

# A Distributed One-Step Estimator

Cheng Huang<sup>\*</sup> and Xiaoming Huo<sup>†</sup>

Georgia Institute of Technology

November 4, 2015

## Abstract

Distributed statistical inference has recently attracted enormous attention. Many existing work focuses on the averaging estimator, e.g., Zhang et al. [2013b] together with many others. We propose a one-step approach to enhance a simple-averaging based distributed estimator. We derive the corresponding asymptotic properties of the newly proposed estimator. We find that the proposed one-step estimator enjoys the same asymptotic properties as the centralized estimator. The proposed one-step approach merely requires one additional round of communication in relative to the averaging estimator; so the extra communication burden is insignificant. In finite sample cases, numerical examples show that the proposed estimator outperforms the simple averaging estimator with a large margin in terms of the mean squared errors. A potential application of the one-step approach is that one can use multiple machines to speed up large scale statistical inference with little compromise in the quality of estimators. The proposed method becomes more valuable when data can only be available at distributed machines with limited communication bandwidth.

## 1 Introduction

In many important contemporary applications, data are often partitioned across multiple servers. For example, a search engine company may have data coming from a large number of locations, and each location collects tera-bytes of data per day [Corbett et al., 2012]. On a different setting, high volumn of data (like videos) have to be stored distributively, instead of on a centralized server [Mitra et al., 2011]. Given the modern “data deluge”, it is often the case that centralized methods are no longer possible to implement. It has also been notified by various researchers (e.g., Jaggi et al. [2014]) that the speed of local processors can be thousands time faster than the rate of data transmission in a modern network. Consequently it is evidently advantageous to develop communication-efficient method, instead of transmitting data to a central location and then apply a global estimator.

In statistical inference, estimators are introduced to infer some important hidden quantities. In ultimate generality, a statistical estimator of a parameter  $\theta \in \Theta$  is a measurable function of the data, taking values in the parameter space  $\Theta$ . Many statistical inference problems could be solved by finding the maximum likelihood estimators (MLE), or more generally, M-estimators. In either case, the task is to minimize an objective function, which is the average of a criterion function over the entire data, which is typically denoted by  $S = \{X_1, X_2, \dots, X_N\}$ , where  $N$  is called the sample size. Here we choose a capitalized  $N$  to distinguish from a lower  $n$  that will be used later. Traditional centralized setting requires access to entire data set  $S$  simultaneously. However, due to the explosion of data size, it may be infeasible to store all the data in a single machine like we did during past several decades. Distributed (sometimes, it is called *parallel*) statistical inference would be an indispensable approach for solving these large-scale problems.

At a high level, there are at least two types of distributed inference problems. In the first type, each sample  $X_i$  is completely observed at one location; at the same time, different samples (i.e.,  $X_i$  and  $X_j$  for  $i \neq j$ ) may be stored at different locations. This paper will focus to this type of problems. On the other hand, it is possible that for the same sample  $X_i$ , different parts are available at different locations, and they are *not* available in a centralized

---

<sup>\*</sup>c.huang@gatech.edu

<sup>†</sup>huo@gatech.edu

fashion. The latter has been studied in the literature (see Gamal and Lai [2015] and references therein). This paper will not study the second type.

For distributed inference in the first type of the aforementioned setting, data are split into several subsets and each subset is assigned to a processor. This paper will focus on the M-estimator framework, in which an estimator is obtained by solving a distributed optimization problem. The objective in the distributed optimization problem may come from an M-estimator framework (or more particularly from the maximum likelihood principle), empirical risk minimization, and/or penalized version of the above. Due to the type 1 setting, we can see that the objective functions in the corresponding optimization problem are separable; in particular, the global objective function is a summation of functions such that each of them only depends on data reside on one machine. The exploration in this paper will base on this fact. As mentioned earlier, a distributed inference algorithm should be communication-efficient because of high communication cost between different machines or privacy concerns (such as sensitive personal information or financial data). It is worth noting that even if the data could be handled by a single machine, distributed inference would still be beneficial for reducing computing time.

Our work has been inspired by recent progress in distributed optimization. We review some noticeable progress in numerical approaches and their associated theoretical analysis. Plenty of research work has been done in distributed algorithms for large scale optimization problems during recent years. Boyd et al. [2011] suggests to use Alternating Direction Method of Multipliers (ADMM) to solve distributed optimization problems in statistics and machine learning. Using a trick of *consistency* (or sometimes called *consensus*) constraints on local variables and a global variable, ADMM can be utilized to solve a distributed version of the Lasso problem [Tibshirani, 1996, Chen, Donoho, and Saunders, 1998]. ADMM has also been adopted in solving distributed logistic regression problem, and many more. ADMM is feasible for a wide range of problems, but it requires iterative communication between local machines and the center. In comparison, we will propose a method that only requires two times iteration. Zinkevich et al. [2010] proposes a parallelized stochastic gradient descent method for empirical risk minimization and proves its convergence. The established contractive mappings technique seems to be a powerful method to quantify the speed of convergence of the derived estimator to its limit. Shamir et al. [2014] presents the Distributed Approximate Newton-type Method (DANE) for distributed statistical optimization problems. Their method firstly averages the local gradients then follows by averaging all local estimators in each iteration until convergence. They prove that this method enjoys linear convergence rate for quadratic objectives. For non-quadratic objectives, it has been showed that the value of objective function has geometric convergence rate. Jaggi et al. [2014] proposes a communication-efficient method for distributed optimization in machine learning, which uses local computation with randomized dual coordinate descent in a primal-dual setting. They also prove the geometric convergence rate of their method. The above works focused on the properties of numerical solutions to the corresponding optimization problems. Nearly all of them require more than two rounds of communication. Due to different emphasis, they did not study the statistical asymptotic properties (such as convergence in probability, asymptotic normality, Fisher information bound) of the resulting estimators.

Now we switch the gear to statistical inference. Distributed inference has been studied in many existing works, and various proposals have been made in different settings. To the best of our knowledge, the distributed one-step estimator has *not* been studied in any of these existing works. We review a couple of state-of-the-art approaches in the literature. Our method builds on a closely related recent line of work of Zhang et al. [2013b], which presents a straight forward approach to solve large scale statistical optimization problem, where the local empirical risk minimizers are simply averaged. They showed that this averaged estimator achieves mean squared error that decays as  $O(N^{-1} + (N/k)^{-2})$ , where  $N$  stands for the total number of samples and  $k$  stands for the total number of machines. They also showed that the mean squared error could be even reduced to  $O(N^{-1} + (N/k)^{-3})$  with one more bootstrapping sub-sampling step. Obviously, there exists efficiency loss in their method since the centralized estimator could achieve means squared error  $O(N^{-1})$ . Liu and Ihler [2014] proposes an inspiring two-step approach: firstly find local maximum likelihood estimators, then subsequently combine them by minimizing the total Kullback-Leibler divergence (KL-divergence). They proved the exactness of their estimator as the global MLE for the full exponential family. They also estimated the mean squared errors of the proposed estimator for a curved exponential family. Due to the adoption of the KL-divergence, the effectiveness of this approach heavily depends on the parametric form of the underlying model. Chen and Xie [2014] proposes a split-and-conquer approach for a penalized regression problem (in particular, a model with the canonical exponential

distribution) and show that it enjoys the same oracle property as the method that uses the entire data set in a single machine. Their approach is based on a majority voting, followed by a weighted average of local estimators, which somewhat resembles a one-step estimator however is different. In addition, their theoretical results requires  $k \leq O(N^{\frac{1}{5}})$ , where  $k$  is the number of machines and  $N$  is the total number of samples; this is going to be different from our needed condition for theoretical guarantees. Their work considers a high-dimensional however sparse parameter vector, which is not considered in this paper. Rosenblatt and Nadler [2014] analyzes the error of averaging estimator in distributed statistical learning under two scenarios. The number of machines is fixed in the first one and the number of machines grows in the same order with the number of samples per machine. They presented asymptotically exact expression for estimator error in both scenarios and showed that the error grows linearly with the number of machines in the latter case. Their work does not consider the one-step updating that will be studied in this paper. Although it seems that their work proves the asymptotic optimality of the simple averaging, our simulations will demonstrate the additional one-step updating can improve over the simple averaging, at least in some interesting finite sample cases. Battey et al. [2015] study the distributed parameter estimation method for penalized regression and establish the oracle asymptotic property of an averaging estimator. They also discussed hypotheses testing, which is not covered in this paper. Precise upper bounds on the errors of their proposed estimator have been developed. We benefited from reading the technical proofs of their paper; however unlike our method, their method is restricted to linear regression problems with penalty and requires the number of machine  $k = o(\sqrt{N})$ . Lee et al. [2015] devise a one-shot approach, which averages “debiased” lasso estimators, to distributed sparse regression in the high-dimensional setting. They show that their approach converges at the same order of rate as the Lasso when the data set is not split across too many machines.

It is worth noting that near all existing distributed estimator are *averaging* estimators. The idea of applying one additional updating, which correspondingly requires one additional round of communication, has not be explicitly proposed. We may notice some precursor of this strategy. For example, in Shamir et al. [2014], an approximate Newton direction was estimated at the central location, and then broadcasted to local machines. Another occurrence is that in Lee et al. [2015], some intermediate quantities are estimated in a centralized fashion, and then distributed to local machines. None of them explicitly described what we will propose.

In the theory on maximum likelihood estimators (MLE) and M-estimators, there is a one-step method, which could make a consistent estimator as efficient as MLE or M-estimators with a single Newton-Raphson iteration. (Here, efficiency stands for the relative efficiency converges to 1.) See van der Vaart [2000] for more details. There have been numerous papers utilizing this method. See Bickel [1975], Fan and Chen [1999] and Zou and Li [2008]. One-step estimator enjoys the same asymptotic properties as the MLE or M-estimators as long as the initial estimators are  $\sqrt{n}$ -consistent. A  $\sqrt{n}$ -consistent estimator is much easier to find than the MLE or an M-estimator. For instance, the simple averaging estimator (e.g., the one proposed by Zhang et al. [2013b]) is good enough as a starting point for a one-step estimator.

In this paper, we propose a one-step estimator for distributed statistical inference. The proposed estimator is built on the well-analyzed simple averaging estimator. We show that the proposed one-step estimator enjoys the same asymptotic properties (including convergence and asymptotic normality) as the centralized estimator, which would utilize the entire data. Given the amount of knowledge we had on the distributed estimators, the above result may not be surprising. However, when we derive an upper bound for the error of the proposed one-step estimator, we found that we can achieve a slightly better one than those in the existing literature. We also perform a detailed evaluation of our one-step method, comparing with simple averaging method and centralized method using synthetic data. The numerical experiment is much more encouraging than the theory predicts: in nearly all cases, the one-step estimator outperformed the simple averaging one with a clear margin. We also observe that the one-step estimator achieves the comparable performance as the global estimator at a much faster rate than the simple averaging estimator. Our work may indicate that in practice, it is better to apply a one-step distributed estimator, than a simple-average one.

This paper is organized as follows. Section 2 describes details of our problem setting and two methods—the simple averaging method and the proposed one-step method. In Section 3, we study the asymptotic properties of the one-step estimator in the M-estimator framework and analyze the upper bound of its estimation error. Section 4 provides some numerical examples of distributed statistical inference with synthetic data. We conclude in Section 5. When appropriate, detailed proofs are relegated to the appendix.

## 2 Problem Formulation

### 2.1 Notations

In this subsection, we will introduce some notations that will be used in this paper. Let  $\{m(x; \theta) : \theta \in \Theta \subset \mathbb{R}^d\}$  denote a collection of criterion functions, which should have continuous second derivative. Consider a data set  $S$  consisting of  $N = nk$  samples, which are drawn i.i.d. from  $p(x)$  (for simplicity, we assume that the sample size  $N$  is a multiple of  $k$ ). This data set is divided evenly at random and stored in  $k$  machines. Let  $S_i$  denote the subset of data assigned to machine  $i$ ,  $i = 1, \dots, k$ , which is a collection of  $n$  samples drawn i.i.d. from  $p(x)$ . Note that any two subsets in those  $S_i$ 's are not overlapping.

For each  $i \in \{1, \dots, k\}$ , let the local empirical criterion function that is based on the local data set on machine  $i$  and the corresponding maximizer be denoted by

$$M_i(\theta) = \frac{1}{|S_i|} \sum_{x \in S_i} m(x; \theta) \quad \text{and} \quad \theta_i = \arg \max_{\theta \in \Theta} M_i(\theta). \quad (2.1)$$

Let the global empirical criterion function be denoted by

$$M(\theta) = \frac{1}{k} \sum_{i=1}^k M_i(\theta). \quad (2.2)$$

And let the population criterion function and its maximizer be denoted by

$$M_0(\theta) = \int_{\mathcal{X}} m(x; \theta) p(x) dx \quad \text{and} \quad \theta_0 = \arg \max_{\theta \in \Theta} M_0(\theta), \quad (2.3)$$

where  $\mathcal{X}$  is the sample space. Note that  $\theta_0$  is the parameter of interest. The gradient and Hessian matrix of  $m(x; \theta)$  with respect to  $\theta$  are denoted by

$$\dot{m}(x; \theta) = \frac{\partial m(x; \theta)}{\partial \theta}, \quad \ddot{m}(x; \theta) = \frac{\partial^2 m(x; \theta)}{\partial \theta \partial \theta^T}. \quad (2.4)$$

We also let the gradient and Hessian of local empirical criterion function be denoted by

$$\dot{M}_i(\theta) = \frac{\partial M_i(\theta)}{\partial \theta} = \frac{1}{|S_i|} \sum_{x \in S_i} \frac{\partial m(x; \theta)}{\partial \theta}, \quad \ddot{M}_i(\theta) = \frac{\partial^2 M_i(x; \theta)}{\partial \theta \partial \theta^T} = \frac{1}{|S_i|} \sum_{x \in S_i} \frac{\partial^2 m(x; \theta)}{\partial \theta \partial \theta^T}, \quad (2.5)$$

where  $i \in \{1, 2, \dots, k\}$ , and let the gradient and Hessian of global empirical criterion function be denoted by

$$\dot{M}(\theta) = \frac{\partial M(\theta)}{\partial \theta}, \quad \ddot{M}(\theta) = \frac{\partial^2 M(\theta)}{\partial \theta \partial \theta^T}. \quad (2.6)$$

Similarly, let the gradient and Hessian of population criterion function be denoted by

$$\dot{M}_0(\theta) = \frac{\partial M_0(\theta)}{\partial \theta}, \quad \ddot{M}_0(\theta) = \frac{\partial^2 M_0(\theta)}{\partial \theta \partial \theta^T}. \quad (2.7)$$

The vector norm  $\|\cdot\|$  for  $a \in \mathbb{R}^d$  that we use in this paper is the usual Euclidean norm  $\|a\| = (\sum_{j=1}^d a_j^2)^{\frac{1}{2}}$ . And we also use  $\|\cdot\|$  to denote a norm for matrix  $A \in \mathbb{R}^{d \times d}$ , which is defined as its maximal singular value, i.e., we have

$$\|A\| = \sup_{u: u \in \mathbb{R}^d, \|u\| \leq 1} \|Au\|.$$

The aforementioned matrix norm will be the major matrix norm that is used throughout the paper. The only exception is that we will also use Frobenius norm in Appendix A. And the Euclidean norm is the only vector norm that we use throughout this paper.

