Optimal Subsample Selection for Massive Logistic Regression with Distributed Data

Lulu Zuo[1], Haixiang Zhang[1*], HaiYing Wang[2] and Liuquan Sun[3]

[1] *Center for Applied Mathematics, Tianjin University, Tianjin 300072, China*

[2] *Department of Statistics, University of Connecticut, Storrs, Mansfield, CT 06269, USA*

[3] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Abstract. With the emergence of big data, it is increasingly common that the data are distributed. i.e., the data are stored at many distributed sites (machines or nodes) owing to data collection or business operations, etc. We propose a distributed subsampling procedure in such a setting to efficiently approximate the maximum likelihood estimator for the logistic regression. We establish the consistency and asymptotic normality of the subsample estimator given the full data. The optimal subsampling probabilities and optimal allocation sizes are explicitly obtained. We develop a two-step algorithm to approximate the optimal subsampling procedure. Numerical simulations and an application to airline data are presented to evaluate the performance of our subsampling method.

Keywords: Allocation size; Big data; Distributed and massive data; Subsample estimator; Subsampling probabilities.

*Corresponding author. Email: haixiang.zhang@tju.edu.cn (H. Zhang)

1 Introduction

With the development of technologies, big data or massive data have become ubiquitous in many scientific fields. Due to the incredible sizes of massive data, it is very challenging to perform standard statistical inference. A major bottleneck is that the huge dataset exceeds the available computational capability at hand. Hence, there is an urgent need for developing new statistical methods to analyze massive datasets. Recently, many efforts have been made on building both methodologies and algorithms for big data analysis. For example, Zhao et al. (2016) proposed a partially linear framework for massive heterogeneous data. Battey et al. (2018) studied the topics on hypothesis testing and parameter estimation with massive data. Shi et al. (2018) introduced a cubic-rate estimator under massive data framework. Jordan et al. (2019) presented a communication-efficient surrogate likelihood method for distributed statistical inference. Volgushev et al. (2019) proposed a distributed inference approach for quantile regression. Besides, Ma et al. (2015) proposed an algorithmic leveraging-based subsampling procedure. Wang et al. (2018) and Wang (2019) developed some optimal subsampling methods for logistic regression. Wang et al. (2019) provided a novel information-based subdata selection approach in the context of linear models. Zuo et al. (2020) introduced a subsample-based estimation method for massive survival data with additive hazards model. Ai et al. (2020) studied optimal subsampling for the big data generalized linear models, among others.

Nowadays, it is increasingly common that the data are inherently distributed. The term "inherent" means that the data are stored in a lot of distributed sites (machines or nodes) due to data collection or business operations, etc. For example, a search engine company may own data coming from a lot of locations, and each location collects huge datasets (Corbett et al. 2013). Faced with this kind of

massive data, we propose a distributed subsampling method in the context of logistic regression, which aims to select informative subsamples and construct effective subsample-based estimators. The main advantages of our method are as follows: First, we establish the convergence rate of the subsample-based estimator, which ensures the consistency of our proposed method. Second, the asymptotic normality of our subsample-based estimator is presented, which is useful for conducting statistical inference in the framework of distributed data. Third, the computational speed of our subsampling method is much faster than the full data approach.

The remainder of this article is organized as follows. In Section 2, we give some notations and assumptions. A distributed subsampling algorithm is presented. Asymptotic properties of the subsample-based estimator are established. In Section 3, we introduce a subsampling strategy with optimal subsampling probablities and optimal allocation sizes. In Section 4, a two-step subsampling procedure is proposed for practical application. In Section 5, simulations and a real data example are provided. Concluding remarks are presented in Section 6. All proof details are given in the Appendix.

2 Methods

2.1 Model and Notation

We consider the logistic regression model

$$P(Y_{ik}=1|\mathbf{X}_{ik})=\frac{\exp(\boldsymbol{\beta}^T\mathbf{X}_{ik})}{1+\exp(\boldsymbol{\beta}^T\mathbf{X}_{ik})}, i=1,\cdots,n_k, \text{ and } k=1,\cdots,K, \tag{2.1}$$

where $\mathbf{X}_{ik}\in\mathbb{R}^d$ is the covariate, $Y_{ik}\in\{0,1\}$ is the response, $\boldsymbol{\beta}\in\mathbb{R}^d$ is a vector of regression coefficients. Here n_k is the sample size of the kth dataset, $n=\sum_{k=1}^K n_k$

is the total sample size, and K is the number of distributed datasets. Denote the full data as $\mathcal{F}_n = \{(Y_{ik}, \mathbf{X}_{ik}), i = 1, \cdots, n_k; k = 1, \cdots, K\}$. We assume that these distributed data satisfy the logistic model in (2.1), i.e., we need the logistic regression model to be true, but the covariate distributions can be heterogeneous. Ideally, if a central computer with super capacity is available, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ is obtained by maximizing the log-likelihood function

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{k=1}^{K} \sum_{i=1}^{n_k} [Y_{ik} \log P_{ik}(\boldsymbol{\beta}) + (1 - Y_{ik}) \log\{1 - P_{ik}(\boldsymbol{\beta})\}],$$

where $P_{ik}(\boldsymbol{\beta}) = \frac{\exp(\mathbf{X}_{ik}^T \boldsymbol{\beta})}{1+\exp(\mathbf{X}_{ik}^T \boldsymbol{\beta})}$. Note that there is no closed-form solution for $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, and a Newton's method is adopted with the following iterative formula,

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - \Big\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} w_{ik}(\hat{\boldsymbol{\beta}}^{(t)}) \mathbf{X}_{ik} \mathbf{X}_{ik}^T \Big\}^{-1} \frac{\partial \ell(\hat{\boldsymbol{\beta}}^{(t)})}{\partial \boldsymbol{\beta}},$$

where $w_{ik}(\boldsymbol{\beta}) = P_{ik}(\boldsymbol{\beta})\{1 - P_{ik}(\boldsymbol{\beta})\}$. Our aim is to construct a subsample-based estimator, which can be used to effectively approximate the full data estimator $\hat{\boldsymbol{\beta}}_{\text{MLE}}$.

2.2 Subsampling Algorithm and Asymptotic Properties

In this section, we propose a distributed subsampling algorithm to approximate the $\hat{\boldsymbol{\beta}}_{\text{MLE}}$. Meanwhile, the consistency and asymptotic normality of the subsample estimator are established. First we propose a subsampling method in Algorithm 1, which can reasonably select a subsample from distributed data. Then, a subsample-based estimator is presented.

Algorithm 1 Distributed Subsampling Algorithm

• *Sampling*: Assign subsampling probabilities $\{\pi_{ik}\}_{i=1}^{n_k}$ for the kth dataset $\mathcal{D}_k = \{(\mathbf{X}_{ik}, Y_{ik}), i = 1, \cdots, n_k\}$ with $\sum_{i=1}^{n_k} \pi_{ik} = 1$, where $k = 1, \cdots, K$. Given r, draw a random subsample of size r_k with replacement from $\mathcal{D}_k$ according to $\{\pi_{ik}\}_{i=1}^{n_k}$, where $\{r_k\}_{k=1}^{K}$ are allocation sizes with $\sum_{k=1}^{K} r_k = r$. For $i = 1, \cdots, n_k$ and $k = 1, \cdots, K$, we denote the corresponding responses, covariates, and subsampling probabilities as Y_{ik}^*, $\mathbf{X}_{ik}^*$ and π_{ik}^*, respectively.

• *Estimation*: Based on the subsamples $\{(Y_{ik}^*, \mathbf{X}_{ik}^*, \pi_{ik}^*) : i = 1, \cdots, r_k\}_{k=1}^{K}$, we maximize the following weighted log-likelihood function to get a subsample-based estimate $\tilde{\boldsymbol{\beta}}$.

$$\ell^*(\boldsymbol{\beta}) = \sum_{k=1}^{K} \frac{1}{r_k} \sum_{i=1}^{r_k} \frac{1}{\pi_{ik}^*} [Y_{ik}^* \log P_{ik}^*(\boldsymbol{\beta}) + (1 - Y_{ik}^*) \log\{1 - P_{ik}^*(\boldsymbol{\beta})\}],$$

where $P_{ik}^*(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^T \mathbf{X}_{ik}^*)}{1+\exp(\boldsymbol{\beta}^T \mathbf{X}_{ik}^*)}$.

In order to characterize asymptotic properties of the subsample-based estimator $\tilde{\boldsymbol{\beta}}$, we need the following regularity assumptions:

(**A.1**) The parameter space $J_B \subset \mathbb{R}^d$ is a compact convex set, and $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ is in the interior of J_B.

(**A.2**) $\frac{1}{n^2} \sum_{k=1}^{K} \frac{1}{r_k} \sum_{i=1}^{n_k} \frac{\|\mathbf{X}_{ik}\|^l}{\pi_{ik}} = O_P(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k})$ for l=2 and 4, where $\|\cdot\|$ denotes the Euclidean norm of a vector.

(**A.3**) As $n \to \infty$, the matrix $\mathcal{H}_X = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{w_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}}) \mathbf{X}_{ik} \mathbf{X}_{ik}^T}{\pi_{ik}}$ converges to a positive definite matrix in probability.

(**A.4**) $\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|\mathbf{X}_{ik}\|^6 = O_p(1)$.

(**A.5**) $\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} = o_P(1)$.

(**A.6**) $\frac{1}{n^3} \sum_{k=1}^{K} \frac{1}{r_k^2} \sum_{i=1}^{n_k} \frac{\|\mathbf{X}_{ik}\|^3}{\pi_{ik}^2} = O_P(\sum_{k=1}^{K} \frac{n_k^3}{n^3 r_k^2})$.

Assumption (A.1) is a standard condition in the proofs. Assumptions (A.2) and (A.6) are two conditions on the subsampling probabilities, allocation sizes and covariates distribution. For uniform subsampling with $\pi_{ik} = 1/n_k$ and $r_k = rn_k/n$, the sufficient condition for those assumptions is $E\|\mathbf{X}\|^4 < \infty$. Assumptions (A.3) and (A.4) impose two conditions on the covariates. (A.3) holds if $E(\mathbf{X}\mathbf{X}^T)$ is positive definite, and (A.4) holds if $E\|\mathbf{X}\|^6 < \infty$. Assumption (A.5) is reasonable for uniform allocation sizes with $\{r_k = rn_k/n\}_{k=1}^K$, and it holds as $r \to \infty$.

