More Efficient Estimation for Logistic Regression with Optimal Subsample

HaiYing Wang
Department of Statistics, University of Connecticut
haiying.wang@uconn.edu

Abstract

Facing large amounts of data, subsampling is a practical technique to extract useful information. For this purpose, Wang et al. (2017) developed an Optimal Subsampling Method under the A-optimality Criterion (OSMAC) for logistic regression that samples more informative data points with higher probabilities. However, the original OSMAC estimator use inverse of optimal subsampling probabilities as weights in the likelihood function. This reduces contributions of more informative data points and the resultant estimator may lose efficiency. In this paper, we propose a more efficient estimator based on OSMAC subsample without weighting the likelihood function. Both asymptotic results and numerical results show that the new estimator is more efficient. In addition, our focus in this paper is inference for the true parameter, while Wang et al. (2017) focuses on approximating the full data estimator. We also develop a new algorithm based on Poisson sampling, which does not require to approximate the optimal subsampling probabilities all at once. This is computationally advantageous when available random-access memory is not enough to hold the full data. Interestingly, asymptotic distributions also show that Poisson sampling produces more efficient estimator if the sampling rate, the ratio of the subsample size to the full data sample size, does not converge to 0. We also obtain the unconditional asymptotic distribution for the estimator based on Poisson sampling.

Keywords: Asymptotic Distribution; Logistic Regression; Massive Data; Optimal Subsampling, Poisson Sampling.

1 Introduction

Extraordinary amounts of data that are collected offer unparalleled opportunities for advancing complicated scientific problems. However, the incredible sizes of big data bring new challenges for data analysis. A major challenge of big data analysis lies with the thirst for computing resources. Faced with this, subsampling has been widely used to reduce the computational burden, in which intended calculations are carried out on a subsample that is drawn from the full data, see Drineas, Kannan and Mahoney (2006*a*,*b*,*c*); Mahoney and Drineas (2009); Drineas et al. (2011); Mahoney (2011); Halko et al. (2011); Clarkson and Woodruff (2013); Kleiner et al. (2014); McWilliams et al. (2014); Yang et al. (2016), among others.

A key to success of a subsampling method is to specify nonuniform sampling probabilities so that more informative data points are sampled with higher probabilities. For this purpose, normalized statistical leverage scores or its variants are often used as subsampling probabilities in the context of linear regression, and this approach is termed *algorithmic leveraging* (Ma et al., 2014; Ma and Sun, 2015; Ma et al., 2015). It has demonstrated remarkable performance in making better use of a fixed amount of computing power (Avron et al., 2010; Meng et al., 2014). Statistical leverage scores only contain information in the covariate and do not take into account the information contained in the observed responses. Wang et al. (2017) derived optimal subsampling probabilities that minimize the asymptotic mean squared error (MSE) of the subsampling-based estimator in the context of logistic regression. The optimal subsampling probabilities directly depend on both the covariates and the responses to take more informative subsample. However, Wang et al. (2017) used a weighted maximum likelihood estimator based on the subsample, and more informative data points are assigned smaller weights in the likelihood function. Thus, the resultant estimator may not be efficient.

In this paper, we propose more efficient estimators based on subsample taken randomly according to the optimal subsampling probabilities. Asymptotic distributions will be derived which show that asymptotic variance-covariance matrices of the new estimators are smaller, in Loewner-ordering, than that of the weighted estimator in Wang et al. (2017). We also consider to use Poisson sampling instead of sampling with replacement. Asymptotic distributions show that Poisson sampling is more efficient in parameter estimation when the subsample size is proportional to the full data sample size. It is also computationally beneficial to use Poisson sampling because there is no need to generate a large amount of random numbers all at once.

Before presenting the framework of the paper, we give a brief review of the emerging field of subsampling-based methods. For linear regression, Drineas, Mahoney and Muthukrishnan (2006) developed a subsampling method and focused on finding influential data units for the least squares (LS) estimates. Drineas et al. (2011) developed an algorithm by processing the data with randomized Hadamard transform and then using uniform subsampling to approximate LS estimates. Drineas et al. (2012) developed an algorithm to approximate statistical leverage scores that are used for algorithmic leveraging. Yang et al. (2015) showed that using normalized square roots of statistical leverage scores as subsampling probabilities yields better approximation than using original statistical leverage scores, if they are very nonuniform. The aforementioned studies focused on developing algorithms for fast approximation of LS estimates. Ma et al. (2014, 2015) and Ma and Sun (2015) considered the statistical properties of algorithmic leveraging. They derived biases and variances of leverage-based subsampling estimators in linear regression and proposed a shrinkage algorithmic leveraging method to improve the performance. Raskutti and Mahoney (2016) considered both the algorithmic and statistical aspects of solving large-scale LS problems using random sketching. Wang et al. (2018) developed an information-based optimal subdata selection method to select subsample deterministically for ordinary LS in linear regression. The aforesaid results were obtained exclusively within the context of linear models. Fithian and Hastie (2014) proposed a computationally efficient local case-control subsampling method for logistic regression with large imbalanced data. Recently, Wang et al. (2017) developed an Optimal Subsampling Method under the A-optimality Criterion (OSMAC) for logistic regression. Although they derived optimal subsampling probabilities, the inference procedure based on weighted likelihood is not efficient.

This paper focuses on logistic regression models, which are widely used for statistical inference in many disciplines, such as business, computer science, education, and genetics, among others (Hosmer Jr et al., 2013). Based on optimal subsample taken according to OSMAC developed in Wang et al. (2017), more efficient methods, in terms of both parameter estimation and numerical computation, will be proposed. The reminder of the paper is organized as follows. Model setups and notations are introduced in Section 2. The OSMAC will also be briefly reviewed in this section. Section 3 presents the more efficient estimator and its asymptotic properties. Section 4 considers Poisson sampling. Section 5 discusses issues related to practical implementation and summaries the methods from Sections 3 and 4 into two practical algorithms. Section 6 gives unconditional asymptotic distributions for the estimator from Poisson sampling. Section 7 evaluates the practical performance of the proposed methods using numerical experiments. Section 8 concludes, and the appendix contains proofs and technical details.

2 Model setup and optimal subsampling

Let $y \in \{0, 1\}$ be a binary response variable and $\mathbf{x}$ be a d dimensional covariate. A logistic regression model describes the conditional probability of $y = 1$ given $\mathbf{x}$, and it has the following form.

$$\mathbb{P}(y = 1|\mathbf{x}) = p(\boldsymbol{\beta}) = \frac{e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}}{1 + e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}}, \tag{1}$$

where $\boldsymbol{\beta}$ is a $d \times 1$ vector of unknown regression coefficients belonging to a compact subset of $\mathbb{R}^d$.

With independent full data of size N from Model (1), say, $\mathcal{D}_N = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, the unknown parameter $\boldsymbol{\beta}$ is often estimated by the maximum likelihood estimator (MLE), denoted as $\hat{\boldsymbol{\beta}}_f$. It is the maximizer of the log-likelihood function, namely,

$$\hat{\boldsymbol{\beta}}_f = \arg\max_{\boldsymbol{\beta}} \ \ell_f(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left\{ y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i}\right) \right\}. \tag{2}$$

Since there is no general closed-form solution to the MLE, Newton's method or iteratively reweighted least squares method (McCullagh and Nelder, 1989) is often adopted to find it numerically. This typically takes $O(\zeta N d^2)$ time, where ζ is the number of iterations in the optimization procedure (Wang et al., 2017). For super-large data, the computing time $O(\zeta N d^2)$ may be too long to afford, and iterative computation is infeasible if the data volume is larger than the available random-access memory (RAM). To overcome this computational bottleneck for the application of logistic regression on massive data, Wang et al. (2017) developed the OSMAC under the subsampling framework.

Let π_1, ..., π_N be subsampling probabilities such that $\sum_{i=1}^{N} \pi_i = 1$. Using subsampling with replacement, draw a random subsample of size n ($\ll N$), according to the probabilities $\{\pi_i\}_{i=1}^{N}$, from the full data. We use * to indicate any quantity for subsample, namely, denote the covariates, responses, and subsampling probabilities in the subsample as $\mathbf{x}_i^*$, y_i^*, and π_i^*, respectively, for $i = 1, ..., n$. Wang et al. (2017) define the subsample estimator $\hat{\boldsymbol{\beta}}_w^{\pi}$ to be the weighted MLE, i.e.,

$$\hat{\boldsymbol{\beta}}_w^{\pi} = \arg\max_{\boldsymbol{\beta}} \ell_w^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \frac{y_i^* \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*}\right)}{\pi_i^*}. \tag{3}$$

The key to success here is how to specify the values for π_i's so that more informative data points are sampled with higher probabilities. Wang et al. (2017) derived optimal subsampling probabilities that minimize the asymptotic MSE of $\hat{\boldsymbol{\beta}}_w^{\pi}$. They first show that $\hat{\boldsymbol{\beta}}_w^{\pi}$ is asymptotically normal. Specifically, for large n and N, the conditional distribution of

$\sqrt{n}(\hat{\boldsymbol{\beta}}_w^{\pi} - \hat{\boldsymbol{\beta}}_f)$ given the full data $\mathcal{D}_N$ can be approximated by a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{V}_N = \mathbf{M}_N^{-1}(\hat{\boldsymbol{\beta}}_f)\mathbf{V}_{Nc}\mathbf{M}_N^{-1}(\hat{\boldsymbol{\beta}}_f)$, in which

$$\mathbf{M}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\phi_i(\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}, \quad \mathbf{V}_{Nc} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - p_i(\hat{\boldsymbol{\beta}}_f)|^2\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{N\pi_i},$$

and $\phi_i(\boldsymbol{\beta}) = p_i(\boldsymbol{\beta})\{1 - p_i(\boldsymbol{\beta})\}$ with $p_i(\boldsymbol{\beta}) = e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}/(1 + e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}})$. Based on this asymptotic distribution, they derive the following two optimal subsampling probabilities

$$\pi_i^{\mathrm{mMSE}}(\hat{\boldsymbol{\beta}}_f) = \frac{|y_i - p_i(\hat{\boldsymbol{\beta}}_f)|\|\mathbf{M}_N^{-1}(\hat{\boldsymbol{\beta}}_f)\mathbf{x}_i\|}{\sum_{j=1}^{N}|y_j - p_j(\hat{\boldsymbol{\beta}}_f)|\|\mathbf{M}_N^{-1}(\hat{\boldsymbol{\beta}}_f)\mathbf{x}_j\|}, \quad i = 1, ..., N; \tag{4}$$

$$\pi_i^{\mathrm{mVc}}(\hat{\boldsymbol{\beta}}_f) = \frac{|y_i - p_i(\hat{\boldsymbol{\beta}}_f)|\|\mathbf{x}_i\|}{\sum_{j=1}^{N}|y_j - p_j(\hat{\boldsymbol{\beta}}_f)|\|\mathbf{x}_j\|}, \quad i = 1, ..., N. \tag{5}$$

Here, π_i^{mMSE} minimize the trace of $\mathbf{V}_N$, $\mathrm{tr}(\mathbf{V}_N)$, and π_i^{mVc} minimize the trace of $\mathbf{V}_{Nc}$, $\mathrm{tr}(\mathbf{V}_{Nc})$. These subsampling probabilities have a lot of nice properties and meaningful interpretations. More details can be found in Section 3 of Wang et al. (2017).

For ease of presentation, use the following general notation to denote subsampling probabilities

$$\pi_i^{\mathrm{OS}}(\boldsymbol{\beta}) = \frac{|y_i - p_i(\boldsymbol{\beta})|h(\mathbf{x}_i)}{\sum_{j=1}^{N}|y_j - p_j(\boldsymbol{\beta})|h(\mathbf{x}_j)}, \quad i = 1, ..., N, \tag{6}$$

where $h(\mathbf{x})$ is a univariate function of $\mathbf{x}$. If $h(\mathbf{x}) = \|\mathbf{x}\|$, then $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_f)$ become $\pi_i^{\mathrm{mVc}}(\hat{\boldsymbol{\beta}}_f)$. If $h(\mathbf{x}) = \|\mathbf{M}^{-1}\mathbf{x}\|$ where $\mathbf{M}$ is the limit of $\mathbf{M}_N(\hat{\boldsymbol{\beta}}_f)$, then $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_f)$ approximate $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_f)$. If $h(\mathbf{x}) = 1$, then $\pi_i^{\mathrm{OS}}(\boldsymbol{\beta})$ are proportional to the local case-control subsampling probabilities (Fithian and Hastie, 2014).

Note that $\pi_i^{\mathrm{OS}}(\boldsymbol{\beta})$ depend on the unknown $\boldsymbol{\beta}$, so a pilot estimate of $\boldsymbol{\beta}$ is required to approximate them. Let $\hat{\boldsymbol{\beta}}_1$ be a pilot estimator from the pilot sample, for which we will provide more details in Section 5. The original weighted OSMAC estimator is

$$\hat{\boldsymbol{\beta}}_w = \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^{n}\frac{y_i^*\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*}\right)}{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)^*}. \tag{7}$$

In Wang et al. (2017), $\hat{\boldsymbol{\beta}}_w$ have exceptional performance because $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)$ are able to include more informative data points in the subsample. However, the weighting scheme adopted in (7) prevent $\hat{\boldsymbol{\beta}}_w$ from being the most efficient estimator. Intuitively, a larger $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)$ means that the data point $(\mathbf{x}_i, y_i)$ contains more information about $\boldsymbol{\beta}$, but it has a smaller weight in the log-likelihood in (7). This reduces contributions of more informative data points in the log-likelihood function for parameter estimation.

The weighted MLE in (7) is used because $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)$ dependent on the responses y_i's and an un-weighted MLE is biased. If the bias can be corrected, then the resultant estimator can be more efficient in parameter estimation. This is a major goal of this paper. Interestingly, for the subsampling probabilities in (7), the bright idea proposed in Fithian and Hastie (2014) can be used to correct the bias.

3 More efficient estimator with optimal subsampling

Let $\{(\mathbf{x}_1^*, y_1^*), ..., (\mathbf{x}_n^*, y_n^*)\}$ be a random subsample of size n taken from the full data using sampling with replacement according to the probabilities $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$ defined in (6). We present a more efficient estimation procedure based on subsample. Denote the more efficient estimator as $\hat{\boldsymbol{\beta}}_r$, where the subscript $_r$ stands for sampling with **r**eplacement. Remember that a pilot estimate is required, and we use $\hat{\boldsymbol{\beta}}_1$ to denote it. We will discuss how to obtain it and how to use it to improve the estimation efficiency in Section 5. Here, we focus the discussion on $\hat{\boldsymbol{\beta}}_r$ and assume that a consistent $\hat{\boldsymbol{\beta}}_1$ is available. The following procedure describes how to obtain $\hat{\boldsymbol{\beta}}_r$.

Calculate

$$\tilde{\boldsymbol{\beta}}_r = \arg\max_{\boldsymbol{\beta}} \ell_r^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* y_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*}\right)\right\}, \tag{8}$$

and let

$$\hat{\boldsymbol{\beta}}_r = \tilde{\boldsymbol{\beta}}_r + \hat{\boldsymbol{\beta}}_1. \tag{9}$$

The naive un-weighted MLE $\tilde{\boldsymbol{\beta}}_r$ in (8) is biased, and the bias is corrected in (9) using $\hat{\boldsymbol{\beta}}_1$. We will show in the following that $\hat{\boldsymbol{\beta}}_r$ is asymptotically unbiased. This, together with the fact that $\hat{\boldsymbol{\beta}}_1$ is consistent, shows the interesting fact that $\tilde{\boldsymbol{\beta}}_r$ converges to $\mathbf{0}$ in probability as $n \to \infty$ and $N \to \infty$.

To investigate the asymptotic properties, we use $\boldsymbol{\beta}_t$ to denote the true value of $\boldsymbol{\beta}$, and summarize some regularity conditions that are required in the following.

Assumption 1. *The matrix $\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$ is finite and positive-definite.*

Assumption 2. *The covariate $\mathbf{x}$ and function $h(\cdot)$ satisfy that $\mathbb{E}\{\|\mathbf{x}\|^2 h^2(\mathbf{x})\} < \infty$, and $\mathbb{E}\{\|\mathbf{x}\|^2 h(\mathbf{x})\} < \infty$.*

Assumption 3. *As $n \to \infty$, $n\mathbb{E}\{h(\mathbf{x})I(\|\mathbf{x}\|^2 > n)\} \to 0$, where $I()$ is the indicator function.*

Assumption 1 is required to establish the asymptotic normality. This is a commonly used assumption, e.g., in Fithian and Hastie (2014); Wang et al. (2017), among others. Assumptions 2 and 3 impose moment conditions on the covariate distribution and the function

$h(\mathbf{x})$. When $h(\mathbf{x}) = 1$, if $\mathbb{E}\|\mathbf{x}\|^2 < \infty$, then both the two conditions in Assumption 2 and the condition in Assumption 3 hold. Thus, the assumptions required in this paper are not stronger than those required by Fithian and Hastie (2014). When $h(\mathbf{x}) = \|\mathbf{x}\|$, by Hölder's inequality,

$$n\mathbb{E}\{h(\mathbf{x})I(\|\mathbf{x}\|^2 > n)\} \le n(\mathbb{E}\|\mathbf{x}\|^3)^{1/3}\{\mathbb{E}I(\|\mathbf{x}\|^2 > n)\}^{2/3} = (\mathbb{E}\|\mathbf{x}\|^3)^{1/3}\{n^{3/2}\mathbb{P}(\|\mathbf{x}\|^3 > n^{3/2})\}^{2/3}.$$

Therefore, if $\mathbb{E}\|\mathbf{x}\|^3 < \infty$, Assumption 3 holds. This shows that $\mathbb{E}\|\mathbf{x}\|^4 < \infty$ implies all the three conditions required in Assumptions 2 and 3. Note that Wang et al. (2017) requires that $\mathbb{E}e^{\mathbf{v}^{\mathrm{T}}\mathbf{x}} < \infty$ for any $\mathbf{v} \in \mathbb{R}^d$ in order to establish the asymptotic properties when a pilot estimate is used to approximate optimal subsampling probabilities. Thus, the required conditions in this paper are weaker than those required in Wang et al. (2017). Assumptions 1 and 2 are required in all the theorems in this paper while Assumption 3 is only required in Theorem 1.

Theorem 1. *Under assumptions 1-3, conditional on $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$, as $n \to \infty$ and $N \to \infty$,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_{wf}) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big), \tag{10}$$

in distribution; furthermore, if $n/N \to 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big) \tag{11}$$

in distribution. Here

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left[\frac{\mathbb{E}\{\phi(\boldsymbol{\beta})h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{4\Phi(\boldsymbol{\beta})}\right]^{-1}, \qquad \Phi(\boldsymbol{\beta}) = \mathbb{E}\{\phi(\boldsymbol{\beta})h(\mathbf{x})\} \tag{12}$$

and

$$\hat{\boldsymbol{\beta}}_{wf} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} |y_i - p_i(\hat{\boldsymbol{\beta}}_1)| h(\mathbf{x}_i)\big[y_i\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1) - \log\{1 + e^{\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)}\}\big], \tag{13}$$

which satisfies that, conditional on $\hat{\boldsymbol{\beta}}_1$,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{wf} - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{wf}\big), \tag{14}$$

in distribution, and

$$\boldsymbol{\Sigma}_{wf} = [\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1}\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}\}[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1}. \tag{15}$$

Remark 1. *The result in* (10) *of Theorem 1 also implies that $\hat{\boldsymbol{\beta}}_r$ is $\sqrt{n}$-consistent in a sense that is similar to Theorem 5 of Wang et al. (2017), namely, given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$ in probability,*

$$\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_{wf} = O_{P|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1}(n^{-1/2}). \tag{16}$$

The $O_{P|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1}(n^{-1/2})$ expression in (16) means that for any $\epsilon > 0$, there exist a δ_ϵ such that

$$\mathbb{P}\Big\{\sup_n \mathbb{P}(\|\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_{wf}\| > n^{-1/2}\delta_\epsilon|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1) \le \epsilon\Big\} \to 1 \tag{17}$$

as $n, N \to \infty$. Xiong and Li (2008) showed that if a sequence is bounded in conditional probability, then it is bounded in unconditional probability, i.e., if $a_n = O_{P|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1}(1)$, then $a_n = O_P(1)$. Therefore, (16) implies that $\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_{wf} = O_P(n^{-1/2})$. Similarly, (14) implies that $\hat{\boldsymbol{\beta}}_{wf} - \boldsymbol{\beta}_t = O_P(N^{-1/2})$. Thus, $\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_t = O_P(n^{-1/2} + N^{-1/2}) = O_P(n^{-1/2})$, showing the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\beta}}_r$ under the unconditional distribution.

