Journal of Machine Learning Research 25 (2024) 1-41 Submitted 6/22; Revised 9/23; Published 3/24

Distributed Estimation on Semi-Supervised Generalized Linear Model

Jiyuan Tu TUJY.19@GMAIL.COM
School of Mathematics
Shanghai University of Finance and Economics, Shanghai, 200433, China

Weidong Liu WEIDONGL@SJTU.EDU.CN
School of Mathematical Sciences
MoE Key Lab of Artificial Intelligence
Shanghai Jiao Tong University, Shanghai, 200240, China

Xiaojun Mao MAOXJ@SJTU.EDU.CN
School of Mathematical Sciences
Ministry of Education Key Laboratory of Scientific and Engineering Computing
Shanghai Jiao Tong University, Shanghai, 200240, China

Editor: Sanmi Koyejo

Abstract

Semi-supervised learning is devoted to using unlabeled data to improve the performance of machine learning algorithms. In this paper, we study the semi-supervised generalized linear model (GLM) in the distributed setup. In the cases of single or multiple machines containing unlabeled data, we propose two distributed semi-supervised algorithms based on the distributed approximate Newton method. When the labeled local sample size is small, our algorithms still give a consistent estimation, while fully supervised methods fail to converge. Moreover, we theoretically prove that the convergence rate is greatly improved when sufficient unlabeled data exists. Therefore, the proposed method requires much fewer rounds of communications to achieve the optimal rate than its fully-supervised counterpart. In the case of the linear model, we prove the rate lower bound after one round of communication, which shows that rate improvement is essential. Finally, several simulation analyses and real data studies are provided to demonstrate the effectiveness of our method.

Keywords: Distributed Learning, Semi-supervised Learning, Generalized Linear Model

1. Introduction

Distributed machine learning has attracted much attention in statistics in recent years. A variety of distributed algorithms have been proposed for statistical inference problems. Many interesting properties are investigated, and distributed machine learning is enlarged remarkably.

As one of the most fundamental ideas in distributed learning, the divide-and-conquer (DC) method often behaves in impressive performances in statistical models. For quantile

Weidong Liu and Xiaojun Mao are the co-corresponding authors.

regression (Volgushev et al., 2019; Chen et al., 2019), kernel ridge regression (Zhang et al., 2015; Lin et al., 2017) and kernel density estimation (Li et al., 2013), the optimal rates of DC depend on the number of machines and the scale of samples. For high dimensional linear regression and support vector machine, Lee et al. (2017); Lian and Fan (2018); Battey et al. (2018); Zhao et al. (2020) proved that the accuracy of DC can be remarkably improved by a well-known debiasing technique in statistics. For iterative algorithms like Distributed Approximate NEwton (DANE) method (Shamir et al. (2014)), Jordan et al. (2019); Jianqing Fan and Wang (2023) proved that with an appropriate initial estimator, DANE and its variants could achieve the optimal statistical rate within a constant round of communications. There are also many other distributed learning algorithms like alternating direction method of multipliers (Boyd et al., 2011), subsampled average mixture algorithm (Zhang et al., 2013), local stochastic gradient descent (Stich, 2019; Yu et al., 2019), distributed dual coordinate ascent method (Jaggi et al., 2014; Smith et al., 2017), lazily aggregated gradient method (Chen et al., 2018), and so on.

In this article, we focus on distributed semi-supervised estimation for the generalized linear model (GLM). There are various practical problems, especially in the era of big data, involving the analysis of massive unlabeled datasets (e.g., electronic medical records data, textual data). Semi-supervised learning (SSL) has demonstrated considerable success in machine learning problems in terms of prediction precision. In non-distributed settings, a growing body of literature demonstrates SSL can be of great use in classification (see, *e.g.* , Ando and Zhang (2005, 2007); Blum and Mitchell (1998); Vapnik (1999); Wang and Shen (2007); Wang et al. (2007, 2009); Zhu (2005); Zhu and Goldberg (2009)). The semi-supervised estimation problems, with continuous labels, have also been studied by Wasserman and Lafferty (2007); Johnson and Zhang (2008); Chakraborty and Cai (2018); Zhang et al. (2019); Cai and Guo (2020); Zhang and Bradic (2021); Hou et al. (2023); Azriel et al. (2022). In recent years, distributed SSL has emerged as an exciting new area of research in machine learning. A novel study of kernel ridge regression by Chang et al. (2017) has shown that the semi-supervised DC method will benefit significantly from the unlabeled data and allows a more flexible number of machines than the ordinary DC. Work done along this line include distributed SSL for the kernel-based gradient descent algorithms (Lin and Zhou, 2018), bias-corrected kernel ridge regression (Guo et al., 2017), learning with multi-penalty regularization (Guo et al., 2019) and flexible Gaussian kernels (Hu and Zhou, 2021). A common feature of these prior works in distributed SSL is that they all focused on one-shot algorithms. It is well known that iterative algorithms can lead to more accurate estimates than one-shot algorithms. In this paper, we study the distributed SSL for generalized linear models using the approximate Newton iteration method.

Within this paper, we concentrate specifically on the scenario where data is distributed across several computing units, some of which possess supplementary unlabeled data. This setup aligns with numerous practical situations; for instance, in the medical field, multiple hospitals collaborate to enhance the efficacy of the model through sharing medical data, which typically encompass ample unlabeled covariates (Hou et al., 2023; Liu et al., 2022; Cai et al., 2022). Throughout this paper, we direct our focus towards the generalized linear model (GLM), a widely utilized statistical model that has been implemented across diverse fields such as healthcare (Parikh et al., 2019; Guo et al., 2022), genetic analysis (Ma et al., 2022), and economics (Theodossiou, 1998). One of the reasons we choose to consider GLM is not only due to its broad range of applications but also due to the unique form of its loss functions, which will be delved into in more detail in Section 2.1.

In the following, we introduce the generalized linear model. Given a function $\psi : \mathbb{R} \to \mathbb{R}$, the observations $(\boldsymbol{X}, Y) \in \mathbb{R}^{p+1}$ are generated according to the following conditional probability function

$$\mathbb{P}(Y \mid \boldsymbol{X}) = \widetilde{c} \exp\left\{\frac{Y\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^* - \psi(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)}{c(\sigma)}\right\}, \tag{1}$$

where $\widetilde{c}$ and $c(\sigma)$ are scale constants, and $\boldsymbol{\beta}^*$ is the true model parameter. Here ψ', the derivative of ψ, is called the canonical link function of the generalized linear model. To estimate $\boldsymbol{\beta}^*$, the most straightforward estimator is the solution of the (empirical) negative log-likelihood function

$$f(\boldsymbol{X}, Y, \boldsymbol{\beta}) = -Y\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \psi(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}). \tag{2}$$

In SSL, a large unlabeled dataset only contains information on predictors $\boldsymbol{X}$. It is thus of great interest to develop a distributed SSL algorithm using both labeled datasets and unlabeled datasets. However, as we will prove, naive incorporation of the unlabeled dataset in DANE with the loss function (2) will lead to degradation of the estimation precision. The primary reason is that the unlabeled predictor does not contain the information of $\boldsymbol{\beta}^*$ in lack of the corresponding response and may result in higher uncertainty. On the other hand, in GLM, the predictors are directly related to the Hessian matrix of the loss (2). In other words, the unlabeled dataset is particularly useful in estimating the Hessian matrix and reducing the number of iterations of iterative algorithms. In this paper, we propose a novel loss function (called the semi-supervised surrogate loss function) based on a variance correction technique to avoid higher uncertainty. The proposed surrogate loss has a more concentrated Hessian matrix, and the resulting estimator is as efficient as the supervised estimator. Further, we combine it with the Newton iteration method and propose a Semi-Supervised Distributed Approximate NEwton estimator (SSDANE) for GLM.

Compared with supervised DANE, our SSDANE has the following advantages. First, SSDANE converges faster to the optimal neighborhood of $\boldsymbol{\beta}^*$. Second, SSDANE requires fewer rounds of communication between the master machine and local machines. This is particularly important when communication cost becomes a bottleneck in distributed computing. Third, SSDANE allows higher dimensions for $\boldsymbol{\beta}^*$. The supervised DANE typically requires a dimension smaller than the labeled local sample sizes. In contrast, SSDANE can still have a nearly optimal rate when the dimension exceeds the labeled local sample sizes. Finally, extensive numerical experiments are performed to demonstrate the better performance of SSDANE. Our experimental results suggest that the performance of the semi-supervised algorithm is consistently superior to that of the supervised algorithm, its performance improves as the amount of unlabeled data increases. Based on our findings, we recommend the use of our SSDANE over DANE, particularly when a larger amount of unlabeled data is available.

The rest of the paper is organized as follows. Section 2 introduces the semi-supervised surrogate loss function and SSDANE. In Section 3, we present theoretical results for the proposed methods. Numerical experiments are provided in Section 4. Finally, we conclude our work in Section 5. The proofs of theoretical results are relegated to Appendix. Given a vector $\boldsymbol{v} = (v_1, ..., v_p)^T$, we denote $|\boldsymbol{v}|_1 = \sum_{l=1}^p |v_l|$, $|\boldsymbol{v}|_2 = \sqrt{\sum_{l=1}^p v_l^2}$ and $|\boldsymbol{v}|_\infty = \sup_{1\le l\le p} |v_l|$.

For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, define $\|\boldsymbol{A}\|$ as the spectral norm, $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ as the largest and smallest eigenvalues of $\boldsymbol{A}$ respectively.

2. Methodology

In this section, we first introduce the semi-supervised surrogate loss in a single-machine setting. Then we apply it in a distributed setup and develop two distributed semi-supervised algorithms.

2.1 Semi-Supervised Surrogate Loss

To motivate the construction of the semi-supervised surrogate loss, we shall consider the single-machine setting. Denote $\mathcal{D}$ as the index set of the labeled pairs $\{(\boldsymbol{X}_i, Y_i)\}$, $\mathcal{D}^*$ as the index set of unlabeled covariates $\{X_i\}$, and $\mathcal{H} = \mathcal{D} \cup \mathcal{D}^*$. We let $|\mathcal{D}| = n$ and $|\mathcal{H}| = n^*$. For the generalized linear model (1) with link function $\psi'(\cdot)$, recall the empirical loss function based on the labelled data $\mathcal{D}$ is

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i\in\mathcal{D}} Y_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \frac{1}{n}\sum_{i\in\mathcal{D}} \psi(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}). \tag{3}$$

With this target function, one may estimate the true parameters by the optimization

$$\widehat{\boldsymbol{\beta}}_{\mathcal{D}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\mathcal{L}(\boldsymbol{\beta})\}. \tag{4}$$

It is well known that the computation time (the number of iterations) of a first/second order method on solving this optimization depends on the condition number of the Hessian matrix

$$\widehat{\boldsymbol{H}}_{\mathcal{D}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i\in\mathcal{D}} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} \psi''(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

When the sample size $|\mathcal{D}|$ is small, or the dimension p is large, the condition number can be large. For example, when $p \sim n$, the condition number of the sample covariance matrix can go to infinity (see Bickel and Levina (2008)). In order to get a more stable and concentrated Hessian matrix, a direct way is to make use of the unlabelled data by noting that the Hessian matrix only depends on the covariates. That is, one may consider the following loss

$$\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i\in\mathcal{D}} Y_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \frac{1}{n^*}\sum_{i\in\mathcal{H}} \psi(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}), \tag{5}$$

where the unlabeled data is used and leads to a more stable Hessian matrix once n^* is much larger than n:

$$\widehat{\boldsymbol{H}}_{\mathcal{H}}^{\mathrm{ss}}(\boldsymbol{\beta}) = \frac{1}{n^*}\sum_{i\in\mathcal{H}} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} \psi''(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

However, due to the unbalance between the sample sizes of $\mathcal{D}$ and $\mathcal{H}$, the asymptotic variance of the estimator $\widehat{\beta}_{\mathcal{H}} = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta})\}$ increases. This results in a great loss in statistical efficiency on estimating $\boldsymbol{\beta}^*$. In fact, the covariance of the gradient at $\boldsymbol{\beta}^*$ satisfies

$$
\begin{aligned}
&\mathrm{Cov}\Big(-\frac{1}{n}\sum_{i\in\mathcal{D}} Y_i\boldsymbol{X}_i+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\Big)\\
=&\frac{1}{n}\mathrm{Cov}\Big(-Y\boldsymbol{X}+\boldsymbol{X}\psi^{'}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\Big)+\mathrm{Cov}\Big(\frac{1}{n^*}\sum_{i\in\mathcal{H}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)-\frac{1}{n}\sum_{i\in\mathcal{D}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\Big)\\
=&\frac{1}{n}\mathrm{Cov}\Big(-Y\boldsymbol{X}+\boldsymbol{X}\psi^{'}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\Big)+\frac{n^*-n}{nn^*}\mathrm{Cov}\Big(\boldsymbol{X}\psi^{'}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\Big),
\end{aligned}
\tag{6}
$$

where the first matrix on the right-hand side of the above equation is the covariance matrix of the gradient of $\mathcal{L}(\boldsymbol{\beta})$ in (3). The unbalance between $\mathcal{D}$ and $\mathcal{H}$ leads to the second extra covariance matrix, which has the same order as the first one.

In order to eliminate the second term of (6) while maintaining the stability of the Hessian matrix, we consider the following two-step method. Suppose that we have an initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ for $\boldsymbol{\beta}^*$. Then we can subtract

$$
\nabla\mathcal{L}^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)})-\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}^{(0)})=\frac{1}{n^*}\sum_{i\in\mathcal{H}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(0)})-\frac{1}{n}\sum_{i\in\mathcal{D}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(0)})
$$

to reduce the extra term in the gradient of $\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta})$. That is, we consider the following corrected estimating equation

$$
\nabla\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta})-\Big(\nabla\mathcal{L}^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)})-\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}^{(0)})\Big)=0. \tag{7}
$$

It is easy to see that if the initial value $\widehat{\boldsymbol{\beta}}^{(0)}$ is close to $\boldsymbol{\beta}^*$, then the covariance of the above equation (at value $\boldsymbol{\beta}^*$) is asymptotically equivalent to that of $\mathcal{L}(\boldsymbol{\beta})$. Therefore, the estimator satisfies equation (7) will have almost the same statistical efficiency as $\mathcal{L}(\boldsymbol{\beta})$. On the other hand, the Hessian matrix of (7) is $\widehat{\boldsymbol{H}}^{\mathrm{ss}}_{\mathcal{H}}(\beta)$, which is more stable than $\widehat{\boldsymbol{H}}_{\mathcal{D}}(\beta)$.

The estimating equation in (7) corresponds to the following loss function

$$
\widetilde{\mathcal{L}}(\boldsymbol{\beta})=\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta})-\Big\langle\nabla\mathcal{L}^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)})-\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}^{(0)}),\boldsymbol{\beta}\Big\rangle. \tag{8}
$$

We call $\widetilde{\mathcal{L}}(\boldsymbol{\beta})$ as the semi-supervised surrogate loss. This surrogate loss can be easily extended to the distributed setting. We will show rigorously that the number of iterations and hence the communication cost are much smaller than the algorithm with only labeled data.

2.2 Distributed Semi-Supervised Learning

In this section, we consider semi-supervised learning in the distributed setup. More specifically, assume we have $N=mn$ i.i.d. pairs of observations $\{(\boldsymbol{X}_i,Y_i)\}$ from the generalized linear model (1) evenly stored in m different machines $\{\mathcal{H}_1,\ldots,\mathcal{H}_m\}$. Let $\mathcal{H}_1$ be the master machine that is in charge of data update and transmission. We discuss the cases where unlabeled data are stored in a single machine and multiple machines separately. In both cases, we develop algorithms that achieve a faster convergence rate than their fully-supervised counterparts.

2.2.1 Unlabeled Data on Master Machine

We first assume the sample size of additional $n^* - n$ unlabeled covariates $\boldsymbol{X}_i$ is not too large so that they can be stored in a single machine (i.e. $\mathcal{H}_1$, the master machine). We denote $\mathcal{D}_j$ as the indices of labelled observations $\{(\boldsymbol{X}_i, Y_i)\}$, and $\mathcal{D}_1^*$ as the unlabeled covariates $\{\boldsymbol{X}_i\}$ on $\mathcal{H}_1$. Then we have that $|\mathcal{D}_j| = n$, $|\mathcal{D}_1^*| = n^* - n$ and $|\mathcal{H}_1| = n^*$.

To conduct distributed learning with the assistance of the unlabeled data, we use the semi-supervised surrogate loss in (8) and replace the local gradient $\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}^{(0)})$ by the global gradient $\frac{1}{m}\sum_{j=1}^m \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(0)})$, where

$$\mathcal{L}_j(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i\in\mathcal{D}_j} Y_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \frac{1}{n}\sum_{i\in\mathcal{D}_j}\psi(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}) \tag{9}$$

is the local loss in the j-th machine. After the master machine receives all the gradients from workers, it solves the following optimization and updates the initial value $\widehat{\boldsymbol{\beta}}^{(0)}$ to $\widehat{\boldsymbol{\beta}}^{(1)}$:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{(1)} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\{\widetilde{\mathcal{L}}^{(1)}(\boldsymbol{\beta})\}\\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{\mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}) - \left\langle \nabla\mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)}) - \frac{1}{m}\sum_{j=1}^m \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(0)}), \boldsymbol{\beta}\right\rangle\right\}.\end{aligned} \tag{10}$$

We call this the one-step Semi-Supervised Distributed Approximate NEwton (SSDANE) estimator, with respect to the DANE proposed by Shamir et al. (2014).[1] The one-step SSDANE can be directly extended to multi-round SSDANE which is stated in Algorithm 1.

