First-order Newton-type Estimator for Distributed Estimation and Inference

Xi Chen[*] Weidong Liu[†] Yichen Zhang[‡]

Abstract

This paper studies distributed estimation and inference for a general statistical problem with a convex loss that could be non-differentiable. For the purpose of efficient computation, we restrict ourselves to stochastic first-order optimization, which enjoys low per-iteration complexity. To motivate the proposed method, we first investigate the theoretical properties of a straightforward Divide-and-Conquer Stochastic Gradient Descent (DC-SGD) approach. Our theory shows that there is a restriction on the number of machines and this restriction becomes more stringent when the dimension p is large. To overcome this limitation, this paper proposes a new multi-round distributed estimation procedure that approximates the Newton step only using stochastic subgradient. The key component in our method is the proposal of a computationally efficient estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$, where $\boldsymbol{\Sigma}$ is the population Hessian matrix and $\boldsymbol{w}$ is any given vector. Instead of estimating $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Sigma}^{-1}$) that usually requires the second-order differentiability of the loss, the proposed First-Order Newton-type Estimator (FONE) directly estimates the vector of interest $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ as a whole and is applicable to non-differentiable losses. Our estimator also facilitates the inference for the empirical risk minimizer. It turns out that the key term in the limiting covariance has the form of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$, which can be estimated by FONE.

1 Introduction

The development of modern technology has enabled data collection of unprecedented size, which poses new challenges to many statistical estimation and inference problems. For example, given N samples, a classical estimation approach usually formulates a maximum likelihood estimation (MLE) problem and then solves the MLE by a deterministic optimization method (e.g., gradient descent or Newton method). However, when the sample size N is excessively large, there are two

[*]Stern School of Business, New York University, Email: xichen3@stern.nyu.edu. Research supported by Adobe research award, Alibaba innovation research award, and Bloomberg data science research grant.

[†]Institute of Natural Sciences and Department of Mathematics, Shanghai Jiao Tong University, Email: weidongl@sjtu.edu.cn. Research supported by NSFC, Grant No. 11825104, 11431006 and 11690013, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Youth Talent Support Program, 973 Program (2015CB856004) and a grant from Australian Research Council.

[‡]Stern School of Business, New York University, Email: yzhang@stern.nyu.edu

major obstacles when adopting this approach. First, a standard machine does not have enough memory to load the entire dataset all at once. Second, a deterministic optimization approach is computationally expensive. To address the storage and computation issues, distributed computing methods, originated from computer science literature, has been recently introduced into statistics. A general distributed computing scheme partitions the entire dataset into L parts, and then loads each part into the memory to compute a local estimator. The final estimator will be obtained via some communication and aggregation among local estimators.

Second, to further accelerate the computation, we consider stochastic first-order methods (e.g., stochastic gradient/subgradient descent (SGD)), which have been widely adopted in practice. There are a few significant advantages of SGD. First, as a first-order method, it only requires the subgradient information. As compared to second-order Newton-type approaches, it is not only computationally efficient and more scalable but also has a wider range of applications to problems where the empirical Hessian matrix does not exist. Second, a stochastic approach is much more efficient than its deterministic counterpart. For example, in a typical regression problem with p-dimensional predictors, the mini-batch SGD enjoys a low per-iteration time complexity of $O(mp)$, where m is the mini-batch size. In contrast, the deterministic gradient descent, which evaluates the gradient on the entire dataset with n samples at each iteration, has a per-iteration complexity of $O(np)$, where n is much larger than m. Although SGD has been widely studied in machine learning and optimization (see Section 2), using SGD for the purpose of statistical inference has not been sufficiently explored.

This paper studies a general statistical estimation and inference problem under the distributed computing setup. As we mentioned, to achieve an efficient computation, we restrict ourselves to the use of only stochastic subgradient information. In particular, consider a general statistical estimation problem in the following risk minimization form,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min}\, F(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi}\sim\Pi} f(\boldsymbol{\theta}, \boldsymbol{\xi}), \tag{1}$$

where $f(\cdot, \boldsymbol{\xi}) : \mathbb{R}^p \to \mathbb{R}$ is a convex loss function that can be non-differentiable (e.g., in quantile regression), and $\boldsymbol{\xi}$ denotes the random sample from a probability distribution Π (e.g., $\boldsymbol{\xi} = (Y, \boldsymbol{X})$ in a regression setup). Our goal is to estimate $\boldsymbol{\theta}^* \in \mathbb{R}^p$ under the *diverging dimension case*, where the dimensionality p is allowed to go to infinity as the sample size grows (but p grows at a slower rate than the sample size). This regime is more challenging than the fixed p case. On the other hand, since this work does not make any sparsity assumption, the high dimensional setting where p could be potentially larger than the sample size is beyond our scope. For the ease of illustration, we will use two motivating examples throughout the paper: (1) logistic regression with $f(\boldsymbol{\theta}, \boldsymbol{\xi}) = \log(1 + \exp(-Y\boldsymbol{X}'\boldsymbol{\theta}))$ (differentiable loss), and (2) quantile regression with $f(\boldsymbol{\theta}, \boldsymbol{\xi}) = (Y - \boldsymbol{X}'\boldsymbol{\theta})(\tau - I\{Y \le \boldsymbol{X}'\boldsymbol{\theta}\})$ (non-differentiable loss), where τ is the quantile level and $I\{\cdot\}$ is the indicator function.

Given n *i.i.d.* samples $\{\boldsymbol{\xi}_i\}_{i=1}^n$, a traditional non-distributed approach for estimating $\boldsymbol{\theta}^*$ is to minimize the empirical risk via a deterministic optimization:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}, \boldsymbol{\xi}_i). \tag{2}$$

Moreover, let $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ be the gradient (when $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ is differentiable) or a subgradient (when $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ is non-differentiable) of $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ at $\boldsymbol{\theta}$. For many popular statistical models, the empirical risk minimizer (ERM) $\widehat{\boldsymbol{\theta}}$ has an asymptotic normal distribution. That is, under some regularity conditions, for a fixed unit length vector $\boldsymbol{w} \in \mathbb{R}^p$, as $n, p \to \infty$,

$$\frac{\sqrt{n}\boldsymbol{w}'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}{\sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}} \to \mathcal{N}(0, 1), \tag{3}$$

where

$$\boldsymbol{\Sigma} := \nabla_{\boldsymbol{\theta}}\mathbb{E}g(\boldsymbol{\theta}, \boldsymbol{\xi})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \quad \boldsymbol{A} = \mathsf{Cov}(g(\boldsymbol{\theta}^*, \boldsymbol{\xi})) = \mathbb{E}\left[g(\boldsymbol{\theta}^*, \boldsymbol{\xi})g(\boldsymbol{\theta}^*, \boldsymbol{\xi})'\right]. \tag{4}$$

Under this framework, the main goal of our paper is twofold:

1. **Distributed estimation**: Develop a distributed stochastic first-order method for estimating $\boldsymbol{\theta}^*$ in the case of diverging p, with the aim to achieve the best possible convergence rate (i.e., the rate of the pooled ERM estimator $\widehat{\boldsymbol{\theta}}$). The method should be applicable to non-differentiable loss $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ and only requires the local strong convexity of $F(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ (instead of the strong convexity of $F(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$).

2. **Distributed inference**: Develop a consistent estimator of the limiting variance $\boldsymbol{w}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ to facilitate the inference. We note that the term $\boldsymbol{A}$ can be easily estimated via replacing the expectation by its sample version. However, it is challenging to estimate $\boldsymbol{\Sigma}$ when f is non-differentiable (and thus g will be discontinuous and the empirical Hessian matrix will not exist). To this end, instead of estimating $\boldsymbol{\Sigma}$, we aim to develop a stochastic first-order based approach that directly estimates $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ for any fixed given $\boldsymbol{w}$.

Let us first focus on the distributed estimation problem. We will first investigate the theoretical proprieties of a straightforward method that combines the stochastic subgradient descent (SGD) and divide-and-conquer (DC) scheme and discuss the theoretical limitation of this method. To overcome the theoretical limitation, we propose a new method called the distributed First-Order Newton-type Estimator (FONE), where the key idea is to approximate the Newton step only using stochastic subgradient information in a distributed setting.

In a distributed setting, the divide-and-conquer (DC) strategy has been recently adopted in many statistical estimation problems (Li et al., 2013; Chen and Xie, 2014; Battey et al., 2018; Zhao et al., 2016; Shi et al., 2018; Banerjee et al., 2018; Volgushev et al., 2018). A standard DC approach estimates a local estimator for each local machine[1] and then aggregates the local estimators to

[1] In a common single machine setup with excessively large data, each "local machine" corresponds to one partition of the data that can fit into the memory.

obtain the final estimator. Combining the idea of DC with the mini-batch SGD naturally leads to a divide-and-conquer SGD (DC-SGD) approach, where we run SGD on each local machine and then aggregate the obtained solutions by an averaging operation. The DC-SGD enjoys a very low communication cost: the communication is one-round with an $O(p)$ vector transmitted from each local machine. Despite the simplicity and wide applicability of the DC-SGD, the theoretical investigation of the asymptotic properties of this approach, especially in the diverging p case, is still quite limited. In fact, our theoretical analysis reveals several interesting phenomena of the mini-batch and DC-SGD when p is diverging, which also leads to useful practical guidelines when implementing DC-SGD. First, a natural starting point in a standard mini-batch SGD is random initialization. However, we show that when p diverges to infinity, a random initialized SGD will no longer converge to $\boldsymbol{\theta}^*$, with the L_2-estimation error being a polynomial of p (see Proposition 4.2). To address the challenge arising from $p \to \infty$, a consistent initial estimator $\widehat{\boldsymbol{\theta}}_0$ is both sufficient and necessary to ensure the convergence of SGD (see Theorem 4.1 and Proposition 4.2). Given a consistent initialization (which can be easily constructed running a deterministic optimization on a small batch of data), we can establish the estimation error rate of the DC-SGD in a distributed environment (see Theorem 4.3). For this DC-SGD to achieve the optimal convergence rate, the number of machines L has to be $O(\sqrt{N/p})$ (see Section 4.1.2), where N is the total number of samples across L machines. The condition could be restrictive when the size of the entire dataset is excessively large as compared to the limited memory size or when the number of machines is large but each local machine has a limited storage (e.g., in a large-scale sensor network). Moreover, as compared to the standard condition $L = O(\sqrt{N})$ in a fixed p setting, the condition $L = O(\sqrt{N/p})$ becomes more stringent when p diverges.

To relax this condition on L and further improve the performance of DC-SGD, this paper proposes a new approach called distributed first-order Newton-type estimator, which successively refines the estimator by multi-round aggregations. The starting point of our method is the "one-step estimator", which is an effective approach to improve the statistical efficiency of a consistent initial estimator $\widehat{\boldsymbol{\theta}}_0$. In particular, the "one-step estimator" essentially performs the following Newton-type step based on $\widehat{\boldsymbol{\theta}}_0$:

$$\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\Sigma}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i) \right), \tag{5}$$

where $\boldsymbol{\Sigma}$ is the population Hessian matrix $\boldsymbol{\Sigma}$ and $\left(\frac{1}{n}\sum_{i=1}^{n} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)\right)$ is the subgradient vector. As we mentioned before, the estimation of $\boldsymbol{\Sigma}$ is not easy when f is non-differentiable and the empirical Hessian matrix does not exist. To address this issue, our key idea is that instead of estimating $\boldsymbol{\Sigma}$ and computing its inverse, we propose an estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w} \in \mathbb{R}^p$ for any given vector $\boldsymbol{w} \in \mathbb{R}^p$, which solves (5) as a special case (with $\boldsymbol{w} = \frac{1}{n}\sum_{i=1}^{n} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)$). In fact, the estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ kills two birds with one stone: it not only constructs a Newton-type estimator of $\boldsymbol{\theta}^*$ but also provides an estimator for the asymptotic variance in (3), which facilitates the inference. In particular, the proposed FONE estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ is an iterative procedure that only utilizes the mini-batches of subgradient to approximate the Newton step.

Based on FONE, we further develop a multi-round distributed version of FONE which successively refines the estimator and does not impose any strict condition on the number of machines L. Theoretically, we show that for a smooth loss, when the number of rounds K exceeds a constant threshold K_0, the obtained distributed FONE $\widehat{\boldsymbol{\theta}}_K$ achieves the optimal convergence rate. For a non-smooth loss, such as quantile regression, our convergence rate only depends on the sample size of one local machine with the largest sub-sample size. This condition is weaker than the case of DC-SGD since the bottleneck in the convergence of DC-SGD is the local machine with the smallest sub-sample size. Therefore, one can improve the performance of distributed FONE for non-smooth losses by gathering more samples on a specific local machine. This is not hard to implement in practice since it is easier to equip only one local machine with more memory and computation power.

In summary, this paper studies the distributed estimation and inference based on stochastic subgradient information in the case of diverging p. To achieve this goal, we start from a simple DC-SGD and then propose our distributed FONE approach. Along the development of FONE, we identify a key problem of estimating $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ and propose a computationally efficient estimator. We summarize our main contributions as follows:

1. We establish the theoretical properties of the DC-SGD in the case of diverging p. In particular, we first show that a consistent initial estimator is almost necessary to guarantee the consistency of the obtained solution from a standard mini-batch SGD (see Proposition 4.2). This is essentially different from the case that p is fixed. Then, we establish the convergence rate of DC-SGD and characterize the restriction on the number of machines (see Theorem 4.3).

2. We develop a general First-Order Newton-type Estimator (FONE) for $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$, which is computationally efficient since it only utilizes the first-order information and is applicable to non-differentiable and/or non-strongly-convex losses (see Algorithm 2). We further extend FONE to the distributed setting (see Algorithm 3).

3. We provide the theoretical properties of the FONE for distributed estimation and inference problems. In particular, we establish the convergence rates of the distributed FONE for both smooth and non-smooth losses (see Theorems 4.5 and 4.7). Second, we prove that the FONE provides a consistent estimator of the limiting variance for the purpose of inference (see Theorems 4.8 and 4.9).

1.1 Notations and organization of the paper

The remainder of the paper is organized as follows: In Section 2, we review the related literature on recent works on distributed estimation and stochastic optimization. Section 3.1 describes the mini-batch SGD algorithm with diverging dimension and the DC-SGD estimator. We further propose FONE and distributed FONE in Section 3.2. Section 4 presents the theoretical results. In Section 5, we demonstrate the performance of the proposed estimators by simulation experiments, followed by conclusions in Section 6. The proofs are provided in Appendix.

In this paper, we will heavily use the asymptotic notations $O(\cdot)$ and $o(\cdot)$. Roughly speaking, $f(n) = O(g(n))$ means that f is bounded above by g (up to constant factor) asymptotically; and $f(n) = o(g(n))$ means that $f(n)/g(n)$ converges to zero and n goes to infinity. For a set of random variables X_n and a corresponding set of positive numbers a_n, $X_n = O_p(a_n)$ means that X_n/a_n is stochastically bounded and $X_n = o_p(a_n)$ means that X_n/a_n converges to zero in probability as n goes to infinity. Finally, denote the Euclidean norm for a vector $\boldsymbol{x} \in \mathbb{R}^p$ by $\|\boldsymbol{x}\|_2$, and denote the spectral norm for a matrix $\boldsymbol{X}$ by $\|\boldsymbol{X}\|$. For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrsim b_n$ if $a_n \geq cb_n$ holds for all n and some absolute constant $c > 0$, $a_n \lesssim b_n$ if $b_n \gtrsim a_n$ holds, and $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. We will use $c, c_0, c_1, \ldots$ and $C, C_0, C_1, \ldots$ to denote constants, whose values can change from place to place.

In addition, since the distributed estimation and inference usually involves quite a few notations, we briefly summarize them here. We use N, L, $n = N/L$, and m to denote the total number of samples, the number of machines (or the number of data partitions), the sample size on each local machine (when evenly distributed), and the batch size for mini-batch SGD, respectively. When we discuss a problem in the classical single machine setting, we will also use n to denote the sample size. We will use $\boldsymbol{\theta}^*$, $\widehat{\boldsymbol{\theta}}$, and $\widehat{\boldsymbol{\theta}}_0$ to denote the minimizer of the popular risk, the ERM, and the initial estimator, respectively. The random sample will be denoted by $\boldsymbol{\xi}$ and in a regression setting $\boldsymbol{\xi} = (Y, \boldsymbol{X})$.

2 Related Works

Our work is closely related to two lines of research—distributed estimation and stochastic optimization. We will review each topic in this section.

In recent years, the divide-and-conquer (DC) approach has been widely applied to statistical estimation problems. Examples include density parameter estimation (Li et al., 2013), generalized linear regression with non-convex penalties (Chen and Xie, 2014), kernel ridge regression (Zhang et al., 2015), high-dimensional sparse linear regression (Lee et al., 2017), high-dimensional generalized linear models (Battey et al., 2018), semi-parametric partial linear models (Zhao et al., 2016), quantile regression processes (Volgushev et al., 2018; Chen et al., 2019), principle component analysis (Fan et al., 2018), one-step estimator (Huang and Huo, 2015), M-estimators with cubic rate

(Shi et al., 2018), and some non-standard problems where rates of convergence are slower than $n^{1/2}$ and limit distributions are non-Gaussian (Banerjee et al., 2018). The DC approach enjoys a low communication cost since it only requires *one-round aggregation*, e.g., averaging local estimators to obtain the global estimator. This is also the idea behind our DC-SGD approach. However, since the averaging only reduces the variance but not the bias term, all these types of results involve a constraint on the number of machines, which aims to make the variance the dominating term.

Jordan et al. (2018) recently proposed a multi-round distributed estimation method by approximating the Newton step in an iterative aggregation scheme. A similar approach was also independently developed by Wang et al. (2017). In particular, the key idea behind the method of Jordan et al. (2018) is that instead of computing the Hessian matrix on the entire dataset, one can approximate the Newton step by using the local Hessian matrix computed on a single machine (see Algorithm 1 in Jordan et al. (2018)). On the other hand, to compute the local Hessian matrix, their method requires the second-order differentiability on the loss function and thus is not applicable to problems such as quantile regression. In contrast, our approach approximates the Newton step via stochastic subgradient and thus can handle the non-differentiability in the loss function. In sum, the methods in Jordan et al. (2018) and Wang et al. (2017) still belong to second-order approaches while our method only utilizes stochastic first-order information.

The second field of related research is stochastic first-order optimization. One of the most popular stochastic optimization methods is stochastic gradient descent (SGD), which dates back to Robbins and Monro (1951). Due to its wide applicability in machine learning, there is a large body of literature on SGD (see, e.g., Zhang (2004); Nesterov and Vial (2008); Bach and Moulines (2011); Dekel et al. (2012); Ghadimi and Lan (2012); Xiao and Zhang (2014); Toulis et al. (2017); Liang and Su (2017); Su and Zhu (2018) and references therein). Here, we will not be able to provide a detailed survey on SGD but only highlight several key differences between our work and the existing literature on SGD. First, many existing works either assume differentiability or *uniform* strong convexity of loss functions. Instead, we do not require any of these assumptions. We only assume $F(\boldsymbol{\theta}) = \mathbb{E}f(\boldsymbol{\theta}, \boldsymbol{\xi})$ (i.e., the objective function in (1)) is strongly convex around the population risk minimizer $\boldsymbol{\theta}^*$. Second, in the majority of the optimization literature on SGD, the goal is to establish the convergence rate of the expected objective value, i.e., $\mathbb{E}(F(\widetilde{\boldsymbol{\theta}}) - F(\boldsymbol{\theta}^*))$, where $\widetilde{\boldsymbol{\theta}}$ is the solution of SGD. In contrast to the convergence from an optimization perspective, we focus on the statistical estimation error $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$, where $\widetilde{\boldsymbol{\theta}}$ is the solution of the DC-SGD in the diverging p case. Moreover, our goal is to derive the estimation error and quantify both bias and variance with an *explicit dependence* on the dimension p, the mini-batch size m, and the number of machines L. The explicit dependence on these parameters cannot be easily found in the existing literature.

In addition, our FONE of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ is related to a recently developed stochastic first-order approach—stochastic variance reduced gradient (SVRG, see e.g., Johnson and Zhang (2013); Lee et al. (2017); Wang and Zhang (2017) and references therein). Our method subsumes SVRG as a special case.

Indeed, when the $\boldsymbol{w} = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\theta}, \boldsymbol{\xi}_i)$, our iterative algorithm (non-distributed version) essentially reduces to SVRG. On the other hand, we allow a general $\boldsymbol{w}$ vector, which does not need to be an averaged gradient (e.g., for the purpose of inference in (3)). Moreover, the theoretical development of SVRG requires the unbiasedness of the stochastic gradient with respect to the averaged gradient $\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\theta}, \boldsymbol{\xi}_i)$, the differentiability, and uniform strong convexity of the loss function f. In contrast, our theoretical results do not require any of these conditions. In fact, the motivation for our procedure is fundamentally different from that for SVRG: our method is to provide an estimator $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ with the population matrix $\boldsymbol{\Sigma}$ for any $\boldsymbol{w}$; while most SVRG literature aims to solve a finite-sum optimization problem $\min \frac{1}{n}\sum\limits_{i=1}^{n} f(\boldsymbol{\theta}, \boldsymbol{\xi}_i)$ for a differentiable strongly-convex f.

Due to space limitations, we are not able to provide a comprehensive survey on SVRG in this paper; instead, we briefly mention two papers that we consider most relevant. Lee et al. (2017) recently developed a distributed SVRG method. However, to ensure the unbiasedness of the stochastic gradient, it has a complicated data reallocation procedure across different machines (see Algorithm 1 in Lee et al. (2017)), which is not required in our procedure. Our distributed FONE computes the mini-batch stochastic subgradient using samples from only one local machine, which also leads to a lower communication cost. Lee et al. (2017) also requires the loss function to be differentiable and strongly-convex. The other work is a recent paper by Li et al. (2018), which adopts SVRG for the inference based on ERM. The main part of the paper considers a fixed p setup, under which the following limiting distribution result holds: $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to \mathcal{N}(0, \boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1})$. Under the assumption of the second-order differentiability of the loss function f, the method in Li et al. (2018) estimates $\boldsymbol{\Sigma}$ by the empirical Hessian matrix $\widehat{\boldsymbol{\Sigma}}$. However, the estimation of $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\Sigma}}$ could suffer from a slow convergence when p is large. Moreover, Li et al. (2018) has not studied the distributed estimation of $\boldsymbol{\theta}^*$. Our idea is different from Li et al. (2018): in the diverging p case, we avoid estimating $\boldsymbol{\Sigma}$ but directly construct an estimator of the vector $\boldsymbol{\Sigma}^{-1}\boldsymbol{w} \in \mathbb{R}^p$ as a whole.

3 Methodology

In this section, we introduce the DC-SGD algorithm and then describe the proposed FONE and its distributed version.

3.1 Divide-and-conquer SGD (DC-SGD) algorithm

Before we introduce our DC-SGD algorithm, we first present the mini-batch SGD algorithm for solving the stochastic optimization in (1) on a single machine with total n samples. In particular, we consider the setting when the dimension $p \to \infty$ but at a slower rate than n, i.e., $p \leq n^{\kappa}$ for some $\kappa \in (0, 1)$. Given n *i.i.d.* samples $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$, we partition the index set $\{1, \ldots, n\}$ into s disjoint mini-batches $H_1, ..., H_s$, where each mini-batch has the size $|H_i| = m$ (for $i = 1, 2, \ldots, s$), and $s = n/m$ is the number of mini-batches. The mini-batch SGD algorithm starts from a consistent initial estimator $\widehat{\boldsymbol{\theta}}_0$ of $\boldsymbol{\theta}^*$. Let $\boldsymbol{z}_0 = \widehat{\boldsymbol{\theta}}_0$. The mini-batch SGD iteratively updates $\boldsymbol{z}_i$ from $\boldsymbol{z}_{i-1}$ as follows and outputs $\widehat{\boldsymbol{\theta}}_{\text{SGD}} = \boldsymbol{z}_s$ as its final estimator,

$$\boldsymbol{z}_i = \boldsymbol{z}_{i-1} - \frac{r_i}{m} \sum_{j \in H_i} g(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j), \quad \text{for} \quad i = 1, 2, \ldots, s, \tag{6}$$

where we set the step-size $r_i = c_0 / \max(i^{\alpha}, p)$ for some $0 < \alpha \leq 1$ and c_0 is a positive constant. It is worthwhile that a typical choice of r_i in the literature is $r_i = c_0 \cdot i^{-\alpha}$ (Polyak and Juditsky, 1992). Since we are considering a diverging p case, our step-size incorporates the dimension p. As one can see, this mini-batch SGD algorithm only uses one pass of the data and enjoys a low per-iteration complexity.

The bias and L_2-estimation error of the mini-batch SGD will be provided in Theorem 4.1 (see Section 4.1). We provide two examples on logistic regression and quantile regression to illustrate the subgradient function $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ in our mini-batch SGD and will refer to these examples throughout the paper.

Example 3.1 (Logistic regression)**.** *Consider a logistic regression model with the response* $Y \in \{-1, 1\}$*, where*

$$\mathbb{P}(Y = 1 | \boldsymbol{X}) = 1 - \mathbb{P}(Y = -1 | \boldsymbol{X}) = \frac{1}{1 + \exp(-\boldsymbol{X}'\boldsymbol{\theta}^*)},$$

and $\boldsymbol{\theta}^* \in \mathbb{R}^p$ *is the true model parameter. Define* $\boldsymbol{\xi} = (Y, \boldsymbol{X})$*. We have the smooth loss function* $f(\boldsymbol{\theta}, \boldsymbol{\xi}) = \log(1 + \exp(-Y\boldsymbol{X}'\boldsymbol{\theta}))$ *and its gradient* $g(\boldsymbol{\theta}, \boldsymbol{\xi}) = -Y\boldsymbol{X}\big(1 + \exp(Y\boldsymbol{X}'\boldsymbol{\theta})\big)^{-1}$.

Example 3.2 (Quantile regression)**.** *Consider a quantile regression model* $Y = \boldsymbol{X}'\boldsymbol{\theta}^* + \epsilon$*, where we assume that* $\boldsymbol{X} = (1, X_1, ..., X_{p-1})'$ *and* $\mathbb{P}(\epsilon \leq 0 | \boldsymbol{X}) = \tau$ *is the so-called quantile level. Define* $\boldsymbol{\xi} = (Y, \boldsymbol{X})$*. We have the non-smooth quantile loss function* $f(\boldsymbol{\theta}, \boldsymbol{\xi}) = \ell_\tau(Y - \boldsymbol{X}'\boldsymbol{\theta})$ *and* $\ell_\tau(x) = x(\tau - I\{x \leq 0\})$*. A subgradient of the quantile loss is given by* $g(\boldsymbol{\theta}, \boldsymbol{\xi}) = \boldsymbol{X}(I\{Y \leq \boldsymbol{X}'\boldsymbol{\theta}\} - \tau)$*.*

Given the mini-batch SGD, we are ready to introduce the divide-and-conquer SGD (DC-SGD). For the ease of illustration, suppose that the entire sample with the size N is evenly distributed on L machines (or split into L parts) with the sub-sample size $n = N/L$ on each local machine. For the ease of presentation, we assume that N/L is a positive integer. On each machine $k = 1, 2, \ldots, L$, we run the mini-batch SGD with the batch size m in (6). Let $\mathcal{H}_k$ be the indices of the data points on the k-th machine, which is further split into s mini-batches $\{H_{k,i}, i = 1, 2, \ldots, s\}$ with $|H_{k,i}| = m$ and $s = n/m$. On the k-th machine, we run our mini-batch SGD in (6) and obtain the local estimator $\widehat{\boldsymbol{\theta}}^{(k)}_{\text{SGD}}$. The final estimator is aggregated by averaging the local estimators from L machines, i.e.,

$$\widehat{\boldsymbol{\theta}}_{\text{DC}} = \frac{1}{L} \sum_{k=1}^{L} \widehat{\boldsymbol{\theta}}^{(k)}_{\text{SGD}}. \tag{7}$$

Algorithm 1 DC-SGD algorithm

Input: The initial estimator $\widehat{\boldsymbol{\theta}}_0 \in \mathbb{R}^p$, the step-size sequence $r_i = c_0/\max(i^\alpha, p)$ for some $0 < \alpha \leq 1$, the mini-batch size m.