Theorem 1 shows that, asymptotically, the distribution of $\hat{\boldsymbol{\beta}}_r$ given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$ is centered around $\hat{\boldsymbol{\beta}}_{wf}$ with variance-covariance matrix $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$, and the distribution of $\hat{\boldsymbol{\beta}}_{wf}$ is centered around $\boldsymbol{\beta}_t$ with variance-covariance matrix $N^{-1}\boldsymbol{\Sigma}_{wf}$. Thus, both $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ and $N^{-1}\boldsymbol{\Sigma}_{wf}$ should be considered in accessing the quality of $\hat{\boldsymbol{\beta}}_r$ for estimating the true parameter $\boldsymbol{\beta}_t$. However, in subsampling setting, it is expected that $n \ll N$; otherwise, the computational benefit is minimum. Thus, $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ is the dominating term in quantifying the variation of $\hat{\boldsymbol{\beta}}_r$. If $n/N \to 0$, then the variation of $\hat{\boldsymbol{\beta}}_{wf}$ can be ignored as stated in (11).

Now we compare the estimation efficiency of $\hat{\boldsymbol{\beta}}_r$ with that of the weighted estimator $\hat{\boldsymbol{\beta}}_w$. With the optimal subsampling probabilities $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_f)$, the asymptotic variance-covariance matrix (scaled by n), $\mathbf{V}_N$, for the weighted estimator $\hat{\boldsymbol{\beta}}_w$ has a form of $\mathbf{V}_N^{\mathrm{OS}} = \mathbf{M}_N^{-1}(\hat{\boldsymbol{\beta}}_f)\mathbf{V}_{Nc}^{\mathrm{OS}}\mathbf{M}_N^{-1}(\hat{\boldsymbol{\beta}}_f)$, in which

$$\mathbf{V}_{Nc}^{\mathrm{OS}} = \bigg\{\frac{1}{N}\sum_{i=1}^{N}|y_i - p_i(\hat{\boldsymbol{\beta}}_f)|h(\mathbf{x}_i)\bigg\}\bigg\{\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - p_i(\hat{\boldsymbol{\beta}}_f)|\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{h(\mathbf{x}_i)}\bigg\}.$$

Note that the full data MLE $\hat{\boldsymbol{\beta}}_f$ is consistent under Assumptions 1-2. If $\mathbb{E}\{\|\mathbf{x}\|^2/h(\mathbf{x})\} < \infty$, then from Lemma 1 in the appendix and the law of large numbers, $\mathbf{V}_N^{\mathrm{OS}}$ converges in probability to $\mathbf{V}^{\mathrm{OS}} = \mathbf{M}^{-1}\mathbf{V}_c^{\mathrm{OS}}\mathbf{M}^{-1}$, where

$$\mathbf{M} = \mathbb{E}\{\phi(\boldsymbol{\beta}_t)\mathbf{x}\mathbf{x}^{\mathrm{T}}\} \quad \text{and} \quad \mathbf{V}_c^{\mathrm{OS}} = 4\Phi(\boldsymbol{\beta}_t)\mathbb{E}\bigg\{\frac{\phi(\boldsymbol{\beta}_t)\mathbf{x}\mathbf{x}^{\mathrm{T}}}{h(\mathbf{x})}\bigg\}.$$

Note that the asymptotically distribution of $\hat{\boldsymbol{\beta}}_w$ given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$ is centered around $\hat{\boldsymbol{\beta}}_f$. It can be shown that under Assumptions 1-2,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}\big(\mathbf{0}, \mathbf{M}^{-1}\big). \tag{18}$$

Thus, both $n^{-1}\mathbf{V}^{\mathrm{OS}}$ and $N^{-1}\mathbf{M}^{-1}$ should be considered in accessing the quality of $\hat{\boldsymbol{\beta}}_w$ for estimating the true parameter $\boldsymbol{\beta}_t$. However, similar to the case for $\hat{\boldsymbol{\beta}}_r$, $N^{-1}\mathbf{M}^{-1}$ is small compared with $n^{-1}\mathbf{V}^{\mathrm{OS}}$ if $n \ll N$, and it is negligible if $n/N \to 0$. Therefore, the relative performance between $\hat{\boldsymbol{\beta}}_r$ and $\hat{\boldsymbol{\beta}}_w$ are mainly determined by the relative magnitude between $\mathbf{V}^{\mathrm{OS}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$. We have the following result comparing $\mathbf{V}^{\mathrm{OS}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$.

Proposition 1. *Suppose that* $\mathbf{M}$, $\mathbf{V}_c^{\mathrm{OS}}$, *and* $\mathbf{\Sigma}_{\boldsymbol{\beta}_t}$ *are finite and positive definite matrices. We have that*

$$\mathbf{\Sigma}_{\boldsymbol{\beta}_t} \leq \mathbf{V}^{\mathrm{OS}}, \tag{19}$$

where the inequality is in the Loewner ordering, i.e., for positive semi-definite matrices $\mathbf{A}$ *and* $\mathbf{B}$, $\mathbf{A} \geq \mathbf{B}$ *if and only if* $\mathbf{A} - \mathbf{B}$ *is positive semi-definite. If* $h(\mathbf{x}) = 1$, *then the equality in* (19) *holds.*

This proposition shows that $\hat{\boldsymbol{\beta}}_r$ is typically more efficient than $\hat{\boldsymbol{\beta}}_w$ in estimating $\boldsymbol{\beta}_t$. The numerical results in Section 7 also confirm this. Note that (19) holds if $h(\mathbf{x}) = 1$. This indicates that for subsample obtained from local case-control subsampling with replacement, the weighted and unweighted estimators have the same conditional asymptotic distribution.

4 Poisson sampling

For the more efficient estimator $\hat{\boldsymbol{\beta}}_r$ in Section 3 as well as the weighted estimator $\hat{\boldsymbol{\beta}}_w$ in Wang et al. (2017), the subsampling procedure used is sampling with replacement, which is faster to compute than sampling without replacement for a fixed sample size. In addition, the resultant subsample are independent and identically distributed (i.i.d.) conditional on the full data. However, to implement sampling with replacement, subsampling probabilities $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$ need to be calculated all at once, and a large amount of random numbers need to be generated all at once. This may reduce the computational efficiency, and it may require a large RAM to implement the method. Furthermore, since a data point may be included multiple times in the subsample, the resultant estimator may not be the most efficient.

To enhance the computation and estimation efficiency of the subsample estimator, we consider Poisson sampling, which is also fast to compute and the resultant subsample can be independent without conditioning on the full data. Note that for subsampling with replacement, a resultant subsample is in general not independent, although it is i.i.d conditional on the full data. As another advantage with Poisson sampling, there is no need to calculate subsampling probabilities all at once, nor to generate a large amount of random numbers all at once. Furthermore, a data point cannot be included in the subsample for more than one time. A limitation of Poisson sampling is that the subsample size is always random. Due to this, we abuse the notation in this section and use n to denote the expected subsample size.

Note that $\pi_i^{\mathrm{OS}}(\boldsymbol{\beta})$ depend on the full data through the term in the denominator, $\sum_{i=1}^N |y_i - p_i(\boldsymbol{\beta})|h(\mathbf{x}_i)$. Write $\Psi_N(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^N |y_i - p_i(\boldsymbol{\beta})|h(\mathbf{x}_i)$, and denote its limit as $\Psi(\boldsymbol{\beta}) = \mathbb{E}\{|y - p(\boldsymbol{\beta})|h(\mathbf{x})\}$. Note that $\Psi(\boldsymbol{\beta}_t) = 2\Phi(\boldsymbol{\beta}_t)$. The pilot sample can be used to obtain an estimator of $\Psi(\boldsymbol{\beta}_t)$ to approximate $\Psi_N(\boldsymbol{\beta})$. Let $\hat{\Psi}_1$ be a pilot estimator of $\Psi(\boldsymbol{\beta}_t)$. Here, we focus on the Poisson sampling procedure and assume that such $\hat{\Psi}_1$ is available and consistent. We will provide more details on this in the next section.

With $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$ available, the Poisson subsampling procedure is described as the following. For $i = 1, ..., N$, calculate $\pi_i^p = |y_i - p_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)/(N\hat{\Psi}_1)$, generate $u_i \sim U(0,1)$, and include $(\mathbf{x}_i, y_i, \pi_i^p)$ in the subsample if $u_i \leq n\pi_i^p$. For the obtained subsample, say $\{(\mathbf{x}_1^*, y_1^*, \pi_1^{p*}), ..., (\mathbf{x}_{n^*}^*, y_{n^*}^*, \pi_{n^*}^{p*})\}$, calculate

$$\tilde{\boldsymbol{\beta}}_p = \arg\max_{\boldsymbol{\beta}} \ell_p^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n^*} (n\pi_i^{p*} \vee 1)\big\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* y_i^* + \log(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*})\big\}, \tag{20}$$

and let $\hat{\boldsymbol{\beta}}_p = \tilde{\boldsymbol{\beta}}_p + \hat{\boldsymbol{\beta}}_1$. Note that here the actual subsample size n^* is random.

Poisson sampling does not require to calculate π_i^p's all at once; each π_i^p can be calculated on the go for each individual data point when scanning through the full data. Thus, one pass through the data finishes the sampling. For the estimation step, if π_i^p is large so that $n\pi_i^p > 1$, then this more informative data point will be given a larger weight, $n\pi_i^p$, in the log-likelihood in (20). The following theorem describes asymptotic properties of $\hat{\boldsymbol{\beta}}_p$.

Theorem 2. *Under assumptions 1-2, conditional on $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$, as $n \to \infty$ and $N \to \infty$, if $n/N \to 0$*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}), \tag{21}$$

in distribution; if $n/N \to \rho \in (0,1)$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \hat{\boldsymbol{\beta}}_{wf}) \longrightarrow \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_\rho\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}), \tag{22}$$

in distribution, where

$$\boldsymbol{\Lambda}_\rho = \frac{\mathbb{E}\big[\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\{\Phi(\boldsymbol{\beta}_t) - \rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\}_+\mathbf{x}\mathbf{x}^{\mathrm{T}}\big]}{4\Phi^2(\boldsymbol{\beta}_t)}, \tag{23}$$

and $()_+$ means the positive part of the quantity, i.e., $a_+ = aI(a > 0)$.

Remark 2. *Similar to the case of Theorem 1, Theorem 2 implies that $\hat{\boldsymbol{\beta}}_p$ is $\sqrt{n}$-consistent to $\boldsymbol{\beta}_t$ under the unconditional distribution.*

Theorem 2 shows that with Poisson sampling, the asymptotic variance-covariance matrices may differ for different sampling ratios n/N. In addition, comparing Theorems 1 and 2, we know that $\hat{\boldsymbol{\beta}}_r$ and $\hat{\boldsymbol{\beta}}_p$ have the same asymptotic distribution if $n/N \to 0$. This is intuitive because if the sampling rate n/N is small, sampling with replacement is not very different from sampling without replacement. However, if the sampling rate n/N does not converge to zero, $\hat{\boldsymbol{\beta}}_r$ and $\hat{\boldsymbol{\beta}}_p$ have the same asymptotic bias but different asymptotic variance-covariance matrices. The following result compares the two asymptotic variance covariance matrices.

Proposition 2. *If $\mathbf{\Sigma}_{\boldsymbol{\beta}_t}$ is a finite and positive definite matrix and $\rho > 0$, then*

$$\mathbf{\Sigma}_{\boldsymbol{\beta}_t}\mathbf{\Lambda}_\rho\mathbf{\Sigma}_{\boldsymbol{\beta}_t} < \mathbf{\Sigma}_{\boldsymbol{\beta}_t}, \tag{24}$$

under the Loewner ordering.

This proposition shows that Poisson sampling is more efficient than sampling with replacement if the expected subsample size is proportional to the full data size.

5 Pilot estimate and practical implementation

Since $\pi_i^{\mathrm{OS}}(\boldsymbol{\beta})$ depend on the unknown $\boldsymbol{\beta}$, a pilot estimate of $\boldsymbol{\beta}$ is required to approximate them. The pilot estimate can be obtained by taking a pilot sample using uniform subsampling or case-control subsampling (Wang et al., 2017). For uniform subsampling, all subsampling probabilities are equal, while for case-control subsampling, the subsampling probability for the cases ($y_i = 1$) are different from that for the controls ($y_i = 0$). Let the subsampling probabilities used to take the pilot sample be

$$\pi_{1i} = \frac{c_0(1-y_i) + c_1 y_i}{N}, \tag{25}$$

where c_0 and c_1 are two constants that can be used to balance the numbers of 0's and 1's in the responses for the pilot subsample. If $c_0 = c_1 = 1$, then $\pi_{1i} = N^{-1}$ are the uniform subsampling probability. This choice is recommended due to its simplicity if the proportion of 1's is close to 0.5 (Wang et al., 2017). If $c_0 \neq c_1$, then π_{1i} are the case control subsampling probabilities. This choice is recommended for imbalanced full data. Often, some prior information about the marginal probability $\mathbb{P}(y = 1)$ is available. If p_{pr} is the prior marginal probability, we can choose $c_0 = \{2(1 - p_{pr})\}^{-1}$ and $c_1 = (2p_{pr})^{-1}$. The pilot estimate $\hat{\boldsymbol{\beta}}_1$ can be obtained using the pilot subsample. For uniform subsampling, weighted and unweighted estimators are the same. For case-control subsampling, we use the unweighted estimators for both sampling with replacement and Poisson sampling.

In Wang et al. (2017), the pilot sample are combined with the second stage sample taken using approximated optimal subsampling probabilities to obtain a final estimator. While this does not make a difference asymptotically since n_1 is typically a small term compared with n, i.e., $n_1 = o(n)$, using the pilot sample helps to improve the finite sample performance in practical application. However, combing the raw samples may not be the most computationally efficient way of utilizing the pilot sample. Since $\hat{\boldsymbol{\beta}}_1$ is already calculated, we can use it directly to improve the second stage estimator using a divide-and-conquer method. This avoids iterative calculation on the pilot sample for the second time.

For Poisson subsampling, the pilot sample can also be used to construct $\hat{\Psi}_1$ to approximate $\Psi_N(\boldsymbol{\beta})$. We use the following expression to obtain $\hat{\Psi}_1$.

$$\hat{\Psi}_1 = \frac{1}{N}\sum_{i=1}^{n_1^*} \frac{|y_i^* - p_i^*(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i^*)}{n\pi_{1i}^* \wedge 1}, \tag{26}$$

where $p_i^*(\boldsymbol{\beta}) = e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*}/(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*})$. It can be verified that $\hat{\Psi}_1$ defined in (26) converges in probability to $\Psi(\boldsymbol{\beta}_t)$.

Another practical issue one has to consider is whether the full data can be loaded into available RAM or not. If so, subsampling probabilities can be calculated in RAM and subsampling with replacement can be implemented directly. Otherwise, special considerations have to be given in practical implementation. If the full data is larger than available RAM while subsampling probabilities $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$ can still be loaded in available RAM, one can calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$ by scanning the data from hard drive line-by-line or block-by-block, generate row indexes for a subsample, and then scan the data line-by-line or block-by-block to take the subsample. To be specific, one can draw a subsample, say $\{idx_1, ..., idx_n\}$, from $\{1, ..., N\}$, sort the indexes to have $\{idx_{(1)}, ..., idx_{(n)}\}$, and then use the following algorithm to scan the data line-by-line or block-by-block in order to obtain the subsample.

Algorithm 1 Obtain subsample with given indexes by scanning through the full data

Input: data file, subsample indexes $\{idx_{(1)}, ..., idx_{(n)}\}$.
$i \leftarrow 1$
$j \leftarrow 1$
while $i \leq N$ **and** $j \leq n$ **do**
  readline(data file)
  if $i == idx_{(j)}$ **then**
    include the ith data point into the subsample
    while $i == idx_{(j)}$ **do**
      $j \leftarrow j + 1$
    end while
  end if
  $i \leftarrow i + 1$
end while

Remark 3. *Clearly, Algorithm 1 takes no more than linear time to run. We assume that a generic function* `readline()` *reads a single line (or multiple lines) from the data file and stop at the beginning of the next line (or next block) in the data file. No calculation is performed* *on a data line if it is not included in the subsample. Such functionality are provided by most programming languages. For example, Julia (Bezanson et al., 2017) and Python (van Rossum, 1995) has a function* `readline()` *that read a file line-by-line; R (R Core Team, 2017) has a function* `readLines()` *that read one or multiple lines; C (Kernighan and Ritchie, 1988) and C++ (Stroustrup, 1986) has a function* `getline()` *to read one line at a time.*

Taking into account all aforementioned issues in this section, including how to obtain the pilot estimates, how to combine it with the second stage estimates, as well as how to process data file line-by-line, we summarize practical implementation procedures in Algorithm 2 and Algorithm 3, where Algorithm 2 is for sampling with replacement and Algorithm 3 is for Poisson sampling. The algorithms are presented for the scenario that data volume is larger than the size the the RAM. When the full data can be loaded into the RAM, the procedure of sampling can be done directly in the RAM and the method is easier to implement for that scenario.

Remark 4. *In Algorithm 2 and Algorithm 3, if $n_1 = o(n)$, then the result for $\hat{\boldsymbol{\beta}}_r$ in Theorem 1 hold for $\check{\boldsymbol{\beta}}_r$ and the result for $\hat{\boldsymbol{\beta}}_p$ in Theorem 2 hold for $\check{\boldsymbol{\beta}}_p$ as well. This is because $\{\ddot{\ell}_r^{*_1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_r)\}^{-1}\ddot{\ell}_r^{*_1}(\tilde{\boldsymbol{\beta}}_1)\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_t) = O_p(\sqrt{n_1}/\sqrt{n}) = o_P(1)$ and $\{\ddot{\ell}_r^{*_1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_r)\}^{-1}\ddot{\ell}_r(\tilde{\boldsymbol{\beta}}_r) \to 1$ in probability. The reason for $\check{\boldsymbol{\beta}}_p$ is similar.*

Remark 5. *In Algorithm 2 and Algorithm 3, to combine the two stage estimates using the second derivative of the likelihood, the inconsistent estimators $\tilde{\boldsymbol{\beta}}_1$, and $\tilde{\boldsymbol{\beta}}_r$ or $\tilde{\boldsymbol{\beta}}_p$ should be used, because their limits correspond to the terms in the asymptotic variance-covariance matrices of the more efficient estimators. This is an advantage of the proposed estimators for implementation using existing software that fit logistic regression. One can use the inverse of the estimated variance-covariance matrix from the software output to replace the second derivative of the likelihood.*

Remark 6. *The variance-covariance estimators $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_r)$ in (31) and $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p)$ in (35) can be replaced the following simplified estimators,*

$$\hat{\mathbb{V}}_s(\check{\boldsymbol{\beta}}_p) = \{\ddot{\ell}_p^{*_1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_p^{*}(\tilde{\boldsymbol{\beta}}_p)\}^{-1} \qquad \text{and} \qquad \hat{\mathbb{V}}_s(\check{\boldsymbol{\beta}}_r) = \{\ddot{\ell}_r^{*_1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_r)\}^{-1}, \tag{27}$$

respectively. If the subsampling ratio n/N is much smaller than one, then $\hat{\mathbb{V}}_s(\check{\boldsymbol{\beta}}_r)$ and $\hat{\mathbb{V}}_s(\check{\boldsymbol{\beta}}_p)$ perform very similarly $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_r)$ and $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p)$, respectively.