Theorem 1. *Under Assumptions (A.1)−(A.5), as $n \to \infty$, for any $\epsilon > 0$, with probability approaching one, there exists a finite Δ_ϵ and r_ϵ, such that*

$$P\left(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\| \geq \left\{\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k}\right\}^{1/2} \Delta_\epsilon \Big| \mathcal{F}_n\right) < \epsilon, \tag{2.2}$$

for all $r_k \geq r_\epsilon$, and $k = 1, \cdots K$.

Under Assumption (A.5), the convergence rate in (2.2) ensures that we can consistently approximate $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ with $\tilde{\boldsymbol{\beta}}$. For practical application, we suggest to use $\tilde{\boldsymbol{\beta}}$ rather than $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ for reducing computational burden. Next, we establish the asymptotic normality of $\tilde{\boldsymbol{\beta}}$, which is given in the following theorem.

Theorem 2. *If Assumptions (A.1)−(A.6) hold, conditional on $\mathcal{F}_n$, and as $n \to \infty$ and $r \to \infty$, with probability tending to one, we have*

$$\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) \xrightarrow{d} N(0, \mathbf{I}), \tag{2.3}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\boldsymbol{\Sigma} = \mathcal{H}_X^{-1}\boldsymbol{\Gamma}\mathcal{H}_X^{-1}$ with

$$\boldsymbol{\Gamma} = \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{n_k}\frac{\{Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\pi_{ik}}. \tag{2.4}$$

3 Optimal Subsampling Strategy

We consider how to specify the subsampling probablities $\{\pi_{ik}\}_{i=1}^{n_k}$, and the allocation sizes $\{r_k\}_{k=1}^{K}$ for given r. A naive choice is the uniform subsampling strategy with and $\{\pi_{ik} = 1/n_k\}_{i=1}^{n_k}$ and $\{r_k = [rn_k/n]\}_{k=1}^{K}$, where $[\ \cdot\ ]$ denotes the rounding operation. However, this uniform subsampling method is not optimal. A nonuniform subsampling strategy may has a better performance (Wang et al., 2018). Our idea is to determine the optimal allocation sizes and optimal subsampling probabilities by minimizing the asymptotic variance matrix $\mathbf{\Sigma}$ in Theorem 2. However, because $\mathbf{\Sigma}$ is a matrix, the meaning of "minimizing" needs to be carefully defined. For this purpose, we adopt the idea of A-optimality from optimal design of experiments, and use the trace to induce a complete ordering of the asymptotic variance matrix (Kiefer, 1959). In this case, the asymptotic mean squared error (AMSE) of $\tilde{\boldsymbol{\beta}}$ is equal to the trace of $\mathbf{\Sigma}$, i.e.,

$$AMSE(\tilde{\boldsymbol{\beta}}) = tr(\mathbf{\Sigma}), \tag{3.1}$$

where $tr(\cdot)$ denotes the trace of a matrix.

As mentioned above, the optimal allocation sizes and subsampling probabilities require the calculation of $\mathcal{H}_X^{-1}$ if we determine them by minimizing $tr(\mathbf{\Sigma})$, which takes substantial time in the case of big n. Note that $\mathcal{H}_X$ and $\mathbf{\Gamma}$ are nonnegative definite, and $\mathbf{\Sigma} = (\mathcal{H}_X)^{-1}\mathbf{\Gamma}(\mathcal{H}_X)^{-1}$. Simple matrix algebra yields that $tr(\mathbf{\Sigma}) = tr(\mathbf{\Gamma}\mathcal{H}_X^{-2}) \leq \lambda_{max}(\mathcal{H}_X^{-2})\ tr(\mathbf{\Gamma})$, where $\lambda_{max}(\cdot)$ denotes the maximum eigenvalue of a matrix. The minimizer of $tr(\mathbf{\Gamma})$ minimizes an upper bound of $tr(\mathbf{\Sigma})$. In fact, $\mathbf{\Sigma}$ depends on r_k and π_{ik} only through $\mathbf{\Gamma}$, and $\mathcal{H}_X$ is free of r_k and π_{ik}. Hence, we suggest to determine the optimal allocation sizes and optimal subsampling probabilites by directly minimizing $tr(\mathbf{\Gamma})$ rather than $tr(\mathbf{\Sigma})$, which can effectively speed up our subsampling algorithm.

Theorem 3. *In Algorithm 1, if the subsampling probabilities and allocation sizes are chosen as*

$$\pi_{ik}^{m\Gamma} = \frac{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}{\sum_{i=1}^{n_k} |Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}, i = 1, \cdots, n_k, \tag{3.2}$$

and

$$r_k^{m\Gamma} = r \cdot \frac{\sum_{i=1}^{n_k} |Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} |Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}, k = 1, \cdots, K, \tag{3.3}$$

then $tr(\boldsymbol{\Gamma})$ *attains its minimum.*

Remark: *In practice, we can use* $[r_k^{m\Gamma}]$ *as the optimal allocation sizes for* $k = 1, \cdots, K$*, where* $[\ \cdot\ ]$ *denotes the rounding operation.*

4 Two-Step Algorithm

The optimal subsampling probabilities and allocation sizes in (3.2) and (3.3) depend on the unavailable $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$. To deal with this problem, we use a pilot estimator $\tilde{\boldsymbol{\beta}}_0$ to replace $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ in (3.2) and (3.3). Below, we propose a two-step subsampling procedure in Algorithm 2.

Algorithm 2 Two-Step Algorithm

• *Step 1*: Given r_0, we run Algorithm 1 with subsampling size $r_k = [r_0 n_k/n]$ to obtain a pilot estimator $\tilde{\boldsymbol{\beta}}_0$, using either $\pi_{ik} = 1/n_k$ or $\pi_{ik} = 1/(2n_{0k})$ if $i \in S_{0k}$ and $\pi_{ik} = 1/(2n_{1k})$ if $i \in S_{1k}$, where $[\,\cdot\,]$ denotes the rounding operation. Here n_{0k} and n_{1k} are the numbers of elements in $S_{0k} = \{i : Y_{ik} = 0\}$ and $S_{1k} = \{i : Y_{ik} = 1\}$, respectively. Replace $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ with $\tilde{\boldsymbol{\beta}}_0$ in (3.2) and (3.3) to get the allocation sizes $r_k(\tilde{\boldsymbol{\beta}}_0)$ and subsampling probabilities $\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)$, for $i = 1, \cdots, n_k$ and $k = 1, \cdots, K$, respectively.

• *Step 2*: Based on the $\{r_k(\tilde{\boldsymbol{\beta}}_0)\}_{k=1}^K$ and $\{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)\}_{i=1}^{n_k}$ in Step 1, we can select a subsample $\{(Y_{ik}^*, \mathbf{X}_{ik}^*, \pi_{ik}^*) : i = 1, \cdots, r_k\}_{k=1}^K$ from the full data $\mathcal{F}_n$. Maximize the following weighted log-likelihood function to get a two-step subsample estimate $\breve{\boldsymbol{\beta}}$.

$$\ell_{\tilde{\boldsymbol{\beta}}_0}^*(\boldsymbol{\beta}) = \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{1}{\pi_{ik}^*(\tilde{\boldsymbol{\beta}}_0)} [Y_{ik}^* \log P_{ik}^*(\boldsymbol{\beta}) + (1 - Y_{ik}^*) \log\{1 - P_{ik}^*(\boldsymbol{\beta})\}],$$

where $P_{ik}^*(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^T \mathbf{X}_{ik}^*)}{1+\exp(\boldsymbol{\beta}^T \mathbf{X}_{ik}^*)}$.

For the subsample-based estimator $\breve{\boldsymbol{\beta}}$ in Algorithm 2, we need the following assumption in order to derive its asymptotic properties, and a similar assumption was also required by Wang et al. (2018).

(**A.7**) $E(e^{4\lambda\|\mathbf{X}\|}) < \infty$, where $\lambda = sup_{\boldsymbol{\beta} \in J_B} \|\boldsymbol{\beta}\|$.

Theorem 4. *Under Assumptions (A.1), (A.4) and (A.7), if the pilot estimate* $\tilde{\boldsymbol{\beta}}_0$ *exists, then as* $r_0 \to \infty$, $r \to \infty$, *and* $n \to \infty$, *for any* $\epsilon > 0$, *with probability approaching one, there exists a finite* Δ_ϵ *and* r_ϵ, *such that*

$$P(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\| \geq r^{1/2} \Delta_\epsilon | \mathcal{F}_n) < \epsilon, \tag{4.1}$$

for all $r \geq r_\epsilon$.

Based on Theorem 4, as long as the pilot estimate $\tilde{\boldsymbol{\beta}}_0$ exists, the two-step Algorithm 2 produces a consistent subsample-based estimator $\breve{\boldsymbol{\beta}}$. Its asymptotic normality is given in the following theorem.

Theorem 5. *If Assumptions (A.1), (A.4) and (A.7) hold, conditional on* $\tilde{\boldsymbol{\beta}}_0$ *and* $\mathcal{F}_n$, *as* $r_0 \to \infty$, $r \to \infty$, *and* $n \to \infty$, *with probability tending to one, we have*

$$\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) \xrightarrow{d} N(0, \mathbf{I}), \tag{4.2}$$

where $\xrightarrow{d}$ *denotes convergence in distribution,* $\boldsymbol{\Sigma} = \mathcal{H}_X^{-1}\boldsymbol{\Gamma}\mathcal{H}_X^{-1}$ *with*

$$\boldsymbol{\Gamma} = \frac{1}{rn^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{X}_{ik}\|\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\|\mathbf{X}_{ik}\|}. \tag{4.3}$$

In order to estimate the standard errors for each component of $\breve{\boldsymbol{\beta}}$, a simple way is to replace $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ with $\breve{\boldsymbol{\beta}}$ in the asymptotic variance matrix $\boldsymbol{\Sigma}$. However, it involves the full data with heavy calculation burden. To solve this issue, we propose to estimate the covariance matrix of $\breve{\boldsymbol{\beta}}$ with a subsample,

$$\breve{\boldsymbol{\Sigma}} = (\breve{\mathcal{H}}_X)^{-1}\breve{\boldsymbol{\Gamma}}(\breve{\mathcal{H}}_X)^{-1}, \tag{4.4}$$

where

$$\breve{\mathcal{H}}_X = \frac{1}{n}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{r_k}\frac{w_{ik}^*(\breve{\boldsymbol{\beta}})\mathbf{X}_{ik}^*\mathbf{X}_{ik}^{*T}}{\pi_{ik}^*},$$
$$\breve{\boldsymbol{\Gamma}} = \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k^2}\sum_{i=1}^{r_k}\frac{\{Y_{ik}^* - P_{ik}^*(\breve{\boldsymbol{\beta}})\}^2\mathbf{X}_{ik}^*\mathbf{X}_{ik}^{*T}}{\pi_{ik}^{*2}}.$$

From the above formulas, if $\breve{\boldsymbol{\beta}}$ is replaced by $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, then $\breve{\mathcal{H}}_X$ and $\breve{\boldsymbol{\Gamma}}$ are unbiased estimators of $\mathcal{H}_X$ and $\boldsymbol{\Gamma}$, respectively. The standard errors of components in $\breve{\boldsymbol{\beta}}$ are obtained by the square roots of diagonal elements of $\breve{\boldsymbol{\Gamma}}$. We will evaluate the performance of (4.4) by numerical studies in Section 5.