## 2.2 Review on M-estimators

In this paper, we will study the distributed scheme for large-scale statistical inference. To make our conclusions more general, we consider M-estimators, which could be regarded as a generalization of the Maximum Likelihood Estimators (MLE). The M-estimator  $\hat{\theta}$  could be obtained by maximizing empirical criterion function, which means

$$\hat{\theta} = \arg \max_{\theta \in \Theta} M(\theta) = \arg \max_{\theta \in \Theta} \frac{1}{|S|} \sum_{x \in S} m(x; \theta).$$

Note that, when the criterion function is the log likelihood function, i.e.,  $m(x; \theta) = \log f(x; \theta)$ , the M-estimator is exactly the MLE. Let us recall that  $M_0(\theta) = \int_{\mathcal{X}} m(x; \theta) p(x) dx$  is the population criterion function and  $\theta_0 = \arg \max_{\theta \in \Theta} M_0(\theta)$  is the maximizer of population criterion function. It is known that  $\hat{\theta}$  is a consistent estimator for  $\theta_0$ , i.e.,  $\hat{\theta} - \theta_0 \xrightarrow{P} 0$ . See Chapter 5 of van der Vaart [2000].

## 2.3 Simple Averaging Estimator

Let us recall that  $M_i(\theta)$  is the local empirical criterion function on machine  $i$ ,

$$M_i(\theta) = \frac{1}{|S_i|} \sum_{x \in S_i} m(x; \theta).$$

And,  $\theta_i$  is the local M-estimator on machine  $i$ ,

$$\theta_i = \arg \max_{\theta \in \Theta} M_i(\theta).$$

Then as mentioned in Zhang et al. [2013b], the simplest and most intuitive method is to take average of all local M-estimators. Let  $\theta^{(0)}$  denote the average of these local M-estimators, we have

$$\theta^{(0)} = \frac{1}{k} \sum_{i=1}^k \theta_i, \quad (2.8)$$

which is referred as the simple averaging estimator in the rest of this paper.

## 2.4 One-step Estimator

Under the problem setting above, starting from the simple averaging estimator  $\theta^{(0)}$ , we can obtain the one-step estimator  $\theta^{(1)}$  by performing a single Newton-Raphson update, i.e.,

$$\theta^{(1)} = \theta^{(0)} - [\ddot{M}(\theta^{(0)})]^{-1} [\dot{M}(\theta^{(0)})], \quad (2.9)$$

where  $M(\theta) = \frac{1}{k} \sum_{i=1}^k M_i(\theta)$  is the global empirical criterion function,  $\dot{M}(\theta)$  and  $\ddot{M}(\theta)$  are the gradient and Hessian of  $M(\theta)$ , respectively. The whole process to compute one-step estimator can be summarized as follows.

- (1) For each  $i \in \{1, 2, \dots, k\}$ , machine  $i$  compute the local M-estimator with its local data set,

$$\theta_i = \arg \max_{\theta \in \Theta} M_i(\theta) = \arg \max_{\theta \in \Theta} \frac{1}{|S_i|} \sum_{x \in S_i} m(x; \theta).$$

- (2) All local M-estimators are averaged to obtain simple averaging estimator,

$$\theta^{(0)} = \frac{1}{k} \sum_{i=1}^k \theta_i.$$

Then  $\theta^{(0)}$  is sent back to each local machine.

- (3) For each  $i \in \{1, 2, \dots, k\}$ , machine  $i$  compute the gradient and Hessian matrix of its local empirical criterion function  $M_i(\theta)$  at  $\theta = \theta^{(0)}$ . Then send  $\dot{M}_i(\theta^{(0)})$  and  $\ddot{M}_i(\theta^{(0)})$  to the central machine.
- (4) Upon receiving all gradients and Hessian matrices, the central machine computes gradient and Hessian matrix of  $M(\theta)$  by averaging all local information,

$$\dot{M}(\theta^{(0)}) = \frac{1}{k} \sum_{i=1}^k \dot{M}_i(\theta^{(0)}), \quad \ddot{M}(\theta^{(0)}) = \frac{1}{k} \sum_{i=1}^k \ddot{M}_i(\theta^{(0)}).$$

Then the central machine would perform a Newton-Raphson iteration to obtain a one-step estimator,

$$\theta^{(1)} = \theta^{(0)} - [\ddot{M}(\theta^{(0)})]^{-1} [\dot{M}(\theta^{(0)})].$$

Note that  $\theta^{(1)}$  is not necessarily the maximizer of empirical criterion function  $M(\theta)$  but it shares the same asymptotic properties with the corresponding global maximizer (M-estimator) under some mild conditions, i.e., we will show

$$\theta^{(1)} \xrightarrow{P} \theta_0, \quad \sqrt{N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, \Sigma), \quad \text{as } N \rightarrow \infty,$$

where the covariance matrix  $\Sigma$  will be specified later.

The one-step estimator has advantage over simple averaging estimator in terms of estimation error. In Zhang et al. [2013b], it is showed both theoretically and empirically that the MSE of simple averaging estimator  $\theta^{(0)}$  grows significantly with the number of machines  $k$  when the total number of samples  $N$  is fixed. More precisely, there exists some constant  $C_1, C_2 > 0$  such that

$$\mathbb{E}[\|\theta^{(0)} - \theta_0\|^2] \leq \frac{C_1}{N} + \frac{C_2 k^2}{N^2} + O(kN^{-2}) + O(k^3 N^{-3}).$$

Fortunately, one-step method  $\theta^{(1)}$  could achieve a lower upper bound of MSE with only one additional step. we will show the following in Section 3:

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{C_1}{N} + O(N^{-2}) + O(k^4 N^{-4}).$$

### 3 Main Results

At first, some assumptions will be introduced in Section 3.1. After that, we will study the asymptotic properties of one-step estimator in Section 3.2, i.e., convergence, asymptotic normality and mean squared error (MSE). In Section 3.3, we will consider the one-step estimator under the presence of information loss.

#### 3.1 Assumptions

Throughout this paper, we impose some regularity conditions on the criterion function  $m(x; \theta)$ , the local empirical criterion function  $M_i(\theta)$  and population criterion function  $M_0(\theta)$ . We use the similar assumptions in Zhang et al. [2013b]. Those conditions are also standard in classical statistical analysis of M-estimators (cf. van der Vaart [2000]).

First assumption restricts the parameter space to be compact, which is reasonable and not rigid in practice. One reason is that the possible parameters lie in a finite scope for most cases. Another justification is that the largest number that computers could cope with is always limited.

**Assumption 3.1** (parameter space). *The parameter space  $\Theta \in \mathbb{R}^d$  is a compact convex set. And let  $D \triangleq \max_{\theta, \theta' \in \Theta} \|\theta - \theta'\|$  denote the diameter of  $\Theta$ .*

We also assume that  $m(x; \theta)$  is concave with respect to  $\theta$  and  $M_0(\theta)$  has some curvature around the unique optimal point  $\theta_0$ , which is a standard assumption for any method requires consistency.

**Assumption 3.2** (invertibility). *The Hessian of population criterion function  $M_0(\theta)$  at  $\theta_0$  is a nonsingular matrix, which means  $\ddot{M}(\theta_0)$  is negative definite and there exists some  $\lambda > 0$  such that  $\sup_{u \in \mathbb{R}^d: \|u\| < 1} u^t \ddot{M}(\theta_0) u \leq -\lambda$ .*

In addition, we require the criterion function  $m(x; \theta)$  to be smooth enough, at least in the neighborhood of the optimal point  $\theta_0$ ,  $B_\delta = \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta\}$ . So, we impose some regularity conditions on the first and second derivative of  $m(x; \theta)$ . We assume the gradient of  $m(x; \theta)$  is bounded in moment and the difference between  $\dot{m}(x; \theta)$  and  $\dot{M}_0(\theta)$  is also bounded in moment. Moreover, we assume that  $\ddot{m}(x; \theta)$  has Lipschitz continuity in  $B_\delta$ .

**Assumption 3.3** (smoothness). *There exist some constants  $G$  and  $H$  such that*

$$\mathbb{E}[\|\dot{m}(X; \theta)\|^8] \leq G^8 \text{ and } \mathbb{E}\left[\left\|\ddot{m}(X; \theta) - \ddot{M}_0(\theta)\right\|^8\right] \leq H^8, \forall \theta \in B_\delta.$$

For any  $x \in \mathcal{X}$ , the Hessian matrix  $\ddot{m}(x; \theta)$  is  $L(x)$ -Lipschitz continuous,

$$\left\|\ddot{m}(x; \theta) - \ddot{m}(x; \theta')\right\| \leq L(x)\|\theta - \theta'\|, \forall \theta, \theta' \in B_\delta,$$

where  $L(x)$  satisfies

$$\mathbb{E}[L(X)^8] \leq L^8 \text{ and } \mathbb{E}[(L(X) - \mathbb{E}[L(X)])^8] \leq L^8,$$

for some finite constant  $L > 0$ .

By Theorem 8.1 in Chapter XIII of Lang [1993],  $m(x; \theta)$  enjoys interchangeability between differentiation on  $\theta$  and integration on  $x$ , which means the following two equations hold:

$$\dot{M}_0(\theta) = \frac{\partial}{\partial \theta} \int_{\mathcal{X}} m(x; \theta) p(x) dx = \int_{\mathcal{X}} \frac{\partial m(x; \theta)}{\partial \theta} p(x) dx = \int_{\mathcal{X}} \dot{m}(x; \theta) p(x) dx,$$

and,

$$\ddot{M}_0(\theta) = \frac{\partial^2}{\partial \theta^t \partial \theta} \int_{\mathcal{X}} m(x; \theta) p(x) dx = \int_{\mathcal{X}} \frac{\partial^2 m(x; \theta)}{\partial \theta^t \partial \theta} p(x) dx = \int_{\mathcal{X}} \ddot{m}(x; \theta) p(x) dx.$$

### 3.2 Asymptotic Properties and Mean Squared Error (MSE) Bound

Our main result is that one-step estimator enjoys oracle asymptotic properties and has mean squared error of  $O(N^{-1})$  under some mild conditions.

**Theorem 3.4.** *Let  $\Sigma = \ddot{M}_0(\theta_0)^{-1} \mathbb{E}[\dot{m}(x; \theta_0) \dot{m}(x; \theta_0)^t] \ddot{M}_0(\theta_0)^{-1}$ , where the expectation is taken with respect to  $p(x)$ . Under Assumption 3.1, 3.2, and 3.3, when the number of machines  $k$  satisfies  $k = O(\sqrt{N})$ ,  $\theta^{(1)}$  is consistent and asymptotically normal, i.e., we have*

$$\theta^{(1)} - \theta_0 \xrightarrow{P} 0 \text{ and } \sqrt{N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, \Sigma) \text{ as } N \rightarrow \infty.$$

See Appendix C for a proof. The above theorem indicates that the one-step estimator is asymptotically equivalent to the centralized M-estimator.

**Remark.** *It is worth noting that the condition  $\|\sqrt{N}(\theta^{(0)} - \theta_0)\| = O_P(1)$  suffices for our proof to Theorem 3.4. Let  $\tilde{\theta}^{(0)}$  denote another starting point for the one-step update, then the following estimator*

$$\tilde{\theta}^{(1)} = \tilde{\theta}^{(0)} - \ddot{M}(\tilde{\theta}^{(0)})^{-1} \dot{M}(\tilde{\theta}^{(0)})$$

*also enjoys the same asymptotic properties with  $\theta^{(1)}$  (and the centralized M-estimator  $\hat{\theta}$ ) as long as  $\sqrt{N}(\tilde{\theta}^{(0)} - \theta_0)$  is bounded in probability. Therefore, we can replace  $\theta^{(0)}$  with any estimator  $\tilde{\theta}^{(0)}$  that satisfies*

$$\|\sqrt{N}(\tilde{\theta}^{(0)} - \theta_0)\| = O_P(1).$$

**Theorem 3.5.** Under Assumption 3.1, 3.2, and 3.3, the mean squared error of the one-step estimator  $\theta^{(1)}$  is bounded by

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{2\text{Tr}[\Sigma]}{N} + O(N^{-2}) + O(k^4 N^{-4}).$$

When the number of machines  $k$  satisfies  $k = O(\sqrt{N})$ , we have

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{2\text{Tr}[\Sigma]}{N} + O(N^{-2}).$$

See Appendix D for a proof.

In particular, when we choose the criterion function to be the log likelihood function,  $m(x; \theta) = \log f(x; \theta)$ , the one-step estimator has the same asymptotic properties with the maximum likelihood estimator (MLE), which is described below.

**Corollary 3.6.** If  $m(x; \theta) = \log f(x; \theta)$  and  $k = O(\sqrt{N})$ , one-step estimator  $\theta^{(1)}$  is a consistent and asymptotic efficient estimator of  $\theta_0$ ,

$$\theta^{(1)} - \theta_0 \xrightarrow{P} 0 \text{ and } \sqrt{N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, I(\theta_0)^{-1}), \text{ as } N \rightarrow \infty,$$

where  $I(\theta_0)$  is the Fisher's information at  $\theta = \theta_0$ . And the mean squared error of  $\theta^{(1)}$  is bounded as follows:

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{2\text{Tr}[I^{-1}(\theta_0)]}{N} + O(N^{-2}) + O(k^4 N^{-4}).$$

*Proof.* It follows immediately from Theorem 3.4, 3.5 and the definition of the Fisher's information.  $\square$

### 3.3 Under the Presence of Communication Failure

In practice, it is possible that the information (local estimator, local gradient and local Hessian) from a local machine *cannot* be received by the central machine due to various causes (for instance, network problem or hardware crash). We assume that the communication failure on each local machine occurs independently.

We now derive a distributed estimator under the scenario with possible information loss. We will also present the corresponding theoretical results. We use  $a_i \in \{0, 1\}$ ,  $i = 1, \dots, k$ , to denote the status of local machines: when machine  $i$  successfully sends all its local information to central machine, we have  $a_i = 1$ ; when machine  $i$  fails, we have  $a_i = 0$ . The corresponding simple averaging estimator is computed as

$$\theta^{(0)} = \frac{\sum_{i=1}^k a_i \theta_i}{\sum_{i=1}^k a_i}.$$

And one-step estimator is as follows

$$\theta^{(1)} = \theta^{(0)} - \left[ \sum_{i=1}^k a_i \ddot{M}_i(\theta^{(0)}) \right]^{-1} \left[ \sum_{i=1}^k a_i \dot{M}_i(\theta^{(0)}) \right].$$

**Corollary 3.7.** Suppose  $r$  is the probability (or rate) that a local machine fails to send its information to the central machine. When  $n = N/k \rightarrow \infty$ ,  $k \rightarrow \infty$  and  $k = O(\sqrt{N})$ , the one-step estimator is asymptotically normal:

$$\sqrt{(1-r)N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, \Sigma).$$

And more precisely, unless all machines fail, we have

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] \leq \frac{2\text{Tr}[\Sigma]}{N(1-r)} + \frac{6\text{Tr}[\Sigma]}{Nk(1-r)^2} + O(N^{-2}(1-r)^{-2}) + O(k^2 N^{-2}).$$

See Appendix E for a proof. Note that the probability that all machines fail is  $r^k$ , which is negligible when  $r$  is small and  $k$  is large.