Distribute the initial estimator $\widehat{\boldsymbol{\theta}}_0$ to each local machine $k = 1, 2, \ldots, L$.
for each local machine $k = 1, 2, \ldots, L$ **do**
  Set the starting point $\boldsymbol{z}_0^{(k)} = \widehat{\boldsymbol{\theta}}_0$.
  for each iteration $i = 1, \ldots, s$ **do**
    Update

$$\boldsymbol{z}_i^{(k)} = \boldsymbol{z}_{i-1}^{(k)} - \frac{r_i}{m} \sum_{j \in H_{k,i}} g(\boldsymbol{z}_{i-1}^{(k)}, \boldsymbol{\xi}_j),$$

  end for
  Set $\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}^{(k)} = \boldsymbol{z}_s^{(k)}$ as the local SGD estimator on the machine k.
end for
Aggregate the local estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}^{(k)}$ by averaging and compute the final estimator:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}} = \frac{1}{L} \sum_{k=1}^{L} \widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}^{(k)}.$$

Output: $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$.

Note that the DC-SGD algorithm only involves one round of aggregation. The details of the DC-SGD are presented in Algorithm 1.

In Theorem 4.3, we establish the convergence rate of the DC-SGD in terms of the dimension p, the number of machines L, the total sample size N and the mini-batch size m. Moreover, we show that for the DC-SGD estimator to achieve the same rate as the mini-batch SGD running on the entire dataset, it requires a condition on the number of machines L.

3.2 First-Order Newton-type Estimator (FONE)

To relax the condition on the number of machines L, one idea is to perform a Newton-type step in (5). However, as we have pointed out, the estimation of $\boldsymbol{\Sigma}$ requires the second-order differentiability of the loss function. Moreover, a typical Newton method successively refines the estimator of $\boldsymbol{\Sigma}$ based on the current estimate of $\boldsymbol{\theta}^*$ and thus requires the computation of matrix inversion in (5) for multiple iterations, which could be computationally expensive when p is large.

In this section, we propose a new First-Order Newton-type Estimator (FONE) that directly estimates $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ (for any given vector $\boldsymbol{a}$) only using the stochastic first-order information. Then for

Algorithm 2 First-Order Newton-type Estimator (FONE) of $\mathbf{\Sigma}^{-1}\boldsymbol{a}$

Input: Dataset $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n\}$, the initial estimator $\widehat{\boldsymbol{\theta}}_0$, step-size η, the batch-size m, and a given vector $\boldsymbol{a} \in \mathbb{R}^p$.

1: Set $\boldsymbol{z}_0 = \widehat{\boldsymbol{\theta}}_0$.
2: **for** each $t = 1, 2, \ldots, T$ **do**
3: Choose B_t to be m distinct elements uniformly from $\{1, 2, ..., n\}$.
4: Calculate
$$g_{B_t}(\boldsymbol{z}_{t-1}) = \frac{1}{m}\sum_{i \in B_t} g(\boldsymbol{z}_{t-1}, \boldsymbol{\xi}_i), \qquad g_{B_t}(\boldsymbol{z}_0) = \frac{1}{m}\sum_{i \in B_t} g(\boldsymbol{z}_0, \boldsymbol{\xi}_i).$$
5: Update
$$\boldsymbol{z}_t = \boldsymbol{z}_{t-1} - \eta\{g_{B_t}(\boldsymbol{z}_{t-1}) - g_{B_t}(\boldsymbol{z}_0) + \boldsymbol{a}\}.$$
6: **end for**
7: **Output:**
$$\widehat{\boldsymbol{\theta}}_{\text{FONE}} = \widehat{\boldsymbol{\theta}}_0 - \boldsymbol{z}_T. \tag{8}$$

---

a given initial estimator $\widehat{\boldsymbol{\theta}}_0$, we can perform the Newton-type step in (5) as

$$\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_0 - \widehat{\mathbf{\Sigma}^{-1}\boldsymbol{a}}, \qquad \boldsymbol{a} = \left(\frac{1}{n}\sum_{i=1}^{n} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)\right), \tag{9}$$

where $\widehat{\mathbf{\Sigma}^{-1}\boldsymbol{a}}$ is our estimator of $\mathbf{\Sigma}^{-1}\boldsymbol{a}$.

To estimate $\mathbf{\Sigma}^{-1}\boldsymbol{a}$, we note that $\mathbf{\Sigma}^{-1}\boldsymbol{a} = \sum_{i=0}^{\infty}(1 - \eta\mathbf{\Sigma})^i\eta\boldsymbol{a}$, for some small enough η such that $\|\eta\mathbf{\Sigma}\| < 1$. Then we can use the following iterative procedure $\{\widetilde{\boldsymbol{z}}_t\}$ to approximate $\mathbf{\Sigma}^{-1}\boldsymbol{a}$:

$$\widetilde{\boldsymbol{z}}_t = \widetilde{\boldsymbol{z}}_{t-1} - \eta(\mathbf{\Sigma}\widetilde{\boldsymbol{z}}_{t-1} - \boldsymbol{a}), \ 1 \le t \le T, \tag{10}$$

where η here can be viewed as a constant step-size. To see that (10) leads to an approximation of $\mathbf{\Sigma}^{-1}\boldsymbol{a}$, when T is large enough, we have

$$\begin{aligned}
\widetilde{\boldsymbol{z}}_T &= \widetilde{\boldsymbol{z}}_{T-1} - \eta(\mathbf{\Sigma}\widetilde{\boldsymbol{z}}_{T-1} - \boldsymbol{a}) = (I - \eta\mathbf{\Sigma})\widetilde{\boldsymbol{z}}_{T-1} + \eta\mathbf{\Sigma}\boldsymbol{a} \\
&= (I - \eta\mathbf{\Sigma})^2\widetilde{\boldsymbol{z}}_{T-2} + (I - \eta\mathbf{\Sigma})\eta\boldsymbol{a} + \eta\boldsymbol{a} \\
&= (I - \eta\mathbf{\Sigma})^{T-1}\widetilde{\boldsymbol{z}}_1 + \sum_{i=0}^{T-2}(I - \eta\mathbf{\Sigma})^i\eta\boldsymbol{a} \approx \mathbf{\Sigma}^{-1}\boldsymbol{a}.
\end{aligned}$$

As the iterate $\widetilde{\boldsymbol{z}}_t$ approximates $\mathbf{\Sigma}^{-1}\boldsymbol{a}$, let us define $\boldsymbol{z}_t = \widehat{\boldsymbol{\theta}}_0 - \widetilde{\boldsymbol{z}}_t$, which is the quantity of interest (see the left-hand side of the Newton-type step in (9)). To avoid estimating $\mathbf{\Sigma}$ in the recursive update in (10), we adopt the following first-order approximation:

$$-\boldsymbol{\Sigma}\widetilde{\boldsymbol{z}}_{t-1} = \boldsymbol{\Sigma}(\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0) \approx g_{B_t}(\boldsymbol{z}_{t-1}) - g_{B_t}(\widehat{\boldsymbol{\theta}}_0), \tag{11}$$

where $g_{B_t}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i\in B_t} g(\boldsymbol{\theta}, \boldsymbol{\xi}_i)$ is the averaged stochastic subgradient over a subset of the data indexed by $B_t \subseteq \{1, 2, \ldots, n\}$. Here B_t is randomly chosen from $\{1, \ldots, n\}$ with replacement for every iteration.

Given (11), we construct our FONE of $\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ by the following recursive update from $t = 1, 2, \ldots, T$:

$$\boldsymbol{z}_t = \boldsymbol{z}_{t-1} - \eta\{g_{B_t}(\boldsymbol{z}_{t-1}) - g_{B_t}(\widehat{\boldsymbol{\theta}}_0) + \boldsymbol{a}\}, \quad \boldsymbol{z}_0 = \widehat{\boldsymbol{\theta}}_0. \tag{12}$$

The obtained $\boldsymbol{z}_T$, as an estimator of $\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ can be directly used in the Newton-type step in (9). The choices of the input parameters and the convergence rate of our FONE will be proved in Propositions 4.4 and 4.6. Also note that for constructing the estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$, we can simply use $\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{z}_T$ and the procedure is summarized in Algorithm 2.

Remark 3.3. *Our FONE of $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ in (12) is related to the stochastic variance reduced gradient (SVRG), see, e.g., Johnson and Zhang (2013); Lee et al. (2017); Wang and Zhang (2017) and references therein. Our method can be viewed as a more generalized version of SVRG. In particular, suppose that the loss function $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ is differentiable and let $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ be its gradient. The SVRG, which aims to solve the optimization problem $\min \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}, \boldsymbol{\xi}_i)$, iteratively updates:*

$$\boldsymbol{z}_t = \boldsymbol{z}_{t-1} - \eta_n\Big(g(\boldsymbol{z}_{t-1}, \boldsymbol{\xi}_{i_t}) - g(\boldsymbol{z}_0, \boldsymbol{\xi}_{i_t}) + \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_0, \boldsymbol{\xi}_i)\Big)$$

where i_t is randomly drawn from $\{1, 2, \ldots, n\}$. On the other hand, our FONE aims to provide a consistent estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$. As one can see, when $\boldsymbol{a} = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_0, \boldsymbol{\xi}_i)$, our FONE reduces to a mini-batch version of SVRG.

In addition, there are two major technical differences between our FONE and SVRG. First of all, existing theoretical development of SVRG heavily relies on the unbiasedness of the stochastic gradient $g(\boldsymbol{z}_0, \boldsymbol{\xi}_{i_t})$, i.e., requiring $\mathbb{E}_{i_t} g(\boldsymbol{z}_0, \boldsymbol{\xi}_{i_t}) = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_0, \boldsymbol{\xi}_i)$. However, the unbiasedness condition is not necessary in our method, which makes FONE directly applicable to distributed settings and also to an arbitrary vector $\boldsymbol{a}$. For example, by choosing $\boldsymbol{a}$ to be a unit length vector $\boldsymbol{w}$, Algorithm 2 can be used for estimating the limiting variance in (3) of the empirical risk minimizer (see Section 4.3 below for more details). Second, our FONE applies to non-smooth loss functions and thus it applies to Newton-type approximation in (9) for quantile regression.

3.3 Distributed FONE for estimating $\boldsymbol{\theta}^*$

Based on the FONE for $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$, we present a distributed FONE for estimating $\boldsymbol{\theta}^*$. Suppose the entire dataset with N samples is distributed on L local machines $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_L\}$ (not necessarily evenly distributed). Our distributed FONE is a multi-round approach with K rounds, where K is a pre-specified constant. For each round $j = 1, 2, \ldots, K$, with the initialization $\widehat{\boldsymbol{\theta}}_{j-1}$, we first calculate $\boldsymbol{a} = \frac{1}{N}\sum_{i=1}^{N} g(\widehat{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\xi}_i)$ by averaging the subgradients from each local machine. Then we apply FONE (Algorithm 2) with $\boldsymbol{a}$ on the local machine with the largest sub-sample size. Since FONE is performed on one local machine, this iterative procedure does not incur any extra communication cost. The detailed algorithm is given in Algorithm 3. In fact, the presented Algorithm 3 is essentially estimating $\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ with $\boldsymbol{a} = \frac{1}{N}\sum_{i=1}^{N} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)$ and $\widehat{\boldsymbol{\theta}}_0$ is a pre-given initial estimator.

It is worthwhile noting that in contrast to DC-SGD where each local machine plays the same role, distributed FONE performs the update in (14) only on one local machine. The convergence rate of distributed FONE will depend on the sub-sample size of this machine (see Theorems 4.5 and 4.7). Therefore, to achieve the best convergence rate, we perform the update in (14) on the machine with the largest sub-sample size and *index it by the first machine* without loss of generality.

4 Theoretical Results

In this section, we provide theoretical results for mini-batch SGD in the diverging p case, DC-SGD, the newly proposed FONE and its distributed version. We first note that in most cases, the minimizer $\boldsymbol{\theta}^*$ in (1) is also a solution of the following estimating equation:

$$\mathbb{E}g(\boldsymbol{\theta}^*, \boldsymbol{\xi}) = 0, \tag{15}$$

where $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ is the gradient or a subgradient of $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ at $\boldsymbol{\theta}$. We will assume that (15) holds throughout our paper. In fact, we can introduce (15) as our basic model (instead of (1)) as in the literature from stochastic approximation (see, e.g., Lai (2003)). However, we choose to present the minimization form in (1) as it is more commonly used in statistical learning literature.

Now let us first establish the theory for the DC-SGD approach in the diverging p case.

4.1 Theory for mini-batch SGD and DC-SGD

To establish the theory for DC-SGD, we first state our assumptions. The first assumption is on the relationship among the dimension p, the sample size n, and the mini-batch size m. Recall that α is the decaying rate in the step-size of SGD (see the input of Algorithm 1).

(C1). Suppose that $p \to \infty$ and $p = O(n^{\kappa_1})$ for some $0 < \kappa_1 < 1$. The mini-batch size m satisfies $p \log n = o(m)$ and $n^{\tau_1} \le m \le n/p^{1/\alpha + \tau_2}$ for some $0 < \tau_1, \tau_2 < 1$.

The remaining assumptions are on the continuity of the subgradient $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ and its expectation $G(\boldsymbol{\theta}) := \mathbb{E}g(\boldsymbol{\theta}, \boldsymbol{\xi})$.

(C2). Suppose that $G(\boldsymbol{\theta})$ is differentiable on $\boldsymbol{\theta}$ and denote by $\boldsymbol{\Sigma}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta})$. For some constant $C_1 > 0$, we have

$$\left\|\boldsymbol{\Sigma}(\boldsymbol{\theta}_1) - \boldsymbol{\Sigma}(\boldsymbol{\theta}_2)\right\| \le C_1 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \quad \text{for any } \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p. \tag{16}$$

Algorithm 3 Distributed FONE for Estimating $\boldsymbol{\theta}^*$ in (1)

Input: The total sample size N, the entire data $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N\}$ is distributed into L machines/parts $\{\mathcal{H}_k\}$ for $k = 1, 2 \ldots, L$ with $|\mathcal{H}_k| = n_k$. Initial estimator $\widehat{\boldsymbol{\theta}}_0 \in \mathbb{R}^p$, the batch size m, step-size η. Number of rounds K.

for each round $j = 1, 2, \ldots, K$ **do**
  for each local machine $k = 1, 2, \ldots, L$ **do**
    Calculate $\sum_{i \in \mathcal{H}_k} g(\widehat{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\xi}_i)$.
  end for
  Collect $\sum_{i \in \mathcal{H}_k} g(\widehat{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\xi}_i)$ from each local machine to compute their average:

$$\boldsymbol{a} = \frac{1}{N} \sum_{k=1}^{L} \sum_{i \in \mathcal{H}_k} g(\widehat{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\xi}_i) = \frac{1}{N} \sum_{i=1}^{N} g(\widehat{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\xi}_i). \tag{13}$$

  Send $\boldsymbol{a}$ to the first machine (the local machine with the largest sub-sample size).
  Set $\boldsymbol{z}_0 = \widehat{\boldsymbol{\theta}}_{j-1}$
  for each $t = 1, 2, \ldots, T$ **do**
    Choose B_t to be m distinct elements uniformly drawn from the data on the first machine $\mathcal{H}_1$.
    Calculate

$$g_{B_t}(\boldsymbol{z}_{t-1}) = \frac{1}{m} \sum_{i \in B_t} g(\boldsymbol{z}_{t-1}, \boldsymbol{\xi}_i), \qquad g_{B_t}(\boldsymbol{z}_0) = \frac{1}{m} \sum_{i \in B_t} g(\boldsymbol{z}_0, \boldsymbol{\xi}_i).$$

    Update

$$\boldsymbol{z}_t = \boldsymbol{z}_{t-1} - \eta\{g_{B_t}(\boldsymbol{z}_{t-1}) - g_{B_t}(\boldsymbol{z}_0) + \boldsymbol{a}\}. \tag{14}$$

  end for
  Set $\widehat{\boldsymbol{\theta}}_j = \boldsymbol{z}_T$.
end for
Output: $\widehat{\boldsymbol{\theta}}_K$.

Furthermore, let $\lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ and $\lambda_{\max}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ be the minimum and maximum eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, respectively. We assume that $c_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)) \leq \lambda_{\max}(\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)) \leq c_1^{-1}$ for some constant $c_1 > 0$.

It is worthwhile to note that $\boldsymbol{\Sigma}$ defined in (4) is a brief notation for $\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)$. The minimum eigenvalue condition on $\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)$ (i.e., $\lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)) \geq c_1$) ensures that the population risk $F(\boldsymbol{\theta})$ is locally strongly convex at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Throughout this paper, we define a loss function f to be smooth when f is continuously differentiable. We give two separate conditions for smooth and non-smooth loss functions, respectively.

(C3). (For smooth loss function f) For $\boldsymbol{v} \in \mathbb{R}^p$, assume that

$$\left|\boldsymbol{v}'[g(\boldsymbol{\theta}_1, \boldsymbol{\xi}) - g(\boldsymbol{\theta}_2, \boldsymbol{\xi})]\right| \le U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,$$

where $U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ satisfies that

$$\sup_{\|\boldsymbol{v}\|_2=1} \sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \mathbb{E} \exp\left(t_0 U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right) \le C, \qquad \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \le p^{c_2},$$

for some $c_2, t_0, C > 0$. Moreover, *one of the following two conditions* on $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ holds,

1. $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \sup_{\boldsymbol{\theta}} \exp(t_0|\boldsymbol{v}' g(\boldsymbol{\theta}, \boldsymbol{\xi})|) \le C$ for some $t_0, C > 0$;
2. $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \exp(t_0|\boldsymbol{v}' g(\boldsymbol{\theta}^*, \boldsymbol{\xi})|) \le C$ and $c_1 \le \lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta})) \le \lambda_{\max}(\boldsymbol{\Sigma}(\boldsymbol{\theta})) \le c_1^{-1}$ uniformly in $\boldsymbol{\theta}$ for some $t_0, c_1, C > 0$.

Next, we consider the setting that $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ is non-differentiable (e.g., quantile regression) such that $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ is its subgradient. In this case, the subgradient $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ may be discontinuous, which violates Condition (C3). To this end, we propose an alternative condition (C3*) as follows.

(C3*). (For non-smooth loss function f) Suppose that for some constant $c_2, c_3, c_4 > 0$,

$$\sup_{\boldsymbol{\theta}_1 : \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_2 \le c_4} \mathbb{E}\Big\{ \sup_{\boldsymbol{\theta}_2 : \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \le n^{-M},\ \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*\|_2 \le c_4} \|g(\boldsymbol{\theta}_1, \boldsymbol{\xi}) - g(\boldsymbol{\theta}_2, \boldsymbol{\xi})\|_2^4 \Big\} \le p^{c_2} n^{-c_3 M}$$

for any large $M > 0$. Also

$$\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}(\boldsymbol{v}'(g(\boldsymbol{\theta}_1, \boldsymbol{\xi}) - g(\boldsymbol{\theta}_2, \boldsymbol{\xi}))^2 \exp\{t_0|\boldsymbol{v}'(g(\boldsymbol{\theta}_1, \boldsymbol{\xi}) - g(\boldsymbol{\theta}_2, \boldsymbol{\xi}))|\} \le C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$$

and $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \sup_{\boldsymbol{\theta}} \exp(t_0|\boldsymbol{v}' g(\boldsymbol{\theta}, \boldsymbol{\xi})|) \le C$ for some t_0, $C > 0$.

Conditions (C2), (C3) and (C3*) can be easily verified in our two motivating examples of logistic regression and quantile regression (see Appendix D). In Condition (C3), we only require either one of the two bullets of the conditions on $g(\boldsymbol{\theta}, \boldsymbol{\xi})$ to hold. The second bullet in Condition (C3) requires the moment condition on the subgradient $g(\boldsymbol{\theta}^*, \boldsymbol{\xi})$ holds at the true parameter $\boldsymbol{\theta}^*$, and weakens the uniform moment condition in the first bullet. On the other hand, it imposes an extra condition on the uniform eigenvalue bound of the matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, which is equivalent to assuming that the loss function f is strongly convex on its entire domain. It is also worthwhile noting that the second bullet covers the case of linear regression where $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{X}'/n$ for all $\boldsymbol{\theta}$.

4.1.1 Theory of mini-batch SGD

Given these assumptions, we first provide some theoretical results of the mini-batch SGD in the diverging p case. In particular, let $\mathbb{E}_0(\cdot)$ be the expectation to $\{\boldsymbol{\xi}_i, 1 \le i \le n\}$ given the initial estimator $\widehat{\boldsymbol{\theta}}_0$. Let us denote the solution of mini-batch SGD in (6) with $s = n/m$ iterations by $\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$. We obtain the consistency result of the mini-batch SGD in the diverging p case. Recall that $r_i = c_0/\max(i^\alpha, p)$ for some $0 < \alpha \leq 1$ and c_0 is a sufficiently large constant.

Theorem 4.1. *Assume (C1), (C2), (C3) or (C3*) hold and the initial estimator $\widehat{\boldsymbol{\theta}}_0$ is independent to $\{\boldsymbol{\xi}_i, i = 1, 2, \ldots, n\}$. On the event $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$ with $d_n \to 0$, the mini-batch SGD estimator satisfies*

$$\mathbb{E}_0\|\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}} - \boldsymbol{\theta}^*\|_2^2 = O(\frac{p}{m^{1-\alpha}n^\alpha}) \quad \text{and} \quad \|\mathbb{E}_0(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}) - \boldsymbol{\theta}^*\|_2 = O(\frac{p}{m^{1-\alpha}n^\alpha}).$$

Furthermore, if $\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 = o_{\mathbb{P}}(1)$, then $\|\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}} - \boldsymbol{\theta}^*\|_2^2 = O_{\mathbb{P}}(\frac{p}{m^{1-\alpha}n^\alpha})$.

Theorem 4.1 characterizes both the mean squared error and the bias of the obtained estimator from SGD. When the decaying rate of the step-size $\alpha = 1$, the convergence rate is not related to m, and it achieves the same rate as the ERM $\widehat{\boldsymbol{\theta}}$ in (2) (i.e., $O(\sqrt{p/n})$).

We note that Theorem 4.1 requires a consistent initial estimator $\widehat{\boldsymbol{\theta}}_0$. In practice, we can always use a small separate subset of samples to construct the initial estimator by minimizing the empirical risk.

In contrast to the fixed p setting where an arbitrary initialization can be used, a consistent initial estimator is almost necessary to ensure the convergence in the diverging p case, which is shown in the following proposition:

Proposition 4.2. *Assume that the initial estimator $\widehat{\boldsymbol{\theta}}_0$ is independent to $\{\boldsymbol{\xi}_i, i = 1, 2, \ldots, n\}$ and satisfies $\mathbb{E}\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2^2 \geq p^{2\nu}$ for some $\nu > 0$, the step-size $r_i \leq C/i^\alpha$ for some $0 < \alpha \leq 1$ and the batch size $m \geq 1$. Suppose that $\sup_{\|\boldsymbol{v}\|_1=1} \sup_{\boldsymbol{\theta}} \mathbb{E}(\boldsymbol{v}' g(\boldsymbol{\theta}, \boldsymbol{\xi}))^2 \leq C$. We have $\mathbb{E}\|\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}} - \boldsymbol{\theta}^*\|_2^2 \geq Cp^{2\nu}$ for all $n \leq m\exp(o(p^\nu))$ when $\alpha = 1$ and for all $n = o(mp^{\nu/(1-\alpha)})$ when $0 < \alpha < 1$.*

We note that Proposition 4.2 provides a lower bound result, which shows that in the diverging p case, a standard mini-batch SGD with a random initialization will not converge with high probability. Indeed, a random initial estimator $\widehat{\boldsymbol{\theta}}_0$ will incur an error $\mathbb{E}\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2^2 \approx p$. When $p = n^\kappa$ for some $\kappa > 0$, the exponential relationship $n \leq m\exp(o(p^\nu))$ holds with $\nu = 0.5$ and thus Proposition 4.2 implies that $\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$ has a large mean squared error that is at least on the order of p. Proposition 4.2 indicates that a good initialization is crucial for SGD when p is diverging along with n.

4.1.2 Theory of DC-SGD

With the theory of mini-batch SGD in place, we provide the convergence result of the DC-SGD estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$ in (7) (see Algorithm 1). For the ease of presentation, we assume that the data are evenly distributed, where each local machine has $n = N/L$ samples.

Theorem 4.3. *Assume (C1), (C2), (C3) or (C3*) hold, suppose the initial estimator* $\widehat{\boldsymbol{\theta}}_0$ *is independent to* $\{\boldsymbol{\xi}_i, i = 1, 2, \ldots, N\}$*. On the event* $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$ *with* $d_n \to 0$*, the DC-SGD estimator achieves the following convergence rate:*

$$\mathbb{E}_0\|\widehat{\boldsymbol{\theta}}_{\mathrm{DC}} - \boldsymbol{\theta}^*\|_2^2 = O\left(\frac{p}{L^{1-\alpha}m^{1-\alpha}N^{\alpha}} + \frac{p^2L^{2\alpha}}{m^{2-2\alpha}N^{2\alpha}}\right). \tag{17}$$

The convergence rate in (17) contains two terms. The first term comes from the variance of the DC-SGD estimator, while the second one comes from the squared bias term. Note that $n = N/L$, the squared bias term in (17) can be written as $\left(\frac{p}{m^{1-\alpha}n^{\alpha}}\right)^2$, which is the same as the square of the bias from the mini-batch SGD on one machine (see Theorem 4.1). This is because the averaging of the local estimators from L machines cannot reduce the bias term. On the other hand, the variance term is reduced by a factor of $1/L$ by averaging over L machines. Therefore, when L is not too large, the variance will become the dominating term and gives the optimal convergence rate. An upper bound on L is a universal condition in the divide-and-conquer (DC) scheme to achieve the optimal rate in a statistical estimation problem (see, e.g., Li et al. (2013); Chen and Xie (2014); Zhang et al. (2015); Huang and Huo (2015); Battey et al. (2018); Zhao et al. (2016); Lee et al. (2017); Volgushev et al. (2018)). In particular, let us consider the optimal step-size r_i where $\alpha = 1$. When the number of machines $L = O(\sqrt{N/p})$, the rate in (17) becomes $O(p/N)$, which is a classical optimal rate when using all the N samples.

We next show on the two motivating examples that the constraint on the number of machines $L = O(\sqrt{N/p})$ is necessary to achieve the optimal rate by DC-SGD. To this end, we provide the lower bounds on our two examples for the bias of the SGD estimator on each local machine.

