Journal of Machine Learning Research 20 (2019) 1-59 Submitted 9/18; Revised 7/19; Published 8/19

More Efficient Estimation for Logistic Regression with Optimal Subsamples

HaiYing Wang HAIYING.WANG@UCONN.EDU
Department of Statistics
University of Connecticut
Storrs, CT 06269, USA

Editor: Tong Zhang

Abstract

In this paper, we propose improved estimation method for logistic regression based on subsamples taken according the optimal subsampling probabilities developed in Wang et al. (2018). Both asymptotic results and numerical results show that the new estimator has a higher estimation efficiency. We also develop a new algorithm based on Poisson subsampling, 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 subsampling produces a more efficient estimator if the sampling ratio, the ratio of the subsample size to the full data sample size, does not converge to zero. We also obtain the unconditional asymptotic distribution for the estimator based on Poisson subsampling. Pilot estimators are required to calculate subsampling probabilities and to correct biases in un-weighted estimators; interestingly, even if pilot estimators are inconsistent, the proposed method still produce consistent and asymptotically normal estimators.

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 et al. (2006a,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. (2017), 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., 2015). It has demonstrated remarkable performance in better using of a fixed amount of computing power (Avron et al., 2010; Meng et al., 2014). Statistical leverage scores only contain information in the covariates and do not take into account

the information contained in the observed responses. Wang et al. (2018) 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 subsamples. Wang et al. (2018) used a inverse probability weighted estimator based on the optimal subsample, where more informative data points are assigned smaller weights in the objective function. Thus, we can improve the estimation efficiency based on the optimal subsample by using a better weighting scheme.

In this paper, we propose more efficient estimators based on subsamples taken randomly according to the optimal subsampling probabilities. We will derive asymptotic distributions to 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. (2018). We also consider to use Poisson subsampling. Asymptotic distributions show that Poisson subsampling 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 subsampling because there is no need to calculate and use subsampling probabilities for all data points simultaneously.

Before presenting the framework of the paper, we give a brief review of the emerging field of subsampling-based methods. For linear regression, Drineas et al. (2006d) 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. (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. (2019) and Wang (2019) 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. Han et al. (2019) developed a local uncertainty sampling approach for multi-class logistic regression. Recently, Wang et al. (2018) developed an Optimal Subsampling Method under the A-optimality Criterion (OSMAC) for logistic regression; Yao and Wang (2019) and Ai et al. (2019) extended this method to include multi-class logistic regression and generalized linear regression models, respectively. Although they derived optimal subsampling probabilities, they did not investigate whether a better weighting scheme can further improve the estimation efficiency.

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 subsamples taken according to OSMAC developed in Wang et al. (2018), 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 subsampling. 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 subsampling. Section 7 discusses asymptotic distributions with pilot and model misspecifications. Section 8 evaluates the practical performance of the proposed methods using numerical experiments. Section 9 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,

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

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 $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$. It is the maximizer of the log-likelihood function, namely,

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

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. For super-large data set, 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 to massive data, Wang et al. (2018) 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 according to the probabilities $\{\pi_i\}_{i=1}^{N}$ from the full data. We use * to indicate quantities for a subsample, namely, denote the covariates, responses, and subsampling probabilities in a subsample as $\mathbf{x}_i^*$, y_i^*, and π_i^*, respectively, for $i = 1, ..., n$. Wang et al. (2018) define the weighted subsample estimator $\hat{\boldsymbol{\beta}}_w^{\pi}$ to be

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

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. (2018) derived optimal subsampling probabilities that minimize the asymptotic MSE of $\hat{\boldsymbol{\beta}}_w^{\pi}$. They first showed that $\hat{\boldsymbol{\beta}}_w^{\pi}$ is asymptotically normal. Specifically, for large n and N, the conditional distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_w^{\pi} - \hat{\boldsymbol{\beta}}_{\text{MLE}})$ given the full data $\mathcal{D}_N$ can be approximated by a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{V}_N = \mathbf{M}_N^{-1}\mathbf{V}_{Nc}\mathbf{M}_N^{-1}$, in which

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

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

$$\pi_i^{\text{Aopt}}(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = \frac{|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{M}_N^{-1}\mathbf{x}_i\|}{\sum_{j=1}^{N}|y_j - p(\mathbf{x}_j, \hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{M}_N^{-1}\mathbf{x}_j\|}, \quad i = 1, ..., N; \tag{2}$$

$$\pi_i^{\text{Lopt}}(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = \frac{|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{x}_i\|}{\sum_{j=1}^{N}|y_j - p(\mathbf{x}_j, \hat{\boldsymbol{\beta}}_{\text{MLE}})|\|\mathbf{x}_j\|}, \quad i = 1, ..., N. \tag{3}$$

Here, $\{\pi_i^{\text{Aopt}}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}_{i=1}^N$ minimize $\text{tr}(\mathbf{V}_N)$, the trace of $\mathbf{V}_N$, and this is the A-optimality criterion in optimum experimental designs (Atkinson et al., 2007); $\{\pi_i^{\text{Lopt}}(\hat{\boldsymbol{\beta}}_{\text{MLE}})\}_{i=1}^N$ minimize $\text{tr}(\mathbf{V}_{Nc})$, and this is a choice of the L-optimality criterion. These subsampling probabilities have a lot of nice properties and meaningful interpretations. More details can be found in Section 3 of Wang et al. (2018).

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

$$\pi_i^{\text{OS}}(\boldsymbol{\beta}) = \frac{|y_i - p(\mathbf{x}_i, \boldsymbol{\beta})|h(\mathbf{x}_i)}{\sum_{j=1}^{N}|y_j - p(\mathbf{x}_j, \boldsymbol{\beta})|h(\mathbf{x}_j)}, \quad i = 1, ..., N, \tag{4}$$

where $h(\mathbf{x})$ is a univariate function of $\mathbf{x}$. We provide some intuitions on choosing $h(\mathbf{x})$. Let L be a matrix with d columns. Choosing $h(\mathbf{x}) = \|L\mathbf{M}_N^{-1}\mathbf{x}\|$ minimizes the trace of $L\mathbf{V}_N L^{\text{T}}$, which is the conditional asymptotic variance-covariance matrix of $L\hat{\boldsymbol{\beta}}_w^{\pi}$ (scaled by n) given the full data $\mathcal{D}_N$. Two special choices of $h(\mathbf{x})$ correspond to $L = \mathbf{I}$ (the identity matrix) and $L = \mathbf{M}_N$. If $L = \mathbf{I}$, then $h(\mathbf{x}) = \|\mathbf{M}_N^{-1}\mathbf{x}\|$ and $\pi_i^{\text{OS}}(\boldsymbol{\beta})$ becomes $\pi_i^{\text{Aopt}}(\boldsymbol{\beta})$; if $L = \mathbf{M}_N$, then $h(\mathbf{x}) = \|\mathbf{x}\|$ and $\pi_i^{\text{OS}}(\boldsymbol{\beta})$ becomes $\pi_i^{\text{Lopt}}(\boldsymbol{\beta})$. If one is interested in a specific component of $\boldsymbol{\beta}$, say β_j, then L can be chosen as a row vector with the j-th element being one and all other elements being zero. With this choice, $h(\mathbf{x}) = \|\mathbf{M}_{N,\,\boldsymbol{.}j}^{-1}\mathbf{x}\|$ where $\mathbf{M}_{N,\,\boldsymbol{.}j}^{-1}$ means the j-th row of $\mathbf{M}_N^{-1}$, and the asymptotic variance of $\hat{\boldsymbol{\beta}}_{w,j}^{\pi}$ is minimized. If $h(\mathbf{x}) = 1$, then $\pi_i^{\text{OS}}(\boldsymbol{\beta})$'s are proportional to the local case-control subsampling probabilities (Fithian and Hastie, 2014).

Note that $\{\pi_i^{\text{OS}}(\boldsymbol{\beta})\}_{i=1}^N$ depend on the unknown $\boldsymbol{\beta}$, so a pilot estimate of $\boldsymbol{\beta}$ is required to approximate them. Let $\hat{\boldsymbol{\beta}}_0$ be a pilot estimator from a pilot subsample taken from the full data, for which we will provide more details in Section 5. The original weighted OSMAC estimator is

$$\hat{\boldsymbol{\beta}}_w = \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^{n}\frac{y_i^*\boldsymbol{\beta}^{\text{T}}\mathbf{x}_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\text{T}}\mathbf{x}_i^*}\right)}{\pi_i^{\text{OS}}(\hat{\boldsymbol{\beta}}_0)^*}. \tag{5}$$

In Wang et al. (2018), $\hat{\boldsymbol{\beta}}_w$ has exceptional performance because $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ are able to include more informative data points in the subsample. However, we can improve the weighting scheme adopted in (5). Intuitively, a larger $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)$ means that the data point $(\mathbf{x}_i, y_i)$ contains more information about $\boldsymbol{\beta}$, but it has a smaller weight in the objective function in (5). This reduces contributions of more informative data points to the objective function for parameter estimation.

The weighted estimator in (5) is used because $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ depend on the responses y_i's and an un-weighted estimator is biased. If the bias can be corrected, then the resultant estimator can be more efficient in parameter estimation, because an un-weighted estimator often has a smaller variance-covariance matrix compared with an inverse probability weighted estimator. Intuitively, if some data points with very small values of $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)$ are selected in the subsample, then the target function in (5) would be dominated by these data points. As a result, the variance-covariance matrix of the weighted estimator would be inflated by small values of $\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)$. Note that π_i's appear in the denominator of $\mathbf{V}_{Nc}$ in the asymptotic variance-covariance matrix of the weighted estimator. A major goal of this paper is to develop un-weighted estimation procedures. Interestingly, for the subsampling probabilities in (4), the bright idea proposed in Fithian and Hastie (2014) can be used to correct the bias of the un-weighted estimator.

3. More efficient estimator

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 $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ defined in (4). Using this subsample, we present a more efficient estimation procedure based on un-weighted estimator with bias correction. Remember that a pilot estimate is required, and we use $\hat{\boldsymbol{\beta}}_0$ to denote it. Here, we focus the discussion on the new estimation procedure and assume that $\hat{\boldsymbol{\beta}}_0$ is obtained based on a pilot subsample of size n_0 and it is consistent. More details about this pilot estimator will be provided in Section 5, and the scenario that $\hat{\boldsymbol{\beta}}_0$ is inconsistent will be investigated in Section 7.1. The following procedure describes how to obtain the un-weighted estimator with bias correction, denoted as $\hat{\boldsymbol{\beta}}_{uw}$.

Calculate the naive un-weighted estimator

$$\tilde{\boldsymbol{\beta}}_{uw} = \arg\max_{\boldsymbol{\beta}} \ell_{uw}^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* y_i^* - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^*}\right) \right\}, \tag{6}$$

and then let

$$\hat{\boldsymbol{\beta}}_{uw} = \tilde{\boldsymbol{\beta}}_{uw} + \hat{\boldsymbol{\beta}}_0. \tag{7}$$

The naive un-weighted estimator $\tilde{\boldsymbol{\beta}}_{uw}$ in (6) is biased, and the bias is corrected in (7) using $\hat{\boldsymbol{\beta}}_0$. We will show in the following that $\hat{\boldsymbol{\beta}}_{uw}$ is asymptotically unbiased. This, together with the fact that $\hat{\boldsymbol{\beta}}_0$ is consistent, shows the interesting fact that $\tilde{\boldsymbol{\beta}}_{uw}$ converges to $\mathbf{0}$ in probability as n_0, n, and N go to infinity.

To investigate the asymptotic properties, we use $\boldsymbol{\beta}_t$ to denote the true value of $\boldsymbol{\beta}$, and summarize some regularity conditions 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. (2018), 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,

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

Note that $n^{3/2}I(\|\mathbf{x}\|^3 > n^{3/2}) \le \|\mathbf{x}\|^3$ and $I(\|\mathbf{x}\|^3 > n^{3/2}) \to 0$ in probability. Thus, if $\mathbb{E}(\|\mathbf{x}\|^3) < \infty$, then $n^{3/2}\mathbb{P}(\|\mathbf{x}\|^3 > n^{3/2}) = \mathbb{E}\{n^{3/2}I(\|\mathbf{x}\|^3 > n^{3/2})\} \to 0$ (see Theorem 1.3.6 of Serfling, 1980). 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. (2018) 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. (2018). Assumptions 1 and 2 are required in all the theorems in this paper while Assumption 3 is only required in Theorems 1, 18, and 24.

Theorem 1 *Under Assumptions 1-3, conditional on* $\mathcal{D}_N$, *if* $\hat{\boldsymbol{\beta}}_0$ *is consistent, then as* n_0, n, *and* N *go to infinity,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big), \tag{8}$$

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big) \tag{9}$$

in distribution, where

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

and $\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}$ *is a weighted MLE based on the full data defined as*

$$\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} |y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)| h(\mathbf{x}_i)\big[y_i\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0) - \log\{1 + e^{\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)}\}\big]. \tag{10}$$

Here $\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}$ *satisfies that*

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

in distribution if $\hat{\boldsymbol{\beta}}_0$ is obtained from a uniform pilot subsample of size n_0 such that $n_0/\sqrt{N} = o(1)$ or if $\hat{\boldsymbol{\beta}}_0$ is independent of $\mathcal{D}_N$, where

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

Remark 2 *Theorem 1 shows that the un-weighted estimator $\hat{\boldsymbol{\beta}}_{uw}$ is $\sqrt{n}$-consistent to $\hat{\boldsymbol{\beta}}_{\text{wMLE}}$, a weighted MLE based on the full data in conditional probability, while Theorem 5 of Wang et al. (2018) shows that the weighted estimator $\hat{\boldsymbol{\beta}}_w$ is $\sqrt{n}$-consistent to $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, the un-weighted MLE based on the full data in conditional probability. Specifically,* (8) *implies that given $\mathcal{D}_N$ in probability,*

$$\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\text{wMLE}} = O_{P|\mathcal{D}_N}(n^{-1/2}). \tag{12}$$

The $O_{P|\mathcal{D}_N}(n^{-1/2})$ expression in (12) *means that for any $\epsilon > 0$, there exist a δ_ϵ such that as $n, N \to \infty$,*

$$\mathbb{P}\Big\{ \sup_n \mathbb{P}(\|\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\text{wMLE}}\| > n^{-1/2}\delta_\epsilon|\mathcal{D}_N) \le \epsilon \Big\} \to 1.$$

Note that if a sequence is bounded in conditional probability, then it is bounded in unconditional probability, i.e., if $a_n = O_{P|\mathcal{D}_N}(1)$, then $a_n = O_P(1)$ (Xiong and Li, 2008; Cheng and Huang, 2010). Therefore, (12) *implies that $\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\text{wMLE}} = O_P(n^{-1/2})$. Similarly,* (11) *implies that $\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_t = O_P(N^{-1/2})$. Thus, $\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_t = O_P(n^{-1/2} + N^{-1/2}) = O_P(n^{-1/2})$, showing the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\beta}}_{uw}$ to the true parameter under the unconditional distribution.*

Remark 3 *For $\hat{\boldsymbol{\beta}}_{\text{wMLE}}$, if $\hat{\boldsymbol{\beta}}_0$ is fixed, say $\hat{\boldsymbol{\beta}}_0 = \boldsymbol{\beta}_0$, then the population log-likelihood for the objective function in* (10) *is*

$$\mathbb{E}\Big(a(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\beta}_0)\Big[p(\mathbf{x}, \boldsymbol{\beta} - \boldsymbol{\beta}_0)\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \log\{1 + e^{\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}\}\Big]\Big),$$

where $a(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\beta}_0) = \big[p(\mathbf{x}, \boldsymbol{\beta})\{1 - p(\mathbf{x}, \boldsymbol{\beta}_0)\} + \{1 - p(\mathbf{x}, \boldsymbol{\beta})\}p(\mathbf{x}, \boldsymbol{\beta}_0)\big]h(\mathbf{x})$. If $h(\mathbf{x}) = 1$, then this population log-likelihood is identical to that for the local case-control subsampling estimator. For general $h(\mathbf{x})$, since it does not rely on the response variable, we expect that $\hat{\boldsymbol{\beta}}_{\text{wMLE}}$ inherits the main properties of the the local case-control subsampling estimator, including those under model misspecification. Indeed this is the case, and more details for the scenarios of misspecifications will be presented in Section 7.

Theorem 1 shows that, asymptotically, the distribution of $\hat{\boldsymbol{\beta}}_{uw}$ given $\mathcal{D}_N$ is centered around $\hat{\boldsymbol{\beta}}_{\text{wMLE}}$ with variance-covariance matrix $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$, and the distribution of $\hat{\boldsymbol{\beta}}_{\text{wMLE}}$ is centered around $\boldsymbol{\beta}_t$ with variance-covariance matrix $N^{-1}\boldsymbol{\Sigma}_{\text{wMLE}}$. Thus, both $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ and $N^{-1}\boldsymbol{\Sigma}_{\text{wMLE}}$ should be considered in accessing the quality of $\hat{\boldsymbol{\beta}}_{uw}$ 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 $\hat{\boldsymbol{\beta}}_{uw}$. If $n/N \to 0$, then the variation of $\hat{\boldsymbol{\beta}}_{\text{wMLE}}$ can be ignored as stated in (9).

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

$$\mathbf{V}_{Nc}^{\mathrm{OS}} = \left\{\frac{1}{N}\sum_{i=1}^{N}|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|h(\mathbf{x}_i)\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\mathrm{MLE}})|\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{h(\mathbf{x}_i)}\right\}.$$

Note that the full data MLE $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ is consistent under Assumptions 1-2. If $\mathbb{E}\{\|\mathbf{x}\|^2/h(\mathbf{x})\} < \infty$, then from Lemma 28 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

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

Note that the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{w}$ given $\mathcal{D}_N$ is centered around $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$. It can be shown that under Assumptions 1-2,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}\big(\mathbf{0}, \mathbf{M}^{-1}\big),$$

in distribution. 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 $\hat{\boldsymbol{\beta}}_{uw}$, $N^{-1}\mathbf{M}^{-1}$ 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 $\hat{\boldsymbol{\beta}}_{uw}$ 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 4 *If $\mathbf{M}$, $\mathbf{V}_c^{\mathrm{OS}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ are finite and positive definite matrices, then*

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \leq \mathbf{V}^{\mathrm{OS}}. \tag{13}$$

Here, 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 (13) holds. Furthermore, note that the asymptotic variance-covariance matrix (scaled by n) for uniform subsampling estimator is $\mathbf{M}^{-1}$. If $\boldsymbol{\beta}_t \neq \mathbf{0}$ and $h(\mathbf{x}) = \|L\mathbf{M}^{-1}\mathbf{x}\|$ for some matrix L, then

$$\mathrm{tr}(L\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}L^{\mathrm{T}}) \leq \mathrm{tr}(L\mathbf{V}^{\mathrm{OS}}L^{\mathrm{T}}) \leq \mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}\mathrm{tr}(L\mathbf{M}^{-1}L^{\mathrm{T}}) < \mathrm{tr}(L\mathbf{M}^{-1}L^{\mathrm{T}}). \tag{14}$$

Remark 5 *This proposition shows that $\hat{\boldsymbol{\beta}}_{uw}$ is typically more efficient than $\hat{\boldsymbol{\beta}}_{w}$ in estimating $\boldsymbol{\beta}_t$. The numerical results in Section 8 also confirm this. Assume that $n/N \to \rho$. For the un-weighted estimator, the variation of $\sqrt{N}(\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\mathrm{wMLE}})$ is measured by $\rho^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ and the variation of $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}} - \boldsymbol{\beta}_t)$ is measured by $\boldsymbol{\Sigma}_{\mathrm{wMLE}}$, while for the weighted estimator the variation of $\sqrt{N}(\hat{\boldsymbol{\beta}}_{w} - \hat{\boldsymbol{\beta}}_{\mathrm{MLE}})$ is measured by $\rho^{-1}\mathbf{V}^{\mathrm{OS}}$ and the variation of $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} - \boldsymbol{\beta}_t)$ is measured by $\mathbf{M}^{-1}$. Note that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$, $\boldsymbol{\Sigma}_{\mathrm{wMLE}}$, $\mathbf{V}^{\mathrm{OS}}$, and $\mathbf{M}^{-1}$ are all fixed constant matrices that do not depend on ρ, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \leq \mathbf{V}^{\mathrm{OS}}$, and $\boldsymbol{\Sigma}_{\mathrm{wMLE}} = \boldsymbol{\Sigma}_{\mathrm{MLE}}$ if $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} = \mathbf{V}^{\mathrm{OS}}$. Thus, if ρ is small enough, $\hat{\boldsymbol{\beta}}_{uw}$ is more efficient than $\hat{\boldsymbol{\beta}}_{w}$ in estimating $\boldsymbol{\beta}_t$, and we do not need to require that $n/N \to 0$.*

Since the equality in (13) holds if $h(\mathbf{x}) = 1$, this indicates that for subsample obtained from local case-control subsampling with replacement, the weighted and un-weighted estimators have the same conditional asymptotic distribution.