## 4 Numerical Examples

In this section, we will discuss the results of simulation studies comparing the performance of the simple averaging estimator  $\theta^{(0)}$  and the one-step estimator  $\theta^{(1)}$ , as well as the centralized M-estimator  $\hat{\theta}$ , which maximizes the global empirical criterion function  $M(\theta)$  when the entire data are available centrally. Besides, we will also study the resampled averaging estimator, which is proposed by Zhang et al. [2013b]. The main idea of a resampled averaging estimator is to resample  $\lfloor sn \rfloor$  observations from each local machine to obtain another averaging estimator  $\theta_1^{(0)}$ . Then the resampled averaging estimator can be constructed as follows:

$$\theta_{re}^{(0)} = \frac{\theta^{(0)} - s\theta_1^{(0)}}{1 - s}.$$

In our numerical examples, the resampling ratio  $s$  is chosen to be  $s = 0.1$  based on past empirical studies. We shall implement these estimators for logistic regression, Beta distribution and Gaussian Distribution. We will also study the parameter estimation for Beta distribution with occurrence of communication failures, in which some local machines could fail to send their local information to the central machine.

### 4.1 Logistic Regression

In this example, we simulate the data from the following logistic regression model:

$$y \sim \text{Bernoulli}(p), \text{ where } p = \frac{\exp(x^t \theta)}{1 + \exp(x^t \theta)} = \frac{\exp(\sum_{j=1}^d x_j \theta_j)}{1 + \exp(\sum_{j=1}^d x_j \theta_j)}. \quad (4.10)$$

In this model,  $y \in \{0, 1\}$  is a binary response,  $x \in R^d$  is a continuous predictor and  $\theta \in R^d$  is the parameter of interest.

In each single experiment, we choose a fixed vector  $\theta$  with each entry  $\theta_j, j = 1, \dots, d$ , drawn from  $\text{Unif}(-1, 1)$  independently. Entry  $x_j, j = 1, \dots, d$  of  $x \in R^d$  is sampled from  $\text{Unif}(-1, 1)$ , independent from parameters  $\theta_j$ 's and other entries. After generating parameter  $\theta$  and predictor  $x$ , we can compute the value of probability  $p$  and generate  $y$  according to (4.10). We fix the number of observed samples  $N = 2^{17} = 131,072$  in each experiment, but vary the number of machines  $k$ . The target is to estimate  $\theta$  with different number of parallel splits  $k$  of the data. The experiment is repeated for  $K = 50$  times to obtain reliable average error. And the criterion function is the log-likelihood function,

$$m(x, y; \theta) = yx^t \theta - \log(1 + \exp(x^t \theta)).$$

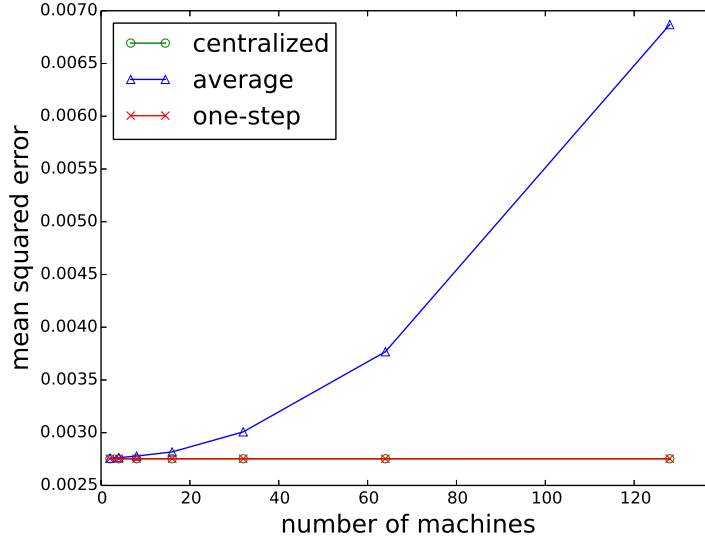
The goal of each experiment is to estimate parameter  $\theta_0$  maximizing population criterion function

$$M_0(\theta) = \mathbb{E}_{x,y}[m(x, y; \theta)] = \mathbb{E}_{x,y}[yx^t \theta - \log(1 + \exp(x^t \theta))].$$

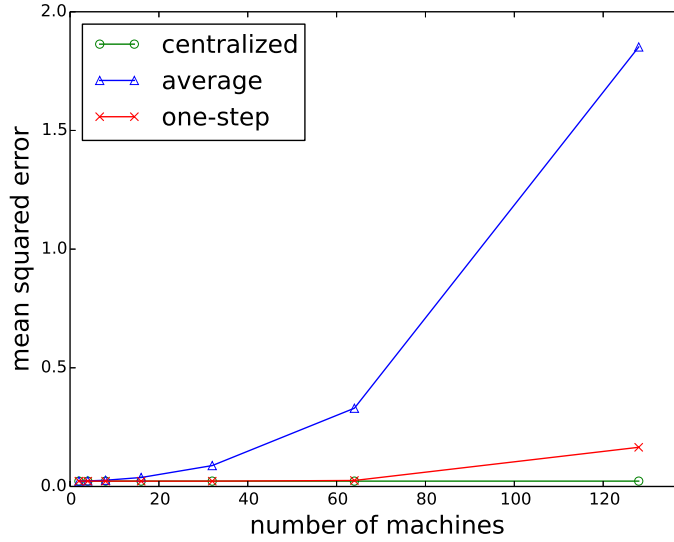
In this particular case,  $\theta_0$  is exactly the same with the true parameter.

In each experiment, we split the data into  $k = 2, 4, 8, 16, 32, 64, 128$  non-overlapping subsets of size  $n = N/k$ . We compute a local estimator  $\theta_i$  from each subset. And simple averaging estimator is obtained by taking all local estimators,  $\theta^{(0)} = \frac{1}{k} \sum_{i=1}^k \theta_i$ . Then the one-step estimator  $\theta^{(1)}$  could be computed by applying a Newton-Raphson update to  $\theta^{(0)}$ , i.e., equation (2.9).

The dimension is chosen to be  $d = 20$  and  $d = 100$ , which could help us understand the performance of those estimators in both low and high dimensional cases. In Fig. 1, we plot the mean squared error of each estimator versus the number of machines  $k$ . As we expect, the mean squared error of simple averaging estimator grows rapidly with the number of machines. But, the mean squared error of one-step estimator remains the same with the mean squared error of oracle estimator when the number of machines  $k$  is not very large. Even when the  $k = 128$  and the dimension of predictors  $d = 100$ , the performance of one-step estimator is significantly better than simple averaging estimator. As we can easily find out from Fig. 1, the mean squared error of simple averaging estimator is about 10 times of that of one-step estimator when  $k = 128$  and  $d = 100$ . Detailed values of mean squared error are listed in Table 1 and 2. From the tables, we can easily figure out that the standard deviation of the error of one-step estimator is significantly smaller than that of simple averaging, especially when the number of machines  $k$  is large, which means one-step estimator is more stable.



(a)  $d = 20$



(b)  $d = 100$

Figure 1: Logistic Regression: The mean squared error  $\|\hat{\theta} - \theta_0\|^2$  versus number of machines, with fifty simulations. The “average” is  $\theta^{(0)}$  and the “one-step” is  $\theta^{(1)}$ . The “centralized” denotes the oracle estimator with entire data.

## 4.2 Beta Distribution

In this example, we use data simulated from Beta distribution  $\text{Beta}(\alpha, \beta)$ , whose p.d.f. is as follows:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}.$$

In each experiment, we generate the value of parameter as  $\alpha \sim \text{Unif}(1, 3)$  and  $\beta \sim \text{Unif}(1, 3)$ , independently. Once  $(\alpha, \beta)$  is determined, we can simulate samples from the above density. In order to examine the performance of two distributed methods when  $k$  is extremely large, we choose to use a data set with relatively small size  $N = 2^{13} = 8192$  and let number of machines vary in a larger range  $k = 2, 4, 8, \dots, 256$ . And the objective is to estimate parameter  $(\alpha, \beta)$  from the observed data. The experiment is again repeated for  $K = 50$  times. The

no. of machines	2	4	8	16	32	64	128
simple avg ( $\times 10^{-4}$ )	28.036 (7.982)	28.066 (7.989)	28.247 (8.145)	28.865 (8.443)	30.587 (9.812)	38.478 (14.247)	69.898 (27.655)
one-step ( $\times 10^{-4}$ )	28.038 (7.996)	28.038 (7.996)	28.038 (7.996)	28.038 (7.996)	28.038 (7.996)	28.035 (7.998)	28.039 (8.017)
centralized ( $\times 10^{-4}$ )	28.038 (7.996)						

Table 1: Logistic Regression ( $d = 20$ ): Detailed values of squared error  $\|\hat{\theta} - \theta_0\|^2$ . In each cell, the first number is the mean of squared error in  $K = 50$  experiments and the number in the brackets is the standard deviation of the squared error.

no. of machines	2	4	8	16	32	64	128
simple avg ( $\times 10^{-3}$ )	23.066 (4.299)	23.818 (4.789)	26.907 (6.461)	38.484 (10.692)	87.896 (22.782)	322.274 (67.489)	1796.147 (324.274)
one-step ( $\times 10^{-3}$ )	22.787 (4.062)	22.784 (4.060)	22.772 (4.048)	22.725 (3.998)	22.612 (3.835)	24.589 (4.651)	151.440 (43.745)
centralized ( $\times 10^{-3}$ )	22.787 (4.063)						

Table 2: Logistic Regression ( $d = 100$ ): Detailed values of squared error  $\|\hat{\theta} - \theta_0\|^2$ . In each cell, the first number is the mean of squared error in  $K = 50$  experiments and the number in the brackets is the standard deviation of squared error.

criterion function is  $m(x; \theta) = \log f(x; \alpha, \beta)$ , which implies that the centralized estimator is the MLE.

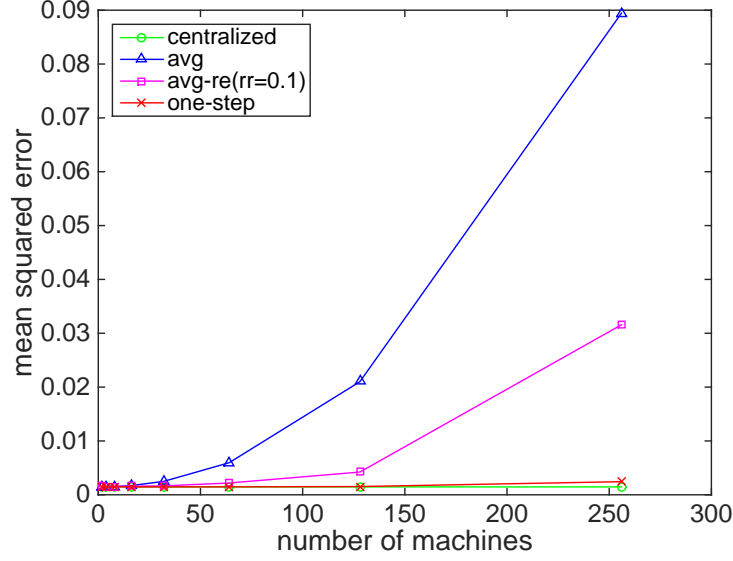
Figure 2 and Table 3 show that the one-step estimator has almost the same performance with centralized estimator in terms of MSE and standard deviation when the number of machines  $k \leq \sqrt{N}$  (i.e., when  $k \leq 64$ ). However, the one-step estimator performs worse than centralized estimator when  $k > \sqrt{N}$  (i.e., when  $k = 128$  or 256), which confirms the necessity of condition  $k = O(\sqrt{N})$  in Theorem 3.4. In addition, we can easily find out that both simple averaging estimator and resampled averaging estimator are worse than the proposed one-step estimator regardless of the value of  $k$ .

number of machines	simple avg ( $\times 10^{-3}$ )	resampled avg ( $\times 10^{-3}$ )	one-step ( $\times 10^{-3}$ )	centralized ( $\times 10^{-3}$ )
2	1.466 (1.936)	1.616 (2.150)	1.466 (1.943)	1.466 (1.943)
4	1.480 (1.907)	1.552 (2.272)	1.466 (1.943)	
8	1.530 (1.861)	1.545 (2.177)	1.466 (1.943)	
16	1.704 (1.876)	1.594 (2.239)	1.466 (1.946)	
32	2.488 (2.628)	1.656 (2.411)	1.468 (1.953)	
64	5.948 (5.019)	2.184 (3.529)	1.474 (1.994)	
128	21.002 (11.899)	4.221 (7.198)	1.529 (2.199)	
256	89.450 (35.928)	31.574 (36.518)	2.435 (3.384)	

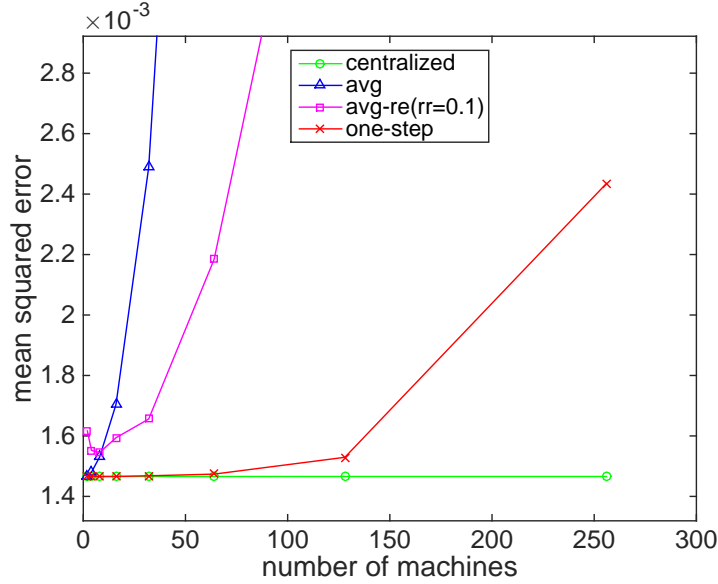
Table 3: Beta Distribution: Detailed values of squared error  $\|\hat{\theta} - \theta_0\|^2$ . In each cell, the first number is the mean squared error with  $K = 50$  experiments and the number in the brackets is the standard deviation of the squared error.

