One-Shot Federated Learning for non-Convex Loss Functions

Saber Salehkaleybar SALEH@SHARIF.EDU

Arsalan Sharifnassab A.SHARIFNASSAB@GMAIL.COM

S. Jamaloddin Golestani GOLESTANI@SHARIF.EDU

Department of Electrical Engineering
Sharif University of Technology
Tehran, Iran

Editor:

Abstract

We consider the problem of federated learning in a one-shot setting where we have m machines who has access to n samples from an unknown distribution. Each machine constructs a signal of limited length B and sends it to a main server. The sever receives all the signals and outputs values for a model with d parameters which minimizes the expected non-convex loss function. We propose a distributed learning algorithm, called Multi-Resolution Estimator for Non-Convex loss function (MRE-NC), whose expected loss function is bounded by $(\log^3(mn)\sqrt{d})/((mB)^{1/d}\sqrt{2n})$ with respect to the optimal one with high probability. Experiments on synthetic data show the effectiveness of MRE-NC in distributed learning of model's parameters for the non-convex loss functions.

Keywords: Federated learning, Distributed learning, Few shot learning, Communication efficiency, non-Convex Optimization.

1. Introduction

Consider a set of m machines where each machine has access to n samples drawn from an unknown distribution. Each machine sends a single message with a limited length to a server which collects the received messages and output values for a model's parameters minimizing an empirical loss function over the whole data resides in machines.

The above setting with a single interaction between machines and the server is one of the scenarios that occurs in a machine learning paradigm which is commonly called "Federated Learning". With the advances in smart phones or tablets, these devices can collect unprecedented amount of data from interactions of users with mobile applications. This huge amount of data can be exploited to improve the performance of machine learning models running in smart devices. Due to the sensitive nature of the data, 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 the 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 interactions.

Several work have studied the problem of minimizing a convex loss function in the context of distributed learning where the machines are allowed to send one message with specific number of bits to the server. We call this scheme of communication between machines and the server, "one-shot federated learning". Earlier work (Zhang et al., 2012) proposed averaging method in which each machine obtains the optimal values for the parameters over its own data and sends it to the server. Afterwards, the server returns the average of received parameters from machines as the output. Unfortunately, the estimation error of averaging method does not go to zeros as the number of machines increases. Recently, for the convex loss functions in the setting of one-shot federated learning, Salehkaleybar et al. (2019b) proposed a lower bound on the estimation error achievable by any algorithm with a constraint on the number of bits sent by machines. They also proposed an estimator whose expected error meets the mentioned lower bound up to a logarithmic factor and therefore, it is order optimal. Although studying the case of non-convex loss function is more desirable in training deep neural networks, there is a little work in the literature of one-shot federated learning. Our aim in this paper is to propose an estimator for this setting with non-convex loss function which can obtain values for the parameters whose loss function is guaranteed to be close to the optimal one. In particular, propose an estimator, called Multi-Resolution Estimator for Non-Convex loss function (MRE-NC), whose expected loss function is bounded by $(\log^3(mn)\sqrt{d})/((mB)^{1/d}\sqrt{2n})$ with respect to the optimal one with high probability where d and B are dimension of model's parameters and signal length in bits, respectively.

1.1 Related Work

McMahan et al. (2017) 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. They proposed "FedAvg" algorithm to address some of the challenges which executes in several synchronous rounds. In each round, the server selects a fraction of machines randomly 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 $10 - 100\times$ with respect to the stochastic gradient descent (SGD) algorithm. Although there is no theoretical guarantee for the convergence of the proposed solution and it might diverge in some practical scenarios.