Example 3.1 (Continued). *For a logistic regression model with* $\boldsymbol{\xi} = (Y, \boldsymbol{X})$*, let* $\boldsymbol{X} = (1, X_1, ..., X_{p-1})'$ *with* $\mathbb{E}X_i = 0$ *for all* $1 \leq i \leq p-1$ *and* $\boldsymbol{\theta}^* = (1, 0, ..., 0)$*. Suppose that* $\mathbb{E}\|\boldsymbol{X}\|_2^2 \geq cp$ *for some* $c > 0$ *and* $\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}\exp(t_0|\boldsymbol{v}'\boldsymbol{X}|) \leq C$*. Suppose the initial estimator* $\widehat{\boldsymbol{\theta}}_0$ *is independent to* $\{\boldsymbol{X}_i, i = 1, 2, \ldots, n\}$*. On the event* $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$ *with* $d_n \to 0$*, we have* $\|\mathbb{E}_0(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}) - \boldsymbol{\theta}^*\|_2 \geq \frac{cp}{m^{1-\alpha}n^{\alpha}}$*.*

Example 3.2 (Continued). *For a quantile regression model, assume that* ϵ *is independent with* $\boldsymbol{X}$ *and* $\mathbb{E}X_i = 0$ *for all* $1 \leq i \leq p-1$*. Let* $F(x)$ *be the cumulative distribution function of* ϵ*. Suppose that* $\mathbb{E}\|\boldsymbol{X}\|_2^2 \geq cp$ *for some* $c > 0$ *and* $\sup_{\|\boldsymbol{v}\|_2=1}\mathbb{E}\exp(t_0|\boldsymbol{v}'\boldsymbol{X}|) \leq C$*. Suppose the initial estimator* $\widehat{\boldsymbol{\theta}}_0$ *is independent to* $\{\boldsymbol{X}_i, i = 1, 2, \ldots, n\}$*, and assume that* $F(\cdot)$ *has bounded third-order derivatives and* $F'(0)$*,* $F''(0)$ *are positive. On the event* $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$ *with* $d_n \to 0$*, we have* $\|\mathbb{E}_0(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}) - \boldsymbol{\theta}^*\|_2 \geq \frac{cp}{m^{1-\alpha}n^{\alpha}}$*.*

For the DC-SGD estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$, it is easy to see that the mean squared error $\mathbb{E}_0\|\widehat{\boldsymbol{\theta}}_{\mathrm{DC}} - \boldsymbol{\theta}^*\|_2^2 \geq \|\mathbb{E}_0(\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}) - \boldsymbol{\theta}^*\|_2^2$ (the squared bias of $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$). Recall that the bias of $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$ is the average over local machines, and each local machine induces the same bias $\|\mathbb{E}_0(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}) - \boldsymbol{\theta}^*\|_2$ (see the bias in the above two examples). Therefore, for logistic regression and quantile regression, when $\alpha = 1$ and

$\sqrt{N/p} = o(L)$, we have

$$\frac{\mathbb{E}_0\|\widehat{\boldsymbol{\theta}}_{\mathrm{DC}} - \boldsymbol{\theta}^*\|_2^2}{p/N} \geq \frac{\|\mathbb{E}_0(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}) - \boldsymbol{\theta}^*\|_2^2}{p/N} \geq \frac{c^2p^2/n^2}{p/N} = c^2\frac{L^2}{N/p} \to \infty.$$

This shows that when the number of machines L is much larger than $\sqrt{N/p}$, the convergence rate of DC-SGD will no longer be optimal.

4.2 Theory for First-order Newton-type Estimator (FONE)

We provide our main theoretical results on FONE for estimating $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ and the distributed FONE for estimating $\boldsymbol{\theta}^*$. The smooth loss and non-smooth loss functions are discussed separately in Section 4.2.1 and Section 4.2.2.

Recall that n denotes the sample size used in FONE in the single machine setting (see Algorithm 2). In our theoretical results, we denote the step-size in FONE by η_n (instead of η in Algorithms 2 and 3) to highlight the dependence of the step-size on n. For the FONE method, Condition (C1) can be further weakened to the following condition:

(C1*). Suppose that $p \to \infty$ and $p = O(n^{\kappa_1})$ for some $0 < \kappa_1 < 1$. The mini-batch size m satisfies $p \log n = o(m)$ with $m = O(n^{\kappa_2})$ for some $0 < \kappa_2 < 1$.

4.2.1 Smooth loss function f

To establish the convergence rate of our distributed FONE, we first provide a consistency result for $\widehat{\boldsymbol{\theta}}_{\mathrm{FONE}}$ in (8).

Proposition 4.4 (On $\widehat{\boldsymbol{\theta}}_{\mathrm{FONE}}$ for $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ for smooth loss function f)**.** *Assume (C1*), (C2) and (C3) hold. Suppose that the initial estimator satisfies $\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(d_n)$, and $\|\boldsymbol{a}\|_2 = O(\tau_n)$ (or $O_{\mathbb{P}}(\tau_n)$ for the random case). The iteration number T and step-size η_n satisfy $\log n = o(\eta_n T)$ and $T = O(n^A)$ for some $A > 0$. We have*

$$\|\widehat{\boldsymbol{\theta}}_{\mathrm{FONE}} - \boldsymbol{\Sigma}^{-1}\boldsymbol{a}\|_2 = O_{\mathbb{P}}\big(\tau_n d_n + \tau_n^2 + \sqrt{\frac{p\log n}{n}}\tau_n + \sqrt{\eta_n}\tau_n + n^{-\gamma}\big) \tag{18}$$

for any large $\gamma > 0$.

The relationship between η_n and T (i.e., $\log n = o(\eta_n T)$) is intuitive since when the step-size η_n is small, Algorithm 2 requires more iterations T to converge. The consistency of the estimator requires that the length of the vector $\boldsymbol{a}$ goes to zero, i.e., $\tau_n = o(1)$, since τ_n^2 appears in the convergence rate in (18). In Section 4.3, we further discuss a slightly modified FONE that deals with any vector $\boldsymbol{a}$, which applies to the estimation of the limiting variance of $\widehat{\boldsymbol{\theta}}$ in (3). When $\tau_n = o(1)$, $d_n = o(1)$, and $\eta_n = o(1)$, each term in $O_{\mathbb{P}}$ in (18) goes to zero and thus the proposition guarantees that $\widehat{\boldsymbol{\theta}}_{\mathrm{FONE}}$ is a consistent estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$. Moreover, since Proposition 4.4 will be used as an intermediate step for establishing the convergence rate of the distributed FONE, to facilitate the ease of use of Proposition 4.4, we leave d_n, τ_n, and η_n unspecified here and discuss their magnitudes in Theorem 4.5. A practical choice of η_n is further discussed in the experimental section.

Given Proposition 4.4, we now provide the convergence result for the multi-round distributed FONE for estimating $\boldsymbol{\theta}^*$ and approximating $\widehat{\boldsymbol{\theta}}$ in Algorithm 3. To this end, let us first provide some intuitions on the improvement for one-round distributed FONE from the initial estimator $\widehat{\boldsymbol{\theta}}_0$ to $\widehat{\boldsymbol{\theta}}_1$. For the first round in Algorithm 3, the algorithm essentially estimates $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ with $\boldsymbol{a} = \frac{1}{N}\sum_{i=1}^{N} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)$. When $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ is differentiable and noting that $\frac{1}{N}\sum_{i=1}^{N} g(\widehat{\boldsymbol{\theta}}, \boldsymbol{\xi}_i) = 0$ (where $\widehat{\boldsymbol{\theta}}$ is the ERM in (2)), we can prove that (see more details in the proof of Theorem 4.5),

$$\begin{aligned}
\boldsymbol{a} &= \frac{1}{N}\sum_{i=1}^{N} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i) - g(\widehat{\boldsymbol{\theta}}, \boldsymbol{\xi}_i) \\
&= G(\widehat{\boldsymbol{\theta}}_0) - G(\widehat{\boldsymbol{\theta}}) + \frac{1}{N}\sum_{i=1}^{N}\{[g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i) - g(\widehat{\boldsymbol{\theta}}, \boldsymbol{\xi}_i)] - [G(\widehat{\boldsymbol{\theta}}_0) - G(\widehat{\boldsymbol{\theta}})]\} \\
&= \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}) + O_{\mathbb{P}}(1)\left(\|\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}\|_2\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 + \|\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}\|_2^2\right) \\
&\quad + O_{\mathbb{P}}(1)\Big(\sqrt{\frac{p\log N}{N}}\|\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}\|_2 + N^{-\gamma}\Big),
\end{aligned} \tag{19}$$

for any $\gamma > 0$. Note that in Algorithm 3, the FONE procedure is executed on the first machine. For the ease of plugging the result in Proposition 4.4 on FONE, we let $n := n_1$ to denote the sub-sample size on the first machine.

Assuming that the initial estimator $\widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\theta}}$ satisfy $\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(n^{-\delta_1})$ for some $\delta_1 > 0$, then by (19), we have $\|\boldsymbol{a}\|_2 = O_{\mathbb{P}}(n^{-\delta_1})$ (i.e., the length $\tau_n = O(n^{-\delta_1})$ in Proposition 4.4). Moreover, we can further choose d_n in Proposition 4.4 to be $d_n = O(n^{-\delta_1})$. Let the step-size $\eta_n = n^{-\delta_2}$ for some $\delta_2 > 0$. After one round of distributed FONE in Algorithm 3, by Proposition 4.4, we can obtain that $\|\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}\|_2 = O_{\mathbb{P}}(n^{-\delta_1-\delta_0})$ with $\delta_0 = \min(\delta_1, \delta_2/2, (1-\kappa_1)/2)$, where κ_1 is the parameter in our assumption $p = O(n^{\kappa_1})$ (see Condition (C1*)). Therefore, one can see that, with one round of distributed FONE, the rate of convergence improves from $O(n^{-\delta_1})$ to $O(n^{-\delta_1-\delta_0})$. Therefore, by implementing distributed FONE for K rounds, we will have $\|\widehat{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}}\|_2 = O(n^{-\delta_1-K\delta_0})$. This convergence result of distributed FONE is formally stated in the next theorem.

Theorem 4.5 (distributed FONE for smooth loss function f)**.** *Assume (C1*), (C2) and (C3) hold, $N = O(n^A)$ for some $A > 0$. Suppose that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 + \|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(n^{-\delta_1})$ for some $\delta_1 > 0$. Let $\eta_n = n^{-\delta_2}$ for some $\delta_2 > 0$, $\log n = o(\eta_n T)$, $T = O(n^A)$ for some $A > 0$, and $p\log n = o(m)$. For any $\gamma > 0$, there exists $K_0 > 0$ such that for any $K \geq K_0$, we have $\|\widehat{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}}\|_2 = O_{\mathbb{P}}(n^{-\gamma})$.*

Again, we recall that $n = n_1$ denotes the number of samples on the first machine. Note that γ in Theorem 4.5 can be arbitrarily large. Hence our estimator $\widehat{\boldsymbol{\theta}}_K$ can asymptotically approximate the ideal solution $\widehat{\boldsymbol{\theta}}$ with a fast rate. Under some regular conditions, it is typical that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(\sqrt{p/N})$. Therefore, for a smooth loss function f, our distributed FONE achieves the optimal rate $O_{\mathbb{P}}(\sqrt{p/N})$. Note that it does not need any condition on the number of machines L. Given the step-size $\eta_n = n^{-\delta_2}$, by the condition $\log n = o(\eta_n T)$, we can choose the number of iterations $T = n^{\delta_2}(\log n)^2$ in the distributed FONE. Therefore, the computation complexity of distributed FONE is $O(np + n^{\delta_2}(\log n)^2 mp)$ for each round, on the first machine. We also note that n is the sub-sample on the first machine, which is much smaller than the total sample size N. In terms of the communication cost, each machine only requires to transmit an $O(p)$ vector for each round.

4.2.2 Non-smooth loss function f

For a non-smooth loss, we provide the following convergence rate of the FONE of $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ under Conditions (C1*), (C2) and (C3*).

Proposition 4.6 (On $\widehat{\boldsymbol{\theta}}_{\mathrm{FONE}}$ for $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ for non-smooth loss function f)**.** *Assume the conditions in Proposition 4.4 hold with (C3) being replaced by (C3*). We have*

$$\|\widehat{\boldsymbol{\theta}}_{\mathrm{FONE}} - \boldsymbol{\Sigma}^{-1}\boldsymbol{a}\|_2 = O_{\mathbb{P}}\Big(\tau_n d_n + \tau_n^2 + \sqrt{\frac{p\log n}{n}}\sqrt{\tau_n} + \frac{p\log n}{m}\sqrt{\eta_n} + \sqrt{\eta_n}\tau_n + \frac{p\log n}{n}\Big). \tag{20}$$

With Proposition 4.6 in hand, we now provide the convergence rate of the distributed FONE in Algorithm 3 under Condition (C3*). It is worthwhile noting that when $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ is non-differentiable, then $\frac{1}{N}\sum_{i=1}^N g(\widehat{\boldsymbol{\theta}}, \boldsymbol{\xi}_i)$ can be nonzero due to the discontinuity in the function $g(\boldsymbol{\theta}, \boldsymbol{\xi})$, where $\widehat{\boldsymbol{\theta}}$ is the ERM in (2). Therefore we need to assume that

$$\sum_{i=1}^N g(\widehat{\boldsymbol{\theta}}, \boldsymbol{\xi}_i) = O_{\mathbb{P}}(q_N) \tag{21}$$

with $q_N = O(N^{\kappa_3})$ for some $0 < \kappa_3 < 1$. For example, for a quantile regression, $q_N = O(p^{3/2}\log N)$ (He and Shao, 2000), which satisfies this condition when $p = o(N^{\kappa_4})$ with $0 < \kappa_4 < 2/3$.

Given Conditions (C1*), (C2), (C3*) and (21), we have the following convergence rate of $\widehat{\boldsymbol{\theta}}_K$:

Theorem 4.7 (distributed FONE for non-smooth loss function f)**.** *Suppose that (C1*), (C2), (C3*) and (21) hold, $N = O(n^A)$ and $T = O(n^A)$ for some $A > 0$. Suppose that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 + \|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(n^{-\delta_1})$ for some $\delta_1 > 0$. Let $\eta_n = n^{-\delta_2}$ for some $\delta_2 > 0$, $\log n = o(\eta_n T)$, and $p\log n = o(m)$. For any $0 < \gamma < 1$, there exists $K_0 > 0$ such that for any $K \geq K_0$, we have*

$$\|\widehat{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}}\|_2 = O_{\mathbb{P}}\Big(\frac{q_N}{N} + \sqrt{\eta_n}\frac{p\log n}{m} + \big(\frac{p\log n}{n}\big)^{\gamma}\Big). \tag{22}$$

As one can see from (22), the distributed FONE has a faster convergence rate when the sub-sample size on the first machine n_1 is large (recall that $n := n_1$). In practice, it is usually affordable to increase the memory and computational resources for only one local machine. This is different from the case of DC-SGD, where the convergence rate actually depends on the smallest sub-sample size among local machines.[2]

Let us provide more discussion on the convergence rate in (22). When the number of rounds K gets larger, the parameter γ in the exponent of the third term can be arbitrarily close to 1. Therefore, for the ease of discussion, let us treat γ as 1. For the second term in the right-hand side of (22), we can choose the step-size η_n and the batch size m such that $\sqrt{\eta_n}/m \leq 1/(n \log n)$, and the second term will be dominated by the third one. So we only need to consider the first term q_N/N and the third term $(p \log n)/n$. Due to q_N in (21), the relationship between these two terms depends on the specific model. Usually, under some conditions on the dimension p, $\|\widehat{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}}\|_2$ achieves a faster rate than $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$, which makes $\widehat{\boldsymbol{\theta}}_K$ attain the optimal rate for estimating $\boldsymbol{\theta}^*$. Let us take the quantile regression as an example, where the ERM $\widehat{\boldsymbol{\theta}}$ has an error rate of $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(\sqrt{p/N})$ and $q_N = O(p^{3/2} \log N)$ (He and Shao, 2000). Assuming that $p = O(\sqrt{N}/\log N)$ and $n \geq c\sqrt{Np} \log N$, both the first and the third terms will be $O(\sqrt{p/N})$, such that $\|\widehat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(\sqrt{p/N})$.

Similar to the smooth case, the computation complexity is $O(np+n^{\delta_2}(\log n)^2 mp)$ for each round, on the first machine. Assuming the second term of (22) is dominated by the third term, we may specify $m = \sqrt{\eta_n} n \log n$ and the corresponding computation complexity becomes $O(n^{1+\delta_2/2}(\log n)^3 p)$. Again, each machine only requires to transmit an $O(p)$ vector for each round.

4.3 Application to the estimation of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ with $\|\boldsymbol{w}\|_2 = 1$

Another important application of the proposed FONE is to conduct the inference of $\boldsymbol{\theta}^*$ in the diverging p case. Recall the limiting distribution result in (3). To estimate the limiting variance, we note that $\boldsymbol{A}$ can be easily estimated by $\widehat{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i) g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)'$. Therefore, we only need to estimate $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$. The challenge here is the theory of our Propositions 4.4 and 4.6 only applies to the case $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ with $\|\boldsymbol{a}\|_2 = o(1)$ or $o_{\mathbb{P}}(1)$. However, in the inference application, we have $\|\boldsymbol{w}\|_2 = 1$. To address this challenge, given the unit length vector $\boldsymbol{w}$, we define $\boldsymbol{a} = \tau_n \boldsymbol{w}$, where $\|\boldsymbol{a}\|_2 = \tau_n = o(1)$ and its rate will be specified later in our theoretical results in Theorems 4.8 and 4.9. We run Algorithm 2 and its output $\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{z}_T$ is an estimator of $\tau_n \boldsymbol{\Sigma}^{-1}\boldsymbol{w}$. Then the estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ can be naturally constructed as,

$$\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}} = \frac{\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{z}_T}{\tau_n}, \quad \text{where in Algorithm 2} \quad \boldsymbol{a} = \tau_n \boldsymbol{w}. \tag{23}$$

We note that the initial estimator $\widehat{\boldsymbol{\theta}}_0$ for estimating $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ needs to be close to the targeting parameter $\boldsymbol{\theta}^*$. In a non-distributed setting, we could choose the ERM $\widehat{\boldsymbol{\theta}}$ as $\widehat{\boldsymbol{\theta}}_0$ for inference, while in the distributed setting, we use $\widehat{\boldsymbol{\theta}}_K$ from distributed FONE in Algorithm 3 with a sufficiently large K.

[2]Noting that although we present the evenly distributed setting for DC-SGD for the ease of illustration, one can easily see the convergence rate is actually determined by the smallest sub-sample size from the proof.

We next provide the theoretical results of the estimator in (23) for two cases: f is smooth and f is non-smooth. We note that for the purpose of asymptotic valid inference, we only need $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$ in (23) to be a consistent estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$. To show the consistency of our estimator, we provide the convergence rates in the following Theorems 4.8 and 4.9 for smooth and non-smooth loss functions, respectively:

Theorem 4.8 (Estimating $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ for a smooth loss function f)**.** *Under the conditions of Proposition 4.4, let* $\tau_n = \sqrt{(p\log n)/n}$*. Assuming that* $\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(d_n)$ *and* $\log n = o(\eta_n T)$*, we have*

$$\|\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}} - \boldsymbol{\Sigma}^{-1}\boldsymbol{w}\|_2 = O_{\mathbb{P}}\Big(\sqrt{\frac{p\log n}{n}} + \sqrt{\eta_n} + d_n\Big). \tag{24}$$

From Theorem 4.8, the estimator $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$ is consistent as long as $d_n = o(1)$ and the step-size $\eta_n = o(1)$. Let us further provide some discussion on the convergence rate in (24). If we choose a good initiation such that $d_n = O(\sqrt{(p\log n)/n})$, the term d_n in (24) will be a smaller order term. For example, the initialization rate $d_n = O(\sqrt{(p\log n)/n})$ can be easily satisfied by using either the ERM $\widehat{\boldsymbol{\theta}}$ or $\widehat{\boldsymbol{\theta}}_K$ from distributed FONE with a sufficiently large K. Moreover, we can specify η_n to be small (e.g., $\eta_n = O((p\log n)/n)$). Then the rate in (24) is $\sqrt{(p\log n)/n}$, which almost matches the parametric rate for estimating a p dimensional vector.

For non-smooth loss function f, we have the following convergence rate of $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$:

Theorem 4.9 (Estimating $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ for non-smooth loss function f)**.** *Under the conditions of Proposition 4.6, let* $\tau_n = \big((p\log n)/n\big)^{1/3}$*. Assuming that* $\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 = O_{\mathbb{P}}(d_n)$ *and* $\log n = o(\eta_n T)$*, we have*

$$\|\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}} - \boldsymbol{\Sigma}^{-1}\boldsymbol{w}\|_2 = O_{\mathbb{P}}\Big(\big(\frac{p\log n}{n}\big)^{1/3} + \sqrt{\eta_n}\big(\frac{n^{1/3}(p\log n)^{2/3}}{m} + 1\big) + d_n\Big). \tag{25}$$

To make d_n a smaller order term in the rate in (25), we choose a good initiation such that $d_n = O((p\log n)/n)^{1/3})$. As long as the step-size η_n is small such that $\eta_n = \min\big(\frac{(p\log n)^{2/3}}{n^{2/3}}, \frac{m^2}{(p\log n)^{2/3}n^{4/3}}\big)$, the convergence rate in (25) is $O_{\mathbb{P}}\big(((p\log n)/n)^{1/3}\big)$, which implies that $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$ is a consistent estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$.

Furthermore, we briefly comment on an efficient implementation for computing the limiting variance $\boldsymbol{w}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$. Instead of explicitly constructing the estimator of $\boldsymbol{A}$ by a $p \times p$ matrix $\widehat{\boldsymbol{A}} = \frac{1}{n}\sum_{i=1}^{n} g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)'$, we can directly compute the estimator by

$$(\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}})'\widehat{\boldsymbol{A}}(\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}) = \frac{1}{n}\sum_{i=1}^{n}\Big(g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)'\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}\Big)^2, \tag{26}$$

where $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$ is pre-computed by FONE. The implementation in (26) only incurs a computation cost of $O(np)$.

5 Experimental Results

In this section, we provide simulation studies to illustrate the performance of our methods on two statistical estimation problems in Examples 3.1–3.2, i.e., logistic regression and quantile regression (QR). For regression problems in the two motivating examples, let $\boldsymbol{\xi}_i = (Y_i, \boldsymbol{X}_i)$ for $i = 1, 2, \ldots, N$, where $\boldsymbol{X}_i = (1, X_{i,1}, X_{i,2}, \ldots, X_{i,p-1})' \in \mathbb{R}^p$ is a random covariate vector and N is the total sample size. Here $(X_{i,1}, X_{i,2}, \ldots, X_{i,p-1})$ follows a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_{p-1})$, where $\mathbf{I}_{p-1}$ is a $p-1$ dimensional identity matrix. We also provide the simulation studies with correlated design $\boldsymbol{X}$, which are relegated to Appendix E.1). The true coefficient $\boldsymbol{\theta}^*$ follows a uniform distribution $\mathrm{Unif}([-0.5, 0.5]^p)$. For QR in Example 3.2, we follow the standard approach (see, e.g., Pang et al. (2012)) that first generates the data from a linear regression model $Y_i = \boldsymbol{X}_i'\boldsymbol{\theta} + \epsilon_i$, where $\boldsymbol{\theta}$ follows a uniform distribution $\mathrm{Unif}([-0.5, 0.5]^p)$ and $\epsilon_i \sim N(0, 1)$. For each quantile level τ, we need to compute the true QR coefficient $\boldsymbol{\theta}^*$ by shifting ϵ_i such that $\Pr(\epsilon_i \le 0) = \tau$. Thus, the true QR coefficient $\boldsymbol{\theta}^* = \boldsymbol{\theta} + (\Phi^{-1}(\tau), 0, 0, \ldots, 0)'$, where Φ is the CDF of the standard normal distribution. In our experiment, we set the quantile level $\tau = 0.25$. All of the data points are evenly distributed on L machines with sub-sample size $n = n_i = N/L$ for $i = 1, 2, \ldots, L$. We further discuss the imbalanced situation in Section 5.4.

In the following experiments, we evaluate the DC-SGD estimator (see Algorithm 1) and distributed FONE (Dis-FONE, see Algorithm 3) by their L_2-estimation errors. In particular, we report the L_2-distance to the true coefficient $\boldsymbol{\theta}^*$ as well as the L_2-distance to the ERM $\widehat{\boldsymbol{\theta}}$ in (2). We also compare the methods with mini-batch SGD in (6) on the entire dataset in a non-distributed setting, which can be considered as a special case of DC-SGD when the number of machines $L = 1$. For all these methods, it is required to provide a consistent initial estimator $\widehat{\boldsymbol{\theta}}_0$. In our experiments below, we compute the initial estimator by minimizing the empirical risk function in (2) with a small batch of fresh samples. It is clear that as dimension p grows, it requires more samples to achieve the desired accuracy of the initial estimator. Therefore, we specify the size of the fresh samples as $n_0 = 10p$. We also discuss the effect of the accuracy of the initial estimator $\widehat{\boldsymbol{\theta}}_0$ by varying n_0 (see Appendix E.2).

For DC-SGD, the step-size is set to $r_i = c_0/\max(i^\alpha, p)$ with $\alpha = 1$, and c_0 is a positive scaling constant. We use an intuitive data-driven approach to choose c_0. We first specify a set $\mathcal{C}$ of candidate choices for c_0 ranging from 0.001 to 1000. We choose the best c_0 that achieves the smallest objective function in (2) with $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(1)}_{\mathrm{SGD}}$ using data points from the first machine (see Algorithm 1), i.e., $c_0 = \arg\min_{c \in \mathcal{C}} \frac{1}{n}\sum_{i=1}^n f(\widehat{\boldsymbol{\theta}}^{(1)}_{\mathrm{SGD}}, \boldsymbol{\xi}_i^{(1)})$, where $\{\boldsymbol{\xi}_i^{(1)}, i = 1, 2, \ldots, n\}$ denotes the samples on the first machine. For Dis-FONE, the step-size is set to $\eta = c_0' m/n$, where c_0' is also selected from a set $\mathcal{C}$ of candidate constants. Similarly, we choose the best tuning constant that achieves the smallest objective in (2) with $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_1$ and samples from the first machine. Here, $\widehat{\boldsymbol{\theta}}_1$ is the output of Dis-FONE after the first round of the algorithm. That is, $c_0' = \arg\min_{c \in \mathcal{C}} \frac{1}{n}\sum_{i=1}^n f(\widehat{\boldsymbol{\theta}}_1, \boldsymbol{\xi}_i^{(1)})$. Moreover, given Condition (C1), we set the mini-batch size in DC-SGD (or the size of B_t in Dis-

Table 1: Logistic regression: comparisons of L_2-errors when varying the total sample size N and dimension p changes. Here the number of machines $L = 20$. Denote by $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$ the DC-SGD estimator and $\widehat{\boldsymbol{\theta}}_K$ the Dis-FONE with $K = 20$.

p	L_2-distance to the truth $\boldsymbol{\theta}^*$					L_2-distance to ERM $\widehat{\boldsymbol{\theta}}$		
	$\widehat{\boldsymbol{\theta}}_0$	$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$	$\widehat{\boldsymbol{\theta}}_K$	$\widehat{\boldsymbol{\theta}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$	$\widehat{\boldsymbol{\theta}}_K$
$N = 10^5$								
100	1.251	0.447	0.148	0.103	0.093	0.445	0.116	0.038
200	1.899	1.096	0.523	0.168	0.153	1.091	0.494	0.049
500	4.509	3.853	3.111	0.338	0.301	3.748	3.021	0.085
$N = 2 \times 10^5$								
100	1.303	0.390	0.100	0.072	0.067	0.386	0.074	0.025
200	2.094	1.248	0.315	0.115	0.109	1.235	0.286	0.034
500	4.717	3.920	2.189	0.222	0.211	3.891	2.133	0.045
$N = 5 \times 10^5$								
100	1.342	0.313	0.081	0.046	0.042	0.304	0.069	0.018
200	1.833	0.874	0.169	0.073	0.068	0.868	0.152	0.023
500	4.835	3.885	1.006	0.141	0.130	3.859	0.989	0.036

FONE, see Algorithm 3) as $m = \lfloor p \log n \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. For Dis-FONE, we first set the number of the iterations T in each round as $T = 20$ and the number of rounds $K = 20$ for logistic regression and $K = 80$ for quantile regression. Note that due to the non-smoothness in the loss function of quantile regression, it requires more rounds of iterations K to ensure the convergence. We carefully evaluate the effect of T and K (by considering different values of T and K) in Section 5.3. All results reported below are based on the average of 100 independent runs of simulations.

Furthermore, we also conduct the simulation studies for the estimator of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ and the estimator for the limiting variance in (26) described in Section 4.3. Due to space limitations, the results are provided in Appendix E.3.