### 4.3 Beta Distribution with Possibility of Losing Information

Now, we would like to compare the performance of simple averaging estimator and one-step estimator under a more practical scenario, in which each single local machine could fail to send its information to central machine. We assume those failures would occur independently with probability  $r = 0.05$ . The simulation settings are similar to previous example in Section 4.2, however, we will generate  $N = 409600$  samples from Beta distribution



(a) overview



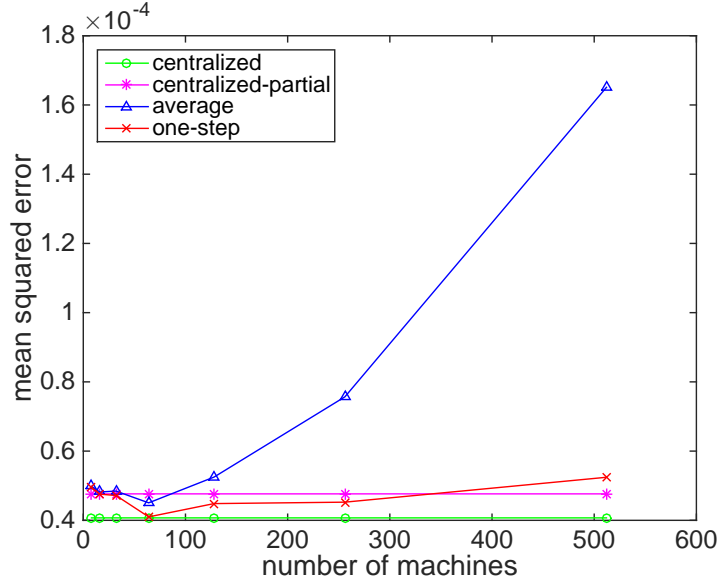
(b) detailed view

Figure 2: Beta Distribution: The error  $\|\theta - \theta_0\|^2$  versus the number of machines, with fifty simulations, where  $\theta_0$  is the true parameter. The “avg” is  $\theta^{(0)}$ , the “avg-re” is  $\theta_{re}^{(0)}$  with resampling ratio  $rr = 10\%$  and the “one-step” is  $\theta^{(1)}$ . The “centralized” denotes maximum likelihood estimator with the entire data.

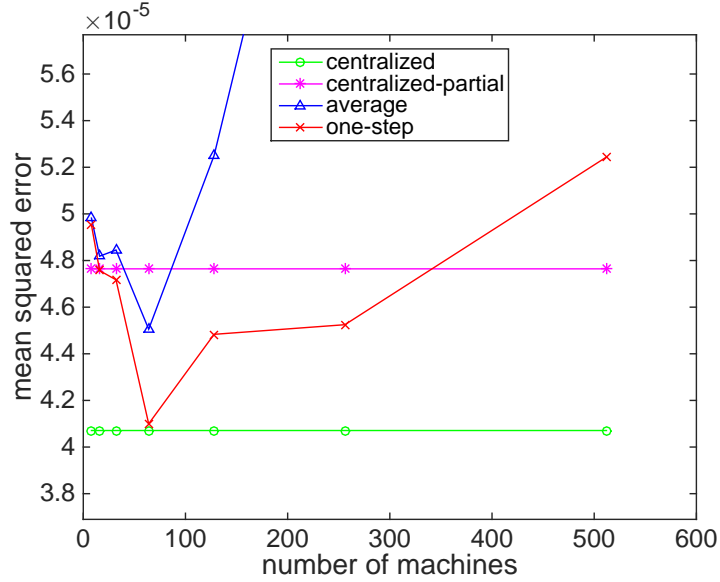
$\text{Beta}(\alpha, \beta)$ , where  $\alpha$  and  $\beta$  are chosen from  $\text{Unif}(1, 3)$ , independently. And the goal of experiment is to estimate parameter  $(\alpha, \beta)$ . In each experiment, we let the number of machines vary  $k = 8, 16, 32, 64, 128, 256, 512$ . We also compare the performance of the centralized estimator with entire data and centralized estimator with  $(1 - r) \times 100\% = 95\%$  of entire data. This experiment is repeated for  $K = 50$  times.

In Figure 3 (a), we plot the MSE of each estimator against the number of machines. As expected, the MSE of simple averaging estimator grows significantly with the number of machines while the other three remains nearly the same. We can easily find out that performance of simple averaging estimator is far worse than others, especially when the number of machines is large (for instance, when  $k = 256$  or  $512$ ). If we take a closer look at the other three estimators from Fig. 3 (b), we will find that the performance of one-step estimator is volatile but always remains in a reasonable range. And as expected, the error of one-step estimator converges to the error of

oracle estimator with partial data when number of machines  $k$  is large.



(a) overview



(b) detailed view

Figure 3: Beta Distribution with Possibility of Losing Information: The error  $\|\theta - \theta_0\|^2$  versus the number of machines, with fifty simulations, where  $\theta_0$  is the true parameter. The “average” is  $\theta^{(0)}$  and the “one-step” is  $\theta^{(1)}$ . The “centralized” denotes maximum likelihood estimator with the entire data. And the “centralized-partial” denotes the maximum likelihood estimator with  $(1 - r) \times 100\% = 95\%$  of data.

#### 4.4 Gaussian Distribution with Unknown Mean and Variance

In this part, we will compare the performance of the simple averaging estimator, the resampled averaging estimator and the one-step estimator when fixing the number of machines  $k = \sqrt{N}$  and letting the value of  $N$  increase. We draw  $N$  samples from  $N(\mu, \sigma^2)$ , where  $\mu \sim \text{Unif}(-2, 2)$  and  $\sigma^2 \sim \text{Unif}(0.25, 9)$ , independently. We let  $N$  vary in  $\{4^3, \dots, 4^9\}$  and repeat the experiment for  $K = 50$  times for each  $N$ . We choose the criterion function to be

number of machines	simple avg ( $\times 10^{-5}$ )	one-step ( $\times 10^{-5}$ )	centralized ( $\times 10^{-5}$ )	centralized(95%) ( $\times 10^{-5}$ )
8	4.98 (10.76)	4.95 (10.62)	4.07 (6.91)	4.76 (9.81)
16	4.82 (7.61)	4.75 (7.40)		
32	4.85 (9.65)	4.72 (9.31)		
64	4.51 (7.89)	4.10 (7.04)		
128	5.25 (9.16)	4.48 (7.77)		
256	7.57 (12.26)	4.52 (7.70)		
512	16.51 (20.15)	5.24 (8.02)		

Table 4: Beta Distribution with Possibility of Losing Information: Detailed values of squared error  $\|\hat{\theta} - \theta_0\|^2$ . In each cell, the first number is the mean of squared error in  $K = 50$  experiments and the number in the brackets is the standard deviation of squared error.

the log-likelihood function

$$m(x; \mu, \sigma^2) = -\frac{(x - \mu)^2}{2\sigma^2} - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log \sigma^2.$$

Figure 4 and Table 5 show that one-step estimator is asymptotically efficient while simple averaging estimator is absolutely not. It is worth noting that the resampled averaging estimator is not asymptotic efficient though it is better than simple averaging estimator. When the number of samples  $N$  is relatively small, the one-step estimator is worse than centralized estimator. When the number of samples  $N$  grows large, the differences between the one-step estimator and the centralized estimator become minimal in terms of both mean squared error and standard deviation. However, the error of the simple averaging estimator is significant larger than both the one-step estimator and the centralized estimator. When the sample size  $N = 4^9 \approx 250,000$ , the mean squared error of the simple averaging estimator is more than twice of that of the one-step and the centralized estimator.

no. of machines	no. of samples	simple avg	resampled avg	one-step	centralized
8	64	3.022104 (4.385627)	2.153958 (3.458645)	1.694668 (2.882794)	1.388959 (2.424813)
16	256	0.739784 (1.209734)	0.392389 (0.739390)	0.318765 (0.621990)	0.286175 (0.566140)
32	1024	0.118766 (0.151695)	0.041050 (0.053808)	0.034494 (0.046586)	0.032563 (0.045779)
64	4096	0.026839 (0.046612)	0.016519 (0.030837)	0.014255 (0.029258)	0.014414 (0.030533)
128	16384	0.010996 (0.019823)	0.004542 (0.009089)	0.004329 (0.009453)	0.004357 (0.009315)
256	65536	0.002909 (0.005785)	0.001158 (0.002733)	0.001105 (0.002779)	0.001099 (0.002754)
512	262144	0.000843 (0.001426)	0.000461 (0.000744)	0.000376 (0.000596)	0.000376 (0.000595)

Table 5: Gaussian Distribution with Unknown Mean and Variance: Detailed values of squared error  $\|\hat{\theta} - \theta_0\|^2$ . In each cell, the first number is the mean of squared error in  $K = 50$  experiments and the number in the brackets is the standard deviation of squared error.

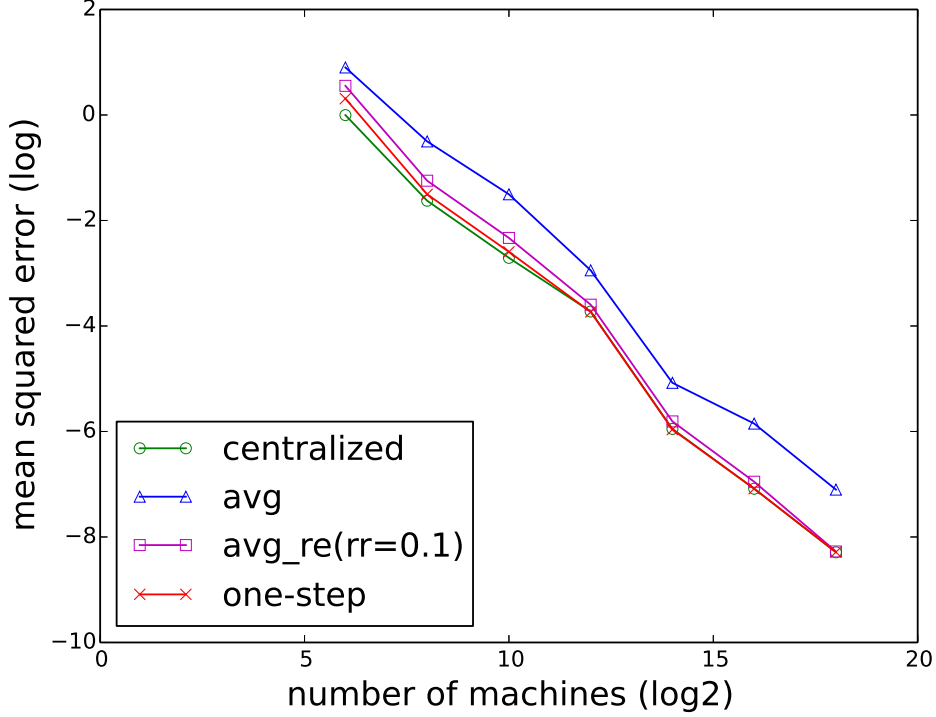


Figure 4: Gaussian Distribution with Unknown Mean and Variance: The log error  $\log \|\theta - \theta_0\|^2$  versus the log number of machines ( $\log_2 k$ ), with fifty repeated experiments for each  $N$ , where  $\theta_0$  is the true parameter. The “avg”, “avg-re” and “one-step” denote  $\theta^{(0)}$ ,  $\theta_{re}^{(0)}$  with resampling ratio  $rr = 10\%$  and  $\theta^{(1)}$ , respectively. The “centralized” denotes the maximum likelihood estimator with the entire data. The sample size is fixed to be  $N = k^2$ .

## 5 Conclusion

The M-estimator is a fundamental and high-impact methodology in statistics. The classic M-estimator theory is based on the assumption that the entire data are available at a central location, and can be processed/computed without considering communication issues. In many modern estimation problems arising in contemporary sciences and engineering, the classical notion of asymptotic optimality suffers from a significant deficiency: it requires access to all data. The asymptotic property when the data has to be dealt with distributively is underdeveloped. In this paper, we close this gap by considering a distributed one-step estimator.

Our one-step estimator builds on the existing *averaging* estimator. In a nutshell, after obtaining an averaging estimator, this initial estimate is broadcasted to local machines, to facilitate their computation of gradients and Hessians of their objective functions. By doing so, the data do *not* need to be transmitted to the central machine. The central machine then collects the locally estimated gradients and Hessians, to produce a global estimate of the overall gradient and overall Hessian. Consequently, a one-step update of the initial estimator is implemented. Just like the one-step approach has improved the estimator in the classical (non-distributed) setting, we found that the one-step approach can improve the performance of an estimator under the distributed setting, both theoretically and numerically.

Besides the works that have been cited earlier, there are many other results that are in the relevant literature, however they may not be directly technically linked to what’s been done here. We discuss their influence and insights in the next few paragraphs.

An interesting split-and-merge Bayesian approach for variable selection under linear models is proposed in Song and Liang [2015]. The method firstly split the ultrahigh dimensional data set into a number of lower dimensional subsets and select relevant variables from each of the subsets, and then aggregate the variables selected from

each subset and then select relevant variables from the aggregated data set. Under mild conditions, the authors show that the proposed approach is consistent, i.e., the underlying true model will be selected in probability 1 as the sample size becomes large. This work differs from all the other approaches that we discussed in this paper: it splits the variables, while all other approaches that we referenced (including ours) split the data according to observations. This paper certainly is in line with our research, however takes a very distinct angle.

An interesting piece of work that combines distributed statistical inference and information theory in communication is presented in Zhang et al. [2013a]. Their current results need to rely on special model settings: uniform location family  $\mathcal{U} = \{P_\theta, \theta \in [-1, 1]\}$ , where  $P_\theta$  denotes the uniform distribution on the interval  $[\theta - 1, \theta + 1]$ , or Gaussian location families  $N_d([-1, 1]^d) = \{N(\theta, \sigma^2 I_{d \times d}) \mid \theta \in \Theta = [-1, 1]^d\}$ . It will be interesting to see whether or not more general results are feasible.

Nowak [2003] proposed a distributed expectation-maximization (EM) algorithm for density estimation and clustering in sensor networks. Though the studied problem is technically different from ours, it provides an inspiring historic perspective: distributed inference has been studied more than ten years ago.

Neiswanger et al. [2013] propose an asymptotically exact, embarrassingly parallel MCMC method by approximating each sub-posterior with Gaussian density, Gaussian kernel or weighted Gaussian kernel. They prove the asymptotic correctness of their estimators and bound rate of convergence. Our paper does not consider the MCMC framework. The analytical tools that they used in proving their theorems are of interests.

Wang et al. [2014] propose a distributed variable selection algorithm, which accepts a variable if more than half of machines select that variable. They give upper bounds for the success probability and Mean Squared Error (MSE) of estimator. This work bears similarity with Song and Liang [2015] and Chen and Xie [2014], however with somewhat different emphases.

Kleiner et al. [2014] propose a scalable bootstrap (named ‘bag of little bootstraps’ (BLB)) for massive data to assess the quality of estimators. They also demonstrate its favorable statistical performance through both theoretical analysis and simulation studies. A comparison with this work will be interesting, however not included here.

Zhao et al. [2014] consider a partially linear framework for massive heterogeneous data and propose an aggregation type estimator for the commonality parameter that possesses the minimax optimal bound and asymptotic distribution when number of sub-populations does not grow too fast.