After introducing the setting of federated learning by McMahan et al. (2017), several research work addressed its challenges such as communication constraints, system heterogeneity (different computational and communication capabilities of 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 (Konečný et al., 2016), updating the model from a restricted space (Konečný et al., 2016), or utilizing lossless encodings (Sattler et al., 2019). To resolve system heterogeneity issues such as stragglers, asynchronous communication schemes with the assumption of bounded delay have been devised between the server and machines (Zinkevich et al., 2010; Ho et al., 2013). There are also several works provided convergence guarantee for the case of non-i.i.d samples

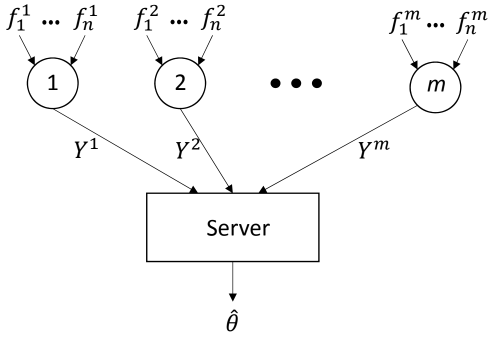

Figure 1: A distributed system of m machines, each having access to n independent sample functions from an unknown distribution P. Each machine sends a signal to a server based on its observations. The server receives all signals and output an estimate $\hat{\theta}$ for the optimization problem in (2).

distributed among the machines (Li et al., 2020; Wang et al., 2019). Moreover, some notions of privacy can be preserved by utilizing differential privacy techniques (Abadi et al., 2016; Geyer et al., 2017) or secure multi-party computation (Chen et al., 2019).

A similar setting to the one in 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 and the number of machines are much less than the one in federated learning setting. Moreover, 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 machine, then each of them can send their own data to the server. Hence, we can achieve the estimation of error a centralized solution who has access to entire data. However, the problem becomes non-trivial if each machine can only send a limited number of bits to the server. In the one-shot setting, Zhang et al. (2012) proposed a simple averaging method where each machine obtains the optimal values of the parameters for 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 of $O(1/\sqrt{mn} + 1/n)$. This bound can be improved to $O(1/\sqrt{mn} + 1/n^{1.5})$ with boot-strap method (Zhang et al., 2012) or $O(1/\sqrt{mn} + 1/n^{9/4})$ by optimizing a surrogate loss function using Taylor series expansion (Jordan et al., 2018). Recently, Salehkaleybar et al. (2019b) proposed a lower bound on the estimation error of any algorithm and also presented an algorithm which is order-optimal. For the case of multi-shot setting, the main 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 (Bottou, 2010; Lian et al., 2015; McMahan et al., 2017). In fact, FedAvg algorithm (McMahan et al., 2017) 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 function, there is no theoretical guarantee on the quality of the output. Moreover, in the one-shot setting, the problems becomes more challenging since these gradient descent based methods cannot be adopted easily to this setting.

1.2 Outline

The paper is organized as follows. We begin with a detailed model and problem definition in Section 2. In Section 3, we present the MRE-NC algorithm and its error upper bound. Afterwards, we report our numerical experiments in Section 4. Finally, in Section 5, we conclude the paper. All proofs are relegated to the appendix for improved readability.

2. Problem Definition

Consider a positive integer d and a collection $\mathcal{F}$ of real-valued functions over $[-1,1]^d$. Let P be an unknown probability distribution over the functions in $\mathcal{F}$. Consider the expected loss function

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

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

$$\theta^* = \operatorname*{argmin}_{\theta\in[-1,1]^d} F(\theta). \tag{2}$$

The expected loss is to be minimized in a distributed fashion, as follows. We consider a distributed system comprising m identical machines and a server. Each machine i has access to a set of n independently and identically distributed samples $\{f_1^i, \cdots, f_n^i\}$ drawn from the probability distribution P. Based on these observed functions, machine i then sends a signal Y^i to the server. We assume that the length of each signal is limited to b bits. The server then collects signals $Y^1, \ldots, Y^m$ and outputs an estimation of θ^*, which we denote by $\hat{\theta}$. See Fig. 1 for an illustration of the system model.[1]

We let the following assumptions be in effect throughout the paper:

Assumption 1 (Lipschitz Continuity) *We assume:*

- *Each $f \in \mathcal{F}$ is Lipschitz continuous. More concretely, for any $f \in \mathcal{F}$ and any $\theta, \theta' \in [-1,1]^d$, we have $\|f(\theta) - f(\theta')\| \le \|\theta - \theta'\|$.*
- *The minimizer of F lies in the interior of the cube $[-1,1]^d$.*