4. Poisson subsampling

For the more efficient estimator $\hat{\boldsymbol{\beta}}_{uw}$ in Section 3 as well as the weighted estimator $\hat{\boldsymbol{\beta}}_{w}$, 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 $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ need to be used 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 for 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 subsampling, 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 generally not independent, although it is i.i.d conditional on the full data. As another advantage with Poisson subsampling, 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 subsampling is that the subsample size is always random. Due to this, we use n^* to denote the actual subsample size, and abuse the notation in this section to use n to denote the expected subsample size, i.e., $\mathbb{E}(n^*) = n$.

Note that $\{\pi_i^{\mathrm{OS}}(\boldsymbol{\beta})\}_{i=1}^N$ depend on the full data through the term in the denominator, $\sum_{i=1}^N |y_i - p(\mathbf{x}_i, \boldsymbol{\beta})|h(\mathbf{x}_i)$. Write $\Psi_N(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^N |y_i - p(\mathbf{x}_i, \boldsymbol{\beta})|h(\mathbf{x}_i)$, and denote its limit as $\Psi(\boldsymbol{\beta}) = \mathbb{E}\{|y - p(\mathbf{x}, \boldsymbol{\beta})|h(\mathbf{x})\}$. Note that $\Psi(\boldsymbol{\beta}_t) = 2\Phi(\boldsymbol{\beta}_t)$. The pilot subsample can be used to obtain an estimator of $\Psi(\boldsymbol{\beta}_t)$ to approximate $\Psi_N(\boldsymbol{\beta})$. Let $\hat{\Psi}_0$ be a pilot estimator of $\Psi(\boldsymbol{\beta}_t)$. Here, we focus on the Poisson subsampling procedure and assume that such $\hat{\Psi}_0$ is available and consistent. We will provide more details on $\hat{\Psi}_0$ in Section 5 and Section 7.

With $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ available, the Poisson subsampling procedure is described as the following. For $i = 1, ..., N$, calculate $\pi_i^p = |y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)/(N\hat{\Psi}_0)$, 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

$$\tilde{\boldsymbol{\beta}}_p = \arg\max_{\boldsymbol{\beta}} \ell_p^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n^*} (n\pi_i^{p*} \vee 1)\big\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* y_i^* + \log(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^*})\big\}, \tag{15}$$

and let $\hat{\boldsymbol{\beta}}_p = \tilde{\boldsymbol{\beta}}_p + \hat{\boldsymbol{\beta}}_0$. Note that here the actual subsample size n^* is random.

Poisson subsampling does not require to calculate π_i^p's all at once; each π_i^p is calculated 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 objective function in (15). The following theorem describes asymptotic properties of $\hat{\boldsymbol{\beta}}_p$.

Theorem 6 *Under Assumptions 1-2 and assume that $\hat{\boldsymbol{\beta}}_0$ is consistent, conditional on $\mathcal{D}_N$, as n_0, n, and N go to infinity, if $n/N \to 0$, then*

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

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \hat{\boldsymbol{\beta}}_{\text{wMLE}}) \longrightarrow \mathbb{N}(0,\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_\rho\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}), \tag{16}$$

in distribution, where

$$\boldsymbol{\Lambda}_\rho = \frac{\mathbb{E}\big[|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\{\Psi(\boldsymbol{\beta}_t) - \rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\}_+\mathbf{x}\mathbf{x}^{\mathrm{T}}\big]}{4\Psi^2(\boldsymbol{\beta}_t)}$$

with $\psi(\boldsymbol{\beta}) = y - p(\mathbf{x},\boldsymbol{\beta})$ and $\Psi(\boldsymbol{\beta}) = \mathbb{E}\{|y - p(\mathbf{x},\boldsymbol{\beta})|h(\mathbf{x})\}$, and $()_+$ means the positive part of the quantity, i.e., $a_+ = aI(a > 0)$.

Remark 7 *Similar to the case of Theorem 1,* (16) *implies that given $\mathcal{D}_N$ in probability, $\hat{\boldsymbol{\beta}}_p - \hat{\boldsymbol{\beta}}_{\text{wMLE}} = O_{P|\mathcal{D}_N}(n^{-1/2})$, which implies that $\hat{\boldsymbol{\beta}}_p - \hat{\boldsymbol{\beta}}_{\text{wMLE}} = O_P(n^{-1/2})$ unconditionally because if a sequence is stochastically bounded in conditional probability, then it is also stochastically bounded in unconditional probability (Xiong and Li, 2008; Cheng and Huang, 2010). Since $\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_t = O_P(N^{-1/2})$, we have $\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t = O_P(n^{-1/2} + N^{-1/2}) = O_P(n^{-1/2})$, showing that $\hat{\boldsymbol{\beta}}_p$ is $\sqrt{n}$-consistent to $\boldsymbol{\beta}_t$ unconditionally on the full data.*

Theorem 6 shows that with Poisson subsampling, the asymptotic variance-covariance matrices may differ for different sampling ratios n/N. In addition, comparing Theorems 1 and 6, we know that $\hat{\boldsymbol{\beta}}_{uw}$ and $\hat{\boldsymbol{\beta}}_p$ have the same asymptotic distribution if $n/N \to 0$. This is intuitive because if the sampling ratio n/N is small, sampling with replacement has close performance to sampling without replacement. However, if the sampling ratio n/N does not converge to zero, then $\hat{\boldsymbol{\beta}}_{uw}$ and $\hat{\boldsymbol{\beta}}_p$ have the same asymptotic mean but different asymptotic variance-covariance matrices. The following result compares the two asymptotic variance-covariance matrices.

Proposition 8 *If $\rho > 0$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ is a finite and positive definite matrix, then*

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

under the Loewner ordering.

This proposition shows that Poisson subsampling is more efficient than sampling with replacement.

5. Pilot estimate and practical implementation

Since $\{\pi_i^{\text{OS}}(\boldsymbol{\beta})\}_{i=1}^N$ 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 subsample using uniform subsampling or case-control subsampling. For uniform subsampling, all subsampling probabilities are equal, while for case-control subsampling, the subsampling probability for the

cases ($y_i = 1$) is different from that for the controls ($y_i = 0$). Let the subsampling probabilities used to take the pilot subsample be

$$\pi_{0i} = \frac{c_0(1 - y_i) + c_1 y_i}{N}, \tag{17}$$

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 $\pi_{0i} = N^{-1}$ corresponds to the uniform subsampling. This choice is recommended due to its simplicity if the proportion of 1's is close to 0.5 (Wang et al., 2018). If $c_0 \neq c_1$, then π_{0i}'s 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 $\hat{\boldsymbol{\beta}}_0$ can be obtained using the pilot subsample. For uniform subsampling, weighted and un-weighted estimators are the same. For case-control subsampling, we use un-weighted estimators with bias correction for both sampling with replacement and Poisson subsampling.

To obtain a final estimator, Wang et al. (2018) pooled the pilot subsample with the second stage subsample taken using approximated optimal subsampling probabilities. While this does not make a difference asymptotically since n_0 is typically a small term compared with n, i.e., $n_0 = o(n)$, using the pilot subsample helps to improve the finite sample performance in practical applications. However, pooling the raw samples may not be the most computationally efficient way of utilizing the pilot subsample. Since $\hat{\boldsymbol{\beta}}_0$ is already calculated, we can use it directly to improve the second stage estimator using the aggregation procedure in the divide-and-conquer method (Lin and Xie, 2011; Schifano et al., 2016). This avoids iterative calculations on the pilot subsample for the second time.

For subsampling with replacement, when the full data cannot be loaded into available RAM, special considerations have to be given in practical implementation. If the full data is larger than available RAM while subsampling probabilities $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ can still be loaded in available RAM, one can calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ by reading 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. A detailed procedure is provided in Section A of the appendix.

For Poisson subsampling, the pilot subsample can also be used to construct $\hat{\Psi}_0$ to approximate $\Psi_N(\boldsymbol{\beta})$. We use the following expression to obtain $\hat{\Psi}_0$.

$$\hat{\Psi}_0 = \frac{1}{N} \sum_{i=1}^{n_0^*} \frac{|y_i^{*0} - p(\mathbf{x}_i^{*0}, \hat{\boldsymbol{\beta}}_0)| h(\mathbf{x}_i^{*0})}{(n_0 \pi_{0i}^*) \wedge 1}, \tag{18}$$

where $(\mathbf{x}_i^{*0}, y_i^{*0})$'s are observations in the pilot subsample. If $h(\mathbf{x}) = \|L\mathbf{M}_N^{-1}\mathbf{x}\|$ for some L, then the pilot subsample is used to approximate $\mathbf{M}_N$ through

$$\hat{\mathbf{M}}_0 = \frac{1}{N} \sum_{i=1}^{n_0^*} \frac{\phi_i^{*0}(\hat{\boldsymbol{\beta}}_0) \mathbf{x}_i^{*0} (\mathbf{x}_i^{*0})^{\mathrm{T}}}{(n_0 \pi_{0i}^*) \wedge 1},$$

where $\phi_i^{*0}(\boldsymbol{\beta}) = p(\mathbf{x}_i^{*0}, \boldsymbol{\beta})\{1 - p(\mathbf{x}_i^{*0}, \boldsymbol{\beta})\}$. It can be verified that $\hat{\Psi}_0$ and $\hat{\mathbf{M}}_0$ converge in probability to $\Psi(\boldsymbol{\beta}_t)$ and $\mathbf{M}$, respectively.

Taking into account all aforementioned issues in this section, including how to obtain pilot estimates, how to combine them with the second stage estimates, as well as how to process data file line-by-line, we summarize practical implementation procedures in Algorithm 1 for sampling with replacement and in Algorithm 2 for Poisson subsampling.

Algorithm 1 More efficient estimation based on subsampling with replacement

Step 1: obtain the pilot $\hat{\boldsymbol{\beta}}_0$

(1) Take pilot subsample $(\mathbf{x}_i^{*0}, y_i^{*0})$, $i = 1, ..., n_0$ using sampling with replacement according to subsampling probabilities $\{\pi_{0i}\}_{i=1}^N$ in (17).

(2) Calculate

$$\tilde{\boldsymbol{\beta}}_0 = \arg\max_{\boldsymbol{\beta}} \ell_{uw}^{*0}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n_0} \left\{ \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^{*0} y_i^{*0} - \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^{*0}}\right) \right\},$$

and let $\hat{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\beta}}_0 + \mathbf{b}$, where $\mathbf{b} = \{\log(c_0/c_1), 0, ..., 0\}^{\mathrm{T}}$.

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

(1) Calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ defined in equation (4); take subsample $(\mathbf{x}_i^*, y_i^*)$, $i = 1, ..., n$ according to sampling probabilities $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ using sampling with replacement.

(2) Calculate

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

and let $\hat{\boldsymbol{\beta}}_{uw} = \tilde{\boldsymbol{\beta}}_{uw} + \hat{\boldsymbol{\beta}}_0$.

Step 3: combine the two estimators $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_{uw}$

Calculate

$$\check{\boldsymbol{\beta}}_{uw} = \{\ddot{\ell}_{uw}^{*0}(\tilde{\boldsymbol{\beta}}_0) + \ddot{\ell}_{uw}^{*}(\tilde{\boldsymbol{\beta}}_{uw})\}^{-1} \{\ddot{\ell}_{uw}^{*0}(\tilde{\boldsymbol{\beta}}_0)\hat{\boldsymbol{\beta}}_0 + \ddot{\ell}_{uw}^{*}(\tilde{\boldsymbol{\beta}}_{uw})\hat{\boldsymbol{\beta}}_{uw}\},$$

where $\ddot{\ell}_{uw}^{*0}(\tilde{\boldsymbol{\beta}}_0) = \sum_{i=1}^{n_0} \phi_i^{*0}(\tilde{\boldsymbol{\beta}}_0)\mathbf{x}_i^{*0}(\mathbf{x}_i^{*0})^{\mathrm{T}}$, $\ddot{\ell}_{uw}^{*}(\tilde{\boldsymbol{\beta}}_{uw}) = \sum_{i=1}^{n} \phi_i^{*}(\tilde{\boldsymbol{\beta}}_{uw})\mathbf{x}_i^{*}(\mathbf{x}_i^{*})^{\mathrm{T}}$, and $\phi_i^*(\boldsymbol{\beta}) = p(\mathbf{x}_i^*, \boldsymbol{\beta})\{1 - p(\mathbf{x}_i^*, \boldsymbol{\beta})\}$. .

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

$$\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{uw}) = \{\ddot{\ell}_{uw}^{*0}(\tilde{\boldsymbol{\beta}}_0) + \ddot{\ell}_{uw}^{*}(\tilde{\boldsymbol{\beta}}_{uw})\}^{-1} \Bigg[\sum_{i=1}^{n_0} \{\psi_i^{*0}(\tilde{\boldsymbol{\beta}}_0)\}^2 \mathbf{x}_i^{*0}(\mathbf{x}_i^{*0})^{\mathrm{T}} + \sum_{i=1}^{n} \{\psi_i^{*}(\tilde{\boldsymbol{\beta}}_{uw})\}^2 \mathbf{x}_i^{*}(\mathbf{x}_i^{*})^{\mathrm{T}} \Bigg] \{\ddot{\ell}_{uw}^{*0}(\tilde{\boldsymbol{\beta}}_0) + \ddot{\ell}_{uw}^{*}(\tilde{\boldsymbol{\beta}}_{uw})\}^{-1}, \quad (19)$$

where $\psi_i^*(\boldsymbol{\beta}) = y_i^* - p(\mathbf{x}_i^*, \boldsymbol{\beta})$.

Algorithm 2 More efficient estimation based on Poisson subsampling

Step 1: obtain the pilots $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$

(1) For $i = 1, ..., N$, calculate $\pi_{0i} = \frac{c_0(1-y_i)+c_1 y_i}{N}$, generate $u_{0i} \sim U(0,1)$, and add $(\mathbf{x}_i, y_i, \pi_{i1})$ in the subsample if $u_{0i} \leq n_0\pi_{0i}$.

(2) For the obtained subsample, say $(\mathbf{x}_i^{*0}, y_i^{*0}, \pi_{0i}^{*0})$, $i = 1, ..., n_0^*$, calculate

$$\tilde{\boldsymbol{\beta}}_0 = \arg\max_{\boldsymbol{\beta}} \ell_p^{*0}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n_0^*} (n\pi_{0i}^{*0} \vee 1)\big\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^{*0} y_i^{*0} + \log\big(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^{*0}}\big)\big\},$$

let $\hat{\boldsymbol{\beta}}_0 = \tilde{\boldsymbol{\beta}}_0 + \mathbf{b}$, and then calculate $\hat{\Psi}_0$ in equation (18).

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

(1) For $i = 1, ..., N$, calculate $\pi_i^p = \frac{|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}{N\hat{\Psi}_0}$, 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

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

and let $\hat{\boldsymbol{\beta}}_p = \tilde{\boldsymbol{\beta}}_p + \hat{\boldsymbol{\beta}}_0$.

Step 3: combine the two estimators $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_p$

Calculate

$$\check{\boldsymbol{\beta}}_p = \{\ddot{\ell}_p^{*0}(\tilde{\boldsymbol{\beta}}_0) + \ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p)\}^{-1}\{\ddot{\ell}_p^{*0}(\tilde{\boldsymbol{\beta}}_0)\hat{\boldsymbol{\beta}}_0 + \ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p)\hat{\boldsymbol{\beta}}_p\},$$

where $\ddot{\ell}_p^{*0}(\tilde{\boldsymbol{\beta}}_0) = \sum_{i=1}^{n_0^*} \phi_i^{*0}(\tilde{\boldsymbol{\beta}}_0)\mathbf{x}_i^{*0}(\mathbf{x}_i^{*0})^{\mathrm{T}}$ and $\ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p) = \sum_{i=1}^{n^*} \phi_i^*(\tilde{\boldsymbol{\beta}}_p)\mathbf{x}_i^*(\mathbf{x}_i^*)^{\mathrm{T}}$.

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

$$\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p) = \{\ddot{\ell}_p^{*0}(\tilde{\boldsymbol{\beta}}_0) + \ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p)\}^{-1}\Bigg[\sum_{i=1}^{n_0^*}\{\psi_i^{*0}(\tilde{\boldsymbol{\beta}}_0)\}^2\mathbf{x}_i^{*0}(\mathbf{x}_i^{*0})^{\mathrm{T}} + \sum_{i=1}^{n^*}\{\psi_i^*(\tilde{\boldsymbol{\beta}}_p)\}^2\mathbf{x}_i^*(\mathbf{x}_i^*)^{\mathrm{T}}\Bigg]\{\ddot{\ell}_p^{*0}(\tilde{\boldsymbol{\beta}}_0) + \ddot{\ell}_p^*(\tilde{\boldsymbol{\beta}}_p)\}^{-1}. \quad (20)$$

Remark 9 *In Algorithm 1 and Algorithm 2, if $n_0 = o(n)$, then the results for $\hat{\boldsymbol{\beta}}_{uw}$ in Theorem 1 hold for $\check{\boldsymbol{\beta}}_{uw}$ and the results for $\hat{\boldsymbol{\beta}}_p$ in Theorem 6 hold for $\check{\boldsymbol{\beta}}_p$ as well. This is because $\{\ddot{\ell}^{*0}_{uw}(\tilde{\boldsymbol{\beta}}_0)+\ddot{\ell}^{*}_{uw}(\tilde{\boldsymbol{\beta}}_{uw})\}^{-1}\ddot{\ell}^{*0}_{uw}(\tilde{\boldsymbol{\beta}}_0)\sqrt{n}(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_t) = O_p(\sqrt{n_0}/\sqrt{n}) = o_P(1)$ and $\{\ddot{\ell}^{*0}_{uw}(\tilde{\boldsymbol{\beta}}_0)+\ddot{\ell}^{*}_{uw}(\tilde{\boldsymbol{\beta}}_{uw})\}^{-1}\ddot{\ell}_{uw}(\tilde{\boldsymbol{\beta}}_{uw}) \to 1$ in probability. The reason for $\check{\boldsymbol{\beta}}_p$ is similar.*

Remark 10 *In Algorithm 1 and Algorithm 2, to combine the two stage estimates using the second derivative of the objective functions, the inconsistent estimators $\tilde{\boldsymbol{\beta}}_0$, and $\tilde{\boldsymbol{\beta}}_{uw}$ 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 objective function.*

Remark 11 *The variance-covariance estimators $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{uw})$ in (19) and $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p)$ in (20) can be replaced by the following simplified estimators,*

$$\hat{\mathbb{V}}_s(\check{\boldsymbol{\beta}}_{uw}) = \{\ddot{\ell}^{*0}_{uw}(\tilde{\boldsymbol{\beta}}_0) + \ddot{\ell}^{*}_{uw}(\tilde{\boldsymbol{\beta}}_{uw})\}^{-1} \qquad \text{and} \qquad \hat{\mathbb{V}}_s(\check{\boldsymbol{\beta}}_p) = \{\ddot{\ell}^{*0}_{p}(\tilde{\boldsymbol{\beta}}_0) + \ddot{\ell}^{*}_{p}(\tilde{\boldsymbol{\beta}}_p)\}^{-1},$$

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

Remark 12 *The time complexity of Algorithm 1 is the same as that of Algorithm 2 in Wang et al. (2018). The major computing time is to calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ in Step 2, but it does not require iterative calculations on the full data. Once $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ are available, the calculations of $\hat{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_{uw}$ are fast because they are done on the subsamples only. To calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$, the required time varies. For π_i^{Lopt}, the required time is $O(Nd)$; for π_i^{Aopt}, the required time is $O(Nd^2)$. Thus, the time complexity of Algorithm 1 with π_i^{Lopt} is $O(Nd)$ and the time complexity with π_i^{Aopt} is $O(Nd^2)$, if the sampling ratio n/N is much smaller than one.*

6. Unconditional distribution

Asymptotic distributional results in Sections 3 and 4, as well as in Wang et al. (2018), are about conditional distributions, i.e., they are about conditional distributions of subsample-based estimators given the full data. We investigate the unconditional distribution of $\hat{\boldsymbol{\beta}}_p$ in this section.

Theorem 13 *Under Assumptions 1 and 2, if the pilot estimators are obtained from a uniform subsample of sample size $n_0 = o(\sqrt{N})$ 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 $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ are independent of the data $\mathcal{D}_N$, then as n_0, n, and N go to infinity such that $n/N \to \rho \in [0,1)$, we have*

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

in distribution, where

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

Remark 14 *If the pilot estimators $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ are obtained through the full data $\mathcal{D}_N$, 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^{Lopt}, π_i^{Aopt}, and the local case-control subsampling. In general, if $h(\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|^a$ for some matrix $\mathbf{A}$ and constant $a \geq 0$, then the four additional moment conditions reduce to one requirement of $\mathbb{E}\{h^3(\mathbf{x})\|\mathbf{x}\|^3\} < \infty$.*

Remark 15 *If $\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x}) \leq \Psi(\boldsymbol{\beta}_t)$ almost surely, then $\boldsymbol{\Lambda}_u$ reduced to $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}$ and as a result $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_u\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ reduces to $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$. Furthermore, if the subsampling probabilities are propositional to the local case-control subsampling probabilities, i.e., $h(\mathbf{x}) = 1$, then $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_u\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ reduces to $4\mathbb{E}\{\phi(\boldsymbol{\beta})\}\mathbf{M}^{-1}$. For the uniform Poisson subsampling estimator, the unconditional asymptotic variance-covariance matrix (scaled by n) is $\mathbf{M}^{-1}$. From (14), with the same expected subsample size, the proposed method has a higher estimation efficiency than subsampling proportional to the local case-control subsampling probabilities, which is more efficient than the uniform Poisson subsampling approach.*

Remark 16 *Fithian and Hastie (2014)'s investigation corresponds to the case of $h(\mathbf{x}) = 1$ and $\rho = 2\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}$. For this scenario in Theorem 13, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_p$ reduces to $2N^{-1}\mathbf{M}^{-1}$, which is the same as obtained in Fithian and Hastie (2014). This result is particularly neat in the fact that this asymptotic variance-covariance matrix is proportional to that from the full data MLE with a multiplier of 2. The result in Theorem 13 is more general. It shows that if $h(\mathbf{x}) = 1$, then as long as $\rho|\psi(\boldsymbol{\beta}_t)| \leq 2\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}$ (which is satisfied if $\rho = 2\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}$), the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_p$ can be written as*

$$\frac{4\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}}{\rho N}\mathbf{M}^{-1},$$

which is proportional to that of the full data MLE with a multiplier of $4\rho^{-1}\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}$. We need to emphasize that this simple representation holds only when $\rho|\psi(\boldsymbol{\beta}_t)| \leq 2\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}$ almost surely. If the subsampling ratio ρ gets closer to one, the asymptotic variance-covariance matrix in (21) may not be simplified.