5 Numerical Studies

5.1 Simulation

In this section, we conduct simulations to verify our proposed method. The true parameter is $\boldsymbol{\beta} = (-1, -0.5, 0, 0.5, 1)^T$ with $d = 5$. We consider the following four cases for the covariate $\mathbf{X}$,

Case I : $\mathbf{X} \sim N(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{|i-j|}$.

Case II: $\mathbf{X} \sim N(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{I(i \neq j)}$.

Case III: $\mathbf{X} \sim t_5(0, \Sigma)$, i.e., $\mathbf{X}$ follows a multivariate t distribution with degree 5, and covariance matrix $\Sigma_{ij} = 0.5^{|i-j|}$.

Case IV: $\mathbf{X} = (X_1, \cdots, X_5)^T$, where X_i are independent exponential random variables with probability density function $f(x) = 2e^{-2x} I(x > 0)$, $i = 1, \cdots, 5$.

Note that in Cases I $-$ III the covariate distributions are symmetric, while in Case IV the covariate distribution is skewed. For Case IV, there could exist potential outliers in $\mathbf{X}$ due to the skewness of covariate distribution. We carry out computations on a server with 128GB memory using R software. All the simulation are based on 1000 replications. We set the sample size of each datasets as $\{n_k = [nu_k / \sum_{k=1}^{K} u_k]\}_{k=1}^{K}$, where u_k are generated from the uniform distribution over $(1, 2)$ with $K = 5$ and 100, respectively.

In Tables 1 and 2, we report the simulation results on subsample-based estimator for β_1 (other β_i's are similar and omitted), including the estimated bias (Bias) given by the sample mean of estimates minus $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, the mean of estimated standard errors (ESE) of the estimates, the sampling standard error (SSE) of the estimates, and the empirical 95% coverage probability (CP), where $r_0 = 200$, $n = 10^6$ and 10^8, respectively. The subsample sizes $r = 200$, 400, 600, 800 and 1000, respectively. It

can be seen from the results that the subsample-based estimator is unbiased. The ESE and SSE are close to each other, and the coverage probabilities are satisfactory. In all cases, the performance of our subsample-based estimator becomes better as r increases.

For comparison, we consider the uniform subsampling method (Uniform) with $\pi_{ik} = 1/n_k$, and $r_k = [rn_k/n]$, for $i = 1, \cdots, n_k$ and $k = 1, \cdots, K$. Let $MSE = \frac{1}{B}\sum_{b=1}^{B} \|\breve{\boldsymbol{\beta}}^{(b)} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|^2$, where $\breve{\boldsymbol{\beta}}^{(b)}$ is for the bth subsample, $b = 1, \cdots, B$. Figures 1 and 2 present the $MSEs$ of each method for $K = 5$, $n = 10^6$ and $K = 100$, $n = 10^8$, where $B = 1000$. From the results, we can see that the $MSEs$ of our method (Proposed) are much smaller than those of Uniform.

We conduct the second simulation to evaluate the computational efficiency of our two-step subsampling algorithm, where the mechanism of data generation is the same as the above-mentioned situation. For fair comparison, we count the CPU time with one core based on the mean calculation time of 1000 repetitions of each subsample-based method. In Table 3, we report the results for Case I with $n = 10^6, K = 5, r_0 = 200$, r=200, 400, 600, 800 and 1000, respectively. The computing time for the full data method is also given in the last row. Note that the uniform subsampling requires the least computing time, because its subampling probabilities $\pi_{ik} = 1/n_k$ and allocation sizes $r_k = [rn_k/n]$, do not take time to compute. Our subsampling algorithm has great computation advantage over the full data method. To further investigate the computational gain of the subsampling approach, we increase the dimension d to $d = 30$ with the true parameter $\boldsymbol{\beta}$ being a 30×1 vector of 0.5 entries. Table 4 records the computing time for Case I with $r_0 = 200, r = 1000, K = 5$, $n = 10^4, 10^5, 10^6, 10^7$ and 10^8, respectively. It is clear that both subsampling methods take significantly less computing times than the full data approach.

We conduct the third simulation to assess our method when the covariates have

different distributions towards corresponding distributed datasets. For $K = 5$, the distributed datasets $\mathcal{D}_1, \cdots, \mathcal{D}_5$ are generated similar to the first simulation, except that the covariates in $\mathcal{D}_1$ follow from $N_5(0, \mathbf{I})$, the covariates in $\mathcal{D}_2, \cdots, \mathcal{D}_5$ are generated from Cases I, II, III and IV, respectively. In Table 5, we present the Bias, SSE, ESE and CP for the proposed subsample estimator $\breve{\beta}_1$ with $n = 10^6$ and 10^7 (other β_i's are similar and omitted). Moreover, the MSEs of uniform and non-uniform subsampling methods are given in Figure 3. The results indicate that our method also works well with heterogeneous covariates, i.e., the covariates can have different distributions in different data blocks.

5.2 A Real Data Example

We apply our method to an example about airline data (Schifano et al., 2016), which are publicly available at *http://stat-computing.org/dataexpo/2009/* . The data consist of flight arrival and departure details for all commercial flights within the USA from October 1987 to April 2008, which are stored within 22 files year by year ($K = 22$; see Table 6). For analysis, the response variable Y denotes whether an airline is arrival delayed more than 15 minutes (1=yes, 0= otherwise). The vector of covariates $\mathbf{X} = (X_1, X_2, X_3)^T$, where X_1 is the day/night flight status (binary; 1 if departure between 8 p.m. and 5 a.m., 0 otherwise), X_2 is the departure delay time (continuous, in minutes) and X_3 is the distance (continuous, in thousands of miles). There are totally $n = 119,793,199$ observations with completed information on Y and $\mathbf{X}$.

For comparison, we also report the full data estimate $\hat{\boldsymbol{\beta}}_{\text{MLE}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^T$ with $\hat{\beta}_1 = -1.1120$, $\hat{\beta}_2 = 0.1284$ and $\hat{\beta}_3 = -2.4974$, respectively. Table 7 gives the mean of subsample estimates, the mean of estimated standard errors (ESE) and the sampling

standard error (SSE) of estimates based on 1000 subsamples with $r = 400, 600, 800$ and 1000, respectively. It can be seen from Table 6 that the subsample estimators are close to $\hat{\boldsymbol{\beta}}_{\text{MLE}}$. In Figure 4, we present the MSE of both subsampling methods based on $B = 1000$. We can see that the MSEs of our method (Proposed) are much smaller than those of Uniform. Moreover, an illustrative example about the allocation size r_k with $r = 1000$ is reported in Table 6.

6 Concluding Remarks

We have studied the statistical properties of a subsampling algorithm for the logistic regression model with distributed and massive data. We derived the optimal subsampling probabilities and optimal allocation sizes. The asymptotic properties of the subsample estimator were established. Some simulations and a real data example were provided to check the performance of our method.

There are several topics to be studied in the future. First, the simulations indicated that our method works well with potential outliers in the covariates. In the case of mislabels (outliers in the responses), it requires further research. Second, our method relies on the homogeneous structure of distributed datasets. In practice, it is important to consider the impact of the stratified heterogeneity of the strata on the sampling and regression, which is out of the scope of this manuscript. Third, our distributed subsampling approach can be extended to the big data generalized linear models (Ai et al., 2020; Zhang et al., 2020).

Acknowledgement

The authors would like to thank the Editor, an Associate Editor and three reviewers for their constructive and insightful comments that greatly improved the manuscript. The work of Wang was supported by National Science Foundation (NSF), USA grant DMS-1812013. The work of Sun was supported in part by the National Natural Science Foundation of China (Grant Nos. 11771431, 11690015 and 11926341) and Key Laboratory of RCSDS, CAS (No. 2008DP173182).

7 Appendix

Lemma 1. *If Assumptions (A.1)−(A.3) hold, then conditional on $\mathcal{F}_n$, we have*

$$\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} = O_{P|\mathcal{F}_n}\left(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\right)^{1/2}, \tag{7.1}$$

and

$$\tilde{\mathcal{H}}_X^{-1} = O_{P|\mathcal{F}_n}(1), \tag{7.2}$$

where $\tilde{\mathcal{H}}_X = \frac{\partial^2 \ell^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$, and the probability measure in $O_{P|\mathcal{F}_n}(\cdot)$ is conditional measure given $\mathcal{F}_n$.