3. MRE-NC Algorithm and its Error Upper Bound

Here, we propose an order optimal estimator under general communication budget B, for $B \ge d\log mn$. The high level idea is that to obtain an approximation of F over the domain and then letting $\hat{\theta}$ be the minimizer of these approximations. For efficient function approximation, transmitted signals are designed such that the server can construct a multi-resolution view of function $F(\theta)$ 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 observes n functions and sends a signal Y^i comprising $\lceil B/(d\log mn)\rceil$ sub-signals of length $\lfloor d\log mn\rfloor$. Each sub-signal has four parts of the form $(p, \Delta, \theta^i, \Delta^i)$. The four parts p, Δ, θ^i, Δ^i are as follows.

[1] The considered model here is similar to the one in (Salehkaleybar et al., 2019a).

- Part p: Let

$$\delta \triangleq \frac{\log^2(mn)}{(mB)^{1/d}}. \tag{3}$$

Let $t = \log(1/\delta)$. Without loss of generality we assume that t is a non-negative integer.[2] Let C be a d-dimensional cube with edge size one centered at $s = (0, \cdots, 0)$. Consider a sequence of $t+1$ grids on C as follows. For each $l = 0, \ldots, t$, we partition the cube C into 2^{ld} smaller equal sub-cubes with edge size 2^{-l}. The lth grid G^l comprises 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 select p, we randomly choose an l from $0, \ldots, t$ with probability

$$\Pr(l) = \frac{2^{(d-2)l}}{\sum_{j=0}^{t} 2^{(d-2)j}}. \tag{4}$$

We then let p be a uniformly chosen random grid point in G^l. The level l and point p chosen in different sub-signals of a machine are independent and have the same distribution. Note that $O(d\log(1/\delta)) = O(d\log(mn))$ bits suffice to identify p uniquely.

- Part Δ: We let

$$\hat{F}^i(\theta) \triangleq \frac{2}{n}\sum_{j=1}^{n/2} f_j^i(\theta), \tag{5}$$

and refer to it as the empirical function of the ith machine. For each sub-signal, if the selected p in the previous part is in G^0, i.e., $p = (0, \cdots, 0)$, then we set Δ to the value of $\hat{F}^i$ at $\theta = s$. Otherwise, if p is in G^l for $l \geq 1$, we let

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

where $p' \in G^{l-1}$ is the parent of p. Note that Δ is in the range over $\big(2^{-l}\sqrt{d}\big) \times \big[-1, +1\big]$. This is due to the Lipschitz continuity of the functions in $\mathcal{F}$ (see Assumption 1) and the fact that $\|p - p'\| = 2^{-l}\sqrt{d}$. Hence, $O(d\log(mn))$ bits suffice to represent Δ within accuracy δ.

- Part θ^i, Δ^i: If machine i is selected a point p in G^t, it also sends another two sub-signals θ^i, Δ^i (Otherwise, it sends dummy messages for these parts). It finds the minimizer of $\frac{2}{n}\sum_{j=n/2+1}^{n} f_j^i(\theta)$ in the cell in G^t containing the point p and considers it as θ^i. Moreover, it sets $\Delta^i = \hat{F}^i(\theta^i) - \hat{F}^i(p)$. It can be seen that these sub-signals can be sent with $O(d\log(mn))$ bits within accuracy δ.

At the server, we approximate the function F over C as follows. We first eliminate redundant sub-signals so that no two surviving sub-signals from a same machine have the same p-parts (consequently, for each machine, the surviving sub-signals are distinct). We call this process

[2] 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.

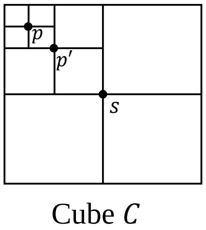

Figure 2: An illustration of cube C centered at point s for $d = 2$. The point p belongs to G^2 and p' is the parent of p.