From Theorems 6 and 13, 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 ratio 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 ratio 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 $\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_t = O_P(N^{-1/2})$. 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 matrix for the unconditional distribution to be larger than that for the conditional distribution. Indeed this is true, and we present it in the following proposition.

Proposition 17 *If $\rho > 0$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ is a finite and positive definite matrix, then*

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_u\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \geq \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} > \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\boldsymbol{\Lambda}_\rho\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}, \tag{22}$$

under the Loewner ordering. Furthermore, if $\mathbb{P}\{\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x}) > \Psi(\boldsymbol{\beta}_t)\} > 0$*, then the "*$\geq$*" sign in* (22) *can be replaced by ">", the strict great sign.*

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 13 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, $h(\mathbf{x}) = 1$, and the additional moment requirement is that $\mathbb{E}(\|\mathbf{x}\|^3) < \infty$.

7. Misspecifications

In this section, we discuss the effect when the pilot estimates are misspecified or when the model is misspecified. Pilot estimates misspecification often occurs when they are from other data sources or when they are calculated based on convenient subsamples, e.g., using the first n_0 observations in the full data to calculate them. In these cases, it is reasonable to assume that the pilot estimates are independent of $\mathcal{D}_N$ and we use this assumption in this section.

7.1. Pilot estimates misspecification

Here, we assume that the model is correctly specified but the pilot estimates $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ converge to limits that are different from the true parameters for the current data. Interestingly, in this case, the proposed estimators are still consistent and no specific convergence rate is required for $\hat{\boldsymbol{\beta}}_0$ or $\hat{\Psi}_0$.

The following theorem describes the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{uw}$, the estimator based on subsampling with replacement. Note that $\hat{\Psi}_0$ is not required by $\hat{\boldsymbol{\beta}}_{uw}$.

Theorem 18 *When the logistic regression model in* (1) *is correctly specified and the pilot estimator* $\hat{\boldsymbol{\beta}}_0$ *that is independent of* $\mathcal{D}_N$ *is inconsistent, i.e.,* $\hat{\boldsymbol{\beta}}_0 \to \boldsymbol{\beta}_0$ *in probability for some* $\boldsymbol{\beta}_0$ *that is different from* $\boldsymbol{\beta}_t$*, then under Assumptions 1-3, conditional on* $\mathcal{D}_N$*, as* n*, and* N *go to infinity,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\text{wMLE}}) \longrightarrow \mathbb{N}\big\{\mathbf{0},\ \Psi(\boldsymbol{\beta}_0)\boldsymbol{\varsigma}_a^{-1}\big\},$$

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}\big\{\mathbf{0},\ \Psi(\boldsymbol{\beta}_0)\boldsymbol{\varsigma}_a^{-1}\big\},$$

in distribution, where $\Psi(\boldsymbol{\beta}_0) = \mathbb{E}\{|\psi(\boldsymbol{\beta}_0)|h(\mathbf{x})\}$ *and*

$$\boldsymbol{\varsigma}_a = \mathbb{E}[\{1 - p(\mathbf{x}, \boldsymbol{\beta}_t)\}p(\mathbf{x}, \boldsymbol{\beta}_0)p(\mathbf{x}, \boldsymbol{\beta}_t - \boldsymbol{\beta}_0)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}].$$

Here $\hat{\boldsymbol{\beta}}_{\text{wMLE}}$ *satisfies that,*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\varsigma}_a^{-1}\boldsymbol{\varsigma}_b\boldsymbol{\varsigma}_a^{-1}\big),$$

in distribution, where

$$\boldsymbol{\varsigma}_b = \mathbb{E}\big\{\phi(\boldsymbol{\beta}_0)\phi(\boldsymbol{\beta}_t - \boldsymbol{\beta}_0)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big\}.$$

Remark 19 *If $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_t$, then direct calculations show that $\Psi(\boldsymbol{\beta}_0)\boldsymbol{\varsigma}_a^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$ and $\boldsymbol{\varsigma}_a^{-1}\boldsymbol{\varsigma}_b\boldsymbol{\varsigma}_a^{-1} = \boldsymbol{\Sigma}_{\mathrm{wMLE}}$, that is, the results in Theorem 18 reduce to the same results in Theorem 1.*

Remark 20 *If the pilot estimator $\hat{\boldsymbol{\beta}}_0$ is very wrong such that $\boldsymbol{\beta}_t^{\mathrm{T}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 < 0$, i.e., $p(\mathbf{x}, \boldsymbol{\beta}_t) > 0.5 > p(\mathbf{x}, \boldsymbol{\beta}_0)$ or $p(\mathbf{x}, \boldsymbol{\beta}_t) < 0.5 < p(\mathbf{x}, \boldsymbol{\beta}_0)$, then it can be shown that*

$$\Psi(\boldsymbol{\beta}_0)\boldsymbol{\varsigma}_a^{-1} > \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}.$$

Detailed proof for this result is presented in Section B.4.1 of the appendix.

The following theorem describes the asymptotic distribution of $\hat{\boldsymbol{\beta}}_p$, the estimator based on Poisson subsampling. Note that $\hat{\boldsymbol{\beta}}_p$ requires both $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$.

Theorem 21 *Assume that the logistic regression model is correctly specified, and the pilot estimators $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ are independent of $\mathcal{D}_N$ and they are inconsistent, i.e., $\hat{\boldsymbol{\beta}}_0 \to \boldsymbol{\beta}_0$ and $\hat{\Psi}_0 \to \Psi_0$ in probability for some $\boldsymbol{\beta}_0$ and Ψ_0, respectively. Under Assumptions 1-2, conditional on $\mathcal{D}_N$, as n and N go to infinity, if $n/N \to 0$, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \Psi_0\boldsymbol{\varsigma}_a^{-1}\big),$$

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \Psi_0\boldsymbol{\varsigma}_a^{-1}\boldsymbol{\varsigma}_c\boldsymbol{\varsigma}_a^{-1}\big),$$

in distribution, where

$$\boldsymbol{\varsigma}_c = \mathbb{E}\Big[|\psi(\boldsymbol{\beta}_0)|\big\{1 - \rho\Psi_0^{-1}|\psi(\boldsymbol{\beta}_0)|h(\mathbf{x})\big\}_+\psi^2(\boldsymbol{\beta}_t - \boldsymbol{\beta}_0)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big].$$

Remark 22 *If $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_t$ and $\Psi_0 = \Psi(\boldsymbol{\beta}_t)$, then direct calculations show that $\Psi_0^{-1}\boldsymbol{\varsigma}_c = \boldsymbol{\Lambda}_\rho$, and thus the results in Theorem 21 reduce to the same results in Theorem 6.*

Remark 23 *We have a result similar to that in Proposition 8. By direct calculation, we know that*

$$\boldsymbol{\varsigma}_c < \mathbb{E}\Big[|\psi(\boldsymbol{\beta}_0)|\psi^2(\boldsymbol{\beta}_t - \boldsymbol{\beta}_0)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big] = \boldsymbol{\varsigma}_a,$$

under the Loewner ordering, which indicates that

$$\Psi_0\boldsymbol{\varsigma}_a^{-1}\boldsymbol{\varsigma}_c\boldsymbol{\varsigma}_a^{-1} < \Psi_0\boldsymbol{\varsigma}_a^{-1}.$$

Thus, when the pilot estimators are misspecified, Poisson subsampling still has a higher estimation efficiency compared with subsampling with replacement if $\Psi_0 = \mathbb{E}\{|\psi(\boldsymbol{\beta}_0)|h(\mathbf{x})\}$, which is the case if $\hat{\Psi}_0$ is constructed from a pilot subsample.

7.2. Model misspecification

In this section, we consider the case when the logistic regression model is misspecified, namely, the model in (1) is not correct. Instead, we assume that the true probability of $y = 1$ given $\mathbf{x}$ is

$$\mathbb{P}(y = 1|\mathbf{x}) = p_t(\mathbf{x}),$$

for some unknown function $p_t(\mathbf{x})$. When the logistic regression model is misspecified, we need to define the meaning of consistency because there is no true $\boldsymbol{\beta}$ any more. In this case, consistency often means that the estimator converges to a limit that minimizes expected loss with respect to a specified loss function. Here, if we denote the limit as $\boldsymbol{\beta}_l$ and define it to be the minimizer of

$$\mathbb{E}\Big\{ - p_t(\mathbf{x})h(\mathbf{x})\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + h(\mathbf{x}) \log\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}}\right)\Big\},$$

then $\boldsymbol{\beta}_l$ satisfies

$$\mathbb{E}\big[\big\{p_t(\mathbf{x}) - p(\mathbf{x}, \boldsymbol{\beta}_l)\big\}h(\mathbf{x})\mathbf{x}\big] = \mathbf{0},$$

where $p(\mathbf{x}, \boldsymbol{\beta}) = e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}(1 + e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}})^{-1}$.

Now we investigate the asymptotic properties of the proposed estimators under model misspecification. In this case, we need to assume that the pilot estimators are consistent which is also required in the local case-control subsampling method. In addition, to investigate the asymptotic normality, we also need an additional assumption on the convergence rate of the pilot estimator $\hat{\boldsymbol{\beta}}_0$.

The following theorem describes the asymptotic behavior of the estimator $\hat{\boldsymbol{\beta}}_{uw}$ based on subsampling with replacement.

Theorem 24 *Assume that the pilot sample is independent of $\mathcal{D}_N$ and the pilot estimator $\hat{\boldsymbol{\beta}}_0$ satisfies that $\sqrt{n_0}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_l) \to N(\mathbf{0}, \boldsymbol{\Sigma}_0)$ in distribution. Under Assumptions 1-3, if $n_0/N \to \rho_0$ and $n/N \to \rho$ with $\rho_0, \rho \in (0, 1)$, then conditional on $\mathcal{D}_N$, as n_0, n, and N go to infinity,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \omega\boldsymbol{\kappa}_a^{-1}\big) \tag{23}$$

in distribution, where

$$\boldsymbol{\kappa}_a = \frac{1}{4}\mathbb{E}\big[\{p_t(\mathbf{x}) - 2p_t(\mathbf{x})p(\mathbf{x}, \boldsymbol{\beta}_l) + p(\mathbf{x}, \boldsymbol{\beta}_l)\}h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big],$$
$$\omega = \mathbb{E}\big[\{p_t(\mathbf{x}) - 2p_t(\mathbf{x})p(\mathbf{x}, \boldsymbol{\beta}_l) + p(\mathbf{x}, \boldsymbol{\beta}_l)\}h(\mathbf{x})\big],$$

and $\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}$ satisfies that

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}} - \boldsymbol{\beta}_l) \longrightarrow \mathbb{N}\Big\{\mathbf{0},\ \boldsymbol{\kappa}_a^{-1}\big(\boldsymbol{\kappa}_b + \rho_0^{-1}\boldsymbol{\kappa}_c\boldsymbol{\Sigma}_0\boldsymbol{\kappa}_c\big)\boldsymbol{\kappa}_a^{-1}\Big\}, \tag{24}$$

in distribution, with

$$\boldsymbol{\kappa}_b = \frac{1}{4}\mathbb{E}\big[\{p_t(\mathbf{x}) - 2p_t(\mathbf{x})p(\mathbf{x}, \boldsymbol{\beta}_l) + p^2(\mathbf{x}, \boldsymbol{\beta}_l)\}h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big],\ \textit{and}$$
$$\boldsymbol{\kappa}_c = \frac{1}{4}\mathbb{E}\big[\{1 - 2p(\mathbf{x}, \boldsymbol{\beta}_l)\}\{p_t(\mathbf{x}) - p(\mathbf{x}, \boldsymbol{\beta}_l)\}h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big].$$

Remark 25 *If the model is correctly specified, i.e.,* $p_t(\mathbf{x}) = p(\mathbf{x}, \boldsymbol{\beta}_t)$*, then* $\omega\boldsymbol{\kappa}_a^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}$*,* $\boldsymbol{\kappa}_b = \frac{1}{4}\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$ *and* $\boldsymbol{\kappa}_c = \mathbf{0}$*, and therefore the results in Theorem 24 reduce to the same expressions as those in Theorem 1.*

From Theorem 24, with model misspecification, it is critical to have a good pilot estimator $\hat{\boldsymbol{\beta}}_0$. Note that the pilot sample size is typically much smaller than the full data sample size, so ρ_0 can be close to zero. From (24), we see that the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}$ can be inflated by a small pilot sample size.

The following theorem presents asymptotic results for the estimator based on Poisson subsampling.

Theorem 26 *Assume that the pilot sample is independent of* $\mathcal{D}_N$ *and the pilot estimators satisfy that* $\sqrt{n_0}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_l) \to N(\mathbf{0}, \boldsymbol{\Sigma}_0)$ *in distribution and* $\hat{\Psi}_0 \to \omega$ *in probability. Under Assumptions 1-2, if* $n_0/N \to \rho_0$ *and* $n/N \to \rho$ *with* $\rho_0, \rho \in (0, 1)$*, then conditional on* $\mathcal{D}_N$*, as* n_0*,* n*, and* N *go to infinity,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}) \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\kappa}_a^{-1}\boldsymbol{\kappa}_d\boldsymbol{\kappa}_a^{-1}\big).$$

in distribution, where $\boldsymbol{\kappa}_a$ *and* $\boldsymbol{\kappa}_b$ *are defined in Theorem 24, and*

$$\boldsymbol{\kappa}_d = \frac{1}{4}\mathbb{E}\Big[|\psi(\boldsymbol{\beta}_l)|\{1 - \rho\omega^{-1}|\psi(\boldsymbol{\beta}_l)|h(\mathbf{x})\}_+ h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big].$$

Remark 27 *Similarly to Proposition 8, we have that*

$$\boldsymbol{\kappa}_a^{-1}\boldsymbol{\kappa}_d\boldsymbol{\kappa}_a^{-1} < \boldsymbol{\kappa}_a^{-1}$$

under the Loewner ordering, indicating that the estimator based on Poisson subsampling has a smaller conditional variance-covariance matrix.

For Poisson subsampling, compared with $\hat{\boldsymbol{\beta}}_0$, we require a much weaker assumption on $\hat{\Psi}_0$; we only need it to converge without specifying certain convergence rate. The reason is that the effect of $\hat{\Psi}_0$ on all the subsampling probabilities are the same and it mainly controls the expected subsample size, while $\hat{\boldsymbol{\beta}}_0$ affects individual subsampling probabilities differently corresponding to different values of $\mathbf{x}_i$ and y_i.

8. Numerical evaluations

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

8.1. 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. (2018). 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. (2018).

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

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

where $\check{\boldsymbol{\beta}}_{\text{new}} = \check{\boldsymbol{\beta}}_{uw}$ for the subsampling with replacement estimator described in Algorithm 1 and $\check{\boldsymbol{\beta}}_{\text{new}} = \check{\boldsymbol{\beta}}_p$ for Poisson subsampling estimator described in Algorithm 2. We calculate empirical MSEs from $S = 1000$ subsamples using

$$\text{MSE}(\check{\boldsymbol{\beta}}) = \frac{1}{S}\sum_{s=1}^{S} \|\check{\boldsymbol{\beta}}^{(s)} - \boldsymbol{\beta}_t\|^2, \tag{25}$$

where $\check{\boldsymbol{\beta}}^{(s)}$ is the estimate from the s-th subsample. We fixed the first step sample size $n_0 = 200$ and choose n to be 100, 200, 400, 600, 800, and 1000. This is the same setup used in Wang et al. (2018).

Figure 1 presents the relative efficiency of $\check{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_p$ based on two different choices of π_i^{OS}: π_i^{Aopt} and π_i^{Lopt}. It is seen that in general $\check{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_p$ are more efficient than $\check{\boldsymbol{\beta}}_w$. Among the six cases, the only case that $\check{\boldsymbol{\beta}}_w$ can be more efficient is when $\mathbf{x}$ has a T_3 distribution and π^{Lopt} is used, but the difference is not very significant. For all other cases, $\check{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_p$ are more efficient. For example, when $\mathbf{x}$ has the nzNormal distribution, $\check{\boldsymbol{\beta}}_p$ can be 250% as efficient as $\check{\boldsymbol{\beta}}_w$ if π^{Aopt} is used. Between $\check{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_p$, $\check{\boldsymbol{\beta}}_p$ is more efficient than $\check{\boldsymbol{\beta}}_{uw}$ 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.

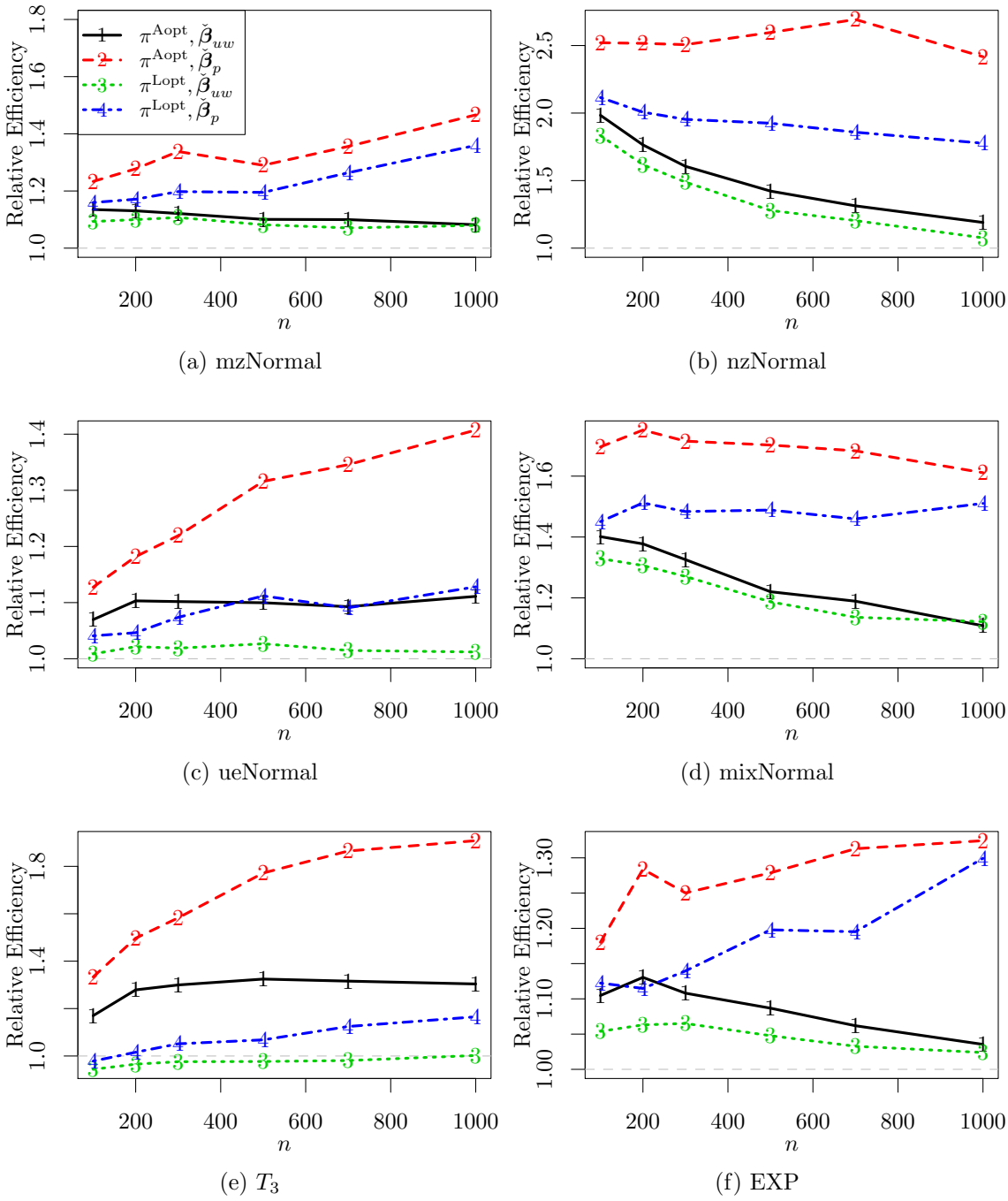
To evaluate the performance of the proposed method with different choices of the subsampling probabilities for subsampling with replacement and Poisson subsampling, Figure 2 plots empirical MSEs of using π^{Aopt}, π^{Lopt}, π^{lcc} (local case-control), and the uniform subsampling probability. In general, π^{Aopt} with Poisson subsampling has the smallest empirical MSEs while uniform subsampling with replacement has the worst estimation efficiency. This agrees with our theoretical results: 1) π^{Aopt} minimizes the asymptotic MSE of the parameter estimator which corresponds to the empirical MSE defined in (25) for the experiments, while π^{Lopt} minimizes the asymptotic MSE of a transformed parameter estimator, and 2) Poisson subsampling has a higher estimation efficiency compared with subsampling with replacement.