2.2.2 Unlabeled Data on Multiple Machines

In this section, we consider the case that the sample size of unlabeled data is large so that they are separately stored in multiple machines. To be more specific, we denote $\mathcal{U}$ as the set of machines having unlabeled dataset and let $\mathcal{D}_j$ and $\mathcal{D}_j^*$ be the index sets of labelled observations $\{(\boldsymbol{X}_i, Y_i)\}$ and unlabeled covariates $\{\boldsymbol{X}_i\}$ on $\mathcal{H}_j$ respectively. Then for $j\in\mathcal{U}$, there is $|\mathcal{D}_j| = n$, $|\mathcal{D}_j^*| = n^* - n$ and $|\mathcal{H}_j| = n^*$. Suppose we have an initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$, for $j\in\mathcal{U}$, we apply SSDANE on the j-th machine to obtain

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_j^{(1)} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{\widetilde{\mathcal{L}}_j^{(1)}(\boldsymbol{\beta})\right\}\\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{\mathcal{L}_j^{\mathrm{ss}}(\boldsymbol{\beta}) - \left\langle \nabla\mathcal{L}_j^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)}) - \frac{1}{m}\sum_{k=1}^m \nabla\mathcal{L}_k(\widehat{\boldsymbol{\beta}}^{(0)}), \boldsymbol{\beta}\right\rangle\right\},\end{aligned}$$

where $\mathcal{L}_j^{\mathrm{ss}}(\boldsymbol{\beta})$ denotes the local semi-supervised empirical loss on the j-th machine, namely,

$$\mathcal{L}_j^{\mathrm{ss}}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i\in\mathcal{D}_j} Y_i\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \frac{1}{n^*}\sum_{i\in\mathcal{H}_j}\psi(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

[1] The original DANE is based on the average of local solutions for all workers. For the sake of comparison, we refer the solution in a local machine to DANE and the average of them to DANE-Avg

Algorithm 1 Semi-Supervised Distributed Approximate NEwton Method (SSDANE)

```
Input: Labeled data $\{(\boldsymbol{X}_i, Y_i) \mid i \in \mathcal{D}_j\}$ on machine $\mathcal{H}_j$ for $j = 1, ..., m$, and unlabeled data $\{\boldsymbol{X}_i \mid i \in \mathcal{D}_1^*\}$ on the master machine $\mathcal{H}_1$, the number of iterations $T$.
The master machine $\mathcal{H}_1$ obtains the initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ by minimizing the local empirical loss function on $\mathcal{H}_1$.
for $t = 1, \ldots, T$ do
    The master machine broadcasts the parameter $\widehat{\boldsymbol{\beta}}^{(t-1)}$ to each worker machine.
    for $j = 1, \ldots, m$ do
        The $j$-th machine computes the local gradient
            $\nabla \mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(t-1)}) = -\frac{1}{n}\sum_{i \in \mathcal{D}_j} Y_i \boldsymbol{X}_i^{\mathrm{T}} + \frac{1}{n}\sum_{i \in \mathcal{D}_j} \psi'(\boldsymbol{X}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^{(t-1)}) \boldsymbol{X}_i,$
        and sends back to the master machine $\mathcal{H}_1$.
    end for
    The master machine updates the parameter by solving
        $\widehat{\boldsymbol{\beta}}^{(t)} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \{\widetilde{\mathcal{L}}^{(t)}(\boldsymbol{\beta})\}$
        $= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}) - \left\langle \nabla \mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(t-1)}) - \frac{1}{m}\sum_{j=1}^{m} \nabla \mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(t-1)}), \boldsymbol{\beta} \right\rangle \right\}.$    (11)
end for
Output: The final estimator $\widehat{\boldsymbol{\beta}}^{(T)}$.
```

Then we take the average over these local estimators to obtain a more accurate one,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{Avg}}^{(1)} = \frac{1}{|\mathcal{U}|}\sum_{j \in \mathcal{U}} \widehat{\boldsymbol{\beta}}_j^{(1)}. \tag{12}$$

The multi-round realization of the averaged SSDANE (SSDANE-Avg) method is provided in Algorithm 2.

3. Theoretical Results

In this section, we investigate the theoretical properties of the proposed algorithms. First, we need the following regular assumptions.

Assumption 1 *Let $\boldsymbol{H} = \mathbb{E}\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}$, then there is a constant $\rho > 0$ such that*

$$\rho \le \lambda_{\min}(\boldsymbol{H}) \le \lambda_{\max}(\boldsymbol{H}) \le \rho^{-1}.$$

Assumption 2 *There exist positive numbers $\eta_0 > 0$, C_0 such that*

$$\max\left\{ \sup_{\boldsymbol{v} \in \mathbb{S}^{p-1}} \mathbb{E}\Big[\exp\Big\{\eta_0 |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}|^2\Big\}\Big], \mathbb{E}\Big[\exp\{\eta_0 |\psi'(\boldsymbol{X}\boldsymbol{\beta}^*)|^2\}\Big], \mathbb{E}\Big[\exp\big(\eta_0 Y^2\big)\Big] \right\} \le C_0.$$

Algorithm 2 Semi-Supervised Distributed Approximate NEwton with Average (SSDANE-Avg)

```
Input: Labeled data $\{(\boldsymbol{X}_i, Y_i) \mid i \in \mathcal{D}_j\}$ on machine $\mathcal{H}_j$ for $j = 1, ..., m$; unlabeled data $\{\boldsymbol{X}_i \mid i \in \mathcal{D}_j^*\}$ on the each machine $\mathcal{H}_j$, where $j \in \mathcal{U}$; the number of iterations $T$.
The master machine $\mathcal{H}_1$ obtains the initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ by minimizing the local empirical loss function on $\mathcal{H}_1$.
for $t = 1, \ldots, T$ do
    The master machine broadcasts the parameter $\widehat{\boldsymbol{\beta}}^{(t-1)}$ to each worker machine.
    for $j = 1, ..., m$ do
        The $j$-th machine computes the local gradient
            $$\nabla \mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(t-1)}) = -\frac{1}{n} \sum_{i \in \mathcal{D}_j} Y_i \boldsymbol{X}_i^{\mathrm{T}} + \frac{1}{n} \sum_{i \in \mathcal{D}_j} \psi'(\boldsymbol{X}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^{(t-1)}) \boldsymbol{X}_i,$$
        and sends back to the master machine $\mathcal{H}_1$.
    end for
    The master machine computes $m^{-1} \sum_{j=1}^m \nabla \mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(t-1)})$ and broadcasts to worker machine with unlabeled data, that is, $\mathcal{H}_j$ for $j \in \mathcal{U}$.
    for $j \in \mathcal{U}$ do
        The $j$-th machine updates the parameter by solving
            $$\begin{aligned} \widehat{\boldsymbol{\beta}}_j^{(t)} &= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \widetilde{\mathcal{L}}_j^{(t)}(\boldsymbol{\beta}) \right\} \\ &= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \mathcal{L}_j^{\mathrm{ss}}(\boldsymbol{\beta}) - \left\langle \nabla \mathcal{L}_j^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(t-1)}) - \frac{1}{m} \sum_{k=1}^m \nabla \mathcal{L}_k(\widehat{\boldsymbol{\beta}}^{(t-1)}), \boldsymbol{\beta} \right\rangle \right\}, \end{aligned} \qquad (13)$$
        and sends back to the master machine $\mathcal{H}_1$.
    end for
    The master machine updates the parameter by $\widehat{\boldsymbol{\beta}}_{\mathrm{Avg}}^{(t)} = |\mathcal{U}|^{-1} \sum_{j \in \mathcal{U}} \widehat{\boldsymbol{\beta}}_j^{(t)}$.
end for
Output: The final estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{Avg}}^{(T)}$.
```

Assumption 3 *The canonical link function $\psi'(\cdot)$ is twice differentiable, and there exists a uniform constant c_ψ such that*

$$\sup_{x \in \mathbb{R}} \max\{|\psi''(x)|, |\psi'''(x)|\} \leq c_\psi.$$

The above assumptions are standard conditions for proving estimation consistency. Assumption 1 assumes the well-posedness of the Hessian matrix. Assumption 2 guarantees that the covariate $\boldsymbol{X}$ and the label Y admit the sub-Gaussian distribution. In Assumption 3 we assume the canonical link function $\psi'(x)$ has uniformly bounded first-order and second-order derivatives.

3.1 Convergence Rate of Semi-Supervised Distributed Estimation

3.1.1 Unlabeled Data on Master Machine

We first give the convergence rate of the estimator in Algorithm 1 when all unlabeled data are stored in the master machine.

Theorem 1 *Suppose Assumption 1 to 3 hold, assume the initial estimator* $\widehat{\boldsymbol{\beta}}^{(0)}$ *satisfies* $|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*|_2 = O_{\mathbb{P}}(r_n)$. *Moreover, the dimension of the parameter* p *satisfies* $p = o(\min\{(n^*)^{2/3}/\log n^*, mn/\log n^*\})$ *and* $r_n = o(p^{-1/2})$. *The* T*-th round* SSDANE *estimator satisfies*

$$\left|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*\right|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}} + r_n\Big(\frac{p\log n^*}{n^*}\Big)^{T/2} + \frac{1}{\sqrt{p}}(\sqrt{p}r_n)^{2^T}\right). \tag{14}$$

This theorem provides an upper bound for the SSDANE method. The first term is a nearly optimal statistical convergence rate with all labeled data. The second and third terms represent the improved convergence rate by each iteration in the algorithm. For example, the rate in the second term was improved by the factor $\sqrt{p(\log n^*)/n^*}$ with one iteration. To achieve the nearly optimal statistical rate $\sqrt{\frac{p\log n^*}{mn}}$, it suffices to take

$$T \geq \frac{\log n + \log m - \log p - \log\log n^*}{\log n^* - 2\log p - \log\log n^*}. \tag{15}$$

Note that n^* is the number of total samples in the master. It indicates that the information of unlabeled data indeed helps improve the rate of the algorithm. In fact, if the loss $\mathcal{L}_1(\boldsymbol{\beta})$ with only labeled data is used in DANE, i.e. replacing $\mathcal{L}_1^{ss}(\boldsymbol{\beta})$ by $\mathcal{L}_1(\boldsymbol{\beta})$ in (11),

$$\widehat{\boldsymbol{\beta}}_o^{(t)} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\mathcal{L}_1(\boldsymbol{\beta}) - \left\langle \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_o^{(t-1)}) - \frac{1}{m}\sum_{j=1}^m \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_o^{(t-1)}), \boldsymbol{\beta}\right\rangle\right\}, \tag{16}$$

then the second term of (14) becomes $r_n\Big(\frac{p\log n}{n}\Big)^{T/2}$, which is much slower than the one with unlabeled data as long as $n^* \gg n$. A slower rate means more iterations and more rounds of communication between the master machine and workers.

The improvement of the rate by the unlabeled data is essential. Consider the one-round estimator with $T = 1$ and for the linear model. Proposition 7 below shows a tight lower bound for DANE, $\widehat{\boldsymbol{\beta}}_o^{(1)}$ is $C\Big(\sqrt{\frac{p}{mn}} + r_n\sqrt{\frac{p}{n}}\Big)$. Except for the theoretical evidence, we will conduct extensive numerical analysis to verify such improvement.

To further see the advantage of our semi-supervised surrogate loss in distributed estimation, we consider the case that $p > n$ but p is much smaller than n^* and mn. Since $p > n$, then the Hessian matrix of $\mathcal{L}_1(\boldsymbol{\beta})$ is not positive definite. Hence $\widehat{\boldsymbol{\beta}}_o^{(t)}$ performs poorly on estimation of $\boldsymbol{\beta}^*$. For example, for the linear model, (16) is reduced to solve a linear equation $\widehat{\Sigma}_1\boldsymbol{\beta} = \boldsymbol{b}$, where $\widehat{\Sigma}_1 = \frac{1}{n}\sum_{i\in\mathcal{D}_1}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}$ is not full rank so that the solution is not unique. On the other hand, the Hessian matrix of our semi-supervised surrogate loss is still positive definite as $n^* \gg p$. In this case, $\widehat{\boldsymbol{\beta}}^{(T)}$ is well defined and can still have the nearly optimal rate.

Next, we present the asymptotic normality result for our proposed SSDANE estimator.

Theorem 2 *(Asymptotic normality of SSDANE) Under the same conditions as in Theorem 1, and additionally, we assume the dimension of the parameter* p *satisfies* $p = o(\min\{(n^*)^{1/2}/\log n^*, (mn/\log^2 n^*)^{1/3}\})$. *Then when the iteration round* T *satisfies* (15), *for any vector* $\boldsymbol{v} \in \mathbb{S}^{p-1}$, *we have that*

$$\frac{\sqrt{mn}}{\sigma(\boldsymbol{v})}\boldsymbol{v}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\begin{aligned} \big\{\sigma(\boldsymbol{v})\big\}^2 &= \boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}^{-1}\boldsymbol{\mathcal{C}}\boldsymbol{H}^{-1}\boldsymbol{v}, \\ \text{here} \quad \boldsymbol{H} &= \mathbb{E}\big[\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big], \\ \boldsymbol{\mathcal{C}} &= c(\sigma)\mathbb{E}\left[\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right]. \end{aligned} \tag{17}$$

The result shows that the proposed semi-supervised surrogate loss keeps the same statistical efficiency compared to $\mathcal{L}(\boldsymbol{\beta})$ with all labeled data.

Remark 3 *Regarding the statistical efficiency, we also note that the unlabeled data sometimes helps reduce the asymptotic variance in some works of semi-supervised learning. However, our methods have different settings and targets from these works. In particular, most results enjoy variance reduction due to additional information. For example, in Hou et al. (2023), the author proposed the imputation method with the assistance of additional surrogate information. In Cai and Guo (2020), the author considered estimating the explained variance with the unlabeled data. Chakrabortty and Cai (2018) and Azriel et al. (2022) estimated the model parameter for the miss-specified linear model. In contrast, we focus on estimating the model parameter in the distributed setting. It would also be interesting to study the aforementioned problem in the distributed setting by extending our method. And there should be no technical difficulties in showing a similar efficiency-enhancement effect.*

3.1.2 Unlabeled Data on Multiple Machines

In this section, we provide similar theoretical results when the unlabeled data are separately stored in multiple machines.

Theorem 4 *Suppose Assumption 1 to 3 hold, assume the initial estimator* $\widehat{\boldsymbol{\beta}}^{(0)}$ *satisfies* $|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*|_2 = O_{\mathbb{P}}(r_n)$. *Moreover, there are rate constraints* $p = o(\min\{(n^*)^{2/3}/\log n^*, mn/\log n^*\})$ *and* $r_n = o(p^{-1/2})$. *Then the* T*-th round* SSDANE-Avg *estimator satisfies*

$$\left|\widehat{\boldsymbol{\beta}}^{(T)}_{\mathrm{Avg}} - \boldsymbol{\beta}^*\right|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}} + r_n\Big(\frac{p\log n^*}{n^*}\Big)^{T} + r_n\Big(\frac{p\log n^*}{|\mathcal{U}|n^*}\Big)^{T/2} + \frac{1}{\sqrt{p}}(\sqrt{p}r_n)^{2^T}\right), \tag{18}$$

where $|\mathcal{U}|$ *denotes the cardinality of the set* $\mathcal{U}$.

We now compare the rates between (18) and (14). It is obvious that the more machines with unlabeled data, the faster the convergence rate of the algorithm. Note that the total number of unlabeled data is $|\mathcal{U}|(n^*-n)$. If the $|\mathcal{U}|(n^*-n)$ data are all stored in the master machine, then according to Theorem 1, the rate of SSDANE $\widehat{\boldsymbol{\beta}}^{(T)}$ is

$$O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}}+r_n\Big(\frac{p\log n^*}{|\mathcal{U}|(n^*-n)+n}\Big)^{T/2}+\frac{1}{\sqrt{p}}(\sqrt{p}r_n)^{2^T}\right).$$

This rate is slower than or the same as (18) when $|\mathcal{U}|\leq n^*/(p\log n^*)$, relying on magnitude of n^*-n. Therefore, when the unlabeled data are stored in multiple workers, the average SSDANE can have a better convergence rate. Furthermore, it has a faster rate than SSDANE by an extra factor $1/|\mathcal{U}|$ when n^* is close to n (c.f. $n^*=n$). On the other hand, the average of SSDANE requires $|\mathcal{U}|$ workers to solve the optimization (13) simultaneously, and hence needs a more homogeneous distributed system.