A recent work [Arjevani and Shamir, 2015] shed interesting new light into the distributed inference problem. The authors studied the fundamental limits to communication-efficient distributed methods for convex learning and optimization, under different assumptions on the information available to individual machines, and the types of functions considered. The current problem formulation is more numerical than statistical properties. Their idea may lead to interesting counterparts in statistical inference.

Besides estimation, other distributed statistical technique may be of interests, such as the distributed principal component analysis Balcan et al. [2014]. This paper does not touch this line of research.

Various researchers have studied communication-efficient algorithms for statistical estimation (e.g., see the papers Dekel et al. [2012], Balcan et al. [2012], Wainwright [2014], McDonald et al. [2010] and references therein). They were not discussed in details here, because they are pretty much discussed/compared in other references of this paper.

There is now a rich and well-developed body of theory for bounding and/or computing the minimax risk for various statistical estimation problems, e.g., see Yang and Barron [1999] and references therein. In several cited references, researchers have started to derive the optimal minimax rate for estimators under the distributed inference setting. This will be an exciting future research direction.

## References

- Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. Technical report, ArXiv, June 2015. <http://arxiv.org/abs/1506.01900>.
- Maria-Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. Arxiv, Georgia Institute of Technology, April 2012.



- Maria-Florina Balcan, Vandana Kanchanapally, Yingyu Liang, and David Woodruff. Improved distributed principal component analysis. Technical report, ArXiv, August 2014. <http://arxiv.org/abs/1408.5823>.
- Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*, 2015.
- Peter J. Bickel. One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975.
- Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- Shaobin Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.
- Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684, 2014.
- James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, JJ Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, and et al. Spanner: Googles globally distributed database. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, 2012.
- Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.
- Jianqing Fan and Jianwei Chen. One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):927–943, 1999.
- Mostafa El Gamal and Lifeng Lai. Are Slepian-Wolf rates necessary for distributed parameter estimation? Technical report, arXiv, August 2015. <http://arxiv.org/abs/1508.02765>.
- Martin Jaggi, Virginia Smith, Martin Takáč, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I. Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 3068–3076, 2014.
- Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.
- Serge Lang. *Real and functional analysis*, volume 142. Springer Science & Business Media, 1993.
- Jason D. Lee, Yuekai Sun, Qiang Liu, and Jonathan E. Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015.
- Qiang Liu and Alex T. Ihler. Distributed estimation, information loss and exponential families. In *Advances in Neural Information Processing Systems*, pages 1098–1106, 2014.
- Ryan McDonald, Keith Hall, and Gideon Mann. Distributed training strategies for the structured perceptron. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.
- Siddharth Mitra, Mayank Agrawal, Amit Yadav, Niklas Carlsson, Derek Eager, and Anirban Mahanti. Characterizing web-based video sharing workloads. *ACM Transactions on the Web*, 5(2), May 2011.
- Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*, 2013.
- Robert D. Nowak. Distributed EM algorithms for density estimation and clustering in sensor networks. *IEEE Transactions on Signal Processing*, 51(8):2245–2253, 2003.

- Jonathan Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *arXiv preprint arXiv:1407.2724*, 2014.
- Haskell P. Rosenthal. On the subspaces of  $L^p$  ( $p > 2$ ) spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.
- Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1000–1008, 2014.
- Qifan Song and Faming Liang. A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *J. R. Statist. Soc. B*, 77, Part 5:947–972, 2015.
- Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- Aad W. van der Vaart. *Asymptotic statistics (Cambridge series in statistical and probabilistic mathematics)*. Cambridge University Press, 2000.
- Martin Wainwright. Constrained forms of statistical minimax: Computation, communication, and privacy. In *Proceedings of International Congress of Mathematicians*, 2014.
- Xiangyu Wang, Peichao Peng, and David B. Dunson. Median selection subset aggregation for parallel inference. In *Advances in Neural Information Processing Systems*, pages 2195–2203, 2014.
- Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.
- Yuchen Zhang, John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. Technical report, UC Berkeley, 2013a. Presented at the NIPS Conference 2013.
- Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013b.
- Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *arXiv preprint arXiv:1410.8570*, 2014.
- Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J. Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2010.
- Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.

## Appendix Overview

The appendix is organized as follows. In Section A, we analyze the upper bounds of sum of i.i.d. random vectors and random matrices, which will be useful in later proofs. In Section B, we derive the upper bounds of the local M-estimators and the simple averaging estimator. We present the proofs to Theorem 3.4 and Theorem 3.5 in Section C and Section D, respectively. A proof of Corollary 3.7 will be in Section E.

## A Bounds on Gradient and Hessian

In order to establish the convergence of gradients and Hessians of the empirical criterion function to those of population criterion function, which is essential for the later proofs, we will present some results on the upper bound of sums of i.i.d. random vectors and random matrices. We start with stating a useful inequality on the sum of independent random variables from Rosenthal [1970].

**Lemma A.1** (Rosenthal's Inequality, Rosenthal [1970], Theorem 3). *For  $q > 2$ , there exists constant  $C(q)$  depending only on  $q$  such that if  $X_1, \dots, X_n$  are independent random variables with  $\mathbb{E}[X_j] = 0$  and  $\mathbb{E}[|X_j|^q] < \infty$  for all  $j$ , then*

$$(\mathbb{E}[\sum_{j=1}^n X_j^q])^{1/q} \leq C(q) \max \left\{ (\sum_{j=1}^n \mathbb{E}[|X_j|^q])^{1/q}, (\sum_{j=1}^n \mathbb{E}[|X_j|^2])^{1/2} \right\}$$

Equipped with the above lemma, we can bound the moments of mean of random vectors.

**Lemma A.2.** *Let  $X_1, \dots, X_n \in \mathbb{R}^d$  be i.i.d. random vectors with  $\mathbb{E}[X_i] = \mathbf{0}$ . And there exists some constants  $G > 0$  and  $q_0 \geq 2$  such that  $\mathbb{E}[\|X_i\|^{q_0}] < G^{q_0}$ . Let  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ , then for  $1 \leq q \leq q_0$ , we have*

$$\mathbb{E}[\|\bar{X}\|^q] \leq \frac{C_v(q, d)}{n^{q/2}} G^q,$$

where  $C(q, d)$  is a constant depending solely on  $q$  and  $d$ .

*Proof.* The main idea of this proof is to transform the sum of random vectors into the sum of random variables and then apply Lemma A.1. Let  $X_{i,j}$  denote the  $j$ -th component of  $X_i$  and  $\bar{X}_j$  denote the  $j$ -th component of  $\bar{X}$ .

(1) Let us start with a simpler case in which  $q = 2$ .

$$\begin{aligned} \mathbb{E}[\|\bar{X}\|^2] &= \sum_{j=1}^d \mathbb{E}[\bar{X}_j^2] = \sum_{j=1}^d \sum_{i=1}^n \mathbb{E}[X_{i,j}^2/n^2] \\ &= \sum_{j=1}^d \mathbb{E}[X_{1,j}^2]/n = n^{-1} \mathbb{E}[\|X_1\|^2] \leq n^{-1} G^2, \end{aligned}$$

The last inequality holds because  $\mathbb{E}[\|X_1\|^q] \leq (\mathbb{E}[\|X_1\|^{q_0}])^{q/q_0} \leq G^q$  for  $1 \leq q \leq q_0$  by Hölder's inequality.

(2) When  $1 \leq q < 2$ , we have

$$\mathbb{E}[\|\bar{X}\|^q] \leq (\mathbb{E}[\|\bar{X}\|^2])^{q/2} \leq n^{-q/2} G^q.$$

(3) For  $2 < q \leq q_0$ , with some simple algebra, we have

$$\begin{aligned} \mathbb{E}[\|\bar{X}\|^q] &= \mathbb{E} \left[ \left( \sum_{j=1}^d |\bar{X}_j|^2 \right)^{q/2} \right] \\ &\leq \mathbb{E} \left[ \left( d \max_{1 \leq j \leq d} |\bar{X}_j|^2 \right)^{q/2} \right] = d^{q/2} \mathbb{E} \left[ \max_{1 \leq j \leq d} |\bar{X}_j|^q \right] \\ &\leq d^{q/2} \mathbb{E} \left[ \sum_{j=1}^d |\bar{X}_j|^q \right] = d^{q/2} \sum_{j=1}^d \mathbb{E} [|\bar{X}_j|^q]. \end{aligned}$$

As a continuation, we have

$$\begin{aligned}
\mathbb{E}[\|\bar{X}\|^q] &\leq d^{q/2} \sum_{j=1}^d \mathbb{E}[\|\bar{X}_j\|^q] \\
&\leq d^{q/2} \sum_{j=1}^d [C(q)]^q \max \left\{ \sum_{i=1}^n \mathbb{E}[|X_{i,j}/n|^q], \left( \sum_{i=1}^n \mathbb{E}[|X_{i,j}/n|^2] \right)^{q/2} \right\} \quad (\text{Lemma A.1}) \\
&= d^{q/2} [C(q)]^q \sum_{j=1}^d \max \left\{ \frac{\mathbb{E}[\|X_{1,j}\|^q]}{n^{q-1}}, \frac{(\mathbb{E}[\|X_{1,j}\|^2])^{q/2}}{n^{q/2}} \right\} \\
&\leq d^{q/2+1} [C(q)]^q \max \left\{ \frac{\mathbb{E}[\|X_1\|^q]}{n^{q-1}}, \frac{(\mathbb{E}[\|X_1\|^2])^{q/2}}{n^{q/2}} \right\} \quad (\text{since } \mathbb{E}[|X_{1,j}|^q] \leq \mathbb{E}[\|X_1\|^q]) \\
&\leq d^{q/2+1} [C(q)]^q \max \left\{ \frac{G^q}{n^{q-1}}, \frac{G^q}{n^{q/2}} \right\} \quad (\text{Hölder's inequality}) \\
&= \frac{d^{q/2+1} [C(q)]^q}{n^{q/2}} G^q. \quad (q-1 > q/2 \text{ when } q > 2)
\end{aligned}$$

To complete this proof, we just need to set  $C_v(q, d) = d^{q/2+1} [C(q)]^q$ .

□

To bound the moment of the mean of i.i.d. random matrices, let us consider another matrix norm – Frobenius norm  $\|\cdot\|_F$ , i.e.,

$$\|A\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}, \forall A \in \mathbb{R}^{d \times d}.$$

Note that

$$\|A\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2} = \sqrt{\text{trace}(A^t A)} \geq \sqrt{\sup_{u \in \mathbb{R}^d: \|u\| \leq 1} \|A^t A u\|} = \|A\|,$$

and

$$\|A\|_F \leq \sqrt{d \sup_{u \in \mathbb{R}^d: \|u\| \leq 1} \|A^t A u\|} = \sqrt{d} \|A\|.$$

With Frobenius norm, we can regard a random matrix  $X \in \mathbb{R}^{d \times d}$  as a random vector in  $\mathbb{R}^{d^2}$  and apply Lemma A.2 to obtain the following lemma.

**Lemma A.3.** *Let  $X_1, \dots, X_n \in \mathbb{R}^{d \times d}$  be i.i.d. random matrices with  $\mathbb{E}[X_i] = \mathbf{0}_{d \times d}$ . Let  $\|X_i\|$  denote the norm of  $X_i$ , which is defined as its maximal singular value. Suppose  $\mathbb{E}[\|X_i\|^{q_0}] \leq H^{q_0}$ , where  $q_0 \geq 2$  and  $H > 0$ . Then for  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  and  $1 \leq q \leq q_0$ , we have*

$$\mathbb{E}[\|\bar{X}\|^q] \leq \frac{C_m(q, d)}{n^{q/2}} H^q,$$

where  $C_m(q, d)$  is a constant depending on  $q$  and  $d$  only.

*Proof.* By the fact  $\|A\|_F \leq \sqrt{d} \|A\|$ , we have

$$\mathbb{E}[\|X_i\|_F^{q_0}] \leq \mathbb{E}[\|\sqrt{d} X_i\|^{q_0}] \leq (\sqrt{d} H)^{q_0}.$$

Then by the fact  $\|A\| \leq \|A\|_F$  and Lemma A.2, we have

$$\mathbb{E}[\|\bar{X}\|^q] \leq \mathbb{E}[\|\bar{X}\|_F^q] \leq \frac{C_v(q, d^2)}{n^{q/2}} (\sqrt{d} H)^q = \frac{C_v(q, d^2) d^{\frac{q}{2}}}{n^{q/2}} H^q.$$

In the second inequality, we treat  $\bar{X}$  as a  $d^2$ -dimensional random vector and then apply Lemma A.2. Then the proof can be completed by setting  $C_m(q, d) = C_v(q, d^2) d^{\frac{q}{2}}$ . □

## B Error Bound of Local M-estimator and Simple Averaging Estimator

Since the simple averaging estimator is the average of all local estimators and the one-step estimator is just a single Newton-Raphson update from the simple averaging estimator. Thus, it is natural to study the upper bound of the mean squared error (MSE) of a local M-estimator and the upper bound of the MSE of the simple averaging estimator. The main idea in the following proof is similar to the thread in the proof of **Theorem 1** in Zhang et al. [2013b], but the conclusions are different. Besides, in the following proof, we use a correct analogy of mean value theorem for vector-valued functions.

### B.1 Bound the Error of Local M-estimators $\theta_i, i = 1, \dots, k$

In this subsection, we would like to analyze the mean squared error of a local estimator  $\theta_i = \arg \max_{\theta \in \Theta} M_i(\theta)$ ,  $i = 1, \dots, k$  and prove the following lemma in the rest of this subsection.

**Lemma B.1.** *Let  $\Sigma = \ddot{M}_0(\theta_0)^{-1} \mathbb{E}[\dot{m}(X; \theta_0) \dot{m}(X; \theta_0)^t] \ddot{M}_0(\theta_0)^{-1}$ , where the expectation is taken with respect to  $X$ . Under Assumption 3.1, 3.2 and 3.3, for each  $i = 1, \dots, k$ , we have*

$$\mathbb{E}[\|\theta_i - \theta_0\|^2] \leq \frac{2}{n} \text{Tr}(\Sigma) + O(n^{-2}).$$

Since  $\dot{M}_i(\theta_i) = 0$ , by Theorem 4.2 in Chapter XIII of Lang [1993], we have

$$\begin{aligned} 0 &= \dot{M}_i(\theta_i) = \dot{M}_i(\theta_0) + \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho [\theta_i - \theta_0] \\ &= \dot{M}_i(\theta_0) + \ddot{M}_0(\theta_0)[\theta_i - \theta_0] + \left[ \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_0(\theta_0) \right] [\theta_i - \theta_0] \\ &= \dot{M}_i(\theta_0) + \ddot{M}_0(\theta_0)[\theta_i - \theta_0] + \left[ \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right] [\theta_i - \theta_0] \\ &\quad + [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0] \quad (\text{subtract and add } \ddot{M}_i(\theta_0)), \end{aligned}$$

**Remark.** *Here, it is worth noting that there is no analogy of mean value theorem for vector-valued functions, which implies that there does not necessarily exist  $\theta'$  lying on the line between  $\theta_i$  and  $\theta_0$  satisfying  $\dot{M}_i(\theta_i) - \dot{M}_i(\theta_0) = \ddot{M}_i(\theta')(\theta_i - \theta_0)$ . Numerous papers make errors by claiming such  $\theta'$  lies between  $\theta_i$  and  $\theta_0$ .*

If last two terms in above equation are reasonably small, this lemma follows immediately. So, our strategy is as follows. First, we show that the mean squared error of both  $[\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0)][\theta_i - \theta_0]$  and  $[\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0]$  is small under some “good” events. Then we will show the probability of “bad” events is small enough. And Lemma B.1 will follow by the fact that  $\Theta$  is compact.