To assess the performance of $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{uw})$ in (19) and $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p)$ in (20), we use $\text{tr}\{\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{uw})\}$ and $\text{tr}\{\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_p)\}$ to estimate the MSEs of $\check{\boldsymbol{\beta}}_{uw}$ 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 not be appropriate if n/N is not small enough. Figure 3 presents the results for using π^{Lopt}. Results for using π^{Aopt} are similar and thus are omitted. Note

that our purpose here is to evaluate the quality of $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{uw})$ in (19) and $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{p})$ in (20), so in this figure we plot the original empirical and estimated MSEs without scaling then using the MSEs of $\check{\boldsymbol{\beta}}_{w}$. Here, a closer value between the estimated MSE and the empirical MSE indicates a better performance of $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{uw})$ or $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{p})$. From Figure 3, the estimated MSEs are very close to the empirical MSEs, except for the case of nzNormal covariate for subsampling with replacement. In 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. (2018) also has a similar problem of underestimation. For Poisson subsampling, the problem of underestimation from $\hat{\mathbb{V}}(\check{\boldsymbol{\beta}}_{p})$ is not significant.

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 (Dua and Karra Taniskidou, 2017). The data set contains a binary response variable indicating whether a process produces 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 4 gives the relative efficiency of $\check{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_{p}$ to $\check{\boldsymbol{\beta}}_{w}$ for both π_i^{Lopt} and π_i^{Aopt}. It is seen that when π_i^{Aopt} are used, $\check{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_{p}$ always outperform $\check{\boldsymbol{\beta}}_{w}$. When π_i^{Lopt} are used, $\check{\boldsymbol{\beta}}_{uw}$ 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 $\check{\boldsymbol{\beta}}_{uw}$ and π^{Aopt} dominates π^{Lopt} in estimation efficiency.

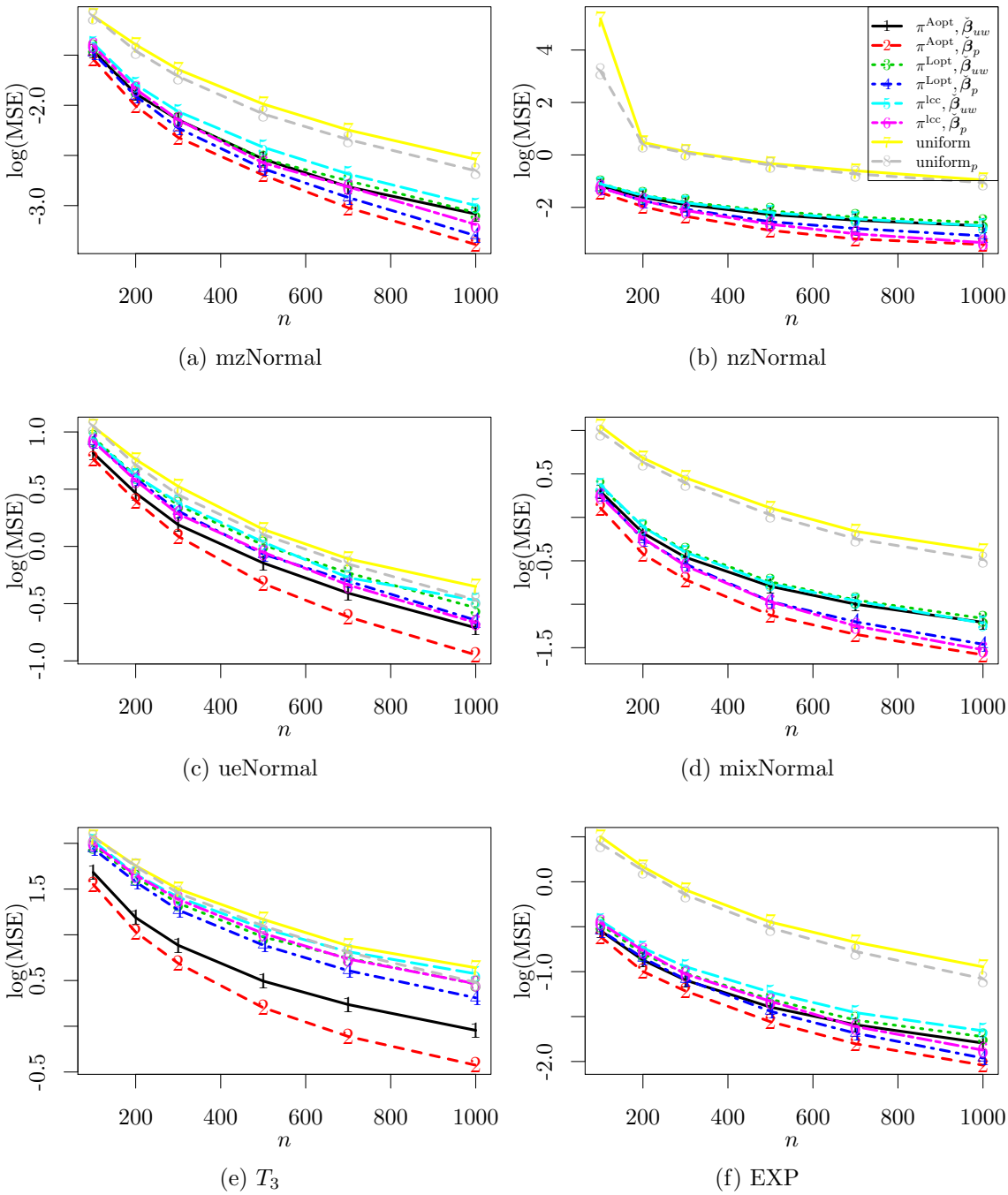


(a) mzNormal (b) nzNormal

(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 $n_0 = 200$. A relative efficiency larger than one means the associate method is more efficient than the original OSMAC estimator.

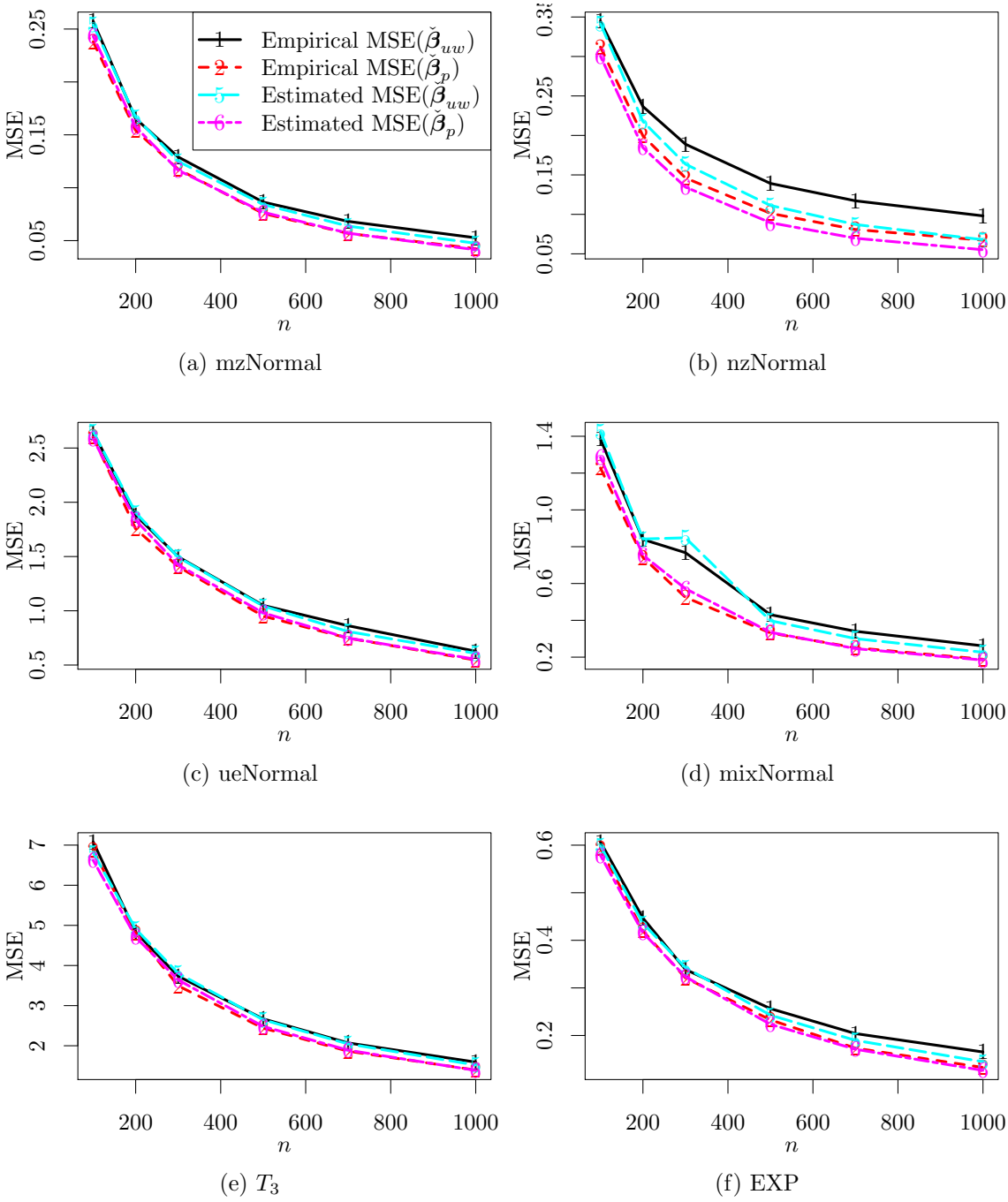


(a) mzNormal (b) nzNormal

(c) ueNormal (d) mixNormal

(e) T_3 (f) EXP

Figure 2: MSE for different subsampling probabilities with second step subsample size n and a fixed first step subsample size $n_0 = 200$. Logarithm is taken on MSEs for better presentation.



(a) mzNormal (b) nzNormal

(c) ueNormal (d) mixNormal

(e) T_3 (f) EXP

Figure 3: Empirical and estimated MSEs for different second step subsample size n based on π^{Lopt} with the first step subsample size being fixed at $n_0 = 200$.

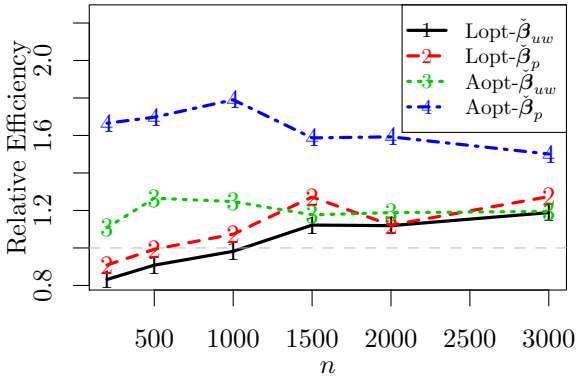


Figure 4: Relative efficiency for the SUSY data set with $n_0 = 200$ and different second step subsample size n. The gray horizontal dashed line is the reference line when relative efficiency is one.

8.2. 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 complexity, so they should have similar computational efficiency as the weighted estimator. For Poisson subsampling, 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 $n_0 = 200$ and $n = 1000$.

Table 1 gives the required CPU times (in seconds) to obtain $\check{\boldsymbol{\beta}}_w$, $\check{\boldsymbol{\beta}}_{uw}$, and $\check{\boldsymbol{\beta}}_p$, using π_i^{Lopt} and π_i^{Aopt}. The computing times for using the full data (Full) are also given for comparisons. It is seen that $\check{\boldsymbol{\beta}}_{uw}$ and $\check{\boldsymbol{\beta}}_p$ are a little faster than $\check{\boldsymbol{\beta}}_w$ but the advantages are not very significant. The reason is that the original OSMAC estimator $\check{\boldsymbol{\beta}}_{uw}$ pools the pilot subsample with the second stage subsample and performs iterative calculations on the pilot subsample twice, while the proposed method combines the pilot estimator with the second stage estimator which only requires iterative calculations on the pilot subsample once. Since the pilot subsample size is small, the difference is not 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 generated and kept in the RAM. Here $n_0 = 200$, $n = 1000$, 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
π_i^{Lopt}, $\check{\boldsymbol{\beta}}_w$	0.14	0.13	0.45	5.24
π_i^{Lopt}, $\check{\boldsymbol{\beta}}_{uw}$	0.08	0.11	0.41	3.71
π_i^{Lopt}, $\check{\boldsymbol{\beta}}_p$	0.08	0.11	0.43	3.88
π_i^{Aopt}, $\check{\boldsymbol{\beta}}_w$	0.13	0.32	3.31	35.15
π_i^{Aopt}, $\check{\boldsymbol{\beta}}_{uw}$	0.12	0.31	3.29	34.98
π_i^{Aopt}, $\check{\boldsymbol{\beta}}_p$	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^{Aopt} does not cost much more time than using π^{Lopt}. 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 subsampling 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 $n_0 = 200$, $n = 1000$, 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
π_i^{Lopt}, $\check{\boldsymbol{\beta}}_w$	4.26	41.60	441.46	4374.94
π_i^{Lopt}, $\check{\boldsymbol{\beta}}_{uw}$	4.13	41.42	413.09	4384.99
π_i^{Lopt}, $\check{\boldsymbol{\beta}}_p$	2.77	27.58	272.32	2699.13
π_i^{Aopt}, $\check{\boldsymbol{\beta}}_w$	4.43	41.75	434.96	4393.38
π_i^{Aopt}, $\check{\boldsymbol{\beta}}_{uw}$	4.10	41.83	417.55	4369.04
π_i^{Aopt}, $\check{\boldsymbol{\beta}}_p$	2.88	27.93	273.24	2719.51
Full	139.46	1411.78	14829.63	138134.69

9. Summary

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

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

In addition, we have derive the unconditional asymptotic distribution for the proposed estimator based on Poisson subsampling. Interestingly, if the subsampling ratio 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 ratio does not converge to zero, the unconditional asymptotic distribution has a larger variance-covariance matrix. Our results also include the local 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.

Furthermore, we have proved consistency and asymptotic normality for the proposed estimators under two types of misspecifications: one is that pilot estimators are inconsistent, and the other is that the logistic regression model is misspecified.

Acknowledgments

The author would like to thank the reviewers for their insightful comments and suggestions, which greatly helped improve the paper and strengthen the proposed methods. This work was partially supported by the NSF grant DMS-1812013 and a UConn REP grant.

Appendix A. Subsampling with replacement from hard drive

If the full data can be loaded into available RAM, 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 $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ can still be loaded in available RAM, one can calculate $\{\pi_i^{\mathrm{OS}}(\hat{\boldsymbol{\beta}}_0)\}_{i=1}^N$ 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 Algorithm 3 to scan the data line-by-line or block-by-block in order to obtain the subsample.

Algorithm 3 Obtain the subsample with the 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 i-th data point into the subsample
    while $i == idx_{(j)}$ **do**
      $j \leftarrow j + 1$
    end while
  end if
  $i \leftarrow i + 1$
end while

Clearly, Algorithm 3 takes no more than linear time to run. Here, 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 is 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.

Appendix B. 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 B.1; technical details related to Poisson subsampling in Section 4 are presented in Section B.2; technical details related to unconditional results in Section 6 are presented in Section B.3; and technical details related to model misspecification in Section 7 are presented in Section B.4.1.

B.1. Proofs for subsampling with replacement

In this section we prove the results in Section 3. For ease of presentation, we use notation λ to denote the log-likelihood shifted by $\hat{\boldsymbol{\beta}}_0$. For the subsample, $\lambda^*_{uw}(\boldsymbol{\beta}) = \ell^*_{uw}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)$. Denote the first and second derivatives of $\lambda^*_{uw}(\boldsymbol{\beta})$ as $\dot{\lambda}^*_{uw}(\boldsymbol{\beta}) = \partial\lambda^*_{uw}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\ddot{\lambda}^*_{uw}(\boldsymbol{\beta}) = \partial^2\lambda^*_{uw}(\boldsymbol{\beta})/(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}})$.

Note that from Xiong and Li (2008); Cheng and Huang (2010), the fact that a sequence converges to 0 in conditional probability is equivalent to the fact that it converges to 0 in unconditional probability. This can also be proved directly by using the fact the probability measure is bounded by 1. Thus, in the following, we will use $o_P(1)$ to denote a sequence converging to 0 in probability without stating whether the underlying probability measure is conditional or unconditional.

We first present some lemmas that will be used to prove Theorem 1, and provide their proofs in Sections B.1.2 - B.1.5.

Lemma 28 *Let $\mathbf{v}_1, ..., \mathbf{v}_N$ be i.i.d. random vectors with the same distribution of $\mathbf{v}$. Let g_{1N} be a bounded function and g_2 be a fixed function that does not depend on N. If $g_{1N}(\mathbf{v}) = o_P(1)$ and $\mathbb{E}|g_2(\mathbf{v})| < \infty$, then*

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

Lemma 29 *Let $\boldsymbol{\eta}_i = |\psi_i(\hat{\boldsymbol{\beta}}_0)|\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i$, where $\psi_i(\boldsymbol{\beta}) = y_i - p(\mathbf{x}_i, \boldsymbol{\beta})$. Under Assumptions 1 and 2, conditional on the consistent $\hat{\boldsymbol{\beta}}_0$, if $n_0/\sqrt{N} \to 0$, then*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}} - \boldsymbol{\beta}_t) = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}}{2\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\}}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\eta}_i + o_P(1),$$

which converges in distribution to a normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1}\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}\}[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1}$, as n_0 and N go to infinity.

Lemma 30 *Let*

$$\dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t) = \sum_{i=1}^{n}\{y^*_i - p^*_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^*_i.$$

Under Assumptions 1 and 2, conditional on $\mathcal{D}_N$ and the consistent $\hat{\boldsymbol{\beta}}_0$, as n_0, n, and N go to infinity,

$$\frac{\dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t)}{\sqrt{n}} - \frac{\sqrt{n}\sum_{i=1}^{N}\boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)} \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}_t}\big),$$

in distribution.

Lemma 31 *Under Assumptions 1-3, as n_0, n, and N go to infinity, for any $\mathbf{s}_n \to 0$ in probability,*

$$\frac{1}{n}\sum_{i=1}^{n}\phi^*_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \mathbf{s}_n)\|\mathbf{x}^*_i\|^2 - \sum_{i=1}^{N}\pi_i(\hat{\boldsymbol{\beta}}_0)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\|\mathbf{x}_i\|^2 = o_P(1).$$

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

$$\lambda^*_{uw}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^{\mathrm{T}} \mathbf{x}_i^* y_i^* - \log \left\{ 1 + e^{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^{\mathrm{T}} \mathbf{x}_i^*} \right\} \Big],$$

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

$$\gamma(\mathbf{s}) = \frac{1}{\sqrt{n}} \mathbf{s}^{\mathrm{T}} \dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t) + \frac{1}{2n} \sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n}) (\mathbf{s}^{\mathrm{T}} \mathbf{x}_i^*)^2,$$

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 31,

$$\frac{1}{n} \sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n}) \mathbf{x}_i^* (\mathbf{x}_i^*)^{\mathrm{T}} - \sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_0) \phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} = o_P(1).$$

From Lemma 28 and the law of large numbers,

$$\sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_0) \phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} = \frac{\frac{1}{N} \sum_{i=1}^{N} |\psi_i(\hat{\boldsymbol{\beta}}_0)| h(\mathbf{x}_i) \phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0) \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\hat{\boldsymbol{\beta}}_0)} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} + o_P(1).$$

Combining the above two equations, we have that $n^{-1} \sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n}) \mathbf{x}_i^* (\mathbf{x}_i^*)^{\mathrm{T}}$ converges in probability to $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1}$, a positive definite matrix. In addition, from Lemma 29 and Lemma 30, $\dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t)/\sqrt{n}$ is stochastically bounded. Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), the maximizer of $\gamma(\mathbf{s})$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_t)$, satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_t) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \frac{1}{\sqrt{n}} \dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t) + o_P(1)$$

given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_0$. Thus,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \left\{ \frac{1}{\sqrt{n}} \dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t) - \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}} - \boldsymbol{\beta}_t) \right\} + o_P(1). \tag{26}$$

From Lemma 29,

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}} - \boldsymbol{\beta}_t) = \frac{\sqrt{n} \sum_{i=1}^{N} \boldsymbol{\eta}_i}{2N \mathbb{E}\{\phi(\boldsymbol{\beta}_t) h(\mathbf{x})\}} = \frac{\sqrt{n} \sum_{i=1}^{N} \boldsymbol{\eta}_i}{N \Psi_N(\hat{\boldsymbol{\beta}}_0)} + o_P(1), \tag{27}$$

Combining equations (26) and (27), Lemma 30, Slutsky's theorem, and the fact that a conditional probability is bounded by one, Theorem 1 follows. ■

B.1.1. Proof of Proposition 4

Proof of Proposition 4. To prove that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} \leq \mathbf{V}^{\mathrm{OS}} = \mathbf{M}^{-1}\mathbf{V}_c^{\mathrm{OS}}\mathbf{M}^{-1}$, we just need to show that

$$\boldsymbol{\Sigma}^{-1}_{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}} \geq \mathbf{M}(\mathbf{V}_c^{\mathrm{OS}})^{-1}\mathbf{M}.$$

From the strong law of large numbers,

$$\begin{aligned}
\mathbf{M} &= \frac{1}{N}\sum_{i=1}^{N}\phi_i(\boldsymbol{\beta}_t)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + o(1),\\
\mathbf{V}_c^{\mathrm{OS}} &= 4\Phi(\boldsymbol{\beta}_t)\frac{1}{N}\sum_{i=1}^{N}\frac{\phi_i(\boldsymbol{\beta}_t)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{h(\mathbf{x}_i)} + o(1),\\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} &= 4\Phi(\boldsymbol{\beta}_t)\left\{\frac{1}{N}\sum_{i=1}^{N}\phi_i(\boldsymbol{\beta}_t)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right\}^{-1} + o(1),
\end{aligned}$$

almost surely. Thus, we only need to verify that

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

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

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

which is true if

$$\mathbf{H} \geq \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}} \tag{28}$$

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,

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

which implies (28). If $h(\mathbf{x}) = 1$, the equality can be verified directly.