For the asymptotic distribution of $\widehat{\boldsymbol{\beta}}^{(T)}_{\mathrm{Avg}}$, we can establish a similar result with the same asymptotic variance as Theorem 2. Therefore, the average of SSDANE further accelerates the algorithm while keeping the statistical efficiency. We further note that, suppose the loss $\mathcal{L}^{\mathrm{ss}}_j(\boldsymbol{\beta})$ with unlabeled data were used trivially in DANE, i.e. let

$$\widehat{\boldsymbol{\beta}}^{(t)}_{\mathrm{ss},j}=\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\left\{\mathcal{L}^{\mathrm{ss}}_1(\boldsymbol{\beta})-\Big\langle\nabla\mathcal{L}^{\mathrm{ss}}_1(\widehat{\boldsymbol{\beta}}^{(t-1)}_{\mathrm{ss}})-\frac{1}{m}\sum_{j=1}^{m}\nabla\mathcal{L}^{\mathrm{ss}}_j(\widehat{\boldsymbol{\beta}}^{(t-1)}_{\mathrm{ss}}),\boldsymbol{\beta}\Big\rangle\right\},$$

and $\widehat{\boldsymbol{\beta}}^{(t)}_{\mathrm{ss,Avg}}=\frac{1}{|\mathcal{U}|}\sum_{j\in\mathcal{U}}\widehat{\boldsymbol{\beta}}^{(t)}_{\mathrm{ss},j}$. We can prove that the asymptotic variance of $\widehat{\boldsymbol{\beta}}^{(t)}_{\mathrm{ss,Avg}}$ is

$$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}^{-1}\Big(\boldsymbol{\mathcal{C}}+\frac{(n^*-n)|\mathcal{U}|}{mn^*}\boldsymbol{\mathcal{C}}^{\mathrm{ss}}\Big)\boldsymbol{H}^{-1}\boldsymbol{v},$$

where $\boldsymbol{\mathcal{C}}^{\mathrm{ss}}=\mathrm{C}ov\big\{\psi'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\big\}$. That is, a trivial use of unlabeled data in DANE can lead to a larger variance when $|\mathcal{U}|$ is proportional to m.

Remark 5 *We note that our current theoretical framework is founded on the premise that the data has comparable sample sizes across local sources, primarily for the sake of clarity in presentation. Specifically, for the* SSDANE *method, we conduct minimization of the surrogate loss (defined in* (10)*) solely on the master machine $\mathcal{H}_1$. In this context, we allow for imbalanced local sample sizes, provided that the master machine contains a sufficient number of covariates. Conversely, in the case of the* SSDANE-Avg *method, we take the average of all* SSDANE *estimators from each machine in $\mathcal{U}$. In this case, when sample sizes across machines are imbalanced, the statistical rate becomes contingent on the smallest local sample size, denoted as* $\min\{n^*_j|j\in\mathcal{U}\}$. *In simpler terms, the presence of smaller local sample sizes can potentially impact the theoretical performance of* SSDANE-Avg *negatively. In practice, we suggest using the* SSDANE *method when the sample size across machines is extremely imbalanced. Notably, recent research efforts have delved into the realm of distributed training in the context of imbalanced data (see, e.g., Duan et al. (2021a,b); Chen et al. (2023)). It would indeed be intriguing to explore the fusion of these methodologies with our distributed semi-supervised approach.*

3.2 Two Examples

In this section, we provide two specific examples of the generalized linear model. The first one is linear regression. We use this example to illustrate the lower bound of the rate for DANE with labeled data and support the claims on better performance of our SSDANE in Section 2.2.1.

Linear regression Assume the observation pair $(\boldsymbol{X}, Y)$ comes from the following model

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon, \tag{19}$$

where ϵ denotes the Gaussian noise. In this case, the canonical link function is $\psi'(x) = x$. Therefore $\psi(x) = x^2/2$, which has a quadratic form. The third-order derivative of $\psi(x)$ vanishes for the linear model.

Proposition 6 *In the linear model, suppose the covariate $\boldsymbol{X}$ follows the sub-Gaussian distribution. Then the T-th round* SSDANE *estimator $\widehat{\boldsymbol{\beta}}^{(T)}$ satisfies that*

$$|\widehat{\boldsymbol{\beta}}^{(T)} - \boldsymbol{\beta}^*|_2 = O_{\mathbb{P}}\left\{\sqrt{\frac{p\log n^*}{mn}} + r_n\Big(\frac{p\log n^*}{n^*}\Big)^{T/2}\right\}.$$

For linear regression, the third term in (14) is disappeared when the canonical link function is in linear form. In the following, we provide the lower bound of one-step SSDANE.

Proposition 7 *In the linear model, assume the covariate $\boldsymbol{X}$ follows the sub-Gaussian distribution. Moreover, we assume that*

$$\inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \lambda_{\min}\left[\mathrm{C}ov\Big\{(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{\Sigma})\boldsymbol{v}\Big\}\right] \geq \rho_1.$$

Let $\widehat{\boldsymbol{\beta}}^{(0)}$ be an initial estimator which is independent to $\boldsymbol{X}_i$'s and ϵ_i's, then there exists a constant c, such that the one-step SSDANE *estimator $\widehat{\boldsymbol{\beta}}^{(1)}$ satisfies*

$$\mathbb{P}\left(|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*|_2 \geq c\Big(\sqrt{\frac{p}{mn}} + r_n\sqrt{\frac{p}{n^*}}\Big)\right) \geq \frac{1}{2}, \tag{20}$$

where $r_n = |\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*|_2$.

Suppose there is no additional unlabeled data and we have $n^* = n$. In this case, SSDANE becomes the DANE and $\widehat{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}_o^{(1)}$. Therefore, the DANE with only labeled data has the lower bound $c\Big(\sqrt{\frac{p}{mn}} + r_n\sqrt{\frac{p}{n}}\Big)$.

Logistic regression Assume the observation $(\boldsymbol{X}, Y) \in \mathbb{R}^{p+1}$ admits the following conditional probability function

$$\mathbb{P}(Y \mid \boldsymbol{X}) = \frac{\exp(Y\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)}{1 + \exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)}, \tag{21}$$

where the response variable Y only takes value in $\{0, 1\}$. Then the canonical link function $\psi'(\cdot)$ is defined as $\psi(x) = e^x/(1 + e^x)$. It is not hard to verify that ψ fulfills Assumption 3. Therefore, the theorems above hold for Logistic regression.

3.3 Theory on non-distributed setting

To show a trivial case involving unlabeled data will decline the statistical efficiency, we provide a result on asymptotic distribution in the non-distributed setting. Recall the loss function

$$\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i\in\mathcal{D}} Y_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \frac{1}{n^*}\sum_{i\in\mathcal{H}} \psi(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

Define

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ss}} \triangleq \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \Big\{\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta})\Big\}. \tag{22}$$

Proposition 8 *(Asymptotic normality of the semi-supervised GLM estimator) Suppose Assumption 1 to 3 hold, and additionally, we assume the rate constraint $p = o((n/\log n^*)^{2/3})$. Then for any vector $\boldsymbol{v} \in \mathbb{S}^{p-1}$, we have that*

$$\frac{\sqrt{n}}{\sigma(\boldsymbol{v})}\boldsymbol{v}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}} - \boldsymbol{\beta}^*) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\begin{aligned} &\big\{\sigma(\boldsymbol{v})\big\}^2 = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}^{-1}\Big(\boldsymbol{\mathcal{C}} + \frac{n^*-n}{n^*}\boldsymbol{\mathcal{C}}^{\mathrm{ss}}\Big)\boldsymbol{H}^{-1}\boldsymbol{v}, \qquad (23)\\ \textit{here}\quad &\boldsymbol{\mathcal{C}} = c(\sigma)\mathbb{E}\left[\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right], \quad \boldsymbol{\mathcal{C}}^{\mathrm{ss}} = \mathit{Cov}\big\{\psi'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\big\},\\ &\boldsymbol{H} = \mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}. \end{aligned}$$

Let $n^* = n$ in Proposition 8. In this case, $\mathcal{D} = \mathcal{H}$ and hence $\widehat{\boldsymbol{\beta}}^{\mathrm{ss}} = \widehat{\boldsymbol{\beta}}_{\mathcal{D}}$. That is, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ is $\boldsymbol{H}^{-1}\boldsymbol{\mathcal{C}}\boldsymbol{H}^{-1}$. While the unlabeled data is used, the covariance matrix would become to be $\boldsymbol{H}^{-1}\Big(\boldsymbol{\mathcal{C}} + \frac{n^*-n}{n^*}\boldsymbol{\mathcal{C}}^{\mathrm{ss}}\Big)\boldsymbol{H}^{-1}$ and hence has lower statistical efficiency.

4. Empirical Analysis

In the empirical analysis, we conduct several experiments to show the effectiveness of our proposed methods.

4.1 Simulation Studies on Synthetic Dataset

In this section, we show the performance of our proposed semi-supervised estimators on linear regression and logistic regression.

Parameter Settings In both models, we assume the i.i.d. covariate vectors $\boldsymbol{X}_i = (X_{i,1}, ..., X_{i,p})^{\mathrm{T}}$ are drawn from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for $i = 1, ..., N$. Here the covariance matrix $\boldsymbol{\Sigma}$ is a $p \times p$ Toeplitz matrix with its (i,j)-th entry $\Sigma_{ij} = 0.5^{|i-j|}$, where $1 \le i, j \le p$. We fix dimension $p = 20$ and the true coefficient

$$\boldsymbol{\beta}^* = (1, 0.95, 0.9, ..., 0.1, 0.05).$$

We repeat 100 independent simulations and report the averaged estimation error and the corresponding standard error. We mainly compare the SSDANE (Algorithm 1) and SSDANE-Avg (Algorithm 2) with the following six methods:

- DANE: Supervised version of distributed approximate Newton estimator. We run Algorithm 1 without using the additional unlabeled dataset.
- DANE-Avg: Supervised version of averaged distributed approximate Newton estimator. We run Algorithm 2 without using the additional unlabeled dataset.
- DANEimp: Imputation-based distributed approximate Newton estimator. We interpolate the unlabeled data by the current estimator $\widehat{\boldsymbol{\beta}}$, and run DANE for all pairs of data.
- DANEimp-Avg: Imputation-based averaged distributed approximate Newton estimator. We interpolate the unlabeled data by the current estimator $\widehat{\boldsymbol{\beta}}$, and run DANE-Avg for all pairs of data.
- Local estimator: We minimize the empirical loss function $\mathcal{L}_1(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i\in\mathcal{D}_1} Y_i \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \frac{1}{n}\sum_{i\in\mathcal{D}_1}\psi(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})$ with the data on the master machine $\mathcal{H}_1$.
- Pooled estimator: We collect all labeled pairs into one machine and minimize the empirical loss function.

In particular, we denote SSDANE-Avg(α) (DANE-Avg(α)) as the estimator which is the average of the local SSDANE (DANE) estimators from α fraction of machines. For the choice of the initial estimator, we uniformly use the local estimator on the master machine $\mathcal{H}_1$.

4.1.1 Results for Linear Regression

We first consider linear regression, where the observation $(\boldsymbol{X}, Y)$ is generated from the linear model (19).

Effect of the Number of Machines and Local Unlabeled Data To investigate the effect of the number of machines and local unlabeled data, we fix the labeled local sample size n to be 100, and vary the number of machines m from $\{20, 50, 100\}$, and the unlabeled local sample size n^* from $\{100, 200, 500\}$. We compare the one-step SSDANE and SSDANE-Avg with different averaging fractions. The results of the ℓ_2-errors and their standard errors are reported in Table 1.

As we can see from above, for every fixed m, more unlabeled samples help reduce the ℓ_2-error of the semi-supervised estimators. The averaged estimator SSDANE-Avg always has a lower error than SSDANE, which only solves the problem (11) on the master machine. Moreover, the more fraction of machines we average from, the smaller the estimation error we get. All of these findings coincide with the theoretical results in Theorem 1 and Theorem 4.

Covariance of the gradient at $\boldsymbol{\beta}^*$ From the proof of Theorem 2 and Proposition 8, we observe that the asymptotic variance of the estimators differs only in the covariance of the gradient of the loss function $\mathcal{L}$. Therefore, to demonstrate the statistical efficiency of our method in comparison with the naive semi-supervised estimator, we compare the rescaled covariance of the gradient at the true parameter $\boldsymbol{\beta}^*$. Specifically, for each round, we generate N samples and compute $N\nabla\mathcal{L}(\boldsymbol{\beta}^*)(\nabla\mathcal{L}(\boldsymbol{\beta}^*))^{\mathrm{T}}$. Then, we repeat this process 100 times and compute their average, which provides the covariance of the gradient $\nabla\mathcal{L}(\boldsymbol{\beta}^*)$

Table 1: The ℓ_2-errors and their standard errors (in parentheses) of the one-step SSDANE, SSDANE-Avg(0.2), SSDANE-Avg(0.5) and SSDANE-Avg(1) under labeled local sample size $n = 100$. Data are generated from a linear model with parameter dimension $p = 20$.

m	n^*	SSDANE	SSDANE-Avg(0.2)	SSDANE-Avg(0.5)	SSDANE-Avg(1)
20	100	0.588(0.226)	0.310(0.083)	0.236(0.053)	0.213(0.042)
	200	0.321(0.109)	0.193(0.041)	0.163(0.030)	0.154(0.026)
	500	0.203(0.040)	0.156(0.026)	0.145(0.028)	0.142(0.028)
50	100	0.586(0.228)	0.227(0.056)	0.199(0.041)	0.187(0.035)
	200	0.299(0.103)	0.131(0.029)	0.118(0.022)	0.112(0.020)
	500	0.173(0.039)	0.101(0.019)	0.096(0.017)	0.094(0.016)
100	100	0.593(0.235)	0.199(0.049)	0.184(0.038)	0.178(0.036)
	200	0.295(0.104)	0.106(0.023)	0.097(0.020)	0.095(0.019)
	500	0.161(0.039)	0.073(0.015)	0.070(0.014)	0.069(0.013)

Table 2: The rescaled trace of the covariance $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ of the $\mathcal{L}$, $\mathcal{L}^{\mathrm{ss}}$, $\widetilde{\mathcal{L}}^{(1)}$ and $\widetilde{\mathcal{L}}^{(5)}$ under labeled local sample size $n = 100$. Data are generated from a linear model with parameter dimension $p = 20$.

m	n^*	$\mathcal{L}$	$\mathcal{L}^{\mathrm{ss}}$	$\widetilde{\mathcal{L}}^{(1)}$	$\widetilde{\mathcal{L}}^{(5)}$
20	100	0.997	0.997	6.097	2891.910
	200	0.997	12.443	3.720	1.763
	500	0.997	18.476	2.196	1.046
50	100	0.979	0.979	13.591	9386.730
	200	0.979	12.530	7.425	3.062
	500	0.979	19.100	3.542	1.048
100	100	0.944	0.944	26.512	22642.400
	200	0.944	12.228	13.755	5.829
	500	0.944	18.367	5.822	0.998

(denoted as $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$). Here the loss function is chosen as $\mathcal{L}$ (fully supervised empirical loss), $\mathcal{L}^{\mathrm{ss}}$ (semi-supervised empirical loss), $\widetilde{\mathcal{L}}^{(1)}$ (1-step semi-supervised surrogate loss), and $\widetilde{\mathcal{L}}^{(5)}$ (5-step semi-supervised surrogate loss). We present the trace of the covariance (Table 2), the difference $\|\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L}) - \boldsymbol{C}\|$ (Table 3), and draw the distribution of eigenvalues of $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ (Figure 1)

From Table 2, it can be observed that fully supervised surrogate loss ($n^* = n$) has the largest covariance trace. This is because the supervised estimator has a large bias. As n^* grows, the semi-supervised empirical loss exhibits a larger covariance trace than that of $\mathcal{L}$, which aligns with the result in Proposition 8. On the contrary, the semi-supervised surrogate loss has a smaller covariance trace. In particular, for large iteration numbers and unlabeled sample sizes, the semi-supervised surrogate loss has a covariance trace similar to that of $\mathcal{L}$. Table 3 reveals that $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ approximates $\boldsymbol{C}$ for SSDANE when the local unlabeled sample size is large. The naive semi-supervised loss has a larger difference as the unlabeled sample size increases. Figure 1 demonstrates similar implications.