Suppose  $S_i = \{x_1, \dots, x_n\}$  is the data set on local machine  $i$ . Let us define some good events:

$$\begin{aligned} E_1 &= \left\{ \frac{1}{n} \sum_{j=1}^n L(x_j) \leq 2L \right\}, \\ E_2 &= \left\{ \left\| \ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0) \right\| \leq \lambda/4 \right\}, \\ E_3 &= \left\{ \left\| \dot{M}_i(\theta_0) \right\| \leq \frac{\lambda}{2} \delta' \right\}, \end{aligned}$$

where  $\delta' = \min(\delta, \frac{\lambda}{8L})$ ,  $\lambda$  is the constant in Assumption 3.2 and  $L$  and  $\delta$  are the constants in Assumption 3.3. We will show that event  $E_1$  and  $E_2$  ensure that  $M_i(\theta)$  is strictly concave at a neighborhood of  $\theta_0$ . And we will also show that in event  $E_3$ ,  $\theta_i$  is fairly close to  $\theta_0$ . Let  $E = E_1 \cap E_2 \cap E_3$ , then we have the following lemma:

**Lemma B.2.** *Under event  $E$ , we have*

$$\|\theta_i - \theta_0\| \leq \frac{4}{\lambda} \|\dot{M}_i(\theta_0)\|$$

*Proof.* First, we will show  $\ddot{M}_i(\theta)$  is a negative definite matrix over a ball centered at  $\theta_0$ :  $B_{\delta'} = \{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta'\} \subset B_\delta$ . For any fixed  $\theta \in B_{\delta'}$ , we have

$$\begin{aligned} \left\| \ddot{M}_i(\theta) - \ddot{M}_0(\theta_0) \right\| &\leq \left\| \ddot{M}_i(\theta) - \ddot{M}_i(\theta_0) \right\| + \left\| \ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0) \right\| \\ &\leq 2L\|\theta - \theta_0\| + \frac{\lambda}{4} \leq \lambda/4 + \lambda/4 = \lambda/2, \end{aligned}$$

where we apply event  $E_1$ , Assumption 3.3 and the fact that  $\delta' = \min(\delta, \frac{\lambda}{8L})$  on the first term and event  $E_2$  on the second term. Since  $\ddot{M}_0(\theta_0)$  is negative definite by Assumption 3.2, above inequality implies that  $\ddot{M}_i(\theta)$  is negative definite for all  $\theta \in B_{\delta'}$  and

$$\sup_{u \in \mathbb{R}^d: \|u\| \leq 1} u^t \ddot{M}_i(\theta) u \leq -\lambda/2. \quad (\text{B.11})$$

With negative definiteness of  $\ddot{M}_i(\theta)$ ,  $\theta \in B_{\delta'}$ , event  $E_3$  and concavity of  $M_i(\theta)$ ,  $\theta \in \Theta$ , we have

$$\frac{\lambda}{2} \delta' \stackrel{E_3}{\geq} \|\dot{M}_i(\theta_0)\| = \|\dot{M}_i(\theta_0) - \dot{M}_i(\theta_i)\| \stackrel{(\text{B.11})}{\geq} \frac{\lambda}{2} \|\theta_i - \theta_0\|.$$

Thus, we know  $\|\theta_i - \theta_0\| \leq \delta'$ , or equivalently,  $\theta_i \in B_{\delta'}$ . Then by applying Taylor's Theorem on  $M_i(\theta)$  at  $\theta_0$ , we have

$$M_i(\theta_i) \stackrel{(\text{B.11})}{\leq} M_i(\theta_0) + \dot{M}_i(\theta_0)^t(\theta_i - \theta_0) - \frac{\lambda}{4} \|\theta_i - \theta_0\|^2.$$

Thus, as  $M_i(\theta_0) \leq M_i(\theta_i)$  by definition,

$$\begin{aligned} \frac{\lambda}{4} \|\theta_i - \theta_0\|^2 &\leq M_i(\theta_0) - M_i(\theta_i) + \dot{M}_i(\theta_0)^t(\theta_i - \theta_0) \\ &\leq \|\dot{M}_i(\theta_0)\| \|\theta_i - \theta_0\|, \end{aligned}$$

which implies

$$\|\theta_i - \theta_0\| \leq \frac{4}{\lambda} \|\dot{M}_i(\theta_0)\|.$$

□

For  $1 \leq q \leq 8$ , we can bound  $\mathbb{E}[\|\dot{M}_i(\theta_0)\|^q]$  by Lemma A.2 and Assumption 3.3,

$$\mathbb{E}[\|\dot{M}_i(\theta_0)\|^q] \leq \frac{C_v(q, d)}{n^{q/2}} G^q,$$

where  $C_v(q, d)$  is a constant depending on  $q$  and  $d$  only. Then by conditioning on event  $E$ , we have

$$\begin{aligned} \mathbb{E}[\|\theta_i - \theta_0\|^q] &= \mathbb{E}[\|\theta_i - \theta_0\|^q 1_{(E)}] + \mathbb{E}[\|\theta_i - \theta_0\|^q 1_{(E^c)}] \\ &\leq \frac{4^q}{\lambda^q} \mathbb{E}[\|\dot{M}_i(\theta_0)\|^q] + D^q \Pr(E^c) \\ &\leq \frac{4^q}{\lambda^q} \frac{C_v(q, d)}{n^{q/2}} G^q + D^q \Pr(E^c). \end{aligned}$$

If we can show  $\Pr(E^c) = O(n^{-\frac{q}{2}})$ , then  $\mathbb{E}[\|\theta_i - \theta_0\|^q] = O(n^{-\frac{q}{2}})$  follows immediately.

**Lemma B.3.** *Under Assumption 3.3, we have*

$$\Pr(E^c) = O(n^{-4}).$$

*Proof.* Under Assumption 3.3, by applying Lemma A.2 and A.3, we can bound the moments of  $\dot{M}_i(\theta_0)$  and  $\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)$ . Rigorously, for  $1 \leq q \leq 8$ , we have

$$\begin{aligned}\mathbb{E} \left[ \|\dot{M}_i(\theta_0)\|^q \right] &\leq \frac{C_v(q, d)}{n^{q/2}} G^q, \\ \mathbb{E} \left[ \left\| \ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0) \right\|^q \right] &\leq \frac{C_m(q, d)}{n^{q/2}} H^q.\end{aligned}$$

Therefore, by Markov's inequality, we have

$$\begin{aligned}\Pr(E^c) &= \Pr(E_1^c \cup E_2^c \cup E_3^c) \leq \Pr(E_1^c) + \Pr(E_2^c) + \Pr(E_3^c) \\ &\leq \frac{\mathbb{E} \left[ \left| \frac{1}{n} \sum_{j=1}^n L(x_j) - \mathbb{E}[L(x)] \right|^8 \right]}{L^8} + \frac{\mathbb{E} \left[ \left\| \ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0) \right\|^8 \right]}{(\lambda/4)^8} + \frac{\mathbb{E} \left[ \|\dot{M}_i(\theta_0)\|^8 \right]}{(\lambda\delta'/2)^8} \\ &\leq O\left(\frac{1}{n^4}\right) + O\left(\frac{1}{n^4}\right) + O\left(\frac{1}{n^4}\right) = O(n^{-4}).\end{aligned}$$

□

Now, we have showed that for  $1 \leq q \leq 8$ ,

$$\mathbb{E}[\|\theta_i - \theta_0\|^q] \leq \frac{4^q C_v(q, d)}{\lambda^q n^{q/2}} G^q + O(n^{-4}) = O(n^{-\frac{q}{2}}). \quad (\text{B.12})$$

Until now,  $[\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0]$  has been well bounded. Next, we will consider the moment bound of  $\int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0)$ .

**Lemma B.4.** Under assumption 3.3, for  $1 \leq q \leq 4$ ,

$$\mathbb{E} \left[ \left\| \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right\|^q \right] \leq L^q \frac{4^q}{\lambda^q} \frac{\sqrt{C_v(2q, d)}}{n^{q/2}} G^q + O(n^{-2}) = O(n^{-q/2}).$$

*Proof.* By Minkowski's integral inequality, we have

$$\begin{aligned}\mathbb{E} \left[ \left\| \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right\|^q \right] &\leq \mathbb{E} \left[ \int_0^1 \left\| \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) - \ddot{M}_i(\theta_0) \right\|^q d\rho \right] \\ &= \int_0^1 \mathbb{E} \left[ \left\| \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) - \ddot{M}_i(\theta_0) \right\|^q \right] d\rho.\end{aligned}$$

For simplicity of notation, we use  $\theta' = (1-\rho)\theta_0 + \rho\theta_i$  in this proof. When event  $E$  holds, we have

$$\|\theta' - \theta_0\| = \|\rho(\theta_i - \theta_0)\| \leq \rho\delta' \leq \delta,$$

which means that  $\theta' \in B_\delta, \forall \rho \in [0, 1]$ . Thus, because of the convexity of the matrix norm  $\|\cdot\|$ , we can apply Jensen's inequality and Assumption 3.3 and get

$$\left\| \ddot{M}_i(\theta') - \ddot{M}_i(\theta_0) \right\|^q \leq \frac{1}{n} \sum_{j=1}^n \left\| \ddot{m}(x_j; \theta') - \ddot{m}(x_j; \theta_0) \right\|^q \leq \frac{1}{n} \sum_{j=1}^n L(x_j)^q \|\theta' - \theta_0\|^q.$$

Then apply Hölder's inequality,

$$\begin{aligned}\mathbb{E} \left[ \left\| \ddot{M}_i(\theta') - \ddot{M}_i(\theta_0) \right\|^q 1_{(E)} \right] &\leq \left\{ \mathbb{E} \left[ \left( \frac{1}{n} \sum_{j=1}^n L(x_j)^q \right)^2 \right] \right\}^{1/2} \left\{ \mathbb{E}[\|\theta' - \theta_0\|^{2q}] \right\}^{1/2} \\ &\stackrel{\text{Jensen's}}{\leq} C(q) L^q \rho^q \left\{ \mathbb{E}[\|\theta_i - \theta_0\|^{2q}] \right\}^{1/2} \\ &\stackrel{(\text{B.12})}{\leq} C(q) L^q \frac{4^q \sqrt{C(2q, d)} G^q}{\lambda^q n^{q/2}} + O(n^{-2}).\end{aligned}$$

When event  $E$  does not hold, we know that  $\left\| \ddot{M}_i(\theta') - \ddot{M}_i(\theta_0) \right\|^q$  must be finite by the assumption that  $\Theta$  is compact and  $\ddot{M}_i(\theta)$  is continuous. By Lemma B.3, the probability that event  $E$  does not hold is bounded by  $O(n^{-4})$ , which implies,

$$\mathbb{E} \left[ \left\| \ddot{M}_i(\theta') - \ddot{M}_i(\theta_0) \right\|^q \right] \leq C(q) L^q \frac{4^q \sqrt{C(2q, d)} G^q}{\lambda^q n^{q/2}} + O(n^{-2}) + O(n^{-4}) = O(n^{-q/2}).$$

Therefore, we have

$$\begin{aligned} \mathbb{E} \left[ \left\| \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right\|^q \right] &\leq \int_0^1 \mathbb{E} \left[ \left\| \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) - \ddot{M}_i(\theta_0) \right\|^q \right] d\rho \\ &\leq C(q) L^q \frac{4^q \sqrt{C(2q, d)} G^q}{\lambda^q n^{q/2}} + O(n^{-2}) + O(n^{-4}) = O(n^{-q/2}). \end{aligned}$$

□

Now, recall that we have

$$\begin{aligned} 0 = \dot{M}_i(\theta_0) + \ddot{M}_0(\theta_0)[\theta_i - \theta_0] &+ \left[ \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right] [\theta_i - \theta_0] \\ &+ [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0]. \end{aligned} \quad (\text{B.13})$$

For the sum of last two terms, we have

$$\begin{aligned} &\mathbb{E} \left[ \left\| \left[ \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right] [\theta_i - \theta_0] + [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0] \right\|^2 \right] \\ &\leq 2\mathbb{E} \left[ \left\| \left[ \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right] [\theta_i - \theta_0] \right\|^2 \right] \\ &\quad + 2\mathbb{E} \left[ \left\| [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)][\theta_i - \theta_0] \right\|^2 \right] \quad (\text{since } (a+b)^2 \leq 2a^2 + 2b^2) \\ &\leq 2(\mathbb{E} \left[ \left\| \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right\|^4 \right])^{1/2} (\mathbb{E}[\|\theta_i - \theta_0\|^4])^{1/2} \\ &\quad + 2(\mathbb{E}[\|\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)\|^4])^{1/2} (\mathbb{E}[\|\theta_i - \theta_0\|^4])^{1/2} \quad (\text{Hölder's inequality}) \\ &= O(n^{-2}) + O(n^{-2}) \quad (\text{Lemma A.3 \& B.4 and (B.12)}) \\ &= O(n^{-2}). \end{aligned}$$

Until now, we have established the upper bound for the mean squared error of local M-estimators,

$$\mathbb{E}[\|\theta_i - \theta_0\|^2] \leq \frac{2}{n} \text{Tr}(\Sigma) + O(n^{-2}),$$

for  $i = 1, \dots, k$ .

## B.2 Bound the Error of Simple Averaging Estimator $\theta^{(0)}$

Next, we will study the mean squared error of simple averaging estimator,

$$\theta^{(0)} = \frac{1}{k} \sum_{i=1}^k \theta_i.$$

We start with a lemma, which bounds the bias of local M-estimator  $\theta_i, i = 1, \dots, k$ .



**Lemma B.5.** *There exists some constant  $\tilde{C} > 0$  such that for  $i = 1, \dots, k$ , we have*

$$\|\mathbb{E}[\theta_i - \theta_0]\| \leq \frac{\tilde{C}}{n} + O(n^{-2}),$$

where  $\tilde{C} = 16[C_v(4, d)]^{\frac{1}{4}} \sqrt{C_v(2, d)} \lambda^{-3} G^2 L + 4 \sqrt{C_m(2, d)} \sqrt{C_v(2, d)} \lambda^{-2} G H$ .