The first inequality in (14) can be verified directly using the result in (13). For the second inequality in (14), noting that $h(\mathbf{x}) = \|L\mathbf{M}^{-1}\mathbf{x}\|$, by the CauchySchwarz inequality, we have

$$\begin{aligned}
\mathrm{tr}(L\mathbf{V}^{\mathrm{OS}}L^{\mathrm{T}}) &= \mathrm{tr}(L\mathbf{M}^{-1}\mathbf{V}_c^{\mathrm{OS}}\mathbf{M}^{-1}L^{\mathrm{T}})\\
&= 4\Phi(\boldsymbol{\beta}_t)\mathrm{tr}\big[L\mathbf{M}^{-1}\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h^{-1}(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big\}\mathbf{M}^{-1}L^{\mathrm{T}}\big]\\
&= 4\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\}\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h^{-1}(\mathbf{x})\|L\mathbf{M}^{-1}\mathbf{x}\|^2\big\}\\
&= 4\big[\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)\|L\mathbf{M}^{-1}\mathbf{x}\|\big\}\big]^2\\
&\leq 4\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)\|L\mathbf{M}^{-1}\mathbf{x}\|^2\big\}\\
&= 4\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}\mathrm{tr}\big[L\mathbf{M}^{-1}\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)\mathbf{x}\mathbf{x}^{\mathrm{T}}\big\}\mathbf{M}^{-1}L^{\mathrm{T}}\big]\\
&= 4\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\}\mathrm{tr}(L\mathbf{M}^{-1}L^{\mathrm{T}})\\
&< \mathrm{tr}(L\mathbf{M}^{-1}L^{\mathrm{T}}),
\end{aligned}$$

which finishes the proof. ■

B.1.2. Proof of Lemma 28

Proof of Lemma 28. Let B be a bound for g_{1N} i.e., $|g_{1N}| \leq B$. For any $\epsilon > 0$, by Markov's inequality,

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

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 $g_{1N}(\mathbf{v}_i) \leq B$ and $g_{1N}(\mathbf{v}_i) = o_P(1)$ imply that $\mathbb{E}|g_{1N}(\mathbf{v})| = o(1)$. Thus, there is a N_ζ such that $\mathbb{E}|g_{1N}(\mathbf{v})| < \zeta\epsilon/(2K)$ when $N > N_\zeta$. Therefore, for any $\zeta > 0$, $\mathbb{P}\{|N^{-1}\sum_{i=1}^{N} g_{1N}(\mathbf{v}_i)g_2(\mathbf{v}_i)| > \epsilon\} < \zeta$ for sufficiently large N. This finishes the proof. ■

B.1.3. Proof of Lemma 29

Proof of Lemma 29. Since $\hat{\boldsymbol{\beta}}_{\text{wMLE}}$ is the maximizer of

$$\lambda_{\text{wMLE}}(\boldsymbol{\beta}) = \sum_{i=1}^{N} |y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\big[y_i\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0) - \log\{1 + e^{\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)}\}\big],$$

$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_t)$ is the maximizer of $\gamma_{\text{wMLE}}(\mathbf{s}) = \lambda_{\text{wMLE}}(\boldsymbol{\beta}_t + \mathbf{s}/\sqrt{N}) - \lambda_{\text{wMLE}}(\boldsymbol{\beta}_t)$. By Taylor's expansion,

$$\gamma_{\text{wMLE}}(\mathbf{s}) = \frac{1}{\sqrt{N}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_{\text{wMLE}}(\boldsymbol{\beta}_t) - \frac{1}{2N}\sum_{i=1}^{N} |y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{N})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2$$

where

$$\dot{\lambda}_{\text{wMLE}}(\boldsymbol{\beta}_t) = \sum_{i=1}^{N} \boldsymbol{\eta}_i = \sum_{i=1}^{N} |\psi_i(\hat{\boldsymbol{\beta}}_0)|\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i, \tag{29}$$

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

Since that $n_0/\sqrt{N} \to 0$ and $\|\boldsymbol{\eta}_i\|$ is bounded by $\|h(\mathbf{x}_i)\mathbf{x}_i\|$, we can ignore the data points that are used in obtaining the pilot $\hat{\boldsymbol{\beta}}_0$. The following discussion focus on $\boldsymbol{\eta}_i$'s for which $(\mathbf{x}_i, y_i)$'s are not included in the pilot subsample. Note that for such $\boldsymbol{\eta}_i$'s, $\mathbb{E}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0) = \mathbf{0}$

because

$$\begin{aligned}\mathbb{E}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0,\mathbf{x}_i) =&\mathbb{E}\Big[|\psi_i(\hat{\boldsymbol{\beta}}_0)|\big\{y_i-p(\mathbf{x}_i,\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)\big\}h(\mathbf{x}_i)\mathbf{x}_i\Big|\hat{\boldsymbol{\beta}}_0,\mathbf{x}_i\Big]\\ =&-p(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)p(\mathbf{x}_i,\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)\{1-p(\mathbf{x}_i,\boldsymbol{\beta}_t)\}h(\mathbf{x}_i)\mathbf{x}_i\\ &+\{1-p(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)\}\{1-p(\mathbf{x}_i,\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)\}p(\mathbf{x}_i,\boldsymbol{\beta}_t)h(\mathbf{x}_i)\mathbf{x}_i\\ =&-\frac{e^{\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0}}\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_t-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_t-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0}}\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_t}}h(\mathbf{x}_i)\mathbf{x}_i\\ &+\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0}}\frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_t-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0}}\frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_t}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_t}}h(\mathbf{x}_i)\mathbf{x}_i=0. \end{aligned}\tag{30}$$

This also gives that

$$\mathbb{V}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0)=\mathbb{E}\big\{\mathbb{V}(\boldsymbol{\eta}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)\big|\hat{\boldsymbol{\beta}}_0\big\}+\mathbb{V}\big\{\mathbb{E}(\boldsymbol{\eta}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)\big|\hat{\boldsymbol{\beta}}_0\big\}=\mathbb{E}\big\{\mathbb{V}(\boldsymbol{\eta}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)\big|\hat{\boldsymbol{\beta}}_0\big\}.$$

Now, since

$$\begin{aligned}\mathbb{V}(\boldsymbol{\eta}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0) =&\mathbb{E}\Big[|y_i-p(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)|^2\big\{y_i-p(\mathbf{x}_i,\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)\big\}^2h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\Big|\hat{\boldsymbol{\beta}}_0,\mathbf{x}_i\Big]\\ =&p(\mathbf{x}_i,\boldsymbol{\beta}_t)\{1-p(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)\}^2\big\{1-p(\mathbf{x}_i,\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)\big\}^2h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\ &+\{1-p(\mathbf{x}_i,\boldsymbol{\beta}_t)\}\{p(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)\}^2\big\{p(\mathbf{x}_i,\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)\big\}^2h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\ =&\phi_i(\hat{\boldsymbol{\beta}}_0)\phi_i(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_t)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}},\end{aligned}$$

we have

$$\mathbb{V}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0)=\mathbb{E}\Big\{\phi(\hat{\boldsymbol{\beta}}_0)\phi(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_t)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\Big|\hat{\boldsymbol{\beta}}_0\Big\}.$$

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 $\mathbb{V}(\boldsymbol{\eta}_i|\boldsymbol{\beta}_t)=\mathbb{E}\{\phi(\boldsymbol{\beta}_t)\phi(\boldsymbol{\beta}_t-\boldsymbol{\beta}_t)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\}=0.25\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\}$. Notice that $\big|\phi(\hat{\boldsymbol{\beta}}_0)\phi(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_t)-0.25\phi_i(\boldsymbol{\beta}_t)\big|h^2(\mathbf{x}_i)\|\mathbf{x}_i\|^2$ converges to 0 in probability and it is bounded by $h^2(\mathbf{x}_i)\|\mathbf{x}_i\|^2$, an integrable random variable under Assumption 2. Thus,

$$\mathbb{E}\big\|\mathbb{V}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0)-\mathbb{V}(\boldsymbol{\eta}_i|\boldsymbol{\beta}_t)\big\|\le\mathbb{E}\big\{\big|\{\phi(\hat{\boldsymbol{\beta}}_0)\phi(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_t)-0.25\phi_i(\boldsymbol{\beta}_t)\big|h^2(\mathbf{x}_i)\|\mathbf{x}_i\|^2\big\}=o(1).$$

This implies that

$$\mathbb{V}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0)=\mathbb{V}(\boldsymbol{\eta}_i|\boldsymbol{\beta}_t)+o_P(1)=0.25\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i\big\}+o_P(1).$$

For $\boldsymbol{\eta}_i$'s that $(\mathbf{x}_i,y_i)$'s are not included in the pilot subsample, conditional on $\hat{\boldsymbol{\beta}}_0$, $\boldsymbol{\eta}_i$'s are i.i.d. with mean $\mathbf{0}$ and variance $\mathbb{V}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0)$. Since for any $\epsilon>0$,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\Big\{\|\boldsymbol{\eta}_i\|^2I(\|\boldsymbol{\eta}_i\|>\sqrt{N}\epsilon)\Big|\hat{\boldsymbol{\beta}}_0\Big\}\le&\frac{1}{N}\sum_{i=1}^N\mathbb{E}\Big\{\|h(\mathbf{x}_i)\mathbf{x}_i\|^2I(\|h(\mathbf{x}_i)\mathbf{x}_i\|>\sqrt{N}\epsilon)\Big|\hat{\boldsymbol{\beta}}_0\Big\}\\ =&\mathbb{E}\{\|h(\mathbf{x})\mathbf{x}\|^2I(\|h(\mathbf{x})\mathbf{x}\|>\sqrt{N}\epsilon)\}\to0,\end{aligned}$$

the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998) applies conditional on $\hat{\boldsymbol{\beta}}_0$. Thus, we have, conditional on $\hat{\boldsymbol{\beta}}_0$,

$$\frac{\dot{\lambda}_{\text{wMLE}}(\boldsymbol{\beta}_t)}{\sqrt{N}} \longrightarrow \mathbb{N}\bigg[\mathbf{0},\ \frac{\mathbb{E}\big\{\phi(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}\big\}}{4}\bigg],$$

in distribution. From Lemma 28, conditional on $\hat{\boldsymbol{\beta}}_0$,

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}|y_i - p(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{N})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\ &= \frac{1}{4}\mathbb{E}\{|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\} + o_{P|\hat{\boldsymbol{\beta}}_0}(1) = \frac{1}{2}\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\} + o_P(1).\end{aligned}$$

Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), the maximizer of $\gamma_{\text{wMLE}}(\mathbf{s})$, $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_t)$, satisfies

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_t) = 2[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1}\frac{1}{\sqrt{N}}\dot{\lambda}_{\text{wMLE}}(\boldsymbol{\beta}_t) + o_P(1). \tag{31}$$

Note that

$$[\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}]^{-1} = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}}{4\Phi(\boldsymbol{\beta}_t)}. \tag{32}$$

Combining equations (29), (31), and (32), we have

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

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

■

B.1.4. Proof of Lemma 30

Proof of Lemma 30. Note that given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_0$, $\{y_i^* - p_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i^*$ 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 $\hat{\boldsymbol{\beta}}_0$. For the expectation, we have,

$$\mathbb{E}\big[\{y^* - p(\mathbf{x}_i^*,\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^*\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0\big] = \sum_{i=1}^{N}\pi_i(\hat{\boldsymbol{\beta}}_0)\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i = \frac{\sum_{i=1}^{N}\boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)},$$

where $\Psi_N(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^{N}|y_i - p(\mathbf{x}_i,\boldsymbol{\beta})|h(\mathbf{x}_i)$. From Lemma 29 and its proof, $\sum_{i=1}^{N}\boldsymbol{\eta}_i = O_P(\sqrt{N})$ conditional on $\hat{\boldsymbol{\beta}}_0$ in probability, i.e., for any $\epsilon > 0$, there exits a K such that $\mathbb{P}\{\mathbb{P}(\sum_{i=1}^{N}\boldsymbol{\eta}_i/\sqrt{N} > K\big|\hat{\boldsymbol{\beta}}_0) < \epsilon\} \to 1$ as $n_0, N \to \infty$. From Xiong and Li (2008), we know that $\sum_{i=1}^{N}\boldsymbol{\eta}_i = O_P(\sqrt{N})$ unconditionally. Thus, for the expectation, we have

$$\Delta = \mathbb{E}\big[\{y^* - p(\mathbf{x}_i^*,\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^*\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0\big] = O_P(1/\sqrt{N}).$$

For the variance,

$$\begin{aligned}
&\mathbb{V}\big[\{y^* - p(\mathbf{x}_i^*, \boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^* | \mathcal{D}_N, \hat{\boldsymbol{\beta}}_0\big] \\
=&\sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_0)\{y_i - p(\mathbf{x}_i, \boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}^2 \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} - \Delta^2 \\
=&\frac{\frac{1}{N}\sum_{i=1}^{N} |\psi_i(\hat{\boldsymbol{\beta}}_0)| \psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0) h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\hat{\boldsymbol{\beta}}_0)} - O_P(1/N) \\
=&\frac{\frac{1}{N}\sum_{i=1}^{N} |\psi_i(\boldsymbol{\beta}_t)| (y_i - 0.5)^2 h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\boldsymbol{\beta}_t)} + o_P(1) \\
=&\frac{1}{4}\frac{\mathbb{E}\{|\psi(\boldsymbol{\beta}_t)| h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{\Psi(\boldsymbol{\beta}_t)} + o_P(1) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} + o_P(1)
\end{aligned}$$

where the third equality is from Lemma 28 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 conditional distribution. Denote $\dot{\lambda}_{ri}^* = \{y_i^* - p_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i^*$.

$$\begin{aligned}
&\mathbb{E}\frac{1}{n}\sum_{i=1}^{n} \big\{\|\dot{\lambda}_{ri}^*\|^2 I(\|\dot{\lambda}_{ri}^*\| > \sqrt{n}\epsilon) \big| \mathcal{D}_N, \hat{\boldsymbol{\beta}}_0\big\} \\
\leq&\mathbb{E}\big\{\|\mathbf{x}^*\|^2 I(\|\mathbf{x}\| > \sqrt{n}\epsilon) \big| \mathcal{D}_N, \hat{\boldsymbol{\beta}}_0\big\} \\
=&\sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_0)\big\{\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\| > \sqrt{n}\epsilon)\big\} \\
\leq&\frac{\frac{1}{N}\sum_{i=1}^{N}\big\{h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\| > \sqrt{n}\epsilon)\big\}}{\Psi_N(\hat{\boldsymbol{\beta}}_0)} \\
\leq&\frac{\frac{1}{N}\sum_{i=1}^{N}\big\{h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\| > \sqrt{n}\epsilon)\big\}}{\Psi_N(\hat{\boldsymbol{\beta}}_0)} = o_P(1),
\end{aligned}$$

by Lemma 28 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. ■

B.1.5. Proof of Lemma 31

Proof of Lemma 31. We begin with the following partition,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \mathbf{s}_n)\|\mathbf{x}_i^*\|^2 \\
=&\frac{1}{n}\sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \mathbf{s}_n)\|\mathbf{x}_i^*\|^2 I(\|\mathbf{x}_i^*\|^2 \leq n) + \frac{1}{n}\sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \mathbf{s}_n)\|\mathbf{x}_i^*\|^2 I(\|\mathbf{x}_i^*\|^2 > n) \\
\equiv&\Delta_1 + \Delta_2.
\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}(\Delta_2|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0) &= \sum_{i=1}^{N}\pi_i(\hat{\boldsymbol{\beta}}_0)\phi_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0+\mathbf{s}_n)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\|^2>n)\\ &\le \frac{\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\|^2>n)}{\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}\\ &\le \frac{\frac{1}{N}\sum_{i=1}^{N}h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\|\mathbf{x}_i\|^2>n)}{\Psi_N(\hat{\boldsymbol{\beta}}_0)} = o_P(1)\end{aligned}$$

as $n, N \to \infty$, where the last step is from Lemma 28.

Similarly, we can show that

$$\mathbb{E}(\Delta_1|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0) - \sum_{i=1}^{N}\pi_i(\hat{\boldsymbol{\beta}}_0)\phi_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)\|\mathbf{x}_i\|^2 = o_P(1).$$

Thus, we only need to show that $\Delta_1 - \mathbb{E}(\Delta_1|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0) = o_P(1)$. For this, we show that the conditional variance of Δ_1 goes to 0 in probability. Notice that

$$\begin{aligned}&\mathbb{V}(\Delta_1|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0)\\ =&\frac{1}{n}\mathbb{V}\big\{\phi^*(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)\|\mathbf{x}^*\|^2 I(\|\mathbf{x}^*\|^2\le n)\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0\big\}\\ \le&\frac{1}{16n}\mathbb{E}\big\{\|\mathbf{x}^*\|^4 I(\|\mathbf{x}^*\|^2\le n)\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0\big\}\\ =&\frac{1}{16n}\sum_{i=1}^{n}\mathbb{E}\big\{\|\mathbf{x}^*\|^4 I(i-1<\|\mathbf{x}^*\|^2\le i)\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0\big\}\\ \le&\frac{1}{16n}\sum_{i=1}^{n}i^2\mathbb{E}\big\{I(i-1<\|\mathbf{x}^*\|^2\le i)\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0\big\}\\ \le&\frac{1}{16n}\sum_{i=1}^{n}i^2\big\{\mathbb{P}(\|\mathbf{x}^*\|^2>i-1\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0)-\mathbb{P}(\|\mathbf{x}^*\|^2>i\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0)\big\}\\ =&\frac{1}{16n}\left\{\mathbb{P}(\|\mathbf{x}^*\|^2>0\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0)-n^2\mathbb{P}(\|\mathbf{x}^*\|^2>n\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0)+\sum_{i=1}^{n-1}(2i+1)\mathbb{P}(\|\mathbf{x}^*\|^2>i\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0)\right\}\\ \le&\frac{1}{16n}\left\{1+\sum_{i=1}^{n-1}3i\mathbb{P}(\|\mathbf{x}^*\|^2>i\big|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0)\right\}\end{aligned}$$

This is $o_P(1)$ because

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} i\mathbb{P}(\|\mathbf{x}^*\|^2 > i|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0) = \frac{1}{n}\sum_{i=1}^{n} i\sum_{j=1}^{n}\pi_j(\hat{\boldsymbol{\beta}}_0)I(\|\mathbf{x}_j\|^2 > i)\\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} i\pi_j(\hat{\boldsymbol{\beta}}_0)I(\|\mathbf{x}_j\|^2 > i)\\
&= \frac{\frac{1}{N}\sum_{j=1}^{n}\frac{1}{n}\sum_{i=1}^{n} i|\psi_j(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_j)I(\|\mathbf{x}_j\|^2 > i)}{\Psi_N(\hat{\boldsymbol{\beta}}_0)}\\
&\le \frac{\frac{1}{N}\sum_{j=1}^{n}\frac{1}{n}\sum_{i=1}^{n} ih(\mathbf{x}_j)I(\|\mathbf{x}_j\|^2 > i)}{\Psi_N(\hat{\boldsymbol{\beta}}_0)},
\end{aligned}$$

and the numerator is non-negative and has an expectation

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

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

B.2. Proofs for Poisson subsampling

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

Define $\delta_i^{\hat{\boldsymbol{\beta}}_0} = I\{u_i \le n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\}$, and use notation λ_p to denote the log-likelihood shifted by $\hat{\boldsymbol{\beta}}_0$, i.e., $\lambda_p(\boldsymbol{\beta}) = \ell_p^*(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)$. Using these notations, the estimator $\hat{\boldsymbol{\beta}}_p$ is the maximizer of

$$\lambda_p(\boldsymbol{\beta}) = \sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\big[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\mathbf{x}_i y_i - \log\{1 + e^{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\mathbf{x}_i}\}\big], \tag{33}$$

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 30 and 31 are derived below which will be used to prove Theorem 6. We will prove these two lemmas in Sections B.2.2 and B.2.3.