5.1 Effect of N and p

In Tables 1–2, we fix the number of machines $L = 20$ and vary the total sample size N from $\{10^5, 2\times 10^5, 5\times 10^5\}$ and dimension $p \in \{100, 200, 500\}$. Results for logistic regression are reported in Table 1 and results for quantile regression are in Table 2. In both tables, the left columns provide the L_2 estimation errors (with respect to the truth $\boldsymbol{\theta}^*$) of the DC-SGD estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$, SGD estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$, Dis-FONE $\widehat{\boldsymbol{\theta}}_K$, and the ERM $\widehat{\boldsymbol{\theta}}$. For reference, we also report L_2-errors of the initial estimator $\widehat{\boldsymbol{\theta}}_0$. The right columns report the L_2-distances to the benchmark ERM $\widehat{\boldsymbol{\theta}}$.

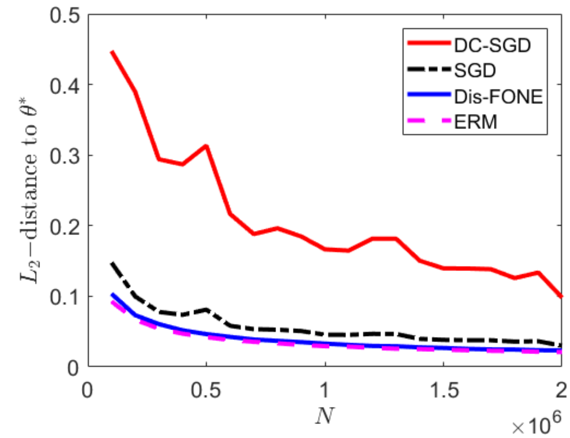
Table 2: Quantile regression: comparisons of L_2-errors when varying the total sample size N and dimension p. Here the number of machines $L = 20$. Denote by $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$ the DC-SGD estimator and $\widehat{\boldsymbol{\theta}}_K$ the Dis-FONE with $K = 80$.

p	L_2-distance to the truth $\boldsymbol{\theta}^*$					L_2-distance to ERM $\widehat{\boldsymbol{\theta}}$		
	$\widehat{\boldsymbol{\theta}}_0$	$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$	$\widehat{\boldsymbol{\theta}}_K$	$\widehat{\boldsymbol{\theta}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$	$\widehat{\boldsymbol{\theta}}_K$
$N = 10^5$								
100	0.450	0.079	0.063	0.047	0.043	0.073	0.050	0.020
200	0.715	0.114	0.109	0.082	0.071	0.106	0.097	0.035
500	1.278	0.198	0.176	0.144	0.126	0.176	0.142	0.062
$N = 2 \times 10^5$								
100	0.450	0.070	0.037	0.035	0.030	0.067	0.021	0.015
200	0.726	0.101	0.067	0.059	0.054	0.098	0.037	0.027
500	1.287	0.176	0.118	0.098	0.076	0.157	0.065	0 .046
$N = 5 \times 10^5$								
100	0.451	0.043	0.030	0.029	0.025	0.037	0.017	0.014
200	0.719	0.067	0.047	0.041	0.037	0.064	0.15	0.020
500	1.294	0.105	0.076	0.074	0.057	0.276	0.99	0.035

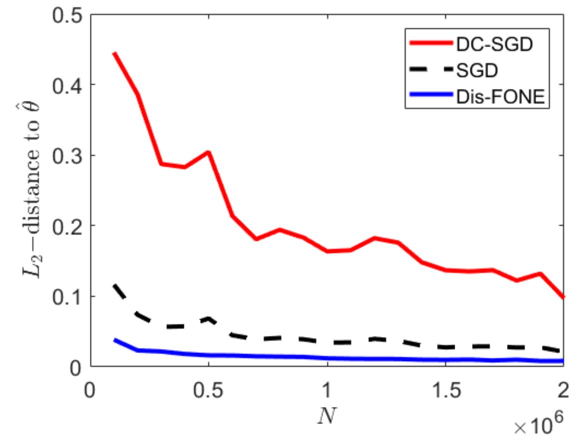
From Tables 1–2, we can see that the proposed Dis-FONE $\widehat{\boldsymbol{\theta}}_K$ achieves similar errors as the ERM $\widehat{\boldsymbol{\theta}}$ in all cases, and outperforms DC-SGD and SGD especially when p is large. We also provide Figure 1 that captures the performance of the estimators in terms of their L_2-errors when the total sample size N increases. From Figure 1, we can see that the estimation error for each method decreases as N increases. Moreover, the L_2-error of Dis-FONE is very close to the ERM as N increases, while there is a significant gap between DC-SGD and the ERM .

5.2 Effect on the number of machines L

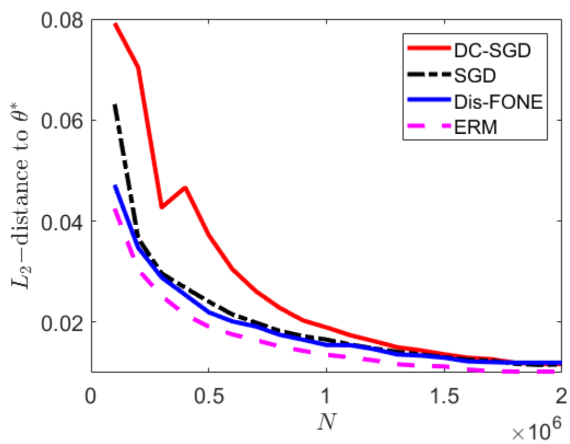
For the effect on the number of machines L, we fix the total sample size $N = 10^5$ and the dimension $p = 100$ and vary the number of machines L from 1 to 100, and plot the L_2-errors in Figure 2. From Figure 2, the L_2-error of DC-SGD increases as L increases (i.e., each machine has fewer samples). In contrast, the L_2-error of Dis-FONE versus L is almost flat, and is very close to ERM even when L is large. This is consistent with our theoretical result that DC-SGD will fail when L is large. The SGD estimator, which is the $L = 1$ case of DC-SGD (and thus its error is irrelevant of L and is presented by a horizontal line), provides moderate accuracy.



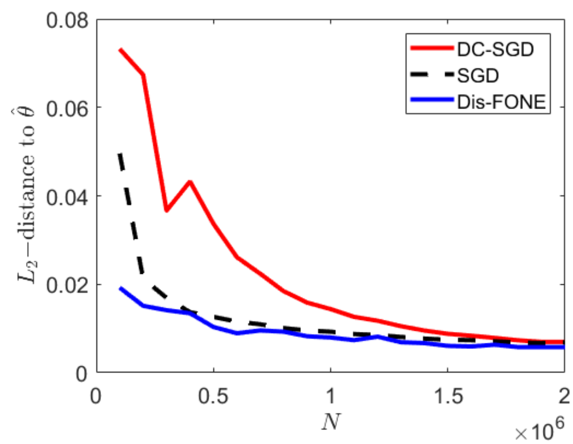
(a) Logistic regression: L_2-distance to $\boldsymbol{\theta}^*$



(b) Logistic regression: L_2-distance to $\widehat{\boldsymbol{\theta}}$



(c) Quantile regression: L_2-distance to $\boldsymbol{\theta}^*$

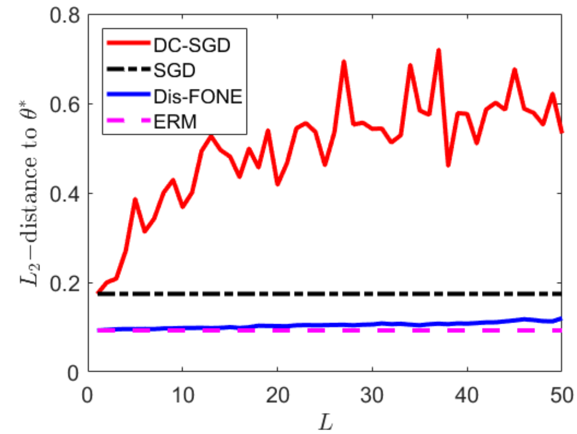


(d) Quantile regression: L_2-distance to $\widehat{\boldsymbol{\theta}}$

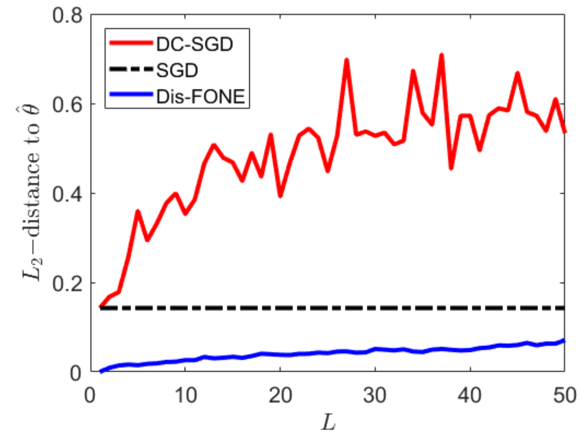
Figure 1: Comparison of L_2-errors when N increases. The left column reports the L_2-errors with respect to the truth $\boldsymbol{\theta}^*$ and the right column reports the L_2-errors with respect to the ERM $\widehat{\boldsymbol{\theta}}$. Here the dimension $p = 100$ and the number of machines $L = 20$. In Dis-FONE, we set $K = 20$ in the logistic regression case and $K = 80$ in the quantile regression case.

5.3 Effect of K and T in Dis-FONE

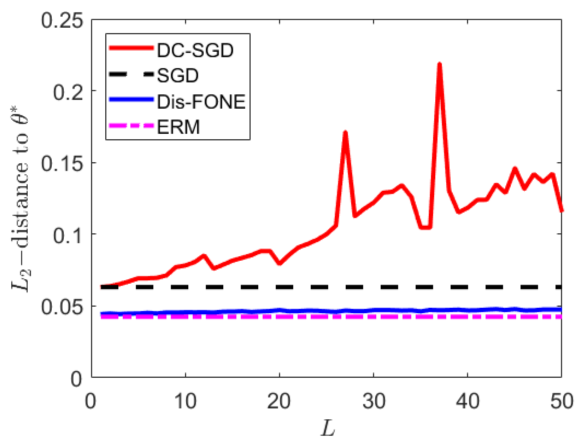
For Dis-FONE, we provide the comparison of the estimator errors with different numbers of rounds K and numbers of inner iterations T. In Figure 3, we fix the total sample size $N = 10^5$, the dimension $p = 100$, the number of machines $L = 20$ and vary T from $\{5, 20, 100\}$. The x-axis in Figure 3 is the number of rounds K. For all three cases of T, the performance of Dis-FONE is quite desirable and reaches the accuracy of the ERM when K becomes larger. When T is smaller, it requires a larger K for Dis-FONE to converge. In other words, we need to perform more rounds



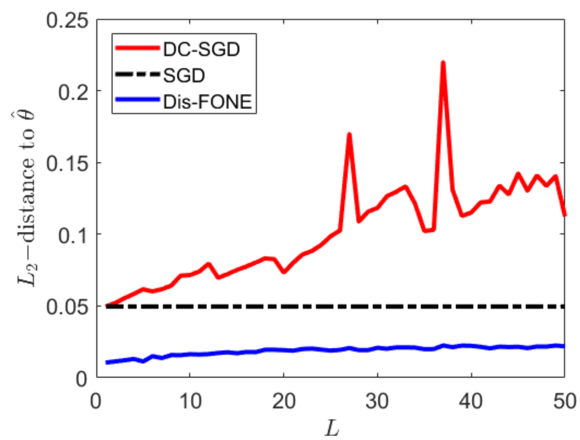
(a) Logistic regression: L_2-distance to $\boldsymbol{\theta}^*$



(b) Logistic regression: L_2-distance to $\widehat{\boldsymbol{\theta}}$



(c) Quantile regression: L_2-distance to $\boldsymbol{\theta}^*$



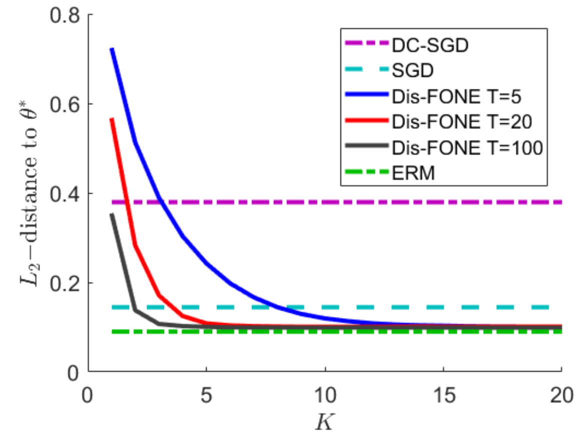
(d) Quantile regression: L_2-distance to $\widehat{\boldsymbol{\theta}}$

Figure 2: Comparison of L_2-errors when the number of machines L increases. Here the total sample size $N = 10^5$ and the dimension $p = 100$. Denote by $\widehat{\boldsymbol{\theta}}_K$ the Dis-FONE with $K = 20$ in the logistic regression case and $K = 80$ in the quantile regression case.

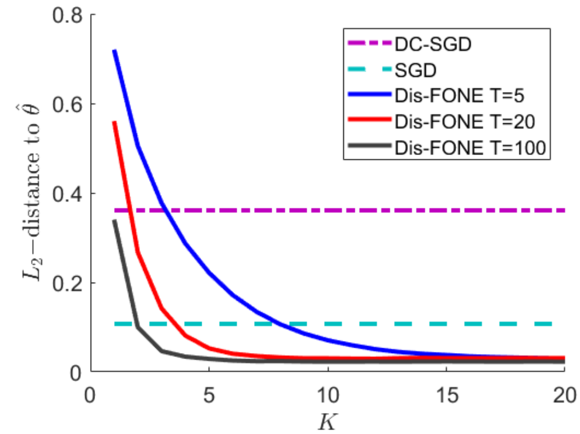
of Dis-FONE to achieve the same accuracy.

5.4 Effect on the sub-sample size of the first machine n_1 in Dis-FONE

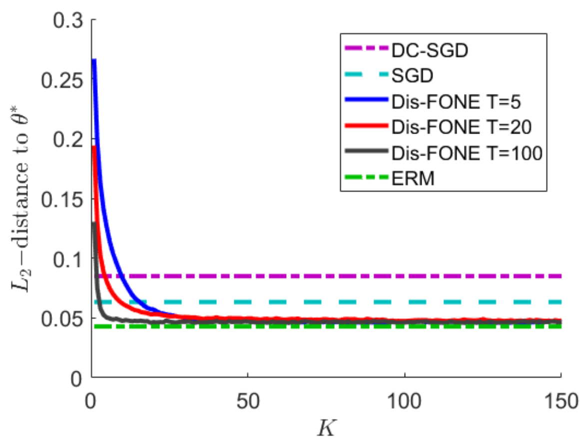
In previous simulation studies, the entire dataset is evenly separated on different machines. As one can see from Algorithm 3 and Theorem 4.7, the sub-sample size on the first machine n_1 plays a different role than that on the other machines $n_2, n_3, \ldots, n_L$ in Dis-FONE. In Figure 4, we investigate the effect of n_1 by varying n_1 from N/L (the case of evenly distributed) to $10 \times N/L$. Let the remaining data points be evenly distributed on the other machines, i.e., $n_2 = n_3 = \cdots =$



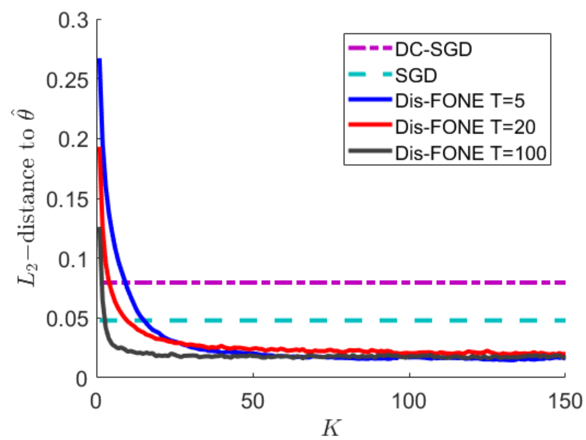
(a) Logistic regression: L_2-distance to $\boldsymbol{\theta}^*$



(b) Logistic regression: L_2-distance to $\widehat{\boldsymbol{\theta}}$



(c) Quantile regression: L_2-distance to $\boldsymbol{\theta}^*$

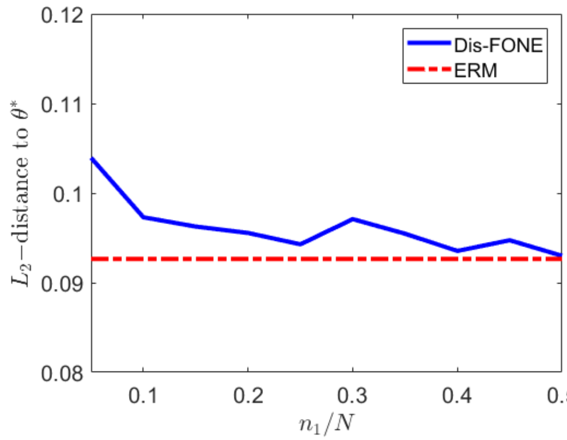


(d) Quantile regression: L_2-distance to $\widehat{\boldsymbol{\theta}}$

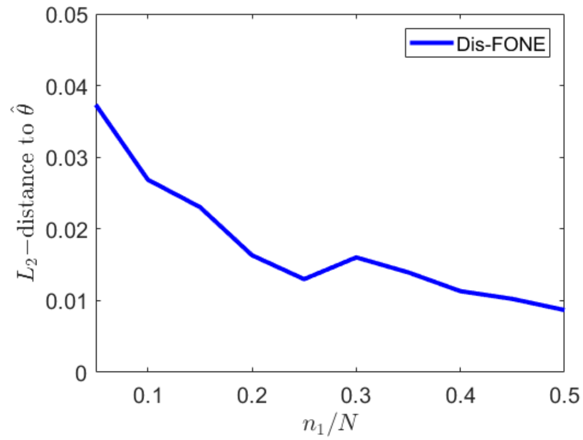
Figure 3: Comparison of L_2-errors when the number of rounds K in Dis-FONE increases. The x-axis is the number of rounds K in Dis-FONE. Here the total sample size $N = 10^5$, the dimension $p = 100$, and the number of machines $L = 20$. The errors of DC-SGD, SGD, and ERM are presented by the horizontal lines since their performance is irrelevant of K.

$n_L = (N - n_1)/(L - 1)$. We set $N = 10^5$ and $L = 20$. From Figure 4, the L_2-error of Dis-FONE gets much closer to ERM $\widehat{\boldsymbol{\theta}}$ in (2) when the largest sub-sample size n_1 increases, which is consistent with our theoretical results.

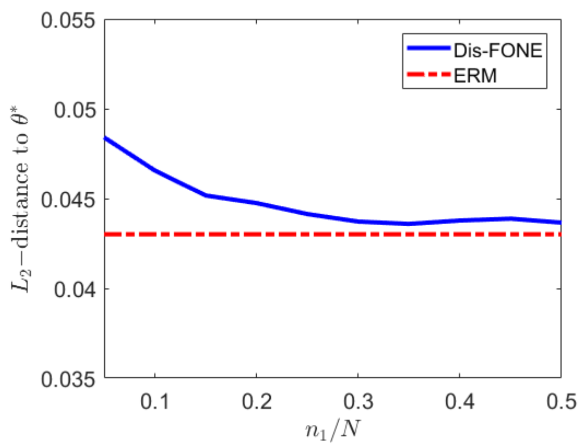
In Appendix E, we further investigate the case of correlated design, the effect of the quality of the initial estimator, as well as the performance of the estimator of limiting variance.



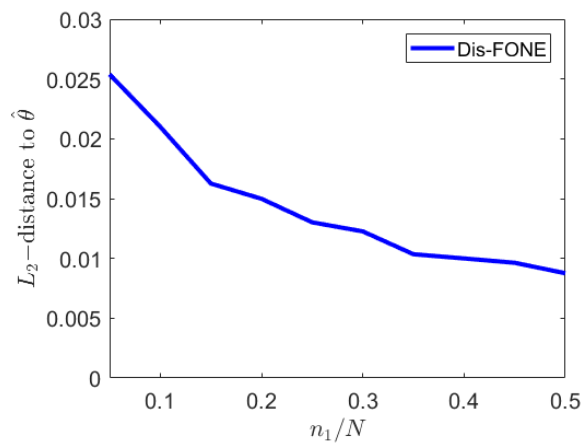
(a) Logistic regression: L_2-distance to $\boldsymbol{\theta}^*$



(b) Logistic regression: L_2-distance to $\widehat{\boldsymbol{\theta}}$



(c) Quantile regression: L_2-distance to $\boldsymbol{\theta}^*$



(d) Quantile regression: L_2-distance to $\widehat{\boldsymbol{\theta}}$

Figure 4: Comparison of L_2-errors when the sub-sample size of the first machine n_1 in Dis-FONE increases. The x-axis is the ratio of n_1 to the total sample size N. Here the total sample size $N = 10^5$, the dimension $p = 100$, and the number of machines $L = 20$.

6 Conclusions

This paper studies general distributed estimation and inference problems based on stochastic subgradient descent. We propose an efficient First-Order Newton-type Estimator (FONE) for estimating $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ and its distributed version. The key idea behind our method is to use stochastic gradient information to approximate the Newton step. We further characterize the theoretical properties when using FONE for distributed estimation and inference with both smooth and non-smooth loss functions. We also conduct simulation studies to demonstrate the performance of the proposed distributed FONE. The proposed FONE of $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ is general a estimator, which could find applications to other statistical estimation problems.

A Technical Lemmas

In this section, we will give some technical lemmas which are used to prove the main results.

Lemma A.1. *Let $\boldsymbol{\zeta}_1, ..., \boldsymbol{\zeta}_n$ be independent p-dimensional random vectors with $\mathbb{E}\boldsymbol{\zeta}_i = \mathbf{0}$ and*

$$\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}(\boldsymbol{v}'\boldsymbol{\zeta}_i)^2 \exp(t_0|\boldsymbol{v}'\boldsymbol{\zeta}_i|) < \infty$$

for some $t_0 > 0$. Let B_n be a sequence of positive numbers such that

$$\sup_{\|\boldsymbol{v}\|_2=1} \sum_{i=1}^{n} \mathbb{E}(\boldsymbol{v}'\boldsymbol{\zeta}_i)^2 \exp(t_0|\boldsymbol{v}'\boldsymbol{\zeta}_i|) \le B_n^2.$$

Then for $x > 0$ and $2\sqrt{p + x^2} \le B_n$, we have

$$\mathbb{P}\Big(\Big\|\sum_{i=1}^{n} \boldsymbol{\zeta}_i\Big\|_2 \ge C_{t_0} B_n \sqrt{p + x^2}\Big) \le e^{-x^2},$$

where C_{t_0} is a positive constant depending only on t_0.

Proof. Let $S_{1/2}^{p-1}$ be a $1/2$ net of the unit sphere S^{p-1} in the Euclidean distance in $\mathbb{R}^p$. By the proof of Lemma 3 in Cai et al. (2010), we have $d_p :=\text{Card}(S_{1/2}^{p-1}) \le 5^p$. So there exist d_p points $\boldsymbol{v}_1, ..., \boldsymbol{v}_{d_p}$ in S^{p-1} such that for any $\boldsymbol{v}$ in S^{p-1}, we have $\|\boldsymbol{v} - \boldsymbol{v}_j\|_2 \le 1/2$ for some j. Therefore, for any vector $\boldsymbol{u} \in \mathbb{R}^p$, $\|\boldsymbol{u}\|_2 \le \sup_{j \le d_p} |\boldsymbol{v}_j'\boldsymbol{u}| + \|\boldsymbol{u}\|_2/2$. That is, $\|\boldsymbol{u}\|_2 \le 2\sup_{j \le d_p} |\boldsymbol{v}_j'\boldsymbol{u}|$. Therefore,

$$\begin{aligned}
\mathbb{P}\Big(\Big\|\sum_{i=1}^{n} \boldsymbol{\zeta}_i\Big\|_2 \ge C_{t_0} B_n \sqrt{p + x^2}\Big) &\le \mathbb{P}\Big(\sup_{j \le d_p} \Big|\sum_{i=1}^{n} \boldsymbol{v}_j'\boldsymbol{\zeta}_i\Big| \ge 2^{-1} C_{t_0} B_n \sqrt{p + x^2}\Big) \\
&\le 5^p \max_j \mathbb{P}\Big(\Big|\sum_{i=1}^{n} \boldsymbol{v}_j'\boldsymbol{\zeta}_i\Big| \ge 2^{-1} C_{t_0} B_n \sqrt{p + x^2}\Big) \\
&\le e^{-x^2},
\end{aligned}$$

where we let $C_{t_0} = 4(t_0 + t_0^{-1})$. The last inequality follows from Lemma 1 in Cai and Liu (2011), by noting that $2\sqrt{p + x^2} \le B_n$. □

Let $h(\boldsymbol{u}, \boldsymbol{\xi})$ be a q-dimensional random vector with zero mean. For some constant $c_4 > 0$, define

$$\Theta_0 = \{\boldsymbol{u} \in \mathbb{R}^q : \|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \le c_4\}, \tag{27}$$

where $\boldsymbol{u}_0$ is a point in $\mathbb{R}^q$. Assume the following conditions hold.

(B1). $\mathbb{E}\sup_{\boldsymbol{u} \in \Theta_0} \|h(\boldsymbol{u}, \boldsymbol{\xi})\|_2 \le q^c$ for some $c > 0$.

(B2). For $\boldsymbol{u} \in \Theta_0$, assume $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}(\boldsymbol{v}'h(\boldsymbol{u}, \boldsymbol{\xi}))^2 \le b(\boldsymbol{u})$ and $b(\boldsymbol{u})$ satisfies $|b(\boldsymbol{u}_1) - b(\boldsymbol{u}_2)| \le q^c\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2^\gamma$ for some $c, \gamma > 0$, uniformly in $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \Theta_0$.

(B3). Assume that for some $t_0 > 0$ and $0 \le \alpha \le 1$,

$$\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}(\boldsymbol{v}'h(\boldsymbol{u},\boldsymbol{\xi}))^2 \exp\Big(t_0\Big|\frac{\boldsymbol{v}'h(\boldsymbol{u},\boldsymbol{\xi})}{b^{\alpha/2}(\boldsymbol{u})}\Big|\Big) \le Cb(\boldsymbol{u})$$

for some constant $C > 0$, uniformly in $\boldsymbol{u} \in \Theta_0$.

(B4). $\mathbb{E}\sup_{\boldsymbol{u}_1,\boldsymbol{u}_2\in\Theta_0,\|\boldsymbol{u}_1-\boldsymbol{u}_2\|_2\le n^{-M}} \|h(\boldsymbol{u}_1,\boldsymbol{\xi}) - h(\boldsymbol{u}_2,\boldsymbol{\xi})\|_2 \le q^{c_2}n^{-c_3M}$ for any $M \ge M_0$ with some $M_0 > 0$ and some $c_2, c_3 > 0$.

(B4*) We have

$$\sup_{\boldsymbol{u}_1\in\Theta_0} \mathbb{E} \sup_{\boldsymbol{u}_2\in\Theta_0:\|\boldsymbol{u}_1-\boldsymbol{u}_2\|_2\le n^{-M}} \Big\|\frac{h(\boldsymbol{u}_1,\boldsymbol{\xi}) - h(\boldsymbol{u}_2,\boldsymbol{\xi})}{b^{\alpha/2}(\boldsymbol{u}_2)}\Big\|_2^4 \le q^{c_2}n^{-c_3M}$$

for some $c_2, c_3 > 0$, and

$$\sup_{\boldsymbol{u}_1\in\Theta_0} \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \sup_{\boldsymbol{u}_2\in\Theta_0:\|\boldsymbol{u}_1-\boldsymbol{u}_2\|_2\le n^{-M}} \exp\Big(t_0\Big|\frac{\boldsymbol{v}'[h(\boldsymbol{u}_1,\boldsymbol{\xi}) - h(\boldsymbol{u}_2,\boldsymbol{\xi})]}{b^{\alpha/2}(\boldsymbol{u}_2)}\Big|\Big) \le C$$

for any $M \ge M_0$ with some $M_0 > 0$ and some $t_0, C > 0$.