Remark 7. *The time complexity of Algorithm 2 is the same as that of Algorithm 2 in Wang et al. (2017). The major computing time is to calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$ in Step 2, but it does not require iterative calculations on the full data. Once $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$ are available, it takes*

$O(N)$ time to obtain the subsample using Algorithm 1 and the calculations of $\hat{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_{uw}$ are fast because they are done on the subsamples only. To calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$, the required time varies. For π_i^{mVc}, the required time is $O(Nd)$; for π_i^{mMSE}, the required time is $O(Nd^2)$. Thus, the time complexity of Algorithm 2 with π_i^{mVc} is $O(Nd)$ and the time complexity with π_i^{mMSE} is $O(Nd^2)$, if the sampling ratio n/N is much smaller than one.

Algorithm 2 More efficient estimation based on subsampling with replacement

Step 1: obtain the pilot $\hat{\boldsymbol{\beta}}_1$

(1) Take pilot subsample indexes of size n_1 from $\{1, ..., N\}$ using sampling with replacement according to subsampling probabilities π_{1i} in (25), and use Algorithm 1 to obtain a pilot subsample, say, $(\mathbf{x}_i^{*1}, y_i^{*1})$, $i = 1, ..., n_1$.

(2) Calculate

$$\tilde{\boldsymbol{\beta}}_1 = \arg\max_{\boldsymbol{\beta}} \ell_r^{*1}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n_1} \left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^{*1} y_i^{*1} - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^{*1}}\right)\right\}, \tag{28}$$

and let $\hat{\boldsymbol{\beta}}_1 = \tilde{\boldsymbol{\beta}}_1 + \mathbf{b}$, where $\mathbf{b} = \{\log(c_0/c_1), 0, ..., 0\}^{\mathrm{T}}$.

Step 2: obtain the more efficient estimator $\hat{\boldsymbol{\beta}}_r$

(1) Calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$ defined in equation (6); take subsample indexes of size n from $\{1, ..., N\}$ according to sampling probabilities $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_1)\}_{i=1}^N$ using sampling with replacement; and use Algorithm 1 to obtain a subsample, $(\mathbf{x}_i^*, y_i^*)$, $i = 1, ..., n$.

(2) Calculate

$$\tilde{\boldsymbol{\beta}}_r = \arg\max_{\boldsymbol{\beta}} \ell_r^{*}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^{*} y_i^{*} - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i}\right)\right\}, \tag{29}$$

and let $\hat{\boldsymbol{\beta}}_r = \tilde{\boldsymbol{\beta}}_r + \hat{\boldsymbol{\beta}}_1$.

Step 3: combine the two estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_r$

Calculate

$$\check{\boldsymbol{\beta}}_r = \{\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_r)\}^{-1}\{\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1)\hat{\boldsymbol{\beta}}_1 + \ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_r)\hat{\boldsymbol{\beta}}_r\} \tag{30}$$

where $\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1) = \sum_{i=1}^{n_1} \phi^{*1}(\tilde{\boldsymbol{\beta}}_1)\mathbf{x}_i^{*1}(\mathbf{x}_i^{*1})^{\mathrm{T}}$ and $\ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_r) = \sum_{i=1}^{n} \phi^{*}(\tilde{\boldsymbol{\beta}}_r)\mathbf{x}_i^{*}(\mathbf{x}_i^{*})^{\mathrm{T}}$.

The variance-covariance matrix of $\check{\boldsymbol{\beta}}_r$ can be estimated by

$$\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_r) = \{\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_r)\}^{-1}\left[\sum_{i=1}^{n_1}\{\psi_i^{*1}(\tilde{\boldsymbol{\beta}}_1)\}^2\mathbf{x}_i^{*1}(\mathbf{x}_i^{*1})^{\mathrm{T}}\times \right.$$
$$\left. + \sum_{i=1}^{n}\{\psi_i^{*}(\tilde{\boldsymbol{\beta}}_r)\}^2\mathbf{x}_i^{*}(\mathbf{x}_i^{*})^{\mathrm{T}}\right]\{\ddot{\ell}_r^{*1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_r^{*}(\tilde{\boldsymbol{\beta}}_r)\}^{-1}, \tag{31}$$

where $\psi_i^*(\boldsymbol{\beta}) = y_i^* - p_i^*(\boldsymbol{\beta})$.

Algorithm 3 More efficient estimation based on Poisson sampling

Step 1: obtain the pilots $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$

(1) For $i = 1, ..., N$, calculate $\pi_{1i} = \frac{c_0(1-y_i)+c_1 y_i}{N}$, generate $u_{1i} \sim U(0,1)$, and add $(\mathbf{x}_i, y_i, \pi_{i1})$ in the subsample if $u_{1i} \leq n_1 \pi_{1i}$.

(2) For the obtained subsample, say $(\mathbf{x}_i^{*_1}, y_i^{*_1}, \pi_{1i}^{*_1})$, $i = 1, ..., n_1^*$, calculate

$$\tilde{\boldsymbol{\beta}}_1 = \arg\max_{\boldsymbol{\beta}} \ell_p^{*_1}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n_1^*} (n\pi_{1i}^{*_1} \vee 1)\big\{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^{*_1} y_i^{*_1} + \log\big(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^{*_1}}\big)\big\}, \tag{32}$$

let $\hat{\boldsymbol{\beta}}_1 = \tilde{\boldsymbol{\beta}}_1 + \mathbf{b}$, and then calculate $\hat{\Psi}_1$ in equation (26).

Step 2: obtain the more efficient estimator $\hat{\boldsymbol{\beta}}_p$

(1) For $i = 1, ..., N$, calculate $\pi_i^p = \frac{|y_i - p_i(\hat{\boldsymbol{\beta}}_1)| h(\mathbf{x}_i)}{N \hat{\Psi}_1}$, generate $u_i \sim U(0,1)$, and if $u_i \leq n\pi_i^p$ add $(\mathbf{x}_i, y_i, \pi_i^p)$ in the subsample.

(2) For the obtained subsample, say $\{(\mathbf{x}_1^*, y_1^*, \pi_1^{p*}), ..., (\mathbf{x}_{n^*}^*, y_{n^*}^*, \pi_{n^*}^{p*})\}$, calculate

$$\tilde{\boldsymbol{\beta}}_p = \arg\max_{\boldsymbol{\beta}} \ell_p^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n^*} (n\pi_i^{p*} \vee 1)\big\{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* y_i^* + \log(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^*})\big\}, \tag{33}$$

and let $\hat{\boldsymbol{\beta}}_p = \tilde{\boldsymbol{\beta}}_p + \hat{\boldsymbol{\beta}}_1$.

Step 3: combine the two estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_p$

Calculate

$$\check{\boldsymbol{\beta}}_p = \{\ddot{\ell}_p^{*_1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p)\}^{-1} \{\ddot{\ell}_p^{*_1}(\tilde{\boldsymbol{\beta}}_1)\hat{\boldsymbol{\beta}}_1 + \ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p)\hat{\boldsymbol{\beta}}_p\} \tag{34}$$

where $\ddot{\ell}_p^{*_1}(\tilde{\boldsymbol{\beta}}_1) = \sum_{i=1}^{n_1^*} \phi^{*_1}(\tilde{\boldsymbol{\beta}}_1) \mathbf{x}_i^{*_1} (\mathbf{x}_i^{*_1})^{\mathrm{T}}$ and $\ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p) = \sum_{i=1}^{n^*} \phi^*(\tilde{\boldsymbol{\beta}}_p) \mathbf{x}_i^* (\mathbf{x}_i^*)^{\mathrm{T}}$.

The variance-covariance matrix of $\check{\boldsymbol{\beta}}_p$ can be estimated by

$$\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p) = \{\ddot{\ell}_p^{*_1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p)\}^{-1} \Bigg[\sum_{i=1}^{n_1^*} \{\psi_i^{*_1}(\tilde{\boldsymbol{\beta}}_1)\}^2 \mathbf{x}_i^{*_1} (\mathbf{x}_i^{*_1})^{\mathrm{T}} \times$$
$$+ \sum_{i=1}^{n^*} \{\psi_i^*(\tilde{\boldsymbol{\beta}}_p)\}^2 \mathbf{x}_i^* (\mathbf{x}_i^*)^{\mathrm{T}}\Bigg] \{\ddot{\ell}_p^{*_1}(\tilde{\boldsymbol{\beta}}_1) + \ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p)\}^{-1}. \tag{35}$$

6 Unconditional distribution

Asymptotic distributional results in Sections 3 and 4, as well as in Wang et al. (2017), are about conditional distributions, i.e., they are about conditional distributions of subsample-based estimators given the full data and the pilot estimate. We investigate the unconditional distribution of $\hat{\boldsymbol{\beta}}_p$ in this section.

Theorem 3. *Assume that Assumptions 1 and 2 hold. If the pilot estimators are obtained with subsampling probabilities in* (25) *and* $\mathbb{E}\{h^3(\mathbf{x})\|\mathbf{x}\|^3\}$, $\mathbb{E}\{h^3(\mathbf{x})\|\mathbf{x}\|^2\}$, $\mathbb{E}\{h(\mathbf{x})\|\mathbf{x}\|^3\}$, *and* $\mathbb{E}\{h^2(\mathbf{x})\}$ *are finite, or if* $\hat{\boldsymbol{\beta}}_1$ *and* $\hat{\Psi}_1$ *are independent of the data, then the following results hold.*

As $n \to \infty$ *and* $N \to \infty$, *if* $n/N \to 0$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}), \tag{36}$$

in distribution; if $n/N \to \rho \in (0, 1)$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_u\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}), \tag{37}$$

in distribution, where

$$\boldsymbol{\Lambda}_u = \frac{\mathbb{E}[\phi(\boldsymbol{\beta}_t)\{\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x}) \vee \Phi(\boldsymbol{\beta}_t)\}h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}]}{4\Phi^2(\boldsymbol{\beta}_t)}. \tag{38}$$

Remark 8. *If the pilot estimators* $\hat{\boldsymbol{\beta}}_1$ *and* $\hat{\Psi}_1$ *are obtained through the data, stronger moment conditions are required. Note that* $h(\mathbf{x})$ *is often a function of the norm of* $\mathbf{x}$, *such as in* π_i^{mVc}, π_i^{mMSE}, *and the local case control subsampling. In general, if* $h(\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|^a$ *for some constant matrix* $\mathbf{A}$ *and constant* $a \geq 0$, *then the four additional moment conditions reduce to one requirement of* $\mathbb{E}\{h^3(\mathbf{x})\|\mathbf{x}\|^3\} < \infty$.

Fithian and Hastie (2014) obtained unconditional distribution of local case-control estimator by assuming that the pilot estimate is independent of the data. Our Theorem 3 includes this scenario, and the required assumptions are the same as those required in Fithian and Hastie (2014). In practice, a consistent pilot estimator that is independent of the data may not be available and a pilot subsample from the full data is required to construct it. For this scenario, a pilot estimator is dependent on the data, and we need a stronger moment condition to establish the asymptotic normality. For local case-control subsampling, $h(\mathbf{x}) = 1$, so the additional moment requirement is that $\mathbb{E}(\|\mathbf{x}\|^3) < \infty$.

From Theorems 2 and 3, the conditional asymptotic distribution and unconditional asymptotic distribution of $\hat{\boldsymbol{\beta}}_p$ are the same if $n/N \to 0$. This is intuitive, because if the sampling rate n/N is small, the variation of $\hat{\boldsymbol{\beta}}_p$ due to the variation of the full data is small compared with the variation due to the variation of the subsampling.

However, if the sampling rate n/N does not converge to zero, then the conditional asymptotic distribution and unconditional asymptotic distribution of $\hat{\boldsymbol{\beta}}_p$ are quite different. First, we notice that under the unconditional distribution, $\hat{\boldsymbol{\beta}}_p$ is asymptotically unbiased to $\boldsymbol{\beta}_t$, while under the conditional distribution, $\hat{\boldsymbol{\beta}}_p$ is asymptotically biased with the bias being $\hat{\boldsymbol{\beta}}_{wf} - \boldsymbol{\beta}_t = O_P(N^{-1/2})$. Second, since the variation of $\hat{\boldsymbol{\beta}}_p$ due to the variation of the full data is not negligible, we expect that the asymptotic variance covariance matrices for the unconditional distribution is larger than that for the conditional distribution. Indeed this is true, and we present it in the following proposition.

Proposition 3. *If $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ is a finite and positive definite matrix and $\rho > 0$, then*

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_u\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \geq \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} > \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_\rho\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}, \tag{39}$$

under the Loewner ordering. Furthermore, if $\mathbb{P}\{\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x}) > \Phi(\boldsymbol{\beta}_t)\} > 0$, then the "$\geq$" sign in (39) can be replaced by "$>$", the strict great sign.

7 Numerical evaluations

We evaluate the performance of the more efficient estimators in terms of both estimation efficiency and computational efficiency in this section.

7.1 Estimation efficiency

In this section, we use numerical experiments based on simulated and real data sets to evaluate the estimators proposed in this paper. For simulation, to compare with the original OSMAC estimator, we use exactly the same setup used in Section 5.1 of Wang et al. (2017). Specifically, the full data sample size $N = 10,000$ and the true value of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_t$, is a 7×1 vector of 0.5. The following 6 distributions of $\mathbf{x}$ are considered: multivariate normal distribution with mean zero (mzNormal), multivariate normal distribution with nonzero mean (nzNormal), multivariate normal distribution with mean zero and unequal variances (ueNormal), mixture of two multivariate normal distributions with different means (mixNormal), multivariate t distribution with degrees of freedom 3 (T_3), and exponential distribution (EXP). Detailed explanations of these distributions can be found in Section 5.1 of Wang et al. (2017).

To evaluate the estimation performance of the new estimators compared with the original weighted OSMAC estimator, $\check{\boldsymbol{\beta}}_w$, we define the estimation efficiency of $\check{\boldsymbol{\beta}}_{\text{new}}$ relative to $\check{\boldsymbol{\beta}}_w$

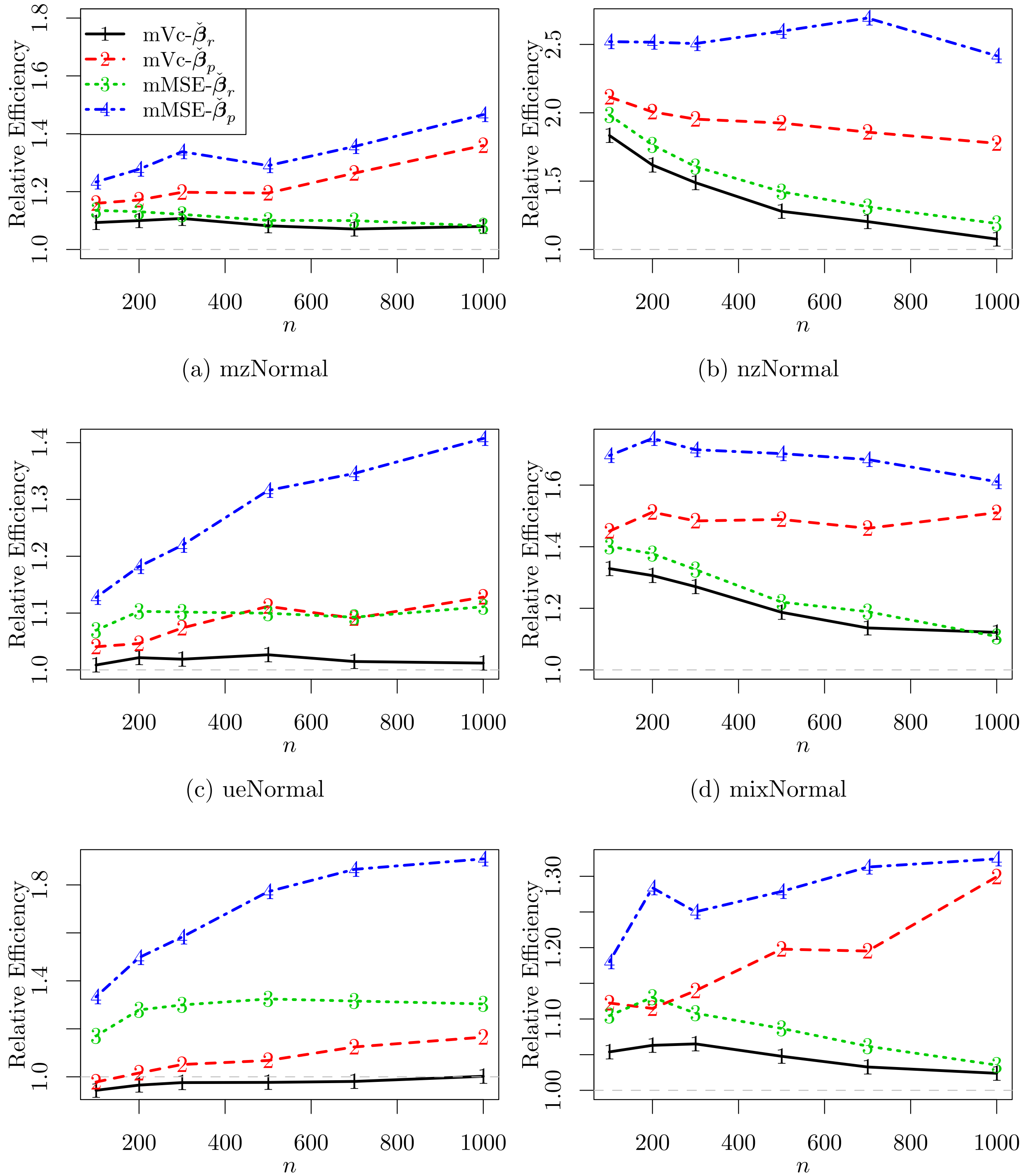
as

$$\text{Relative Efficiency} = \frac{\text{MSE}(\check{\boldsymbol{\beta}}_w)}{\text{MSE}(\check{\boldsymbol{\beta}}_{\text{new}})}, \tag{40}$$

where $\check{\boldsymbol{\beta}}_{\text{new}} = \check{\boldsymbol{\beta}}_r$ for the subsampling with replacement estimator described in Algorithm 2 and $\check{\boldsymbol{\beta}}_{\text{new}} = \check{\boldsymbol{\beta}}_p$ for Poisson subsampling estimator described in Algorithm 3. We calculate empirical MSEs from $S = 1000$ subsamples using $\text{MSE}(\check{\boldsymbol{\beta}}) = S^{-1}\sum_{s=1}^{S}\|\check{\boldsymbol{\beta}}^{(s)} - \boldsymbol{\beta}_t\|^2$, where $\check{\boldsymbol{\beta}}^{(s)}$ is the estimate from the sth subsample. We fixed the first step sample size $n_1 = 200$ and choose n to be 100, 200, 400, 600, 800, and 1000. This is the same setup used in Wang et al. (2017).

Figure 1 presents the relative efficiency of $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$ based on two different choices of π_i^{OS}: π_i^{mMSE} and π_i^{mVc}. It is seen that in general $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$ are more efficient than $\check{\boldsymbol{\beta}}_w$. Among the six cases, the only case for $\check{\boldsymbol{\beta}}_w$ to be more efficient is when $\mathbf{x}$ has a T_3 distribution and π^{mVc} is used, but the difference is not very significant. For all other cases, $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$ are more efficient. For example, when $\mathbf{x}$ has a the nzNormal distribution, $\check{\boldsymbol{\beta}}_p$ can be 250% as efficient as $\check{\boldsymbol{\beta}}_w$ if π^{mMSE} is used. Between $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$, $\check{\boldsymbol{\beta}}_p$ is more efficient than $\check{\boldsymbol{\beta}}_r$ for all cases. We also calculate the empirical unconditional MSE by generating the full data in each repetition of the simulation. The results are similar and thus are omitted.