Lemma 32 *Let*

$$\dot{\lambda}_p(\boldsymbol{\beta}_t) = \sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\{y_i - p(\mathbf{x}_i, \boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i.$$

Under Assumptions 1 and 2, conditional on $\mathcal{D}_N$, the consistent estimator $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$, if $n = o(N)$, then

$$\frac{\dot{\lambda}_p(\boldsymbol{\beta}_t)}{\sqrt{n}} - \frac{\sqrt{n}\sum_{i=1}^{N}\boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)} \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}\big),$$

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

$$\frac{\dot{\lambda}_p(\boldsymbol{\beta}_t)}{\sqrt{n}} - \frac{\sqrt{n}\sum_{i=1}^N \boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)} \longrightarrow \mathbb{N}\big(\mathbf{0},\ \boldsymbol{\Lambda}_\rho\big),$$

in distribution.

Lemma 33 *Under Assumptions 1 and 2, as* n_0*,* n*, and* N *go to infinity, for any* $\mathbf{s}_n \to 0$ *in probability,*

$$\frac{1}{n}\sum_{i=1}^N \delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \mathbf{s}_n)\|\mathbf{x}_i\|^2 - \sum_{i=1}^N \pi_i^p(\hat{\boldsymbol{\beta}}_0)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\|\mathbf{x}_i\|^2 = o_P(1).$$

Proof of Theorem 6. The estimator $\hat{\boldsymbol{\beta}}_p$ is the maximizer of (33), so $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$ is the maximizer of $\gamma_p(\mathbf{s}) = \lambda_p(\boldsymbol{\beta}_t + \mathbf{s}/\sqrt{n}) - \lambda_p(\boldsymbol{\beta}_t)$. By Taylor's expansion,

$$\gamma_p(\mathbf{s}) = \frac{1}{\sqrt{n}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) + \frac{1}{2n}\sum_{i=1}^N \delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2$$

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

From Lemmas 32 and 33, conditional on $\mathcal{D}_N$, and $\hat{\boldsymbol{\beta}}_0$,

$$\frac{1}{n}\sum_{i=1}^N \delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} + o_P(1).$$

In addition, from Lemma 32, conditional on $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$, $\dot{\lambda}_p(\boldsymbol{\beta}_t)/\sqrt{n}$ converges in distribution to a normal limit. Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), the maximizer of $\gamma_p(\mathbf{s})$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$, satisfies

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

given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$. Combining this with Lemma 32, Slutsky's theorem, and the fact that a conditional probability is bounded, Theorem 6 follows. ■

B.2.1. Proof of Proposition 8

Proof of Proposition 8. 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

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

■

B.2.2. Proof of Lemma 32

Proof of Lemma 32. Note that, $\delta_i^{\hat{\boldsymbol{\beta}}_0} = I\{u_i \le n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\}$, where u_i are i.i.d. with the standard uniform distribution. Thus, given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$, $\dot{\lambda}_p(\boldsymbol{\beta}_t)$ is a sum of N independent random vectors. We now exam the mean and variance of $\dot{\lambda}_p(\boldsymbol{\beta}_t)$. Recall that $\boldsymbol{\eta}_i = |\psi_i(\hat{\boldsymbol{\beta}}_0)|\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i$, and $\psi_i(\boldsymbol{\beta}) = y_i - p(\mathbf{x}_i, \boldsymbol{\beta})$. For the mean, we have,

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbb{E}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big\} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{N}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \wedge 1\}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{N} n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i = \frac{\sqrt{n}}{\sqrt{N}}\frac{\sum_{i=1}^{N}\boldsymbol{\eta}_i}{\hat{\Psi}_0\sqrt{N}} = O_P(\sqrt{n/N}),
\end{aligned}
$$

where the last equality is from Lemma 29.

For the variance,

$$
\begin{aligned}
&\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big\} \\
=&\frac{1}{n}\sum_{i=1}^{N}[\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \wedge 1\} - \{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \wedge 1\}^2]\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}^2\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \\
=&\sum_{i=1}^{N}\pi_i^p(\hat{\boldsymbol{\beta}}_0)\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - n\sum_{i=1}^{N}\{\pi_i^p(\hat{\boldsymbol{\beta}}_0)\}^2\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \\
=&\frac{\frac{1}{N}\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_0} \\
&- \frac{n}{N}\frac{\frac{1}{N}\sum_{i=1}^{N}\psi_i^2(\hat{\boldsymbol{\beta}}_0)\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_0^2} \\
\equiv&\Delta_3 - \Delta_4
\end{aligned}
\tag{34}
$$

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 28, if $n/N \to \rho$,

$$
\Delta_4 \to \rho\frac{\mathbb{E}\{\psi^2(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{4\Psi^2(\boldsymbol{\beta}_t)}, \tag{35}
$$

in probability.

For the term Δ_3 in (34), it is equal to

$$\begin{aligned}\Delta_3 &= \frac{1}{\hat{\Psi}_0^2}\frac{1}{N}\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}{N}\vee\hat{\Psi}_0\Big\}\psi_i^2(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\ &= \frac{1}{\hat{\Psi}_0^2}\frac{n}{N^2}\sum_{i=1}^{N}\psi_i^2(\hat{\boldsymbol{\beta}}_0)\psi_i^2(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}I\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}{N}>\hat{\Psi}_0\Big\}\\ &\quad+\frac{1}{\hat{\Psi}_0}\frac{1}{N}\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|\psi_i^2(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}I\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}{N}\le\hat{\Psi}_0\Big\}.\end{aligned}$$

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 28, if $n/N\to\rho$, as n_0, n, and N go to infinity,

$$\begin{aligned}\Delta_3 \to&\frac{\rho\mathbb{E}\big[\psi^2(\boldsymbol{\beta}_t)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}I\big\{\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\ge\Psi(\boldsymbol{\beta}_t)\big\}\big]}{4\Psi^2(\boldsymbol{\beta}_t)}\\ &+\frac{\mathbb{E}\big[|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}I\big\{\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\le\Psi(\boldsymbol{\beta}_t)\big\}\big\}\big]}{4\Psi(\boldsymbol{\beta}_t)}\\ =&\frac{\mathbb{E}\Big(|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big[\{\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\}\vee\Psi(\boldsymbol{\beta}_t)\}\big]\Big)}{4\Psi^2(\boldsymbol{\beta}_t)},\end{aligned}\tag{36}$$

in probability. From, (34), (35), and (36), if $n/N\to\rho$,

$$\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0,\hat{\Psi}_0\big\}=\frac{\mathbb{E}\big[|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\{\Psi(\boldsymbol{\beta}_t)-\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x})\}_+\big]}{4\Psi^2(\boldsymbol{\beta}_t)}+o_P(1).$$

Specifically, when $\rho=0$,

$$\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0,\hat{\Psi}_0\big\}=\boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}+o_P(1).$$

Now we check the Lindeberg-Feller condition (Section *2.8 of van der Vaart, 1998) under the condition distribution. Denote $\dot{\lambda}_{pi} = \delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i$. For any $\epsilon > 0$

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\Big\{\|\dot{\lambda}_{pi}\|^2 I(\|\dot{\lambda}_{pi}\| > \sqrt{n}\epsilon)\Big|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big\} \\
&\leq \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\Big[\|\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\mathbf{x}_i\|^2 I(\|\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\mathbf{x}_i\| > \sqrt{n}\epsilon)\Big|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big] \\
&= \sum_{i=1}^{N}\pi_i^p(\hat{\boldsymbol{\beta}}_0)\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\|\mathbf{x}_i\|^2 I(\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\|\mathbf{x}_i\| > \sqrt{n}\epsilon) \\
&\leq \frac{|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\{n/N|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i) + \hat{\Psi}_0\}\|\mathbf{x}_i\|^2 I(\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) + 1\}\|\mathbf{x}_i\| > \sqrt{n}\epsilon)}{\hat{\Psi}_0^2} \\
&\leq \frac{\frac{1}{N}\sum_{i=1}^{N}h^2(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\{h(\mathbf{x}_i)/\hat{\Psi}_0 + 1\}\|\mathbf{x}_i\| > \sqrt{n}\epsilon)}{\hat{\Psi}_0^2} \\
&+ \frac{\frac{1}{N}\sum_{i=1}^{N}h(\mathbf{x}_i)\|\mathbf{x}_i\|^2 I(\{h(\mathbf{x}_i)/\hat{\Psi}_0 + 1\}\|\mathbf{x}_i\| > \sqrt{n}\epsilon)}{\hat{\Psi}_0} = o_P(1),
\end{aligned}
$$

where the last equality is from Lemma 28. Thus, applying the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998) finishes the proof. ■

B.2.3. Proof of Lemma 33

Proof of Lemma 33. Note that, from Lemma 28,

$$
\sum_{i=1}^{N}\pi_i^p(\hat{\boldsymbol{\beta}}_0)\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \frac{1}{\hat{\Psi}_0 N}\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t} + o_P(1);
$$

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

$$
\Delta_5 \equiv \frac{1}{n}\sum_{i=1}^{N}\Big|\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \mathbf{s}_n) - \delta_i^{\boldsymbol{\beta}_t}\{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t)\Big|\|\mathbf{x}_i\|^2 = o_P(1),
$$

then the result in Lemma 33 follows. Noting that Δ_5 is nonnegative, we prove $\Delta_5 == o_P(1)$ by showing that $\mathbb{E}(\Delta_5|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0) = o_P(1)$. Note that given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$, the only

random terms in Δ_5 are $\delta_i^{\hat{\boldsymbol{\beta}}_0} = I\{u_i \le n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\}$ and $\delta_i^{\boldsymbol{\beta}_t} = I\{u_i \le n\pi_i^p(\boldsymbol{\beta}_t)\}$. We have that

$$\begin{aligned}
&\mathbb{E}(\Delta_5|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0) \\
\le&\frac{1}{n}\sum_{i=1}^N \{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \wedge n\pi_i^p(\boldsymbol{\beta}_t) \wedge 1\} \\
&\qquad \times \left|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \mathbf{s}_n) - \{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t)\right| \|\mathbf{x}_i\|^2 \\
&+ \frac{1}{n}\sum_{i=1}^N |n\pi_i^p(\hat{\boldsymbol{\beta}}_0) - n\pi_i^p(\boldsymbol{\beta}_t)|\left|n\pi_i^p(\hat{\boldsymbol{\beta}}_0) + n\pi_i^p(\boldsymbol{\beta}_t) + 2\right| \|\mathbf{x}_i\|^2 \\
\equiv&\Delta_6 + \Delta_7.
\end{aligned}$$

Note that $n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \wedge n\pi_i^p(\boldsymbol{\beta}_t) \wedge 1 \le n\pi_i^p(\hat{\boldsymbol{\beta}}_0)$. Thus Δ_6 is bounded by

$$\frac{1}{\hat{\Psi}_0}\frac{1}{N}\sum_{i=1}^N \left|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \mathbf{s}_n) - \{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t)\right| h(\mathbf{x}_i)\|\mathbf{x}_i\|^2,$$

which is $o_P(1)$ by Lemma 28 if $|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\} - \{n\pi_i^p(\boldsymbol{\beta}_t)\vee 1\}| = o_P(1)$. This is true because

$$\begin{aligned}
|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\} - \{n\pi_i^p(\boldsymbol{\beta}_t) \vee 1\}| \le& n|\pi_i^p(\hat{\boldsymbol{\beta}}_0) - \pi_i^p(\boldsymbol{\beta}_t)| \\
\le& \frac{nh(\mathbf{x}_i)}{N}\left|\frac{|\psi_i(\hat{\boldsymbol{\beta}}_0)|}{\hat{\Psi}_0} - \frac{|\psi_i(\boldsymbol{\beta}_t)|}{\Psi_N(\boldsymbol{\beta}_t)}\right| = o_P(1).
\end{aligned}$$

The term Δ_7 is bounded by

$$\frac{1}{N}\sum_{i=1}^N \left|\frac{|\psi_i(\hat{\boldsymbol{\beta}}_0)|}{\hat{\Psi}_0} - \frac{|\psi_i(\boldsymbol{\beta}_t)|}{\Psi_N(\boldsymbol{\beta}_t)}\right|\left|\frac{|\psi_i(\hat{\boldsymbol{\beta}}_0)|}{\hat{\Psi}_0} + \frac{|\psi_i(\boldsymbol{\beta}_t)|}{\Psi_N(\boldsymbol{\beta}_t)} + \frac{2}{h(\mathbf{x}_i)}\right| h^2(\mathbf{x}_i)\|\mathbf{x}_i\|^2 = o_P(1),$$

where the last equality is from Lemma 28 and the fact that $\mathbb{E}\{h^2(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$. ■

B.3. Proofs for unconditional distribution

In this section we prove Theorem 13 in Section 6. A lemma similar to Lemma 32 is presented below and will be proved later in this section. Lemma 33 can be used in the proof of Theorem 13 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 subsample taken according to the subsampling probabilities π_{0i} in (17), we define $\delta_i^{(1)} = I\{u_{0i} \le \frac{c_0(1-y_i)+c_1 y_i}{N}\}$, where u_{0i} are i.i.d. standard uniform random variables. With this notation, the estimator $\hat{\Psi}_0$ defined in (18) can be written as

$$\hat{\Psi}_0 = \frac{1}{N}\sum_{i=1}^N \frac{\delta_i^{(1)}|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}{n\pi_{0i} \wedge 1}. \tag{37}$$

Lemma 34 *Let $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ be constructed according to Step 1 of Algorithm 2, respectively. For*

$$\dot{\lambda}_p(\boldsymbol{\beta}_t) = \sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\beta}}_0} \{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\{y_i - p(\mathbf{x}_i, \boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i,$$

under the same assumptions of Theorem 13, if $n = o(N)$, then

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

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

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

in distribution.

Proof of Theorem 13. The proof of this theorem is similar to that of Theorem 6. The key difference is that Lemma 34 is about asymptotic distribution unconditionally.

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

$$\lambda_p(\boldsymbol{\beta}) = \sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\beta}}_0} \{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\big[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\mathbf{x}_i y_i - \log\{1 + e^{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\mathbf{x}_i}\}\big],$$

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

$$\gamma_p(\mathbf{s}) = \frac{1}{\sqrt{n}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) + \frac{1}{2n}\sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\beta}}_0} \{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2$$

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

From Lemma 33,

$$\frac{1}{n}\sum_{i=1}^{N} \delta_i^{\hat{\boldsymbol{\beta}}_0} \{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i(\mathbf{x}_i)^{\mathrm{T}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_t}^{-1} + o_P(1).$$

In addition, from Lemma 34, $\dot{\lambda}_p(\boldsymbol{\beta}_t)/\sqrt{n}$ converges in distribution to a normal limit. Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), the maximizer of $\gamma_p(\mathbf{s})$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$, satisfies

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

Combining this with Lemma 34 and Slutsky's theorem, Theorem 13 follows. ■

B.3.1. Proof of Proposition 17

Proof of Proposition 17 To prove (22), we just need to show that $\mathbf{\Lambda}_u \geq \mathbf{\Sigma}_{\boldsymbol{\beta}_t}^{-1} > \mathbf{\Lambda}_\rho$. From Proposition 8, we know that $\mathbf{\Sigma}_{\boldsymbol{\beta}_t}^{-1} > \mathbf{\Lambda}_\rho$. To show that $\mathbf{\Lambda}_u \geq \mathbf{\Sigma}_{\boldsymbol{\beta}_t}^{-1}$, we notice that

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

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., $\mathbb{P}\{\rho\phi(\boldsymbol{\beta}_t)h(\mathbf{x}) > \Phi(\boldsymbol{\beta}_t)\} > 0$. ■

B.3.2. Proof of Lemma 34

Proof of Lemma 34.

We first proof the case when the pilot estimates $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ depend on the data. For any $l \in \mathbb{R}^d$, denote $\tau_{Ni} = \sqrt{N/n}\hat{\Psi}_0\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i^{\mathrm{T}}l$, $i = 1, ..., N$, where $\delta_i^{\hat{\boldsymbol{\beta}}_0} = I\{u_i \leq n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\}$, 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. Again, since $n_0 = o(\sqrt{N})$, we can focus on τ_{Ni}'s that $(\mathbf{x}_i, y_i)$'s are not included in the pilot subsample. We now exam the mean and variance of these τ_{Ni}'s. For the mean, based on calculation similar to that in (30), we have,

$$\mathbb{E}\big(\tau_{Ni}\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big) = \sqrt{nN}\hat{\Psi}_0\mathbb{E}\big\{\pi_i^p(\hat{\boldsymbol{\beta}}_0)\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i^{\mathrm{T}}l\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big\} = \frac{\sqrt{n}\mathbb{E}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0)}{\sqrt{N}} = 0,$$

which implies that

$$\mathbb{E}\tau_{Ni} = 0.$$

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

$$\begin{aligned}\mathbb{E}\big(\tau_{Ni}^2\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big) =& N\hat{\Psi}_0^2\mathbb{E}\Big[\pi_i^p(\hat{\boldsymbol{\beta}}_0)\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)(\mathbf{x}_i^{\mathrm{T}}l)^2\Big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big] \\ =& \mathbb{E}\Big[|\psi_i(\hat{\boldsymbol{\beta}}_0)|\Big\{\frac{n}{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i) \vee \hat{\Psi}_0\Big\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2\Big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big].\end{aligned}$$

If we let

$$\Upsilon_{Ni} = |\psi_i(\hat{\boldsymbol{\beta}}_0)|\Big\{\frac{n}{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i) \vee \hat{\Psi}_0\Big\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2,$$

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

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

in probability. We now show that

$$\mathbb{E}(\Upsilon_{Ni}) \to \mathbb{E}(\Upsilon_i) = 0.25\mathbb{E}[|\psi(\boldsymbol{\beta}_t)|\{\rho|\psi(\boldsymbol{\beta}_t)|h(\mathbf{x}) \vee \Psi(\boldsymbol{\beta}_t)\}h(\mathbf{x})(\mathbf{x}^{\mathrm{T}}l)^2].$$

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

$$\begin{aligned}|\mathbb{E}(\Upsilon_{Ni}) - \mathbb{E}(\Upsilon_i)| &\le \mathbb{E}\{\Xi_i I(\Xi_i > \epsilon)\} + \mathbb{E}\{\Xi_i I(\Xi_i \le \epsilon)\} \\ &\le \mathbb{E}\big[\{h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}} l)^2 + \Upsilon_i + \hat{\Psi}_0 h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}} l)^2\} I(\Xi_i > \epsilon)\big] + \epsilon\end{aligned}$$

We know that $\mathbb{E}\big[\{h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}} l)^2 + \Upsilon_i\} I(\Xi_i > \epsilon)\big] \to 0$ 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 $I(\Xi_i > \epsilon)$ is bounded and is $o_P(1)$. Similarly, $\mathbb{E}\{\hat{\Psi}_0 h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}} l)^2 I(\Xi_i > \epsilon)\} \le \mathbb{E}\{h(\mathbf{x})\}\mathbb{E}\{h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}} l)^2 I(\Xi_i > \epsilon)\} \to 0$. Thus, $\mathbb{E}(\Upsilon_{Ni}) - \mathbb{E}(\Upsilon_i) \to 0$, and we have finished proving that

$$\mathbb{V}(\tau_{Ni}) \to \mathbb{E}(\Upsilon_i). \tag{38}$$

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

$$\mathbb{E}|\tau_{Ni}|^3 = o(\sqrt{N}). \tag{39}$$

For the conditional expectation,

$$\begin{aligned}&\mathbb{E}\big(|\tau_{Ni}|^3 \big| \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big) \\ &= N\sqrt{N/n}\hat{\Psi}_0^3 \mathbb{E}\Big[\pi_i^p(\hat{\boldsymbol{\beta}}_0)\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}^2 \psi_i^3(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)(\mathbf{x}_i^{\mathrm{T}} l)^3 \Big| \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big] \\ &= \sqrt{N/n}\mathbb{E}\Big[|\psi_i(\hat{\boldsymbol{\beta}}_0)| \Big\{\frac{n}{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)| h(\mathbf{x}_i) \vee \hat{\Psi}_0\Big\}^2 \psi_i^3(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0) h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}} l)^3 \Big| \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big] \\ &\le 2\|l\|^3 \sqrt{N/n}\mathbb{E}\big[\{h^2(\mathbf{x}_i) + \hat{\Psi}_0^2\} h(\mathbf{x}_i)\|\mathbf{x}_i\|^3 \big| \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big] \\ &\le 2\|l\|^3 \sqrt{N/n}\big[\mathbb{E}\{h^3(\mathbf{x}_i)\|\mathbf{x}_i\|^3\} + \mathbb{E}\{\hat{\Psi}_0^2 | \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\}\mathbb{E}\{h(\mathbf{x}_i)\|\mathbf{x}_i\|^3\}\big].\end{aligned}$$