Lemma A.2. *Let $1 \le m \le n$ and $q \le n$. Assume (B1)-(B3) and (B4) (or (B4*)) hold. For any $\gamma_1, \gamma_2 > 0$, there exists a constant $c > 0$ such that*

$$\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_0} \frac{\big\|\frac{1}{m}\sum_{i\in B_t} h(\boldsymbol{\theta},\boldsymbol{\xi}_i)\big\|_2}{\sqrt{b(\boldsymbol{\theta}) + b^{\alpha}(\boldsymbol{\theta})(q\log n)/m + n^{-\gamma_2}}} \ge c\sqrt{\frac{q\log n}{m}}\Big) = O(n^{-\gamma_1}).$$

Proof. Since B_t and $\{\boldsymbol{\xi}_i\}$ are independent, without loss of generality, we can assume that B_t is a fixed set. Let $\{\boldsymbol{\theta}_1 ..., \boldsymbol{\theta}_{s_q}\}$ be s_q points such that for any $\boldsymbol{\theta} \in \Theta_0$, we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}_j\|_2 \le n^{-M}$ for sufficiently large M and some j. It is easy to prove that $s_q \le Cq^{q/2}n^{qM} \le Cn^{2qM}$ for some $C > 0$. For notation briefness, let $\widetilde{b}(\boldsymbol{\theta}) = b(\boldsymbol{\theta}) + b^{\alpha}(\boldsymbol{\theta})(q\log n)/m + n^{-\gamma_2}$. We have

$$\begin{aligned}\frac{\sum_{i\in B_t} h(\boldsymbol{\theta},\boldsymbol{\xi}_i)}{\sqrt{\widetilde{b}(\boldsymbol{\theta})}} - \frac{\sum_{i\in B_t} h(\boldsymbol{\theta}_j,\boldsymbol{\xi}_i)}{\sqrt{\widetilde{b}(\boldsymbol{\theta}_j)}} &= \sum_{i\in B_t} h(\boldsymbol{\theta},\boldsymbol{\xi}_i) \times \frac{\sqrt{\widetilde{b}(\boldsymbol{\theta}_j)} - \sqrt{\widetilde{b}(\boldsymbol{\theta})}}{\sqrt{\widetilde{b}(\boldsymbol{\theta})\widetilde{b}(\boldsymbol{\theta}_j)}} \\ &\quad + \frac{1}{\sqrt{\widetilde{b}(\boldsymbol{\theta}_j)}} \times \Big(\sum_{i\in B_t} h(\boldsymbol{\theta},\boldsymbol{\xi}_i) - \sum_{i\in B_t} h(\boldsymbol{\theta}_j,\boldsymbol{\xi}_i)\Big) \\ &=: \Gamma_1 + \Gamma_2.\end{aligned}$$

By (B1), we can obtain hat

$$\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta_0} \Big\|\sum_{i\in B_t} h(\boldsymbol{\theta},\boldsymbol{\xi}_i)\Big\|_2 = O(n^c)$$

for some $c > 0$. By (B2), we can show that $|\widetilde{b}(\boldsymbol{\theta}) - \widetilde{b}(\boldsymbol{\theta}_j)| \leq Cn^{c-\alpha'\gamma M}$ for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_j\|_2 \leq n^{-M}$, uniformly in j, where $\alpha' = 1$ if $\alpha = 0$ and $\alpha' = \alpha$ if $\alpha > 0$. Therefore

$$\max_j \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} \frac{\left|\sqrt{\widetilde{b}(\boldsymbol{\theta}_j)} - \sqrt{\widetilde{b}(\boldsymbol{\theta})}\right|}{\sqrt{\widetilde{b}(\boldsymbol{\theta})\widetilde{b}(\boldsymbol{\theta}_j)}} \leq Cn^{c+2\gamma_2-\gamma\alpha' M}.$$

This implies that $\mathbb{E}\max_j \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} \|\Gamma_1\|_2 = O(n^{2c+2\gamma_2-\gamma\alpha' M})$.

We first consider the case that (B4) holds. Then $\mathbb{E}\max_j \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} \|\Gamma_2\|_2 = O(n^{\gamma_2/2+1+c_2-c_3 M})$. Hence, by Markov's inequality, for any $\gamma_1 > 0$, by letting M be sufficiently large, we have

$$\mathbb{P}\Big(\max_j \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} \Big\| \frac{\frac{1}{m}\sum_{i\in B_t}(h(\boldsymbol{\theta}, \boldsymbol{\xi}_i) - h(\boldsymbol{\theta}_j, \boldsymbol{\xi}_i))}{\sqrt{\widetilde{b}(\boldsymbol{\theta}_j)}} \Big\|_2 \geq c\sqrt{\frac{q\log n}{m}}\Big)$$
$$= O(n^{-\gamma_1}). \tag{28}$$

We next prove (28) under (B4*). By the proof of Lemma A.1, we have

$$\begin{aligned}\Big\|\sum_{i\in B_t}(h(\boldsymbol{\theta}, \boldsymbol{\xi}_i) - h(\boldsymbol{\theta}_j, \boldsymbol{\xi}_i))\Big\|_2 &\leq 2\max_{1\leq l\leq d_q}\Big|\boldsymbol{v}_l'\sum_{i\in B_t}(h(\boldsymbol{\theta}, \boldsymbol{\xi}_i) - h(\boldsymbol{\theta}_j, \boldsymbol{\xi}_i))\Big| \\ &\leq 2\max_{1\leq l\leq d_q}\Big|\sum_{i\in B_t}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} |\boldsymbol{v}_l' H(\boldsymbol{\theta}, \boldsymbol{\theta}_j, \boldsymbol{\xi}_i)|\Big|,\end{aligned}$$

where $H(\boldsymbol{\theta}, \boldsymbol{\theta}_j, \boldsymbol{\xi}) = h(\boldsymbol{\theta}, \boldsymbol{\xi}) - h(\boldsymbol{\theta}_j, \boldsymbol{\xi})$. It is easy to see from (B4*) that, for sufficiently large M,

$$\max_j \max_{1\leq l\leq s_q} \frac{\left|\sum_{i\in B_t}\mathbb{E}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} |\boldsymbol{v}_l' H(\boldsymbol{\theta}, \boldsymbol{\theta}_j, \boldsymbol{\xi}_i)|\right|}{\sqrt{\widetilde{b}(\boldsymbol{\theta}_j)}} = o(\sqrt{\frac{q\log n}{m}}).$$

Set $\mathcal{H}_{l,j}(\boldsymbol{\xi}_i) = \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} |\boldsymbol{v}_l' H(\boldsymbol{\theta}, \boldsymbol{\theta}_j, \boldsymbol{\xi}_i)|/b^{\alpha/2}(\boldsymbol{\theta}_j)$. By (B4*) and Holder's inequality, we have

$$\max_j \sum_{i\in B_t}\mathbb{E}(\mathcal{H}_{l,j}(\boldsymbol{\xi}_i))^2\exp(t_0\mathcal{H}_{l,j}(\boldsymbol{\xi}_i)/2) \leq mq^{c_2/2}n^{-c_3 M/2}.$$

We now take $B_n^2 = c_5 m\widetilde{b}(\boldsymbol{\theta}_j)/b^{\alpha}(\boldsymbol{\theta}_j)$ and $x^2 = c_5 q\log n$ in Lemma 1 in Cai and Liu (2011), noting that $mq^{c_2/2}n^{-c_3 M/2} \leq B_n^2$ and $x^2 \leq B_n^2$, we have for any $\gamma, M > 0$, there exist $c, c_5 > 0$ such that uniformly in j,

$$\mathbb{P}\Big(\Big|\frac{\sum_{i\in B_t}[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} |\boldsymbol{v}_l' H(\boldsymbol{\theta}, \boldsymbol{\theta}_j, \boldsymbol{\xi}_i)| - \mathbb{E}\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_2 \leq n^{-M}} |\boldsymbol{v}_l' H(\boldsymbol{\theta}, \boldsymbol{\theta}_j, \boldsymbol{\xi}_i)|]}{m\sqrt{\widetilde{b}(\boldsymbol{\theta}_j)}} \geq c\sqrt{\frac{q\log n}{m}}\Big)$$
$$= O(n^{-\gamma q}).$$

This proves (28) under (B4*) by noting that $s_q = O(n^{2qM})$ and $d_q \leq 5^q$.

Now it suffices to show that

$$\mathbb{P}\Big(\max_j \frac{\left\|\frac{1}{m}\sum_{i\in B_t}h(\boldsymbol{\theta}_j, \boldsymbol{\xi}_i)\right\|_2}{\sqrt{\widetilde{b}(\boldsymbol{\theta}_j)}} \geq c\sqrt{\frac{q\log n}{m}}\Big) = O(n^{-\gamma_1}). \tag{29}$$

Let $\boldsymbol{\zeta}_i = h(\boldsymbol{\theta}_j, \boldsymbol{\xi}_i)/b^{\alpha/2}(\boldsymbol{\theta}_j)$, $i \in B_t$. By (B3), it is easy to see that

$$\sup_{\|\boldsymbol{v}\|_2=1} \sum_{i\in B_t} \mathbb{E}(\boldsymbol{v}'\boldsymbol{\zeta}_i)^2 \exp(t_0|\boldsymbol{v}'\boldsymbol{\zeta}_i|) \le C_2 m[b(\boldsymbol{\theta}_j)]^{1-\alpha}$$

for some $C_2 > 0$. Take $x = \sqrt{(\gamma_1 + 2M)q\log n}$ and

$$\begin{aligned} B_n^2 &= 4(C_2+\gamma_1+2M+1)\Big(m[b(\boldsymbol{\theta}_j)]^{1-\alpha} + q\log n + m(b(\boldsymbol{\theta}_j))^{-\alpha}n^{-\gamma_2}\Big) \\ &= 4(C_2+\gamma_1+2M+1)m\widetilde{b}(\boldsymbol{\theta}_j)/b^{\alpha}(\boldsymbol{\theta}_j). \end{aligned}$$

Note that $2\sqrt{q+x^2} \le B_n$. By Lemma A.1, we obtain (29) by letting c be sufficiently large. □

Let $\bar{g}(\boldsymbol{\theta},\boldsymbol{\xi}) = g(\boldsymbol{\theta},\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\theta},\boldsymbol{\xi})$. For some $c_4 > 0$, define

$$\begin{aligned} \mathcal{C}_t &= \Big\{\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le c_4} \Big\|\frac{1}{m}\sum_{i\in B_t}\bar{g}(\boldsymbol{\theta},\boldsymbol{\xi}_i)\Big\|_2 \le c\sqrt{\frac{p\log n}{m}}\Big\}, \\ \mathcal{C} &= \Big\{\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\le c_4} \Big\|\frac{1}{n}\sum_{i=1}^{n}\bar{g}(\boldsymbol{\theta},\boldsymbol{\xi}_i)\Big\|_2 \le c\sqrt{\frac{p\log n}{n}}\Big\}, \end{aligned}$$

where c is sufficiently large.

Lemma A.3. *Under (C3) or (C3*) and $p\log n = o(m)$, for any $\gamma > 0$, there exists a constant $c_4 > 0$ such that*

$$\mathbb{P}(\mathcal{C}_t \cap \mathcal{C}) \ge 1 - O(n^{-\gamma}).$$

The same result holds with B_t being replaced by H_t.

Proof. In Lemma A.2, take $\boldsymbol{u} = \boldsymbol{\theta}$, $\boldsymbol{u}_0 = \boldsymbol{\theta}^*$, $q = p$, $\alpha = 0$ and $h(\boldsymbol{\theta},\boldsymbol{\xi}) = g(\boldsymbol{\theta},\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\theta},\boldsymbol{\xi})$. Then (C3) (or (C3*)) implies that (B1)-(B4) (or (B4*), respectively) hold with $\alpha = 0$, and $b(\boldsymbol{\theta}) = C$ for some large C. So we have

$$\mathbb{P}(\mathcal{C}_t \cap \mathcal{C}) \ge 1 - O(n^{-\gamma})$$

for any large γ. □

Lemma A.4. *Suppose that $p \to \infty$, $r_i = c_0/\max(p, i^{\alpha})$ for $c_0 > 0$ and $0 < \alpha \le 1$. Let $c > 0$, $\tau > 0$ and $d \ge 1$.*

(1) For a positive sequence $\{a_i\}$ that satisfies $a_i \le (1 - cr_i)a_{i-1} + r_i^d b_n$, $1 \le i \le n$, we have $a_i \le C(r_i^{d-1}b_n + i^{-\gamma})$ for any $\gamma > 0$ and all $i \ge p^{1/\alpha+\tau}$ by letting c_0 be sufficiently large.

(2) For a positive sequence $\{a_i\}$ that satisfies $a_i \ge (1 - cr_i)a_{i-1} + r_i^d b_n$, $1 \le i \le n$, we have $a_i \ge Cr_i^{d-1}b_n$ for all $i \ge p^{1/\alpha+\tau}$ by letting c_0 be sufficiently large.

Proof. We first prove the first claim. For $i \geq p^{1/\alpha+\tau}$, we have

$$
\begin{aligned}
a_i &\leq (1-cr_i)a_{i-1} + r_i^d b_n \\
&= a_0 \prod_{j=1}^{i}(1-cr_j) + b_n \sum_{k=1}^{i} r_k^d \prod_{j=k}^{i-1}(1-cr_{j+1}) \\
&\leq a_0 \exp\big(-c\sum_{j=1}^{i} r_j\big) + b_n \sum_{k=1}^{i} r_k^d \exp\big(-c\sum_{j=k}^{i-1} r_{j+1}\big) \\
&\leq a_0 \exp\Big(-\widetilde{c}\big(p_\alpha/p + \frac{1}{2}\int_{p_\alpha}^{i}\frac{1}{x^\alpha}\mathrm{d}x\big)\Big) + b_n \sum_{k=p_\alpha+1}^{i} r_k^d \exp\Big(-\frac{\widetilde{c}}{2}\int_{k}^{i}\frac{1}{x^\alpha}\mathrm{d}x\Big) \\
&+ b_n \sum_{k=1}^{p_\alpha} r_k^d \exp\Big(-\widetilde{c}\big(\frac{p_\alpha - k}{p} + \frac{1}{2}\int_{p_\alpha}^{i}\frac{1}{x^\alpha}\mathrm{d}x\big)\Big) \\
&= a_0 \exp\Big(-\widetilde{c}\big(p_\alpha/p + \frac{1}{2}\int_{p_\alpha}^{i}\frac{1}{x^\alpha}\mathrm{d}x\big)\Big) + c_0^d b_n \sum_{k=p_\alpha+1}^{i} k^{-\alpha d} \exp\Big(-\frac{\widetilde{c}}{2}\int_{k}^{i}\frac{1}{x^\alpha}\mathrm{d}x\Big) \\
&+ c_0^d b_n \sum_{k=1}^{p_\alpha} p^{-d} \exp\Big(-\widetilde{c}\big(\frac{p_\alpha - k}{p} + \frac{1}{2}\int_{p_\alpha}^{i}\frac{1}{x^\alpha}\mathrm{d}x\big)\Big),
\end{aligned} \tag{30}
$$

where $p_\alpha = \lfloor p^{1/\alpha} \rfloor$, $\widetilde{c} = c_0 c$, and α, c_0 are defined in the step-size r_i.

When $\alpha = 1$, we have

$$
\begin{aligned}
(30) \quad &= \quad \frac{a_0 p^{\widetilde{c}/2} \mathrm{e}^{-\widetilde{c}}}{i^{\widetilde{c}/2}} + c_0^d b_n \sum_{k=p+1}^{i} \frac{k^{\widetilde{c}/2-d}}{i^{\widetilde{c}/2}} + c_0^d b_n \sum_{k=1}^{p} \frac{p^{\widetilde{c}/2-d}\exp(\widetilde{c}k/p - \widetilde{c})}{i^{\widetilde{c}/2}} \\
&\leq \quad \frac{a_0 p^{\widetilde{c}/2} \mathrm{e}^{-\widetilde{c}}}{i^{\widetilde{c}/2}} + c_0^d b_n i^{1-d} + \frac{c_0^d b_n p^{\widetilde{c}/2-d+1}}{i^{\widetilde{c}/2}} \\
&\leq \quad C(r_i^{d-1} b_n + i^{-\gamma}),
\end{aligned}
$$

when c_0 is large enough such that $\widetilde{c} = c_0 c \geq 2\max(d,\gamma)(1+1/\tau)$.

When $\alpha < 1$, for any $\kappa > 0$ and $1 \leq u < i$, we have

$$
\begin{aligned}
&\int_u^i x^{-\alpha d} \exp\big(\frac{\kappa x^{1-\alpha}}{1-\alpha}\big)\mathrm{d}x \\
&\qquad = \frac{1}{\kappa} x^{-\alpha d+\alpha} \exp\big(\frac{\kappa x^{1-\alpha}}{1-\alpha}\big)\Big|_u^i - \int_u^i \frac{\alpha - \alpha d}{\kappa} x^{-\alpha d+\alpha-1} \exp\big(\frac{\kappa x^{1-\alpha}}{1-\alpha}\big)\mathrm{d}x \\
&\qquad \leq \frac{1}{\kappa} x^{-\alpha d+\alpha} \exp\big(\frac{\kappa x^{1-\alpha}}{1-\alpha}\big)\Big|_u^i + u^{\alpha-1}\int_u^i \frac{\alpha(d-1)}{\kappa} x^{-\alpha d} \exp\big(\frac{\kappa x^{1-\alpha}}{1-\alpha}\big)\mathrm{d}x.
\end{aligned}
$$

Therefore, we have

$$
\int_u^i x^{-\alpha d} \exp\big(\frac{\kappa x^{1-\alpha}}{1-\alpha}\big)\mathrm{d}x \leq \frac{1}{\kappa - \alpha d + \alpha} x^{-\alpha d+\alpha} \exp\big(\frac{\kappa x^{1-\alpha}}{1-\alpha}\big)\Big|_u^i \tag{31}
$$

for $\kappa > \alpha(d-1)$. By (31), we have for $i \geq p^{1/\alpha+\tau}$,

$$\begin{aligned}
(30) &= a_0 \exp\Big(-\widetilde{c}\big(\frac{p_\alpha}{p} + \frac{i^{1-\alpha} - p_\alpha^{1-\alpha}}{2-2\alpha}\big)\Big) + c_0^d b_n \sum_{k=p_\alpha+1}^{i} k^{-\alpha d} \exp\Big(-\frac{\widetilde{c}\big(i^{1-\alpha} - k^{1-\alpha}\big)}{2-2\alpha}\Big) \\
&\quad + c_0^d b_n \sum_{k=1}^{p_\alpha} p^{-d} \exp\Big(-\widetilde{c}\big(\frac{p_\alpha - k}{p} + \frac{i^{1-\alpha} - p_\alpha^{1-\alpha}}{2-2\alpha}\big)\Big) \\
&\leq a_0 \exp\Big(-\frac{\widetilde{c}(i^{1-\alpha} - p_\alpha^{1-\alpha})}{2-2\alpha}\Big) + c_0^d b_n i^{-\alpha d} \\
&\quad + c_0^d b_n \exp\Big(-\frac{\widetilde{c} i^{1-\alpha}}{2-2\alpha}\Big) \int_{p_\alpha+1}^{i} x^{-\alpha d} \exp\Big(\frac{\widetilde{c} x^{1-\alpha}}{2-2\alpha}\Big) \mathrm{d}x \\
&\quad + c_0^d b_n p_\alpha p^{-d} \exp\Big(-\frac{\widetilde{c}(i^{1-\alpha} - p_\alpha^{1-\alpha})}{2-2\alpha}\big) \\
&\leq a_0 \exp\big(-\frac{\widetilde{c}(i^{1-\alpha} - p_\alpha^{1-\alpha})}{2-2\alpha}\Big) + c_0^d b_n i^{-\alpha d} \\
&\quad + c_0^d b_n \Big(\frac{i^{-\alpha(d-1)}}{\widetilde{c}/2 - \alpha d + \alpha} + p_\alpha p^{-d} \exp\big(-\frac{\widetilde{c}(i^{1-\alpha} - p_\alpha^{1-\alpha})}{2-2\alpha}\big)\Big) \\
&\leq C(r_i^{d-1} b_n + i^{-\gamma})
\end{aligned}$$

for large enough c_0 such that $\widetilde{c} > 2\alpha(d-1)$.

To prove the second claim, we first recall that $p \to \infty$ and $\sup_{i\geq 1} r_i = o(1)$. Hence $1 - cr_j \geq \exp(-2cr_j)$ for all j. Then

$$\begin{aligned}
a_i &\geq a_0 \exp\big(-2c \sum_{j=1}^{i} r_j\big) + b_n \sum_{k=1}^{i} r_k^d \exp\big(-2c \sum_{j=k}^{i-1} r_{j+1}\big) \\
&\geq b_n \sum_{k=1}^{i} r_k^d \exp\Big(-2\widetilde{c} \int_k^i \frac{1}{x^\alpha} \mathrm{d}x\Big). \qquad (32)
\end{aligned}$$

When $\alpha = 1$, we have

$$(32) \geq c_0^d b_n i^{-2\widetilde{c}} \sum_{k=p+1}^{i} k^{2\widetilde{c}-d} \geq \frac{c_0^d b_n (i^{-d+1} - i^{-2\widetilde{c}} p^{2\widetilde{c}-d+1})}{2\widetilde{c} - d + 1} \geq c_1 r_i^{d-1} b_n,$$

for $2\widetilde{c} > d-1$ and $i \geq p^{1+\tau}$.

When $\alpha < 1$, we have for $i \geq p^{1/\alpha+\tau}$,

$$\begin{aligned}
(32) &\geq c_0^d b_n \sum_{k=p_\alpha+1}^{i} k^{-\alpha d} \exp\Big(-\frac{2\widetilde{c}\big(i^{1-\alpha} - k^{1-\alpha}\big)}{1-\alpha}\Big) \\
&\geq c_0^d b_n \exp\Big(-\frac{2\widetilde{c} i^{1-\alpha}}{1-\alpha}\Big) \int_{p_\alpha}^{i} x^{-\alpha d} \exp\Big(\frac{2\widetilde{c} x^{1-\alpha}}{1-\alpha}\Big) \mathrm{d}x \\
&\geq \frac{c_0^d b_n}{2\widetilde{c}} x^{-\alpha d+\alpha} \exp\big(\frac{2\widetilde{c}(x^{1-\alpha} - i^{1-\alpha})}{1-\alpha}\big)\Big|_{p_\alpha}^{i} \\
&= \frac{1}{2c} r_i^{d-1} b_n - \frac{c_0^d}{2\widetilde{c}} p_\alpha^{-\alpha d+\alpha} b_n \exp\big(\frac{\widetilde{c}(p_\alpha^{1-\alpha} - i^{1-\alpha})}{1-\alpha}\big)
\end{aligned}$$

$$\geq \quad Cr_i^{d-1}b_n.$$

The proof is complete. □

B Proofs for results of DC-SGD in Section 4.1

B.1 Proof of Theorem 4.1

By Condition (C1), we have $\sqrt{(p\log n)/m} \to 0$. Without loss of generality, we can assume that $\sqrt{(p\log n)/m} = o(d_n)$. Let $\boldsymbol{\delta}_i = \boldsymbol{z}_i - \boldsymbol{\theta}^*$ and $\bar{g}(\boldsymbol{\theta},\boldsymbol{\xi}) = g(\boldsymbol{\theta},\boldsymbol{\xi}) - \mathbb{E}g(\boldsymbol{\theta},\boldsymbol{\xi})$. Define

$$\Theta_0 = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq c_4\},$$

where c_4 is given in (27). Define the events $\mathcal{F}_i = \{\|\boldsymbol{\delta}_{i-1}\|_2 \leq d_n\}$, and

$$\mathcal{C}_i \quad = \quad \Big\{\sup_{\boldsymbol{\theta}\in\Theta_0} \Big\|\frac{1}{m}\sum_{j\in H_i} \bar{g}(\boldsymbol{\theta},\boldsymbol{\xi}_j)\Big\|_2 \leq C\sqrt{\frac{p\log n}{m}}\Big\},$$

where C is sufficiently large. From the SGD updating rule (6), we have

$$\|\boldsymbol{\delta}_i\|_2^2 = \|\boldsymbol{\delta}_{i-1}\|_2^2 - 2\frac{r_i}{m}\sum_{j\in H_i}\boldsymbol{\delta}_{i-1}' g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j) + \Big\|\frac{r_i}{m}\sum_{j\in H_i} g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\Big\|_2^2. \tag{33}$$

Since $G(\boldsymbol{\theta}) = \mathbb{E}g(\boldsymbol{\theta},\boldsymbol{\xi})$ and $G(\boldsymbol{\theta}^*) = 0$ by (15) and (C2),

$$\begin{aligned}
\frac{r_i}{m}\sum_{j\in H_i}\boldsymbol{\delta}_{i-1}' g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j) &= r_i\boldsymbol{\delta}_{i-1}' G(\boldsymbol{z}_{i-1}) + \frac{r_i}{m}\sum_{j\in H_i}\boldsymbol{\delta}_{i-1}'\bar{g}(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\\
&= r_i\boldsymbol{\delta}_{i-1}'\Big(G(\boldsymbol{z}_{i-1}) - G(\boldsymbol{\theta}^*)\Big) + \frac{r_i}{m}\sum_{j\in H_i}\boldsymbol{\delta}_{i-1}'\bar{g}(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\\
&\geq c_1 r_i\|\boldsymbol{\delta}_{i-1}\|_2^2 - C_1 r_i\|\boldsymbol{\delta}_{i-1}\|_2^3 - r_i\big\|\boldsymbol{\delta}_{i-1}\big\|_2\Big\|\frac{1}{m}\sum_{j\in H_i}\bar{g}(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\Big\|_2.
\end{aligned}$$

Similarly,

$$\frac{r_i}{m}\sum_{j\in H_i} g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j) \quad = \quad r_i\big(G(\boldsymbol{z}_{i-1}) - G(\boldsymbol{\theta}^*)\big) + \frac{r_i}{m}\sum_{j\in H_i}\bar{g}(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)$$

and

$$\begin{aligned}
\Big\|\frac{r_i}{m}\sum_{j\in H_i} g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\Big\|_2^2 &\leq 2r_i^2\|G(\boldsymbol{z}_{i-1}) - G(\boldsymbol{\theta}^*)\|_2^2 + 2\Big\|\frac{r_i}{m}\sum_{j\in H_i}\bar{g}(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\Big\|_2^2\\
&\leq r_i^2 c_1^{-2}\|\boldsymbol{\delta}_{i-1}\|_2^2 + C_1^2 r_i^2\|\boldsymbol{\delta}_{i-1}\|_2^4 + 2\Big\|\frac{r_i}{m}\sum_{j\in H_i}\bar{g}(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\Big\|_2^2.
\end{aligned}$$

Therefore, on $\mathcal{C}_i \cap \mathcal{F}_i$, since $\sup_{i\geq 1} r_i = o(1)$,

$$\|\boldsymbol{\delta}_i\|_2^2 \quad \leq \quad (1 - c_1 r_i)\|\boldsymbol{\delta}_{i-1}\|_2^2 + C\Big(r_i d_n\sqrt{\frac{p\log n}{m}} + r_i^2\frac{p\log n}{m} + r_i d_n^3 + r_i^2 d_n^2 + r_i^2 d_n^4\Big)$$

$$\leq \quad (1 - c_1 r_i/2)d_n^2 + Cr_i \frac{p\log n}{m},$$

where we used the inequality $d_n\sqrt{\frac{p\log n}{m}} \leq td_n^2 + t^{-1}(p\log n)/m$ for any small $t > 0$. Note that $\sqrt{(p\log n)/m} = o(d_n)$. Therefore, $\mathcal{F}_i \cap \mathcal{C}_i \subset \{\|\boldsymbol{\delta}_i\|_2 \leq d_n\} = \mathcal{F}_{i+1}$. Combining the above arguments for $j = 1, 2, \ldots, i$, on the event $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\} \cap (\cap_{k=1}^i \mathcal{C}_k)$, we have $\max_{1\leq j\leq i} \|\boldsymbol{\delta}_j\|_2 \leq d_n$.