Proof. For any $\boldsymbol{\beta} \in J_B$, we can derive that

$$E\left\{\frac{\dot{\ell}^*(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\} = \frac{\dot{\ell}(\boldsymbol{\beta})}{n}. \tag{7.3}$$

For the jth component of $\dot{\ell}^*(\boldsymbol{\beta})$, i.e., $\dot{\ell}_j^*(\boldsymbol{\beta}) = \sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{r_k}\frac{(\{Y_{ik}^* - P_{ik}^*(\boldsymbol{\beta})\}\mathbf{X}_{ik}^*)_j}{\pi_{ik}^*}$,

$$E\left\{\frac{\dot{\ell}_j^*(\boldsymbol{\beta})}{n} - \frac{\dot{\ell}_j(\boldsymbol{\beta})}{n}\middle|\mathcal{F}_n\right\}^2$$

$$
\begin{aligned}
&= E\Big\{ \frac{1}{n}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{r_k}\frac{(\{Y_{ik}^*-P_{ik}^*(\boldsymbol{\beta})\}\mathbf{X}_{ik}^*)_j}{\pi_{ik}^*} - \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\{Y_{ik}-P_{ik}(\boldsymbol{\beta})\}\mathbf{X}_{ik})_j \Big| \mathcal{F}_n \Big\}^2 \\
&= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\Big[\sum_{i=1}^{n_k}\frac{(\{Y_{ik}-P_{ik}(\boldsymbol{\beta})\}\mathbf{X}_{ik})_j^2}{\pi_{ik}} - \Big(\sum_{i=1}^{n_k}(\{Y_{ik}-P_{ik}(\boldsymbol{\beta})\}\mathbf{X}_{ik})_j\Big)^2\Big] \\
&\le \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^2}{\pi_{ik}}.
\end{aligned}
$$

By Assumption (A.2),

$$
E\Big\{ \frac{\dot{\ell}_j^*(\boldsymbol{\beta})}{n} - \frac{\dot{\ell}_j(\boldsymbol{\beta})}{n} \Big| \mathcal{F}_n \Big\}^2 = O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big).
$$

Using the Markov's inequality together with (7.3), we can get

$$
\frac{\dot{\ell}^*(\boldsymbol{\beta})}{n} - \frac{\dot{\ell}(\boldsymbol{\beta})}{n} = O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big)^{1/2}. \tag{7.4}
$$

By Assumption (A.1), we have $\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} - \frac{\dot{\ell}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} = O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big)^{1/2}$. Because $\frac{\dot{\ell}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} = 0$, it follows that (7.1) holds.

To prove (7.2), some direct calculations yield that

$$
E\Big\{ \frac{\partial^2 \ell^*(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \Big| \mathcal{F}_n \Big\} = \frac{\partial^2 \ell(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}. \tag{7.5}
$$

For any component $\frac{\partial^2 \ell_{j_1 j_2}^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$ of $\frac{\partial^2 \ell^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$ with $1 \le j_1, j_2 \le p$, we can derive that

$$
\begin{aligned}
&E\Big\{ \frac{\partial^2 \ell_{j_1 j_2}^*(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} - \frac{\partial^2 \ell_{j_1 j_2}(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \Big| \mathcal{F}_n \Big\}^2 \\
&= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\Big[\sum_{i=1}^{n_k}\frac{\{w_{ik}^2(\boldsymbol{\beta})\mathbf{X}_{ik}\mathbf{X}_{ik}^T\mathbf{X}_{ik}\mathbf{X}_{ik}^T\}_{j_1 j_2}}{\pi_{ik}} - \Big(\sum_{i=1}^{n_k}\{w_{ik}(\boldsymbol{\beta})\mathbf{X}_{ik}\mathbf{X}_{ik}^T\}_{j_1 j_2}\Big)^2\Big] \\
&\le \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^4}{\pi_{ik}}.
\end{aligned}
$$

By Assumption (A.2),

$$E\Big\{\frac{\partial^2\ell^*_{j_1j_2}(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}-\frac{\partial^2\ell_{j_1j_2}(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\Big|\mathcal{F}_n\Big\}^2=O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^K\frac{n_k^2}{n^2r_k}\Big).$$

It follows from the Markov's inequality that

$$\frac{\partial^2\ell^*(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}-\frac{\partial^2\ell(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}=O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^K\frac{n_k^2}{n^2r_k}\Big)^{1/2}. \tag{7.6}$$

Based on Assumptions (A.1) and (A.3), we know (7.2) holds. This ends the proof.

Proof of Theorem 1. Conditional on $\mathcal{F}_n$, the Assumption (A.5), Lemma 1 and (7.4) lead to that $\frac{\dot\ell^*(\boldsymbol{\beta})}{n}-\frac{\dot\ell(\boldsymbol{\beta})}{n}\to 0$ in probability. Note that the parameter space J_B is compact, and $\hat{\boldsymbol{\beta}}_{\rm MLE}$ is the unique solution to $\frac{\dot\ell(\boldsymbol{\beta})}{n}=0$. Thus, it follows from Theorem 5.9 and its remark of van der Vaart (1998) that conditional on $\mathcal{F}_n$, as $n\to\infty$,

$$\|\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\rm MLE}\|=o_{P|\mathcal{F}_n}(1). \tag{7.7}$$

Using the Taylor's theorem (Ferguson, 1996, Chapter 4), we have

$$0=\frac{\dot\ell^*_j(\tilde{\boldsymbol{\beta}})}{n}=\frac{\dot\ell^*_j(\hat{\boldsymbol{\beta}}_{\rm MLE})}{n}+\frac{\partial^2\ell^*_j(\hat{\boldsymbol{\beta}}_{\rm MLE})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\rm MLE})+\frac{1}{n}R_j, \tag{7.8}$$

where

$$R_j=(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\rm MLE})^T\int_0^1\int_0^1\frac{\partial^2\dot\ell^*_j\{\hat{\boldsymbol{\beta}}_{\rm MLE}+uv(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\rm MLE})\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}vdudv(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\rm MLE}).$$

Note that for all $\boldsymbol{\beta}$,

$$\begin{aligned}\Big\|\frac{\partial^2\ell^*_j(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\Big\| &= \Big\|\sum_{k=1}^K\frac{1}{r_k}\sum_{i=1}^{r_k}\frac{P^*_{ik}(\boldsymbol{\beta})(1-P^*_{ik}(\boldsymbol{\beta}))(1-2P^*_{ik}(\boldsymbol{\beta}))}{\pi^*_{ik}}\mathbf{X}^*_{ik}\mathbf{X}^{*T}_{ik}\mathbf{X}^*_{ik}\Big\|\\ &\le \sum_{k=1}^K\frac{1}{r_k}\sum_{i=1}^{r_k}\frac{\|\mathbf{X}^*_{ik}\|^3}{\pi^*_{ik}}.\end{aligned}$$

Thus,

$$\Big\| \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}_j^*\{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} + uv(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} v du dv \Big\| \leq \sum_{k=1}^{K} \frac{1}{2r_k} \sum_{i=1}^{r_k} \frac{\|\mathbf{X}_{ik}^*\|^3}{\pi_{ik}^*} = O_{P|\mathcal{F}_n}(n), \quad (7.9)$$

where the last equality is from the fact that

$$\begin{aligned} & P\Big(\sum_{k=1}^{K} \frac{1}{nr_k} \sum_{i=1}^{r_k} \frac{\|\mathbf{X}_{ik}^*\|^3}{\pi_{ik}^*} \geq \tau \Big| \mathcal{F}_n \Big) \\ & \leq \frac{1}{n\tau} E\Big(\sum_{k=1}^{K} \frac{1}{r_k} \sum_{i=1}^{r_k} \frac{\|\mathbf{X}_{ik}^*\|^3}{\pi_{ik}^*} \Big) \\ & = \frac{1}{n\tau} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|\mathbf{X}_{ik}\|^3 \to 0, \end{aligned}$$

as $\tau \to \infty$ with Assumption (A.4). From (7.8) and (7.9), we have

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = -\mathcal{H}_X^{-1} \Big\{ \frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} + O_{P|\mathcal{F}_n}(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|^2) \Big\}. \quad (7.10)$$

It follows from (7.1) and (7.2), together with (7.7) and (7.10) that

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} \Big)^{1/2} + o_{P|\mathcal{F}_n}(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|).$$

Hence, $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K} \frac{n_k^2}{n^2 r_k} \Big)^{1/2}$. This ends the proof.

Proof of Theorem 2. Note that

$$\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} = \frac{1}{n} \sum_{k=1}^{K} \frac{1}{r_k} \sum_{i=1}^{r_k} \frac{\{Y_{ik}^* - P_{ik}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}\mathbf{X}_{ik}^*}{\pi_{ik}^*} = \sum_{k=1}^{K} \sum_{i=1}^{r_k} \boldsymbol{\eta}_{ik}. \quad (7.11)$$

Given $\mathcal{F}_n$, we know that $\{\boldsymbol{\eta}_{ik} : i = 1, \cdots, n_k, k = 1, \cdots, K\}$ are independent random variables with

$$\sum_{k=1}^{K} \sum_{i=1}^{r_k} Var(\boldsymbol{\eta}_{ik} | \mathcal{F}_n)$$

$$= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{n_k}\frac{\{Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\pi_{ik}} - \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\Big(\sum_{i=1}^{n_k}\{Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}\mathbf{X}_{ik}\Big)^2$$

$$= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{n_k}\frac{\{Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\pi_{ik}} + O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big) \tag{7.12}$$

$$= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{n_k}\frac{\{Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\pi_{ik}} + o_P(1), \tag{7.13}$$

where (7.12) and (7.13) hold by Assumptions (A.2) and (A.5), respectively. Meanwhile, for every $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{i=1}^{r_k} E\{\|\boldsymbol{\eta}_{ik}\|^2 I(\|\boldsymbol{\eta}_{ik}\| > \varepsilon)|\mathcal{F}_n\} \\
&\le \sum_{k=1}^{K}\sum_{i=1}^{r_k} E\Big\{\|\boldsymbol{\eta}_{ik}\|^2 \cdot \frac{\|\boldsymbol{\eta}_{ik}\|}{\varepsilon}\Big|\mathcal{F}_n\Big\} \\
&= \frac{1}{\varepsilon}\sum_{k=1}^{K}\sum_{i=1}^{r_k} E(\|\boldsymbol{\eta}_{ik}\|^3|\mathcal{F}_n) \\
&= \frac{1}{\varepsilon}\sum_{k=1}^{K}\sum_{i=1}^{r_k}\frac{1}{n^3 r_k^3}\sum_{i=1}^{n_k}\frac{|Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|^3\|\mathbf{X}_{ik}\|^3}{\pi_{ik}^2} \\
&\le \frac{1}{\varepsilon}\sum_{k=1}^{K}\frac{1}{n^3 r_k^2}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^3}{\pi_{ik}^2}.
\end{aligned}$$

By Assumptions (A.5) and (A.6), we can derive that

$$\sum_{k=1}^{K}\sum_{i=1}^{r_k} E\{\|\boldsymbol{\eta}_{ik}\|^2 I(\|\boldsymbol{\eta}_{ik}\| > \varepsilon)|\mathcal{F}_n\} \le \frac{1}{\varepsilon}O_P\Big(\sum_{k=1}^{K}\frac{n_k^3}{n^3 r_k^2}\Big) \le \frac{1}{\varepsilon}O_P\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big) = o_P(1).$$