*Proof.* The main idea of this proof is to use equation (B.13) and apply the established error bounds of Hessian and the aforementioned local m-estimators. By equation (B.13) and fact  $\mathbb{E}[M_i(\theta_0)] = 0$ , we have

$$\begin{aligned} & \|\mathbb{E}[\theta_i - \theta_0]\| \\ = & \|\mathbb{E}\{\ddot{M}_0(\theta_0)^{-1} \left[ \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right] [\theta_i - \theta_0] \\ & + \ddot{M}_0(\theta_0)^{-1} [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)] [\theta_i - \theta_0]\}\| \\ \leq & \|\mathbb{E}\{\ddot{M}_0(\theta_0)^{-1} \left[ \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right] [\theta_i - \theta_0]\}\| \\ & + \|\mathbb{E}\{\ddot{M}_0(\theta_0)^{-1} [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)] [\theta_i - \theta_0]\}\| \\ \stackrel{\text{Jensen's}}{\leq} & \mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0)^{-1} \left[ \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right] [\theta_i - \theta_0] \right\|^2 \right] \\ & + \mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0)^{-1} [\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)] [\theta_i - \theta_0] \right\|^2 \right] \\ \stackrel{\text{Assumption 3.2}}{\leq} & \lambda^{-1} \mathbb{E} \left[ \left\| \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right\|^2 \|\theta_i - \theta_0\|^2 \right] \\ & + \lambda^{-1} \mathbb{E} \left[ \|\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)\|^2 \|\theta_i - \theta_0\|^2 \right] \\ \stackrel{\text{Hölder's}}{\leq} & \lambda^{-1} \mathbb{E} \left[ \left\| \int_0^1 \ddot{M}_i((1-\rho)\theta_0 + \rho\theta_i) d\rho - \ddot{M}_i(\theta_0) \right\|^2 \right]^{1/2} \mathbb{E}[\|\theta_i - \theta_0\|^2]^{1/2} \\ & + \lambda^{-1} \mathbb{E}[\|\ddot{M}_i(\theta_0) - \ddot{M}_0(\theta_0)\|^2]^{1/2} \mathbb{E}[\|\theta_i - \theta_0\|^2]^{1/2}. \end{aligned}$$

Then we can apply Lemma A.3 & B.4, and (B.12) to bound each term, thus, we have

$$\begin{aligned} \|\mathbb{E}[\theta_i - \theta_0]\| & \leq \lambda^{-1} \sqrt{L^2 \frac{4^2}{\lambda^2} \frac{\sqrt{C_v(4, d)}}{n} G^2 + O(n^{-2})} \sqrt{\frac{4^2}{\lambda^2} \frac{C_v(2, d)}{n} G^2 + O(n^{-4})} \\ & \quad + \lambda^{-1} \sqrt{\frac{C_m(2, d)}{n} H^2} \sqrt{\frac{4^2}{\lambda^2} \frac{C_v(2, d)}{n} G^2 + O(n^{-4})} \\ & \leq \lambda^{-1} \left[ L \frac{4}{\lambda} \frac{C_v(4, d)^{1/4}}{\sqrt{n}} G + O(n^{-\frac{3}{2}}) \right] \left[ \frac{4}{\lambda} \frac{\sqrt{C_v(2, d)}}{\sqrt{n}} G + O(n^{-\frac{7}{2}}) \right] \\ & \quad + \lambda^{-1} \frac{\sqrt{C_m(2, d)}}{\sqrt{n}} H \left[ \frac{4}{\lambda} \frac{\sqrt{C_v(2, d)}}{\sqrt{n}} G + O(n^{-\frac{7}{2}}) \right] \\ & = L \frac{4^2}{\lambda^3} \frac{C_v(4, d)^{1/4} \sqrt{C_v(2, d)}}{n} G^2 + O(n^{-2}) \\ & \quad + \frac{4}{\lambda^2} \frac{\sqrt{C_m(2, d)} \sqrt{C_v(2, d)}}{n} G H + O(n^{-4}). \end{aligned}$$

Let  $\tilde{C} = 16[C_v(4, d)]^{\frac{1}{4}} \sqrt{C_v(2, d)} \lambda^{-3} G^2 L + 4 \sqrt{C_m(2, d)} \sqrt{C_v(2, d)} \lambda^{-2} G H$ , then we have

$$\|\mathbb{E}[\theta_i - \theta_0]\| \leq \frac{\tilde{C}}{n} + O(n^{-2}).$$

□

Then we can show that the MSE of  $\theta^{(0)}$  could be bounded as follows.

**Lemma B.6.** *There exists some constant  $\tilde{C} > 0$  such that*

$$\mathbb{E}[\|\theta^{(0)} - \theta_0\|^2] \leq \frac{2}{N} \text{Tr}(\Sigma) + \frac{\tilde{C}^2 k^2}{N^2} + O(kN^{-2}) + O(k^3 N^{-3}),$$

where  $\tilde{C} = 16[C_v(4, d)]^{\frac{1}{4}} \sqrt{C_v(2, d)} \lambda^{-3} G^2 L + 4\sqrt{C_m(2, d)} \sqrt{C_v(2, d)} \lambda^{-2} GH$ .

*Proof.* The mean squared error of  $\theta^{(0)}$  could be decomposed into two parts: covariance and bias. Thus,

$$\begin{aligned} \mathbb{E}[\|\theta^{(0)} - \theta_0\|^2] &= \text{Tr}(\text{Cov}[\theta^{(0)}]) + \|\mathbb{E}[\theta^{(0)} - \theta_0]\|^2 \\ &= \frac{1}{k} \text{Tr}(\text{Cov}[\theta_1]) + \|\mathbb{E}[\theta_1 - \theta_0]\|^2 \\ &\leq \frac{1}{k} \mathbb{E}[\|\theta_1 - \theta_0\|^2] + \|\mathbb{E}[\theta_1 - \theta_0]\|^2, \end{aligned}$$

where the first term is well bounded by Lemma B.1 and the second term could be bounded by Lemma B.5. Thus, we know

$$\mathbb{E}[\|\theta^{(0)} - \theta_0\|^2] \leq \frac{2}{N} \text{Tr}(\Sigma) + \frac{\tilde{C}^2 k^2}{N^2} + O(kN^{-2}) + O(k^3 N^{-3}).$$

More generally, for  $1 \leq q \leq 8$ , we have

$$\begin{aligned} &\mathbb{E}[\|\theta^{(0)} - \theta_0\|^q] = \mathbb{E}[\|(\theta^{(0)} - \mathbb{E}[\theta^{(0)}]) + (\mathbb{E}[\theta^{(0)}] - \theta_0)\|^q] \\ &\leq 2^q \mathbb{E}[\|\theta^{(0)} - \mathbb{E}[\theta^{(0)}]\|^q] + 2^q \|\mathbb{E}[\theta^{(0)}] - \theta_0\|^q \quad (\text{since } (a+b)^q \leq 2^q a^q + 2^q b^q) \\ &= 2^q \mathbb{E}[\|\theta^{(0)} - \mathbb{E}[\theta^{(0)}]\|^q] + 2^q \|\mathbb{E}[\theta_1] - \theta_0\|^q \quad (\text{since } \mathbb{E}[\theta^{(0)}] = \mathbb{E}[\theta_1]) \\ &\leq 2^q \mathbb{E}[\|\theta^{(0)} - \theta_0\|^q] + 2^q \frac{\tilde{C}^q}{n^q} + O(n^{-q-1}) \\ &\stackrel{\text{Lemma A.2}}{\leq} 2^q \frac{C_v(q, d)}{k^{q/2}} \mathbb{E}[\|\theta_1 - \theta_0\|^q] + 2^q \frac{\tilde{C}^q}{n^q} + O(n^{-q-1}) \\ &\stackrel{(\text{B.12})}{\leq} 2^q \frac{C_v(q, d)}{k^{q/2}} \left[ \frac{4^q}{\lambda^q} \frac{C(q, d)}{n^{q/2}} G^q + O(n^{-4}) \right] + 2^q \frac{\tilde{C}^q}{n^q} + O(n^{-q-1}) \\ &= 8^q [C_v(q, d)]^2 \lambda^{-q} G^q N^{-\frac{q}{2}} + O(N^{-\frac{q}{2}} n^{\frac{q}{2}-4}) + 2^q \frac{\tilde{C}^q}{n^q} + O(n^{-q-1}). \end{aligned}$$

In summary, we have

$$\mathbb{E}[\|\theta^{(0)} - \theta_0\|^q] \leq O(N^{-\frac{q}{2}}) + \frac{2^q \tilde{C}^q k^q}{N^q} + O(k^{q+1} N^{-q-1}) = O(N^{-\frac{q}{2}}) + O(k^q N^{-q}). \quad (\text{B.14})$$

□

## C Proof of Theorem 3.4

The whole proof could be completed in two steps: first, show simple averaging estimator  $\theta^{(0)}$  is  $\sqrt{N}$ -consistent when  $k = O(\sqrt{N})$ ; then show the consistency and asymptotic normality of the one-step estimator  $\theta^{(1)}$ . In the first step, we need to show the following.

**Lemma C.1.** *Under Assumption 3.1, 3.2 and 3.3, when  $k = O(\sqrt{N})$ , the simple averaging estimator  $\theta^{(0)}$  is  $\sqrt{N}$ -consistent estimator of  $\theta_0$ , i.e.,*

$$\sqrt{N} \|\theta^{(0)} - \theta_0\| = O_P(1) \text{ as } N \rightarrow \infty.$$

*Proof.* If  $k$  is finite and does not grow with  $N$ , the proof is trivial. So, we just need to consider the case that  $k \rightarrow \infty$ . We know that  $\|\mathbb{E}[\sqrt{n}(\theta_i - \theta_0)]\| \leq O(\frac{1}{\sqrt{n}})$  by Lemma B.5 and  $\mathbb{E}[\|\sqrt{n}(\theta_i - \theta_0)\|^2] \leq 2\text{Tr}(\Sigma) + O(n^{-1})$  by Lemma B.1. By applying Lindeberg-Lévy Central Limit Theorem, we have

$$\begin{aligned}\sqrt{N}(\theta^{(0)} - \theta_0) &= \frac{1}{\sqrt{k}} \sum_{i=1}^k \sqrt{n}(\theta_i - \theta_0) = \frac{1}{\sqrt{k}} \sum_{i=1}^k \{\sqrt{n}(\theta_i - \theta_0) - \mathbb{E}[\sqrt{n}(\theta_i - \theta_0)]\} + \sqrt{nk}\mathbb{E}[\theta_1 - \theta_0] \\ &\xrightarrow{d} \mathbf{N}(0, \Sigma) + \lim_{N \rightarrow \infty} \sqrt{nk}\mathbb{E}[\theta_1 - \theta_0],\end{aligned}$$

It suffices to show  $\lim_{N \rightarrow \infty} \sqrt{nk}\mathbb{E}[\theta_1 - \theta_0]$  is finite. By Lemma B.5, we have

$$\|\mathbb{E}[\theta_i - \theta_0]\| = O(\frac{1}{n}), \forall i \in \{1, 2, \dots, k\},$$

which means that  $\|\sqrt{nk}\mathbb{E}[\theta_i - \theta_0]\| = O(1)$  if  $k = O(\sqrt{N}) = O(n)$ . Thus, when  $k = O(\sqrt{N})$ ,  $\sqrt{N}(\theta^{(0)} - \theta_0)$  is bounded in probability.  $\square$

Now, we can prove Theorem 3.4.

*Proof.* By the definition of the one-step estimator

$$\theta^{(1)} = \theta^{(0)} - \ddot{M}(\theta^{(0)})^{-1} \dot{M}(\theta^{(0)}),$$

and by Theorem 4.2 in Chapter XIII of Lang [1993], we have

$$\begin{aligned}\sqrt{N}\ddot{M}(\theta^{(0)})(\theta^{(1)} - \theta_0) &= \ddot{M}(\theta^{(0)})\sqrt{N}(\theta^{(0)} - \theta_0) - \sqrt{N}(\dot{M}(\theta^{(0)}) - \dot{M}(\theta_0)) - \sqrt{N}\dot{M}(\theta_0) \\ &= \ddot{M}(\theta^{(0)})\sqrt{N}(\theta^{(0)} - \theta_0) - \sqrt{N} \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho (\theta^{(0)} - \theta_0) - \sqrt{N}\dot{M}(\theta_0) \\ &= \left[ \ddot{M}(\theta^{(0)}) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho \right] \sqrt{N}(\theta^{(0)} - \theta_0) - \sqrt{N}\dot{M}(\theta_0),\end{aligned}$$

As it is shown in (B.14), for any  $\rho \in [0, 1]$ , when  $k = O(\sqrt{N})$ , we have

$$\|(1-\rho)\theta_0 + \rho\theta^{(0)} - \theta_0\| \leq \rho\|\theta^{(0)} - \theta_0\| \xrightarrow{P} 0.$$

Since  $\ddot{M}(\cdot)$  is a continuous function,  $\left\| \ddot{M}(\theta^{(0)}) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho \right\| \xrightarrow{P} 0$ . Thus,

$$\sqrt{N}\ddot{M}(\theta^{(0)})(\theta^{(1)} - \theta_0) = -\sqrt{N}\dot{M}(\theta_0) + o_P(1).$$

And,  $\ddot{M}(\theta^{(0)}) \xrightarrow{P} \ddot{M}_0(\theta_0)$  because of  $\theta^{(0)} \xrightarrow{P} \theta_0$  and Law of Large Number. Therefore, we can obtain

$$\sqrt{N}(\theta^{(1)} - \theta_0) \xrightarrow{d} \mathbf{N}(0, \Sigma) \text{ as } N \rightarrow \infty$$

by applying Slutsky's Lemma.  $\square$

## D Proof of Theorem 3.5

Let us recall the formula for one-step estimator,

$$\theta^{(1)} = \theta^{(0)} - \ddot{M}(\theta^{(0)})^{-1} \dot{M}(\theta^{(0)}).$$

Then by Theorem 4.2 in Chapter XIII of Lang [1993], we have

$$\begin{aligned}
& \ddot{M}_0(\theta_0)(\theta^{(1)} - \theta_0) = [\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0) + \ddot{M}(\theta^{(0)})(\theta^{(1)} - \theta_0) \\
& = [\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0) + \ddot{M}(\theta^{(0)})(\theta^{(0)} - \theta_0) - [\ddot{M}(\theta^{(0)}) - \ddot{M}(\theta_0)] - \dot{M}(\theta_0) \\
& = [\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0) + \left[ \ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho \right] (\theta^{(0)} - \theta_0) - \dot{M}(\theta_0).
\end{aligned}$$

Then we have

$$\begin{aligned}
\theta^{(1)} - \theta_0 & = -\ddot{M}_0(\theta_0)^{-1}\dot{M}(\theta_0) + \ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0) \\
& \quad + \ddot{M}_0(\theta_0)^{-1} \left[ \ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho \right] (\theta^{(0)} - \theta_0) \quad (\text{D.15})
\end{aligned}$$

We will show the last two terms are small enough. Similar to the proof of Lemma B.1, we define a “good” event:

$$E_4 = \{\|\theta^{(0)} - \theta_0\| \leq \delta\}.$$

The probability of above event is close to 1 when  $N$  is large.