We now assume $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}\sup_{\boldsymbol{\theta}} \exp(t_0|\boldsymbol{v}'g(\boldsymbol{\theta}, \boldsymbol{\xi})|) \leq C$ (by Condition (C3) bullet 1 or (C3*)). We have

$$\|\boldsymbol{\delta}_i\|_2 \leq \|\boldsymbol{\delta}_{i-1}\|_2 + \frac{r_i}{m}\sum_{j\in H_i}\sup_{\boldsymbol{\theta}} \|g(\boldsymbol{\theta}, \boldsymbol{\xi}_j)\|_2 \leq C\frac{1}{m}\sum_{j=1}^{n}\sup_{\boldsymbol{\theta}} \|g(\boldsymbol{\theta}, \boldsymbol{\xi}_j)\|_2.$$

Thus $\mathbb{E}_0\|\boldsymbol{\delta}_i\|_2^6 \leq Cn^6$ and $\mathbb{E}_0\|\boldsymbol{\delta}_i\|_2^8 \leq Cn^8$. Recall that $\mathbb{E}_0(\cdot)$ is denoted by the expectation to $\{\boldsymbol{\xi}_i\}$ given the initial estimator $\widehat{\boldsymbol{\theta}}_0$. By $G(\boldsymbol{\theta}^*) = 0$,

$$\mathbb{E}_0\Big[\frac{1}{m}\sum_{j\in H_i}\boldsymbol{\delta}_{i-1}' g(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\Big] = \mathbb{E}_0\big[\boldsymbol{\delta}_{i-1}' G(\boldsymbol{z}_{i-1})\big] = \mathbb{E}_0\big[\boldsymbol{\delta}_{i-1}'\boldsymbol{\Sigma}(\widetilde{\boldsymbol{z}}_{i-1})\boldsymbol{\delta}_{i-1}\big],$$

where $\widetilde{\boldsymbol{z}}_{i-1} = \widetilde{\alpha}\boldsymbol{z}_{i-1} + (1-\widetilde{\alpha})\boldsymbol{\theta}^*$ for some $\widetilde{\alpha} \in (0,1)$. Then on the event $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$, by (C2), Lemma A.3, and $\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^6 \leq Cn^6$,

$$\begin{aligned}
\mathbb{E}_0\big[\boldsymbol{\delta}_{i-1}'\boldsymbol{\Sigma}(\widetilde{\boldsymbol{z}}_{i-1})\boldsymbol{\delta}_{i-1}\big] \quad &\geq \quad \mathbb{E}_0\big[\boldsymbol{\delta}_{i-1}'\boldsymbol{\Sigma}\boldsymbol{\delta}_{i-1}\big] - C_1 d_n \mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 I\{\cap_{k=1}^i \mathcal{C}_k\} \\
&\quad -\mathbb{E}_0\big[\big|\boldsymbol{\delta}_{i-1}'\big(\boldsymbol{\Sigma}(\widetilde{\boldsymbol{z}}_{i-1}) - \boldsymbol{\Sigma}\big)\boldsymbol{\delta}_{i-1}\big|\big] I\{\{\cap_{k=1}^i \mathcal{C}_k\}^c\} \\
&\geq \quad c_1\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 - C_1 d_n \mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 - C_1\mathbb{E}_0\big[\|\boldsymbol{\delta}_{i-1}\|_2^3 I\{\{\cap_{k=1}^i \mathcal{C}_i\}^c\}\big] \\
&\geq \quad 2^{-1}c_1\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 - C_1\mathbb{E}_0\big[\|\boldsymbol{\delta}_{i-1}\|_2^3 I\{\{\cap_{k=1}^i \mathcal{C}_i\}^c\}\big] \\
&\geq \quad 2^{-1}c_1\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 - Cn^{3-\gamma} \qquad (34)
\end{aligned}$$

for any $\gamma > 0$. Also, on the event $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$,

$$\begin{aligned}
\mathbb{E}_0\Big\|\frac{1}{m}\sum_{j\in H_i} g(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\Big\|_2^2 \quad &= \quad \mathbb{E}_0\|G(\boldsymbol{z}_{i-1})\|_2^2 + \mathbb{E}_0\Big\|\frac{1}{m}\sum_{j\in H_i}\bar{g}(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\Big\|_2^2 \\
&\leq \quad c_1^{-1}\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + C\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^4 + \frac{Cp}{m}. \qquad (35)
\end{aligned}$$

Moreover, on the event $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$,

$$\begin{aligned}
\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^4 \quad &\leq \quad d_n^2\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + (\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^8 \cdot \mathbb{P}_0\{\|\boldsymbol{\delta}_{i-1}\|_2 > d_n\})^{1/2} \\
&\leq \quad d_n^2\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + (\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^8 \cdot \mathbb{P}(\cup_{k=1}^{i-1}\mathcal{C}_k^c))^{1/2} \\
&\leq \quad d_n^2\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + Cn^{4-\gamma}
\end{aligned}$$

for any $\gamma > 0$. Therefore, on the event $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$,

$$\mathbb{E}_0\big[\|\boldsymbol{\delta}_i\|_2^2\big] \quad \leq \quad (1 - cr_i/2)\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + Cr_i^2\frac{p}{m}, \qquad (36)$$

which by Lemma A.4 implies that for any $\tau > 0$, $\gamma > 0$ and all $i \geq p^{1/\alpha+\tau}$, $\mathbb{E}_0\|\boldsymbol{\delta}_i\|_2^2 \leq C_1(p/(i^\alpha m) + i^{-\gamma})$. By (C1), we have $s = n/m \geq p^{1/\alpha+\tau_2}$. That is, $\mathbb{E}_0\|\boldsymbol{\delta}_s\|_2^2 \leq C_1 p/(n^\alpha m^{1-\alpha})$.

Now consider the setting that Condition (C3) holds with bullet 2: $c_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta})) \leq \lambda_{\max}(\boldsymbol{\Sigma}(\boldsymbol{\theta})) \leq c_1^{-1}$ uniformly in $\boldsymbol{\theta}$. Then we have $c_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}(\widetilde{\boldsymbol{z}}_{i-1})) \leq \lambda_{\max}(\boldsymbol{\Sigma}(\widetilde{\boldsymbol{z}}_{i-1})) \leq c_1^{-1}$. Therefore

$$\mathbb{E}_0\Big[\frac{1}{m}\sum_{j\in H_i}\boldsymbol{\delta}_{i-1}' g(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\Big] = \mathbb{E}_0\big[\boldsymbol{\delta}_{i-1}'\boldsymbol{\Sigma}(\widetilde{\boldsymbol{z}}_{i-1})\boldsymbol{\delta}_{i-1}\big] \geq c_1\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2.$$

Also by (C3),

$$\begin{aligned}
\mathbb{E}_0\big\|\bar{g}(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\big\|_2^2 &= \mathbb{E}_0\Big(\mathbb{E}_0\big[\|\bar{g}(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\|_2^2 | \boldsymbol{z}_{i-1}\big]\Big)\\
&\leq 2\mathbb{E}_0\Big(\mathbb{E}_0\big[\big\|\bar{g}(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j) - \bar{g}(\boldsymbol{\theta}^*, \boldsymbol{\xi}_j)\big\|_2^2 | \boldsymbol{z}_{i-1}\big]\Big) + Cp\\
&\leq Cp\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + Cp.
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{E}_0\Big\|\frac{1}{m}\sum_{j\in H_i} g(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\Big\|_2^2 &\leq 2\mathbb{E}_0\|G(\boldsymbol{z}_{i-1})\|_2^2 + 2\mathbb{E}_0\Big\|\frac{1}{m}\sum_{j\in H_i}\bar{g}(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\Big\|_2^2\\
&\leq C\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + C\frac{p}{m}\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + C\frac{p}{m}.
\end{aligned}$$

That is, (36) still holds and $\mathbb{E}_0\|\boldsymbol{\delta}_s\|_2^2 \leq Cp/(n^\alpha m^{1-\alpha})$.

We now consider the bias of $\mathbb{E}_0\boldsymbol{z}_i$. We have

$$\mathbb{E}_0\Big[\frac{1}{m}\sum_{j\in H_i} g(\boldsymbol{z}_{i-1}, \boldsymbol{\xi}_j)\Big] = \mathbb{E}_0\Big(G(\boldsymbol{z}_{i-1}) - G(\boldsymbol{\theta}^*)\Big) = \boldsymbol{\Sigma}\mathbb{E}_0\boldsymbol{\delta}_{i-1} + \mathbb{E}_0(\boldsymbol{\Sigma}(\boldsymbol{z}_i') - \boldsymbol{\Sigma})\boldsymbol{\delta}_{i-1}$$

and $\|(\boldsymbol{\Sigma}(\boldsymbol{z}_i') - \boldsymbol{\Sigma})\boldsymbol{\delta}_{i-1}\|_2 \leq C\|\boldsymbol{\delta}_{i-1}\|_2^2$. Therefore, on the event $\{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \leq d_n\}$, for any $\tau > 0$, $0 < \mu < 1$ and $i \geq \max(p^{1/\alpha+\tau/2}, (n/m)^\mu)$,

$$\begin{aligned}
\|\mathbb{E}_0\boldsymbol{\delta}_i\|_2 &\leq \|\boldsymbol{I} - r_i\boldsymbol{\Sigma}\|\|\mathbb{E}_0\boldsymbol{\delta}_{i-1}\|_2 + Cr_i\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2\\
&\leq (1 - c_1 r_i)\|\mathbb{E}_0\boldsymbol{\delta}_{i-1}\|_2 + Cr_i\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2\\
&\leq (1 - c_1 r_i)\|\mathbb{E}_0\boldsymbol{\delta}_{i-1}\|_2 + Cr_i^2\frac{p}{m} + Ci^{-\gamma}\\
&\leq (1 - c_1 r_i)\|\mathbb{E}_0\boldsymbol{\delta}_{i-1}\|_2 + Cr_i^2\frac{p}{m},
\end{aligned}$$

by noting that $\gamma > 0$ can be arbitrarily large. Let $q_\alpha = \max(p^{1/\alpha+\tau/2}, (n/m)^\mu)$. Then for any $\gamma > 0$,

$$\begin{aligned}
\|\mathbb{E}_0\boldsymbol{\delta}_s\|_2 &\leq \prod_{j=q_\alpha+1}^{s}(1 - c_1 r_j)\|\mathbb{E}_0\boldsymbol{\delta}_{q_\alpha}\|_2 + \frac{p}{m}\sum_{k=q_\alpha+1}^{s} r_k^2\prod_{j=k}^{i-1}(1 - cr_{j+1})\\
&\leq C(q_\alpha/s)^{\widetilde{c}} + Cr_s\frac{p}{m} + Cs^{-\gamma},
\end{aligned}$$

where $\widetilde{c}$ is sufficiently large. Therefore, by the proof of Lemma A.4, $\|\mathbb{E}_0\boldsymbol{\delta}_s\|_2 \leq C_1 p/(n^\alpha m^{1-\alpha})$. □

B.2 Proof of Proposition 4.2

Since $\sup_{\|\boldsymbol{v}\|_2=1}\sup_{\boldsymbol{\theta}}\mathbb{E}(\boldsymbol{v}'g(\boldsymbol{\theta},\boldsymbol{\xi}))^2\le C$, by the independence between $\boldsymbol{\xi}_j$ and $\boldsymbol{z}_{i-1}$, we have

$$\mathbb{E}(\boldsymbol{\delta}_{i-1}'g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j))^2\le C\|\boldsymbol{\delta}_{i-1}\|_2^2.$$

By (33), we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\delta}_i\|_2^2 &\ge \mathbb{E}\|\boldsymbol{\delta}_{i-1}\|_2^2-\frac{2r_i}{m}\sum_{j\in H_i}\mathbb{E}\boldsymbol{\delta}_{i-1}'g(\boldsymbol{\theta}_{i-1},\boldsymbol{\xi}_j)\\
&\ge \mathbb{E}\|\boldsymbol{\delta}_{i-1}\|_2^2-Cr_i\sqrt{\mathbb{E}\|\boldsymbol{\delta}_{i-1}\|_2^2}\\
&\ge \min\Big((1-Cr_i/p^\nu)p^{2\nu},\mathbb{E}\|\boldsymbol{\delta}_{i-1}\|_2^2-Cr_ip^\nu\Big)\\
&\ge \min\Big((1-Cr_i/p^\nu)p^{2\nu},(1-Cr_{i-1}/p^\nu)p^{2\nu}-Cr_ip^\nu,\\
&\qquad \mathbb{E}\|\boldsymbol{\delta}_{i-2}\|_2^2-Cr_ip^\nu-Cr_{i-1}p^\nu\Big)\\
&\ge (1-C/p^\nu)p^{2\nu}-C\sum_{j=1}^{i}r_jp^\nu.
\end{aligned}$$

Note that $\sum_{j=1}^i r_j=O(i^{1-\alpha})$ when $0<\alpha<1$ and $\sum_{j=1}^i r_j=O(\log i)$ when $\alpha=1$, So if $\alpha=1$ and $\log(n/m)=o(p^\nu)$, or if $0<\alpha<1$ and $n/m=o(p^{\nu/(1-\alpha)})$, we have $\mathbb{E}\|\boldsymbol{\delta}_s\|_2^2\ge Cp^{2\nu}$. □

B.3 Proof of Theorem 4.3

Denote by $\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}^{(k)}$ the local mini-batch SGD estimator on machine k. Since $\boldsymbol{\xi}_i$'s are *i.i.d.* and independent to the initial estimator $\widehat{\boldsymbol{\theta}}_0$, by $N=nL$ and Theorem 4.1,

$$\begin{aligned}
\mathbb{E}_0\|\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}-\boldsymbol{\theta}^*\|_2^2 &= \mathbb{E}_0\Big\|\frac{1}{L}\sum_{k=1}^{L}(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}^{(k)}-\boldsymbol{\theta}^*)\Big\|_2^2\\
&\le \mathbb{E}_0\Big\|\frac{1}{L}\sum_{k=1}^{L}\Big\{(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}^{(k)}-\boldsymbol{\theta}^*)-\mathbb{E}_0(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}^{(k)}-\boldsymbol{\theta}^*)\Big\}\Big\|_2^2\\
&\quad +\|\mathbb{E}_0(\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}^{(1)}-\boldsymbol{\theta}^*)\|_2^2\\
&= O\Big(\frac{p}{Lm^{1-\alpha}n^\alpha}\Big)+O\Big((\frac{p}{m^{1-\alpha}n^\alpha})^2\Big)\\
&= O\Big(\frac{p}{L^{1-\alpha}m^{1-\alpha}N^\alpha}+\frac{p^2}{L^{-2\alpha}m^{2-2\alpha}N^{2\alpha}}\Big).
\end{aligned}\tag{37}$$

□

B.4 Proof of the lower bound of bias for Example 3.1

We first provide an upper bound for $\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3$. On the event $\{\|\widehat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}^*\|_2\le d_n\}$,

$$\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3 = \mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3 I\{\cap_{j=1}^i\mathcal{C}_j\}+\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3 I\{\{\cap_{j=1}^i\mathcal{C}_j\}^c\}$$

$$\leq \min(d_n^3, d_n \mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2) + Cn^{3-\gamma}$$

for any $\gamma > 0$. Therefore $\max_{1\leq i\leq s} \mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3 = o(1)$ and

$$\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3 = o(1)\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + O(n^{3-\gamma}). \tag{38}$$

We next prove that $\mathbb{E}_0\|\boldsymbol{\delta}_i\|_2^2 \geq cr_ip/m$ for any $\tau > 0$ and $i \geq p^{1/\alpha+\tau}$. Recall that

$$\begin{aligned}
\|\boldsymbol{\delta}_i\|_2^2 &= \|\boldsymbol{\delta}_{i-1}\|_2^2 - 2\frac{r_i}{m}\sum_{j\in H_i}\boldsymbol{\delta}'_{i-1}g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j) + \Big\|\frac{r_i}{m}\sum_{j\in H_i}g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\Big\|_2^2 \\
&=: \|\boldsymbol{\delta}_{i-1}\|_2^2 - 2r_iU_1 + r_i^2U_2.
\end{aligned} \tag{39}$$

Note that $\mathbb{E}_0\boldsymbol{\delta}'_{i-1}g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j) = \mathbb{E}_0\boldsymbol{\delta}'_{i-1}G(\boldsymbol{z}_{i-1})$. Hence from the proof of (34) and (38),

$$\mathbb{E}_0U_1 = \frac{1}{m}\sum_{j\in H_i}\mathbb{E}_0\boldsymbol{\delta}'_{i-1}g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j) \leq C\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + Cn^{3-\gamma}$$

for any sufficiently large $\gamma > 0$. For U_2,

$$\begin{aligned}
\mathbb{E}_0\Big\|\frac{1}{m}\sum_{j\in H_i}g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\Big\|_2^2 &\geq \mathbb{E}_0\Big\|\frac{1}{m}\sum_{j\in H_i}\bar{g}(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\Big\|_2^2 \\
&= \frac{\sum_{j\in H_i}\Big(\mathbb{E}_0\|g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\|_2^2 - \mathbb{E}_0\|G(\boldsymbol{z}_{i-1})\|_2^2\Big)}{m^2}
\end{aligned}$$

Recall that $G(\boldsymbol{\theta}) = \mathbb{E}g(\boldsymbol{\theta},\boldsymbol{\xi}) = \mathbb{E}\left(\frac{\boldsymbol{X}}{1+e^{-\boldsymbol{X}'\boldsymbol{\theta}^*}} - \frac{\boldsymbol{X}}{1+e^{-\boldsymbol{X}'\boldsymbol{\theta}}}\right)$. We have

$$\|G(\boldsymbol{\theta})\|_2^2 = \Big\|\mathbb{E}\big(\frac{\boldsymbol{X}}{1+e^{-\boldsymbol{X}'\boldsymbol{\theta}^*}} - \frac{\boldsymbol{X}}{1+e^{-\boldsymbol{X}'\boldsymbol{\theta}}}\big)\Big\|_2^2 \leq Cp\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2.$$

So we have $\mathbb{E}_0\|G(\boldsymbol{z}_{i-1})\|_2^2 \leq Cp\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 = o(p)$. Also

$$\mathbb{E}_0\|g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\|_2^2 = \mathbb{E}_0\frac{\|\boldsymbol{X}_j\|_2^2}{(1+e^{-\boldsymbol{X}_j'\boldsymbol{z}_{i-1}})^2} \geq \mathbb{E}\frac{\|\boldsymbol{X}_j\|_2^2}{(1+e^{-\boldsymbol{X}_j'\boldsymbol{\theta}^*})^2} - Cp\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2 \geq Cp.$$

This yields that

$$\mathbb{E}_0\|\boldsymbol{\delta}_i\|_2^2 \geq (1-cr_i)\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + Cr_i^2\frac{p}{m}$$

for some positive constants c and C. Then by Lemma A.4, $\mathbb{E}_0\|\boldsymbol{\delta}_i\|_2^2 \geq c_1r_ip/m$ for all $i \geq p^{1/\alpha+\tau/2}$.

Now by $\boldsymbol{\theta}^* = (1,0,...0)'$, $\mathbb{E}X_i = 0$ for $1 \leq i \leq p-1$ and Taylor's formulation, we have

$$\begin{aligned}
\mathbb{E}_0\delta_{i,1} &= \mathbb{E}_0\delta_{i-1,1} - \frac{r_i}{m}\sum_{j\in H_i}\mathbb{E}_0g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j) \\
&= \mathbb{E}_0\delta_{i-1,1} - r_i\frac{e}{(1+e)^2}\mathbb{E}_0\delta_{i-1,1} + r_i\frac{e^2-e}{2(1+e)^3}\mathbb{E}_0\boldsymbol{\delta}'_{i-1}\boldsymbol{\Sigma}\boldsymbol{\delta}_{i-1} + O(r_i)\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3.
\end{aligned}$$

By (38), we have

$$\mathbb{E}_0\delta_{i,1} \geq (1-cr_i)\mathbb{E}_0\delta_{i-1,1} + Cr_i^2 p/m$$

for some positive c and C and all $i \geq p^{1/\alpha+\tau/2}$. Noting that $\prod_{j=p^{1/\alpha+\tau/2}+1}^{s}(1-cr_j) = O(n^{-\gamma})$ for any $\gamma > 0$ by letting c_0 in r_i be sufficiently large, by the proof of the second claim in Lemma A.4,

$$\begin{aligned}
\mathbb{E}_0\delta_{s,1} &\geq C\frac{p}{m}\sum_{k=p^{1/\alpha+\tau/2}+1}^{s} r_k^2\prod_{j=k}^{i-1}(1-cr_{j+1}) \\
&\quad +\mathbb{E}_0\delta_{p^{1/\alpha+\tau/2},1}\prod_{j=p^{1/\alpha+\tau/2}+1}^{s}(1-cr_j) \\
&\geq Cr_s p/m,
\end{aligned}$$

which completes the proof. □

B.5 Proof of the lower bound of bias for Example 3.2

As above, we can show that $\max_{1\leq i\leq s}\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3 = o(1)$ and $\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^3 = o(1)\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + O(n^{3-\gamma})$. Also, similarly,

$$\mathbb{E}_0 U_1 \leq C\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2 + Cn^{3-\gamma}$$

for any sufficiently large $\gamma > 0$. Note that

$$\mathbb{E}_0\|G(\boldsymbol{z}_{i-1})\|_2^2 \leq \mathbb{E}_0\|\boldsymbol{X}_j\|_2^2\Big(F(\boldsymbol{X}_j'\boldsymbol{\delta}_{i-1})-\tau\Big)^2 \leq C\mathbb{E}_0\|\boldsymbol{X}_j\|_2^2(\boldsymbol{X}_j'\boldsymbol{\delta}_{i-1})^2 \leq Cp\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2^2.$$

Also

$$\begin{aligned}
\mathbb{E}_0\|g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j)\|_2^2 &= \mathbb{E}_0\|\boldsymbol{X}_j\|_2^2(F(\boldsymbol{X}_j'\boldsymbol{\delta}_{i-1}) + \tau^2 - 2\tau F(\boldsymbol{X}_j'\boldsymbol{\delta}_{i-1})) \\
&\geq \tau(1-\tau)\mathbb{E}_0\|\boldsymbol{X}_j\|_2^2 - Cp\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|_2 \\
&\geq Cp.
\end{aligned}$$

Then by Lemma A.4, we have $\mathbb{E}_0\|\boldsymbol{\delta}_i\|_2^2 \geq cr_i p/m$ for all $i \geq p^{1/\alpha+\tau/2}$.

Since $\mathbb{E}X_i = 0$ for $1 \leq i \leq p-1$, we have for $i \geq p^{1/\alpha+\tau/2}$,

$$\begin{aligned}
\mathbb{E}_0\delta_{i,1} &= \mathbb{E}_0\delta_{i-1,1} - \frac{r_i}{m}\sum_{j\in H_i}\mathbb{E}_0 g(\boldsymbol{z}_{i-1},\boldsymbol{\xi}_j) \\
&= \mathbb{E}_0\delta_{i-1,1} - \frac{r_i}{m}\sum_{j\in H_i}\mathbb{E}_0[F(\boldsymbol{X}_j'\boldsymbol{\delta}_{i-1}) - F(0)] \\
&= (1-r_iF'(0))\mathbb{E}_0\delta_{i-1,1} + r_iF''(0)\mathbb{E}_0\boldsymbol{\delta}_{i-1}'\boldsymbol{\Sigma}\boldsymbol{\delta}_{i-1} + O(r_i\mathbb{E}_0\|\boldsymbol{\delta}_{i-1}\|^3) \\
&\geq (1-r_iF'(0))\mathbb{E}_0\delta_{i-1,1} + cF''(0)r_i^2 p/m.
\end{aligned}$$

So we have $\mathbb{E}_0\delta_{s,1} \geq Cr_s p/m$. □

C Proofs for results of FONE in Section 4.2

C.1 Proof of Proposition 4.4

Define

$$\mathcal{E}_t = \Big\{ \sup_{\substack{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_2 \le c_4, \\ \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*\|_2 \le c_4}} \frac{\Big\|\frac{1}{m}\sum_{i\in B_t}[\bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\xi}_i) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\xi}_i)]\Big\|_2}{\sqrt{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 + n^{-\gamma_2}}} \le c\sqrt{\frac{p\log n}{m}} \Big\}.$$

In Lemma A.2, take $\boldsymbol{u} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, $\boldsymbol{u}_0 = (\boldsymbol{\theta}_0', \boldsymbol{\theta}_0')'$, $q = 2p$ and $h(\boldsymbol{u}, \boldsymbol{\xi}) = \bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\xi}) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\xi})$. Then (C3) implies that (B1)–(B4) hold with $\alpha = 1$, $b(\boldsymbol{u}) = C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2$ and $b(\boldsymbol{u})$ satisfies $|b(\boldsymbol{u}_1) - b(\boldsymbol{u}_2)| \le C(1 + \|\boldsymbol{\theta}^*\|_2)\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2 \le C\sqrt{p}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2$. Therefore, by Lemma A.2,

$$\mathbb{P}(\mathcal{E}_t) \ge 1 - O(n^{-\gamma})$$

for any large γ. Now take $m = n$ and define

$$\mathcal{E} = \Big\{ \sup_{\substack{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_2 \le c_4, \\ \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*\|_2 \le c_4}} \frac{\Big\|\frac{1}{n}\sum_{i=1}^{n}[\bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\xi}_i) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\xi}_i)]\Big\|_2}{\sqrt{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 + n^{-\gamma_2}}} \le c\sqrt{\frac{p\log n}{n}} \Big\}.$$

Then for any $\gamma_2, \gamma > 0$,

$$\mathbb{P}(\mathcal{E}) \ge 1 - O(n^{-\gamma}).$$

Recall

$$\boldsymbol{z}_t = \boldsymbol{z}_{t-1} - \eta_t\Big(\frac{1}{m}\sum_{i\in B_t}[g(\boldsymbol{z}_{t-1}, \boldsymbol{\xi}_i) - g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)] + \boldsymbol{a}\Big).$$

Let the event $\mathcal{A} = \{\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 \le d_n, \|\boldsymbol{a}\|_2 \le \tau_n\}$ with $d_n, \tau_n \to 0$, and $\mathcal{B}_t = \{\|\boldsymbol{z}_{t-1} - (\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\Sigma}^{-1}\boldsymbol{a})\|_2 \le b_n\}$ with $b_n \to 0$, $\frac{p\log n}{m} \le b_n^2$ and $\tau_n = o(b_n)$. Note that on $\mathcal{A}\cap\mathcal{B}_t$, we have $\|\boldsymbol{z}_{t-1} - \boldsymbol{\theta}^*\|_2 \le C(b_n + d_n)$. Define $\mathcal{D}_t = \mathcal{A} \cap \cap_{i=1}^t \mathcal{C}_i$, where $\mathcal{C}_i$ is defined in the proof of Theorem 4.1.