"redundancy elimination". We then let N_s be the total number of surviving sub-signals that contain $s = (0, \cdots, 0)$ in their p part, and compute

$$\hat{F}(s) = \frac{1}{N_s} \sum_{\substack{\text{Subsignals of the form} \\ (s,\Delta,\theta^i,\Delta^i) \\ \text{after redundancy elimination}}} \Delta,$$

Then, for any point $p \in G^l$ with $l \geq 1$, we let

$$\hat{F}(p) = \hat{F}(p') + \frac{1}{N_p} \sum_{\substack{\text{Subsignals of the form} \\ (p,\Delta,\theta^i,\Delta^i) \\ \text{after redundancy elimination}}} \Delta, \tag{6}$$

where N_p is the number of signals having point p in their first argument, after redundancy elimination. Moreover, for each cell corresponding to a point p in G^t, we keep only one sub-signals from all signals of the form $(p, \Delta, \theta^i, \Delta^i)$ and consider $\hat{F}(\theta^i) = \hat{F}(p) + \Delta^i$. Finally, the server outputs θ_i with minimum $\hat{F}(\theta^i)$.

Theorem 1 *Let $\hat{\theta}$ be the output of the above algorithm. Then,*

$$\Pr\left(F(\hat{\theta}) > F(\theta^*) + \frac{\log^3(mn)\sqrt{d}}{(mB)^{1/d}\sqrt{2n}}\right) = \exp\Big(-\Omega\big(\log^2(mn)\big)\Big).$$

The proof is given in Appendix B. The proof 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 at all points of G^t, we can find a good estimate of minimum of F in each cell in G^t.

4. Experiments

We evaluated the performance of MRE-NC and compared with two naive approaches: 1- The averaging method: Each machine obtains empricial 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 return as the output.

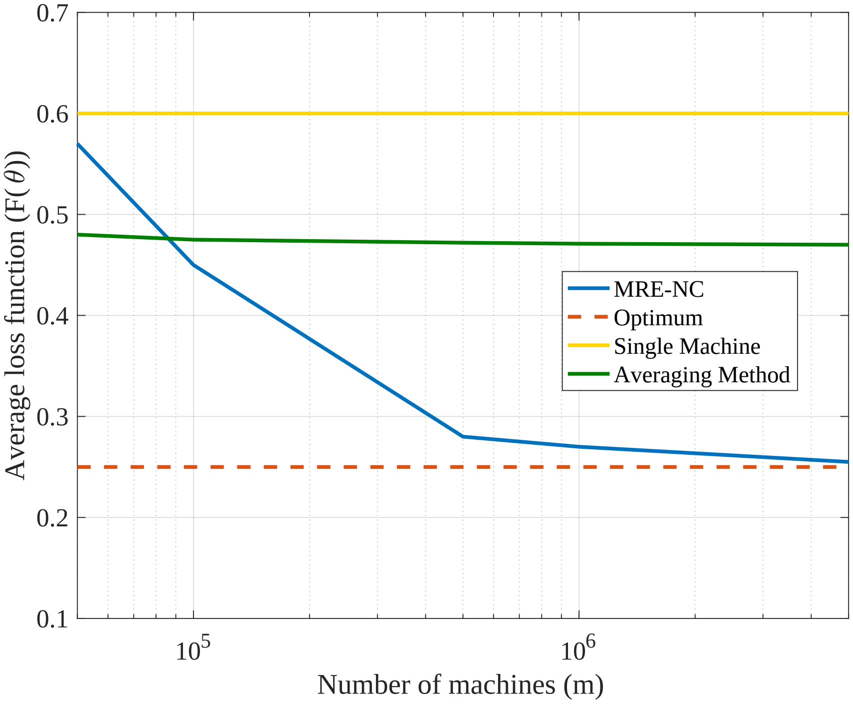

Figure 3: Comparison of the performance of MRE-NC with two naive approaches.

In our experiment, each sample (X, y), $X \in \mathbb{R}^2$, and $y \in R$ is generated according to $y = \theta_2^T ReLU(\theta_1 X) + N$ where the entries $[\theta_1]_{2\times2}$ are drawn randomly from the range $[-2, 2]$ and $\theta_2 = [1, -1]$. Moreover, N is sampled from $N(0, 0.5)$.

In Fig. 3, the value of $F(\theta)$ is depicted versus different 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 performance of MRE-NC improved as the number of machines and got close to the optimal one. However, as expected, two naive approaches failed in returning an output with low average loss function.