In view of (7.11) and (7.13), together with the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart, 1998) and the Slutsky's theorem, conditional on $\mathcal{F}_n$, as $n \to \infty$ and $r_k \to \infty$, we have that

$$\frac{1}{n}\boldsymbol{\Gamma}^{-1/2}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) \xrightarrow{d} N(0, \mathbf{I}). \tag{7.14}$$

From Lemma 1, (7.10) and Theorem 1,

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = -\tilde{\mathcal{H}}_X^{-1}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big). \tag{7.15}$$

It can be checked that

$$\begin{aligned}
&-\tilde{\mathcal{H}}_X^{-1}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} \\
&= -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} - (\tilde{\mathcal{H}}_X^{-1} - \mathcal{H}_X^{-1})\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} \\
&= -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} + [\mathcal{H}_X^{-1}(\tilde{\mathcal{H}}_X - \mathcal{H}_X)\tilde{\mathcal{H}}_X^{-1}]\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} \\
&= -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}(1)O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big)^{1/2} O_{P|\mathcal{F}_n}(1)O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big)^{1/2} \\
&= -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big).
\end{aligned}$$

Hence,

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big). \tag{7.16}$$

By Assumption (A.2), we have

$$\boldsymbol{\Sigma} = \mathcal{H}_X^{-1}\boldsymbol{\Gamma}\mathcal{H}_X^{-1} = O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big). \tag{7.17}$$

Thus, (7.16) and (7.17) yield that

$$\begin{aligned}
&\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) \\
&= -\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big)^{1/2} \\
&= -\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Gamma}^{-1/2}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}\Big(\sum_{k=1}^{K}\frac{n_k^2}{n^2 r_k}\Big)^{1/2}.
\end{aligned}$$

$$= -\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Gamma}^{-1/2}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} + o_P(1). \tag{7.18}$$

Note that

$$\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\boldsymbol{\Gamma}^{1/2}(\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\boldsymbol{\Gamma}^{1/2})^T = \boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Gamma}^{1/2}\mathcal{H}_X^{-1}\boldsymbol{\Sigma}^{-1/2} = \mathbf{I}. \tag{7.19}$$

By (7.17), (7.18) and the Slutsky's theorem, we can get that as $n \to \infty$,

$$\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) \xrightarrow{d} N(0, \mathbf{I}).$$

This ends the proof.

Proof of Theorem 3. It can be shown that

$$\begin{aligned}
tr(\boldsymbol{\Gamma}) &= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{n_k} tr\Big(\frac{\{Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\pi_{ik}}\Big) \\
&= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\sum_{i=1}^{n_k}\frac{\{Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\|\mathbf{X}_{ik}\|^2}{\pi_{ik}} \\
&= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\Big[\sum_{i=1}^{n_k}\pi_{ik}\sum_{i=1}^{n_k}\frac{\{Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\|\mathbf{X}_{ik}\|^2}{\pi_{ik}}\Big] \\
&\geq \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k}\Big[\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{X}_{ik}\|\Big]^2 \qquad (7.20) \\
&= \frac{1}{n^2}\frac{1}{r}\sum_{k=1}^{K}r_k\sum_{k=1}^{K}\frac{\big[\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{X}_{ik}\|\big]^2}{r_k} \\
&\geq \frac{1}{n^2 r}\Big[\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{X}_{ik}\|\Big]^2, \qquad (7.21)
\end{aligned}$$

where (7.20) and (7.21) follows from the Cauchy-Schwarz inequality and the equality hold if and only if $\pi_{ik} \propto |Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{X}_{ik}\|$, and $r_k \propto \sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{X}_{ik}\|$, respectively. This ends the proof.

Next, we establish two lemmas that will be used in the proofs of Theorems 4 and 5.

Lemma 2. *Under Assumptions (A.4) and (A.7), for l=2 and 4,*

$$\frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^l}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} = O_{P|\mathcal{F}_n}(r^{-1}), \tag{7.22}$$

and

$$\frac{1}{n^3}\sum_{k=1}^{K}\frac{1}{r_k^2(\tilde{\boldsymbol{\beta}}_0)}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^3}{\pi_{ik}^2(\tilde{\boldsymbol{\beta}}_0)} = O_{P|\mathcal{F}_n}(r^{-2}). \tag{7.23}$$

Proof. It follows from the expressions of $r_k(\tilde{\boldsymbol{\beta}}_0)$ and $\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)$ that

$$\begin{aligned}
&\frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^l}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)}\\
&= \frac{1}{rn^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik}-P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|}{|Y_{ik}-P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|}\|\mathbf{X}_{ik}\|^l\\
&= \frac{1}{rn}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^{l-1}}{|Y_{ik}-P_{ik}(\tilde{\boldsymbol{\beta}}_0)|}\cdot\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik}-P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|\\
&\le \frac{1}{rn}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^{l-1}(1+e^{\mathbf{X}_{ik}^T\tilde{\boldsymbol{\beta}}_0}+e^{-\mathbf{X}_{ik}^T\tilde{\boldsymbol{\beta}}_0}) \qquad (7.24)\\
&\le \frac{1}{rn}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^{l-1}(1+2e^{\lambda\|\mathbf{X}_{ik}\|})\\
&\le \frac{3}{rn}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^{l-1}e^{\lambda\|\mathbf{X}_{ik}\|}, \qquad (7.25)
\end{aligned}$$

where (7.24) holds by Assumption (A.4). Note that

$$E\{\|\mathbf{X}_{ik}\|^{l-1}e^{\lambda\|\mathbf{X}_{ik}\|}\} \le \{E(\|\mathbf{X}_{ik}\|^{2(l-1)})E(e^{2\lambda\|\mathbf{X}_{ik}\|})\}^{1/2} < \infty. \tag{7.26}$$

Hence, (7.22) follows from (7.25), (7.26) and the law of large numbers. Analogously, we can prove that (7.23) holds. This ends the proof.

Lemma 3. *If Assumptions (A.1), (A.4) and (A.7) hold, conditional on $\mathcal{F}_n$ we have*

$$\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} = O_{P|\mathcal{F}_n}(r^{-1/2}), \tag{7.27}$$

and

$$\{\tilde{\mathcal{H}}_X^{\tilde{\boldsymbol{\beta}}_0}\}^{-1} = O_{P|\mathcal{F}_n}(1), \tag{7.28}$$

where $\tilde{\mathcal{H}}_X^{\tilde{\boldsymbol{\beta}}_0} = \frac{\partial^2 \ell^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$.

Proof. For any $\boldsymbol{\beta} \in J_B$, we can derive that

$$E\Big\{ \frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{n} \Big| \mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0 \Big\} = \frac{\dot{\ell}(\boldsymbol{\beta})}{n}. \tag{7.29}$$

For the jth component $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0 j}(\boldsymbol{\beta})$ of $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})$,

$$\begin{aligned}
&E\Big\{ \frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0 j}(\boldsymbol{\beta})}{n} - \frac{\dot{\ell}_j(\boldsymbol{\beta})}{n} \Big| \mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0 \Big\}^2 \\
&= \frac{1}{n^2} \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \Big[\sum_{i=1}^{n_k} \frac{(\{Y^*_{ik} - P^*_{ik}(\boldsymbol{\beta})\}\mathbf{X}^*_{ik})^2_j}{\pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)} - \Big\{ \sum_{i=1}^{n_k} (\{Y_{ik} - P_{ik}(\boldsymbol{\beta})\}\mathbf{X}_{ik})_j \Big\}^2 \Big] \\
&\le \frac{1}{n^2} \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{n_k} \frac{\|\mathbf{X}_{ik}\|^2}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)}.
\end{aligned}$$

By Lemma 2,

$$E\Big\{ \frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0 j}(\boldsymbol{\beta})}{n} - \frac{\dot{\ell}_j(\boldsymbol{\beta})}{n} \Big| \mathcal{F}_n \Big\}^2 = O_{P|\mathcal{F}_n}(r^{-1}). \tag{7.30}$$

In view of the Markov's inequality and Assumption (A.1), (7.27) follows from (7.29) and (7.30).

In a similar manner, we obtain

$$E\Big\{ \frac{\partial^2 \ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \Big| \mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0 \Big\} = \frac{\partial^2 \ell(\boldsymbol{\beta})}{n\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}. \tag{7.31}$$

For any component $\frac{\partial^2 \ell^{*j_1 j_2}_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ of $\frac{\partial^2 \ell^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ with $1 \le j_1, j_2 \le p$, it can be shown that

$$
\begin{aligned}
& E\Big\{ \frac{\partial^2 \ell^{*j_1 j_2}_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{n \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} - \frac{\partial^2 \ell_{j_1 j_2}(\boldsymbol{\beta})}{n \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big| \mathcal{F}_n \Big\}^2 \\
&= \frac{1}{n^2} \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \Big[\sum_{i=1}^{n_k} \frac{\{ w_{ik}^2(\boldsymbol{\beta}) \mathbf{X}_{ik} \mathbf{X}_{ik}^T \mathbf{X}_{ik} \mathbf{X}_{ik}^T \}_{j_1 j_2}}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} - \Big(\sum_{i=1}^{n_k} \{ w_{ik}(\boldsymbol{\beta}) \mathbf{X}_{ik} \mathbf{X}_{ik}^T \}_{j_1 j_2} \Big)^2 \Big] \\
&\le \frac{1}{n^2} \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{n_k} \frac{\| \mathbf{X}_{ik} \|^4}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} = O_{P|\mathcal{F}_n}(r^{-1}), \qquad (7.32)
\end{aligned}
$$

where (7.32) holds by Lemma 2. From (7.31), (7.32) and the Markov's inequality, we know that (7.28) holds. This ends the proof.