Since $\mathbb{E}\{h^3(\mathbf{x}_i)\|\mathbf{x}_i\|^3\} < \infty$ and $\mathbb{E}\{h(\mathbf{x}_i)\|\mathbf{x}_i\|^3\} < \infty$, (39) follows if $\mathbb{E}\{\hat{\Psi}_0^2\} = O(1)$. This is true because

$$\begin{aligned}\mathbb{E}\{\hat{\Psi}_0^2\} &= \mathbb{E}\bigg\{\frac{1}{N}\sum_{k_1=1}^N \frac{\delta_{k_1}^{(1)}|y_{k_1} - p(\mathbf{x}_{k_1}, \hat{\boldsymbol{\beta}}_0)| h(\mathbf{x}_{k_1})}{n_0/N} \frac{1}{N}\sum_{k_2=1}^N \frac{\delta_{k_2}^{(1)}|y_{k_2} - p(\mathbf{x}_{k_2}, \hat{\boldsymbol{\beta}}_0)| h(\mathbf{x}_{k_2})}{n_0/N}\bigg\} \\ &\le \frac{1}{N^2}\sum_{k_1 \ne k_2}^N \mathbb{E}\bigg\{\frac{\delta_{k_1}^{(1)} h(\mathbf{x}_{k_1})}{n_0/N}\frac{\delta_{k_2}^{(1)} h(\mathbf{x}_{k_2})}{n_0/N}\bigg\} + \frac{1}{N^2}\sum_{k=1}^N \mathbb{E}\bigg\{\frac{\delta_k^{(1)} h^2(\mathbf{x}_k)}{\{n_0/N\}^2}\bigg\} \\ &= \frac{1}{N^2}\sum_{k_1 \ne k_2}^N \mathbb{E}\{h(\mathbf{x}_{k_1})h(\mathbf{x}_{k_2})\} + \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}\bigg\{\frac{1}{n_0/N} h^2(\mathbf{x}_k)\bigg\} \to \mathbb{E}\{h^2(\mathbf{x})\}.\end{aligned}$$

Denote $\nu_{Ni} = \tau_{Ni}\{\mathbb{V}(\tau_{Ni})\}^{-1/2}$. We know that ν_{Ni}'s, for which $(\mathbf{x}_i, y_i)$'s are not included in the pilot subsample, are i.i.d. conditional on $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$. Thus, from Theorem 7.3.2 of Chow and Teicher (2003), they are interchangeable. The fact that $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ are consistent estimators implies that they are a sequence of two estimators, and for each $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$, τ_{Ni} are interchangeable and can be . 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 $i \ne j$,

$$\mathbb{E}(\nu_{Ni}\nu_{Nj}) = \mathbb{E}\{\mathbb{E}(\nu_{Ni}\nu_{Nj} | \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0)\} = 0, \tag{40}$$

and

$$\frac{1}{\sqrt{N}}\mathbb{E}\{|\nu_{Ni}|^3\} = \mathbb{E}|\tau_{Ni}|^3\{\mathbb{V}(\tau_{Ni})\}^{-3/2} \to 0 \tag{41}$$

which follows from (39). We now show that for $i \neq j$,

$$\mathbb{E}\{\nu_{Ni}^2\nu_{Nj}^2\} \to 1. \tag{42}$$

Since $\nu_{Ni} = \tau_{Ni}\{\mathbb{V}(\tau_{Ni})\}^{-1/2}$, from (38), to prove (42), 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 $\mathbb{E}\big(\tau_{Ni}^2\tau_{Nj}^2\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big) = \mathbb{E}\big(\tau_{Ni}^2\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big)\mathbb{E}\big(\tau_{Ni}^2\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big) = \mathbb{E}\big(\Upsilon_{Ni}\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big)\mathbb{E}\big(\Upsilon_{Nj}\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big) = \mathbb{E}\big(\Upsilon_{Ni}\Upsilon_{Nj}\big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big)$, 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 $\Xi_{ij} = |\Upsilon_{Ni}\Upsilon_{Nj} - \Upsilon_i\Upsilon_j|$. For any $\epsilon > 0$,

$$\begin{aligned}
&|\mathbb{E}(\Upsilon_{Ni}\Upsilon_{nj}) - \mathbb{E}(\Upsilon_i\Upsilon_j)| \\
\leq&\mathbb{E}\{\Xi_{ij}I(\Xi_{ij} > \epsilon)\} + \mathbb{E}\{\Xi_{ij}I(\Xi_{ij} \leq \epsilon)\} \\
\leq&\mathbb{E}\big[\{h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2h^2(\mathbf{x}_j)(\mathbf{x}_j^{\mathrm{T}}l)^2 + \Upsilon_i\Upsilon_j\}I(\Xi_{ij} > \epsilon)\big] \\
&+ \mathbb{E}\big[\hat{\Psi}_0^2h(\mathbf{x}_i)\|\mathbf{x}_i\|^2h(\mathbf{x}_j)\|\mathbf{x}_j\|^2I(\Xi_{ij} > \epsilon)\big] \\
&+ \mathbb{E}(\hat{\Psi}_0)\mathbb{E}\big[\{h(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2h^2(\mathbf{x}_j)(\mathbf{x}_j^{\mathrm{T}}l)^2 + h(\mathbf{x}_j)(\mathbf{x}_j^{\mathrm{T}}l)^2h^2(\mathbf{x}_i)(\mathbf{x}_i^{\mathrm{T}}l)^2\}I(\Xi_{ij} > \epsilon)\big] + \epsilon
\end{aligned}$$

Since $i \neq j$, $\mathbb{E}\{h^2(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$, $\mathbb{E}\{h(\mathbf{x})\|\mathbf{x}\|^2\} < \infty$, $\mathbb{E}(\hat{\Psi}_0) < \infty$, and $I(\Xi_{ij} > \epsilon) = o_P(1)$ is bounded, the above results indicates that (42) holds.

Since (40), (41), (42) are satisfied, the central limit theorem in Theorem 2 of Blum et al. (1958) holds for ν_{Ni}, which gives that

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

in distribution. Note that

$$\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\nu_i &= \frac{\hat{\Psi}_0}{\sqrt{n}\{\mathbb{V}(\tau_{Ni})\}^{1/2}}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i^{\mathrm{T}}l \\
&= \frac{\hat{\Psi}_0}{\sqrt{n}\{\mathbb{V}(\tau_{Ni})\}^{1/2}}l^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) = \frac{\Psi}{\sqrt{n}\{\mathbb{V}(\tau_{Ni})\}^{1/2}}l^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) + o_P(1).
\end{aligned}$$

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

$$\frac{1}{\sqrt{n}}l^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) \to \mathbb{N}(0, l^{\mathrm{T}}\boldsymbol{\Lambda}_u l) \tag{43}$$

in distribution, where

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

and the equality holds if $\rho = 0$, i.e., $n/N \to 0$. Based on (43), from the Cramér-Wold theorem, we have that

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

in distribution.

When the pilot estimates $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$ are independent of the data, if we can prove the results in Lemma 34 under the conditional distribution given $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$, 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 32 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 $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$, $\dot{\lambda}_p(\boldsymbol{\beta}_t)$ is a sum of N independent random vectors. We now exam the mean and variance of $\dot{\lambda}_p(\boldsymbol{\beta}_t)$ given $\hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$. For the mean,

$$\frac{1}{\sqrt{n}}\mathbb{E}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big\} = \mathbf{0}.$$

For the variance,

$$\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big\} = \frac{\mathbb{E}\Big[|\psi_i(\hat{\boldsymbol{\beta}}_0)|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\Big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big]}{\hat{\Psi}_0},$$

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 $\epsilon > 0$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\Big\{\big\|\dot{\lambda}_{pi}\big\|^2 I(\big\|\dot{\lambda}_{pi}\big\| > \sqrt{n}\epsilon)\Big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big\} \\
&\leq \frac{\mathbb{E}\Big[\{h(\mathbf{x})\|\mathbf{x}\|^2 + \hat{\Psi}_0\}h(\mathbf{x})I(\{h(\mathbf{x})/\hat{\Psi}_0 + 1\}\|\mathbf{x}\| > \sqrt{n}\epsilon)\Big|\hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\Big]}{\hat{\Psi}_0^2} = o_P(1).
\end{aligned}$$

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

■

B.4. Proofs for cases of misspecifications

B.4.1. Proofs with pilot misspecification

Proof of Theorem 18. By similar arguments used in the proof of Theorem 1, we know that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_t)$ is the maximizer of

$$\gamma(\mathbf{s}) = \frac{1}{\sqrt{n}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_{uw}^*(\boldsymbol{\beta}_t) + \frac{1}{2n}\sum_{i=1}^{n}\phi_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i^*)^2,$$

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

$$\dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t) = \sum_{i=1}^{n} \{y_i^* - p_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i^*.$$

Given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_0$, $\{y_i^* - p_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i^*$ are i.i.d. random vectors. We exam their mean and variance, and check the Lindeberg-Feller condition under the conditional distribution. For the expectation, direct calculations give

$$\mathbb{E}\big[\{y^* - p(\mathbf{x}_i^*, \boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^* \big| \mathcal{D}_N, \hat{\boldsymbol{\beta}}_0\big] = \sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_0)\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i = \frac{\sum_{i=1}^N \boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)}, \tag{44}$$

where $\boldsymbol{\eta}_i = |\psi_i(\hat{\boldsymbol{\beta}}_0)|\psi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i$. Conditional on $\hat{\boldsymbol{\beta}}_0$, $\boldsymbol{\eta}_i$'s are i.i.d., and we still have $\mathbb{E}(\boldsymbol{\eta}_i|\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0) = \mathbf{0}$ and thus $\mathbb{E}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0) = \mathbf{0}$ due to (30). Thus

$$\begin{aligned}\mathbb{V}(\boldsymbol{\eta}_i|\hat{\boldsymbol{\beta}}_0) = \mathbb{E}(\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{\mathrm{T}}|\hat{\boldsymbol{\beta}}_0) &= \mathbb{E}\big\{\psi^2(\hat{\boldsymbol{\beta}}_0)\psi^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big|\hat{\boldsymbol{\beta}}_0\big\} \\ &= \mathbb{E}\big\{\psi^2(\boldsymbol{\beta}_0)\psi^2(\boldsymbol{\beta}_t - \boldsymbol{\beta}_0)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\big\} + o_P(1) = \boldsymbol{\varsigma}_b + o_P(1),\end{aligned}$$

where the third equality is from Lemma 28 and the facts that $\psi^2(\cdot) \le 1$ and $\mathbb{E}\{h^2(\mathbf{x})\|\mathbf{x}|^2\} < \infty$.

Similar to the proof of Lemma 29, the Lindeberg-Feller central limit theorem applies conditional on $\hat{\boldsymbol{\beta}}_0$. Thus, we have that, conditional on $\hat{\boldsymbol{\beta}}_0$,

$$\frac{\sum_{i=1}^N \boldsymbol{\eta}_i}{\sqrt{N}} \longrightarrow \mathbb{N}(\mathbf{0},\ \boldsymbol{\varsigma}_b), \tag{45}$$

in distribution.

From (44) and (45), we have

$$\Delta = \mathbb{E}\big[\{y^* - p(\mathbf{x}_i^*, \boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^* \big| \mathcal{D}_N, \hat{\boldsymbol{\beta}}_0\big] = O_P(1/\sqrt{N}). \tag{46}$$

For the conditional variance of $\{y_i^* - p_i^*(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i^*$, using similar approach to the proof of Lemma 30, we have

$$\begin{aligned}&\mathbb{V}\big[\{y^* - p(\mathbf{x}_i^*, \boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^* | \mathcal{D}_N, \hat{\boldsymbol{\beta}}_0\big] \\ =&\sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_0)\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \Delta^2 \\ =&\frac{\frac{1}{N}\sum_{i=1}^N |\psi_i(\hat{\boldsymbol{\beta}}_0)|\psi_i^2(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\hat{\boldsymbol{\beta}}_0)} - O_P(1/N) \\ =&\frac{\mathbb{E}\{|\psi(\boldsymbol{\beta}_0)|\psi^2(\boldsymbol{\beta}_t - \boldsymbol{\beta}_0)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{\Psi(\boldsymbol{\beta}_0)} + o_P(1) = \frac{\boldsymbol{\varsigma}_a}{\Psi(\boldsymbol{\beta}_0)} + o_P(1)\end{aligned} \tag{47}$$

The Lindeberg-Feller condition under the conditional distribution can be verified similarly to the proof of Lemma 30. Thus, we have

$$\frac{\dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t)}{\sqrt{n}} - \frac{\sqrt{n}\sum_{i=1}^N \boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)} \longrightarrow \mathbb{N}\Big\{\mathbf{0},\ \frac{\boldsymbol{\varsigma}_a}{\Psi(\boldsymbol{\beta}_0)}\Big\}, \tag{48}$$

in conditional distribution.

From Lemmas 28 and 31, and the law of large numbers, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0+\acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i^*(\mathbf{x}_i^*)^{\mathrm{T}}\\
&=\sum_{i=1}^{N}\pi_i(\hat{\boldsymbol{\beta}}_0)\phi_i(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+o_P(1)\\
&=\frac{\frac{1}{N}\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\hat{\boldsymbol{\beta}}_0)}+o_P(1)\\
&=\frac{\mathbb{E}\{|\psi(\boldsymbol{\beta}_0)|\phi(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{\Psi(\boldsymbol{\beta}_0)}+o_P(1)=\frac{\boldsymbol{\varsigma}_a}{\Psi(\boldsymbol{\beta}_0)}+o_P(1).
\end{aligned}$$

Since $\boldsymbol{\varsigma}_a$ is a positive definite matrix, and combining (44), (45), (46) and (47) we know that $\dot{\lambda}_{uw}^*(\boldsymbol{\beta}_t)/\sqrt{n}$ is stochastically bounded, from the Basic Corollary in page 2 of Hjort and Pollard (2011), we have that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw}-\boldsymbol{\beta}_t)=\Psi(\boldsymbol{\beta}_0)\boldsymbol{\varsigma}_a^{-1}\frac{1}{\sqrt{n}}\dot{\lambda}_{uw}^*(\boldsymbol{\beta}_t)+o_P(1) \tag{49}$$

given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_0$.

Now note that $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}-\boldsymbol{\beta}_t)$ is the maximizer of

$$\frac{1}{\sqrt{N}}\mathbf{s}^{\mathrm{T}}\sum_{i=1}^{N}\boldsymbol{\eta}_i-\frac{1}{2N}\sum_{i=1}^{N}|y_i-p(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0+\acute{\mathbf{s}}/\sqrt{N})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2$$

with $\acute{\mathbf{s}}$ between $\mathbf{0}$ and $\mathbf{s}$. From Lemma 28, conditional on $\hat{\boldsymbol{\beta}}_0$,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}|y_i-p(\mathbf{x}_i,\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0+\acute{\mathbf{s}}/\sqrt{N})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\
&=\frac{1}{N}\sum_{i=1}^{N}|y_i-p(\mathbf{x}_i,\boldsymbol{\beta}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+o_P(1)\\
&=\mathbb{E}\{|\psi(\boldsymbol{\beta}_0)|\phi_i(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\}+o_P(1)=\boldsymbol{\varsigma}_a+o_P(1)
\end{aligned}$$

Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), we have

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}}-\boldsymbol{\beta}_t)=\boldsymbol{\varsigma}_a^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\eta}_i+o_P(1). \tag{50}$$

Combining this with (49), we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw}-\hat{\boldsymbol{\beta}}_{\mathrm{wMLE}})=\Psi(\boldsymbol{\beta}_0)\boldsymbol{\varsigma}_a^{-1}\Big\{\frac{1}{\sqrt{n}}\dot{\lambda}_{uw}^*(\boldsymbol{\beta}_t)-\frac{\sqrt{n}}{N\Psi(\boldsymbol{\beta}_0)}\sum_{i=1}^{N}\boldsymbol{\eta}_i\Big\}+o_P(1).$$

Combining the above two equations with (48), Slutsky's theorem, and the fact that a conditional probability is bounded, Theorem 18 follows. ■

Proof of Remark 20 We first observe the following equations by direct calculations.

$$\Psi(\boldsymbol{\beta}_0) = \mathbb{E}\big[\{p(\mathbf{x},\boldsymbol{\beta}_t) + p(\mathbf{x},\boldsymbol{\beta}_0) - 2p(\mathbf{x},\boldsymbol{\beta}_0)p(\mathbf{x},\boldsymbol{\beta}_t)\}h(\mathbf{x})\big], \text{ and}$$
$$p(\mathbf{x},\boldsymbol{\beta}_t - \boldsymbol{\beta}_0) = \frac{p(\mathbf{x},\boldsymbol{\beta}_t)\{1-p(\mathbf{x},\boldsymbol{\beta}_0)\}}{p(\mathbf{x},\boldsymbol{\beta}_t)\{1-p(\mathbf{x},\boldsymbol{\beta}_0)\} + p(\mathbf{x},\boldsymbol{\beta}_0)\{1-p(\mathbf{x},\boldsymbol{\beta}_t)\}}. \tag{51}$$

We need to verify that

$$\frac{\mathbb{E}[\{1-p(\mathbf{x},\boldsymbol{\beta}_t)\}p(\mathbf{x},\boldsymbol{\beta}_0)p(\mathbf{x},\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}]}{\mathbb{E}\big[\{p(\mathbf{x},\boldsymbol{\beta}_t) + p(\mathbf{x},\boldsymbol{\beta}_0) - 2p(\mathbf{x},\boldsymbol{\beta}_0)p(\mathbf{x},\boldsymbol{\beta}_t)\}h(\mathbf{x})\big]} < \frac{\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{4\mathbb{E}\{\phi(\boldsymbol{\beta}_t)h(\mathbf{x})\}}$$

Note that from (51)

$$\begin{aligned} & 2\{1-p(\mathbf{x},\boldsymbol{\beta}_t)\}p(\mathbf{x},\boldsymbol{\beta}_0)p(\mathbf{x},\boldsymbol{\beta}_t-\boldsymbol{\beta}_0) < \phi(\boldsymbol{\beta}_t) \\ \Leftrightarrow\ & \{p(\mathbf{x},\boldsymbol{\beta}_t) - p(\mathbf{x},\boldsymbol{\beta}_0)\}\{1-2p(\mathbf{x},\boldsymbol{\beta}_0)\} > 0, \end{aligned}$$

and

$$\begin{aligned} & p(\mathbf{x},\boldsymbol{\beta}_t) + p(\mathbf{x},\boldsymbol{\beta}_0) - 2p(\mathbf{x},\boldsymbol{\beta}_0)p(\mathbf{x},\boldsymbol{\beta}_t) > 2\phi(\boldsymbol{\beta}_t) \\ \Leftrightarrow\ & \{p(\mathbf{x},\boldsymbol{\beta}_t) - p(\mathbf{x},\boldsymbol{\beta}_0)\}\{2p(\mathbf{x},\boldsymbol{\beta}_t) - 1\} > 0. \end{aligned}$$

Thus the inequality holds if

$$\begin{aligned} & p(\mathbf{x},\boldsymbol{\beta}_t) > 0.5 > p(\mathbf{x},\boldsymbol{\beta}_0) \quad \text{or} \quad p(\mathbf{x},\boldsymbol{\beta}_t) < 0.5 < p(\mathbf{x},\boldsymbol{\beta}_0) \\ \Leftrightarrow\ & \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t > 0 > \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \quad \text{or} \quad \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t < 0 < \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \\ \Leftrightarrow\ & \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_t\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 < 0. \end{aligned}$$

This finishes the proof. ■

Proof of Theorem 21. Similarly to the proof of Theorem 6, $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$ is the maximizer of

$$\frac{1}{\sqrt{n}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_t) + \frac{1}{2n}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\phi_i(\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2,$$

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

$$\dot{\lambda}_p(\boldsymbol{\beta}_t) = \sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\{y_i - p(\mathbf{x}_i,\boldsymbol{\beta}_t - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i.$$

Similarly to the proof of Lemma 32, we first notice that given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$, $\dot{\lambda}_p(\boldsymbol{\beta}_t)$ is a sum of N independent random vectors. We now exam the mean and variance of $\dot{\lambda}_p(\boldsymbol{\beta}_t)$. For the mean, we have,

$$\frac{1}{\sqrt{n}}\mathbb{E}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0,\hat{\Psi}_0\big\} = \frac{\sqrt{n}}{\sqrt{N}}\frac{\sum_{i=1}^{N}\boldsymbol{\eta}_i}{\hat{\Psi}_0\sqrt{N}} = O_P(\sqrt{n/N}),$$

where the last equality is due to (45).