5. Conclusion

In this paper, we studied the problem of federated learning in a one-shot setting where we have m machines who has access to n samples from an unknown distribution. Each machine constructs a signal of limited length B and sends it to a main server. We propose MRE-NC whose expected loss function is bounded by $(\log^3(mn)\sqrt{d})/((mB)^{1/d}\sqrt{2n})$ with respect to the optimal one with high probability. Experiments on synthetic data showed the effectiveness of MRE-NC in distributed learning of model's parameters for the non-convex loss functions.

Acknowledgments

This research was supported by INSF under contract No. 97012846.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT*, pages 177–186. Springer, 2010.

Valerie Chen, Valerio Pastro, and Mariana Raykova. Secure computation for machine learning with spdz. *arXiv preprint arXiv:1901.00329*, 2019.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B Gibbons, Garth A Gibson, Greg Ganger, and Eric P Xing. More effective distributed ml via a stale synchronous parallel parameter server. In *Advances in neural information processing systems*, pages 1223–1231, 2013.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, pages 1–14, 2018.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge university press, 1995.

Saber Salehkaleybar, Arsalan Sharifnassab, and S. Jamaloddin Golestani. One-shot distributed learning: theoretical limits and algorithms to achieve them. *arXiv preprint arXiv:1905.04634v1*, 2019a.

Saber Salehkaleybar, Arsalan Sharifnassab, and S Jamaloddin Golestani. One-shot federated learning: theoretical limits and algorithms to achieve them. *arXiv preprint arXiv:1905.04634*, 2019b.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 2019.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

Yuchen Zhang, Martin J Wainwright, and John C Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23:2595–2603, 2010.

Appendices

Appendix A. Preliminaries

In this appendix, we review some preliminaries that will be used in the proofs of our main results.

A.1 Concentration inequalities

We collect two well-known concentration inequalities in the following lemma.

Lemma 2 *(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 $\alpha > 0$,

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

(b) (Theorem 4.2 in Motwani and Raghavan (1995)) 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 $\alpha \in (0,1]$,

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

Appendix B. Proof of Theorem 1

We first show that for any $l \leq t$ and 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 a good estimate of minimum of F in each cell in G^t. Let

$$\epsilon \triangleq \frac{\log^3(mn)\sqrt{d}}{(mB)^{1/d}\sqrt{2n}}. \tag{7}$$

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 any $l \leq t$ and any $p \in G^l$, we have

$$N_p \geq \frac{1}{d} 2^{-2l} \log^{2d-6}(mn)\,(mB)^{2/d}. \tag{8}$$

Then,

Lemma 3 $\Pr\big(\mathcal{E}\big) = 1 - \exp\big(-\Omega(\log^2(mn))\big)$.

The proof is based on the concentration inequality in Lemma 2 (b), and is given in Appendix C.

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

$$\big|\hat{F}(p) - F(p)\big| < \frac{\epsilon}{8}. \tag{9}$$

Lemma 4 $\Pr\big(\mathcal{E}'\big) \;=\; 1 - \exp\big(-\Omega(\log^2(mn))\big)$.

The proof is given in Appendix D and relies on Hoeffding's inequality and the lower bound on the number of received signals for each grid point, driven in Lemma 3. Let $\mathcal{E}''$ be the event that for any machine i, for any $p \in G^t$ and θ in the domain, we have:

$$\Big|\big(\hat{F}^i(\theta) - \hat{F}^i(p)\big) - \big(F(\theta) - F(p)\big)\Big| < \frac{\epsilon}{8}. \tag{10}$$

Lemma 5 $\Pr\big(\mathcal{E}''\big) \;=\; 1 - \exp\big(-\Omega(\log^2(mn))\big)$.

In the remainder of the proof, we assume that $\mathcal{E}'$ and $\mathcal{E}''$ hold. Before proceeding, we need the following lemma.