Proof of Theorem 4. It follows from (7.29) and (7.30) that given $\mathcal{F}_n$,

$$
\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})}{n} - \frac{\dot{\ell}(\boldsymbol{\beta})}{n} \to 0,
$$

Thus, conditional on $\mathcal{F}_n$,

$$
\| \breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} \| = o_P(1), \qquad (7.33)
$$

which ensures that $\breve{\boldsymbol{\beta}}$ is close to $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ as long as r is large enough. Using the Taylor's theorem (Ferguson, 1996, Chapter 4),

$$
0 = \frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0 j}(\breve{\boldsymbol{\beta}})}{n} = \frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0 j}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n} + \frac{\partial^2 \ell^*_{\tilde{\boldsymbol{\beta}}_0 j}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}{n \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} (\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) + \frac{1}{n} R_{\tilde{\boldsymbol{\beta}}_0 j}, \qquad (7.34)
$$

where

$$
R_{\tilde{\boldsymbol{\beta}}_0 j} = (\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}})^T \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0 j} \{ \hat{\boldsymbol{\beta}}_{\mathrm{MLE}} + uv(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}) \}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} v du dv (\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}).
$$

Note that for all $\boldsymbol{\beta}$,

$$
\Big\| \frac{\partial^2 \ell^*_{\tilde{\boldsymbol{\beta}}_0 j}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big\| = \Big\| \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{P^*_{ik}(\boldsymbol{\beta}) \{ 1 - P^*_{ik}(\boldsymbol{\beta}) \} \{ 1 - 2P^*_{ik}(\boldsymbol{\beta}) \}}{\pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)} \mathbf{X}^*_{ik} \mathbf{X}^{*T}_{ik} \mathbf{X}^*_{ik} \Big\|
$$

$$\leq \sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{\|\mathbf{X}^*_{ik}\|^3}{\pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)},$$

and by Assumption (A.4),

$$P\Big(\frac{1}{n}\sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{\|\mathbf{X}^*_{ik}\|^3}{\pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)} \geq \tau \Big| \mathcal{F}_n\Big) \leq \frac{\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^3}{\tau} \to 0$$

in probability as $\tau \to \infty$. Thus,

$$\Big\| \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0 j}\{\hat{\boldsymbol{\beta}}_{\text{MLE}} + uv(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} v du dv \Big\| = O_{P|\mathcal{F}_n}(n). \tag{7.35}$$

By (7.34) and (7.35),

$$\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = -\tilde{\mathcal{H}}_X^{\tilde{\boldsymbol{\beta}}_0 - 1}\Big\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} + O_{P|\mathcal{F}_n}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|^2)\Big\}. \tag{7.36}$$

Based on (7.27), (7.28), (7.33) and (7.36), we have

$$\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = O_{P|\mathcal{F}_n}(r^{-1/2}) + o_{P|\mathcal{F}_n}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}\|).$$

Hence, $\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = O_{P|\mathcal{F}_n}(r^{-1/2})$. This ends the proof.

Proof of Theorem 5. Let

$$\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} = \frac{1}{n}\sum_{k=1}^{K} \frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)} \sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \frac{\{Y^*_{ik} - P^*_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}\mathbf{X}^*_{ik}}{\pi^*_{ik}(\tilde{\boldsymbol{\beta}}_0)} = \sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} \boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}. \tag{7.37}$$

Given $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, we know that $\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}$ are independent random variables with

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} Var(\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0) &= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)}\sum_{i=1}^{n_k}\frac{\{Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} \\ &\quad -\frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)}\Big(\sum_{i=1}^{n_k}\{Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}\mathbf{X}_{ik}\Big)^2. \end{aligned} \tag{7.38}$$

Note that

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)}\Big(\sum_{i=1}^{n_k}\{Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}\mathbf{X}_{ik}\Big)^2\\
&=\frac{1}{rn^2}\sum_{k=1}^{K}\frac{(\sum_{i=1}^{n_k}\{Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}\mathbf{X}_{ik})^2}{\sum_{i=1}^{n_k}|Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|\\
&\le\frac{1}{rn^2}\sum_{k=1}^{K}\frac{(\sum_{i=1}^{n_k}|Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|)^2}{\sum_{i=1}^{n_k}|Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|\\
&=\frac{1}{rn^2}\Big(\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|\Big)^2\\
&\le\frac{1}{r}\Big(\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|\Big)^2\\
&=O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}
$$

By (7.38) and as $r\to\infty$,

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{i=1}^{r_k}Var(\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}|\mathcal{F}_n,\tilde{\boldsymbol{\beta}}_0) &= \frac{1}{n^2}\sum_{k=1}^{K}\frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)}\sum_{i=1}^{n_k}\frac{\{Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\}^2\mathbf{X}_{ik}\mathbf{X}_{ik}^T}{\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)}+O_{P|\mathcal{F}_n}(r^{-1})\\
&= \boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0}+o_P(1). \qquad (7.39)
\end{aligned}
$$

Meanwhile, for every $\varepsilon>0$,

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)}E\{\|\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}\|^2I(\|\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}\|>\varepsilon)|\mathcal{F}_n,\tilde{\boldsymbol{\beta}}_0\}\\
&\le\sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)}E\Big\{\|\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}\|^2\cdot\frac{\|\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}\|}{\varepsilon}\Big|\mathcal{F}_n,\tilde{\boldsymbol{\beta}}_0\Big\}\\
&=\frac{1}{\varepsilon}\sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)}E(\|\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}\|^3|\mathcal{F}_n,\tilde{\boldsymbol{\beta}}_0)\\
&=\frac{1}{\varepsilon}\sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)}\frac{1}{n^3r_k^3(\tilde{\boldsymbol{\beta}}_0)}\sum_{i=1}^{n_k}\frac{|Y_{ik}-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|^3\|\mathbf{X}_{ik}\|^3}{\pi_{ik}^2(\tilde{\boldsymbol{\beta}}_0)}
\end{aligned}
$$

$$\leq \frac{1}{\varepsilon}\frac{1}{n^3}\sum_{k=1}^{K}\frac{1}{r_k^2(\tilde{\boldsymbol{\beta}}_0)}\sum_{i=1}^{n_k}\frac{\|\mathbf{X}_{ik}\|^3}{\pi_{ik}^2(\tilde{\boldsymbol{\beta}}_0)}.$$

By Lemma 2, as $r \to \infty$, we have

$$\sum_{k=1}^{K}\sum_{i=1}^{r_k(\tilde{\boldsymbol{\beta}}_0)} E\{\|\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}\|^2 I(\|\boldsymbol{\eta}_{ik}^{\tilde{\boldsymbol{\beta}}_0}\| > \varepsilon)|\mathcal{F}_n, \tilde{\boldsymbol{\beta}}_0\} \leq \frac{1}{\varepsilon}O_{P|\mathcal{F}_n}(r^{-2}) = o_P(1). \tag{7.40}$$

It follows from (7.37) and (7.39), together with the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart, 1998) and the Slutsky's theorem, we know that conditional on $\mathcal{F}_n$, as $n \to \infty$ and $r \to \infty$,

$$\frac{1}{n}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{-1/2}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}}) \xrightarrow{d} N(0, \mathbf{I}). \tag{7.41}$$

By Lemma 3, (7.36) and Theorem 5, we get that

$$\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}} = -\tilde{\mathcal{H}}_X^{\tilde{\boldsymbol{\beta}}_0 - 1}\Big\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n} + O_{P|\mathcal{F}_n}(r^{-1})\Big\} \tag{7.42}$$

Note that

$$\begin{aligned}
&-\tilde{\mathcal{H}}_X^{\tilde{\boldsymbol{\beta}}_0 - 1}\Big\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} \\
&= -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} - (\tilde{\mathcal{H}}_X^{\tilde{\boldsymbol{\beta}}_0 - 1} - \mathcal{H}_X^{-1})\Big\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} \\
&= -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} + \big[\mathcal{H}_X^{-1}(\tilde{\mathcal{H}}_X^{\tilde{\boldsymbol{\beta}}_0} - \mathcal{H}_X)\tilde{\mathcal{H}}_X^{\tilde{\boldsymbol{\beta}}_0 - 1}\big]\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} \\
&= -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}(1)O_{P|\mathcal{F}_n}(r^{-1/2})O_{P|\mathcal{F}_n}(1)O_{P|\mathcal{F}_n}(r^{-1/2}) \\
&= -\mathcal{H}_X^{-1}\Big\{\frac{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}(r^{-1}).
\end{aligned}$$

Hence, (7.42) and (7.22) yield that

$$\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MLE}}) = -\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{-1/2}\Big\{\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MLE}})}{n}\Big\} + O_{P|\mathcal{F}_n}(r^{-1/2}).$$

It can be proved that

$$\begin{aligned}
&\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}\{\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}\}^T \\
&= \boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}\mathcal{H}_X^{-1}\boldsymbol{\Sigma}^{-1/2} \\
&= \boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0}\mathcal{H}_X^{-1}\boldsymbol{\Sigma}^{-1/2} \\
&= \boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}\boldsymbol{\Gamma}\mathcal{H}_X^{-1}\boldsymbol{\Sigma}^{-1/2} + \boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0} - \boldsymbol{\Gamma})\mathcal{H}_X^{-1}\boldsymbol{\Sigma}^{-1/2} \\
&= \mathbf{I} + \boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0} - \boldsymbol{\Gamma})\mathcal{H}_X^{-1}\boldsymbol{\Sigma}^{-1/2}. \qquad (7.43)
\end{aligned}$$

For the distance between $\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0}$ and $\boldsymbol{\Gamma}$, we have

$$\|\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0} - \boldsymbol{\Gamma}\| \le \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^2\Big|\frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} - \frac{1}{r_k(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\pi_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}\Big|. \qquad (7.44)$$

A straightforward calculation yields that

$$\begin{aligned}
&\Big|\frac{1}{r_k(\tilde{\boldsymbol{\beta}}_0)\pi_{ik}(\tilde{\boldsymbol{\beta}}_0)} - \frac{1}{r_k(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\pi_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})}\Big| \\
&= \left|\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|}{|Y_{ik} - P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|\cdot r} - \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|\cdot r}\right| \\
&\le \frac{1}{r}\left|\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|}{|Y_{ik} - P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|} - \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|}{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}\right| \\
&\quad + \frac{1}{r}\left|\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\tilde{\boldsymbol{\beta}}_0)|\|\mathbf{X}_{ik}\|}{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|} - \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}\right| \\
&\le \frac{1}{r}\left|\frac{1}{|Y_{ik} - P_{ik}(\tilde{\boldsymbol{\beta}}_0)|} - \frac{1}{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|}\right|\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|}{\|\mathbf{X}_{ik}\|} \\
&\quad + \frac{1}{r}\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}|P_{ik}(\tilde{\boldsymbol{\beta}}_0) - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\|\mathbf{X}_{ik}\|}. \qquad (7.45)
\end{aligned}$$