Table 3: The norm of difference $\|\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})-\boldsymbol{\mathcal{C}}\|$ of the $\mathcal{L}$, $\mathcal{L}^{\text{ss}}$, $\widetilde{\mathcal{L}}^{(1)}$ and $\widetilde{\mathcal{L}}^{(5)}$ under labeled local sample size $n = 100$. Data are generated from a linear model with parameter dimension $p = 20$.

m	n^*	$\mathcal{L}$	$\mathcal{L}^{\text{ss}}$	$\widetilde{\mathcal{L}}^{(1)}$	$\widetilde{\mathcal{L}}^{(5)}$
20	100	1.787	1.787	3.211×10^1	3.331×10^4
	200	1.787	7.892×10^1	1.716×10^1	6.037
	500	1.787	1.121×10^2	8.215	1.963
50	100	2.224	2.224	7.789×10^1	1.157×10^5
	200	2.224	8.153×10^1	4.016×10^1	1.634×10^1
	500	2.224	1.262×10^2	1.698×10^1	2.290
100	100	2.138	2.138	1.554×10^2	2.717×10^5
	200	2.138	8.200×10^1	7.753×10^1	3.946×10^1
	500	2.138	1.214×10^2	3.028×10^1	2.202

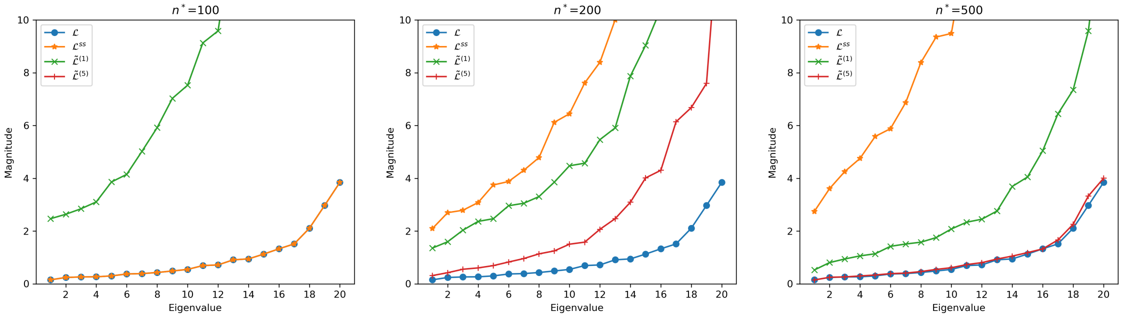

Figure 1: The values of eigenvalues of $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ in increasing order under the linear regression model. The labeled local sample size takes value in 100, the number of machines is 50, the unlabeled local sample size takes value in $\{0, 100, 400\}$, and the dimension p is 20.

Effect of the Number of Iterations Next, we study the effect of iterations. We set the number of machines to be 100, the unlabeled local sample size to be 400, and vary the labeled local sample size from $\{50, 100, 200\}$. The curves of ℓ_2-error over the number of iterations are presented in Figure 2.

From Figure 2, we can see that in all cases, the SSDANE and SSDANE-Avg outperform their fully supervised counterparts. The superiority is more obvious when the labeled local sample size n is small. The imputation-based method DANEimp is similar to SSDANE when only one machine has unlabeled data, but the averaged estimator DANEimp-Avg is inferior to both SSDANE-Avg and DANE-Avg. This indicates that a naive combination of imputation with DANE may lead to a deterioration in performance. Moreover, we find that the fully supervised estimator DANE blows up in several steps when n is small, while the semi-supervised estimators always converge well.

Effect of initialization From Proposition 6, in the linear model, the convergence rate always increases as the iteration number grows. In the following experiment, we examine if

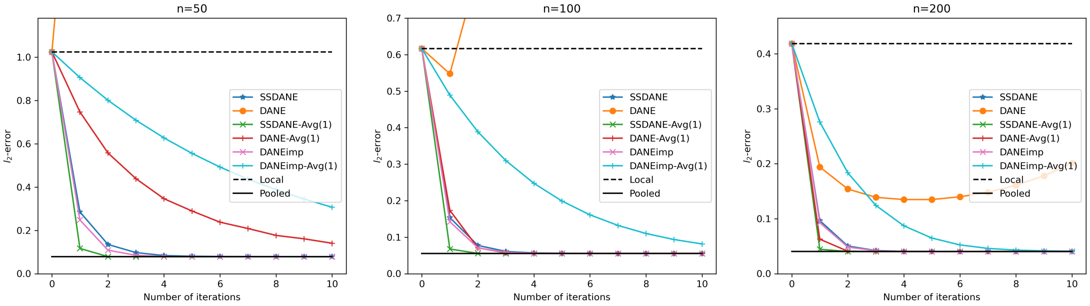

Figure 2: The ℓ_2-error over the number of iterations under linear model. The labeled local sample size takes value in $\{50, 100, 200\}$, the number of machine is 100, the unlabeled local sample size is 400, and the dimension p is 20.

Table 4: The ℓ_2-errors and their standard errors (in parentheses) of the 1-step SSDANE, 5-step SSDANE, 1-step SSDANE-Avg, and 5-step SSDANE-Avg with initialization $\widehat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, under labeled local sample size $n = 100$. Data are generated from a linear model with parameter dimension $p = 20$.

m	n^*	1 Step SSDANE	1 Step SSDANE-Avg	5 Step SSDANE	5 Step SSDANE-Avg
20	100	3.434(1.212)	1.169(0.222)	28.978(37.931)	0.126(0.022)
	200	2.135(0.717)	0.682(0.165)	1.476(1.927)	0.125(0.022)
	500	1.248(0.354)	0.580(0.160)	0.145(0.038)	0.125(0.022)
50	100	3.496(1.273)	1.177(0.221)	33.016(48.324)	0.080(0.016)
	200	2.128(0.722)	0.587(0.132)	1.462(2.004)	0.079(0.015)
	500	1.184(0.338)	0.384(0.107)	0.989(0.032)	0.079(0.015)
100	100	3.492(1.300)	1.177(0.216)	33.780(51.572)	0.056(0.010)
	200	2.138(0.734)	0.558(0.114)	1.475(2.026)	0.056(0.010)
	500	1.175(0.346)	0.307(0.085)	0.077(0.028)	0.056(0.010)

the estimation error is sensitive to the initial point. To see that, we take the initial point as $\widehat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where the noise $\boldsymbol{\epsilon}$ is sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. It is not hard to see that $\mathbb{E}[|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*|_2] \approx \sqrt{p}$, which violate the assumption in the theory. The result is presented in Table 4. As shown in the linear model, the fully supervised estimator ($n^* = n$) is highly sensitive to initialization compared to the semi-supervised estimator ($n^* > n$). Increasing the amount of unlabeled data leads to greater robustness of the estimator.

Moreover, we extend our exploration to encompass the utilization of distributed batch stochastic gradient descent (BSGD) and local stochastic gradient descent (LSGD) methodologies for initialization purposes. In this context, our approach employs a batch size of $0.1n$ for both BSGD and LSGD, while conducting a total of 10 iterations. Within the framework of BSGD, a constant step size is employed, while in alignment with the principles outlined in Stich (2019), the step size in LSGD is determined as $C/(t+1)$. We compare SSDANE and SSDANE-Avg initialized by the local estimator, random initialization ($\widehat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^* + \boldsymbol{\epsilon}$),

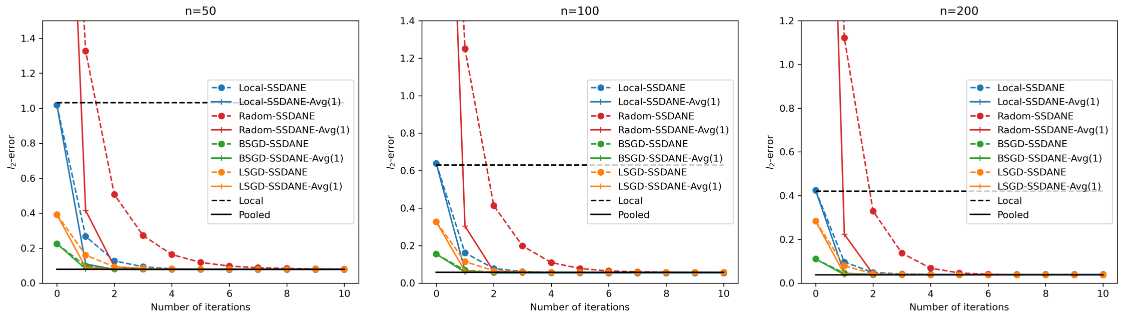

Figure 3: The ℓ_2-error over the number of iterations under a linear model with different initializations. The labeled local sample size takes value in $\{50, 100, 200\}$, the number of machines is 100, the unlabeled local sample size is 400, and the dimension p is 20.

early-stopped BSGD and LSGD, and show how their estimation errors evolve with respect to iterations. The results are elucidated in Figure 3.

As we can see from Figure 3, in the linear model, all these four initialization methods lead to consistent estimators. The SSDANE-Avg method always outperforms its corresponding SSDANE method. Notably, the BSGD and LSGD initialized methods exhibit smaller estimation errors, consequently requiring fewer iterations to attain the optimal rate. In practical applications, the choice of initialization can be tailored to the characteristics of the problem at hand. Specifically, when confronted with scenarios of ample local sample sizes, the employment of a local minimizer for initialization is advocated, considering its intrinsic advantage of circumventing communication overhead. Conversely, for a situation involving diminutive local sample sizes, the adoption of BSGD and LSGD for initialization serves as a prudent approach, aligning with the requisite criterion of achieving $|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*|_2 = o_{\mathbb{P}}(p^{-1/2})$.

4.1.2 Results for Logistic Regression

In this section, we consider logistic regression, where the observation $(\boldsymbol{X}, Y)$ is generated from the model (21). To solve the optimization problem (13) for the logistic regression model, we apply conjugate gradient descent motivated by Minka (2003).

Effect of the Number of Machines and Local Unlabeled Data Similarly, as in the linear model, we fix the labeled local sample size n to be 300 and vary the number of machines m from $\{20, 50, 100\}$, and the unlabeled local sample size n^* from $\{300, 450, 900\}$. We compare the one-step SSDANE and SSDANE-Avg with different averaging fractions. The results are shown in Table 5. We can find similar phenomena as in the linear model, which shows the unlabeled dataset and averaging fraction play important roles in reducing the estimation errors.

Covariance of the gradient at $\boldsymbol{\beta}^*$ In this part, we compare the various quantities of the covariance matrix $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ including the rescaled trace (Table 6), the difference $\|\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L}) - \boldsymbol{\mathcal{C}}\|$ (Table 7), and draw the distribution of eigenvalues of $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ (Figure 4). We can observe a similar phenomenon as in the linear model.

Table 5: The ℓ_2-errors and their standard errors (in parentheses) of the one-step SSDANE, SSDANE-Avg(0.2), SSDANE-Avg(0.5) and SSDANE-Avg(1) under labeled local sample size $n = 300$. Data are generated from a logistic regression model with parameter dimension $p = 20$.

m	n^*	SSDANE	SSDANE-Avg(0.2)	SSDANE-Avg(0.5)	SSDANE-Avg(1)
20	300	0.692(0.267)	0.414(0.105)	0.347(0.074)	0.325(0.061)
	450	0.510(0.154)	0.347(0.076)	0.309(0.064)	0.294(0.053)
	900	0.375(0.080)	0.299(0.053)	0.284(0.052)	0.278(0.049)
50	300	0.656(0.301)	0.280(0.072)	0.247(0.054)	0.239(0.047)
	450	0.455(0.176)	0.226(0.053)	0.204(0.042)	0.199(0.037)
	900	0.309(0.079)	0.189(0.035)	0.178(0.032)	0.177(0.031)
100	300	0.634(0.281)	0.226(0.046)	0.210(0.039)	0.204(0.036)
	450	0.437(0.169)	0.172(0.033)	0.161(0.028)	0.157(0.028)
	900	0.285(0.079)	0.137(0.023)	0.131(0.022)	0.129(0.022)

Table 6: The rescaled of the covariace $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ of $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ of $\mathcal{L}$, $\mathcal{L}^{\mathrm{ss}}$, $\widetilde{\mathcal{L}}^{(1)}$ and $\widetilde{\mathcal{L}}^{(5)}$ under labeled local sample size $n = 300$. Data are generated from a logistic regression model with parameter dimension $p = 20$.

m	n^*	$\mathcal{L}$	$\mathcal{L}^{\mathrm{ss}}$	$\widetilde{\mathcal{L}}^{(1)}$	$\widetilde{\mathcal{L}}^{(5)}$
20	300	0.074	0.074	0.222	0.271
	450	0.074	0.218	0.174	0.091
	900	0.074	0.355	0.130	0.077
50	300	0.073	0.073	0.441	0.693
	450	0.073	0.227	0.325	0.108
	900	0.073	0.348	0.211	0.078
100	300	0.076	0.076	0.809	1.421
	450	0.076	0.213	0.570	0.133
	900	0.076	0.344	0.336	0.078

Effect of the Number of Iterations In this experiment, we fix the number of machines as 100, the unlabeled local sample size as 600, and vary the labeled local sample size from $\{100, 150, 300\}$. Similarly, from Figure 5, we find that the semi-supervised estimators SSDANE and SSDANE-Avg are more stable and accurate, especially when the labeled local sample size n is small. Moreover, DANEimp-Avg always performs worse than SSDANE-Avg.

Effect of initialization In Theorem 1 and 4, our theory suggests that the algorithm converges only when the initial rate $r_n = o(1/\sqrt{p})$. In this experiment, we check if this requirement is necessary. Similarly, as in the linear model, we take the initial point as $\widehat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where the noise $\boldsymbol{\epsilon}$ is sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$. The result of ℓ_2-error is reported in Table 8.

In the table, we denote '/' if the algorithm blows up. We can observe a similar trend as in the linear model, where the semi-supervised estimator shows greater robustness to initialization than its fully supervised counterpart. Additionally, the SSDANE algorithm

Table 7: The norm of difference $\|\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L}) - \boldsymbol{\mathcal{C}}\|$ of $\mathcal{L}$, $\mathcal{L}^{\text{ss}}$, $\widetilde{\mathcal{L}}^{(1)}$ and $\widetilde{\mathcal{L}}^{(5)}$ under labeled local sample size $n = 300$. Data are generated from a logistic regression model with parameter dimension $p = 20$.

m	n^*	$\mathcal{L}$	$\mathcal{L}^{\text{ss}}$	$\widetilde{\mathcal{L}}^{(1)}$	$\widetilde{\mathcal{L}}^{(5)}$
20	300	0.152	0.152	0.937	1.578
	450	0.152	0.935	0.668	0.224
	900	0.152	1.702	0.411	0.163
50	300	0.138	0.138	2.250	5.041
	450	0.138	0.967	1.567	0.349
	900	0.138	1.662	0.908	0.143
100	300	0.146	0.146	4.424	10.456
	450	0.146	0.900	2.975	0.549
	900	0.146	1.610	1.618	0.154

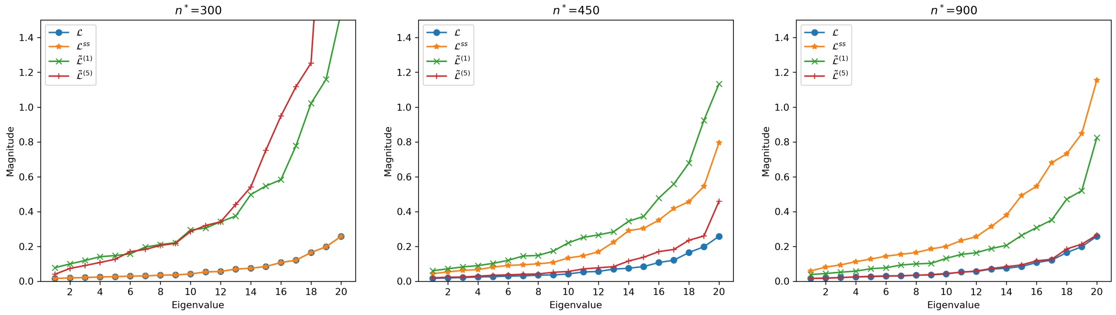

Figure 4: The values of eigenvalues of $\widehat{\boldsymbol{\Sigma}}(\nabla\mathcal{L})$ in increasing order under the logistic regression model. The labeled local sample size is 300, the number of machines is 50, the unlabeled local sample size takes value in $\{0, 150, 600\}$, and the dimension p is 20.

exhibits greater robustness than SSDANE-Avg. These findings suggest that there may be scope for further improvement in Theorem 1 and 4. We can also find that the 5-step SSDANE exhibits a large variance or even blows up when confronted with scenarios of a limited number of covariates n^*. We concur that this behavior could be attributed to unfavorable initialization coupled with the small value of n^*. In the context of our theoretical framework, specifically referencing Theorem 1 and Theorem 4, it is discernible that when $\sqrt{p}r_n$ or $p\log n^*/n^*$ is greater than 1, the statistical rate escalates with an increasing number of steps. This phenomenon potentially elucidates the reasoning behind the occurrence where a 5-step SSDANE is worse than its single-step counterpart in certain instances.

We extend our experiment to investigate the utilization of an array of initialization strategies, including the use of the local minimizer, random input, early-stopped BSGD and LSGD, thereby mirroring the approach adopted in the linear model context. The selection of hyperparameters aligns with the configuration outlined in Section 4.1.1. Con-

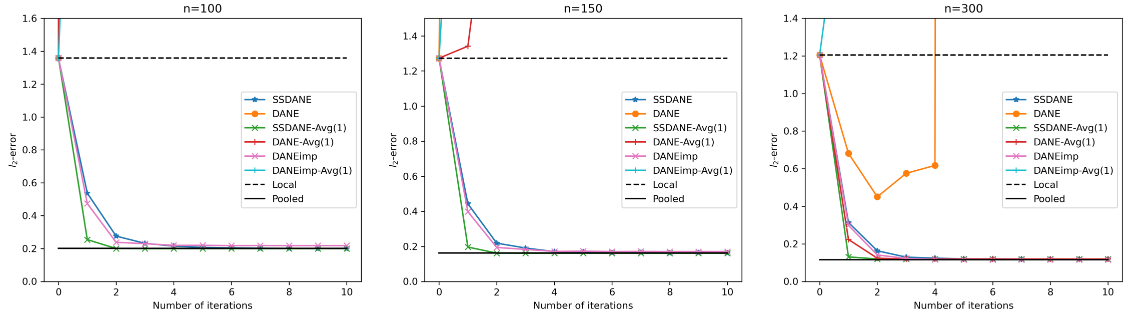

Figure 5: The ℓ_2-error over the number of iterations under the logistic regression model. The labeled local sample size takes value in $\{100, 150, 300\}$, the number of machines is 100, the unlabeled local sample size is 600, and the dimension p is 20.