To assess the performance of $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_r)$ in (31) and $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p)$ in (35), we use $\text{tr}\{\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_r)\}$ and $\text{tr}\{\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p)\}$ to estimate the MSEs of $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$, and compare the average estimated MSEs with the unconditional empirical MSEs. We focus on the unconditional MSE because conditional inference may be appropriate only if $n/N \to 0$. Figure 2 presents the results for using π^{mVc}. Results for using π^{mMSE} are similar and are omitted for clearer presentation. It is seen that the estimated MSEs are very close to the empirical MSEs, except for the case of nzNormal covariate for subsampling with replacement. For this case, the responses are imbalanced with about 95% being 1's. For this scenario, the variance-covariance estimator for $\check{\boldsymbol{\beta}}_w$ proposed in Wang et al. (2017) also has a similar problem of underestimation. For Poisson sampling, the problem of underestimation from $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p)$ is not significant.



(a) mzNormal

(b) nzNormal

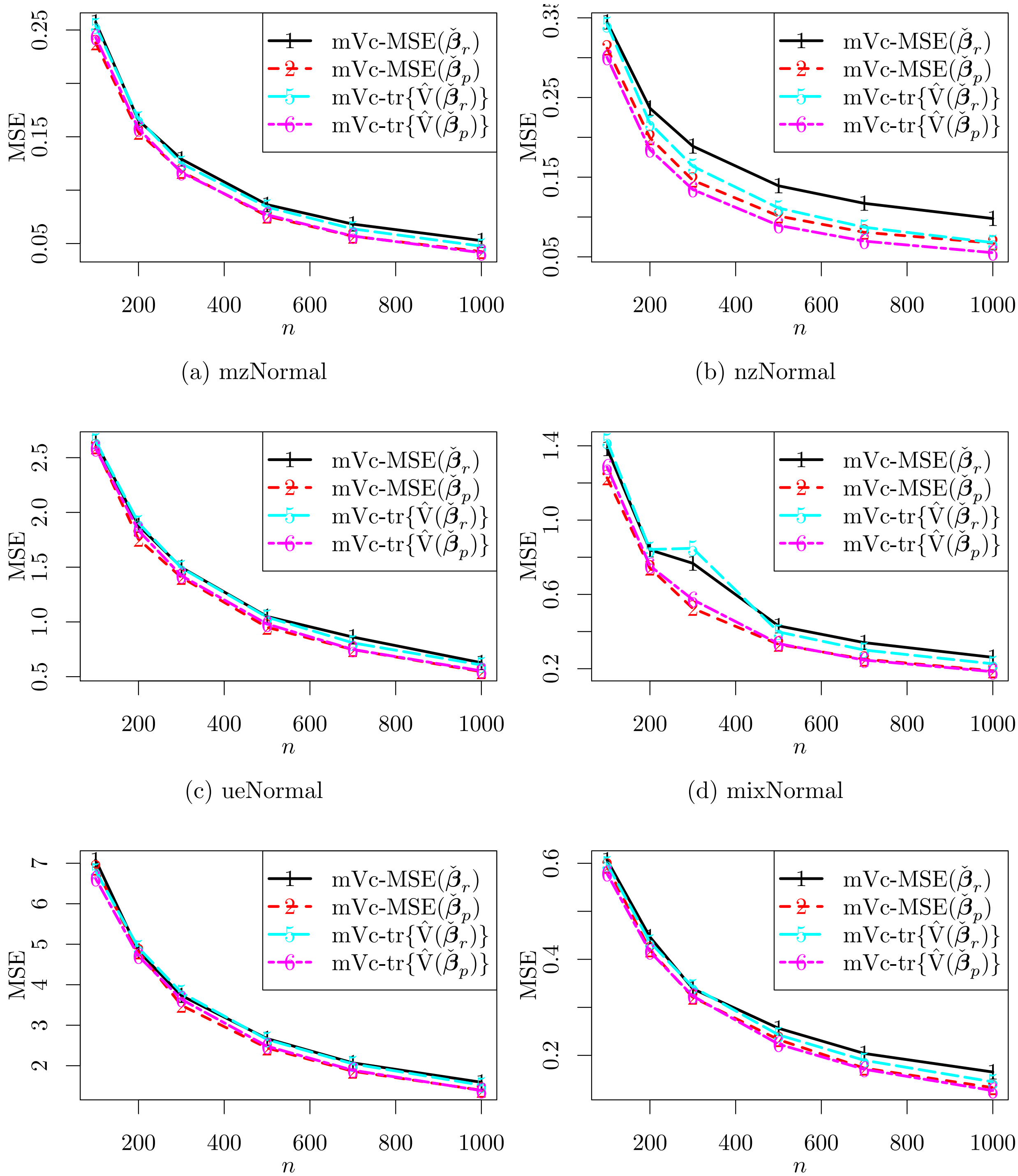
(c) ueNormal

(d) mixNormal

(e) T_3

(f) EXP

Figure 1: Relative efficiency for different second step subsample size n with the first step subsample size being fixed at $n_1 = 200$. A relative efficiency larger than one means the associate method is more efficient than the original OSMAC estimator $\hat{\boldsymbol{\beta}}_w$.



(a) mzNormal (b) nzNormal

(c) ueNormal (d) mixNormal

(e) T_3 (f) EXP

Figure 2: MSE and estimated MSE, $\text{tr}\{\hat{\mathbb{V}}(\breve{\boldsymbol{\beta}})\}$, for different second step subsample size n with the first step subsample size being fixed at $n_1 = 200$.

We also apply the more efficient estimation methods to a supersymmetric (SUSY) benchmark data set (Baldi et al., 2014) available from the Machine Learning Repository (Lichman, 2013). The data contains a binary response variable indicating whether a process produce new supersymmetric particles or not and 18 covariates that are kinematic features about the process. The full sample size is $N = 5,000,000$ and the data file is about 2.4 gigabytes. About 54.24% of the responses in the full data are from the background process. We use the more efficient estimation methods with subsample size n to estimate parameters in logistic regression.

Figures 3 gives the relative efficiency of $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$ to $\check{\boldsymbol{\beta}}_w$ for both π^{mVc} and π^{mMSE}. It is seen that when π^{mMSE} is used, $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$ always outperform $\check{\boldsymbol{\beta}}_w$. When π^{mVc} is used, $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$ may not be as efficient as $\check{\boldsymbol{\beta}}_w$, but they become more efficient when the second stage sample size n gets larger. It is also seen that $\check{\boldsymbol{\beta}}_p$ dominates $\check{\boldsymbol{\beta}}_r$ and π^{mMSE} dominates π^{mVc} in estimation efficiency.

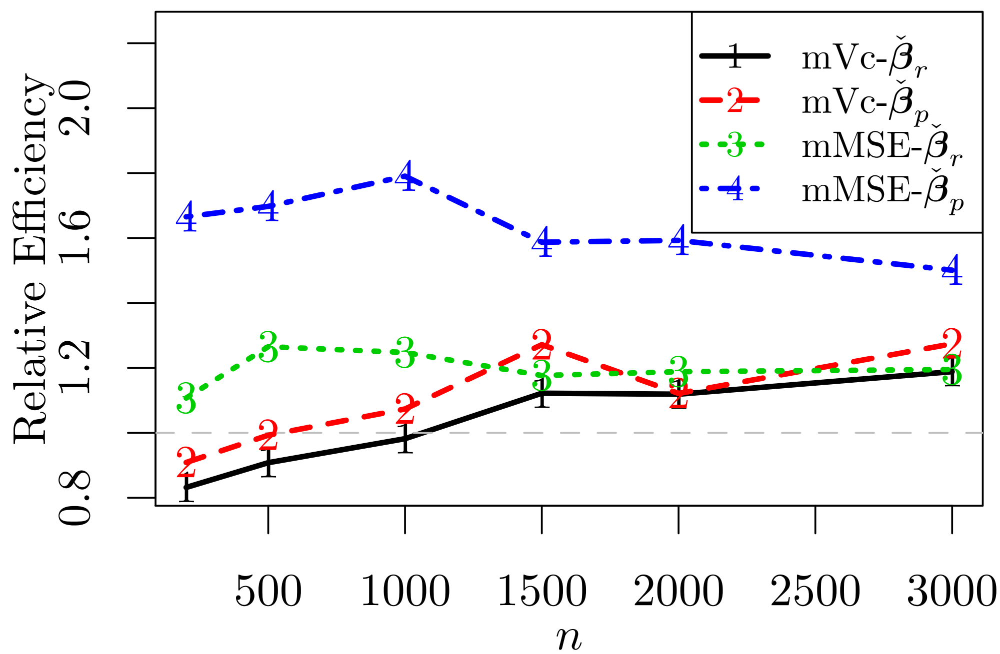


Figure 3: Relative efficiency for the SUSY data set with $n_1 = 200$ and different second step subsample size n. The gray horizontal dashed line is the reference line when relative efficiency is one.

7.2 Computational efficiency

We consider the computational efficiency of the more efficient estimation methods in this section. Note that they have the same order of computational time complicity, so they should have similar computational efficiency. For Poisson sampling, there is no need to calculate

$\{\pi_i^p\}_{i=1}^N$ all at once and random numbers can be generated on the go, so it requires less RAM and may require less CPU times as well. To confirm this, we record the computing time of implementing each of them for the case when $\mathbf{x}$ is mzNormal. All methods are implemented in the R programming language (R Core Team, 2017), and computations are carried out on a desktop running Ubuntu Linux 16.04 with an Intel I7 processor and 16GB RAM. Only one logical CPU is used for the calculation. We set the value of d to $d = 50$, the values of N to be $N = 10^4, 10^5, 10^6$ and 10^7, and the subsample sizes to be $n_1 = 200$ and $n = 1000$.

Table 1 gives the required CPU times (in seconds) to obtain $\check{\boldsymbol{\beta}}_w$, $\check{\boldsymbol{\beta}}_r$, and $\check{\boldsymbol{\beta}}_p$, using π^{mVc} and π^{mMSE}. The computing times for using the full data (Full) are also given for comparisons. It is seen that $\check{\boldsymbol{\beta}}_r$ and $\check{\boldsymbol{\beta}}_p$ are a little more computationally efficient than $\check{\boldsymbol{\beta}}_w$ but the advantages are not very significant. Note that these times are obtained when all the calculations are done in the RAM, and only the CPU times for implementing each method are counted while the time to generate the data is not counted.

Table 1: CPU seconds when the full data are generate and kept in the RAM. Here $n_1 = 200$, $n = 1000$, and the full data size N varies; the covariates are from a $d = 50$ dimensional multivariate normal distribution.

Method	N			
	10^4	10^5	10^6	10^7
mVc, $\check{\boldsymbol{\beta}}_w$	0.14	0.13	0.45	5.24
mVc, $\check{\boldsymbol{\beta}}_r$	0.08	0.11	0.41	3.71
mVc, $\check{\boldsymbol{\beta}}_p$	0.08	0.11	0.43	3.88
mMSE, $\check{\boldsymbol{\beta}}_w$	0.13	0.32	3.31	35.15
mMSE, $\check{\boldsymbol{\beta}}_r$	0.12	0.31	3.29	34.98
mMSE, $\check{\boldsymbol{\beta}}_p$	0.12	0.31	3.29	35.06
Full	0.15	1.62	15.05	247.89

For big data problem, it is common that the full data are larger than the size of the available RAM, and full data can not be loaded into the RAM. For this scenario, one has to load the data into RAM line-by-line or block-by-block. Note that communication between CPU and hard drive is much slower than communication between CPU and RAM. Thus, this will dramatically increase the computing time. To mimic this situation, we store the full data on hard drive and use `readlines()` function to process data 1000 rows each time. We also use a smaller computer with 8GB RAM to implement the method. For the case when $N = 10^7$, the full data is about 9.1GB which is larger than the available RAM.

The computing times when data are scanned from hard drive are reported Table 2. Here the computing times can be over thousand times longer than those when data are loaded into RAM. Note that these computing times can be reduced dramatically if we use some other programming language like C++ (Stroustrup, 1986) or Julia (Bezanson et al., 2017). However, for fair comparisons, we use the same programming language R here. Furthermore, our main purpose here is to demonstrate the computational advantage of subsampling so the real focus is on the relative performance among different methods. From Table 2, it is seen that using π^{mMSE} does not cost much more time than using π^{mVc}. The reason for this observation is that the major computing time is spent in data processing and the computing times used in calculating the subsampling probabilities are short. We also notice that Poisson sampling is more computational efficient than subsampling with replacement since it calculates subsampling probabilities and generates random numbers on the go and requires one time less to scan the full data. Poisson subsampling only used about 2% of the time required by implementing the full data approach.

Table 2: CPU seconds when the full data are scanned from hard drive. Here $n_1 = 200$, $n = 1000$, and the full data size N varies; the covariates are from a $d = 50$ dimensional multivariate normal distribution.

Method	N			
	10^4	10^5	10^6	10^7
mVc, $\breve{\boldsymbol{\beta}}_w$	4.26	41.60	441.46	4374.94
mVc, $\breve{\boldsymbol{\beta}}_r$	4.13	41.42	413.09	4384.99
mVc, $\breve{\boldsymbol{\beta}}_p$	2.77	27.58	272.32	2699.13
mMSE, $\breve{\boldsymbol{\beta}}_w$	4.43	41.75	434.96	4393.38
mMSE, $\breve{\boldsymbol{\beta}}_r$	4.10	41.83	417.55	4369.04
mMSE, $\breve{\boldsymbol{\beta}}_p$	2.88	27.93	273.24	2719.51
Full	139.46	1411.78	14829.63	138134.69

8 Summary

In this paper, we proposed a new unweighted estimator for logistic regression based on OSMAC subsample. We derived conditional asymptotic distribution of the new estimator which has a smaller variance-covariance matrix compared with the weighted estimator.

We also investigate the asymptotic properties if Poisson sampling is used, and showed that the resultant estimator has the same conditional asymptotic distribution if the subsampling rate converges to zero However, if the subsampling rate converges to a positive constant, the estimator based on Poisson sampling has a smaller variance-covariance matrix.

Furthermore, we also derive the unconditional asymptotic distribution for the proposed estimator based on Poisson sampling. Interestingly, if the subsampling rate converges to zero, the unconditional asymptotic distribution is the same as the conditional asymptotic distribution, indicating that the variation of the full data can be ignored. If the subsampling rate does not converge to zero, the unconditional asymptotic distribution has a larger variance-covariance matrix. Our results also include the case-control sampling method. With a stronger moment condition that the third moment of the covariate is finite, we do not require the pilot estimate to be independent of the data.

A Proofs and technical details

In this appendix, we provide proofs for the results in the paper. Technical details related to sampling with replacement in Section 3 are presented in Section A.1; technical details related to Poisson sampling in Section 4 are presented in Section A.2; and technical details related to unconditional results in Section 6 are presented in Section A.3.

A.1 Proofs for subsampling with replacement

In this section we prove the results in Section 3. The proofs require a series of Lemmas, which are presented below and will be proved later in this section.

For easy of presentation, we use notation λ to denote the log-likelihood shifted by $\hat{\boldsymbol{\beta}}_1$. For the subsample, $\lambda_r^*(\boldsymbol{\beta}) = \ell_r^*(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)$. Denote the first and second derivatives of $\lambda_r^*(\boldsymbol{\beta})$ as $\dot{\lambda}_r^*(\boldsymbol{\beta}) = \partial\lambda_r^*(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\ddot{\lambda}_r^*(\boldsymbol{\beta}) = \partial^2\lambda_r^*(\boldsymbol{\beta})/(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}})$.

Note that from Xiong and Li (2008), the fact that a sequence converges to 0 in conditional probability is equivalent to the fact that it converges to 0 in unconditional probability. Thus, in the following, we will not use $o_P(1)$ to denote a sequence converging to 0 in probability without stating whether the underlying probability measure is conditional or unconditional.

Lemma 1. *Let $\mathbf{v}_1, ..., \mathbf{v}_n$ be i.i.d. random vector with the same distribution of $\mathbf{v}$. Let g_1 be a bounded function that may depend on n and other random variables, and g_2 be a fixed function that does not depend on n. If $g_1(\mathbf{v}_i) = o_P(1)$ for each i as $n \to \infty$, and $\mathbb{E}|g_2(\mathbf{v})| < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} g_1(\mathbf{v}_i)g_2(\mathbf{v}_i) = o_P(1).$$

Lemma 2. *Let $\eta_i = |\psi_i(\hat{\boldsymbol{\beta}}_1)|\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)\mathbf{x}_i$, where $\psi_i(\boldsymbol{\beta}) = y_i - p_i(\boldsymbol{\beta})$. Under Assumptions 1 and 2, conditional on $\hat{\boldsymbol{\beta}}_1$,*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{wf} - \boldsymbol{\beta}_t) = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}}{2\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\}} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \eta_i + o_P(1), \tag{A.1}$$

which, as $N \to \infty$, converges in distribution to the a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1}\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}\}[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1}$.

Lemma 3. *Let*

$$\dot{\lambda}_r^*(\boldsymbol{\beta}_t) = \sum_{i=1}^{n} \big\{y_i^* - p_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\big\}\mathbf{x}_i^*. \tag{A.2}$$

Under Assumptions 1 and 2, conditional on $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$ which is assumed to be a consistent estimator of $\boldsymbol{\beta}_t$, as $n \to \infty$ and $N \to \infty$,

$$\frac{\dot{\lambda}_r^*(\boldsymbol{\beta}_t)}{\sqrt{n}} - \frac{\sqrt{n}\sum_{i=1}^{N}\eta_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_1)} \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}\big), \tag{A.3}$$

in distribution.

Lemma 4. *Under Assumptions 1-3, as $n, N \to \infty$, for any $\mathbf{s}_n \to 0$,*

$$\frac{1}{n}\sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \mathbf{s}_n)\|\mathbf{x}_i^*\|^2 - \sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_1)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\|\mathbf{x}_i\|^2 = o_P(1). \tag{A.4}$$

Proof of Theorem 1. The estimator $\hat{\boldsymbol{\beta}}_r$ is the maximizer of

$$\lambda_r^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)^{\mathrm{T}}\mathbf{x}_i^* y_i^* - \log\big\{1 + e^{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_1)^{\mathrm{T}}\mathbf{x}_i^*}\big\}\Big], \tag{A.5}$$

so $\sqrt{n}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_t)$ is the mazimizer of $\gamma(\mathbf{s}) = \lambda_r^*(\boldsymbol{\beta}_t + \mathbf{s}/\sqrt{n}) - \lambda_r^*(\boldsymbol{\beta}_t)$. By Taylor's expansion,

$$\gamma(\mathbf{s}) = \frac{1}{\sqrt{n}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_r^*(\boldsymbol{\beta}_t) + \frac{1}{2n}\sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i^*)^2, \tag{A.6}$$

where $\phi_i^*(\boldsymbol{\beta}) = p_i^*(\boldsymbol{\beta})\{1 - p_i^*(\boldsymbol{\beta})\}$, and $\acute{\mathbf{s}}$ lies between $\mathbf{0}$ and $\mathbf{s}$.