For the variance,

$$\begin{aligned}\frac{1}{n}\mathbb{V}\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0,\hat{\Psi}_0\} =&\frac{\frac{1}{N}\sum_{i=1}^N|\psi_i(\hat{\boldsymbol{\beta}}_0)|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\psi_i^2(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_0}\\ &-\frac{n}{N}\frac{\frac{1}{N}\sum_{i=1}^N\psi_i^2(\hat{\boldsymbol{\beta}}_0)\psi_i^2(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_0^2}\\ \equiv&\Delta_8-\Delta_9\end{aligned}$$

From Lemma 28, if $n/N\to\rho$, using a similar approach used in the proof of Lemma 32, we have

$$\begin{aligned}\Delta_8 &= \frac{1}{\hat{\Psi}_0^2}\frac{1}{N}\sum_{i=1}^N|\psi_i(\hat{\boldsymbol{\beta}}_0)|\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}{N}\vee\hat{\Psi}_0\Big\}\psi_i^2(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\ &= \frac{1}{\Psi_0^2}\mathbb{E}\Big[|\psi(\boldsymbol{\beta}_0)|\big\{\rho|\psi(\boldsymbol{\beta}_0)|h(\mathbf{x})\vee\Psi(\boldsymbol{\beta}_0)\big\}\psi^2(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big]+o_P(1)\end{aligned}$$

and

$$\Delta_9=\rho\frac{\mathbb{E}\{\psi^2(\boldsymbol{\beta}_0)\psi^2(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)h^2(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\}}{\Psi_0^2}=\rho\frac{\boldsymbol{\varsigma}_b}{\Psi_0^2}+o_P(1).$$

Thus,

$$\begin{aligned}&\frac{1}{n}\mathbb{V}\{\dot{\lambda}_p(\boldsymbol{\beta}_t)|\mathcal{D}_N,\hat{\boldsymbol{\beta}}_0,\hat{\Psi}_0\}\\ &=\frac{\mathbb{E}\Big[|\psi(\boldsymbol{\beta}_0)|\big\{1-\Psi_0^{-1}\rho|\psi(\boldsymbol{\beta}_0)|h(\mathbf{x})\big\}_+\psi^2(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big]}{\Psi_0}+o_P(1)=\frac{\boldsymbol{\varsigma}_c}{\Psi_0}+o_P(1).\end{aligned}$$

Applying the Lindeberg-Feller central limit theorem, we have

$$\frac{\dot{\lambda}_p(\boldsymbol{\beta}_t)}{\sqrt{n}}-\frac{\sqrt{n}\sum_{i=1}^N\boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)}\longrightarrow\mathbb{N}\Big(\mathbf{0},\ \frac{\boldsymbol{\varsigma}_c}{\Psi_0^2}\Big),\tag{52}$$

in conditional distribution.

Using a similar approach used to prove Lemma 33, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^N\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\phi_i(\boldsymbol{\beta}_t-\hat{\boldsymbol{\beta}}_0+\mathbf{s}_n)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\\ &=\sum_{i=1}^N\pi_i^p(\hat{\boldsymbol{\beta}}_0)\phi_i(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+o_P(1)\\ &=\frac{\frac{1}{N}\sum_{i=1}^N|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_t-\boldsymbol{\beta}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_0}+o_P(1)=\frac{\boldsymbol{\varsigma}_a}{\Psi_0}+o_P(1).\end{aligned}$$

From (45) and (52), $\dot{\lambda}_p(\boldsymbol{\beta}_t)/\sqrt{n}$ is stochastically bounded. In addition, $\boldsymbol{\varsigma}_a$ is finite and positive-definite. Thus, from the Basic Corollary in page 2 of Hjort and Pollard (2011), $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t)$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_t) = \Psi_0 \boldsymbol{\varsigma}_a^{-1} \frac{1}{\sqrt{n}} \dot{\lambda}_p(\boldsymbol{\beta}_t) + o_P(1),$$

given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$. Combining this with (50), (52), Slutsky's theorem, and the fact that a conditional probability is bounded by one, Theorem 21 follows. ■

B.4.2. Proofs with model misspecification

Proof of Theorem 24. By similar arguments used in the proof of Theorem 1, we know that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_l)$ is the maximizer of

$$\frac{1}{\sqrt{n}} \mathbf{s}^{\mathrm{T}} \dot{\lambda}^*_{uw}(\boldsymbol{\beta}_l) + \frac{1}{2n} \sum_{i=1}^{n} \phi_i^*(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}} \mathbf{x}_i^*)^2,$$

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

$$\dot{\lambda}^*_{uw}(\boldsymbol{\beta}_l) = \sum_{i=1}^{n} \{y_i^* - p(\mathbf{x}_i^*, \boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i^*.$$

We abuse the notation and redefine $\boldsymbol{\eta}_i = |\psi_i(\hat{\boldsymbol{\beta}}_0)|\psi_i(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i$ in this proof. By similar arguments used in the proof of Lemma 30, we have that

$$\mathbb{E}\big[\{y^* - p(\mathbf{x}^*, \boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^* \big| \mathcal{D}_N, \hat{\boldsymbol{\beta}}_0\big] = \sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_0)\psi_i(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i = \frac{\sum_{i=1}^{N} \boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)},$$

and

$$\begin{aligned}
&\mathbb{V}\big[\{y^* - p(\mathbf{x}^*, \boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}^* \big| \mathcal{D}_N, \hat{\boldsymbol{\beta}}_0\big] \\
&= \sum_{i=1}^{N} \pi_i(\hat{\boldsymbol{\beta}}_0)\{y_i - p(\mathbf{x}_i, \boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\}^2 \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} - \Delta^2 \\
&= \frac{\frac{1}{N}\sum_{i=1}^{N} |\psi_i(\boldsymbol{\beta}_l)|(y_i - 0.5)^2 h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\Psi_N(\boldsymbol{\beta}_l)} + o_P(1) = \frac{\boldsymbol{\kappa}_a}{\omega} + o_P(1).
\end{aligned}$$

The Lindeberg-Feller condition under the conditional distribution can be verified using a similar approached used in the proof of Lemma 30. Thus, conditional on $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_0$, as n_0, n, and N go to infinity,

$$\frac{\dot{\lambda}^*_{uw}(\boldsymbol{\beta}_t)}{\sqrt{n}} - \frac{\sqrt{n}\sum_{i=1}^{N} \boldsymbol{\eta}_i}{N\Psi_N(\hat{\boldsymbol{\beta}}_0)} \longrightarrow \mathbb{N}\Big(\mathbf{0},\ \frac{\boldsymbol{\kappa}_a}{\omega}\Big), \tag{53}$$

in conditional distribution. Now we exam $\boldsymbol{\eta}_i$. For the j-th element of $\boldsymbol{\eta}_i$,

$$\begin{aligned}\eta_{ij} =&|\psi_i(\hat{\boldsymbol{\beta}}_0)|\{y_i - p_i(\mathbf{x}_i, \boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\}h(\mathbf{x}_i)x_{ij}\\ =&(2y_i-1)\{y_i - p_i(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)\}\{y_i - p_i(\mathbf{x}_i, \boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\}h(\mathbf{x}_i)x_{ij}\\ =&0.5(2y_i-1)^2\{y_i - p_i(\mathbf{x}_i, \boldsymbol{\beta}_l)\}h(\mathbf{x}_i)x_{ij} + \acute{\eta}_{ij}h(\mathbf{x}_i)x_{ij}\mathbf{x}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_l),\end{aligned} \tag{54}$$

where

$$\begin{aligned}\acute{\eta}_{ij} =&(2y_i-1)\Big[\{y_i - p_i(\mathbf{x}_i, \acute{\boldsymbol{\beta}})\}p_i(\mathbf{x}_i, \acute{\mathbf{s}})\{1 - p_i(\mathbf{x}_i, \acute{\mathbf{s}})\}\\ &- p_i(\mathbf{x}_i, \acute{\boldsymbol{\beta}})\{1 - p_i(\mathbf{x}_i, \acute{\boldsymbol{\beta}})\}\{y_i - p_i(\mathbf{x}_i, \acute{\mathbf{s}})\}\Big],\end{aligned}$$

with $\acute{\mathbf{s}} = \boldsymbol{\beta}_l - \acute{\boldsymbol{\beta}}$, and $\acute{\boldsymbol{\beta}}$ being between $\boldsymbol{\beta}_l$ and $\hat{\boldsymbol{\beta}}_0$. Note that $|\dot{\eta}_{ij}| \leq 2$ and

$$\acute{\eta}_{ij} \to \frac{(2y_i-1)}{4}\big[\{y_i - p_i(\mathbf{x}_i, \boldsymbol{\beta}_l)\} - 2(2y_i-1)p_i(\mathbf{x}_i, \boldsymbol{\beta}_l)\{1 - p_i(\mathbf{x}_i, \boldsymbol{\beta}_l)\}\big],$$

in probability. Thus, from Lemma 28 and direct calculations, we have that

$$\frac{1}{N}\sum_{i=1}^{N} h(\mathbf{x}_i)(\acute{\boldsymbol{\eta}}_i \circ \mathbf{x}_i)\mathbf{x}_i^{\mathrm{T}} = \boldsymbol{\kappa}_c + o_P(1), \tag{55}$$

where $\acute{\boldsymbol{\eta}}_i = (\acute{\eta}_{i1}, ..., \acute{\eta}_{id})^{\mathrm{T}}$ and $\circ$ is the Hadamard product. From (54) and (55), we have that

$$\sum_{i=1}^{N}\boldsymbol{\eta}_i = \frac{1}{2}\sum_{i=1}^{N}(2y_i-1)^2\{y_i - p_i(\mathbf{x}_i, \boldsymbol{\beta}_l)\}h(\mathbf{x}_i)\mathbf{x}_i + \boldsymbol{\kappa}_c N(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_l) + o_P\{N(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_l)\}.$$

Thus, from the central limit theorem and Slutsky's theorem, and the fact that $\hat{\boldsymbol{\beta}}_0$ is independent of $\mathcal{D}_N$,

$$\frac{\sum_{i=1}^{N}\boldsymbol{\eta}_i}{\sqrt{N}} \longrightarrow \mathbb{N}\Big(\mathbf{0},\ \boldsymbol{\kappa}_b + \frac{\boldsymbol{\kappa}_c\boldsymbol{\Sigma}_0\boldsymbol{\kappa}_c}{\rho_0}\Big). \tag{56}$$

From Lemma 28 and using a similar approach to prove Lemma 31, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i^*(\mathbf{x}_i^*)^{\mathrm{T}}\\ &= \sum_{i=1}^{N}\pi_i(\hat{\boldsymbol{\beta}}_0)\phi_i(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + o_P(1)\\ &= \frac{\frac{1}{N}\sum_{i=1}^{N}|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\frac{1}{N}\sum_{i=1}^{N}|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)} + o_P(1) = \frac{\boldsymbol{\kappa}_a}{\omega} + o_P(1).\end{aligned}$$

Since $\boldsymbol{\kappa}_a$ is a positive definite matrix, and $\dot{\lambda}_{uw}^*(\boldsymbol{\beta}_l)/\sqrt{n}$ is stochastically bounded due to (53) and (56), from the Basic Corollary in page 2 of Hjort and Pollard (2011), $\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_l)$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \boldsymbol{\beta}_l) = \omega\boldsymbol{\kappa}_a^{-1}\frac{1}{\sqrt{n}}\dot{\lambda}_{uw}^*(\boldsymbol{\beta}_l) + o_P(1) \tag{57}$$

given $\mathcal{D}_N$ and $\hat{\boldsymbol{\beta}}_0$.

Using similar arguments used in the proof of Lemma 29, we know that $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_l)$ is the maximizer of

$$\frac{1}{\sqrt{N}}\mathbf{s}^{\mathrm{T}}\sum_{i=1}^{N}\boldsymbol{\eta}_i - \frac{1}{2N}\sum_{i=1}^{N}|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{N})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2,$$

where $\acute{\mathbf{s}}$ lies between $\mathbf{0}$ and $\mathbf{s}$. From Lemma 28,

$$\frac{1}{N}\sum_{i=1}^{N}|y_i - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)\phi_i(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{N})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \boldsymbol{\kappa}_a + o_P(1).$$

Thus, from (56) and the Basic Corollary in page 2 of Hjort and Pollard (2011), we know that $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_l)$ satisfies

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{wMLE}} - \boldsymbol{\beta}_l) = \boldsymbol{\kappa}_a^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\eta}_i + o_P(1). \tag{58}$$

From (56), Slutsky's theorem, and the fact that a conditional probability is bounded by one, the result in (24) follows.

From (57) and (58), we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{uw} - \hat{\boldsymbol{\beta}}_{\text{wMLE}}) = \omega\boldsymbol{\kappa}_a^{-1}\left\{\frac{1}{\sqrt{n}}\dot{\lambda}^*_{uw}(\boldsymbol{\beta}_l) - \frac{\sqrt{n}\sum_{i=1}^{N}\boldsymbol{\eta}_i}{\omega N}\right\} + o_P(1).$$

Thus, from (53), Slutsky's theorem, and the fact that a conditional probability is bounded by one, (23) of Theorem 24 follows. ■

Proof of Theorem 26. Using similar arguments used in the proof Theorem 6, we know that $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_l)$ is the maximizer of

$$\frac{1}{\sqrt{n}}\mathbf{s}^{\mathrm{T}}\dot{\lambda}_p(\boldsymbol{\beta}_l) + \frac{1}{2n}\sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\phi_i(\boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0 + \acute{\mathbf{s}}/\sqrt{n})(\mathbf{s}^{\mathrm{T}}\mathbf{x}_i)^2,$$

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

$$\dot{\lambda}_p(\boldsymbol{\beta}_l) = \sum_{i=1}^{N}\delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0) \vee 1\}\{y_i - p(\mathbf{x}_i, \boldsymbol{\beta}_l - \hat{\boldsymbol{\beta}}_0)\}\mathbf{x}_i.$$

Given $\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0$ and $\hat{\Psi}_0$, $\dot{\lambda}_p(\boldsymbol{\beta}_l)$ is a sum of independent variables, and the Lindeberg-Feller condition under the condition distribution can be verified similarly to the proof of Lemma 32. Now we exam the conditional mean and variance of $\dot{\lambda}_p(\boldsymbol{\beta}_l)$. For the mean, from (56) we have,

$$\frac{1}{\sqrt{n}}\mathbb{E}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_l)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big\} = \frac{\sqrt{n}}{\sqrt{N}}\frac{\sum_{i=1}^{N}\boldsymbol{\eta}_i}{\hat{\Psi}_0\sqrt{N}} = O_P(\sqrt{n/N}).$$

For the variance,

$$\begin{aligned}
&\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_l)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big\} \\
=&\frac{\frac{1}{N}\sum_{i=1}^N |\psi_i(\hat{\boldsymbol{\beta}}_0)|\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\psi_i^2(\boldsymbol{\beta}_l-\hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_0} \\
&-\frac{n}{N}\frac{\frac{1}{N}\sum_{i=1}^N \psi_i^2(\hat{\boldsymbol{\beta}}_0)\psi_i^2(\boldsymbol{\beta}_l-\hat{\boldsymbol{\beta}}_0)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\hat{\Psi}_0^2} \equiv \Delta_{10}+\Delta_{11}. \qquad (59)
\end{aligned}$$

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 28, if $n/N\to\rho$,

$$\Delta_{11}\to\rho\frac{\boldsymbol{\kappa}_b}{\omega^2}, \qquad (60)$$

in probability. For the term Δ_{10} in (59), 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 28, if $n/N\to\rho$, as n_0, n, and N go to infinity, by a similar approach used in the proof of Lemma 32, we have

$$\begin{aligned}
\Delta_{10} =& \frac{1}{\hat{\Psi}_0^2}\frac{n}{N^2}\sum_{i=1}^N \psi_i^2(\hat{\boldsymbol{\beta}}_0)\psi_i^2(\boldsymbol{\beta}_l-\hat{\boldsymbol{\beta}}_0)h^2(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} I\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}{N}>\hat{\Psi}_0\Big\} \\
&+\frac{1}{\hat{\Psi}_0}\frac{1}{N}\sum_{i=1}^N |\psi_i(\hat{\boldsymbol{\beta}}_0)|\psi_i^2(\boldsymbol{\beta}_l-\hat{\boldsymbol{\beta}}_0)h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} I\Big\{\frac{n|\psi_i(\hat{\boldsymbol{\beta}}_0)|h(\mathbf{x}_i)}{N}\le\hat{\Psi}_0\Big\} \\
=& \frac{\mathbb{E}\Big(|\psi(\boldsymbol{\beta}_l)|[\{\rho|\psi(\boldsymbol{\beta}_l)|h(\mathbf{x})\}\vee\omega]h(\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}\Big)}{4\omega^2}+o_P(1). \qquad (61)
\end{aligned}$$

From, (59), (60), and (61), if $n/N\to\rho$,

$$\frac{1}{n}\mathbb{V}\big\{\dot{\lambda}_p(\boldsymbol{\beta}_l)|\mathcal{D}_N, \hat{\boldsymbol{\beta}}_0, \hat{\Psi}_0\big\}=\frac{\boldsymbol{\kappa}_d}{\omega}+o_P(1).$$

From the above results, conditional on $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$, we know that

$$\frac{\dot{\lambda}_p(\boldsymbol{\beta}_l)}{\sqrt{n}}-\frac{\sqrt{n}\sum_{i=1}^N\boldsymbol{\eta}_i}{N\hat{\Psi}_0}\longrightarrow\mathbb{N}\Big(\mathbf{0},\ \frac{\boldsymbol{\kappa}_d}{\omega^2}\Big), \qquad (62)$$

in distribution.

In addition, from Lemma 28, using an approach similar to the proof of Lemma 33, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^N \delta_i^{\hat{\boldsymbol{\beta}}_0}\{n\pi_i^p(\hat{\boldsymbol{\beta}}_0)\vee 1\}\phi_i(\boldsymbol{\beta}_l-\hat{\boldsymbol{\beta}}_0+\acute{\mathbf{s}}/\sqrt{n})\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \\
&=\frac{1}{4n}\sum_{i=1}^N \delta_i^{\boldsymbol{\beta}_l}\{n\pi_i^p(\boldsymbol{\beta}_l)\vee 1\}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+o_P(1) \\
&=\frac{1}{4N\hat{\Psi}_0}\sum_{i=1}^N |\psi_i(\boldsymbol{\beta}_l)|h(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+o_P(1)=\frac{\boldsymbol{\kappa}_a}{\omega}+o_P(1). \qquad (63)
\end{aligned}$$

Thus, based on (56), (62), and (63), from the Basic Corollary in page 2 of Hjort and Pollard (2011), $\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_l)$, satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}_l) = \omega \boldsymbol{\kappa}_a \frac{1}{\sqrt{n}} \dot{\lambda}_p(\boldsymbol{\beta}_l) + o_P(1),$$

given $\mathcal{D}_N$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\Psi}_0$. Combining this with (58), (62), Slutsky's theorem, and the fact that a conditional probability is bounded by one, Theorem 26 follows. ■

References

Mingyao Ai, Jun Yu, Huiming Zhang, and HaiYing Wang. Optimal subsampling algorithms for big data generalized linear models. *Statistica Sinica*, 2019. doi: 10.5705/ss.202018.0439.

Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum experimental designs, with SAS*, volume 34. Oxford University Press, 2007.

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

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

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017. doi: 10.1137/141000671.

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

Guang Cheng and Jianhua Huang. Bootstrap consistency for general semiparametric m-estimation. *The Annals of Statistics*, 38(5):2884–2915, 2010.

Yuan Shih Chow and Henry Teicher. *Probability Theory: Independence, Interchangeability, Martingales*. Springer, New York, 2003.

Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2013.

P. Drineas, M.W. Mahoney, S. Muthukrishnan, and T. Sarlos. Faster least squares approximation. *Numerische Mathematik*, 117:219–249, 2011.

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

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

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

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

Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for l_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006d.

Dheeru Dua and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

William Fithian and Trevor Hastie. Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics*, 42(5):1693, 2014.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Lei Han, Kean Ming Tan, Ting Yang, and Tong Zhang. Local uncertainty sampling for large-scale multi-class logistic regression. *Annals of Statistics*, 2019. URL `https://www.e-publications.org/ims/submission/AOS/user/submissionFile/37810?confirm=045881a9`.

Nils Lid Hjort and David Pollard. Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*, 2011.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Brian W. Kernighan and Dennis M. Ritchie. *The C Programming Language*. Prentice Hall Professional Technical Reference, 2nd edition, 1988. ISBN 0131103709.

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

Nan Lin and Ruibin Xie. Aggregated estimating equation estimation. *Statistics and Its Interface*, 4:73–83, 2011.

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

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

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

Peter McCullagh and James A Nelder. *Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability*. Chapman & Hall,, 1989.

Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, pages 415–423, 2014.

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

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

Garvesh Raskutti and Michael Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17:1–31, 2016.

Elizabeth D. Schifano, Jing Wu, Chun Wang, Jun Yan, and Ming-Hui Chen. Online updating of statistical inference in the big data setting. *Technometrics*, 58(3):393–403, 2016.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980.

Bjarne Stroustrup. *The C++ programming language*. Pearson Education India, 1986.

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

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

HaiYing Wang. Divide-and-conquer information-based optimal subdata selection algorithm. *Journal of Statistical Theory and Practice*, 2019. doi: 10.1007/s42519-019-0048-5.

HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525): 393–405, 2019.

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

Tianbao Yang, Lijun Zhang, Rong Jin, and Shenghuo Zhu. An explicit sampling dependent spectral error bound for column subset selection. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 135–143, 2015.

Yun Yang, Mert Pilanci, Martin J Wainwright, et al. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

Yaqiong Yao and HaiYing Wang. Optimal subsampling for softmax regression. *Statistical Papers*, 60(2):235–249, 2019.