We first prove that on $\mathcal{D}_t$, $\max_{1\le i\le t}\|\boldsymbol{z}_i - (\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\Sigma}^{-1}\boldsymbol{a})\|_2 \le b_n$. Let $\widetilde{\boldsymbol{\delta}}_t = \boldsymbol{z}_t - (\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\Sigma}^{-1}\boldsymbol{a})$ and

$$\Delta(\boldsymbol{z}_{t-1}) = \frac{1}{m}\sum_{i\in B_t}[g(\boldsymbol{z}_{t-1}, \boldsymbol{\xi}_i) - g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)] - [G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0)].$$

We have

$$\widetilde{\boldsymbol{\delta}}_t = \widetilde{\boldsymbol{\delta}}_{t-1} - \eta_t\Big(G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0) + \Delta(\boldsymbol{z}_{t-1}) + \boldsymbol{a}\Big)$$

and

$$\begin{aligned} \|\widetilde{\boldsymbol{\delta}}_t\|_2^2 &= \|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 - 2\eta_t\widetilde{\boldsymbol{\delta}}_{t-1}'[G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0)] - 2\eta_t\widetilde{\boldsymbol{\delta}}_{t-1}'(\Delta(\boldsymbol{z}_{t-1}) + \boldsymbol{a}) \\ &\quad + \eta_t^2\Big\|G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0) + \Delta(\boldsymbol{z}_{t-1}) + \boldsymbol{a}\Big\|_2^2. \end{aligned} \tag{40}$$

Note that $G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0) = \boldsymbol{\Sigma}(\boldsymbol{z}^*_{t-1})(\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0)$, where $\boldsymbol{z}^*_{t-1}$ is between $\boldsymbol{z}_{t-1}$ and $\widehat{\boldsymbol{\theta}}_0$ and satisfies $\|\boldsymbol{z}^*_{t-1} - \boldsymbol{\theta}^*\|_2 \le \|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 + \|\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0\|_2$. So we have

$$\begin{aligned}
&\widetilde{\boldsymbol{\delta}}'_{t-1}[G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0)] + \widetilde{\boldsymbol{\delta}}'_{t-1}(\Delta(\boldsymbol{z}_{t-1}) + \boldsymbol{a}) \\
&= \widetilde{\boldsymbol{\delta}}'_{t-1}\boldsymbol{\Sigma}(\boldsymbol{z}^*_{t-1})(\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0) + \widetilde{\boldsymbol{\delta}}'_{t-1}\boldsymbol{a} + \widetilde{\boldsymbol{\delta}}'_{t-1}\Delta(\boldsymbol{z}_{t-1}) \\
&= \widetilde{\boldsymbol{\delta}}'_{t-1}\boldsymbol{\Sigma}(\boldsymbol{z}^*_{t-1})\widetilde{\boldsymbol{\delta}}_{t-1} - \widetilde{\boldsymbol{\delta}}'_{t-1}[\boldsymbol{\Sigma}(\boldsymbol{z}^*_{t-1})\boldsymbol{\Sigma}^{-1} - \mathbf{I}]\boldsymbol{a} + \widetilde{\boldsymbol{\delta}}'_{t-1}\Delta(\boldsymbol{z}_{t-1}).
\end{aligned} \tag{41}$$

On $\mathcal{A}\cap\mathcal{B}_t$, by (C2), we have $2c_1^{-1} \ge \lambda_{\max}(\boldsymbol{\Sigma}(\boldsymbol{z}^*_{t-1})) \ge \lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{z}^*_{t-1})) \ge c_1/2$ since $d_n, b_n \to 0$. Also

$$\begin{aligned}
&\big\|\widetilde{\boldsymbol{\delta}}'_{t-1}[\boldsymbol{\Sigma}(\boldsymbol{z}^*_{t-1})\boldsymbol{\Sigma}^{-1} - \mathbf{I}]\boldsymbol{a}\big\|_2 \\
&\quad \le C_1\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2\|\boldsymbol{z}^*_{t-1} - \boldsymbol{\theta}^*\|_2\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}\|_2 \\
&\quad \le C_1\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2\big(\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 + \|\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0\|_2\big)\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}\|_2 \\
&\quad = C_1\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2\big(\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 + \|\widetilde{\boldsymbol{\delta}}_{t-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{a}\|_2\big)\|\boldsymbol{\Sigma}^{-1}\boldsymbol{a}\|_2 \\
&\quad \le C\Big(\tau_n\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 + \tau_n^2\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2 + \tau_n\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2\Big).
\end{aligned} \tag{42}$$

Furthermore, on $\mathcal{D}_t\cap\mathcal{B}_t$, we have that

$$\|\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta(\boldsymbol{z}_{t-1})\|_2 \le C\sqrt{\frac{p\log n}{m}}\|\widetilde{\boldsymbol{\delta}}'_{t-1}\|_2 \tag{43}$$

and

$$\begin{aligned}
&\big\|G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0) + \Delta(\boldsymbol{z}_{t-1}) + \boldsymbol{a}\big\|_2^2 \\
&\quad \le C\Big(\|\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0\|_2^2 + \|\Delta(\boldsymbol{z}_{t-1})\|_2^2 + \tau_n^2\Big) \\
&\quad \le C(\frac{p\log n}{m} + \tau_n^2) + C\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2.
\end{aligned} \tag{44}$$

Since $\eta_t \le c$ for some small enough $c > 0$, by (40)-(44), we have, on $\mathcal{D}_t\cap\mathcal{B}_t$,

$$\begin{aligned}
\|\widetilde{\boldsymbol{\delta}}_t\|_2^2 \;\le\;& \|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 - \eta_t\widetilde{\boldsymbol{\delta}}'_{t-1}\boldsymbol{\Sigma}(\boldsymbol{z}^*_{t-1})\widetilde{\boldsymbol{\delta}}_{t-1} + C\eta_t^2\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 + \eta_t\tau_n\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 \\
&+C\eta_t\Big(\tau_n\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2 + \tau_n^2\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2 + \sqrt{\frac{p\log n}{m}}\|\widetilde{\boldsymbol{\delta}}'_{t-1}\|_2\Big) \\
&+C_3\eta_t^2(\frac{p\log n}{m} + \tau_n^2) \\
\le\;& \|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 - C_1\eta_t\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 + C_2\eta_t(\tau_n^2\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2^2 + \tau_n^4 + \frac{p\log n}{m}) \\
&+C_3\eta_t^2(\frac{p\log n}{m} + \tau_n^2) \\
\le\;& b_n^2 - C_1\eta_t b_n^2 + C_2\eta_t(\tau_n^2\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2^2 + \tau_n^4 + \frac{p\log n}{m}) \\
&+C_3\eta_t^2(\frac{p\log n}{m} + \tau_n^2).
\end{aligned}$$

Note that

$$\tau_n^2\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\|_2^2 + \tau_n^4 + \frac{p\log n}{m} + \eta_t(\frac{p\log n}{m} + \tau_n^2) = o(b_n^2).$$

So we have on $\mathcal{D}_t \cap \mathcal{B}_t$, $\|\widetilde{\boldsymbol{\delta}}_t\|_2^2 \le b_n^2$. Combining the above arguments,

$$
\begin{aligned}
\{\max_{1\le i\le t} \|\widetilde{\boldsymbol{\delta}}_i\|_2 > b_n\} \cap \mathcal{D}_t \quad & = \{\max_{1\le i\le t} \|\widetilde{\boldsymbol{\delta}}_i\|_2 > b_n, \max_{1\le i\le t-1} \|\widetilde{\boldsymbol{\delta}}_i\| \le b_n\} \cap \mathcal{D}_t \\
& \quad + \{\max_{1\le i\le t} \|\widetilde{\boldsymbol{\delta}}_i\|_2 > b_n, \max_{1\le i\le t-1} \|\widetilde{\boldsymbol{\delta}}_i\|_2 > b_n\} \cap \mathcal{D}_t \\
& \subset \{\max_{1\le i\le t-1} \|\widetilde{\boldsymbol{\delta}}_i\|_2 > b_n\} \cap \mathcal{D}_t \\
& \subset \{\|\widetilde{\boldsymbol{\delta}}_0\|_2 > b_n\} \cap \mathcal{D}_t = \emptyset,
\end{aligned}
$$

where the last inequality follows from $\|\widetilde{\boldsymbol{\delta}}_0\| \le b_n$ due to $\tau_n = o(b_n)$. This proves that $\max_{1\le i\le t} \|\boldsymbol{z}_i - (\widehat{\boldsymbol{\theta}}_0 - \tau_n \boldsymbol{\Sigma}^{-1}\boldsymbol{a})\|_2 \le b_n$ on $\mathcal{D}_t$, i.e., $\mathcal{D}_t \subset \cap_{i=1}^{t+1}\mathcal{B}_i$.

Now let $\mathbb{E}_*(\cdot)$ be the expectation to the random set $\{B_t, t\ge 1\}$ given $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n\}$. Let

$$
\Delta_n(\boldsymbol{z}_{t-1}) = \frac{1}{n}\sum_{i=1}^{n}[g(\boldsymbol{z}_{t-1}, \boldsymbol{\xi}_i) - g(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i)] - [G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0)].
$$

Let $\widetilde{\mathcal{D}}_t = \mathcal{D}_t \cap \mathcal{E} \cap \mathcal{C}$. As in each iteration, B_t, $1 \le t \le T$ are independent, we have

$$
\begin{aligned}
\mathbb{E}_*\Big[\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta(\boldsymbol{z}_{t-1}) I\{\widetilde{\mathcal{D}}_{t-1}\}\Big] &= \mathbb{E}_*\Big[\mathbb{E}_*\big[\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta(\boldsymbol{z}_{t-1}) I\{\widetilde{\mathcal{D}}_{t-1}\} | \{B_i, 1 \le i \le t-1\}\big]\Big] \\
&= \mathbb{E}_*\Big[\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta_n(\boldsymbol{z}_{t-1}) I\{\widetilde{\mathcal{D}}_{t-1}\}\Big].
\end{aligned}
$$

Note that $I\{\widetilde{\mathcal{D}}_t\} = I\{\widetilde{\mathcal{D}}_{t-1}\} - I\{\widetilde{\mathcal{D}}_{t-1} \cap \mathcal{C}_t^c\}$. Thus

$$
\begin{aligned}
&\mathbb{E}_*\Big[\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta(\boldsymbol{z}_{t-1}) I\{\widetilde{\mathcal{D}}_t\}\Big] \\
&\quad = \mathbb{E}_*\Big[\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta_n(\boldsymbol{z}_{t-1}) I\{\widetilde{\mathcal{D}}_{t-1}\}\Big] - \mathbb{E}_*\Big[\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta(\boldsymbol{z}_{t-1}) I\{\widetilde{\mathcal{D}}_{t-1} \cap \mathcal{C}_t^c\}\Big].
\end{aligned} \tag{45}
$$

By (C3), we can get

$$
\mathbb{E} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \le c_4} \|g(\boldsymbol{\theta}, \boldsymbol{\xi})\|_2^2 \le n^{c_5}
$$

for some $c_4, c_5 > 0$. Note that on $\mathcal{D}_{t-1}$, we have $\|\widetilde{\boldsymbol{\delta}}_{t-1}\| \le b_n$ and

$$
\|\Delta(\boldsymbol{z}_{t-1})\|_2 \le 2\frac{1}{m}\sum_{i\in B_t} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \le C(b_n+d_n)} \|g(\boldsymbol{\theta}, \boldsymbol{\xi}_i) - G(\boldsymbol{\theta})\|_2.
$$

Hence

$$
\mathbb{E}\Big|\mathbb{E}_*\Big[\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta(\boldsymbol{z}_{t-1}) I\{\widetilde{\mathcal{D}}_{t-1} \cap \mathcal{C}_t^c\}\Big]\Big| = O(n^{c_5/2-\gamma})
$$

and

$$
\mathbb{E}\Big|\mathbb{E}_*\big[\big\|G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0) + \Delta(\boldsymbol{z}_{t-1}) + \boldsymbol{a}\big\|_2^2 I\{\widetilde{\mathcal{D}}_{t-1} \cap \mathcal{E}_t^c\}\big]\Big| = O(n^{-\gamma})
$$

for any large $\gamma > 0$ (by choosing c in $\mathcal{E}_t$ sufficiently large). On $\mathcal{D}_{t-1} \cap \mathcal{E}$,

$$
\|\widetilde{\boldsymbol{\delta}}'_{t-1}\Delta_n(\boldsymbol{z}_{t-1})\|_2 \quad \le \quad C\sqrt{\frac{p\log n}{n}}\big\|\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0\big\|_2 \|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2 + Cn^{-\gamma_2/2}
$$

$$\leq \quad C\sqrt{\frac{p\log n}{n}}\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 + C\tau_n\sqrt{\frac{p\log n}{n}}\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2 + Cn^{-\gamma_2/2}. \tag{46}$$

Similarly as above, on $\mathcal{D}_t \cap \mathcal{E}_t$, we have

$$\begin{aligned}
&\left\|G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0) + \Delta(\boldsymbol{z}_{t-1}) + \boldsymbol{a}\right\|_2^2 \\
&\leq C(\frac{p\log n}{m}\|\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0\|_2^2 + \tau_n^2) + C\|\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0\|_2^2 + Cn^{-\gamma_2} \\
&\leq C\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 + Cn^{-\gamma_2} + C\tau_n^2.
\end{aligned} \tag{47}$$

By (40)-(42) and (45)-(47),

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\boldsymbol{\delta}}_t\|_2^2 I\{\widetilde{\mathcal{D}}_t\}] \quad &\leq \quad (1 - C_1\eta_n)\mathbb{E}[\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2 I\{\widetilde{\mathcal{D}}_{t-1}\}] \\
&\quad + C_2\eta_n(\tau_n^2 d_n^2 + \tau_n^4 + \frac{p\log n}{n}\tau_n^2 + n^{-\gamma_2/2}) \\
&\quad + C\eta_n^2(n^{-\gamma_2} + \tau_n^2),
\end{aligned}$$

where we used $I\{\widetilde{\mathcal{D}}_t\} \leq I\{\widetilde{\mathcal{D}}_{t-1}\}$. This implies that

$$\begin{aligned}
\mathbb{E}\big[\|\widetilde{\boldsymbol{\delta}}_t\|_2^2 I\{\widetilde{\mathcal{D}}_t\}\big] \quad &\leq \quad (1 - C_1\eta_n)^t\mathbb{E}\big[\|\widetilde{\boldsymbol{\delta}}_0\|_2^2 I\{\mathcal{A}\cap\mathcal{E}\cap\mathcal{C}\}\big] \\
&\quad + \frac{1 - (1 - C_1\eta_n)^t}{C_1\eta_n}\big[C\eta_n(\tau_n^2 d_n^2 + \tau_n^4 + \frac{p\log n}{n}\tau_n^2 + n^{-\gamma_2/2}) \\
&\quad + C\eta_n^2(n^{-\gamma_2} + \tau_n^2)\big].
\end{aligned}$$

Note that $(1 - C_1\eta_n)^t \leq \exp(-C_1\eta_n t)$. Then as long as $\log(n) = o(\eta_n t)$,

$$\mathbb{E}[\|\widetilde{\boldsymbol{\delta}}_t\|_2^2 I\{\widetilde{\mathcal{D}}_t\}] \leq C(\tau_n^2 d_n^2 + \tau_n^4 + \frac{p\log n}{n}\tau_n^2 + n^{-\gamma_2/2}) + C\eta_n(n^{-\gamma_2} + \tau_n^2).$$

Therefore, since $T = O(n^A)$ for some $A > 0$, we have $\mathbb{P}(\{\mathcal{E}\cap\cap_{i=1}^T\mathcal{E}_i\}^c) = O(n^{-\gamma})$ and $\mathbb{P}(\{\mathcal{C}\cap\cap_{i=1}^T\mathcal{C}_i\}^c) = O(n^{-\gamma})$ for any $\gamma > 0$. That is, when $\mathbb{P}(\mathcal{A}^c) = o(1)$, we have

$$\|\widetilde{\boldsymbol{\delta}}_T\|_2 = O_{\mathbb{P}}(\tau_n d_n + \tau_n^2 + \sqrt{\frac{p\log n}{n}}\tau_n + \sqrt{\eta_n}\tau_n + n^{-\gamma_2/4}).$$

This proves the theorem. □

C.2 Proof of Theorem 4.5

Since $f(\boldsymbol{\theta}, \boldsymbol{\xi})$ is differentiable, we have $\frac{1}{N}\sum_{i=1}^N g(\widehat{\boldsymbol{\theta}}, \boldsymbol{\xi}_i) = 0$. Denote by

$$\mathcal{E}_N = \left\{ \sup_{\substack{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_2 \leq c_4, \\ \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*\|_2 \leq c_4}} \frac{\left\|\frac{1}{N}\sum_{i=1}^N[\bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\xi}_i) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\xi}_i)]\right\|_2}{\sqrt{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 + N^{-\gamma_2}}} \leq c\sqrt{\frac{p\log N}{N}} \right\}.$$

By the proof of Proposition 4.4, we have $\mathbb{P}(\mathcal{E}_N) \geq 1 - O(N^{-\gamma})$ for any large γ. Therefore, on the event $\mathcal{E}_N \cap \{\|\widehat{\boldsymbol{\theta}}_{j-1} - \boldsymbol{\theta}^*\|_2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq c_2\}$, in the j-th round in Algorithm 3,

$$\boldsymbol{a} \quad = \quad \frac{1}{N}\sum_{i=1}^N[g(\widehat{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\xi}_i) - g(\widehat{\boldsymbol{\theta}}, \boldsymbol{\xi}_i)]$$

$$
\begin{aligned}
&= G(\widehat{\boldsymbol{\theta}}_{j-1}) - G(\widehat{\boldsymbol{\theta}}) + O\Big(\sqrt{\frac{p\log N}{N}}\|\widehat{\boldsymbol{\theta}}_{j-1} - \widehat{\boldsymbol{\theta}}\|_2 + N^{-\gamma_2/2}\Big) \\
&= \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_{j-1} - \widehat{\boldsymbol{\theta}}) + O\Big(\sqrt{\frac{p\log N}{N}}\|\widehat{\boldsymbol{\theta}}_{j-1} - \widehat{\boldsymbol{\theta}}\|_2 + N^{-\gamma_2/2}\Big) \\
&= \boldsymbol{\Sigma}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_{j-1} - \widehat{\boldsymbol{\theta}}) + O\Big(\sqrt{\frac{p\log N}{N}}\|\widehat{\boldsymbol{\theta}}_{j-1} - \widehat{\boldsymbol{\theta}}\|_2 + N^{-\gamma_2/2}\Big) \\
&\quad + O\big(\|\widehat{\boldsymbol{\theta}}_{j-1} - \widehat{\boldsymbol{\theta}}\|_2^2 + \|\widehat{\boldsymbol{\theta}}_{j-1} - \widehat{\boldsymbol{\theta}}\|_2\|\widehat{\boldsymbol{\theta}}_{j-1} - \boldsymbol{\theta}^*\|_2\big)
\end{aligned}
$$

where $\widetilde{\boldsymbol{\theta}}$ is between $\widehat{\boldsymbol{\theta}}_{j-1}$ and $\widehat{\boldsymbol{\theta}}$ and satisfies $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \le \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 + \|\widehat{\boldsymbol{\theta}}_{j-1} - \boldsymbol{\theta}^*\|_2$. In above and throughout the paper, for a sequence of vector $\{\boldsymbol{x}_n\}$, we write $\boldsymbol{x}_n = O(a_n)$ if $\|\boldsymbol{x}_n\|_2 = O(a_n)$ for simplicity. Now, in the first round of iteration, i.e., $j = 1$, we have $\|\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}\|_2 = O_{\mathbb{P}}(n^{-\delta_1})$. Then we can let $\tau_n = Cn^{-\delta_1}$ with some large constant C. By Proposition 4.4, we have

$$
\|\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}\|_2 = O_{\mathbb{P}}(n^{-2\delta_1} + \sqrt{\frac{p\log n}{n}}n^{-\delta_1} + n^{-\delta_1-\delta_2/2} + n^{-\gamma}).
$$

This yields that $\|\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}\|_2 = O_{\mathbb{P}}(n^{-\delta_1-r} + n^{-\gamma})$ with $r = \min(\delta_1, \delta_2/2, (1-\kappa_1)/2)$. Now in the second round of iteration, we let $d_n = n^{-\delta_1}$ and $\tau_n = C(n^{-\delta_1-r} + n^{-\gamma})$. Then $\|\widehat{\boldsymbol{\theta}}_2 - \widehat{\boldsymbol{\theta}}\|_2 = O_{\mathbb{P}}(n^{-\delta_1-2r} + n^{-\gamma})$. Repeating this argument, we can show that $\||\widehat{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}}\|_2 = O_{\mathbb{P}}(n^{-\delta_1-Kr} + n^{-\gamma})$ which proves the theorem since γ can be arbitrarily large. □

C.3 Proof of Proposition 4.6

Note that (C3*) implies that (B4*) holds with $\alpha = 0$. Define

$$
\mathcal{E}_t = \Big\{ \sup_{\substack{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_2 \le c_4, \\ \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*\|_2 \le c_4}} \frac{\Big\|\frac{1}{m}\sum_{i\in B_t}[\bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\xi}_i) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\xi}_i)]\Big\|_2}{\sqrt{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 + \frac{p\log n}{m}}} \le c\sqrt{\frac{p\log n}{m}} \Big\}
$$

and

$$
\mathcal{E} = \Big\{ \sup_{\substack{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|_2 \le c_4, \\ \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*\|_2 \le c_4}} \frac{\Big\|\frac{1}{n}\sum_{i=1}^{n}[\bar{g}(\boldsymbol{\theta}_1, \boldsymbol{\xi}_i) - \bar{g}(\boldsymbol{\theta}_2, \boldsymbol{\xi}_i)]\Big\|_2}{\sqrt{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 + \frac{p\log n}{n}}} \le c\sqrt{\frac{p\log n}{n}} \Big\}.
$$

We have by Lemma A.2 that $\mathbb{P}(\mathcal{E} \cap \cap_{t=1}^{T}\mathcal{E}_t) \ge 1 - O(n^{-\gamma})$ for any $\gamma > 0$.

On $\mathcal{D}_{t-1} \cap \mathcal{E}$,

$$
\begin{aligned}
\|\widetilde{\boldsymbol{\delta}}_{t-1}'\Delta_n(\boldsymbol{z}_{t-1})\| &\le C\sqrt{\frac{p\log n}{n}}\|\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0\|_2^{1/2}\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2 + C\frac{p\log n}{n}\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2 + Cn^{-\gamma_2/2} \\
&\le C\sqrt{\frac{p\log n}{n}}\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^{3/2} + C(\tau_n^{1/2}\sqrt{\frac{p\log n}{n}} + \frac{p\log n}{n})\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2 + Cn^{-\gamma_2/2}.
\end{aligned}
$$

Similarly, on $\mathcal{D}_{t-1} \cap \mathcal{E}_t$, we have

$$
\Big\|G(\boldsymbol{z}_{t-1}) - G(\widehat{\boldsymbol{\theta}}_0) + \Delta(\boldsymbol{z}_{t-1}) + \boldsymbol{a}\Big\|_2^2
$$

$$\le C\Big(\frac{p\log n}{m}\|\boldsymbol{z}_{t-1}-\widehat{\boldsymbol{\theta}}_0\|_2+\big(\frac{p\log n}{m}\big)^2+\tau_n^2\Big)+C\|\boldsymbol{z}_{t-1}-\widehat{\boldsymbol{\theta}}_0\|_2^2+Cn^{-\gamma_2}$$
$$\le C\|\widetilde{\boldsymbol{\delta}}_{t-1}\|_2^2+\big(\frac{p\log n}{m}\big)^2+C\tau_n^2.$$

So as the proof of Proposition 4.4, we can get

$$\|\widetilde{\boldsymbol{\delta}}_T\|^2=O_{\mathbb{P}}\Big(\tau_n\frac{p\log n}{n}+\big(\frac{p\log n}{n}\big)^2+\eta_n\big(\frac{p\log n}{m}\big)^2+\tau_n^2d_n^2+\eta_n\tau_n^2+\tau_n^4\Big).$$

The proof is complete. □

C.4 Proof of Thoerem 4.7

Note that $\|\widehat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}^*\|_2+\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2=O_{\mathbb{P}}(n^{-\delta_1})$ and thus $d_n=O(n^{-\delta_1})$. For any $0<\delta<1$, by Holder's inequality, we have

$$\tau_n\frac{p\log n}{n}\le\tau_n^{2+2\delta}+\Big(\frac{p\log n}{n}\Big)^{\frac{2+2\delta}{1+2\delta}}.$$

This indicates that

$$\|\widetilde{\boldsymbol{\delta}}_T\|_2=O_{\mathbb{P}}\Big(\tau_n^{1+\delta}+\big(\frac{p\log n}{n}\big)^{\frac{1+\delta}{1+2\delta}}+\sqrt{\eta_n}\frac{p\log n}{m}+\tau_n n^{-r_1}\Big)$$

with $r_1=\min(\delta_1,\delta_2/2)$. Now we estimate τ_n. Let τ_{nj} be the value of τ_n in the j-th round. We have

$$\begin{aligned}
\boldsymbol{a} &= \frac{1}{N}\sum_{i=1}^{N}[g(\widehat{\boldsymbol{\theta}}_{j-1},\boldsymbol{\xi}_i)-g(\widehat{\boldsymbol{\theta}},\boldsymbol{\xi}_i)]+O_{\mathbb{P}}\Big(\frac{q_N}{N}\Big)\\
&= G(\widehat{\boldsymbol{\theta}}_{j-1})-G(\widehat{\boldsymbol{\theta}})+O_{\mathbb{P}}(1)\Big(\sqrt{\frac{p\log N}{N}}\|\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}}\|_2^{1/2}+\frac{q_N+p\log N}{N}\Big)\\
&= \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}})+O_{\mathbb{P}}(1)\Big(\sqrt{\frac{p\log N}{N}}\|\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}}\|_2^{1/2}+\|\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}}\|_2^2\\
&\quad+\|\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}}\|_2\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2+\frac{q_N+p\log N}{N}\Big)\\
&=: \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}})+A_{nj}.
\end{aligned}$$

So on the event

$$E_{j-1}:=\Big\{\|\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}}\|_2\le Cn^{-b_{j-1}}+C\frac{q_N}{N}+C\Big(\frac{p\log n}{n}\Big)^{\frac{1+\delta}{1+2\delta}}+\sqrt{\eta_n}\frac{p\log n}{m}\Big\}$$

for some $b_{j-1}>0$, we have

$$\tau_{nj}\le Cn^{-b_{j-1}}+C\frac{q_N}{N}+C\Big(\frac{p\log N}{N}\Big)^{\frac{1+\delta}{1+2\delta}}+C\sqrt{\eta_n}\frac{p\log n}{m}$$

and

$$\|A_{nj}\|_2\le Cn^{-b_{j-1}(1+\delta)}+Cn^{-b_{j-1}-r_1}+C\frac{q_N}{N}+C\Big(\frac{p\log N}{N}\Big)^{\frac{1+\delta}{1+2\delta}}+\sqrt{\eta_n}\frac{p\log n}{m}$$

by noting that

$$\sqrt{\frac{p\log N}{N}}\|\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}}\|_2^{1/2}\leq\|\widehat{\boldsymbol{\theta}}_{j-1}-\widehat{\boldsymbol{\theta}}\|_2^{1+\delta}+\Big(\frac{p\log N}{N}\Big)^{\frac{1+\delta}{1+2\delta}}.$$

Hence on the event E_{j-1}, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}}_j-\widehat{\boldsymbol{\theta}}\|_2 &\leq \|\widetilde{\boldsymbol{\delta}}_T\|_2+\|A_{nj}\|_2\\ &\leq Cn^{-b_{j-1}(1+\delta)}+Cn^{-b_{j-1}-r_1/2}+C\frac{q_N}{N}+C\Big(\frac{p\log n}{n}\Big)^{\frac{1+\delta}{1+2\delta}}+C\sqrt{\eta_n}\frac{p\log n}{m}.\end{aligned}$$

Note that we can let $b_0=\delta_1$. Then it is easy to see that we can let $b_j\geq\delta_1$ for all j. Hence b_j satisfies $b_j\geq b_{j-1}+\min(\delta\delta_1,r_1/2)$. This proves that

$$\|\widehat{\boldsymbol{\theta}}_K-\widehat{\boldsymbol{\theta}}\|_2=O_{\mathbb{P}}\Big(n^{-\delta_1-K\min(\delta\delta_1,r_1/2)}+\frac{q_N}{N}+\Big(\frac{p\log n}{n}\Big)^{\frac{1+\delta}{1+2\delta}}+C\sqrt{\eta_n}\frac{p\log n}{m}\Big).$$

The proof is complete. □

C.5 Proof of Theorem 4.8

By Proposition 4.4, we have $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}=(\widehat{\boldsymbol{\theta}}_0-\boldsymbol{z}_T)/\tau_n$ and

$$\|\widehat{\boldsymbol{\theta}}_0-\boldsymbol{z}_T-\boldsymbol{\Sigma}^{-1}\tau_n\boldsymbol{w}\|_2=O_{\mathbb{P}}\Big(\tau_nd_n+\tau_n^2+\sqrt{\frac{p\log n}{n}}\tau_n+\sqrt{\eta_n}\tau_n+n^{-\gamma}\Big).$$

Therefore, when $\tau_n=\sqrt{(p\log n)/n}$, we have

$$\big\|\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}-\boldsymbol{\Sigma}^{-1}\boldsymbol{w}\big\|_2=O_{\mathbb{P}}\Big(\sqrt{\frac{p\log n}{n}}+\sqrt{\eta_n}+d_n\Big).$$

□

C.6 Proof of Theorem 4.9

By Proposition 4.6, we have $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}=(\widehat{\boldsymbol{\theta}}_0-\boldsymbol{z}_T)/\tau_n$ and

$$\|\widehat{\boldsymbol{\theta}}_0-\boldsymbol{z}_T-\boldsymbol{\Sigma}^{-1}\tau_n\boldsymbol{w}\|_2=O_{\mathbb{P}}\Big(\tau_nd_n+\tau_n^2+\sqrt{\frac{p\log n}{n}}\sqrt{\tau_n}+\frac{p\log n}{m}\sqrt{\eta_n}+\sqrt{\eta_n}\tau_n+\frac{p\log n}{n}\Big).$$

Therefore, when $\tau_n=\big((p\log n)/n\big)^{1/3}$, we have

$$\begin{aligned}\big\|\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}-\boldsymbol{\Sigma}^{-1}\boldsymbol{w}\big\|_2 &= O_{\mathbb{P}}\Big(\sqrt{\frac{p\log n}{n\tau_n}}+\tau_n+(\frac{p\log n}{\tau_nm}+1)\sqrt{\eta_n}+d_n\Big)\\ &= O_{\mathbb{P}}\Big((\frac{p\log n}{n})^{1/3}+\sqrt{\eta_n}(\frac{n^{1/3}(p\log n)^{2/3}}{m}+1)+d_n\Big).\end{aligned}$$

□

D Verification of Conditions on motivating examples

In this section, we provide verification of the conditions (C2), (C3) and (C3*) on Examples 3.1 and 3.2.