Lemma 6 *Suppose that $\hat{G}$ be a uniform approximate of function G over a domain $\mathcal{W}$ for some $\epsilon > 0$, i.e., $|\hat{G}(w) - G(w)| < \epsilon$ for any $w \in \mathcal{W}$. Assume that w^* is the minimizer of $\hat{G}$ over $\mathcal{W}$. Then, we have: $G(w^*) \leq \inf_{w\in\mathcal{W}} G(w) + 2\epsilon$.*

Let $\hat{G}(\theta) = \hat{F}^i(\theta) - \hat{F}^i(p)$, $G(\theta) = F(\theta) - F(p)$, and $\mathcal{W}$ be the cell in G^t containing the point p. According to above lemma, we have:

$$G(\theta^i) \leq G(\theta^*_{\mathit{cell}}) + 2(\frac{\epsilon}{8}) \rightarrow F(\theta^i) \leq F(\theta^*_{\mathit{cell}}) + \frac{\epsilon}{4}, \tag{11}$$

where θ^*_{cell} is the minimizer of F in the cell containing the point p. Moreover, we know that

$$\begin{aligned}
|\hat{F}(\theta^i) - F(\theta^i)| &= |\hat{F}(p) + F^i(\theta^i) - F^i(p) - F(\theta^i)| \\
&= |\hat{F}(p) - F(p) + F(p) + F^i(\theta^i) - F^i(p) - F(\theta^i)| \\
&\leq |\hat{F}(p) - F(p)| + |(F^i(\theta^i) - F^i(p)) - (F(\theta^i) - F(p))| \\
&\leq \frac{\epsilon}{8} + \frac{\epsilon}{8} = \frac{\epsilon}{4},
\end{aligned} \tag{12}$$

where the last inequality is due to (9) and (10). Thus, from (11) and (12), we have with probability $1 - \exp(-\Omega(\log^2(mn)))$:

$$\hat{F}(\theta^i) \leq F(\theta^*_{\mathit{cell}}) + \frac{\epsilon}{2}. \tag{13}$$

Therefore, we can imply that $\min_i \hat{F}(\theta^i) \leq F(\theta^*) + \epsilon/2$ with probability $1 - \exp(-\Omega(\log^2(mn)))$. By the definition of $\hat{\theta} = \arg\min_i \hat{F}(\theta^i)$ and (12), we can conclude that $F(\hat{\theta}) \leq F(\theta^*) + 3\epsilon/4$ with probability $1 - \exp(-\Omega(\log^2(mn)))$ and Theorem 1 follows.

Appendix C. Proof of Lemma 3

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

$$1 - (1 - x)^k \;\geq\; 1 - e^{kx} \;\geq\; \frac{1}{2}\min\big(kx, 1\big). \tag{14}$$

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 $p \in G^l$,

$$\begin{aligned} Q_p &= 1 - \left(1 - 2^{-dl} \times \frac{2^{(d-2)l}}{\sum_{j=0}^{t} 2^{(d-2)j}}\right)^{\lfloor B/(d\log mn)\rfloor} \\ &\geq \frac{1}{2}\min\left(\frac{2^{-2l}\,\lfloor B/(d\log mn)\rfloor}{\sum_{j=0}^{t} 2^{(d-2)j}},\, 1\right) \\ &\geq \frac{1}{2}\min\left(\frac{2^{-2l}B}{d\log(mn)\,\sum_{j=0}^{t} 2^{(d-2)j}},\, 1\right) \end{aligned}$$

where the equality is due to the probability of a point p in G^l (see (4)) and the number $\lfloor B/(d\log mn)\rfloor$ of sub-signals per machine, and the first inequality is due to (14). We can imply that:

$$\mathbb{E}\big[N_p\big] \;=\; Q_p m \;\geq\; \min\left(\frac{2^{-2l}mB}{d\log(mn)\,\sum_{j=0}^{t} 2^{(d-2)j}},\, \frac{m}{2}\right). \tag{15}$$

For the first term at the right hand side of (15), we have for any $d \geq 2$,

$$\begin{aligned} \sum_{j=0}^{t} 2^{(d-2)j} &\leq t2^{t(d-2)} \\ &\leq \frac{1}{d}\log(mB)\,2^{t(d-2)} \\ &= \frac{1}{d}\log(mB)\,\left(\frac{1}{\delta}\right)^{d-2} \\ &= \frac{1}{d}\log(mB)\,\left(\frac{(mB)^{(d-2)/d}}{\log^{2(d-2)}(mn)}\right), \end{aligned} \tag{16}$$

where the second inequality is due to $t = \log(1/\delta) \leq \log(mB)/d$. The first and second equality is from definition of $t = \log(1/\delta)$ and $\delta = \log^2(mn)/(mB)^{1/d}$.

