\sum_{p\in G^l}\Pr\left(N_p < \frac{2\ln^4(mn)\,2^{-2l}}{n\epsilon^2}\right)\\
&\geq\; 1-t2^{dt}\exp\big(-\ln^2(mn)/(32d)\big)\\
&=\; 1-\ln(1/\delta)\left(\frac{1}{\delta}\right)^d\exp\big(-\ln^2(mn)/(32d)\big)\\
&\geq\; 1-\ln(mn)\,\frac{m^{d/2}}{\ln^d mn}\exp\big(-\ln^2(mn)/32d\big)\\
&\geq\; 1-m^{d/2}\exp\big(-\ln^2(mn)/32d\big),
\end{aligned}$$

where the first equality is by the definition of $\mathcal{E}$, the first inequality is from union bound, the second inequality is due to (140), and the third inequality follows from (135) and the definition of δ in (9). This completes the proof of Lemma 3.

B. Proof of Lemma 4

For any $l\leq t$ and any $p\in G^l$, let

$$\hat{\Delta}(p)=\frac{1}{N_p}\sum_{\substack{\text{Subsignals of the form}\\(p,\Delta,\cdot,\cdot)\\\text{after redundancy elimination}}}\Delta,$$

and let $\Delta^*(p)=\mathbb{E}[\hat{\Delta}(p)]$.

For $l\geq 1$, consider a grid point $p\in G^l$ and let p' be the parent of p. Then, $\|p-p'\|=\sqrt{d}\,2^{-l}$. Furthermore, by definition, for any function $f\in\mathcal{F}$, we have $|f(p)-f(p')|\leq\|p-p'\|$. Therefore, $\hat{\Delta}(p)$ is the average of $N_p\times n/2$ independent variables with absolute values no larger than $\sqrt{d}\,2^{-l}$. Given event $\mathcal{E}$, it then follows from the Hoeffding's inequality that

$$\begin{aligned}
&\Pr\left(\big|\hat{\Delta}(p)-\Delta^*(p)\big|\geq\frac{\epsilon}{8\ln(mn)}\right)\\
&\leq\; 2\exp\left(-nN_p\times\frac{1}{(2\sqrt{d}\,2^{-l})^2}\times\Big(\frac{\epsilon}{8\ln mn}\Big)^2\right)\\
&\leq\; 2\exp\left(-n\times\frac{2\ln^4(mn)\,2^{-2l}}{n\epsilon^2}\times\frac{1}{4d\,2^{-2l}}\times\frac{\epsilon^2}{64\ln^2 mn}\right)\\
&=\; 2\exp\big(-\ln^2(mn)/128d\big),
\end{aligned}$$

Recall from (13) that for $l=1,\ldots,t$ and any $p\in G^l$ with parent p',

$$\hat{F}(p)-F(p)\;=\;\hat{F}(p')-F(p')+\hat{\Delta}(p)-\Delta^*(p).$$

Then,

$$\begin{aligned}
&\Pr\left(|\hat{F}(p)-F(p)|>\frac{l\epsilon}{8\ln mn}\right)\\
&\leq\;\Pr\left(|\hat{F}(p')-F(p')|>\frac{(l-1)\epsilon}{8\ln mn}\right)+\Pr\left(|\hat{\Delta}(p)-\Delta^*(p)|>\frac{\epsilon}{8\ln mn}\right)\\
&\leq\;\Pr\left(|\hat{F}(p')-F(p')|>\frac{(l-1)\epsilon}{8\ln mn}\right)+2\exp\big(-\ln^2(mn)/128d\big).
\end{aligned}$$

Employing an induction on l, we obtain for any $l\leq t$ and any $p\in G^l$,

$$\Pr\left(|\hat{F}(p)-F(p)|>\frac{l\epsilon}{8\ln mn}\right)\;\leq\;2l\,\exp\big(-\ln^2(mn)/128d\big).$$

Therefore,

$$\begin{aligned}
\Pr\left(|\hat{F}(p)-F(p)|>\frac{\epsilon}{8}\right)\;&\leq\;\Pr\left(|\hat{F}(p)-F(p)|>\frac{l\epsilon}{8\ln mn}\right)\\
&\leq\;2\ln(m)\,\exp\big(-\ln^2(mn)/128d\big),
\end{aligned}\tag{141}$$

where the inequalities are due to (135). It then follows from the union bound that