From Lemma 4,

$$\frac{1}{n}\sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i^*(\mathbf{x}_i^*)^{\mathrm{T}} - \sum_{i=1}^{N} \pi(\hat{\boldsymbol{\beta}}_1)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = o_P(1). \tag{A.7}$$

From Lemma 1 and the law of large numbers,

$$\sum_{i=1}^{N} \pi(\hat{\boldsymbol{\beta}}_1)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \frac{\frac{1}{N}\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\hat{\boldsymbol{\beta}}_1)} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} + o_P(1). \tag{A.8}$$

Combining the above two equations, we have that $n^{-1}\sum_{i=1}^{n}\phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i^*(\mathbf{x}_i^*)^{\mathrm{T}}$ converges in probability to $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}$, a positive definite matrix. In addition, from Lemma 2 and Lemma 3, conditional on $\mathcal{D}_N$, and $\hat{\boldsymbol{\beta}}_1$, $\dot{\lambda}_r^*(\boldsymbol{\beta}_t)/\sqrt{n} = O_{P|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1}(1)$. Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), the mazimizer of $\gamma(\mathbf{s})$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_t)$, satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_t) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\frac{1}{\sqrt{n}}\dot{\lambda}_r^*(\boldsymbol{\beta}_t) + O_{P|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1}(1) \tag{A.9}$$

given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$. Thus,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_r - \check{\boldsymbol{\beta}}_w) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\left\{\frac{1}{\sqrt{n}}\dot{\lambda}_r^*(\boldsymbol{\beta}_t) - \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}\sqrt{n}(\check{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_t)\right\} + O_{P|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1}(1). \tag{A.10}$$

From Lemma 2,

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}\sqrt{n}(\hat{\boldsymbol{\beta}}_{wf} - \boldsymbol{\beta}_t) = \frac{\sqrt{n}\sum_{i=1}^{N}\eta_i}{2N\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\}} = \frac{\sqrt{n}\sum_{i=1}^{N}\eta_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_1)} + o_P(1), \tag{A.11}$$

Combining equations (A.10) and (A.11), Lemma 3, and Slutsky's theorem, Theorem 1 follows. □

A.1.1 Proof of Proposition 1

Proof of Proposition 1. To prove that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \le \mathbf{V}^{\mathrm{OS}} = \mathbf{M}^{-1}\mathbf{V}_c^{\mathrm{OS}}\mathbf{M}^{-1}$, we just need to show that

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_f}^{-1} \ge \mathbf{M}(\mathbf{V}_c^{\mathrm{OS}})^{-1}\mathbf{M}. \tag{A.12}$$

From the strong law of large numbers,

$$\mathbf{M} = \frac{1}{N}\sum_{i=1}^{N}\phi_i(\boldsymbol{\beta}_t)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + o(1), \tag{A.13}$$

$$\mathbf{V}_c^{\mathrm{OS}} = 4\Phi(\boldsymbol{\beta}_t)\frac{1}{N}\sum_{i=1}^{N}\frac{\phi_i(\boldsymbol{\beta}_t)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{h(\mathbf{x}_i)} + o(1), \tag{A.14}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} = 4\Phi(\boldsymbol{\beta}_t)\left\{\frac{1}{N}\sum_{i=1}^{N}\phi_i(\boldsymbol{\beta}_t)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right\}^{-1} + o(1), \tag{A.15}$$

almost surely. Thus, we only need to verify that

$$\sum_{i=1}^{N}\phi_i(\boldsymbol{\beta}_t)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \ge \left\{\sum_{i=1}^{N}\phi_i(\boldsymbol{\beta}_t)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right\}\left\{\sum_{i=1}^{N}\frac{\phi_i(\boldsymbol{\beta}_t)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{h(\mathbf{x}_i)}\right\}^{-1}\left\{\sum_{i=1}^{N}\phi_i(\boldsymbol{\beta}_t)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right\}. \tag{A.16}$$

Denote $\mathbf{Z} = \{\sqrt{\phi_1(\boldsymbol{\beta}_t)}\mathbf{x}_1, ..., \sqrt{\phi_N(\boldsymbol{\beta}_t)}\mathbf{x}_N\}^{\mathrm{T}}$, and $\mathbf{H} = \mathrm{diag}\{h(\mathbf{x}_1), ..., h(\mathbf{x}_N)\}$. The inequality in (A.16) can be written as

$$\mathbf{Z}^{\mathrm{T}}\mathbf{H}\mathbf{Z} \geq \mathbf{Z}^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}, \tag{A.17}$$

which is true if

$$\mathbf{H} \geq \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}} \tag{A.18}$$

Note that $(\mathbf{H}^{-1/2}\mathbf{Z})\{(\mathbf{H}^{-1/2}\mathbf{Z})(\mathbf{H}^{-1/2}\mathbf{Z})^{\mathrm{T}}\}^{-1}(\mathbf{H}^{-1/2}\mathbf{Z})^{\mathrm{T}}$ is the projection matrix of $\mathbf{H}^{-1/2}\mathbf{Z}$, so it is, under the Loewner ordering, smaller than or equal to the identity matrix $\mathbf{I}_N$, namely,

$$\mathbf{I}_N \geq \mathbf{H}^{-1/2}\mathbf{Z}(\mathbf{Z}\mathbf{H}^{-1}\mathbf{Z}^{\mathrm{T}})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{H}^{-1/2}, \tag{A.19}$$

which implies (A.18). If $h(\mathbf{x}) = 1$, the equality can be verified directly. □

A.1.2 Proof of Lemma 1

Proof of Lemma 1. Let B be a bound for g_1 i.e., $|g_1| \leq B$. For any $\epsilon > 0$, by Markov's inequality,

$$\begin{aligned}
&\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} g_1(\mathbf{v}_i)g_2(\mathbf{v}_i)\Big| > \epsilon\Big\} \leq \frac{\mathbb{E}|g_1(\mathbf{v})g_2(\mathbf{v})|}{\epsilon} \\
&= \frac{\mathbb{E}[|g_1(\mathbf{v})||g_2(\mathbf{v})|I\{|g_2(\mathbf{v})| \leq K\}]}{\epsilon} + \frac{\mathbb{E}[|g_1(\mathbf{v})||g_2(\mathbf{v})|I\{|g_2(\mathbf{v})| > K\}]}{\epsilon} \\
&\leq \frac{K}{\epsilon}\mathbb{E}|g_1(\mathbf{v})| + \frac{B}{\epsilon}\mathbb{E}\{|g_2(\mathbf{v})|I(|g_2(\mathbf{v})| > K)\}.
\end{aligned}$$

For any $\zeta > 0$, we can choose a K large enough such that $\mathbb{E}\{|g_2(\mathbf{v})|I(|g_2(\mathbf{v})| \leq K)\} < \zeta\epsilon/(2B)$, since $\mathbb{E}|g_2(\mathbf{v})| < \infty$. The facts that $g_1(\mathbf{v}_i) \leq B$ and $g_1(\mathbf{v}_i) = o_P(1)$ imply that $\mathbb{E}|g_1(\mathbf{v})| = o(1)$. Thus, there is a n_ζ such that $\mathbb{E}|g_1(\mathbf{v})| < \zeta\epsilon/(2K)$ when $n > n_\zeta$. Therefore, for any $\zeta > 0$, $\mathbb{P}\{|n^{-1}\sum_{i=1}^{n} g_1(\mathbf{v}_i)g_2(\mathbf{v}_i)| > \epsilon\} < \zeta$ for sufficiently large n. This finishes the proof. □

A.1.3 Proof of Lemma 2

Proof of Lemma 2. Since $\check{\boldsymbol{\beta}}_w$ is the maximizer of

$$\lambda_{wf}(\boldsymbol{\beta}) = \sum_{i=1}^{N} |y_i - p_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)\big[y_i\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1) - \log\{1 + e^{\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_1)}\}\big],$$

$\sqrt{N}(\check{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_t)$ is the mazimizer of $\gamma_{wf}(\mathbf{s}) = \lambda_{wf}(\boldsymbol{\beta}_t + \mathbf{s}/\sqrt{N}) - \lambda_{wf}(\boldsymbol{\beta}_t)$. By Taylor's expansion,

$$\gamma_{wf}(\mathbf{s}) = \frac{1}{\sqrt{N}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_{wf}(\boldsymbol{\beta}_t) - \frac{1}{2N}\sum_{i=1}^{N} |y_i - p_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{N})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2 \tag{A.20}$$

where

$$\dot{\lambda}_{wf}(\boldsymbol{\beta}_t) = \sum_{i=1}^{N} \eta_i = \sum_{i=1}^{N} |\psi_i(\hat{\boldsymbol{\beta}}_1)|\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)\mathbf{x}_i, \tag{A.21}$$

and $\acute{\mathbf{s}}$ lies between $\mathbf{0}$ and $\mathbf{s}$.

Denote $\psi(\boldsymbol{\beta}) = y - p(\boldsymbol{\beta})$ and $\eta = |\psi(\hat{\boldsymbol{\beta}}_1)|\psi(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x})\mathbf{x}$. Note that $\mathbb{E}(\eta|\hat{\boldsymbol{\beta}}_1) = \mathbf{0}$ because

$$\begin{aligned}
\mathbb{E}(\eta|\hat{\boldsymbol{\beta}}_1, \mathbf{x}) =& \mathbb{E}\Big[|\psi(\hat{\boldsymbol{\beta}}_1)|\big\{y - p(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\big\}h(\mathbf{x})\mathbf{x}\Big|\hat{\boldsymbol{\beta}}_1, \mathbf{x}\Big] \\
=& - p(\mathbf{x}, \hat{\boldsymbol{\beta}}_1)p(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\{1 - p(\mathbf{x}, \boldsymbol{\beta}_t)\}h(\mathbf{x})\mathbf{x} \\
&+ \{1 - p(\mathbf{x}, \hat{\boldsymbol{\beta}}_1)\}\{1 - p(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\}p(\mathbf{x}, \boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x} \\
=& - \frac{e^{\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1}}{1 + e^{\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1}} \frac{e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t - \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1}}{1 + e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t - \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1}} \frac{1}{1 + e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t}} h(\mathbf{x})\mathbf{x} \\
&+ \frac{1}{1 + e^{\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1}} \frac{1}{1 + e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t - \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1}} \frac{e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t}}{1 + e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t}} h(\mathbf{x})\mathbf{x} = 0.
\end{aligned} \tag{A.22}$$

This also gives that

$$\mathbb{V}(\eta|\hat{\boldsymbol{\beta}}_1) = \mathbb{E}\big\{\mathbb{V}(\eta|\mathbf{x}, \hat{\boldsymbol{\beta}}_1)\big|\hat{\boldsymbol{\beta}}_1\big\} + \mathbb{V}\big\{\mathbb{E}(\eta|\mathbf{x}, \hat{\boldsymbol{\beta}}_1)\big|\hat{\boldsymbol{\beta}}_1\big\} = \mathbb{E}\big\{\mathbb{V}(\eta|\mathbf{x}, \hat{\boldsymbol{\beta}}_1)\big|\hat{\boldsymbol{\beta}}_1\big\}.$$

Now, since

$$\begin{aligned}
\mathbb{V}(\eta|\mathbf{x}, \hat{\boldsymbol{\beta}}_1) =& \mathbb{E}\Big[|y - p(\hat{\boldsymbol{\beta}}_1)|^2\big\{y - p(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\big\}^2 h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big|\hat{\boldsymbol{\beta}}_1, \mathbf{x}\Big] \\
=& p(\mathbf{x}, \boldsymbol{\beta}_t)\{1 - p(\hat{\boldsymbol{\beta}}_1)\}^2\big\{1 - p(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\big\}^2 h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}} \\
&+ \{1 - p(\mathbf{x}, \boldsymbol{\beta}_t)\}\{p(\hat{\boldsymbol{\beta}}_1)\}^2\big\{p(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\big\}^2 h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}} \\
=& \phi_i(\hat{\boldsymbol{\beta}}_1)\phi_i(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}},
\end{aligned}$$

we have

$$\mathbb{V}(\eta|\hat{\boldsymbol{\beta}}_1) = \mathbb{E}\Big\{\phi(\hat{\boldsymbol{\beta}}_1)\phi(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big|\hat{\boldsymbol{\beta}}_1\Big\}. \tag{A.23}$$

Let $\|\|$ denote the Frobenius norm if applied on a martix, i.e., for a matrix A, $\|A\|^2 = \mathrm{tr}(AA^{\mathrm{T}})$, and denote $\mathbb{V}(\eta|\boldsymbol{\beta}_t) = \mathbb{E}\{\phi(\boldsymbol{\beta}_t)\phi(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\} = 0.25\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$. Notice that $\big|\phi(\hat{\boldsymbol{\beta}}_1)\phi(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_t) - 0.25\phi_i(\boldsymbol{\beta}_t)\big|h^2(\mathbf{x})\|\mathbf{x}\|^2$ converges to 0 in probability and it is bounded by $h^2(\mathbf{x})\|\mathbf{x}\|^2$, an integrable random variable under Assumption 2. Thus,

$$\mathbb{E}\big\|\mathbb{V}(\eta|\hat{\boldsymbol{\beta}}_1) - \mathbb{V}(\eta|\boldsymbol{\beta}_t)\big\| \le \mathbb{E}\big\{\big|\{\phi(\hat{\boldsymbol{\beta}}_1)\phi(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_t) - 0.25\phi_i(\boldsymbol{\beta}_t)\big|h^2(\mathbf{x})\|\mathbf{x}\|^2\big\} = o(1). \tag{A.24}$$

This implies that

$$\mathbb{V}(\eta|\hat{\boldsymbol{\beta}}_1) = \mathbb{V}(\eta|\boldsymbol{\beta}_t) + o_P(1) = 0.25\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}\big\} + o_P(1). \tag{A.25}$$

Conditional on $\hat{\boldsymbol{\beta}}_1$, η_1, ..., η_N are i.i.d. with mean $\mathbf{0}$ and variance $\mathbb{V}(\eta|\hat{\boldsymbol{\beta}}_1)$. Since for any $\epsilon > 0$,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big\{\|\eta_i\|^2 I(\|\eta_i\| > \sqrt{N}\epsilon)\Big|\hat{\boldsymbol{\beta}}_1\Big\} \le& \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big\{\|h(\mathbf{x}_i)\mathbf{x}_i\|^2 I(\|h(\mathbf{x}_i)\mathbf{x}_i\| > \sqrt{N}\epsilon)\Big|\hat{\boldsymbol{\beta}}_1\Big\}\\ =& \mathbb{E}\{\|h(\mathbf{x})\mathbf{x}\|^2 I(\|h(\mathbf{x})\mathbf{x}\| > \sqrt{N}\epsilon)\} \to 0,\end{aligned}$$

the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998) applies conditional on $\hat{\boldsymbol{\beta}}_1$. Thus, we have, conditional on $\hat{\boldsymbol{\beta}}_1$,

$$\frac{\dot{\lambda}_{wf}(\boldsymbol{\beta}_t)}{\sqrt{N}} \longrightarrow \mathbb{N}\bigg[\mathbf{0},\ \frac{\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}\}}{4}\bigg], \tag{A.26}$$

From Lemma 1, conditional on $\hat{\boldsymbol{\beta}}_1$,

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}|y_i - p_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{N})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\ &= \frac{1}{4}\mathbb{E}\{|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\} + o_{P|\hat{\boldsymbol{\beta}}_1}(1) = \frac{1}{2}\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\} + o_P(1).\end{aligned}$$

Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), the mazimizer of $\gamma_{wf}(\mathbf{s})$, $\sqrt{N}(\check{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_t)$, satisfies

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{wf} - \boldsymbol{\beta}_t) = 2[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1}\frac{1}{\sqrt{N}}\dot{\lambda}_w(\boldsymbol{\beta}_t) + o_P(1). \tag{A.27}$$

Note that

$$[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1} = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}}{4\Phi(\boldsymbol{\beta}_t)}. \tag{A.28}$$

Combining equations (A.21), (A.27), and (A.28), we have

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{wf} - \boldsymbol{\beta}_t) = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}}{2\Phi(\boldsymbol{\beta}_t)}\frac{1}{\sqrt{N}}\dot{\lambda}_w(\boldsymbol{\beta}_t) + o_P(1).$$

An application of Slutsky's theorem yields the result for the asymptotic normality.

□

A.1.4 Proof of Lemma 3

Proof of Lemma 3. Note that given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$, $\{y_i^* - p_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\}\mathbf{x}_i^*$ are i.i.d. random vectors. We now exam their mean and variance, and check the Lindeberg-Feller condition (Section *2.8 of van der Vaart, 1998) under the conditional distribution given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_1$. For the expectation, we have,

$$\mathbb{E}\big[\{y^* - p^*(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)\}\mathbf{x}^*|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1\big] = \sum_{i=1}^{N}\pi_i(\hat{\boldsymbol{\beta}}_1)\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i = \frac{\sum_{i=1}^{N}\eta_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_1)}. \tag{A.29}$$

From Lemma 2 and its proof, $\sum_{i=1}^N \eta_i = O_P(\sqrt{N})$ conditional on $\hat{\boldsymbol\beta}_1$ in probability, i.e., for any $\epsilon > 0$, there exits a K such that $\mathbb{P}\{\mathbb{P}(\sum_{i=1}^N \eta_i/\sqrt{N} > K|\hat{\boldsymbol\beta}_1) < \epsilon\} \to 1$ as $n_1, N \to \infty$. From Xiong and Li (2008), we know that $\sum_{i=1}^N \eta_i = O_P(\sqrt{N})$ unconditionally. Thus, for the expectation, we have

$$\Delta = \mathbb{E}\big[\big\{y^* - p^*(\boldsymbol\beta - \hat{\boldsymbol\beta}_1)\big\}\mathbf{x}^*|\mathcal{D}_N, \hat{\boldsymbol\beta}_1\big] = O_P(1/\sqrt{N}).$$

For the variance,

$$\begin{aligned}
&\mathbb{V}\big[\big\{y^* - p^*(\boldsymbol\beta - \hat{\boldsymbol\beta}_1)\big\}\mathbf{x}^*|\mathcal{D}_N, \hat{\boldsymbol\beta}_1\big]\\
=&\sum_{i=1}^N \pi_i\{y_i - p_i(\boldsymbol\beta_t - \hat{\boldsymbol\beta}_1)\}^2\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \Delta^2\\
=&\frac{\frac{1}{N}\sum_{i=1}^N |\psi_i(\hat{\boldsymbol\beta}_1)|\psi_i^2(\boldsymbol\beta_t - \hat{\boldsymbol\beta}_1)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\hat{\boldsymbol\beta}_1)} - O_P(1/N)\\
=&\frac{\frac{1}{N}\sum_{i=1}^N |\psi_i(\boldsymbol\beta_t)|(y_i - 0.5)^2h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\boldsymbol\beta_t)} + o_P(1)\\
=&\frac{1}{4}\frac{\mathbb{E}\{|\psi(\boldsymbol\beta_t)|h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{\Psi(\boldsymbol\beta_t)} + o_P(1) = \boldsymbol\Sigma_{\boldsymbol\beta_t}^{-1} + o_P(1)
\end{aligned}$$

where the third equality is from Lemma 1 and the fact that $\mathbb{E}\{h(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$, and the forth equality is from the law of large numbers.

Now we check the Lindeberg-Feller condition (Section *2.8 of van der Vaart, 1998) under the condition distribution. Denote $\dot\lambda_{ri}^* = \{y_i^* - p_i^*(\boldsymbol\beta - \hat{\boldsymbol\beta}_1)\}\mathbf{x}_i^*$.

$$\mathbb{E}\frac{1}{n}\sum_{i=1}^n \big\{\|\dot\lambda_{ri}^*\|^2 I(\|\dot\lambda_{ri}^*\| > \sqrt{n}\epsilon)\big|\mathcal{D}_N, \hat{\boldsymbol\beta}_1\big\} \tag{A.30}$$

$$\le\mathbb{E}\big\{\|\mathbf{x}^*\|^2 I(\|\mathbf{x}\| > \sqrt{n}\epsilon)\big|\mathcal{D}_N, \hat{\boldsymbol\beta}_1\big\} \tag{A.31}$$

$$=\sum_{i=1}^N \pi(\hat{\boldsymbol\beta}_1)\big\{\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\| > \sqrt{n}\epsilon)\big\} \tag{A.32}$$

$$\le\frac{\frac{1}{N}\sum_{i=1}^N \big\{h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\| > \sqrt{n}\epsilon)\big\}}{\Psi_N(\hat{\boldsymbol\beta}_1)} \tag{A.33}$$

$$\le\frac{\frac{1}{N}\sum_{i=1}^N \big\{h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\| > \sqrt{n}\epsilon)\big\}}{\Psi_N(\hat{\boldsymbol\beta}_1)} = o_P(1), \tag{A.34}$$

by Lemma 1 and the fact that $\mathbb{E}\{h(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$. Thus, applying the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998) finishes the proof. □

A.1.5 Proof of Lemma 4

Proof of Lemma 4. We begin with the following partition,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1+\mathbf{s}_n)\|\mathbf{x}_i^*\|^2 && \text{(A.35)}\\
=&\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1+\mathbf{s}_n)\|\mathbf{x}_i^*\|^2 I(\|\mathbf{x}_i^*\|^2\le n)+\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1+\mathbf{s}_n)\|\mathbf{x}_i^*\|^2 I(\|\mathbf{x}_i^*\|^2> n)\\
\equiv&\Delta_1+\Delta_2. && \text{(A.36)}
\end{aligned}$$

The second term Δ_2 is $o_P(1)$ because it is non-negative and

$$\begin{aligned}
\mathbb{E}(\Delta_2|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1)&=\sum_{i=1}^{N}\pi_i(\hat{\boldsymbol{\beta}}_1)\phi_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1+\mathbf{s}_n)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\|^2>n) && \text{(A.37)}\\
&\le\frac{\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\|^2>n)}{\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)} && \text{(A.38)}\\
&\le\frac{\frac{1}{N}\sum_{i=1}^{N}h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\|^2>n)}{\Psi_N(\hat{\boldsymbol{\beta}}_1)}=o_P(1) && \text{(A.39)}
\end{aligned}$$

as $n,N\to\infty$, where the last step is from Lemma 1.