Table 8: The ℓ_2-errors and their standard errors (in parentheses) of the 1-step SSDANE, 5-step SSDANE, 1-step SSDANE-Avg, and 5-step SSDANE-Avg with initialization $\widehat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, under labeled local sample size $n = 300$. Data are generated from a logistic regression model with parameter dimension $p = 20$.

m	n^*	1 Step SSDANE	1 Step SSDANE-Avg	5 Step SSDANE	5 Step SSDANE-Avg
20	300	10.193(20.198)	115.690(450.732)	/	/
	450	4.715(10.561)	3.239(4.366)	0.431(0.252)	/
	900	1.532(2.203)	1.042(1.651)	0.264(0.053)	/
50	300	10.245(21.135)	133.470(316.598)	/	/
	450	4.963(12.696)	2.994(3.789)	/	/
	900	1.385(1.614)	0.860(1.254)	0.174(0.047)	/
100	300	10.197(21.368)	89.883(180.015)	/	/
	450	4.912(13.116)	7.447(44.892)	/	/
	900	1.400(2.047)	0.757(0.961)	0.127(0.043)	/

trasting with the findings in Figure 3, we can find that random initialization may result in divergence in the SSDANE-Avg method. However, as the local sample size increases, the SSDANE-Avg method exhibits convergence. This observation suggests that both our SSDANE and SSDANE-Avg methods benefit from a well-chosen initialization, and a larger local sample size exerts a positive influence in mitigating sensitivity to the initialization conditions. This phenomenon can be partially elucidated through Theorem 1 and 4.

4.2 Application to Real-World Benchmarks

In this section, we analyze the CelebA dataset[2] from the Kaggle website, which is included in LEAF (Caldas et al., 2018), a standard distributed learning benchmark. Our aim is to train a classifier that distinguishes young people from old ones. In the dataset, each variable has 39 attributes. We take the total sample size as 120000, and randomly partition the dataset

[2] https://www.kaggle.com/datasets/jessicali9530/celeba-dataset

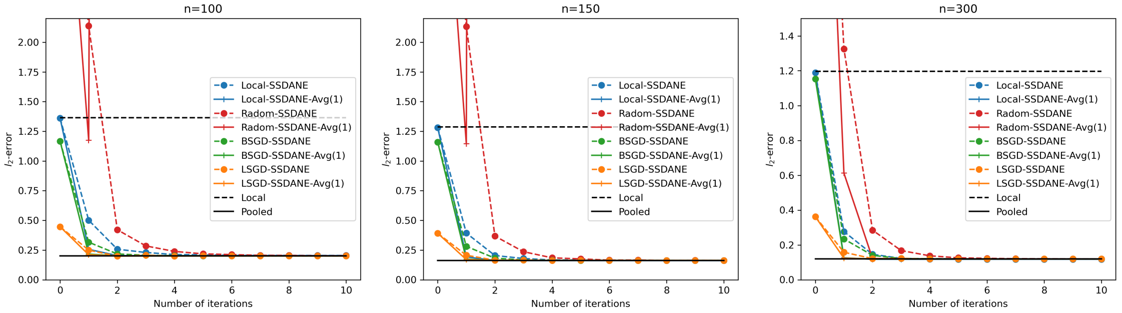

Figure 6: The ℓ_2-error over the number of iterations under a logistic model with different initializations. The labeled local sample size takes value in $\{100, 150, 300\}$, the number of machines is 100, the unlabeled local sample size is 600, and the dimension p is 20.

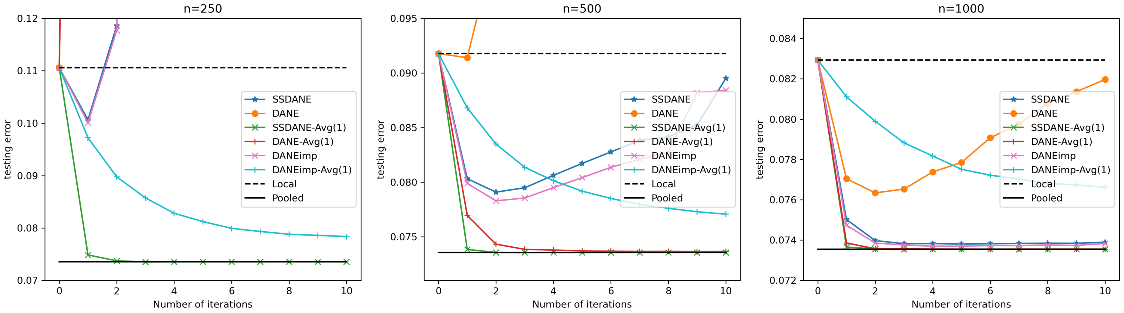

Figure 7: CelebA dataset. The classification error over iterations, under various pairs of (m, n). The total training sample size N is 20000, the unlabeled sample size is 80000, the test sample size is 20000, and the dimension p is 39.

into 20000 testing data, 20000 labeled training data, and 80000 unlabeled training data. We perform 100 random partitions of the dataset and report the averaged classification error on the testing set. We consider three cases where $(m, n) = (20, 1000)$,$(40, 500)$ and $(80, 250)$ respectively. The result is shown in Figure 7. It is not hard to see that the proposed semi-supervised methods always outperform their supervised counterparts, and the superiority of the semi-supervised method is more obvious when the labeled local sample size n is small. While the imputation-based method DANEimp is comparable to SSDANE, the averaged estimator DANEimp-Avg is much worse than SSDANE-Avg. This result also reveals that naively combining imputation with DANE may not improve the training performance.

5. Concluding Remarks and Future Study

In this paper, we study the semi-supervised generalized linear model in the distributed setup. With the assistance of the additional unlabeled data, we theoretically show that the proposed estimators enjoy higher statistical accuracy than their fully supervised counterparts. The superiority is also illustrated by numerical experiments. In the future, there are several directions worth being explored. For example, it is interesting to study distributed semi-supervised learning in a high-dimensional regime. We believe that the unlabeled data can obtain a similar accuracy-enhancement effect when incorporated with the one-shot averaging method (Lee et al., 2017) and the multi-round distributed sparse learning method (Wang et al., 2017; Jordan et al., 2019). Second, while we focus on the scenario where the model is well-specified, it is important to consider the distributed semi-supervised learning for miss-specified models (Bellec et al., 2018; Chakrabortty and Cai, 2018; Zhang and Bradic, 2021; Deng et al., 2020; Azriel et al., 2022). Third, it will be of great importance to explore the distributed semi-supervised ridge regression (Dobriban and Sheng, 2021, 2020; Sheng and Dobriban, 2020). Lastly, of profound significance is the intriguing prospect of synergistically investigating the amalgamation of research endeavors focused on distributed training within the context of imbalanced data (see, e.g., Duan et al. (2021a,b); Chen et al. (2023)) with our distributed semi-supervised approach.

Acknowledgments

Jiyuan Tu's research is supported by "the Fundamental Research Funds for the Central Universities". Weidong Liu's research is supported by NSFC Grant No. 11825104. Xiaojun Mao's research is supported by NSFC Grant No. 12371273, the Shanghai Rising-Star Program 23QA1404600 and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001). The authors would like to thank the action editor and three anonymous referees for their constructive comments, which greatly improves the quality of the paper.

Appendix

The appendix consists of three parts. In Appendix A, we provide proof for the theoretical results of SSDANE and SSDANE-Avg in Section 3.1. In Appendix B, we prove the theories in Section 3.2, especially the rate lower bound of SSDANE for the linear model. Appendix C is devoted to proving the normality result for the semi-supervised generalized linear model in the single machine setting, namely the theorems in Section 3.3.

Appendix A. Proof of Theories in Section 3.1

Proof [Proof of Equation (6)] Notice that

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i\in\mathcal{D}}\big\{-Y_i\boldsymbol{X}_i+\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\big\}\Big]=\mathbf{0}=\mathbb{E}\Big[\frac{1}{n^*}\sum_{i\in\mathcal{H}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)-\frac{1}{n}\sum_{i\in\mathcal{D}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\Big],$$

and

$$Cov\Big(\frac{1}{n}\sum_{i\in\mathcal{D}}\big\{-Y_i\boldsymbol{X}_i+\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\big\},\frac{1}{n^*}\sum_{i\in\mathcal{H}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)-\frac{1}{n}\sum_{i\in\mathcal{D}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\Big)=\mathbf{0},$$

we have that

$$\begin{aligned}&Cov\Big(-\frac{1}{n}\sum_{i\in\mathcal{D}}Y_i\boldsymbol{X}_i+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\Big) \qquad (24)\\ =&\mathrm{Cov}\Big(\frac{1}{n}\sum_{i\in\mathcal{D}}\big\{-Y_i\boldsymbol{X}_i+\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\big\}\Big)+\mathrm{Cov}\Big(\frac{1}{n^*}\sum_{i\in\mathcal{H}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)-\frac{1}{n}\sum_{i\in\mathcal{D}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\Big).\end{aligned}$$

We shall compute them respectively. For the first term, since each $\boldsymbol{X}_i$ are independent to each other, we have that

$$Cov\Big(\frac{1}{n}\sum_{i\in\mathcal{D}}\big\{-Y_i\boldsymbol{X}_i+\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\big\}\Big)=\frac{1}{n}Cov\Big(-Y\boldsymbol{X}+\boldsymbol{X}\psi^{'}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\Big).$$

For the second term in (24), we compute that

$$\begin{aligned}&Cov\Big(\frac{1}{n^*}\sum_{i\in\mathcal{H}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)-\frac{1}{n}\sum_{i\in\mathcal{D}}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\Big)\\ =&\mathrm{C}ov\Big(\sum_{i\in\mathcal{D}^*}\frac{1}{n^*}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)-\sum_{i\in\mathcal{D}}\frac{n^*-n}{nn^*}\boldsymbol{X}_i\psi^{'}(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\Big)\\ =&\Big(\frac{n^*-n}{(n^*)^2}+\frac{n(n^*-n)^2}{(nn^*)^2}\Big)\mathrm{Cov}\Big(\boldsymbol{X}\psi'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\Big)=\frac{n^*-n}{nn^*}\mathrm{C}ov\Big(\boldsymbol{X}\psi'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\Big).\end{aligned}$$

Then (6) can be obtained by substituting the above two formulas into (24). ■

Lemma 9 *Given a convex loss function $\bar{\mathcal{L}}(\boldsymbol{\beta})$, suppose it is second-order differentiable, and there exist constants ρ_0, r_0 such that*

$$\inf_{\boldsymbol{\beta}:|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2\le r_0}\ \inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \boldsymbol{v}^{\mathrm{T}}\nabla^2\bar{\mathcal{L}}(\boldsymbol{\beta})\boldsymbol{v}\ge\rho_0, \tag{25}$$

and $r_0 > 3\rho_0^{-1}|\nabla\bar{\mathcal{L}}(\boldsymbol{\beta}^*)|_2$, then the minimizer $\bar{\boldsymbol{\beta}}$ of $\bar{\mathcal{L}}(\boldsymbol{\beta})$ satisfies that

$$|\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_2\le\frac{3}{\rho_0}|\nabla\bar{\mathcal{L}}(\boldsymbol{\beta}^*)|_2.$$

Proof For simplicity, we denote $b_n = 3\rho_0^{-1}|\nabla\bar{\mathcal{L}}(\boldsymbol{\beta}^*)|_2$, and construct the set $\Theta_1=\{\boldsymbol{\beta}:|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2=b_n\}$ in the parameter space. Then clearly we know that $\Theta_1\subseteq\{\boldsymbol{\beta}:|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2\le r_0\}$. We will show that $\bar{\mathcal{L}}(\boldsymbol{\beta})>\bar{\mathcal{L}}(\boldsymbol{\beta}^*)$ strictly holds uniformly for all $\boldsymbol{\beta}\in\Theta_1$. To show this, we compute that

$$\begin{aligned}
&\bar{\mathcal{L}}(\boldsymbol{\beta})-\bar{\mathcal{L}}(\boldsymbol{\beta}^*)\\
=&\int_0^1(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\nabla\bar{\mathcal{L}}\big\{\boldsymbol{\beta}^*+t(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\big\}\mathrm{d}t\\
=&(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\nabla\bar{\mathcal{L}}(\boldsymbol{\beta}^*)+\int_0^1(1-t)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\nabla^2\bar{\mathcal{L}}\big\{\boldsymbol{\beta}^*+t(\boldsymbol{\beta}-\boldsymbol{\beta})\big\}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\mathrm{d}t\\
\ge&\frac{\rho_0}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2^2-|\nabla\bar{\mathcal{L}}(\boldsymbol{\beta}^*)|_2\cdot|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2\\
=&\frac{\rho_0}{2}\times3\rho_0^{-1}|\nabla\bar{\mathcal{L}}(\boldsymbol{\beta}^*)|_2|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2-|\nabla\bar{\mathcal{L}}(\boldsymbol{\beta}^*)|_2\cdot|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2>0.
\end{aligned}$$

Then by convexity of the loss function $\bar{\mathcal{L}}(\boldsymbol{\beta})$, we have that $|\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|_2\le\frac{3}{\rho_0}|\nabla\bar{\mathcal{L}}(\boldsymbol{\beta}^*)|_2$, which completes the proof. ■

Lemma 10 *Suppose Assumption 1 to 3 hold. Assume $r=o(1)$, then for each $t\ge1$, there is*

$$\inf_{\boldsymbol{\beta}:|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2\le r}\ \inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\boldsymbol{v}^{\mathrm{T}}\nabla^2\widetilde{\mathcal{L}}^{(t)}(\boldsymbol{\beta})\boldsymbol{v}\ge\frac{\rho}{2},$$

holds with probability not less than $1-O((n^*)^{-p\tau})$ for some constant $\tau>0$.

Proof For arbitrary $\boldsymbol{\beta}$ satisfying $|\boldsymbol{\beta}-\boldsymbol{\beta}|_2\le r$, there is

$$\begin{aligned}
&\boldsymbol{v}^{\mathrm{T}}\nabla^2\widetilde{\mathcal{L}}^{(t)}(\boldsymbol{\beta})\boldsymbol{v}\\
=&\frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\psi''\big(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\big)\big(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i\big)^2\\
=&\boldsymbol{v}^{\mathrm{T}}\mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}\boldsymbol{v}+\underbrace{\frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\int_0^1\psi'''\big(\boldsymbol{X}_i^{\mathrm{T}}\{\boldsymbol{\beta}^*+t(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\}\big)\big\{\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\big\}\big(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i\big)^2\mathrm{d}t}_{T_1}\\
&+\boldsymbol{v}^{\mathrm{T}}\Big[\underbrace{\frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\psi''\big(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*\big)\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}-\mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}}_{\boldsymbol{T}_2}\Big]\boldsymbol{v}.
\end{aligned}\tag{26}$$

It left to bound the term $|T_1|$ and $\|\boldsymbol{T}_2\|$ respectively.
Bound of $|T_1|$: From Assumption 3, It is not hard to see that

$$\begin{aligned}
|T_1| =& \left| \frac{1}{n^*} \sum_{i\in\mathcal{H}_1} \int_0^1 \psi'''\big(\boldsymbol{X}_i^{\mathrm{T}}\{\boldsymbol{\beta}^* + t(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\}\big)\big\{\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\big\}\big(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i\big)^2 \mathrm{d}t \right| \\
\le& c_\psi r \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i|^3 \right|.
\end{aligned} \tag{27}$$

Denote $\mathfrak{N}$ as the 1/2-net of $\mathbb{S}^{p-1}$, by Lemma 5.2 of Vershynin (2010) we know that $|\mathfrak{N}| \le 5^p$. Then there is

$$\begin{aligned}
&\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i|^3 \right| \\
\le& 4\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3 \right| + 4\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \sup_{\boldsymbol{v}:|\boldsymbol{v}-\widetilde{\boldsymbol{v}}|_2\le 1/2} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |(\widetilde{\boldsymbol{v}}-\boldsymbol{v})^{\mathrm{T}}\boldsymbol{X}_i|^3 \right| \\
\le& 4\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3 \right| + \frac{1}{2}\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i|^3 \right| \\
\Rightarrow& \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i|^3 \right| \le 8\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3 \right| \\
\le& 8\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3 - \mathbb{E}\big\{|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3\big\} \right| + 8\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\big\{|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3\big\}.
\end{aligned}$$

Note that for each $\widetilde{\boldsymbol{v}}\in\mathfrak{N}$, there is

$$\begin{aligned}
&\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \mathbb{E}\left[\exp\left\{\left(\eta^{3/2}|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3\right)^{2/3}\right\}\right] \\
\le& \sup_{\widetilde{\boldsymbol{v}}\in\mathbb{S}^{p-1}} \mathbb{E}\left[\exp\left\{\eta|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^2\right\}\right] \le C_0.
\end{aligned}$$