Plugging (16) and into (15), it follows that for $l = 0, \ldots, t$ and for any $p \in G^l$,

$$\begin{aligned} \mathbb{E}\big[N_p\big] &\geq \frac{2^{-2l}\,(mB)\,\log^{2(d-2)}(mn)}{\log(mn)\,\log(mB)\,(mB)^{(d-2)/d}} \\ &\geq \frac{2}{d}2^{-2l}\,\log^{2d-6}(mn)\,(mB)^{2/d} \\ &\geq \frac{2}{d}2^{-2t}\,\log^{2d-6}(mn)\,(mB)^{2/d} \\ &= \frac{2}{d}\delta^2\,\log^{2d-6}(mn)\,(mB)^{2/d} \\ &= \frac{2}{d}\log^{2d-2}(mn), \end{aligned} \tag{17}$$

where the first inequality is because of $(mB)^{1/d} < \sqrt{m}$ which implies that: $\log(mB) < d\log(mn)/2$ and the third equality is due to definition of $t = \log(1/\delta)$. Then, for any $l \in 0, \ldots, t$ and any $p \in G^l$,

$$
\begin{aligned}
\Pr\left(N_p \le \frac{1}{d}\log^2(mn)\right) \;&\le\; \Pr\left(N_p \le \frac{\mathbb{E}[N_p]}{2}\right) \\
&\le\; 2^{-(1/2)^2\mathbb{E}[N_p]/2} \\
&\le\; 2^{-\log^2(mn)/(4d)},
\end{aligned}
\tag{18}
$$

where the second inequality is due to Lemma 2 (b), and the last inequality is from (17) for $d \ge 2$. Then,

$$
\begin{aligned}
\Pr\left(\mathcal{E}\right) \;&=\; \Pr\left(N_p \ge \frac{1}{d}2^{-2l}\,\log^{2d-6}(mn)\,(mB)^{2/d}, \quad \forall p \in G^l \text{ and for } l = 0,\dots,t\right) \\
&\ge\; 1 - \sum_{l=0}^{t}\sum_{p\in G^l}\Pr\left(N_p \le \frac{1}{d}\log^2(mn)\right) \\
&\ge\; 1 - t2^{dt}\exp\big(-\log^2(mn)/(2d)\big) \\
&=\; 1 - \log(1/\delta)\left(\frac{1}{\delta}\right)^d\exp\big(-\log^2(mn)/(2d)\big) \\
&\ge\; 1 - \frac{\log(mB)}{d}\,\frac{mB}{\log^{2d}(mn)}\exp\big(-\log^2(mn)/(2d)\big) \\
&=\; 1 - \exp\Big(-\Omega\big(\log^2(mn)\big)\Big),
\end{aligned}
$$

where the first equality is by the definition of $\mathcal{E}$, the first inequality is from union bound and (17), the second inequality due to (18), and the third inequality follows from the definition of δ.

Appendix D. Proof of Lemma 4

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

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

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

We first consider the case of $l = 0$. Note that G^0 consists of a single point $p = s^*$. Moreover, the component Δ in each signal is the average over the gradient of $n/2$ independent functions. Then, $\hat{\Delta}(p)$ is the average over the gradient of $N_p \times n/2$ independent functions. Given event $\mathcal{E}$, it follows from Hoeffding's inequality (Lemma 2 (a)) that

$$\begin{aligned}
&\Pr\left(\left|\hat{F}(s^*) - F(s^*)\right| \geq \frac{\epsilon}{8\log(mn)}\right) \\
&\qquad \leq \exp\left(-N_{s^*}n \times \left(\frac{\epsilon}{8\log(mn)}\right)^2\right) \\
&\qquad \leq \exp\left(-n \times \frac{\log^{2d-6}(mn)(mB)^{2/d}}{d} \times \frac{\epsilon^2}{64\,\log^2(mn)}\right) \qquad (19) \\
&\qquad = \exp\left(\frac{-\log^2(mn)}{128d}\right) \\
&\qquad = \exp\Big(-\Omega\big(\log^2(mn)\big)\Big).
\end{aligned}$$