Similarly, we can show that

$$\mathbb{E}(\Delta_1|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1)-\sum_{i=1}^{N}\pi(\hat{\boldsymbol{\beta}}_1)\phi_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1)\|\mathbf{x}_i\|^2=o_P(1). \tag{A.40}$$

Thus, we only need to show that $\Delta_1-\mathbb{E}(\Delta_1|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1)=o_P(1)$. For this, we show that the conditional variance of Δ_1 goes to 0 in probability. Notice that

$$\begin{aligned}
&\mathbb{V}(\Delta_1|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1)\\
=&\frac{1}{n}\mathbb{V}\big\{\phi^*(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1)\|\mathbf{x}^*\|^2 I(\|\mathbf{x}^*\|^2\le n)\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1\big\}\\
\le&\frac{\|\mathbf{s}\|^4}{16n}\mathbb{E}\big\{\|\mathbf{x}^*\|^4 I(\|\mathbf{x}^*\|^2\le n)\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1\big\}\\
=&\frac{\|\mathbf{s}\|^4}{16n}\sum_{i=1}^{n}\mathbb{E}\big\{\|\mathbf{x}^*\|^4 I(i-1<\|\mathbf{x}^*\|^2\le i)\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1\big\}\\
\le&\frac{\|\mathbf{s}\|^4}{16n}\sum_{i=1}^{n}i^2\mathbb{E}\big\{I(i-1<\|\mathbf{x}^*\|^2\le i)\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1\big\}\\
\le&\frac{\|\mathbf{s}\|^4}{16n}\sum_{i=1}^{n}i^2\big\{\mathbb{P}(\|\mathbf{x}^*\|^2>i-1\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1)-\mathbb{P}(\|\mathbf{x}^*\|^2>i\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_1)\big\}
\end{aligned}$$

$$=\frac{\|\mathbf{s}\|^4}{16n}\bigg\{\mathbb{P}(\|\mathbf{x}^*\|^2>0\big|\mathcal{D}_N,\hat{\boldsymbol\beta}_1)-n^2\mathbb{P}(\|\mathbf{x}^*\|^2>n\big|\mathcal{D}_N,\hat{\boldsymbol\beta}_1)+\sum_{i=1}^{n-1}(2i+1)\mathbb{P}(\|\mathbf{x}^*\|^2>i\big|\mathcal{D}_N,\hat{\boldsymbol\beta}_1)\bigg\}$$
$$\leq\frac{\|\mathbf{s}\|^4}{16n}\bigg\{1+\sum_{i=1}^{n-1}3i\mathbb{P}(\|\mathbf{x}^*\|^2>i\big|\mathcal{D}_N,\hat{\boldsymbol\beta}_1)\bigg\}$$

This is $o_P(1)$ because

$$\frac{1}{n}\sum_{i=1}^{n}i\mathbb{P}(\|\mathbf{x}^*\|^2>i|\mathcal{D}_N,\hat{\boldsymbol\beta}_1)=\frac{1}{n}\sum_{i=1}^{n}i\sum_{j=1}^{n}\pi_j(\hat{\boldsymbol\beta}_1)I(\|\mathbf{x}_j\|^2>i) \tag{A.41}$$
$$=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}i\pi_j(\hat{\boldsymbol\beta}_1)I(\|\mathbf{x}_j\|^2>i) \tag{A.42}$$
$$=\frac{\frac{1}{N}\sum_{j=1}^{n}\frac{1}{n}\sum_{i=1}^{n}i|\psi_j(\hat{\boldsymbol\beta}_1)|h(\mathbf{x}_j)I(\|\mathbf{x}_j\|^2>i)}{\Psi_N(\hat{\boldsymbol\beta}_1)} \tag{A.43}$$
$$\leq\frac{\frac{1}{N}\sum_{j=1}^{n}\frac{1}{n}\sum_{i=1}^{n}ih(\mathbf{x}_j)I(\|\mathbf{x}_j\|^2>i)}{\Psi_N(\hat{\boldsymbol\beta}_1)}, \tag{A.44}$$

and the numerator is non-negative and has an expectation

$$\frac{1}{N}\sum_{j=1}^{n}\frac{1}{n}\sum_{i=1}^{n}i\mathbb{E}\{h(\mathbf{x})I(\|\mathbf{x}\|^2>i)\} \tag{A.45}$$

which is $o(1)$ since $i\mathbb{E}\{h(\mathbf{x})I(\|\mathbf{x}\|^2>i)\}=o(1)$ as $i\to\infty$. □

A.2 Proofs for Poisson subsampling

In this section we prove the results in Section 4 about Poisson subsampling.

Define $\delta_i^{\hat{\boldsymbol\beta}_1}=I\{u_i\leq n\pi_i^p(\hat{\boldsymbol\beta}_1)\}$, and use notation λ_p to denote the log-likelihood shifted by $\hat{\boldsymbol\beta}_1$, i.e., $\lambda_p(\boldsymbol\beta)=\ell_p^*(\boldsymbol\beta-\hat{\boldsymbol\beta}_1)$. Using these notation, the estimator $\hat{\boldsymbol\beta}_p$ is the maximizer of

$$\lambda_p(\boldsymbol\beta)=\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol\beta}_1}\{n\pi_i^p(\hat{\boldsymbol\beta}_1)\vee 1\}\big[(\boldsymbol\beta-\hat{\boldsymbol\beta}_1)^{\mathrm{T}}\mathbf{x}_iy_i-\log\{1+e^{(\boldsymbol\beta-\hat{\boldsymbol\beta}_1)^{\mathrm{T}}\mathbf{x}_i}\}\big], \tag{A.46}$$

Denote the first and second derivatives of $\lambda_p(\boldsymbol\beta)$ as $\dot\lambda_p(\boldsymbol\beta)=\partial\lambda_p(\boldsymbol\beta)/\partial\boldsymbol\beta$ and $\ddot\lambda_p(\boldsymbol\beta)=\partial^2\lambda_p(\boldsymbol\beta)/(\partial\boldsymbol\beta\partial\boldsymbol\beta^{\mathrm{T}})$. Two lemmas similar to Lemmas 3 and 4 are derived below which will be used in the proof of Theorem 2.

Lemma 5. *Let*

$$\dot\lambda_p(\boldsymbol\beta_t)=\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol\beta}_1}\{n\pi_i^p(\hat{\boldsymbol\beta}_1)\vee 1\}\{y_i-p_i(\boldsymbol\beta_t-\hat{\boldsymbol\beta}_1)\}\mathbf{x}_i. \tag{A.47}$$

Under Assumptions 1 and 2, conditional on $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_1$, and $\hat{\Psi}_1$, if $n = o(N)$, then

$$\frac{\dot{\lambda}_p(\boldsymbol{\beta}_t)}{\sqrt{n}} - \frac{\sqrt{n}\sum_{i=1}^N \eta_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_1)} \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big), \tag{A.48}$$

in distribution; if $n/N \to \rho \in (0,1)$, then

$$\frac{\dot{\lambda}_p(\boldsymbol{\beta}_t)}{\sqrt{n}} - \frac{\sqrt{n}\sum_{i=1}^N \eta_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_1)} \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Lambda}_\rho\big), \tag{A.49}$$

in distribution.

Lemma 6. *Under Assumptions 1 and 2, as $n, N \to \infty$, for any $\mathbf{s}_n \to 0$,*

$$\frac{1}{n}\sum_{i=1}^N \delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \mathbf{s}_n)\|\mathbf{x}_i\|^2 - \sum_{i=1}^N \pi_i^p(\hat{\boldsymbol{\beta}}_1)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\|\mathbf{x}_i\|^2 = o_P(1).$$

Proof of Theorem 2. The estimator $\hat{\boldsymbol{\beta}}_p$ is the maximizer of (A.46), so $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$ is the mazimizer of $\gamma_p(\mathbf{s}) = \lambda_p(\boldsymbol{\beta}_t + \mathbf{s}/\sqrt{n}) - \lambda_p(\boldsymbol{\beta}_t)$. By Taylor's expansion,

$$\gamma_p(\mathbf{s}) = \frac{1}{\sqrt{n}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) + \frac{1}{2n}\sum_{i=1}^N \delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2 \tag{A.50}$$

where $\phi_i(\boldsymbol{\beta}) = p_i(\boldsymbol{\beta})\{1 - p_i(\boldsymbol{\beta})\}$, and $\acute{\mathbf{s}}$ lies between $\mathbf{0}$ and $\mathbf{s}$.

From Lemmas 5 and 6, conditional on $\mathcal{D}_N$, and $\hat{\boldsymbol{\beta}}_1$,

$$\frac{1}{n}\sum_{i=1}^N \delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} + o_P(1). \tag{A.51}$$

In addition, from Lemma 5, conditional on $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_1$, and $\hat{\Psi}_1$, $\dot{\lambda}_p(\boldsymbol{\beta}_t)/\sqrt{n}$ converges in distribution to a normal limit. Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), the mazimizer of $\gamma_p(\mathbf{s})$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$, satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\frac{1}{\sqrt{n}}\dot{\lambda}_p(\boldsymbol{\beta}_t) + o_P(1) \tag{A.52}$$

given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_1$, and $\hat{\Psi}_1$. Combining this with Lemma 5 and Slutsky's theorem, Theorem 2 follows. □

A.2.1 Proof of Proposition 2

Proof of Proposition 2. To prove that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_\rho\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} < \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$, we just need to show that $\boldsymbol{\Lambda}_\rho < \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}$. This is true because

$$\begin{aligned}\boldsymbol{\Lambda}_\rho &= \frac{\mathbb{E}\big[\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\{\Phi(\boldsymbol{\beta}_t) - \rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\}_+\mathbf{x}\mathbf{x}^{\mathrm{T}}\big]}{4\Phi^2(\boldsymbol{\beta}_t)}\\ &\le \frac{\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\Phi(\boldsymbol{\beta}_t)\mathbf{x}\mathbf{x}^{\mathrm{T}}\big\}}{4\Phi^2(\boldsymbol{\beta}_t)} = \frac{\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big\}}{4\Phi(\boldsymbol{\beta}_t)} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}.\end{aligned}$$

□

A.2.2 Proof of Lemma 5

Proof of Lemma 5. Note that, $\delta_i^{\hat{\boldsymbol{\beta}}_1} = I\{u_i \le n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\}$, where u_i are i.i.d. with the standard uniform distribution. Thus, given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_1$, and $\hat{\Psi}_1$, $\dot{\lambda}_p(\boldsymbol{\beta}_t)$ is a sum of N independent random vectors. We now exam the mean and variance of $\dot{\lambda}_p(\boldsymbol{\beta}_t)$. Recall that $\eta_i = |\psi_i(\hat{\boldsymbol{\beta}}_1)|\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)\mathbf{x}_i$, and $\psi_i(\boldsymbol{\beta}) = y_i - p_i(\boldsymbol{\beta})$. For the mean, we have,

$$\frac{1}{\sqrt{n}}\mathbb{E}\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\} \tag{A.53}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^N \{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \wedge 1\}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i \tag{A.54}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^N n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i = \frac{\sqrt{n}}{\sqrt{N}}\frac{\sum_{i=1}^N \eta_i}{\hat{\Psi}_1\sqrt{N}} = O_P(\sqrt{n/N}), \tag{A.55}$$

where the last equality is from Lemma 2.

For the variance,

$$\begin{aligned}
&\frac{1}{n}\mathbb{V}\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\}\\
=&\frac{1}{n}\sum_{i=1}^N [\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \wedge 1\} - \{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \wedge 1\}^2]\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}^2\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\
=&\sum_{i=1}^N \pi_i^p(\hat{\boldsymbol{\beta}}_1)\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - n\sum_{i=1}^N \{\pi_i^p(\hat{\boldsymbol{\beta}}_1)\}^2\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\
=&\frac{\frac{1}{N}\sum_{i=1}^N |\psi_i(\hat{\boldsymbol{\beta}}_1)|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_1}\\
&- \frac{n}{N}\frac{\frac{1}{N}\sum_{i=1}^N \psi_i^2(\hat{\boldsymbol{\beta}}_1)\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_1^2}\\
\equiv&\Delta_3 + \Delta_4
\end{aligned} \tag{A.56}$$

Note that $\mathbb{E}\{h(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$, $\mathbb{E}\{h^2(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$, and $|\psi_i(\cdot)|$ are bounded. Thus, from Lemma 1, if $n/N \to \rho$,

$$\Delta_4 \to \rho\frac{\mathbb{E}\{\psi^2(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{4\Psi^2(\boldsymbol{\beta}_t)} = \rho\frac{\mathbb{E}\{\phi^2(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{4\Phi^2(\boldsymbol{\beta}_t)}, \tag{A.57}$$

in probability.

For the term Δ_3 in (A.56), it is equal to

$$\Delta_3 = \frac{1}{\hat{\Psi}_1^2}\frac{1}{N}\sum_{i=1}^N |\psi_i(\hat{\boldsymbol{\beta}}_1)|\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)}{N} \vee \hat{\Psi}_1\Big\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$$

$$= \frac{1}{\hat{\Psi}_1^2} \frac{n}{N^2} \sum_{i=1}^{N} \psi_i^2(\hat{\boldsymbol{\beta}}_1)\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1) h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} I\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)}{N} > \hat{\Psi}_1\Big\}$$
$$+ \frac{1}{\hat{\Psi}_1} \frac{1}{N} \sum_{i=1}^{N} |\psi_i(\hat{\boldsymbol{\beta}}_1)|\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1) h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} I\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)}{N} \le \hat{\Psi}_1\Big\}.$$

Since $\mathbb{E}\{h(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$, $\mathbb{E}\{h^2(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$, and $|\psi_i(\cdot)|$ are bounded, from Lemma 1, if $n/N \to \rho$, as $n \to \infty$ and $N \to \infty$, in probability,

$$\begin{aligned}\Delta_3 \to & \frac{\rho\mathbb{E}\big[\psi^2(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}} I\big\{\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x}) \ge \Psi(\boldsymbol{\beta}_t)\big\}\big]}{4\Psi^2(\boldsymbol{\beta}_t)} \\ & + \frac{\mathbb{E}\big[|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}} I\big\{\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x}) \le \Psi(\boldsymbol{\beta}_t)\}\big\}\big]}{4\Psi(\boldsymbol{\beta}_t)} \\ = & \frac{\mathbb{E}\Big(\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big[\{\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\} \vee \Phi(\boldsymbol{\beta}_t)\}\big]\Big)}{4\Phi^2(\boldsymbol{\beta}_t)}\end{aligned} \tag{A.58}$$

From, (A.56), (A.57), and (A.58), if $n/N \to \rho$,

$$\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big\} = \frac{\mathbb{E}\big[\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\{\Phi(\boldsymbol{\beta}_t) - \rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\}_+\big]}{4\Phi^2(\boldsymbol{\beta}_t)} + o_P(1). \tag{A.59}$$

Specifically, when $\rho = 0$,

$$\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big\} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} + o_P(1). \tag{A.60}$$

Now we check the Lindeberg-Feller condition (Section *2.8 of van der Vaart, 1998) under the condition distribution. Denote $\dot{\lambda}_{pi} = \delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i$. For any $\epsilon > 0$

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\Big\{\big\|\dot{\lambda}_{pi}\big\|^2 I(\big\|\dot{\lambda}_{pi}\big\| > \sqrt{n}\epsilon)\Big|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\Big\} \\ \le & \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\Big[\big\|\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\mathbf{x}_i\big\|^2 I(\big\|\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\mathbf{x}_i\big\| > \sqrt{n}\epsilon)\Big|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\Big] \\ = & \sum_{i=1}^{N}\pi_i^p(\hat{\boldsymbol{\beta}}_1)\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\|\mathbf{x}_i\|^2 I(\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\|\mathbf{x}_i\| > \sqrt{n}\epsilon) \\ \le & \frac{|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)\{n/N|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i) + \hat{\Psi}_1\}\|\mathbf{x}_i\|^2 I(\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) + 1\}\|\mathbf{x}_i\| > \sqrt{n}\epsilon)}{\hat{\Psi}_1^2} \\ \le & \frac{\frac{1}{N}\sum_{i=1}^{N} h^2(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\{h(\mathbf{x}_i)/\hat{\Psi}_1 + 1\}\|\mathbf{x}_i\| > \sqrt{n}\epsilon)}{\hat{\Psi}_1^2} \\ + & \frac{\frac{1}{N}\sum_{i=1}^{N} h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\{h(\mathbf{x}_i)/\hat{\Psi}_1 + 1\}\|\mathbf{x}_i\| > \sqrt{n}\epsilon)}{\hat{\Psi}_1} = o_P(1),\end{aligned}$$

where the last equality is from Lemma 1. Thus, applying the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998) finishes the proof. □

A.2.3 Proof of Lemma 6

Proof of Lemma 6. Note that, from Lemma 1,

$$\sum_{i=1}^{N} \pi_i^p(\hat{\boldsymbol{\beta}}_1)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \frac{1}{\hat{\Psi}_1 N}\sum_{i=1}^{N} |\psi_i(\hat{\boldsymbol{\beta}}_1)|\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} + o_P(1);$$

and from the strong law of large numbers

$$\frac{1}{n}\sum_{i=1}^{N} \delta_i^{\boldsymbol{\beta}_t}\{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} + o_P(1),$$

where $\delta_i^{\boldsymbol{\beta}_t} = I\{u_i \le n\pi_i^p(\boldsymbol{\beta}_t)\}$. Thus, if we show that

$$\Delta_5 \equiv \frac{1}{n}\sum_{i=1}^{N}\left|\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \mathbf{s}_n) - \delta_i^{\boldsymbol{\beta}_t}\{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t)\right|\|\mathbf{x}_i\|^2 = o_P(1), \tag{A.61}$$

then the result in Lemma 6 follows. Noting that Δ_5 is nonnegative, we prove (A.61) by showing that $\mathbb{E}(\Delta_5|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1) = o_P(1)$. Note that given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_1$, and $\hat{\Psi}_1$, the only random terms in Δ_5 are $\delta_i^{\hat{\boldsymbol{\beta}}_1} = I\{u_i \le n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\}$ and $\delta_i^{\boldsymbol{\beta}_t} = I\{u_i \le n\pi_i^p(\boldsymbol{\beta}_t)\}$. We have that

$$\begin{aligned}
&\mathbb{E}(\Delta_5|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1)\\
\le&\frac{1}{n}\sum_{i=1}^{N}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \wedge n\pi_i^p(\boldsymbol{\beta}_t) \wedge 1\}\\
&\qquad\times\left|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \mathbf{s}_n) - \{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t)\right|\|\mathbf{x}_i\|^2\\
&+\frac{1}{n}\sum_{i=1}^{N}|n\pi_i^p(\hat{\boldsymbol{\beta}}_1) - n\pi_i^p(\boldsymbol{\beta}_t)|\left|n\pi_i^p(\hat{\boldsymbol{\beta}}_1) + n\pi_i^p(\boldsymbol{\beta}_t) + 2\right|\|\mathbf{x}_i\|^2\\
\equiv&\Delta_6 + \Delta_7.
\end{aligned} \tag{A.62}$$

Note that $n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \wedge n\pi_i^p(\boldsymbol{\beta}_t) \wedge 1 \le n\pi_i^p(\hat{\boldsymbol{\beta}}_1)$. Thus Δ_6 is bounded by

$$\frac{1}{\hat{\Psi}_1}\frac{1}{N}\sum_{i=1}^{N}\left|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \mathbf{s}_n) - \{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t)\right|h(\mathbf{x}_i)\|\mathbf{x}_i\|^2, \tag{A.63}$$

which is $o_P(1)$ by Lemma 1 if $|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\} - \{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}| = o_P(1)$. This is true because

$$\begin{aligned}
|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\} - \{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}| \le& n|\pi_i^p(\hat{\boldsymbol{\beta}}_1) - \pi_i^p(\boldsymbol{\beta}_t)|\\
\le&\frac{nh(\mathbf{x}_i)}{N}\left|\frac{|\psi_i(\hat{\boldsymbol{\beta}}_1)|}{\hat{\Psi}_1} - \frac{|\psi_i(\boldsymbol{\beta}_t)|}{\Psi_N(\boldsymbol{\beta}_t)}\right| = o_P(1).
\end{aligned}$$

The term Δ_7 is bounded by

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{|\psi_i(\hat{\boldsymbol{\beta}}_1)|}{\hat{\Psi}_1}-\frac{|\psi_i(\boldsymbol{\beta}_t)|}{\Psi_N(\boldsymbol{\beta}_t)}\right|\left|\frac{|\psi_i(\hat{\boldsymbol{\beta}}_1)|}{\hat{\Psi}_1}+\frac{|\psi_i(\boldsymbol{\beta}_t)|}{\Psi_N(\boldsymbol{\beta}_t)}+\frac{2}{h(\mathbf{x}_i)}\right|h^2(\mathbf{x}_i)\|\mathbf{x}_i\|^2=o_P(1),$$

where the equality to $o_P(1)$ is from Lemma 1 and the fact that $\mathbb{E}\{h^2(\mathbf{x})\|\mathbf{x}\|^2\}<\infty$. □

A.3 Proofs for unconditional distribution

In this section we prove Theorem 3 in Section 6. A lemma similar to Lemma 5 is presented below. Lemma 6 can be used in the proof of Theorem 2 because for the problem considered in this paper, convergence to zero in probability is equivalent to convergence to zero in probability under the conditional probability measure (Xiong and Li, 2008).