We know that $|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3$'s are sub-Weibull(2/3) random variables. Therefore, by Theorem 3.1 of Kuchibhotla and Chakrabortty (2022), there exist constants $c_3, \tau > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3 - \mathbb{E}\big\{|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3\big\} \right| \ge c_3\left\{ \sqrt{\frac{p\log n^*}{n^*}} + \frac{(p\log n^*)^{3/2}}{n^*} \right\} \right) \\
\le& 5^p \max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \mathbb{P}\left(\left| \frac{1}{n^*}\sum_{i\in\mathcal{H}_1} |\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3 - \mathbb{E}\big\{|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3\big\} \right| \ge c_3\left\{ \sqrt{\frac{p\log n^*}{n^*}} + \frac{(p\log n^*)^{3/2}}{n^*} \right\} \right) \\
\le& 5^p (n^*)^{-\tau_1 p} \le (n^*)^{-(\tau_1-1)p},
\end{aligned} \tag{28}$$

which tends to 0 when $\tau_1 > 1$. On the other hand, we compute that

$$\begin{aligned}
&\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\big\{|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3\big\} \\
\le&\frac{1}{2\eta_0^{3/2}} \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\big[\exp\big\{\eta_0|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^2\big\}\big] \le \frac{C_0}{2\eta_0^{3/2}},
\end{aligned} \tag{29}$$

where the second line uses the elementary inequality $|x|^3 \le \frac{1}{2}e^{|x|}$. Combining (27), (28) and (29) we have that

$$\begin{aligned}
|T_1| \le& c_\psi r \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \left| \frac{1}{n^*} \sum_{i\in\mathcal{H}_1} |\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i|^3 \right| \\
\le& 8c_\psi r \max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \left| \frac{1}{n^*} \sum_{i\in\mathcal{H}_1} |\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3 - \mathbb{E}\big\{|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3\big\} \right| + 8c_\psi r \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}\big\{|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i|^3\big\} \\
=& r \times O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{n^*}} + \frac{(p\log n^*)^{3/2}}{n^*} + 1 \right) = o_{\mathbb{P}}(1).
\end{aligned}$$

Bound of $\|\boldsymbol{T}_2\|$: Denote $\mathfrak{N}$ as the 1/4-net of $\mathbb{S}^{p-1}$, by Lemma 5.2 of Vershynin (2010) we know that $|\mathfrak{N}| \le 9^p$. Then there is

$$\begin{aligned}
\|\boldsymbol{T}_2\| =& \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \big|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{T}_2\boldsymbol{v}\big| \\
\le& \max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \big|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{T}_2\widetilde{\boldsymbol{v}}\big| + 2\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \sup_{\boldsymbol{v}:|\boldsymbol{v}-\widetilde{\boldsymbol{v}}|_2\le 1/4} \big|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{T}_2(\widetilde{\boldsymbol{v}}-\boldsymbol{v})\big| + \max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \sup_{\boldsymbol{v}:|\boldsymbol{v}-\widetilde{\boldsymbol{v}}|_2\le 1/4} \big|(\widetilde{\boldsymbol{v}}-\boldsymbol{v})^{\mathrm{T}}\boldsymbol{T}_2(\widetilde{\boldsymbol{v}}-\boldsymbol{v})\big| \\
\le& \max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \big|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{T}_2\widetilde{\boldsymbol{v}}\big| + \frac{1}{2}\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \big|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{T}_2\boldsymbol{v}\big| + \frac{1}{16}\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \big|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{T}_2\boldsymbol{v}\big| \\
\Rightarrow& \|\boldsymbol{T}_2\| \le \frac{16}{7}\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \big|\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{T}_2\widetilde{\boldsymbol{v}}\big|.
\end{aligned} \tag{30}$$

For each $\widetilde{\boldsymbol{v}} \in \mathfrak{N}$, it is not hard to see that $\psi''\big(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*\big)(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i)^2$ is sub-exponential random variable. Therefore we can apply Lemma 1 of Cai and Liu (2011) and yield

$$\begin{aligned}
&\mathbb{P}\left(\max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \frac{1}{n^*} \sum_{i\in\mathcal{H}_1} \psi''\big(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*\big)(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i)^2 - \mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X})^2\big\} \ge c_4\sqrt{\frac{p\log n^*}{n^*}} \right) \\
\le& 9^p \max_{\widetilde{\boldsymbol{v}}\in\mathfrak{N}} \mathbb{P}\left(\frac{1}{n^*} \sum_{i\in\mathcal{H}_1} \psi''\big(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*\big)(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i)^2 - \mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X})^2\big\} \ge c_4\sqrt{\frac{p\log n^*}{n^*}} \right) \\
\le& 9^p(n^*)^{-\tau_2 p} \le (n^*)^{-(\tau_2-1)p},
\end{aligned}$$

for some constant $\tau_2 > 0$. Substitute this bound into (30) we have that $\|\boldsymbol{T}_2\| = o(1)$.

Substitute the bound of $|T_1|$ and $\|\boldsymbol{T}_2\|$ into (26) we finally have that

$$\begin{aligned}
&\boldsymbol{v}^{\mathrm{T}}\nabla^2\widetilde{\mathcal{L}}^{(t)}(\boldsymbol{\beta})\boldsymbol{v} \\
\ge& \boldsymbol{v}^{\mathrm{T}}\mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}\boldsymbol{v} - |T_1| - \|\boldsymbol{T}_2\| \ge \frac{\rho}{2},
\end{aligned}$$

hold with probability not less than $1 - (n^*)^{-(\tau_1-1)p} - (n^*)^{-(\tau_2-1)p} = 1 - O((n^*)^{-\tau p})$ by taking $\tau = \min\{\tau_1 - 1, \tau_2 - 1\}$, which proves the lemma. ■

Proof [Proof of Theorem 1] For simplicity, we first consider the convergence rate of $\widehat{\boldsymbol{\beta}}^{(1)}$. By Lemma 10, we know the condition (25) is sufficed with high probability. Now using Lemma 9, we know that it is enough to bound $|\widetilde{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^*)|_2$. Indeed, we can compute that

$$\begin{aligned}
\left|\nabla\widetilde{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^*)\right|_2 =& \left|\nabla\mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)}) + \frac{1}{m}\sum_{j=1}^m \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(0)})\right|_2 \qquad (31)\\
\leq& \Big|\underbrace{\nabla\mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)}) - \mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})}_{T_1}\Big|_2\\
&+ \Big|\underbrace{\frac{1}{m}\sum_{j=1}^m\left[\nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(0)}) - \nabla\mathcal{L}_j(\boldsymbol{\beta}^*) - \mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*)\right]}_{T_2}\Big|_2 + \Big|\underbrace{\frac{1}{m}\sum_{j=1}^m\nabla\mathcal{L}_j(\boldsymbol{\beta}^*)}_{T_3}\Big|_2.
\end{aligned}$$

For the term T_1, there is

$$\begin{aligned}
T_1 =& \frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\left\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*) - \psi'(\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(0)})\right\}\boldsymbol{X}_i - \mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\\
=& \frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\left[\psi''(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - \mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}\right](\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\\
&+ \frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\int_0^1 \psi'''\big(\boldsymbol{X}_i^{\mathrm{T}}\{\widehat{\boldsymbol{\beta}}^{(0)} + t(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\}\big)\boldsymbol{X}_i\big\{\boldsymbol{X}_i(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\big\}^2\mathrm{d}t.
\end{aligned}$$

Following the same strategy as in the proof of Lemma 10, we have that

$$\begin{aligned}
&\left|\frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\left[\psi''(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - \mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}\right](\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\right|_2 = O_{\mathbb{P}}\left(r_n\sqrt{\frac{p\log n^*}{n^*}}\right),\\
&\left|\frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\int_0^1 \psi'''\big(\boldsymbol{X}_i^{\mathrm{T}}\{\widehat{\boldsymbol{\beta}}^{(0)} + t(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\}\big)\boldsymbol{X}_i\big\{\boldsymbol{X}_i(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\big\}^2\mathrm{d}t\right|_2 = O_{\mathbb{P}}\left(\sqrt{p}r_n^2\right).
\end{aligned}$$

Therefore we have

$$|T_1|_2 = O_{\mathbb{P}}\left(r_n\sqrt{\frac{p\log n^*}{n^*}} + \sqrt{p}r_n^2\right). \qquad (32)$$

Similarly, we can show that

$$|T_2|_2 = O_{\mathbb{P}}\left(r_n\sqrt{\frac{p\log n^*}{mn}} + \sqrt{p}r_n^2\right). \qquad (33)$$

T_3 is the average of mn copies of $-Y\boldsymbol{X} - \psi(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}$, thus we can simply apply Lemma 1 of Cai and Liu (2011) and yield

$$|T_3|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}}\right). \tag{34}$$

Plugging (32), (33) and (34) into (31) we have

$$\left|\nabla\widetilde{\mathcal{L}}^{(1)}(\boldsymbol{\beta}^*)\right|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}} + r_n\sqrt{\frac{p\log n^*}{n^*}} + \sqrt{p}r_n^2\right).$$

Therefore, by Lemma 9 we obtain

$$|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}} + r_n\sqrt{\frac{p\log n^*}{n^*}} + \sqrt{p}r_n^2\right).$$

Apply the above formula iteratively we can obtain Theorem 1. ■

Proof [Proof of Theorem 2] Since $|\widehat{\boldsymbol{\beta}}^{(t-1)} - \boldsymbol{\beta}^*|_2 = O_{\mathbb{P}}(\sqrt{p\log n^*/(mn)})$, follow the same strategy as in the proof of Theorem 1, we can show that

$$\begin{aligned}
&|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}}\right),\\
&\left|\nabla\mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(t)}) - \mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(t)})\right|_2\\
=&O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}\frac{p\log n^*}{n^*}} + \sqrt{p}\Big(\sqrt{\frac{p\log n^*}{mn}}\Big)^2\right) = O_{\mathbb{P}}\left(\frac{p\log n^*}{\sqrt{mnn^*}} + \frac{p^{3/2}\log n^*}{mn}\right).
\end{aligned}$$

By optimality condition of $\widehat{\boldsymbol{\beta}}^{(t)}$ we have that

$$\begin{aligned}
&0 = \nabla\widetilde{\mathcal{L}}^{(t)}(\widehat{\boldsymbol{\beta}}^{(t)})\\
=&\nabla\mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(t)}) - \nabla\mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}^*) + \nabla\mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(t-1)}) + \frac{1}{m}\sum_{j=1}^m \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(t-1)})\\
=&\nabla\mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(t)}) - \nabla\mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}^*) + \frac{1}{m}\sum_{j=1}^m \nabla\mathcal{L}_j(\boldsymbol{\beta}^*) + O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}\frac{p\log n^*}{n^*}} + \sqrt{p}\Big(\sqrt{\frac{p\log n^*}{mn}}\Big)^2\right)\\
=&\mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}(\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*) + \frac{1}{m}\sum_{j=1}^m \nabla\mathcal{L}_j(\boldsymbol{\beta}^*) + O_{\mathbb{P}}\left(\frac{p\log n^*}{\sqrt{mnn^*}} + \frac{p^{3/2}\log n^*}{mn}\right),\\
\Rightarrow&\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^* = \left[\mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right\}\right]^{-1}\frac{1}{m}\sum_{j=1}^m \nabla\mathcal{L}_j(\boldsymbol{\beta}^*) + O_{\mathbb{P}}\left(\frac{p\log n^*}{\sqrt{mnn^*}} + \frac{p^{3/2}\log n^*}{mn}\right).
\end{aligned}$$

For each $\widetilde{\boldsymbol{v}} \in \mathbb{S}^{p-1}$, we know

$$\begin{aligned}&\widetilde{\boldsymbol{v}}^{\mathrm{T}} \frac{1}{m} \sum_{j=1}^{m} \nabla \mathcal{L}_j(\boldsymbol{\beta}^*) \\ =& \frac{1}{mn} \sum_{j=1}^{m} \sum_{i \in \mathcal{D}_j} (\psi'(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}^*) - Y_i) \widetilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{X}_i,\end{aligned}$$

which is the average of mn i.i.d. terms. Compute that

$$\mathrm{Var}\Big[(\psi'(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}^*) - Y_i) \widetilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{X}_i\Big] = c(\sigma) \mathbb{E}\left[\psi''(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}^*) (\widetilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{X})^2\right].$$

Therefore, by the central limit theorem, there is

$$\frac{\sqrt{mn}}{\sigma(\widetilde{\boldsymbol{v}})} \widetilde{\boldsymbol{v}}^{\mathrm{T}} \frac{1}{m} \sum_{j=1}^{m} \nabla \mathcal{L}_j(\boldsymbol{\beta}^*) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\begin{aligned}\{\sigma(\widetilde{\boldsymbol{v}})\}^2 &= \widetilde{\boldsymbol{v}}^{\mathrm{T}} \boldsymbol{\mathcal{C}} \widetilde{\boldsymbol{v}}, \\ \text{and } \boldsymbol{\mathcal{C}} &= c(\sigma) \mathbb{E}\left[\psi''(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}}\right].\end{aligned}$$

Next we assume $(p \log n^*)^2 = o(n^*)$ and $p^3 \log^2 n^* = o(mn)$, and replace $\widetilde{\boldsymbol{v}}$ by $\left[\mathbb{E}\left\{\psi''(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}}\right\}\right]^{-1} \boldsymbol{v}$, we can obtain the asymptotic normality result for $\boldsymbol{v}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^*)$. ■

Proof [Proof of Theorem 4] We first consider the convergence rate of $\widehat{\boldsymbol{\beta}}^{(1)}_{\mathrm{Avg}}$. For each $\widehat{\boldsymbol{\beta}}^{(1)}_j = \mathrm{argmin}\, \widetilde{\mathcal{L}}^{(1)}_j(\boldsymbol{\beta})$ (where $j \in \mathcal{U}$), by Theorem 1 we know that

$$\big|\widehat{\boldsymbol{\beta}}^{(1)}_j - \boldsymbol{\beta}^*\big|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p \log n^*}{mn}} + r_n \sqrt{\frac{p \log n^*}{n^*}} + \sqrt{p} r_n^2\right).$$

By the optimality condition, there is

$$\begin{aligned}0 =& \nabla \widetilde{\mathcal{L}}^{(1)}_j(\widehat{\boldsymbol{\beta}}^{(1)}_j) \\ =& \nabla \mathcal{L}^{\mathrm{ss}}_j(\widehat{\boldsymbol{\beta}}^{(1)}_j) - \nabla \mathcal{L}^{\mathrm{ss}}_j(\widehat{\boldsymbol{\beta}}^{(0)}) + \frac{1}{m} \sum_{k=1}^{m} \nabla \mathcal{L}_k(\widehat{\boldsymbol{\beta}}^{(0)}) \\ =& \mathbb{E}\Big\{\psi''(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}}\Big\}(\widehat{\boldsymbol{\beta}}^{(1)}_j - \boldsymbol{\beta}^*) + \nabla \mathcal{L}^{\mathrm{ss}}_j(\widehat{\boldsymbol{\beta}}^{(1)}_j) - \nabla \mathcal{L}^{\mathrm{ss}}_j(\boldsymbol{\beta}^*) - \mathbb{E}\Big\{\psi''(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}}\Big\}(\widehat{\boldsymbol{\beta}}^{(1)}_j - \boldsymbol{\beta}^*) \\ & + \nabla \mathcal{L}^{\mathrm{ss}}_j(\boldsymbol{\beta}^*) - \nabla \mathcal{L}^{\mathrm{ss}}_j(\widehat{\boldsymbol{\beta}}^{(0)}) + \frac{1}{m} \sum_{k=1}^{m} \nabla \mathcal{L}_k(\widehat{\boldsymbol{\beta}}^{(0)}) \\ \Rightarrow& \widehat{\boldsymbol{\beta}}^{(1)}_j - \boldsymbol{\beta}^* = -\Big[\mathbb{E}\Big\{\psi''(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}}\Big\}\Big]^{-1} \Bigg[\nabla \mathcal{L}^{\mathrm{ss}}_j(\widehat{\boldsymbol{\beta}}^{(1)}_j) - \nabla \mathcal{L}^{\mathrm{ss}}_j(\boldsymbol{\beta}^*) - \mathbb{E}\Big\{\psi''(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}^*) \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}}\Big\}(\widehat{\boldsymbol{\beta}}^{(1)}_j - \boldsymbol{\beta}^*) \\ & + \nabla \mathcal{L}^{\mathrm{ss}}_j(\boldsymbol{\beta}^*) - \nabla \mathcal{L}^{\mathrm{ss}}_j(\widehat{\boldsymbol{\beta}}^{(0)}) + \frac{1}{m} \sum_{k=1}^{m} \nabla \mathcal{L}_k(\widehat{\boldsymbol{\beta}}^{(0)})\Bigg].\end{aligned}$$