Note that

$$\begin{aligned}
&\left|\frac{1}{|Y_{ik} - P_{ik}(\tilde{\boldsymbol{\beta}}_0)|} - \frac{1}{|Y_{ik} - P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|}\right| \\
&= \left|e^{(2Y_{ik}-1)\mathbf{X}_{ik}^T\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}} - e^{(2Y_{ik}-1)\mathbf{X}_{ik}^T\tilde{\boldsymbol{\beta}}_0}\right|
\end{aligned}$$

$$\leq e^{\lambda\|\mathbf{X}_{ik}\|}\|\mathbf{X}_{ik}\|\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|+e^{2\lambda\|\mathbf{X}_{ik}\|}\|\mathbf{X}_{ik}\|^2\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|^2, \tag{7.46}$$

and

$$\begin{aligned}
&\left|P_{ik}(\tilde{\boldsymbol{\beta}}_0)-P_{ik}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\right| \\
&=\frac{\left|e^{\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_{ik}}-e^{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\mathbf{X}_{ik}}\right|}{(1+e^{\tilde{\boldsymbol{\beta}}_0^T\mathbf{X}_{ik}})(1+e^{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}^T\mathbf{X}_{ik}})} \\
&\leq e^{\lambda\|\mathbf{X}_{ik}\|}\|\mathbf{X}_{ik}\|\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|+e^{2\lambda\|\mathbf{X}_{ik}\|}\|\mathbf{X}_{ik}\|^2\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|^2.
\end{aligned} \tag{7.47}$$

It follows from (7.44), (7.45), (7.46) and (7.47) that

$$\begin{aligned}
&\|\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0}-\boldsymbol{\Gamma}\| \\
&\leq \frac{1}{rn^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^2e^{\lambda\|\mathbf{X}_{ik}\|}\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\| \\
&\quad+\frac{1}{rn^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^3e^{2\lambda\|\mathbf{X}_{ik}\|}\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|^2\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\| \\
&\quad+\frac{3}{rn^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}e^{\lambda\|\mathbf{X}_{ik}\|}\|\mathbf{X}_{ik}\|\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^2e^{\lambda\|\mathbf{X}_{ik}\|}\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\| \\
&\quad+\frac{3}{rn^2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}e^{\lambda\|\mathbf{X}_{ik}\|}\|\mathbf{X}_{ik}\|\sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\mathbf{X}_{ik}\|^3e^{2\lambda\|\mathbf{X}_{ik}\|}\|\tilde{\boldsymbol{\beta}}_0-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}\|^2 \\
&=O_{P|\mathcal{F}_n}(r^{-1}r_0^{-1/2})+O_{P|\mathcal{F}_n}(r^{-1}r_0^{-1})+O_{P|\mathcal{F}_n}(r^{-1}r_0^{-1/2})+O_{P|\mathcal{F}_n}(r^{-1}r_0^{-1}) \\
&=O_{P|\mathcal{F}_n}(r^{-1}r_0^{-1/2}).
\end{aligned} \tag{7.48}$$

By (7.43) and (7.48),

$$\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}\{\boldsymbol{\Sigma}^{-1/2}\mathcal{H}_X^{-1}(\boldsymbol{\Gamma}^{\tilde{\boldsymbol{\beta}}_0})^{1/2}\}^T=\mathbf{I}+O_{P|\mathcal{F}_n}(r_0^{-1/2}). \tag{7.49}$$

By (7.49) and the Slutsky's theorem, as $r_0\to\infty$, $r\to\infty$ and $n\to\infty$, we can get that

$$\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathrm{MLE}})\xrightarrow{d}N(0,\mathbf{I}).$$

This completes the proof.

References

Ai, M., Yu, J., Zhang, H. and Wang, H. (2020). Optimal subsampling algorithms for big data generalized linear models. *Statistica Sinica*. DOI: 10.5705/ss.202018.0439.

Battey, H., Fan, J., Liu, H., Lu, J. and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46, 1352-1382.

Corbett, J., Dean, J. and Epstein, M. et al. (2013). Spanner: Google's globally distributed database. *ACM Transactions on Computer Systems*, 31, Article No. 8.

Ferguson, T. (1996). *A Course in Large Sample Theory*. New York: Chapman and Hall.

Fleming, T. and Harrington, D.(1991). *Counting Processes and Survival Analysis*. New York: John Wiley and Sons.

Hobza, T., Pardo, L. and Vajda, I. (2008). Robust median estimator in logistic regression. *Journal of Statistical Planning and Inference*, 138, 3822-3840.

Jordan, M., Lee, J. and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114, 668-681.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society, Series B*, 21, 272-319.

Ma, P., Mahoney, M. and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16, 861-911.

Schifano, E., Wu, J., Wang, C., Yan, J. and Chen, M. (2016). Online updating of statistical inference in the big data setting. *Technometrics*, 58, 393-403.

Shi, C., Lu, W., and Song, R. (2018). A massive data framework for M-estimators with cubic-rate. *Journal of the American Statistical Association*, 113, 1698-1709.

van der Vaart, A. (1998). *Asymptotic Statistics*. London: Cambridge University Press.

Volgushev, S., Chao, S. and Cheng, G. (2019). Distributed inference for quantile regression processes. *The Annals of Statistics*, 47, 1634-1662.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsample. *Journal of Machine Learning Research*, 20, 1-59.

Wang, H., Zhu, R. and Ma, P. (2018). Optimal subsampling for large sample Logistic regression. *Journal of the American Statistical Association*, 113, 829-844.

Wang, H., Yang, M. and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114, 393-405.

Zhang, T., Ning, Y. and Ruppert, D. (2020). Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2020.1778483

Zhao, T., Cheng, G., Liu, H. (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics*, 44, 1400-1437.

Zuo, L., Zhang, H., Wang, H., and Liu, L. (2020). Sampling-based estimation for massive survival data with additive hazards model. *Statistics in Medicine*, DOI:10.1002/sim.8783

Table 1.

The proposed subsample estimate of β_1 with $K = 5$ and $n = 10^{6\ddagger}$.

	Case I				Case II			
r	Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
200	0.0114	0.1952	0.2003	0.952	0.0152	0.2329	0.2423	0.940
400	0.0127	0.1346	0.1354	0.946	0.0059	0.1605	0.1594	0.958
600	0.0006	0.1087	0.1130	0.944	0.0102	0.1299	0.1315	0.954
800	0.0030	0.0940	0.0963	0.947	0.0014	0.1121	0.1131	0.958
1000	0.0039	0.0839	0.0838	0.955	0.0013	0.0997	0.1009	0.950
	Case III				Case IV			
r	Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
200	0.0139	0.1724	0.1815	0.946	0.0063	0.3246	0.3429	0.951
400	0.0035	0.1189	0.1227	0.954	0.0160	0.2234	0.2314	0.956
600	0.0032	0.0955	0.0944	0.968	0.0086	0.1805	0.1876	0.943
800	0.0012	0.0827	0.0832	0.941	0.0057	0.1548	0.1537	0.948
1000	0.0044	0.0737	0.0776	0.941	0.0087	0.1382	0.1380	0.952

‡ "Bias" denotes the sample mean of the estimates minus $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$; "ESE" denotes the estimated standard error of the estimates; "SSE" denotes the sampling standard error of the estimates; "CP" denotes the empirical 95% coverage probability towards $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$.

Table 2.

The proposed subsample estimate of β_1 with $K = 100$ and $n = 10^{8\ddagger}$.

	Case I				Case II			
r	Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
200	0.0090	0.2246	0.2194	0.968	0.0189	0.2451	0.2449	0.954
400	0.0063	0.1436	0.1504	0.935	0.0082	0.1599	0.1627	0.950
600	0.0015	0.1131	0.1164	0.941	0.0079	0.1260	0.1273	0.957
800	0.0023	0.0966	0.0958	0.954	0.0053	0.1078	0.1133	0.934
1000	0.0020	0.0856	0.0843	0.959	0.0052	0.0960	0.0980	0.955
	Case III				Case IV			
r	Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
200	0.0186	0.2000	0.2028	0.955	0.0149	0.3613	0.4004	0.940
400	0.0098	0.1274	0.1267	0.949	0.0123	0.2301	0.2309	0.953
600	0.0002	0.0998	0.1026	0.945	0.0022	0.1813	0.1924	0.941
800	0.0051	0.0851	0.0869	0.943	0.0011	0.1550	0.1511	0.958
1000	0.0023	0.0753	0.0786	0.929	0.0013	0.1361	0.1359	0.953

‡ "Bias" denotes the sample mean of the estimates minus $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$; "ESE" denotes the estimated standard error of the estimates; "SSE" denotes the sampling standard error of the estimates; "CP" denotes the empirical 95% coverage probability towards $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$.

Table 3.

The CPU time for Case I with $K = 5$ and $n = 10^6$ (seconds).

	r				
Methods	200	400	600	800	1000
Uniform	0.021	0.022	0.022	0.023	0.024
Proposed	0.254	0.262	0.270	0.282	0.288
Full data	1.239				

Table 4.

The CPU time for Case I with $r = 1000, K = 5$ and $d = 30$ (seconds).

	n				
Methods	10^4	10^5	10^6	10^7	10^8
Uniform	0.037	0.034	0.096	0.423	5.511
Proposed	0.084	0.183	0.580	4.921	70.350
Full data	0.098	0.734	5.619	53.809	768.476

Table 5.

The proposed subsample estimate of β_1 with $K = 5$, $n = 10^6$ and $n = 10^7$‡.

	$n = 10^6$				$n = 10^7$			
r	Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
200	0.0104	0.2075	0.2164	0.948	0.0061	0.2069	0.2198	0.940
400	0.0017	0.1422	0.1385	0.961	0.0062	0.1426	0.1394	0.963
600	0.0075	0.1148	0.1193	0.936	0.0034	0.1157	0.1161	0.950
800	0.0011	0.0989	0.1043	0.946	0.0035	0.0997	0.0995	0.952
1000	0.0003	0.0882	0.0917	0.938	0.0015	0.0889	0.0903	0.941

‡ "Bias" denotes the sample mean of the estimates minus $\hat{\boldsymbol{\beta}}_{\text{MLE}}$; "ESE" denotes the estimated standard error of the estimates; "SSE" denotes the sampling standard error of the estimates; "CP" denotes the empirical 95% coverage probability towards $\hat{\boldsymbol{\beta}}_{\text{MLE}}$.

Table 6.

The number of yearly data and allocation sizes($r = 1000$).