For the pilot sample taken according to the subsampling probabilities π_{1i} in (25), we define $\delta_i^{(1)}=I\{u_{1i}\le\frac{c_0(1-y_i)+c_1y_i}{N}\}$, where u_{1i} are i.i.d. standard uniform random variables. With this notation, the estimator $\hat{\Psi}_1$ defined in (26) can be written as

$$\hat{\Psi}_1=\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^{(1)}|y_i-p_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i)}{n\pi_{1i}\wedge 1}. \tag{A.64}$$

Lemma 7. *Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$ be constructed according to Step 1 of Algorithm 3, respectively. For*

$$\dot{\lambda}_p(\boldsymbol{\beta}_t)=\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\vee 1\}\{y_i-p_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1)\}\mathbf{x}_i, \tag{A.65}$$

under the same assumptions of Theorem 3, if $n=o(N)$, then

$$\frac{\dot{\lambda}_p(\boldsymbol{\beta}_t)}{\sqrt{n}}\longrightarrow\mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big), \tag{A.66}$$

in distribution; if $n/N\to\rho\in(0,1)$, then

$$\frac{\dot{\lambda}_p(\boldsymbol{\beta}_t)}{\sqrt{n}}\longrightarrow\mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Lambda}_u\big), \tag{A.67}$$

in distribution.

Proof of Theorem 3. The proof of this theorem is similar to that of Theorem 2. The key difference is that Lemma 7 is about asymptotic distribution unconditionally.

The estimator $\hat{\boldsymbol{\beta}}_p$ is the maximizer of

$$\lambda_p(\boldsymbol{\beta})=\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\vee 1\}\big[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_1)^{\mathrm{T}}\mathbf{x}_iy_i-\log\{1+e^{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_1)^{\mathrm{T}}\mathbf{x}_i}\}\big], \tag{A.68}$$

so $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$ is the mazimizer of $\gamma_p(\mathbf{s}) = \lambda_p(\boldsymbol{\beta}_t + \mathbf{s}/\sqrt{n}) - \lambda_p(\boldsymbol{\beta}_t)$. By Taylor's expansion,

$$\gamma_p(\mathbf{s}) = \frac{1}{\sqrt{n}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) + \frac{1}{2n}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2 \tag{A.69}$$

where $\acute{\mathbf{s}}$ lies between $\mathbf{0}$ and $\mathbf{s}$.

From Lemma 5,

$$\frac{1}{n}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1 + \acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i(\mathbf{x}_i)^{\mathrm{T}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} + o_P(1). \tag{A.70}$$

In addition, from Lemma 7, $\dot{\lambda}_p(\boldsymbol{\beta}_t)/\sqrt{n}$ converges in distribution to a normal limit. Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), the mazimizer of $\gamma_p(\mathbf{s})$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$, satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\frac{1}{\sqrt{n}}\dot{\lambda}_p(\boldsymbol{\beta}_t) + o_P(1). \tag{A.71}$$

Combining this with Lemma 7 and Slutsky's theorem, Theorem 3 follows. □

A.3.1 Proof of Proposition 3

Proof of Proposition 3. To prove (39), we just need to show that $\boldsymbol{\Lambda}_u \geq \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} > \boldsymbol{\Lambda}_\rho$. From Proposition 2, we know that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} > \boldsymbol{\Lambda}_\rho$. To show that $\boldsymbol{\Lambda}_u \geq \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}$, we notice that

$$\begin{aligned}\boldsymbol{\Lambda}_u &= \frac{\mathbb{E}[\phi(\boldsymbol{\beta}_t)\{\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\vee\Phi(\boldsymbol{\beta}_t)\}h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}]}{4\Phi^2(\boldsymbol{\beta}_t)}.\\ &\geq \frac{\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\Phi(\boldsymbol{\beta}_t)\mathbf{x}\mathbf{x}^{\mathrm{T}}\big\}}{4\Phi^2(\boldsymbol{\beta}_t)} = \frac{\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big\}}{4\Phi(\boldsymbol{\beta}_t)} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1},\end{aligned}$$

where the strict inequality holds if $\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\vee\Phi(\boldsymbol{\beta}_t) \neq \rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x})$ with positive probability, i.e., $\mathbb{P}\{\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x}) > \Phi(\boldsymbol{\beta}_t)\} > 0$. □

A.3.2 Proof of Lemma 7

Proof of Lemma 7. We first proof the case when the pilot estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$ depend on the data. For any $l \in \mathbb{R}^d$, denote $\tau_{Ni} = \sqrt{N/n}\hat{\Psi}_1\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\vee 1\}\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i^{\mathrm{T}}l$, $i = 1, ..., N$, where $\delta_i^{\hat{\boldsymbol{\beta}}_1} = I\{u_i \leq n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\}$, and u_i are i.i.d. standard uniform random variables. Note that τ_{Ni}'s have the same distribution but they are not independent. We now exam the mean and variance of τ_{Ni}. For the mean, based on calculation similar to that in (A.22), we have,

$$\mathbb{E}\big(\tau_{Ni}\big|\hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big) = \sqrt{nN}\hat{\Psi}_1\mathbb{E}\big\{\pi_i^p(\hat{\boldsymbol{\beta}}_1)\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)\mathbf{x}_i^{\mathrm{T}}l\big|\hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big\} = \frac{\sqrt{n}\mathbb{E}(\eta_i|\hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1)}{\sqrt{N}} = 0, \tag{A.72}$$

which implies that

$$\mathbb{E}\tau_{Ni} = 0. \tag{A.73}$$

For the variance, $\mathbb{V}(\tau_{Ni}) = \mathbb{E}(\tau_{Ni}^2)$, we start with the condition expectation,

$$\begin{aligned}\mathbb{E}\big(\tau_{Ni}^2\big|\hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big) =& N\hat{\Psi}_1^2\mathbb{E}\Big[\pi_i^p(\hat{\boldsymbol{\beta}}_1)\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1) \vee 1\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)(\mathbf{x}_i^{\mathrm{T}}l)^2\Big|\hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\Big] \\ =& \mathbb{E}\Big[|\psi_i(\hat{\boldsymbol{\beta}}_1)|\Big\{\frac{n}{N}|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i) \vee \hat{\Psi}_1\Big\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2\Big|\hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\Big].\end{aligned}$$

If we let

$$\Upsilon_{Ni} = |\psi_i(\hat{\boldsymbol{\beta}}_1)|\Big\{\frac{n}{N}|\psi_i(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_i) \vee \hat{\Psi}_1\Big\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2, \tag{A.74}$$

then $\mathbb{V}(\tau_{Ni}) = \mathbb{E}(\Upsilon_{Ni})$. Note that

$$\Upsilon_{Ni} \to \Upsilon_i = 0.25|\psi_i(\boldsymbol{\beta}_t)|\{\rho_1|\psi_i(\boldsymbol{\beta}_t)|h(\mathbf{x}_i) \vee \Psi(\boldsymbol{\beta}_t)\}h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2, \tag{A.75}$$

in probability. We now show that

$$\mathbb{E}(\Upsilon_{Ni}) \to \mathbb{E}(\Upsilon_i) = \mathbb{E}[\phi(\boldsymbol{\beta}_t)\{\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x}) \vee \Phi\phi(\boldsymbol{\beta}_t)\}h(\mathbf{x})(\mathbf{x}^{\mathrm{T}}l)^2]. \tag{A.76}$$

Let $\Xi = |\Upsilon_{Ni} - \Upsilon_i|$. For any ϵ,

$$\begin{aligned}|\mathbb{E}(\Upsilon_{Ni}) - \mathbb{E}(\Upsilon_i)| &= \mathbb{E}\{\Xi I(\Xi > \epsilon)\} + \mathbb{E}\{\Xi I(\Xi \le \epsilon)\} \\ &\le \mathbb{E}\big[\{h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2 + \Upsilon_i + \hat{\Psi}_1 h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2\}I(\Xi > \epsilon)\big] + \epsilon\end{aligned}$$

We know that $\mathbb{E}\big[\{h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2 + \Upsilon_i\}I(\Xi > \epsilon)\big] \to 0$ since $\mathbb{E}\{h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2 + \Upsilon_i\} < \infty$ for any $l \in \mathbb{R}^d$, and $I(\Xi > \epsilon)$ is bounded and is $o_P(1)$. Thus, to prove that $\mathbb{E}(\Upsilon_{Ni}) - \mathbb{E}(\Upsilon_i) \to 0$, we only need to show that

$$\begin{aligned}&\mathbb{E}\big\{\hat{\Psi}_1 h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2 I(\Xi > \epsilon)\big\} \\ &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\bigg\{\frac{\delta_k^{(1)}|y_k - p_k(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_k)}{n_1\pi_k^{(1)} \wedge 1}h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2 I(\Xi > \epsilon)\bigg\} \equiv \frac{1}{N}\sum_{k=1}^{N}\Delta_{8k} \to 0,\end{aligned}$$

which is true if $\Delta_{8k} = o(1)$ for any k. We need to keep in mind that $\delta_k^{(1)} = I\{u_{1k} \le \frac{c_0(1-y_k)+c_1y_k}{N}\}$ and Υ_{Ni} are correlated, and they both depend on the data. For π_{1k} with the expression in (25), we have

$$\frac{1}{n_1\pi_{1k} \wedge 1} \le \frac{N}{n_1}\frac{c_0 + c_1}{c_0c_1} + 1 \quad \text{and} \quad n_1\pi_{1k} \wedge 1 \le \frac{n_1}{N}(c_0 + c_1). \tag{A.77}$$

Using this, we have

$$\Delta_{8k} \le \left(\frac{N}{n_1}\frac{c_0+c_1}{c_0c_1}+1\right)\mathbb{E}\big\{\delta_k^{(1)}h(\mathbf{x}_k)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2I(\Xi>\epsilon)\big\} \tag{A.78}$$

$$= \left(\frac{N}{n_1}\frac{c_0+c_1}{c_0c_1}+1\right)\mathbb{E}\big\{h(\mathbf{x}_k)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2I(\Xi>\epsilon)\big|\delta_k^{(1)}=1\big\}\mathbb{P}(\delta_j^{(1)}=1) \tag{A.79}$$

$$< \frac{(c_0+c_1+c_0c_1)^2}{c_0c_1}\mathbb{E}\big\{h(\mathbf{x}_k)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2I(\Xi>\epsilon)\big|\delta_k^{(1)}=1\big\} \tag{A.80}$$

Note that $\mathbb{E}\big\{h(\mathbf{x}_k)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2\big|\delta_k^{(1)}=1\big\}<\infty$ for any i, k, because $\mathbb{E}\{h^2(\mathbf{x})\|x\|^2\}<\infty$. Thus, $\Delta_{8k}=o(1)$, if $\Xi=o_P(1)$ given $\delta_k^{(1)}=1$. From the expression of Ξ, we only need to show that $\hat{\boldsymbol\beta}_1$ and $\hat\Psi_1$ are consistent given $\delta_k^{(1)}=1$. This is true because $n_1\to\infty$ and fixing the value of one observation does not affect the consistency. For example, for $\hat\Psi_1$, given $\delta_k^{(1)}=1$,

$$\hat\Psi_1=\frac{1}{N}\sum_{j\neq k}^{N}\frac{\delta_j^{(1)}|y_j-p_j(\hat{\boldsymbol\beta}_1)|h(\mathbf{x}_j)}{n_1\pi_j^{(1)}\wedge 1}+\frac{|y_k-p_k(\hat{\boldsymbol\beta}_1)|h(\mathbf{x}_k)}{n_1\{c_0(1-y_k)+c_1y_k\}\wedge N}, \tag{A.81}$$

in which the first term converge to $\Psi(\boldsymbol\beta_t)=2\Phi(\boldsymbol\beta_t)$ and the second term is $o_P(1)$. We have finished proving that

$$\mathbb{V}(\tau_{Ni})\to\mathbb{E}(\Upsilon_i)=\mathbb{E}[\phi(\boldsymbol\beta_t)\{\rho\phi(\boldsymbol\beta_t)h(\mathbf{x})\vee\Phi\phi(\boldsymbol\beta_t)\}h(\mathbf{x})(\mathbf{x}^{\mathrm{T}}l)^2]. \tag{A.82}$$

In the following, we exam the third moment of τ_{Ni} and prove that

$$\mathbb{E}|\tau_{Ni}|^3=o(\sqrt{N}). \tag{A.83}$$

For the conditional expectation,

$$\begin{aligned}
&\mathbb{E}\big(|\tau_{Ni}|^3\big|\hat{\boldsymbol\beta}_1,\hat\Psi_1\big)\\
&=N\sqrt{N/n}\hat\Psi_1^3\mathbb{E}\Big[\pi_i^p(\hat{\boldsymbol\beta}_1)\{n\pi_i^p(\hat{\boldsymbol\beta}_1)\vee 1\}^2\psi_i^3(\boldsymbol\beta_t-\hat{\boldsymbol\beta}_1)(\mathbf{x}_i^{\mathrm{T}}l)^3\Big|\hat{\boldsymbol\beta}_1,\hat\Psi_1\Big]\\
&=\sqrt{N/n}\mathbb{E}\Big[|\psi_i(\hat{\boldsymbol\beta}_1)|\Big\{\frac{n}{N}|\psi_i(\hat{\boldsymbol\beta}_1)|h(\mathbf{x}_i)\vee\hat\Psi_1\Big\}^2\psi_i^3(\boldsymbol\beta_t-\hat{\boldsymbol\beta}_1)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^3\Big|\hat{\boldsymbol\beta}_1,\hat\Psi_1\Big]\\
&\le 2\|l\|^3\sqrt{N/n}\mathbb{E}\big[\{h^2(\mathbf{x}_i)+\hat\Psi_1^2\}h(\mathbf{x}_i)\|\mathbf{x}_i\|^3\big|\hat{\boldsymbol\beta}_1,\hat\Psi_1\big].
\end{aligned}$$

Thus, (A.83) follows if $\mathbb{E}\big[\{h^2(\mathbf{x}_i)+\hat\Psi_1^2\}h(\mathbf{x}_i)\|\mathbf{x}_i\|^3\big]=O(1)$, which is true if

$$\mathbb{E}\{\hat\Psi_1^2h(\mathbf{x}_i)\|\mathbf{x}_i\|^3\}=O(1) \tag{A.84}$$

since $\mathbb{E}\{h^3(\mathbf{x}_i)\|\mathbf{x}_i\|^3\}=\mathbb{E}\{h^3(\mathbf{x})\|\mathbf{x}\|^3\}<\infty$ and it is a constant that does not depend n_1, n, or N. For (A.84),

$$\mathbb{E}\{\hat\Psi_1^2h(\mathbf{x}_i)\|\mathbf{x}_i\|^3\}$$

$$
\begin{aligned}
&= \mathbb{E}\left\{ \frac{1}{N} \sum_{k_1=1}^{N} \frac{\delta_{k_1}^{(1)} |y_{k_1} - p_{k_1}(\hat{\boldsymbol{\beta}}_1)| h(\mathbf{x}_{k_1})}{n_1 \pi_{k_1}^{(1)} \wedge 1} \frac{1}{N} \sum_{k_2=1}^{N} \frac{\delta_{k_2}^{(1)} |y_{k_2} - p_{k_2}(\hat{\boldsymbol{\beta}}_1)| h(\mathbf{x}_{k_2})}{n_1 \pi_{k_2}^{(1)} \wedge 1} h(\mathbf{x}_i) \|\mathbf{x}_i\|^3 \right\} \\
&\leq \frac{1}{N^2} \sum_{k_1 \neq k_2}^{N} \mathbb{E}\left\{ \frac{\delta_{k_1}^{(1)} h(\mathbf{x}_{k_1})}{n_1 \pi_{k_1}^{(1)} \wedge 1} \frac{\delta_{k_2}^{(1)} h(\mathbf{x}_{k_2})}{n_1 \pi_{k_2}^{(1)} \wedge 1} h(\mathbf{x}_i) \|\mathbf{x}_i\|^3 \right\} + \frac{1}{N^2} \sum_{k=1}^{N} \mathbb{E}\left\{ \frac{\delta_k^{(1)} h^2(\mathbf{x}_k)}{\{n_1 \pi_{1k} \wedge 1\}^2} h(\mathbf{x}_i) \|\mathbf{x}_i\|^3 \right\} \\
&= \frac{1}{N^2} \sum_{k_1 \neq k_2}^{N} \mathbb{E}\{h(\mathbf{x}_{k_1}) h(\mathbf{x}_{k_2}) h(\mathbf{x}_i) \|\mathbf{x}_i\|^3\} + \frac{1}{N^2} \sum_{j=1}^{N} \mathbb{E}\left\{ \frac{1}{n_1 \pi_{1k} \wedge 1} h^2(\mathbf{x}_k) h(\mathbf{x}_i) \|\mathbf{x}_i\|^3 \right\} \\
&\leq \frac{1}{N^2} \sum_{k_1 \neq k_2}^{N} \mathbb{E}\{h(\mathbf{x}_{k_1}) h(\mathbf{x}_{k_2}) h(\mathbf{x}_i) \|\mathbf{x}_i\|^3\} + \frac{c_0 + c_1 + c_0 c_1}{c_0 c_1 n_1 N} \sum_{j=1}^{N} \mathbb{E}\{h^2(\mathbf{x}_k) h(\mathbf{x}_i) \|\mathbf{x}_i\|^3\} \\
&= \frac{N-1}{N} \{\mathbb{E} h(\mathbf{x})\}^2 \mathbb{E}\{h(\mathbf{x}) \|\mathbf{x}\|^3\} + \frac{c_0 + c_1 + c_0 c_1}{c_0 c_1 n_1} [\mathbb{E}\{h^2(\mathbf{x})\}][\mathbb{E}\{h(\mathbf{x}) \|\mathbf{x}\|^3\}] = O(1). \quad \text{(A.85)}
\end{aligned}
$$

This finish the proof of (A.84).