Example 3.1. *For a logistic regression model with* $\boldsymbol{\xi} = (Y, \boldsymbol{X})$,

$$\mathbb{P}(Y = 1|\boldsymbol{X}) = 1 - \mathbb{P}(Y = -1|\boldsymbol{X}) = \frac{1}{1 + \exp(-\boldsymbol{X'\theta}^*)}.$$

We have $f(\boldsymbol{\theta}, \boldsymbol{\xi}) = \log(1+\exp(-Y\boldsymbol{X'\theta}))$, *and* $g(\boldsymbol{\theta}, \boldsymbol{\xi}) = \frac{-Y\boldsymbol{X}}{1+\exp(Y\boldsymbol{X'\theta})}$. *Note that* $G(\boldsymbol{\theta}) = \mathbb{E}\left(\frac{\boldsymbol{X}}{1+e^{-\boldsymbol{X'\theta}^*}} - \frac{\boldsymbol{X}}{1+e^{-\boldsymbol{X'\theta}}}\right)$ *is differentiable in* $\boldsymbol{\theta}$. *Moreover, we have,*

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbb{E}\frac{\boldsymbol{XX'}}{[1 + \exp(\boldsymbol{X'\theta})][1 + \exp(-\boldsymbol{X'\theta})]}.$$

Proposition D.1. *In Example 3.1, assume that* $\widetilde{c}_1 \le \lambda_{\min}(\mathbb{E}(\boldsymbol{XX}')) \le \lambda_{\max}(\mathbb{E}(\boldsymbol{XX}')) \le \widetilde{c}_1^{-1}$ *for some* $\widetilde{c}_1 > 0$ *and* $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}|\boldsymbol{v}'\boldsymbol{X}|^3 \le \widetilde{C}_1$ *for some* $\widetilde{C}_1 > 0$.

(1) We have $\lambda_{\max}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ *is bounded uniformly in* $\boldsymbol{\theta}$. *Furthermore, if* $\|\boldsymbol{\theta}^*\|_2 \le \widetilde{C}_2$, *then* $\lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)) \ge c_1$ *for some* $c_1 > 0$ *and (C2) holds.*

(2) If the covariates $\boldsymbol{X}$ *satisfy* $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}\exp(t_0(\boldsymbol{v}'\boldsymbol{X})^2) \le \widetilde{C}_2$ *for some* $t_0, \widetilde{C}_2 > 0$, *then (C3) holds.*

Proof. Note that

$$\|\boldsymbol{\Sigma}(\boldsymbol{\theta})\|_2^2 = \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}\frac{(\boldsymbol{v}'\boldsymbol{X})^2}{[1 + \exp(\boldsymbol{X'\theta})][1 + \exp(-\boldsymbol{X'\theta})]} \le \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}(\boldsymbol{v}'\boldsymbol{X})^2 \le \widetilde{c}_1^{-1}.$$

That is, $\lambda_{\max}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ is bounded uniformly in $\boldsymbol{\theta}$. Also,

$$\begin{aligned}
\lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)) &= \min_{\|\boldsymbol{v}\|_2=1} \mathbb{E}\frac{(\boldsymbol{v}'\boldsymbol{X})^2}{[1 + \exp(\boldsymbol{X'\theta}^*)][1 + \exp(-\boldsymbol{X'\theta}^*)]} \\
&\ge \min_{\|\boldsymbol{v}\|_2=1} \mathbb{E}\frac{(\boldsymbol{v}'\boldsymbol{X})^2}{2(1+e^M)} I\{|\boldsymbol{X}'\boldsymbol{\theta}^*| \le M\} \\
&= \frac{1}{2(1+e^M)} \min_{\|\boldsymbol{v}\|_2=1} \Big(\mathbb{E}(\boldsymbol{v}'\boldsymbol{X})^2 - \mathbb{E}(\boldsymbol{v}'\boldsymbol{X})^2 I\{|\boldsymbol{X}'\boldsymbol{\theta}^*| > M\}\Big) \\
&\ge \frac{1}{2(1+e^M)} \min_{\|\boldsymbol{v}\|_2=1} \left(\mathbb{E}(\boldsymbol{v}'\boldsymbol{X})^2 - \frac{\widetilde{C}_1\widetilde{C}_2}{M}\right) \\
&= \frac{1}{2(1+e^M)} \Big(\widetilde{c}_1 - M^{-1}\widetilde{C}_1\widetilde{C}_2\Big).
\end{aligned}$$

Now let M be a constant that satisfies $M > \widetilde{C}_1\widetilde{C}_2/\widetilde{c}_1$. This yields that $\lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)) \ge c_1$ for some $c_1 > 0$. By noting that the derivative of $(1 + e^x)^{-1}(1 + e^{-x})^{-1}$ is bounded by 3, we have

$$\|\boldsymbol{\Sigma}(\boldsymbol{\theta}_1) - \boldsymbol{\Sigma}(\boldsymbol{\theta}_2)\| \le 3 \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}(\boldsymbol{v}'\boldsymbol{X})^2|\boldsymbol{X}'(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)| \le 3\widetilde{C}_1\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.$$

This proves (C2). Similarly, the derivative of $(1+e^x)^{-1}$ is bounded by 1, and hence for $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$,

$$|\boldsymbol{v}' g(\boldsymbol{\theta}_1, \boldsymbol{\xi}) - \boldsymbol{v}' g(\boldsymbol{\theta}_2, \boldsymbol{\xi})| \leq |\boldsymbol{v}'\boldsymbol{X}||\boldsymbol{X}'(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)| \leq U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,$$

where $U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = |\boldsymbol{v}'\boldsymbol{X}||\boldsymbol{X}'(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)|/\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$. It is easy to see that

$$\sup_{\|\boldsymbol{v}\|_2=1} \sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \mathbb{E} \exp(t_0 U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) \leq \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \exp(t_0 (\boldsymbol{v}'\boldsymbol{X})^2) \leq \widetilde{C}_2,$$

and

$$\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \leq \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \sup_{\|\boldsymbol{\theta}\|_2=1} |\boldsymbol{v}'\boldsymbol{X}||\boldsymbol{\theta}'\boldsymbol{X}| \leq Cp.$$

Therefore, $U(\boldsymbol{v}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ satisfies (C3). At the meantime, since $|\boldsymbol{v}' g(\boldsymbol{\theta}, \boldsymbol{\xi})| \leq |\boldsymbol{v}'\boldsymbol{X}|$ for any $\boldsymbol{\theta}$, and $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \exp(t_0(\boldsymbol{v}'\boldsymbol{X})^2) \leq \widetilde{C}_2$ for some $t_0, \widetilde{C}_2 > 0$, bullet (1) of (C3) holds for $g(\boldsymbol{\theta}, \boldsymbol{\xi})$. □

Example 3.2. *For a quantile regression model,*

$$y = \boldsymbol{X}'\boldsymbol{\theta}^* + \epsilon, \qquad \mathbb{P}(\epsilon \leq 0|\boldsymbol{X}) = \tau.$$

We have the non-smooth quantile loss $f(\boldsymbol{\theta}, \boldsymbol{\xi}) = \ell(y - \boldsymbol{X}'\boldsymbol{\theta})$ *with* $\ell(x) = x(\tau - I\{x \leq 0\})$*), and its subgradient* $g(\boldsymbol{\theta}, \boldsymbol{\xi}) = \boldsymbol{X}(I\{y \leq \boldsymbol{X}'\boldsymbol{\theta}\} - \tau)$*. Then* $G(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{X}(\mathbb{P}(\epsilon \leq \boldsymbol{X}'(\boldsymbol{\theta} - \boldsymbol{\theta}^*)|\boldsymbol{X}) - \tau)]$*. Furthermore, we have* $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}'\rho_{\boldsymbol{X}}(\boldsymbol{X}'(\boldsymbol{\theta} - \boldsymbol{\theta}^*))]$*, where* $\rho_{\boldsymbol{X}}(\cdot)$ *is the density function of* ϵ *given* $\boldsymbol{X}$*.*

Proposition D.2. *Assume that*

$$c_1 \leq \lambda_{\min}(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}'\rho_{\boldsymbol{X}}(0)]) \leq \lambda_{\max}(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}'\rho_{\boldsymbol{X}}(0)]) \leq c_1^{-1}$$

for some $c_1 > 0$ *and* $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}|\boldsymbol{v}'\boldsymbol{X}|^3 \leq \widetilde{C}_1$ *for some* $\widetilde{C}_1 > 0$*. The density function* $\rho_{\boldsymbol{X}}(x)$ *is bounded and satisfies* $|\rho_{\boldsymbol{X}}(x_1) - \rho_{\boldsymbol{X}}(x_2)| \leq \widetilde{C}|x_1 - x_2|$ *for some* $\widetilde{C} > 0$*. Then (C2) holds. Furthermore, if the covariates* $\boldsymbol{X}$ *satisfy* $\sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E} \exp(t_0|\boldsymbol{v}'\boldsymbol{X}|) \leq \widetilde{C}_2$*, then (C3*) holds.*

Proof. By the Lipschitz condition on $\rho_{\boldsymbol{X}}(x)$, we have

$$\|\boldsymbol{\Sigma}(\boldsymbol{\theta}_1) - \boldsymbol{\Sigma}(\boldsymbol{\theta}_2)\|_2 \leq \widetilde{C} \sup_{\|\boldsymbol{v}\|_2=1} \mathbb{E}(\boldsymbol{v}'\boldsymbol{X})^2|\boldsymbol{X}'(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)| \leq \widetilde{C}_1\widetilde{C}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.$$

Hence (C2) holds. Now we prove (C3*). Since $\rho_{\boldsymbol{X}}(x)$ is bounded, we have

$$\begin{aligned}
&\mathbb{E}\big[\sup_{\boldsymbol{\theta}_2: \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \leq n^{-M},\, \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*\|_2 \leq c_4} \|g(\boldsymbol{\theta}_1, \boldsymbol{\xi}) - g(\boldsymbol{\theta}_2, \boldsymbol{\xi})\|_2^4\big] \\
&\leq \mathbb{E}\Big[\|\boldsymbol{X}\|_2^4 I\big\{|\boldsymbol{X}'(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)| - \|\boldsymbol{X}\|_2 n^{-M} \leq \epsilon \leq |\boldsymbol{X}'(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)| + \|\boldsymbol{X}\|_2 n^{-M}\big\}\Big] \\
&\leq 2\widetilde{C}\mathbb{E}\big[\|\boldsymbol{X}\|_2^5 n^{-M}\big] \\
&\leq 2\widetilde{C} p^{5/2} n^{-M}.
\end{aligned}$$

Again,

$$
\begin{aligned}
&\mathbb{E}(\boldsymbol{v}'(g(\boldsymbol{\theta}_1,\boldsymbol{\xi})-g(\boldsymbol{\theta}_2,\boldsymbol{\xi}))^2\exp\{t_0|\boldsymbol{v}'(g(\boldsymbol{\theta}_1,\boldsymbol{\xi})-g(\boldsymbol{\theta}_2,\boldsymbol{\xi}))|\}\\
&\quad\le \mathbb{E}\Big[(\boldsymbol{v}'\boldsymbol{X})^2(I\{\epsilon\le \boldsymbol{X}'(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*)\}-I\{\epsilon\le \boldsymbol{X}'(\boldsymbol{\theta}_2-\boldsymbol{\theta}^*)\})^2\exp\{t_0|\boldsymbol{v}'\boldsymbol{X}|\}\Big]\\
&\quad\le \mathbb{E}\Big[(\boldsymbol{v}'\boldsymbol{X})^2 I\big\{\boldsymbol{X}'(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*)\le\epsilon\le \boldsymbol{X}'(\boldsymbol{\theta}_2-\boldsymbol{\theta}^*)\big\}\exp\{t_0|\boldsymbol{v}'\boldsymbol{X}|\}\Big]\\
&\qquad+\mathbb{E}\Big[(\boldsymbol{v}'\boldsymbol{X})^2 I\big\{\boldsymbol{X}'(\boldsymbol{\theta}_2-\boldsymbol{\theta}^*)\le\epsilon\le \boldsymbol{X}'(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*)\big\}\exp\{t_0|\boldsymbol{v}'\boldsymbol{X}|\}\Big]\\
&\quad\le 2\widetilde{C}\mathbb{E}\Big[(\boldsymbol{v}'\boldsymbol{X})^2|\boldsymbol{X}'(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)|\exp\{t_0|\boldsymbol{v}'\boldsymbol{X}|\}\Big]\\
&\quad\le C\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2,
\end{aligned}
$$

for some $C>0$. This ensures that (C3*) holds. □

E Additional Simulations

In this section, we provide additional simulation studies. We investigate the case of correlated design, the effect of the quality of the initial estimator, as well as the performance of the estimator of limiting variance. The data generating process has been described in Section 5 in the main text.

E.1 Effect of the underlying distribution of covariates $\boldsymbol{X}$

Suppose that $(X_{i,1}, X_{i,2}, \ldots, X_{i,p-1})$ follows a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^0)$ for $i = 1, 2, \ldots, N$. In the previous simulation studies, we adopt the covariance matrix $\mathbf{\Sigma}^0 = \mathbf{I}_{p-1}$. In this section, we consider two different structures of $\mathbf{\Sigma}^0$:

- Toeplitz: $\mathbf{\Sigma}^0_{i,j} = \varsigma^{|i-j|}$,
- Equi Corr: $\mathbf{\Sigma}^0_{i,j} = \varsigma$ for all $i \neq j$, $\Sigma_{i,i} = 1$ for all i.

For both structures, we consider the correlation parameter ς varying from $\{0.3, 0.5, 0.7\}$. In Table 3, we report the L_2-estimation errors of the proposed estimators. In all cases of the covariance matrix, Dis-FONE results are very close to those of the ERM in (2). Meanwhile, the L_2-errors of DC-SGD and SGD increase significantly when the correlation of the design matrix increases.

E.2 Effect of the initial estimator $\widehat{\boldsymbol{\theta}}_0$

Recall that our methods require a consistent initial estimator $\widehat{\boldsymbol{\theta}}_0$ to guarantee the convergence. We investigate the effect on the accuracy of the initial estimator in our methods. In particular, we fix the total sample size $N = 10^5$, the dimension $p = 100$, the number of machines $L = 20$ and varies n_0 from $5p$, $10p$ and $20p$, where n_0 denotes the size of the fresh sample used to construct the initial estimator $\widehat{\boldsymbol{\theta}}_0$. From Table 4, the error of the initial estimator $\widehat{\boldsymbol{\theta}}_0$ decreases as n_0 increases. As a consequence, DC-SGD has a better performance. On the other hand, the L_2-errors of Dis-FONE have already been quite small even when the initial estimator is less accurate.

Table 3: L_2-errors when covariates $\boldsymbol{X}$ are generated from different underlying distributions. Here the total sample size $N = 10^5$ and dimension $p = 100$, and the number of machines $L = 20$. Denote by $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$ the DC-SGD estimator and $\widehat{\boldsymbol{\theta}}_K$ the Dis-FONE.

Model	n_0	L_2-distance to the truth $\boldsymbol{\theta}^*$					L_2-distance to ERM $\widehat{\boldsymbol{\theta}}$		
		$\widehat{\boldsymbol{\theta}}_0$	$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$	$\widehat{\boldsymbol{\theta}}_K$	$\widehat{\boldsymbol{\theta}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$	$\widehat{\boldsymbol{\theta}}_K$
Logistic									
	Identity	1.310	0.467	0.151	0.104	0.092	0.453	0.117	0.037
	Toeplitz (0.3)	1.427	0.535	0.160	0.114	0.100	0.525	0.124	0.045
	Toeplitz (0.5)	1.634	0.694	0.178	0.138	0.117	0.685	0.131	0.055
	Toeplitz (0.7)	1.855	0.990	0.201	0.159	0.143	0.987	0.137	0.057
	Equi Corr (0.3)	1.398	0.548	0.160	0.119	0.103	0.536	0.119	0.039
	Equi Corr (0.5)	2.015	0.807	0.201	0.158	0.137	0.792	0.141	0.050
	Equi Corr (0.7)	2.087	1.279	0.246	0.181	0.163	1.273	0.168	0.061
Quantile									
	Identity	0.455	0.089	0.061	0.048	0.043	0.084	0.046	0.025
	Toeplitz (0.3)	0.500	0.140	0.065	0.055	0.046	0.138	0.048	0.031
	Toeplitz (0.5)	0.589	0.226	0.072	0.066	0.055	0.225	0.051	0.043
	Toeplitz (0.7)	0.775	0.422	0.092	0.100	0.072	0.424	0.065	0.078
	Equi Corr (0.3)	0.542	0.155	0.068	0.055	0.051	0.153	0.047	0.026
	Equi Corr (0.5)	0.637	0.329	0.069	0.064	0.060	0.328	0.040	0.027
	Equi Corr (0.7)	0.814	0.607	0.158	0.084	0.078	0.610	0.141	0.039

Table 4: L_2-errors when varying the size n_0 of the fresh sample used in constructing the initial estimator $\widehat{\boldsymbol{\theta}}_0$. Here the total sample size $N = 10^5$ and dimension $p = 100$, and the number of machines $L = 20$. Denote by $\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$ the DC-SGD estimator and $\widehat{\boldsymbol{\theta}}_K$ the Dis-FONE.

Model	n_0	L_2-distance to the truth $\boldsymbol{\theta}^*$					L_2-distance to ERM $\widehat{\boldsymbol{\theta}}$		
		$\widehat{\boldsymbol{\theta}}_0$	$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$	$\widehat{\boldsymbol{\theta}}_K$	$\widehat{\boldsymbol{\theta}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{DC}}$	$\widehat{\boldsymbol{\theta}}_{\mathrm{SGD}}$	$\widehat{\boldsymbol{\theta}}_K$
Logistic									
	$5p$	3.095	1.211	0.201	0.102	0.093	1.203	0.174	0.040
	$10p$	1.251	0.447	0.148	0.103	0.093	0.445	0.116	0.038
	$20p$	0.791	0.266	0.147	0.102	0.093	0.265	0.113	0.035
Quantile									
	$5p$	0.681	0.109	0.066	0.050	0.044	0.105	0.051	0.027
	$10p$	0.450	0.079	0.063	0.047	0.043	0.073	0.050	0.020
	$20p$	0.311	0.082	0.057	0.048	0.043	0.077	0.040	0.024

Table 5: Left columns: L_2-estimation errors of $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$; Right columns: Square root of the ratio of the estimated variance to the true limiting variance of ERM $\widehat{\boldsymbol{\theta}}$. The sample size $n \in \{10^5, 2\times 10^5, 5\times 10^5\}$ and dimension $p \in \{100, 200, 500\}$. The multiplier $\tau_n = ((p\log n)/n)^{1/2}$, the step-size $\eta_n = (p\log n)/n$ for logistic regression, and $\tau_n = ((p\log n)/n)^{1/3}$, $\eta_n = ((p\log n)/n)^{2/3}$ for quantile regression, respectively.

Model	n	L_2-estimation error			Square root ratio		
		$p=100$	$p=200$	$p=500$	$p=100$	$p=200$	$p=500$
Logistic							
	$n=10^5$	0.198	0.387	0.771	1.043	1.033	1.027
	$n=2\times 10^5$	0.191	0.349	0.725	1.041	1.027	1.019
	$n=5\times 10^5$	0.165	0.325	0.684	1.017	1.017	1.014
Quantile							
	$n=10^5$	0.234	0.256	0.306	1.042	1.023	1.027
	$n=2\times 10^5$	0.214	0.247	0.299	1.007	1.004	1.004
	$n=5\times 10^5$	0.187	0.211	0.251	1.005	1.002	1.003

E.3 Experiments on estimating the limiting variance

In this section, we provide simulation studies for estimating $\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$, where $\boldsymbol{\Sigma}$ is the population Hessian matrix of the underlying regression model and $\|\boldsymbol{w}\|_2 = 1$. As we illustrate in Section 4.3, this estimator plays an important role in estimating the limiting variance of the ERM.

In this experiment, we specify $\boldsymbol{w} = \mathbf{1}_p/\sqrt{p}$, the sample size $n \in \{10^5, 2\times 10^5, 5\times 10^5\}$, dimension $p \in \{100, 200, 500\}$. According to Theorems 4.8 and 4.9, we set the multiplier $\tau_n = ((p\log n)/n)^{1/2}$, the step-size $\eta_n = (p\log n)/n$ for logistic regression, and $\tau_n = ((p\log n)/n)^{1/3}$, $\eta_n = ((p\log n)/n)^{2/3}$ for quantile regression, respectively.

The left columns in Table 5 present the L_2-estimation errors of $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$, i.e., $\big\|\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}} - \boldsymbol{\Sigma}^{-1}\boldsymbol{w}\big\|_2$. Given $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$, we also compute the estimator of limiting variance $\boldsymbol{w}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}$ by (26). In the right columns of Table 5, we report the square root of the ratio between the estimated variance and the true limiting variance, i.e.,

$$\sqrt{(\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}})'\widehat{\boldsymbol{A}}(\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}})\Big/\boldsymbol{w}'\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}.$$

From Table 5, our estimator achieves good performance for both logistic and quantile regression models. As the sample size n increases, the estimation error of $\widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{w}}$ decreases and the ratio of the estimated variance over limiting variance gets closer to 1.

References

Bach, F. and E. Moulines (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*.

Banerjee, M., C. Durot, and B. Sen (2018). Divide and conquer in non-standard problems and the super-efficiency phenomenon. *Ann. Statist.*, To appear.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2018). Distributed estimation and inference with statistical guarantees. *Ann. Statist. 46*(3), 1352–1382.

Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc. 106*(494), 672–684.

Cai, T. T., C.-H. Zhang, H. H. Zhou, et al. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist. 38*(4), 2118–2144.

Chen, X., W. Liu, and Y. Zhang (2019). Quantile regression with memory constraint. *Annals of Statistics*, To appear.

Chen, X. and M. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica 24*(4), 1655–1684.

Dekel, O., R. Gilad-Bachrach, O. Shamir, and L. Xiao (2012). Optimal distributed online prediction using mini-batches. *J. Mach. Learn. Res. 13*(1), 165–202.

Fan, J., D. Wang, K. Wang, and Z. Zhu (2018). Distributed estimation of principal eigenspaces. *Ann. Statist.*, To appear.

Ghadimi, S. and G. Lan (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite pptimization I: a generic algorithmic framework. *SIAM J. Optim. 22*(4), 1469–1492.

He, X. and Q.-M. Shao (2000). On parameters of increasing dimensions. *J. Multivariate Anal. 73*(1), 120–135.

Huang, C. and X. Huo (2015). A distributed one-step estimator. Preprint. Available at arXiv:1511.01443.

Johnson, R. and T. Zhang (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*.

Jordan, M. I., J. D. Lee, and Y. Yang (2018). Communication-efficient distributed statistical inference. *J. Amer. Statist. Assoc.*, To appear.

Lai, T. L. (2003). Stochastic approximation. *Ann. Statist. 31*(2), 391–406.

Lee, J. D., Q. Lin, T. Ma, and T. Yang (2017). Distributed stochastic variance reduced gradient methods by sampling extra data with replacement. *J. Mach. Learn. Res. 18*(1), 4404–4446.

Lee, J. D., Q. Liu, Y. Sun, and J. E. Taylor (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res. 18*(5), 1–30.

Li, R., D. K. Lin, and B. Li (2013). Statistical inference in massive data sets. *Appl. Stoch. Model Bus. 29*(5), 399–409.

Li, T., A. Kyrillidis, L. Liu, and C. Caramanis (2018). Approximate newton-based statistical inference using only stochastic gradients. Preprint. Available at arXiv:1805.08920.

Liang, T. and W. Su (2017). Statistical inference for the population landscape via moment adjusted stochastic gradients. Preprint. Available at arXiv:1712.07519.

Nesterov, Y. and J.-P. Vial (2008). Confidence level solutions for stochastic programming. *Automatica 44*(6), 1559–1568.

Pang, L., W. Lu, and H. J. Wang (2012). Variance estimation in censored quantile regression via induced smoothing. *Comput. Statist. Data Anal. 56*(4), 785–796.

Polyak, B. T. and A. B. Juditsky (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim. 30*(4), 838–855.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *Ann. Math. Statist. 22*(3), 400–407.

Shi, C., W. Lu, and R. Song (2018). A massive data framework for m-estimators with cubic-rate. *J. Amer. Statist. Assoc.*, To appear.

Su, W. and Y. Zhu (2018). Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. Preprint. Available at arXiv:1802.04876.

Toulis, P., E. M. Airoldi, et al. (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Statist. 45*(4), 1694–1727.

Volgushev, S., S.-K. Chao, and G. Cheng (2018). Distributed inference for quantile regression processes. *Ann. Statist.*, To appear.

Wang, J., M. Kolar, N. Srebro, and T. Zhang (2017). Efficient distributed learning with sparsity. In *International Conference on Machine Learning.*

Wang, J. and T. Zhang (2017). Improved optimization of finite sums with minibatch stochastic variance reduced proximal iterations. Preprint. Available at arXiv:1706.07001.

Xiao, L. and T. Zhang (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM J. Optim. 24*(4), 2057–2075.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning.*

Zhang, Y., J. Duchi, and M. Wainwright (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res. 16*, 3299–3340.

Zhao, T., G. Cheng, H. Liu, et al. (2016). A partially linear framework for massive heterogeneous data. *Ann. Statist. 44*(4), 1400–1437.