Take the average over $j \in \mathcal{U}$ and take the norm on both sides, we have that

$$
\begin{aligned}
\Big|\widehat{\boldsymbol{\beta}}_{\mathrm{Avg}}^{(1)} - \boldsymbol{\beta}^*\Big|_2 =& \left|\frac{1}{|\mathcal{U}|}\sum_{j\in\mathcal{U}}\widehat{\boldsymbol{\beta}}_j^{(1)} - \boldsymbol{\beta}^*\right|_2 \\
\le& \rho^{-1}\Bigg[\Big|\frac{1}{|\mathcal{U}|}\sum_{j\in\mathcal{U}}\Big\{\nabla\mathcal{L}_j^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}_j^{(1)}) - \nabla\mathcal{L}_j^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \mathbb{E}\Big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\Big\}(\widehat{\boldsymbol{\beta}}_j^{(1)} - \boldsymbol{\beta}^*)\Big\}\Big|_2 \\
&+ \Big|\frac{1}{|\mathcal{U}|}\sum_{j\in\mathcal{U}}\Big\{\nabla\mathcal{L}_j^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_j^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)}) + \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(0)})\Big\}\Big|_2\Bigg] \\
\le& \rho^{-1}\Bigg[\underbrace{\max_{j\in\mathcal{U}}\Big|\nabla\mathcal{L}_j^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}_j^{(1)}) - \nabla\mathcal{L}_j^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \mathbb{E}\Big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\Big\}(\widehat{\boldsymbol{\beta}}_j^{(1)} - \boldsymbol{\beta}^*)\Big|_2}_{T_1} \\
&+ \underbrace{\Big|\frac{1}{|\mathcal{U}|}\sum_{j\in\mathcal{U}}\Big\{\nabla\mathcal{L}_j^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_j^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)}) + \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(0)})\Big|_2}_{\boldsymbol{T}_2}\Bigg]. \qquad (35)
\end{aligned}
$$

Following the strategy of the proof of Theorem 1, we can prove that

$$
\begin{aligned}
T_1 =& O_{\mathbb{P}}\left(\max_{1\le j\le m}|\widehat{\boldsymbol{\beta}}_j^{(1)} - \boldsymbol{\beta}^*|_2\sqrt{\frac{p\log n^*}{n^*}} + \sqrt{p}\big(\max_{1\le j\le m}|\widehat{\boldsymbol{\beta}}_j^{(1)} - \boldsymbol{\beta}^*|_2\big)^2\right) \\
=& O_{\mathbb{P}}\left(\frac{p\log n^*}{\sqrt{mnn^*}} + r_n\frac{p\log n^*}{n^*} + p^{3/2}r_n^4\right), \\
|\boldsymbol{T}_2|_2 =& O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}} + r_n\sqrt{\frac{p\log n^*}{|\mathcal{U}|n^*}} + \sqrt{p}r_n^2\right).
\end{aligned}
$$

Plug them into (35) we have that

$$
\Big|\widehat{\boldsymbol{\beta}}_{\mathrm{Avg}}^{(1)} - \boldsymbol{\beta}^*\Big|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{mn}} + r_n\frac{p\log n^*}{n^*} + r_n\sqrt{\frac{p\log n^*}{|\mathcal{U}|n^*}} + \sqrt{p}r_n^2\right).
$$

Then Theorem 4 can be proved by applying Theorem 4 inductively. ■

Appendix B. Proof of Theories in Section 3.2

Proof [Proof of Proposition 6] Notice that in the linear regression model, the canonical link function $\psi'(x) = x$ satisfies $\psi'''(x) = 0$. Therefore, in the proof of Theorem 1, we can obtain that

$$
\begin{aligned}
&\Big|\nabla\mathcal{L}_1^{\mathrm{ss}}(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_1^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{(0)}) - \mathbb{E}\big[\psi''\big(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*\big)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big](\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\Big|_2 = O_{\mathbb{P}}\Big(r_n\sqrt{\frac{p\log n^*}{n^*}}\Big) \\
&\Big|\frac{1}{m}\sum_{j=1}^m\nabla\mathcal{L}_j(\boldsymbol{\beta}^*) - \frac{1}{m}\sum_{j=1}^m\nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}^{(0)}) - \mathbb{E}\big[\psi''\big(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*\big)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big](\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}^{(0)})\Big|_2 = O_{\mathbb{P}}\Big(r_n\sqrt{\frac{p\log n^*}{mn}}\Big),
\end{aligned}
$$

where the term $\sqrt{p}r_n^2$ is eliminated. Therefore, the corollary is proved by plugging the above bound into (31) of Theorem 1. ■

Proof [Proof of Proposition 7] For linear model, given the initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$, the one-step SSDANE can be written explicitly as follows

$$\widehat{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}^{(0)} - (\widehat{\boldsymbol{\Sigma}}_1^{\mathrm{ss}})^{-1}\Big\{\frac{1}{mn}\sum_{j=1}^{m}\sum_{i\in\mathcal{D}_j}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(0)} - \frac{1}{mn}\sum_{j=1}^{m}\sum_{i\in\mathcal{D}_j}Y_i\boldsymbol{X}_i\Big\},$$

where $\widehat{\boldsymbol{\Sigma}}_1^{\mathrm{ss}} = (n^*)^{-1}\sum_{i\in\mathcal{H}_1}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}$. Then the statistical error $\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*$ can be written down explicitly as follows

$$\begin{aligned}
&\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^* \\
=&(\widehat{\boldsymbol{\Sigma}}_1^{\mathrm{ss}})^{-1}\Big\{\frac{1}{n^*}\sum_{i\in\mathcal{H}_0}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*) - \frac{1}{mn}\sum_{j=1}^{m}\sum_{i\in\mathcal{D}_j}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(0)} + \frac{1}{mn}\sum_{j=1}^{m}\sum_{i\in\mathcal{D}_j}Y_i\boldsymbol{X}_i\Big\} \\
=&(\widehat{\boldsymbol{\Sigma}}_1^{\mathrm{ss}})^{-1}\Big[\Big(\frac{1}{n^*}\sum_{i\in\mathcal{H}_1}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - \frac{1}{mn}\sum_{j=1}^{m}\sum_{i\in\mathcal{D}_j}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\Big)(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*) + \frac{1}{mn}\sum_{j=1}^{m}\sum_{i\in\mathcal{D}_j}\epsilon_i\boldsymbol{X}_i\Big] \\
=&(\widehat{\boldsymbol{\Sigma}}_1^{\mathrm{ss}})^{-1}\Bigg[\sum_{i\in\mathcal{D}_1^*}\frac{1}{n^*}\Big(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - \boldsymbol{\Sigma}\Big)(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*) + \sum_{i\in\mathcal{D}_1}\Big\{\frac{mn-n^*}{mnn^*}\Big(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - \boldsymbol{\Sigma}\Big)(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*) + \frac{1}{mn}\epsilon_i\boldsymbol{X}_i\Big\} \\
&-\sum_{j=2}^{m}\sum_{i\in\mathcal{D}_j}\frac{1}{mn}\Big\{\Big(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - \boldsymbol{\Sigma}\Big)(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*) + \epsilon_i\boldsymbol{X}_i\Big\}\Bigg] \\
\triangleq&(\widehat{\boldsymbol{\Sigma}}_1^{\mathrm{ss}})^{-1}\boldsymbol{T}. \qquad (36)
\end{aligned}$$

Since we already assume that $\widehat{\boldsymbol{\beta}}^{(0)}$ is independent to all $\boldsymbol{X}_i$'s and ϵ_i's, there are sum of $m(n-1)+n^*$ independent, zero-mean terms in the square bracket. For notational convenience we denote $\boldsymbol{Z}_i = \Big(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} - \boldsymbol{\Sigma}\Big)(\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*)/|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*|_2$ and $\boldsymbol{W}_i = \epsilon_i\boldsymbol{X}_i$, then we have that

$$\mathrm{C}ov(\boldsymbol{Z}_i) = \mathrm{C}ov\left\{\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\frac{\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*}{|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*|_2}\right\}, \quad \mathrm{C}ov(\boldsymbol{W}_i) = \sigma^2\boldsymbol{\Sigma}, \quad \mathrm{C}ov(\boldsymbol{Z}_i, \boldsymbol{W}_i) = 0.$$

Thus we can compute that

$$\begin{aligned}
&\mathrm{C}ov(\boldsymbol{T}) \\
=&\frac{n^*-n}{(n^*)^2}r_n^2\mathrm{C}ov(\boldsymbol{Z}) + \frac{n(mn-n^*)^2}{(mnn^*)^2}r_n^2\mathrm{C}ov(\boldsymbol{Z}) + \frac{n}{(mn)^2}\mathrm{C}ov(\boldsymbol{W}) + \frac{n(m-1)}{(mn)^2}r_n^2\mathrm{C}ov(\boldsymbol{Z}) + \frac{n(m-1)}{(mn)^2}\mathrm{C}ov(\boldsymbol{W}) \\
=&\left(\frac{1}{n^*} + \frac{n^*-2n}{mn^*n}\right)r_n^2\mathrm{C}ov(\boldsymbol{Z}) + \frac{1}{mn}\mathrm{C}ov(\boldsymbol{W}). \qquad (37)
\end{aligned}$$

Moreover, there is

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i\in\mathcal{D}_1^*}\left|\frac{1}{n^*}r_n\boldsymbol{Z}_i\right|_2^3+\sum_{i\in\mathcal{D}_1}\left|\frac{mn-n^*}{mnn^*}r_n\boldsymbol{Z}_i+\frac{1}{mn}\boldsymbol{W}_i\right|_2^3+\sum_{j=2}^{m}\sum_{i\in\mathcal{D}_j}\left|\frac{1}{mn}r_n\boldsymbol{Z}_i+\frac{1}{mn}\boldsymbol{W}_i\right|_2^3\right]\\
\leq&\mathbb{E}\left[\sum_{i\in\mathcal{D}_1^*}\left|\frac{1}{n^*}r_n\boldsymbol{Z}_i\right|_2^3+\sum_{i\in\mathcal{D}_1}4\left(\left|\frac{mn-n^*}{mnn^*}r_n\boldsymbol{Z}_i\right|_2^3+\left|\frac{1}{mn}\boldsymbol{W}_i\right|_2^3\right)+\sum_{j=2}^{m}\sum_{i\in\mathcal{D}_j}4\left(\left|\frac{1}{mn}r_n\boldsymbol{Z}_i\right|_2^3+\left|\frac{1}{mn}\boldsymbol{W}_i\right|_2^3\right)\right]\\
\leq&\left(\frac{16}{(n^*)^2}+\frac{16}{(mn)^2}\right)r_n^3\mathbb{E}\left[|\boldsymbol{Z}|_2^3\right]+\frac{4}{(mn)^2}\mathbb{E}\left[|\boldsymbol{W}|_2^3\right]\leq C_0p^{3/2}\left(\frac{r_n^3}{(n^*)^2}+\frac{1}{(mn)^2}\right),
\end{aligned}
$$

for some constant $C_0 > 0$. Denote $\widetilde{\boldsymbol{T}} = (\mathrm{C}ov(\boldsymbol{T}))^{-1/2}\boldsymbol{T}$, then we can apply multivariate Berry-Esseen theorem (see Theorem 1.1 of Raič (2019)) and yield

$$
\left|\mathbb{P}(\widetilde{\boldsymbol{T}}\in S)-\mathcal{N}(0,\boldsymbol{I}_p)\{S\}\right|\leq C_1p^{7/4}\left\{\min\left(\frac{1}{\sqrt{n^*}},\frac{(r_n\sqrt{mn})^3}{(n^*)^2}\right)+\frac{1}{\sqrt{mn}}\right\}=o(1),
$$

for all measurable convex $S\subseteq\mathbb{R}^p$. Let $\overline{\boldsymbol{T}} = (\overline{T}_1,...,\overline{T}_p)^{\mathrm{T}}$ follows p-dimensional standard Gaussian distribution $\mathcal{N}(0,\boldsymbol{I}_p)$, then we can apply Lemma 1 of Cai and Liu (2011) to the i.i.d. sequence $\overline{T}_i^2-1$ and obtain that

$$
\begin{aligned}
&\mathbb{P}\left(|\overline{\boldsymbol{T}}|_2\geq\sqrt{\frac{p}{2}}\right)=\mathbb{P}\left(\sum_{l=1}^{p}(\overline{T}_l^2-1)\geq-\frac{p}{2}\right)\\
\geq&1-\mathbb{P}\left(\left|\sum_{l=1}^{p}(\overline{T}_l^2-1)\right|\geq\frac{p}{2}\right)\geq\frac{3}{4},
\end{aligned}
$$

for sufficiently large p. Then there is

$$
\mathbb{P}(|\widetilde{\boldsymbol{T}}|\geq\sqrt{p/2})\geq\mathbb{P}(|\overline{\boldsymbol{T}}|\geq\sqrt{p/2})-\left|\mathbb{P}(|\widetilde{\boldsymbol{T}}|\geq\sqrt{p/2})-\mathbb{P}(|\overline{\boldsymbol{T}}|\geq\sqrt{p/2})\right|\geq\frac{5}{8}.
$$

By assumption $\lambda_{\min}(\mathrm{C}ov(\boldsymbol{Z})),\lambda_{\min}(\mathrm{C}ov(\boldsymbol{W}))\geq\rho_1$, then from (37) we know

$$
\lambda_{\min}\{\mathrm{C}ov(\boldsymbol{T})\}\geq C_3\rho_1\left(\frac{r_n^2}{n^*}+\frac{1}{mn}\right).
$$

Then there is

$$
\mathbb{P}\left(|\boldsymbol{T}_2|_2\geq C_3\rho_1\left(\frac{r_n}{\sqrt{n^*}}+\frac{1}{\sqrt{mn}}\right)\sqrt{\frac{p}{2}}\right)\geq\mathbb{P}\left(|\boldsymbol{T}_2|_2\geq\sqrt{\frac{p}{2}}\right)\geq\frac{5}{8}. \tag{38}
$$

Following the strategy, in the proof of Lemma 10 we can show that

$$
\mathbb{P}\left(\lambda_{\min}\{(\widehat{\boldsymbol{\Sigma}}^{\mathrm{ss}})^{-1}\}\leq\frac{\rho}{2}\right)=\mathbb{P}\left(\lambda_{\max}\{\widehat{\boldsymbol{\Sigma}}^{\mathrm{ss}}\}\geq2\rho^{-1}\right)\leq\frac{1}{8}.
$$

Therefore by combining (38) and (36), we know that

$$
\begin{aligned}
&\mathbb{P}\left(|\widehat{\boldsymbol{\beta}}^{(1)}-\boldsymbol{\beta}^*|_2 \geq \frac{\rho}{2} C_3 \rho_1\left(\frac{r_n}{\sqrt{n^*}}+\frac{1}{\sqrt{mn}}\right) \sqrt{\frac{p}{2}}\right) \\
\geq &\mathbb{P}\left(|\boldsymbol{T}_2|_2 \geq C_3 \rho_1\left(\frac{r_n}{\sqrt{n^*}}+\frac{1}{\sqrt{mn}}\right) \sqrt{\frac{p}{2}}\ ;\ \lambda_{\min}\big\{(\widehat{\boldsymbol{\Sigma}}^{\mathrm{ss}})^{-1}\big\} \geq \frac{\rho}{2}\right) \\
\geq &\mathbb{P}\left(|\boldsymbol{T}_2|_2 \geq C_3 \rho_1\left(\frac{r_n}{\sqrt{n^*}}+\frac{1}{\sqrt{mn}}\right) \sqrt{\frac{p}{2}}\right)-\mathbb{P}\left(\lambda_{\min}\big\{(\widehat{\boldsymbol{\Sigma}}^{\mathrm{ss}})^{-1}\big\} \leq \frac{\rho}{2}\right) \\
\geq &\frac{5}{8}-\frac{1}{8}=\frac{1}{2},
\end{aligned}
$$

which proves the theorem.