Years	n_k	r_k	Years	n_k	r_k
1987	1287333	11	1998	5227051	44
1988	5126498	40	1999	5360018	46
1989	4925482	42	2000	5481303	49
1990	5110527	39	2001	4873031	40
1991	4995005	38	2002	5093462	39
1992	5020651	40	2003	6375689	44
1993	4993587	42	2004	6987729	55
1994	5078411	42	2005	6992838	58
1995	5219140	47	2006	7003802	63
1996	5209326	49	2007	7275288	66
1997	5301999	47	2008	6855029	59

Table 7.

Subsample-based estimate $\breve{\boldsymbol{\beta}}$ and (ESE, SSE) in the real data.

	$\boldsymbol{\beta}$	Proposed	Uniform
$r = 400$	β_1	-1.1111 (0.4815, 0.5461)	-1.2216 (0.5502, 0.5873)
	β_2	0.1263 (0.0179, 0.0184)	0.1357 (0.0209, 0.0229)
	β_3	-2.5089 (0.4282, 0.4186)	-2.5836 (0.4382, 0.4639)
$r = 600$	β_1	-1.0752 (0.3677, 0.3900)	-1.1599 (0.4458, 0.4690)
	β_2	0.1281 (0.0139, 0.0142)	0.1339 (0.0170, 0.0218)
	β_3	-2.4871 (0.3572, 0.3846)	-2.5805 (0.3847, 0.3945)
$r = 800$	β_1	-1.0481 (0.3445, 0.3538)	-1.1585 (0.3821, 0.3969)
	β_2	0.1271 (0.0112, 0.0113)	0.1337 (0.0149, 0.0187)
	β_3	-2.4822 (0.2896, 0.2913)	-2.5500 (0.3079, 0.3314)
$r = 1000$	β_1	-1.0883 (0.3120, 0.3236)	-1.1378 (0.3377, 0.3529)
	β_2	0.1286 (0.0098, 0.0099)	0.1323 (0.0133, 0.0182)
	β_3	-2.5110 (0.2625, 0.2688)	-2.5441 (0.2750, 0.2979)

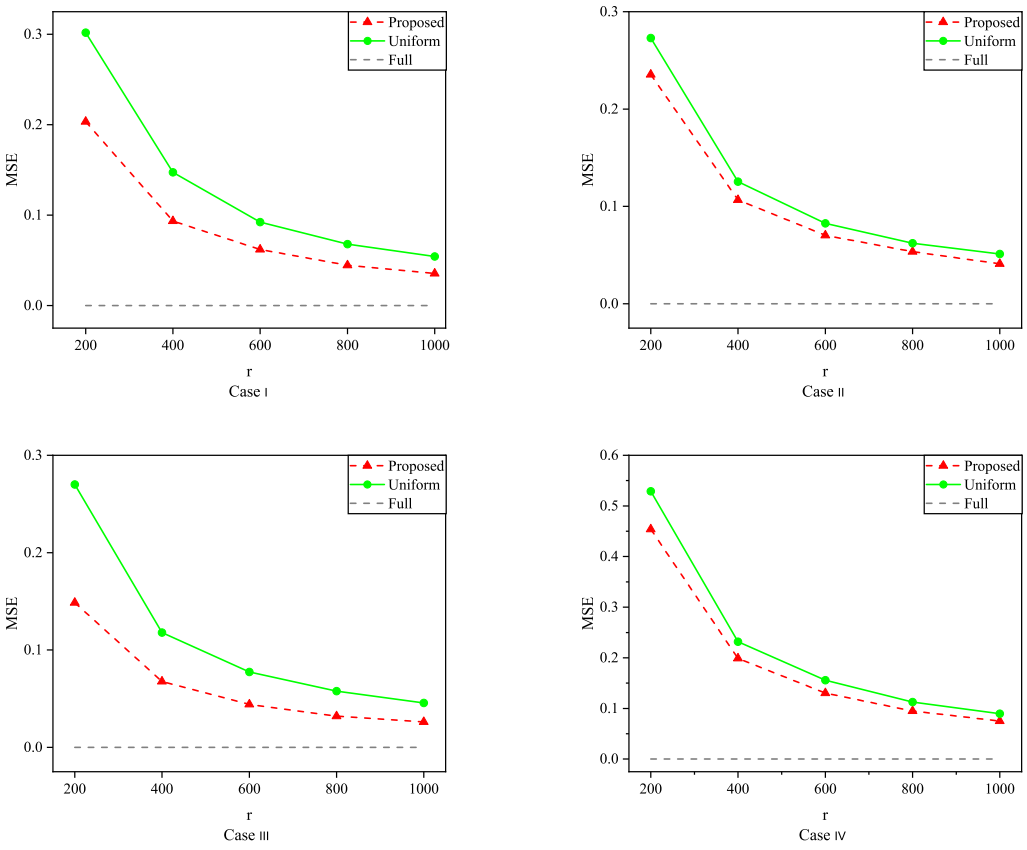

Figure 1. The MSEs for different subsampling methods with $K = 5$ and $n = 10^6$.

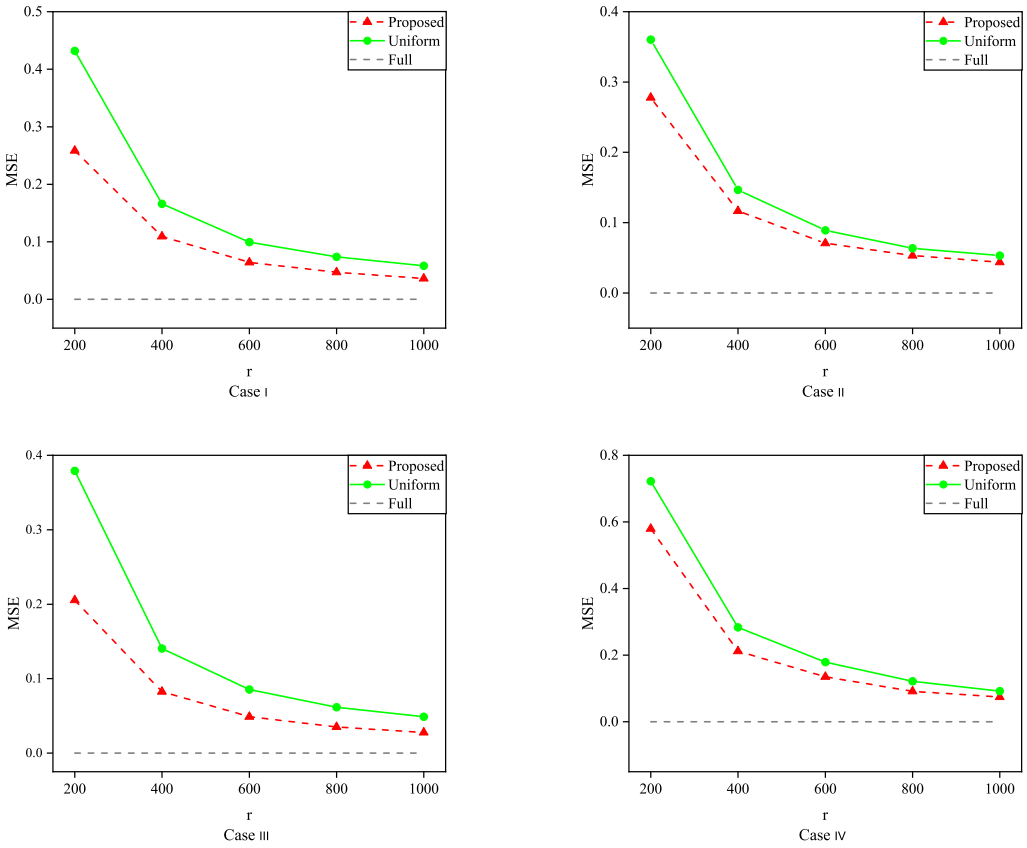

Figure 2. The MSEs for different subsampling methods with $K = 100$ and $n = 10^8$.

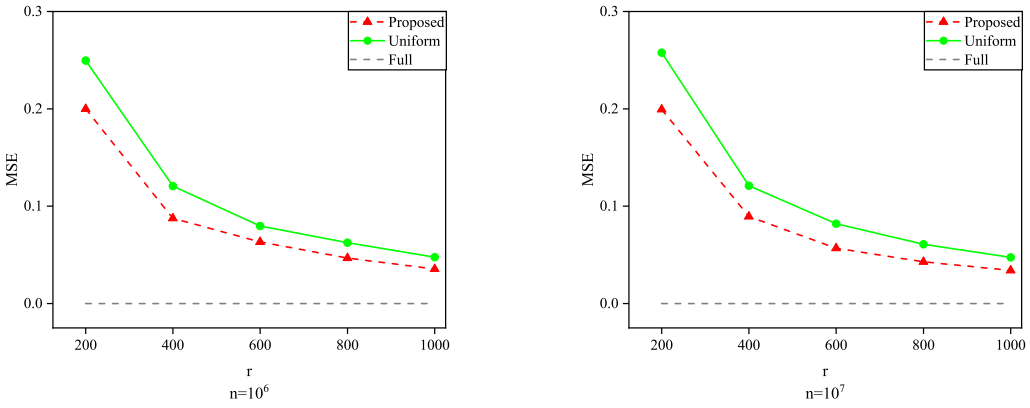

Figure 3. The MSEs for different subsampling methods with $n = 10^6$ and $n = 10^7$.

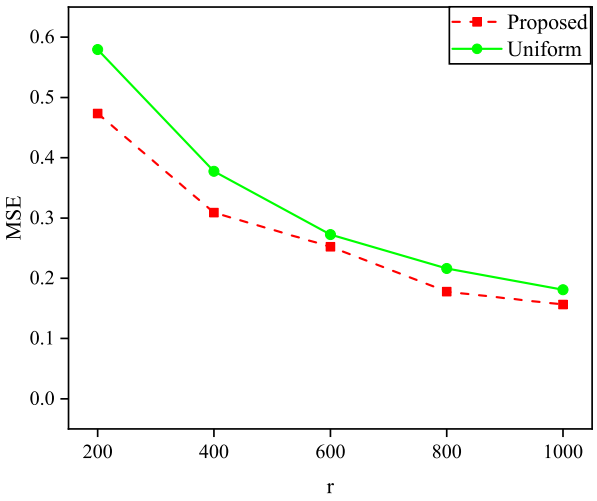

Figure 4. The results of MSE in the real data.