$$\begin{aligned}\Pr\big(\mathcal{E}' \mid \mathcal{E}\big) &\geq 1 - \sum_{l=1}^{t} \sum_{p\in G^l} \Pr\Big(|\hat{F}(p) - F(p)| > \frac{\epsilon}{8}\Big)\\ &\geq 1 - t2^{dt} \times 2\ln(m)\, \exp\big(-\ln^2(mn)/128d\big)\\ &\geq 1 - \ln(m) \times \left(\frac{1}{\delta}\right)^d \times 2\ln(m)\, \exp\big(-\ln^2(mn)/128d\big)\\ &\geq 1 - \ln(m) \times \frac{m^{d/2}}{\ln^d mn} \times 2\ln(m)\, \exp\big(-\ln^2(mn)/128d\big)\\ &\geq 1 - 2m^{d/2}\, \exp\big(-\ln^2(mn)/128d\big),\end{aligned} \tag{142}$$

where the second inequality is due to (141), the third inequality follows from (135), and the fourth inequality is from the definition of δ. On the other hand, we have from Lemma 3 that $\Pr\big(\mathcal{E}\big) = 1 - m^{d/2}\exp\big(-\ln^2(mn)/8d\big)$. Then, $\Pr\big(\mathcal{E}'\big) \geq 1 - m^{d/2}\exp\big(-\ln^2(mn)/32d\big) - 2m^{d/2}\exp\big(-\ln^2(mn)/128d\big)$ and Lemma 4 follows.

C. Proof of Lemma 5

Fix a machine i and let $g(\theta) = (F^i(\theta) - F^i(p)) - (F(\theta) - F(p))$, for all $\theta \in [-1,1]^d$. Note that for any function $f \in \mathcal{F}$, any $p \in G^t$ and any $\theta \in cell_p$, we have $|f(\theta) - f(p)| \leq \|\theta - p\| \leq \sqrt{d}\delta$. Then, $F^i(\theta) - F^i(p)$ is the average over $n/2$ randomly chosen such functions $f(\theta) - f(p)$ with the expected value $F(\theta) - F(p)$. It follows from Hoeffding's inequality (Lemma 7) that:

$$\begin{aligned}\Pr\Big(\big|g(\theta)\big| > \frac{\epsilon}{16}\Big) &= \Pr\Big(\big|(F^i(\theta) - F^i(p)) - (F(\theta) - F(p))\big| > \frac{\epsilon}{16}\Big)\\ &\leq 2\exp\left(-\frac{2\times n/2\times(\epsilon/16)^2}{(2\sqrt{d}\delta)^2}\right)\\ &= 2\exp\left(-n\left(\frac{4\delta\sqrt{d}\ln(mn)}{16\sqrt{n}\times 2\sqrt{d}\delta}\right)^2\right)\\ &= 2\exp\left(-\frac{\ln^2(mn)}{64}\right),\end{aligned} \tag{143}$$

where the first equality is due to the definition of $\epsilon = \delta\sqrt{d}\ln(mn)/\sqrt{2n}$.

Consider a regular grid $\mathcal{D}$ with edge size $\epsilon/16\sqrt{d}$ over $cell_p$. Then,

$$|\mathcal{D}| = \left(\frac{2\delta}{\epsilon/16\sqrt{d}}\right)^d = \left(\frac{32\delta\sqrt{d}\sqrt{n}}{4\delta\sqrt{d}\ln mn}\right)^d = \left(\frac{8\sqrt{n}}{\ln mn}\right)^d \leq n^{d/2},$$

where the second inequality is due to the definition of ϵ, and the last inequality is due to the assumption $\ln mn \geq 8\sqrt{d}$ in (15) It then follows from (143) and the union bound that with probability at least $1 - 2n^{d/2}\exp\big(-\ln^2(mn)/64\big)$, we have

$$|g(\theta)| \leq \frac{\epsilon}{16}, \quad \forall\theta \in \mathcal{D}. \tag{144}$$

On the other hand the function $g(\theta) = (F^i(\theta) - F^i(p)) - (F(\theta) - F(p))$ is the sum of two Lipschitz continuous functions, and is therefore Lipschitz continuous with constant 2. Consider an arbitrary $\theta \in cell_p$ and let θ' be the closest grid point in $\mathcal{D}$ to θ. Then, $\|\theta - \theta'\| \leq \epsilon/32$. Then, assuming (144), we have

$$\begin{aligned}|g(\theta)| &\leq |g(\theta)| + \big|g(\theta') - g(\theta)\big|\\ &\leq \frac{\epsilon}{16} + \big|g(\theta') - g(\theta)\big|\\ &\leq \frac{\epsilon}{16} + 2\big\|\theta' - \theta\big\|\\ &\leq \frac{\epsilon}{16} + \frac{2\epsilon}{32}\\ &= \frac{\epsilon}{8},\end{aligned} \tag{145}$$

where the second inequality is due to (144) and the third inequality follows from the Lipschitz continuity of g with constant 2. Employing union bound over all machines i and all cells $cell_p$ for $p \in G^t$, it follows from (144) and (145) that $\mathcal{E}''$ holds true with probability at least $1 - 2n^{d/2}m^{1+d/2}\exp\big(-\ln^2(mn)/64\big)$. This completes the proof of Lemma 5.