■

Appendix C. Proof of Theories in Section 3.3

Proof [Proof of Proposition 8] We first prove the convergence rate of $\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}$. Notice that the semi-supervised empirical loss $\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta})$ defined in (5) has the same Hessian matrix as the surrogate loss $\widetilde{\mathcal{L}}^{(1)}(\boldsymbol{\beta})$ defined in (10). Therefore by Lemma 10, it holds that

$$
\inf_{\boldsymbol{\beta}:|\boldsymbol{\beta}-\boldsymbol{\beta}^*|_2\leq r}\ \inf_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \boldsymbol{v}^{\mathrm{T}}\nabla^2\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta})\boldsymbol{v} \geq \frac{\rho}{2},
$$

Then by Lemma 9, we only need to bound the term $|\nabla\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}^*)|_2$. Note that for arbitrary unit vector $\boldsymbol{v}\in\mathbb{S}^{p-1}$, there is

$$
\begin{aligned}
&\boldsymbol{v}\nabla\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}^*) \\
=&-\frac{1}{n}\sum_{i\in\mathcal{D}} Y_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}} \\
=&\frac{1}{n}\sum_{i\in\mathcal{D}}\left[-Y_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i+\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right]+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\left[\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}-\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right],
\end{aligned}
$$

Denote $\mathfrak{E}(X,\eta)=\mathbb{E}[X^2\exp(\eta|X|)]$, we compute that

$$
\begin{aligned}
&\mathfrak{E}\left[-Y_{\boldsymbol{i}}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i+\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\},\eta_0\right] \\
\leq&\frac{1}{\eta_0^2}\mathbb{E}\left[\exp\left(2\eta_0\big|-Y_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i+\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\big|\right)\right] \\
\leq&\frac{1}{\eta_0^2}\left\{\mathbb{E}\left[\exp\left(2\eta_0|Y_{\boldsymbol{i}}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i|\right)\right]\right\}^2 \\
\leq&\frac{1}{\eta_0^2}\left\{\mathbb{E}\left[\exp\left(\eta_0|Y_i|^2\right)\exp\left(\eta_0|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i|^2\right)\right]\right\}^2 \leq \frac{1}{\eta_0^2}C_0^4,
\end{aligned}
$$

where the second line uses the elementary inequality $x^2\leq e^x$, the third line uses Jensen's inequality, and the last line uses Cauchy's inequality. Similarly, we have that

$$
\mathfrak{E}\left[\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}-\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\},\eta_0\right] \leq \frac{1}{\eta_0^2}C_0^4.
$$

Then we can apply Lemma 1 of Cai and Liu (2011) and yield that

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{D}}\left[-Y_i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_i+\mathbb{E}\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\}\right]=O_{\mathbb{P}}\left(\sqrt{\frac{\log n^*}{n}}\right),\\
&\frac{1}{n^*}\sum_{i\in\mathcal{H}}\left[\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}-\mathbb{E}\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\}\right]=O_{\mathbb{P}}\left(\sqrt{\frac{\log n^*}{n^*}}\right).
\end{aligned}$$

Therefore, we know that there exists a constant c_1 such that

$$\begin{aligned}
|\nabla\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}^*)|_2=&\sqrt{\sum_{l=1}^{p}\left|\boldsymbol{e}_l^{\mathrm{T}}\nabla\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}^*)\right|^2}\\
\le& c_1\left(\sqrt{\frac{p\log n^*}{n}}+\sqrt{\frac{p\log n^*}{n^*}}\right),
\end{aligned}$$

with probability not less than $1-(n^*)^{-\tau}$, where $\tau>0$ is some positive constant and $\boldsymbol{e}_l$ (where $l=1,...,p$) denotes the l-th coordinate vectors. Therefore, by Lemma 9, we know that

$$\left|\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*\right|_2=O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{n}}+\sqrt{\frac{p\log n^*}{n^*}}\right). \tag{39}$$

On the other hand, by definition of $\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}$ in (22) we have that

$$\begin{aligned}
0=&\nabla\mathcal{L}^{\mathrm{ss}}(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}})=-\frac{1}{n}\sum_{i\in\mathcal{D}}Y_i\boldsymbol{X}_i^{\mathrm{T}}+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{\mathrm{ss}})\boldsymbol{X}_i\\
=&-\frac{1}{n}\sum_{i\in\mathcal{D}}Y_i\boldsymbol{X}_i^{\mathrm{T}}+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}_i+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{\mathrm{ss}})-\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\}\boldsymbol{X}_i\\
=&-\frac{1}{n}\sum_{i\in\mathcal{D}}Y_i\boldsymbol{X}_i^{\mathrm{T}}+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}_i\\
&+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\int_0^1\psi''\big(\boldsymbol{X}_i^{\mathrm{T}}\{\boldsymbol{\beta}^*+t(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)\}\big)\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)\mathrm{d}t\\
\Rightarrow&\frac{1}{n}\sum_{i\in\mathcal{D}}Y_i\boldsymbol{X}_i^{\mathrm{T}}-\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}_i \qquad (40)\\
=&\frac{1}{n^*}\sum_{i\in\mathcal{H}}\int_0^1(1-t)\psi'''\big(\boldsymbol{X}_i^{\mathrm{T}}\{\boldsymbol{\beta}^*+t(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)\}\big)\boldsymbol{X}_i\big\{\boldsymbol{X}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)\big\}^2\mathrm{d}t\\
&+\Big[\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi''(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}-\mathbb{E}\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}\Big](\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)+\mathbb{E}\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*).
\end{aligned}$$

Using the results in the proof of Lemma 10, we know that

$$
\begin{aligned}
&\left|\frac{1}{n^*}\sum_{i\in\mathcal{H}}\int_0^1(1-t)\psi'''\big(\boldsymbol{X}_i^{\mathrm{T}}\{\boldsymbol{\beta}^*+t(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)\}\big)\boldsymbol{X}_i\big\{\boldsymbol{X}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)\big\}^2\mathrm{d}t\right|_2\\
=&O_{\mathbb{P}}\left(\sqrt{p}|\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*|_2^2\right)=O_{\mathbb{P}}\left(\frac{p^{3/2}\log n^*}{n}\right),\\
&\left|\Big[\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi''(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}-\mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}\Big](\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)\right|_2\\
=&O_{\mathbb{P}}\left(\sqrt{\frac{p\log n^*}{n^*}}|\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*|_2\right)=O_{\mathbb{P}}\left(\frac{p\log n^*}{\sqrt{nn^*}}\right).
\end{aligned}
$$

Substitute these bounds into (40), we have that

$$
\begin{aligned}
&\mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)=\frac{1}{n}\sum_{i\in\mathcal{D}}Y_i\boldsymbol{X}_i^{\mathrm{T}}-\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}_i+O_{\mathbb{P}}\left(\frac{p^{3/2}\log n^*}{n}\right)\\
\Rightarrow&\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*=\big[\mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}\big]^{-1}\nabla\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}^*)+O_{\mathbb{P}}\left(\frac{p^{3/2}\log n^*}{n}\right).
\end{aligned}
\tag{41}
$$

For each $\widetilde{\boldsymbol{v}}\in\mathbb{S}^{p-1}$, we know

$$
\begin{aligned}
\widetilde{\boldsymbol{v}}^{\mathrm{T}}\nabla\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}^*)=&-\frac{1}{n}\sum_{i\in\mathcal{D}}Y_i\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i+\frac{1}{n^*}\sum_{i\in\mathcal{H}}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\\
=&\sum_{i\in\mathcal{D}}\left[-\frac{1}{n}Y_i\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i+\frac{1}{n^*}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}+\frac{n^*-n}{nn^*}\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right]\\
&+\sum_{i\in\mathcal{D}^*}\left[\frac{1}{n^*}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}-\frac{1}{n^*}\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right],
\end{aligned}
$$

which is the sum of n^* independent, zero-mean random variables. Compute that

$$
\begin{aligned}
&\mathrm{Var}\left[-\frac{1}{n}Y_i\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i+\frac{1}{n^*}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}+\frac{n^*-n}{nn^*}\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right]\\
=&\frac{1}{n^2}\mathrm{Var}\left[Y_i\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i-\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\right]+\frac{(n^*-n)^2}{(nn^*)^2}\mathrm{Var}\left[\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}-\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right]\\
=&\frac{c(\sigma)}{n^2}\mathbb{E}\left[\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)(\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X})^2\right]+\frac{(n^*-n)^2}{(nn^*)^2}\mathrm{Var}\left[\psi'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}\right],\\
&\mathrm{Var}\left[\frac{1}{n^*}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}-\frac{1}{n^*}\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right]\\
=&\frac{1}{(n^*)^2}\mathrm{Var}\left[\psi'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}\right].
\end{aligned}
$$

Since from Assumption 2 we know both $\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}$ and $\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)$ are sub-Gaussian random variables, we know that the Lindeberg condition for central limit theorem is sufficed. Therefore, by central limit theorem (see, e.g., Theorem 9.1.1 of Chow and Teicher (2012)), we have that

$$\frac{\sqrt{n}}{\sigma(\widetilde{\boldsymbol{v}})}\widetilde{\boldsymbol{v}}^{\mathrm{T}}\nabla\mathcal{L}^{\mathrm{ss}}(\boldsymbol{\beta}^*)\xrightarrow{d}\mathcal{N}(0,1), \tag{42}$$

where

$$\begin{aligned}\big\{\sigma(\widetilde{\boldsymbol{v}})\big\}^2 =& n^2\mathrm{Var}\left[-\frac{1}{n}Y_i\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_i+\frac{1}{n^*}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}+\frac{n^*-n}{nn^*}\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right]\\ &+(n^*-n)n\mathrm{Var}\left[\frac{1}{n^*}\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}-\frac{1}{n^*}\mathbb{E}\big\{\psi'(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{i}}\big\}\right]\\ =&c(\sigma)\mathbb{E}\left[\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)(\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X})^2\right]+\frac{n^*-n}{n^*}\mathrm{Var}\left[\psi'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\widetilde{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{X}\right].\end{aligned}$$

We can also write it in the following way

$$\begin{aligned}\big\{\sigma(\widetilde{\boldsymbol{v}})\big\}^2 =&\widetilde{\boldsymbol{v}}^{\mathrm{T}}\Big(\boldsymbol{\mathcal{C}}+\frac{n^*-n}{n^*}\boldsymbol{\mathcal{C}}^{\mathrm{ss}}\Big)\widetilde{\boldsymbol{v}}\\ \text{where}\quad\boldsymbol{\mathcal{C}} =&c(\sigma)\mathbb{E}\left[\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right],\\ \boldsymbol{\mathcal{C}}^{\mathrm{ss}} =&\mathrm{C}ov\big\{\psi'(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\big\}.\end{aligned}$$

Next we assume $p^{3/2}\log n^* = o(n)$, and replace $\widetilde{\boldsymbol{v}}$ by $\big[\mathbb{E}\big\{\psi''(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}\big]^{-1}\boldsymbol{v}$, we can obtain the asymptotic normality result for $\boldsymbol{v}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{ss}}-\boldsymbol{\beta}^*)$. ■

References

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6(61):1817–1853, 2005.

Rie Kubota Ando and Tong Zhang. Two-view feature generation model for semi-supervised learning. In *Proceedings of the 24th International Conference on Machine Learning*, page 25–32, 2007.

David Azriel, Lawrence D. Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *J. Amer. Statist. Assoc.*, 117(540):2238–2251, 2022.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Ann. Statist.*, 46(3):1352–1382, 2018.

Pierre C. Bellec, Arnak S. Dalalyan, Edwin Grappin, and Quentin Paris. On the prediction loss of the lasso in the partially labeled setting. *Electron. J. Stat.*, 12(2):3443 – 3472, 2018.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577 – 2604, 2008.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, page 92–100, 1998.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found.Trends Mach. Learn.*, 3(1):1–122, 2011.

T. Tony Cai and Zijian Guo. Semisupervised inference for explained variance in high dimensional linear regression and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(2):391–419, 2020.

Tianxi Cai, Molei Liu, and Yin Xia. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *J. Amer. Statist. Assoc.*, 117(540):2105–2119, 2022.

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

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings. *arXiv e-prints*, page arXiv:1812.01097, 2018.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.*, 46(4):1541–1572, 2018.

Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.*, 18(1):1493–1514, 2017.

Dengsheng Chen, Jie Hu, Vince Junkai Tan, Xiaoming Wei, and Enhua Wu. Elastic aggregation for federated optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12187–12197, 2023.

Tianyi Chen, Georgios Giannakis, Tao Sun, and Wotao Yin. Lag: Lazily aggregated gradient for communication-efficient distributed learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *Ann. Statist.*, 47(6):3244 – 3273, 2019.

Y.S. Chow and H. Teicher. *Probability Theory: Independence, Interchangeability, Martingales.* Springer Texts in Statistics. Springer New York, 2012.

Siyi Deng, Yang Ning, Jiwei Zhao, and Heping Zhang. Optimal semi-supervised estimation and inference for high-dimensional linear regression. *arXiv e-prints*, art. arXiv:2011.14185, 2020.

Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *J. Mach. Learn. Res.*, 21(66):1–52, 2020.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *Ann. Statist.*, 49(2):918 – 943, 2021.

Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Trans. Parallel Distrib. Syst.*, 32(1):59–71, 2021a.

Rui Duan, Yang Ning, and Yong Chen. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*, 109(1):67–83, 2021b. ISSN 1464-3510.

Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *J. Mach. Learn. Res.*, 18(118):1–25, 2017.

Zheng-Chu Guo, Shao-Bo Lin, and Lei Shi. Distributed learning with multi-penalty regularization. *Appl. Comput. Harmon. Anal.*, 46(3):478–499, 2019.

Zijian Guo, Prabrisha Rakshit, Daniel S. Herman, and Jinbo Chen. Inference for the case probability in high-dimensional logistic regression. *J. Mach. Learn. Res.*, 22(1), 2022. ISSN 1532-4435.

Jue Hou, Zijian Guo, and Tianxi Cai. Surrogate assisted semi-supervised inference for high dimensional risk prediction. *J. Mach. Learn. Res.*, 24(265):1–58, 2023.

Ting Hu and Ding-Xuan Zhou. Distributed regularized least squares with flexible gaussian kernels. *Appl. Comput. Harmon. Anal.*, 53:349–377, 2021.

Martin Jaggi, Virginia Smith, Martin Takac, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Yongyi Guo Jianqing Fan and Kaizheng Wang. Communication-efficient accurate statistical estimation. *J. Amer. Statist. Assoc.*, 118(542):1000–1010, 2023.

Rie Johnson and Tong Zhang. Graph-based semi-supervised learning and spectral kernel design. *IEEE Trans. Inform. Theory*, 54(1):275–288, 2008.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *J. Amer. Statist. Assoc.*, 114(526):668–681, 2019.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression. *Inf. Inference*, 11(4):1389–1456, 2022. ISSN 2049-8772.

Jason D. Lee, Qiang Liu, Yuekai Sun, and Jonathan E. Taylor. Communication-efficient sparse regression. *J. Mach. Learn. Res.*, 18(1):115–144, 2017.

Runze Li, Dennis K.J. Lin, and Bing Li. Statistical inference in massive data sets. *Appl. Stoch. Models Bus. Ind.*, 29(5):399–409, 2013.

Heng Lian and Zengyan Fan. Divide-and-conquer for debiased l_1-norm support vector machine in ultra-high dimensions. *J. Mach. Learn. Res.*, 18(182):1–26, 2018.

Shao-Bo Lin and Ding-Xuan Zhou. Distributed kernel-based gradient descent algorithms. *Constr. Approx.*, 47(2):249–276, 2018.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *J. Mach. Learn. Res.*, 18(92):1–31, 2017.

Molei Liu, Yin Xia, Kelly Cho, and Tianxi Cai. Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. *J. Mach. Learn. Res.*, 22(1), 2022. ISSN 1532-4435.

Rong Ma, Zijian Guo, T. Tony Cai, and Hongzhe Li. Statistical Inference for Genetic Relatedness Based on High-Dimensional Logistic Regression. *arXiv e-prints*, page arXiv:2202.10007, 2022.

Thomas P Minka. A comparison of numerical optimizers for logistic regression. *Unpublished draft*, pages 1–18, 2003.

Ravi B. Parikh, Christopher Manz, Corey Chivers, Susan Harkness Regli, Jennifer Braun, Michael E. Draugelis, Lynn M. Schuchter, Lawrence N. Shulman, Amol S. Navathe, Mitesh S. Patel, and Nina R. O'Connor. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA Network Open*, 2(10):e1915997–e1915997, 2019. ISSN 2574-3805.

Martin Raič. A multivariate Berry-Esseen theorem with explicit constants. *Bernoulli*, 25 (4A):2824 – 2853, 2019.

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*, volume 32, pages 1000–1008, 2014.

Yue Sheng and Edgar Dobriban. One-shot distributed ridge regression in high dimensions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8763–8772, 2020.

Virginia Smith, Simone Forte, Chenxin Ma, Martin Takáč, Michael I. Jordan, and Martin Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *J. Mach. Learn. Res.*, 18(1):8590–8638, 2017.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *7th International Conference on Learning Representations, ICLR 2019, May 6-9, 2019*, 2019.

I. Theodossiou. The effects of low-pay and unemployment on psychological well-being: A logistic regression approach. *J. Health Econ.*, 17(1):85–104, 1998. ISSN 0167-6296.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 1999.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv e-prints*, art. arXiv:1011.3027, 2010.

Stanislav Volgushev, Shih-Kang Chao, and Guang Cheng. Distributed inference for quantile regression processes. *Ann. Statist.*, 47(3):1634 – 1662, 2019.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3636–3645, 2017.

Junhui Wang and Xiaotong Shen. Large margin semi-supervised learning. *J. Mach. Learn. Res.*, 8(65):1867–1891, 2007.

Junhui Wang, Xiaotong Shen, and Yufeng Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167, 2007.

Junhui Wang, Xiaotong Shen, and Wei Pan. On efficient large margin semi-supervised learning: Method and theory. *J. Mach. Learn. Res.*, 10(25):719–742, 2009.

Larry Wasserman and John Lafferty. Statistical analysis of semi-supervised regression. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5693–5700, 2019.

Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *Ann. Statist.*, 47(5):2538 – 2566, 2019.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.*, 14:3321–3363, 2013.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16(1):3299–3340, 2015.

Yuqian Zhang and Jelena Bradic. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2021.

Weihua Zhao, Fode Zhang, and Heng Lian. Debiasing and distributed estimation for high-dimensional quantile regression. *IEEE Trans. Neural Netw. Learn. Syst.*, 31(7):2569–2577, 2020.

Xiaojin Zhu and Andrew B. Goldberg. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.*, 3(1):1–130, 2009.

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. *Unpublished draft*, 2005.