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, 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

$$\begin{aligned}
&\Pr\left(\left|\hat{\Delta}(p) - \Delta^*(p)\right| \geq \frac{\epsilon}{8\log(mn)}\right) \\
&\leq \exp\left(-nN_p \times \frac{1}{(2\sqrt{d}2^{-l})^2} \times \left(\frac{\epsilon}{8\log(mn)}\right)^2\right) \\
&\leq \exp\left(-n \times \frac{\log^{2d-6}(mn)(mB)^{2/d}2^{-2l}}{d} \times \frac{1}{4d2^{-2l}} \times \frac{\epsilon^2}{64\log^2(mn)}\right) \\
&= \exp\big(-\log^2(mn)/(128d)\big) \\
&= \exp\Big(-\Omega\big(\log^2(mn)\big)\Big),
\end{aligned}$$

Recall from (6) that for any non-zero $l \leq t$ and any $p \in \tilde{G}^l_{s^*}$ with parent p',

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

Then,

$$\begin{aligned}
&\Pr\left(\left\|\hat{F}(p) - F(p)\right\| > \frac{(l+1)\epsilon}{8\log(mn)}\right) \\
&\leq \Pr\left(\left\|\hat{F}(p') - F(p')\right\| > \frac{l\epsilon}{8\log(mn)}\right) \\
&\qquad + \Pr\left(\left\|\hat{\Delta}(p) - \Delta^*(p)\right\| > \frac{\epsilon}{8\log(mn)}\right) \\
&\leq \Pr\left(\left\|\hat{F}(p') - F(p')\right\| > \frac{l\epsilon}{8\log(mn)}\right) + \exp\Big(-\Omega\big(\log^2(mn)\big)\Big).
\end{aligned}$$

Employing an induction on l, we obtain for any $l \le t$,

$$\Pr\left(\big\|\hat{F}(p) - F(p)\big\| > \frac{(l+1)\epsilon}{8\log(mn)}\right) \le \exp\Big(-\Omega\big(\log^2(mn)\big)\Big).$$

Therefore, for any grid point p,

$$\begin{aligned}\Pr\left(\big\|\hat{F}(p) - F(p)\big\| > \frac{\epsilon}{8}\right) &\le \Pr\left(\big\|\hat{F}(p) - F(p)\big\| > \frac{(t+1)\epsilon}{8\log(mn)}\right)\\ &= \exp\Big(-\Omega\big(\log^2(mn)\big)\Big),\end{aligned}$$

where the inequality is because $t+1 = \log(1/\delta) + 1 \le 1/d\log(mB) \le \log(m) \le \log(mn)$. It then follows from the union bound that

$$\begin{aligned}\Pr\big(\mathcal{E}' \mid \mathcal{E}\big) &\ge 1 - \sum_{l=0}^{t} \sum_{p \in \tilde{G}^l_{s^*}} \Pr\left(\big\|\hat{F}(p) - F(p)\big\| > \frac{\epsilon}{8}\right)\\ &\ge 1 - t2^{dt}\exp\Big(-\Omega\big(\log^2(mn)\big)\Big)\\ &= 1 - \log(1/\delta)\left(\frac{1}{\delta}\right)^d \exp\Big(-\Omega\big(\log^2(mn)\big)\Big)\\ &\ge 1 - \frac{\log(mB)}{d}\,\frac{mB}{\log^{2d}(mn)}\exp\Big(-\Omega\big(\log^2(mn)\big)\Big)\\ &= 1 - \exp\Big(-\Omega\big(\log^2(mn)\big)\Big).\end{aligned} \tag{20}$$

On the other hand, we have from Lemma 3 that $\Pr\big(\mathcal{E}\big) = 1 - \exp\big(-\Omega(\log^2(mn))\big)$. Then, $\Pr\big(\mathcal{E}'\big) = 1 - \exp\big(-\Omega(\log^2(mn))\big)$ and Lemma 4 follows.