$$\Pr(E_4^c) \leq \frac{\mathbb{E}[\|\theta^{(0)} - \theta_0\|^8]}{\delta^8} \leq O(N^{-4}) + O(k^8 N^{-8}).$$

**Lemma D.1.** *If event  $E_4$  holds, for  $1 \leq q \leq 4$ , we have*

$$\begin{aligned}
\mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)}) \right\|^q \right] & \leq O(N^{-\frac{q}{2}}) + O(k^q N^{-q}), \\
\mathbb{E} \left[ \left\| \ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho \right\|^q \right] & \leq O(N^{-\frac{q}{2}}) + O(k^q N^{-q}).
\end{aligned}$$

*Proof.* By Lemma A.3, we know

$$\mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0) - \ddot{M}(\theta_0) \right\|^q \right] \leq \frac{C(q, d)}{N^{q/2}} H^q.$$

Under event  $E_4$  and Assumption 3.3, by applying Jensen’s inequality, we have

$$\left\| \ddot{M}(\theta_0) - \ddot{M}(\theta^{(0)}) \right\|^q \leq \frac{1}{N} \sum_{i=1}^k \sum_{x \in S_i} \left\| \ddot{m}(x; \theta^{(0)}) - \ddot{m}(x; \theta_0) \right\|^q \leq \frac{1}{N} \sum_{i=1}^k \sum_{x \in S_i} L(x)^q \|\theta^{(0)} - \theta_0\|^q.$$

Thus, for  $1 \leq q \leq 4$ , we have

$$\begin{aligned}
\mathbb{E} \left[ \left\| \ddot{M}(\theta_0) - \ddot{M}(\theta^{(0)}) \right\|^q \right] & \stackrel{\text{Hölder's}}{\leq} \mathbb{E} \left[ \left( \frac{1}{N} \sum_{i=1}^k \sum_{x \in S_i} L(x)^q \right)^2 \right]^{\frac{1}{2}} \mathbb{E} \left[ \|\theta^{(0)} - \theta_0\|^{2q} \right]^{\frac{1}{2}} \\
& \stackrel{(\text{B.14})}{\leq} O(N^{-\frac{q}{2}}) + \frac{2^q \tilde{C}^q L^q k^q}{N^q} + O\left(\frac{k^{q+1}}{N^{q+1}}\right).
\end{aligned}$$

As a result, we have, for  $1 \leq q \leq 4$ ,

$$\begin{aligned}
\mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)}) \right\|^q \right] & \leq 2^q \mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0) - \ddot{M}(\theta_0) \right\|^q \right] + 2^q \mathbb{E} \left[ \left\| \ddot{M}(\theta_0) - \ddot{M}(\theta^{(0)}) \right\|^q \right] \\
& \leq O(N^{-\frac{q}{2}}) + \frac{4^q \tilde{C}^q L^q k^q}{N^q} + O\left(\frac{k^{q+1}}{N^{q+1}}\right).
\end{aligned}$$

In this proof, we let  $\theta' = (1 - \rho)\theta_0 + \rho\theta^{(0)}$  for the simplicity of notation. Note that  $\theta' - \theta_0 = \rho(\theta^{(0)} - \theta_0)$ , then by event  $E_4$ , Assumption 3.3 and inequality (B.14), we have

$$\begin{aligned} \mathbb{E} \left[ \left\| \ddot{M}(\theta_0) - \ddot{M}(\theta') \right\|^q \right] &\stackrel{\text{Jensen's}}{\leq} \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^k \sum_{x \in S_i} L(x)^q \|\theta' - \theta_0\|^q \right] \\ &\stackrel{\text{Hölder's}}{\leq} \mathbb{E} \left[ \left( \frac{1}{N} \sum_{i=1}^k \sum_{x \in S_i} L(x)^q \right)^2 \right]^{\frac{1}{2}} \rho^q \mathbb{E} \left[ \|\theta^{(0)} - \theta_0\|^{2q} \right]^{\frac{1}{2}} \\ &\leq O(N^{-\frac{q}{2}}) + \frac{2^q \tilde{C}^q L^q k^q}{N^q} + O\left(\frac{k^{q+1}}{N^{q+1}}\right). \end{aligned}$$

So, we have

$$\begin{aligned} \mathbb{E} \left[ \left\| \ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1 - \rho)\theta_0 + \rho\theta^{(0)}) d\rho \right\|^q \right] &\leq \mathbb{E} \left[ \int_0^1 \left\| \ddot{M}(\theta_0) - \ddot{M}((1 - \rho)\theta_0 + \rho\theta^{(0)}) \right\|^q d\rho \right] \\ &= \int_0^1 \mathbb{E} \left[ \left\| \ddot{M}(\theta_0) - \ddot{M}((1 - \rho)\theta_0 + \rho\theta^{(0)}) \right\|^q \right] d\rho \leq O(N^{-\frac{q}{2}}) + \frac{2^q \tilde{C}^q L^q k^q}{N^q} + O\left(\frac{k^{q+1}}{N^{q+1}}\right). \end{aligned}$$

□

Therefore, under event  $E_4$ , for  $1 \leq q \leq 4$ , we can bound  $\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)$  and  $\ddot{M}_0(\theta_0)^{-1}[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1 - \rho)\theta_0 + \rho\theta^{(0)}) d\rho](\theta^{(0)} - \theta_0)$  as follows:

$$\begin{aligned} \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)\|^q] &\leq \lambda^{-q} \mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)}) \right\|^q \right] D^q \\ &\leq O(N^{-\frac{q}{2}}) + O(k^q N^{-q}). \end{aligned}$$

and,

$$\begin{aligned} &\mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0)^{-1}[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1 - \rho)\theta_0 + \rho\theta^{(0)}) d\rho](\theta^{(0)} - \theta_0) \right\|^q \right] \\ &\leq \lambda^{-q} \mathbb{E} \left[ \left\| \ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1 - \rho)\theta_0 + \rho\theta^{(0)}) d\rho \right\|^q \right] D^q \\ &\leq O(N^{-\frac{q}{2}}) + O(k^q N^{-q}). \end{aligned}$$

And by Lemma A.2, for  $1 \leq q \leq 8$ , we have

$$\mathbb{E}[\|\dot{M}(\theta_0)\|^q] = O(N^{-\frac{q}{2}}).$$

Therefore, combining above three bounds and equation (D.15), we have, for  $1 \leq q \leq 4$ ,

$$\begin{aligned} \mathbb{E}[\|\theta^{(1)} - \theta_0\|^q] &\leq 3^q \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)\|^q] \\ &\quad + 3^q \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1} \left[ \ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1 - \rho)\theta_0 + \rho\theta^{(0)}) d\rho \right](\theta^{(0)} - \theta_0)\|^q] \\ &\quad + 3^q \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1} \dot{M}(\theta_0)\|^q] + Pr(E_4^c) D^q \\ &= O(N^{-\frac{q}{2}}) + O(k^q N^{-q}). \end{aligned}$$

Now, we can give tighter bounds for the first two terms in equation (D.15) by Hölder's inequality.

$$\begin{aligned} &\mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)\|^2] \\ &\leq \lambda^{-2} \sqrt{\mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)}) \right\|^4 \right]} \sqrt{\mathbb{E}[\|\theta^{(1)} - \theta_0\|^4]} \\ &= O(N^{-2}) + O(k^4 N^{-4}), \end{aligned}$$

and,

$$\begin{aligned}
& \mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho](\theta^{(0)} - \theta_0)\|^2] \\
& \leq \lambda^{-2} \sqrt{\mathbb{E} \left[ \left\| \ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho \right\|^4 \right]} \sqrt{\mathbb{E}[\|\theta^{(0)} - \theta_0\|^4]} \\
& = O(N^{-2}) + O(k^4 N^{-4}).
\end{aligned}$$

Now, we can finalize our proof by using equation (B.13) again,

$$\begin{aligned}
\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2] & \leq 2\mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}\dot{M}(\theta_0)\|^2] + 4\mathbb{E}[\|\ddot{M}_0(\theta_0)^{-1}[\ddot{M}_0(\theta_0) - \ddot{M}(\theta^{(0)})](\theta^{(1)} - \theta_0)\|^2] \\
& \quad + 4\mathbb{E} \left[ \left\| \ddot{M}_0(\theta_0)^{-1}[\ddot{M}(\theta_0) - \int_0^1 \ddot{M}((1-\rho)\theta_0 + \rho\theta^{(0)})d\rho](\theta^{(0)} - \theta_0) \right\|^2 \right] \\
& \leq 2\frac{\text{Tr}(\Sigma)}{N} + O(N^{-2}) + O(k^4 N^{-4}).
\end{aligned}$$

## E Proof of Corollary 3.7

At first, we will present a lemma on the negative moments of a Binomial random variable, i.e.,  $\mathbb{E} \left[ \frac{1}{Z} 1_{(Z>0)} \right]$  and  $\mathbb{E} \left[ \frac{1}{Z^2} 1_{(Z>0)} \right]$ , where  $Z \sim \mathbf{B}(k, p)$  and  $\mathbf{B}(k, p)$  denotes Binomial distribution with  $k$  independent trials and a success probability  $p$  for each trial. We believe that  $\mathbb{E} \left[ \frac{1}{Z} 1_{(Z>0)} \right]$  and  $\mathbb{E} \left[ \frac{1}{Z^2} 1_{(Z>0)} \right]$  should have been well studied. However, we did not find any appropriate reference on their upper bounds that we need. So, we derive the upper bounds as follows, which will be useful in the proof of Corollary 3.7.

**Lemma E.1.** Suppose  $Z \sim \mathbf{B}(k, p)$ , when  $z > 0$ , we have

$$\mathbb{E} \left[ \frac{1}{Z} 1_{(Z>0)} \right] < \frac{1}{kp} + \frac{3}{k^2 p^2} \quad \text{and} \quad \mathbb{E} \left[ \frac{1}{Z^2} 1_{(Z>0)} \right] < \frac{6}{k^2 p^2}.$$

*Proof.* By definition, we have

$$\begin{aligned}
\mathbb{E} \left[ \frac{1}{Z} 1_{(Z>0)} \right] & = \sum_{z=1}^k \frac{1}{z} \binom{k}{z} p^z (1-p)^{k-z} = \sum_{z=1}^k \frac{1}{z} \frac{k!}{z!(k-z)!} p^z (1-p)^{k-z} \\
& = \sum_{z=1}^k \frac{z+1}{z} \frac{1}{(k+1)p} \frac{(k+1)!}{(z+1)!(k-z)!} p^{z+1} (1-p)^{k-z} \\
& = \sum_{z=1}^k \frac{1}{(k+1)p} \binom{k+1}{z+1} p^{z+1} (1-p)^{k-z} + \sum_{z=1}^k \frac{1}{z} \frac{1}{(k+1)p} \binom{k+1}{z+1} p^{z+1} (1-p)^{k-z} \\
& < \frac{1}{(k+1)p} + \sum_{z=1}^k \frac{z+2}{z} \frac{1}{(k+1)(k+2)p^2} \binom{k+2}{z+2} p^{z+2} (1-p)^{k-z} \\
& < \frac{1}{(k+1)p} + \frac{3}{(k+1)(k+2)p^2} < \frac{1}{kp} + \frac{3}{k^2 p^2}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathbb{E} \left[ \frac{1}{Z^2} 1_{(Z>0)} \right] &= \sum_{z=1}^k \frac{1}{z^2} \binom{k}{z} p^z (1-p)^{k-z} = \sum_{z=1}^k \frac{1}{z^2} \frac{k!}{z!(k-z)!} p^z (1-p)^{k-z} \\
&= \sum_{z=1}^k \frac{(z+1)(z+2)}{z^2} \frac{1}{(k+1)(k+2)p^2} \frac{(k+2)!}{(z+2)!(k-z)!} p^{z+2} (1-p)^{k-z} \\
&\leq \sum_{z=1}^k \frac{6}{(k+1)(k+2)p^2} \binom{k+2}{z+2} p^{z+2} (1-p)^{k-z} \\
&< \frac{6}{(k+1)(k+2)p^2} < \frac{6}{k^2 p^2}.
\end{aligned}$$

□

Now, we can prove Corollary 3.7 could be as follows.

*Proof.* Let the random variable  $Z$  denote the number of machines that successfully communicate with the central machine, which means that  $Z$  follows Binomial distribution,  $B(k, 1-r)$ . By Law of Large Number,  $\frac{Z}{(1-r)k} \xrightarrow{P} 1$  as  $k \rightarrow \infty$ . If  $Z$  is known, the size of available data becomes  $Zn$ . By Theorem 3.4, the one-step estimator  $\theta^{(1)}$  is still asymptotic normal when  $k = O(\sqrt{N})$ ,

$$\sqrt{Zn}(\theta^{(1)} - \theta_0) \xrightarrow{d} N(0, \Sigma) \text{ as } n \rightarrow \infty.$$

Therefore, when  $k \rightarrow \infty$ , we have

$$\sqrt{(1-r)N}(\theta^{(1)} - \theta_0) = \sqrt{\frac{(1-r)N}{Zn}} \sqrt{Zn}(\theta^{(1)} - \theta_0) \xrightarrow{d} \sqrt{\frac{(1-r)N}{Zn}} N(0, \Sigma).$$

Since  $\frac{(1-r)N}{Zn} = \frac{(1-r)k}{Z} \xrightarrow{P} 1$ , by Slutsky's Lemma, we have

$$\sqrt{(1-r)N}(\theta^{(1)} - \theta_0) \xrightarrow{d} N(0, \Sigma).$$

This result indicates that when the local machines could lose communication independently with central machine with probability  $q$ , the one-step estimator  $\theta^{(1)}$  shares the same asymptotic properties with the oracle M-estimator using  $(1-r) \times 100\%$  of the total samples.

Next, we will analyze the mean squared error of one-step estimator with the presence of local machine failures. Note that, when  $Z$  is fixed and known, by Theorem 3.5, we have

$$\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2 | Z] \leq \frac{2\text{Tr}[\Sigma]}{nZ} + O(n^{-2}Z^{-2}) + O(n^{-4}).$$

By Rule of Double Expectation and Lemma E.1,

$$\begin{aligned}
&\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2 1_{(Z>0)}] = \mathbb{E}[\mathbb{E}[\|\theta^{(1)} - \theta_0\|^2 | Z] 1_{(Z>0)}] \\
&\leq \mathbb{E} \left[ \frac{2\text{Tr}[\Sigma]}{nZ} 1_{(Z>0)} \right] + \mathbb{E}[(O(n^{-2}Z^{-2}) + O(n^{-4})) 1_{(Z>0)}] \\
&\leq 2\text{Tr}[\Sigma] \left\{ \frac{1}{nk(1-r)} + \frac{3}{nk^2(1-r)^2} \right\} + O(n^{-2}k^{-2}(1-r)^{-2}) + O(n^{-4}) \\
&= \frac{2\text{Tr}[\Sigma]}{N(1-r)} + \frac{6\text{Tr}[\Sigma]}{Nk(1-r)^2} + O(N^{-2}(1-r)^{-2}) + O(k^2N^{-2}).
\end{aligned}$$

□