Denote $\nu_{Ni} = \tau_{Ni}\{\mathbb{V}(\tau_{Ni})\}^{-1/2}$. We know that ν_{Ni} are i.i.d. conditional on $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$. Thus, from Theorem 7.3.2 of Chow and Teicher (2003), they are interchangeable. The fact that $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$ are consistent estimators implies that they are a sequence of two estimators, and for each $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$, τ_{Ni} are interchangeable. For this setup, the central limit theorem in Theorem 2 of Blum et al. (1958) can be applied to prove the asymptotic normality.

It is evident that ν_{Ni} have mean 0 and variance 1. It is also easy to verify that, for $i \neq j$,

$$
\mathbb{E}(\nu_{Ni}\nu_{Nj}) = \mathbb{E}\{\mathbb{E}(\nu_{Ni}\nu_{Nj} | \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1)\} = 0, \tag{A.86}
$$

and

$$
\frac{1}{\sqrt{N}} \mathbb{E}\{|\nu_{Ni}|^3\} = \mathbb{E}|\tau_{Ni}|^3 \{\mathbb{V}(\tau_{Ni})\}^{-3/2} \to 0 \tag{A.87}
$$

which follows from (A.83). We now show that for $i \neq j$,

$$
\mathbb{E}\{\nu_{Ni}^2 \nu_{Nj}^2\} \to 1. \tag{A.88}
$$

Since $\nu_{Ni} = \tau_{Ni}\{\mathbb{V}(\tau_{Ni})\}^{-1/2}$, from (A.82), to prove (A.88), we only need to show that $\mathbb{E}(\tau_{Ni}^2 \tau_{Nj}^2) \to \mathbb{E}(\Upsilon_i)\mathbb{E}(\Upsilon_j) = \mathbb{E}(\Upsilon_i \Upsilon_j)$, where the equality is because Υ_i and Υ_j are independent. Noting that τ_{Ni}^2 and τ_{Nj}^2 are conditionally independent, we have $\mathbb{E}\big(\tau_{Ni}^2 \tau_{Nj}^2 | \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big) = \mathbb{E}\big(\tau_{Ni}^2 | \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big) \mathbb{E}\big(\tau_{Ni}^2 | \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big) = \mathbb{E}\big(\Upsilon_{Ni} | \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big) \mathbb{E}\big(\Upsilon_{Nj} | \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big) = \mathbb{E}\big(\Upsilon_{Ni}\Upsilon_{Nj} | \hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\big)$, so we know that $\mathbb{E}\big(\tau_{Ni}^2 \tau_{Nj}^2\big) = \mathbb{E}\big(\Upsilon_{Ni}\Upsilon_{Nj}\big)$.

Now we prove that $\mathbb{E}\big(\Upsilon_{Ni}\Upsilon_{Nj}\big) \to \mathbb{E}\big(\Upsilon_i \Upsilon_j\big)$. Let $\Xi_2 = |\Upsilon_{Ni}\Upsilon_{Nj} - \Upsilon_i \Upsilon_j|$. For any $\epsilon > 0$,

$$
\begin{aligned}
&|\mathbb{E}(\Upsilon_{Ni}\Upsilon_{nj}) - \mathbb{E}(\Upsilon_i \Upsilon_j)| \\
\leq &\mathbb{E}\{\Xi_2 I(\Xi_2 > \epsilon)\} + \mathbb{E}\{\Xi_2 I(\Xi_2 \leq \epsilon)\}
\end{aligned}
$$

$$\begin{aligned}
\leq & \mathbb{E}\big[\{h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2h^2(\mathbf{x}_j)(\mathbf{x}_j^{\mathrm{T}}l)^2 + \Upsilon_i\Upsilon_j\}I(\Xi_2 > \epsilon)\big] \\
& + \mathbb{E}\big[\hat{\Psi}_1^2 h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 h(\mathbf{x}_j)\|\mathbf{x}_j\|^2 I(\Xi_2 > \epsilon)\big] \\
& + \mathbb{E}\big[\hat{\Psi}_1\{h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2h^2(\mathbf{x}_j)(\mathbf{x}_j^{\mathrm{T}}l)^2 + h(\mathbf{x}_j)(\mathbf{x}_j^{\mathrm{T}}l)^2h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2\}I(\Xi_2 > \epsilon)\big] + \epsilon \\
& \equiv \Delta_9 + \Delta_{10} + \Delta_{11}
\end{aligned} \tag{A.89}$$

Note that $\mathbb{E}\{h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2h^2(\mathbf{x}_j)(\mathbf{x}_j^{\mathrm{T}}l)^2 + \Upsilon_i\Upsilon_j\} < \infty$, $I(\Xi_2 > \epsilon) \leq 1$, and $\Xi_2 = o_P(1)$, thus $\Delta_9 \to 0$. Now we exam Δ_{10} and Δ_{11} show that they are also $o(1)$. For Δ_{10},

$$\begin{aligned}
\Delta_{10} = & \mathbb{E}\{\hat{\Psi}_1^2 h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 h(\mathbf{x}_j)\|\mathbf{x}_j\|^2 I(\Xi_2 > \epsilon)\} \\
= & \mathbb{E}\bigg\{\frac{1}{N}\sum_{k_1=1}^{N}\frac{\delta_{k_1}^{(1)}|y_{k_1} - p_{k_1}(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_{k_1})}{n_1\pi_{k_1}^{(1)} \wedge 1} \\
& \qquad \times \frac{1}{N}\sum_{k_2=1}^{N}\frac{\delta_{k_2}^{(1)}|y_{k_2} - p_{k_2}(\hat{\boldsymbol{\beta}}_1)|h(\mathbf{x}_{k_2})}{n_1\pi_{k_2}^{(1)} \wedge 1} h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 h(\mathbf{x}_j)\|\mathbf{x}_j\|^2 I(\Xi_2 > \epsilon)\bigg\} \\
\leq & \frac{1}{N^2}\sum_{k_1 \neq k_2}^{N}\mathbb{E}\bigg\{\frac{\delta_{k_1}^{(1)}h(\mathbf{x}_{k_1})}{n_1\pi_{k_1}^{(1)} \wedge 1}\frac{\delta_{k_2}^{(1)}h(\mathbf{x}_{k_2})}{n_1\pi_{k_2}^{(1)} \wedge 1} h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 h(\mathbf{x}_j)\|\mathbf{x}_j\|^2 I(\Xi_2 > \epsilon)\bigg\} \qquad (A.90) \\
& + \frac{1}{N^2}\sum_{k=1}^{N}\mathbb{E}\bigg\{\frac{\delta_k^{(1)}h^2(\mathbf{x}_k)}{\{n_1\pi_{1k} \wedge 1\}^2} h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 h(\mathbf{x}_j)\|\mathbf{x}_j\|^2 I(\Xi_2 > \epsilon)\bigg\}. \qquad (A.91)
\end{aligned}$$

Let $h_{k_1k_2ij}x_{i^2j^2} = h(\mathbf{x}_{k_1})h(\mathbf{x}_{k_2})h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 h(\mathbf{x}_j)\|\mathbf{x}_j\|^2$. From (A.77), we have

$$\begin{aligned}
& \mathbb{E}\bigg\{\frac{\delta_{k_1}^{(1)}}{n_1\pi_{k_1}^{(1)} \wedge 1}\frac{\delta_{k_2}^{(1)}}{n_1\pi_{k_2}^{(1)} \wedge 1} h_{k_1k_2ij}x_{i^2j^2} I(\Xi_2 > \epsilon)\bigg\} \\
\leq & \bigg(\frac{N}{n_1}\frac{c_0 + c_1}{c_0c_1} + 1\bigg)^2 \mathbb{E}\bigg[h_{k_1k_2ij}x_{i^2j^2} I(\Xi_2 > \epsilon)\bigg|\delta_{k_1}^{(1)} = \delta_{k_2}^{(1)} = 1\bigg]\mathbb{P}(\delta_{k_1}^{(1)} = \delta_{k_2}^{(1)} = 1) \\
< & \frac{(c_0 + c_1 + c_0c_1)^2(c_0 + c_1)^2}{c_0^2c_1^2}\mathbb{E}\bigg[h_{k_1k_2ij}x_{i^2j^2} I(\Xi_2 > \epsilon)\bigg|\delta_{k_1}^{(1)} = \delta_{k_2}^{(1)} = 1\bigg].
\end{aligned} \tag{A.92}$$

Note that although k_1 or k_2 may equal i or j, $k_1 \neq k_2$, so we know that $\mathbb{E}(h_{k_1k_2ij}x_{i^2j^2}) < \infty$ since $\mathbb{E}\{h^2(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$. Thus, the term in (A.92) is o(1) if $\Xi_2 = o_P(1)$ given $\delta_j^{(1)} = 1$. This is true because $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$ are consistent given $\delta_{k_1}^{(1)} = \delta_{k_2}^{(1)} = 1$, namely, fixing the values of two observations does not affect the consistency of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$. From (A.92), the term in (A.90) converges to zero. Using a similar approach, it can be shown that if $\mathbb{E}\{h^3(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$, then $\Delta_{11} = o(1)$, and the term in (A.91) converges to zero, which implies that $\Delta_{10} = o(1)$. Thus, (A.88) holds.

Since (A.86), (A.87), (A.88) are satisfied, the central limit theorem in Theorem 2 of Blum et al. (1958) holds for ν_{Ni}, which gives that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\nu_{Ni} \to \mathbb{N}(0, 1), \tag{A.93}$$

in distribution. Note that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\nu_i = \frac{\hat{\Psi}_1}{\sqrt{n}\{\mathbb{V}(\tau_{Ni})\}^{1/2}}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_1}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\vee 1\}\psi_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1)\mathbf{x}_i^{\mathrm{T}}l$$

$$= \frac{\hat{\Psi}_1}{\sqrt{n}\{\mathbb{V}(\tau_{Ni})\}^{1/2}}l^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) = \frac{\Psi}{\sqrt{n}\{\mathbb{V}(\tau_{Ni})\}^{1/2}}l^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t)+o_P(1).$$

Thus, from Slutsky's theorem, for any $l\in\mathbb{R}^d$,

$$\frac{1}{\sqrt{n}}l^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t)\to\mathbb{N}(0,l^{\mathrm{T}}\boldsymbol{\Lambda}_u l) \tag{A.94}$$

in distribution, where

$$\boldsymbol{\Lambda}_u=\frac{\mathbb{V}(\tau_{Ni})}{\Psi^2(\boldsymbol{\beta}_t)}=\frac{\mathbb{E}[\phi(\boldsymbol{\beta}_t)\{\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\vee\Phi\}h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}]}{4\Phi^2(\boldsymbol{\beta}_t)}\geq\frac{\mathbb{E}[\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}]}{4\Phi(\boldsymbol{\beta}_t)}=\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}, \tag{A.95}$$

and the equality holds if $\rho=0$, i.e., $n/N\to 0$. Based on (A.94), from the Cramér-Wold theorem, we have that

$$\frac{1}{\sqrt{n}}\dot{\lambda}_p(\boldsymbol{\beta}_t)\to\mathbb{N}(0,\boldsymbol{\Lambda}_u) \tag{A.96}$$

in distribution.

When the pilot estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$ are independent of the data, if we can prove the results in Lemma 7 under the conditional distribution given $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$, then the result follows unconditionally. We provide the proof under the conditional distribution in the following. The proof is similar to the proof of Lemma 5 and thus we provide only the outline. The difference is we do not conditional on the full data $\mathcal{D}_N$ here.

Note that, given $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$, $\dot{\lambda}_p(\boldsymbol{\beta}_t)$ is a sum of N independent random vectors. We now exam the mean and variance of $\dot{\lambda}_p(\boldsymbol{\beta}_t)$ given $\hat{\boldsymbol{\beta}}_1$ and $\hat{\Psi}_1$. For the mean,

$$\frac{1}{\sqrt{n}}\mathbb{E}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\hat{\boldsymbol{\beta}}_1,\hat{\Psi}_1\big\}=\mathbf{0}. \tag{A.97}$$

For the variance,

$$\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\hat{\boldsymbol{\beta}}_1,\hat{\Psi}_1\big\}=\frac{\mathbb{E}\Big[|\psi_i(\hat{\boldsymbol{\beta}}_1)|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_1)\vee 1\}\psi_i^2(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_1)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\Big|\hat{\boldsymbol{\beta}}_1,\hat{\Psi}_1\Big]}{\hat{\Psi}_1}, \tag{A.98}$$

which, under Assumptions 1 and 2, converges in probability to $\boldsymbol{\Lambda}_u$.

To check the Lindeberg-Feller condition (Section *2.8 of van der Vaart, 1998) under the condition distribution, we note that for any $\epsilon>0$,

$$\frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\Big\{\big\|\dot{\lambda}_{pi}\big\|^2 I(\big\|\dot{\lambda}_{pi}\big\|>\sqrt{n}\epsilon)\Big|\hat{\boldsymbol{\beta}}_1,\hat{\Psi}_1\Big\}$$

$$\leq \frac{\mathbb{E}\Big[\{h(\mathbf{x})\|\mathbf{x}\|^2 + \hat{\Psi}_1\}h(\mathbf{x})I(\{h(\mathbf{x})/\hat{\Psi}_1 + 1\}\|\mathbf{x}\| > \sqrt{n}\epsilon)\Big|\hat{\boldsymbol{\beta}}_1, \hat{\Psi}_1\Big]}{\hat{\Psi}_1^2} = o_P(1).$$

Thus, applying the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998) finishes the proof.

□

References

Avron, H., Maymounkov, P. and Toledo, S. (2010), 'Blendenpik: Supercharging LAPACK's least-squares solver', *SIAM Journal on Scientific Computing* **32**, 1217–1236.

Baldi, P., Sadowski, P. and Whiteson, D. (2014), 'Searching for exotic particles in high-energy physics with deep learning', *Nature Communications* **5**(4308), http://dx.doi.org/10.1038/ncomms5308.

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2017), 'Julia: A fresh approach to numerical computing', *SIAM Review* **59**(1), 65–98.

Blum, J., Chernoff, H., Rosenblatt, M. and Teicher, H. (1958), 'Central limit theorems for interchangeable processes', *Canad. J. Math* **10**, 222–229.

Chow, Y. S. C. and Teicher, H. (2003), *Probability Theory: Independence, Interchangeability, Martingales*, Springer, New York.

Clarkson, K. L. and Woodruff, D. P. (2013), Low rank approximation and regression in input sparsity time, *in* 'Proceedings of the forty-fifth annual ACM symposium on Theory of computing', ACM, pp. 81–90.

Drineas, P., Kannan, R. and Mahoney, M. W. (2006*a*), 'Fast monte carlo algorithms for matrices i: Approximating matrix multiplication', *SIAM Journal on Computing* **36**(1), 132–157.

Drineas, P., Kannan, R. and Mahoney, M. W. (2006*b*), 'Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix', *SIAM Journal on Computing* **36**(1), 158–183.

Drineas, P., Kannan, R. and Mahoney, M. W. (2006*c*), 'Fast monte carlo algorithms for matrices iii: Computing a compressed approximate matrix decomposition', *SIAM Journal on Computing* **36**(1), 184–206.

Drineas, P., Magdon-Ismail, M., Mahoney, M. and Woodruff, D. (2012), 'Faster approximation of matrix coherence and statistical leverage.', *Journal of Machine Learning Research* **13**, 3475–3506.

Drineas, P., Mahoney, M., Muthukrishnan, S. and Sarlos, T. (2011), 'Faster least squares approximation', *Numerische Mathematik* **117**, 219–249.

Drineas, P., Mahoney, M. W. and Muthukrishnan, S. (2006), Sampling algorithms for l_2 regression and applications, *in* 'Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm', Society for Industrial and Applied Mathematics, pp. 1127–1136.

Fithian, W. and Hastie, T. (2014), 'Local case-control sampling: Efficient subsampling in imbalanced data sets', *Annals of statistics* **42**(5), 1693.

Halko, N., Martinsson, P.-G. and Tropp, J. A. (2011), 'Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions', *SIAM review* **53**(2), 217–288.

Hjort, N. L. and Pollard, D. (2011), 'Asymptotics for minimisers of convex processes', *arXiv preprint arXiv:1107.3806* .

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013), *Applied logistic regression*, Vol. 398, John Wiley & Sons.

Kernighan, B. W. and Ritchie, D. M. (1988), *The C Programming Language*, 2nd edn, Prentice Hall Professional Technical Reference.

Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. I. (2014), 'A scalable bootstrap for massive data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(4), 795–816.

Lichman, M. (2013), 'UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science'.
URL: *http://archive.ics.uci.edu/ml*

Ma, P., Mahoney, M. and Yu, B. (2014), A statistical perspective on algorithmic leveraging, *in* 'Proceedings of the 31st International Conference on Machine Learning (ICML-14)', pp. 91–99.

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

Ma, P. and Sun, X. (2015), 'Leveraging for big data regression', *Wiley Interdisciplinary Reviews: Computational Statistics* **7**(1), 70–76.
URL: *http://dx.doi.org/10.1002/wics.1324*

Mahoney, M. W. (2011), 'Randomized algorithms for matrices and data', *Foundations and Trends® in Machine Learning* **3**(2), 123–224.

Mahoney, M. W. and Drineas, P. (2009), 'CUR matrix decompositions for improved data analysis', *Proceedings of the National Academy of Sciences* **106**(3), 697–702.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability*, Chapman & Hall,.

McWilliams, B., Krummenacher, G., Lucic, M. and Buhmann, J. M. (2014), Fast and robust least squares estimation in corrupted linear models, *in* 'Advances in Neural Information Processing Systems', pp. 415–423.

Meng, X., Saunders, M. and Mahoney, M. (2014), 'LSRN: A parallel iterative solver for strongly over- or under- determined systems', *SIAM Journal on Scientific Computing* **36**, C95–C118.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Raskutti, G. and Mahoney, M. (2016), 'A statistical perspective on randomized sketching for ordinary least-squares', *Journal of Machine Learning Research* **17**, 1–31.

Stroustrup, B. (1986), *The C++ programming language*, Pearson Education India.

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

van Rossum, G. (1995), Python tutorial, technical report cs-r9526, Technical report, Centrum voor Wiskunde en Informatica (CWI), Amsterdam.

Wang, H., Yang, M. and Stufken, J. (2018), 'Information-based optimal subdata selection for big data linear regression', *Journal of the American Statistical Association* **accepted**, `https://doi.org/10.1080/01621459.2017.1408468`.

Wang, H., Zhu, R. and Ma, P. (2017), 'Optimal subsampling for large sample logistic regression', *Journal of the American Statistical Association* p. http://dx.doi.org/10.1080/01621459.2017.1292914.

Xiong, S. and Li, G. (2008), 'Some results on the convergence of conditional distributions', *Statistics & Probability Letters* **78**(18), 3249–3253.

Yang, T., Zhang, L., Jin, R. and Zhu, S. (2015), An explicit sampling dependent spectral error bound for column subset selection, *in* 'Proceedings of The 32nd International Conference on Machine Learning', pp. 135–143.

Yang, Y., Pilanci, M. and Wainwright, M. J. (2016), 'Randomized sketches for kernels: Fast and optimal non-parametric regression', *The Annals of Statistics,* p. forthcoming.