Sampling with replacement vs Poisson sampling: a comparative study in optimal subsampling *

Jing Wang †, Jiahui Zou ‡, and HaiYing Wang*

May 26, 2022

Abstract

Faced with massive data, subsampling is a commonly used technique to improve computational efficiency, and using nonuniform subsampling probabilities is an effective approach to improve estimation efficiency. For computational efficiency, subsampling is often implemented with replacement or through Poisson subsampling. However, no rigorous investigation has been performed to study the difference between the two subsampling procedures such as their estimation efficiency and computational convenience. This paper performs a comparative study on these two different sampling procedures. In the context of maximizing a general target function, we first derive asymptotic distributions for estimators obtained from the two sampling procedures. The results show that the Poisson subsampling may have a higher estimation efficiency. Based on the asymptotic distributions for both subsampling with replacement and Poisson subsampling, we derive optimal subsampling probabilities that minimize the variance functions of the subsampling estimators. These subsampling probabilities further reveal the similarities and differences between subsampling with replacement and Poisson subsampling. The theoretical characterizations and comparisons on the two subsampling procedures provide guidance to select a more appropriate subsampling approach in practice. Furthermore, practically implementable algorithms are proposed based on the optimal structural results, which are evaluated through both theoretical and empirical analyses.

keywords: Algorithmic sampling; Asymptotic Distribution; Informative Sample; Massive Data.

*The first two authors contributed equally to this work.

†Department of Statistics, University of Connecticut, USA

‡School of Statistics, Capital University of Economics and Business, Beijing 100070, China

1 Introduction

With fast development of technology, data collecting is becoming easier and easier, and the volumes of available data sets are increasing exponentially. To extract useful information from these massive data, a major challenge lies with the thirst for computing resources. Subsampling is a commonly used technique to reduce computational burden, and it has been an important topic in computer science and statistics with a long standing of literature, such as Drineas et al. (2006*a*,*b*,*c*), Mahoney & Drineas (2009), Drineas et al. (2011), Mahoney (2011), Clarkson & Woodruff (2013), Kleiner et al. (2014), McWilliams et al. (2014), Yang et al. (2016).

To improve the estimation efficiency[1], nonuniform subsampling probabilities are often used so that more informative data points are sampled with higher probabilities. A popular choice is the leverage-based subsampling in which the subsampling distribution is the normalized statistical leverage scores of the design matrix (Drineas et al. 2012, Ma et al. 2015). Yang et al. (2015) showed that if statistical leverage scores are very nonuniform, then using their normalized square roots as the subsampling distribution yields better approximation. For logistic regression, Wang et al. (2018) derived an optimal subsampling distribution that minimizes the asymptotic variance of the subsampling estimator, and Wang (2019) further developed a more efficient estimation approach based on the selected subsample. Ting & Brochu (2018) investigated optimal subsampling with influence functions. Wang et al. (2019) proposed a method called information-based optimal subdata selection which selects data points deterministically for linear regression. The subsampling approach has a close connection to the technique of coreset approximation (Campbell & Broderick 2018, 2019), which also use a subset of the data with associated weights instead of the full data to reduce calculations. The coreset approximation is often used in Bayes analysis and the problem is often to better approximate the objective function in a functional space, while this paper focuses on approximating the full data estimator.

For computational efficiency, subsampling is often implemented with replacement or through Poisson subsampling. Subsampling with replacement needs to use all subsampling probabilities simultaneously to generate random numbers from a multinomial distribution. The resultant subsample observations are independent and identically distributed (i.i.d.) conditional on the full data, but their unconditional distributions are not independent. Pois-

[1]The estimation efficiency is different from that discussed in Chapter 8 of van der Vaart (1998), which focuses on achieving the asymptotic lower bound of regular estimators. Here we focus on taking a subsample that better approximates the full data estimator, and we consider it with computational efficiency simultaneously.

son subsampling considers each data point and determines if it should be included in the subsample by generating a random number from the uniform distribution. If the subsampling probabilities in Poisson subsample are all equal, then the subsampling procedure is also called the Bernoulli subsampling (Särndal et al. 2003). For Poisson subsampling, the resultant subsample observations do not have identical conditional distributions, but their unconditional distributions can be independent.

Although subsampling with replacement and Poisson subsampling are commonly used in practice, no rigorous investigation has been performed to compare them, especially in the context of optimal subsampling. When they perform similarly and when one is preferable to the other? This paper studies this topic, and has the following major contributions. 1) In the context when an estimator is obtained by maximizing a target function, we first derive conditional and unconditional asymptotic distributions for estimators from both subsampling with replacement and Poisson subsampling. These asymptotic distributions accurately characterize the subsampling approximation errors, and we derive general structure results of optimal subsampling probabilities to minimize these errors for the two subsampling procedures. 2) We systematically compare subsampling with replacement and Poisson subsampling, both theoretically and empirically. We identify conditions when the asymptotic distributions for subsampling with replacement and for Poisson subsampling are the same, and when they are different. We also discuss the similarity and difference for the two subsampling procedures in terms of the structural results of optimal subsampling probabilities. 3) Based on the optimal subsampling probabilities, we propose practical algorithms and evaluate their performance through both theoretical analysis and numerical experiments.

It is worth mentioning that our investigation views subsampling as a computational tool and investigates it within a statistical framework. For computer scientists, subsampling is a commonly used randomized device to speed up computing by using a subsample estimator to approximate the full data estimator (e.g., McWilliams et al. 2014, Woodruff et al. 2014), while for statisticians resampling is widely adopted in exchangeable bootstrap schemes to build confidence regions (e.g., Shao & Tu 1995, Politis et al. 1999). This paper lies in the middle of these two communities. We derive asymptotic distributions of subsampling estimators in a similar fashion to existing literature on bootstrap. However, our purpose is not to establish the bootstrap consistency. Instead, we utilize the asymptotic distributions to develop better subsampling probabilities so that the subsample estimator better approximate the full data estimator. In addition, we focus on data dependent subsampling probabilities for which existing investigations and techniques on bootstrap do not apply because they require data independent and exchangeable sampling weights (Præstgaard & Wellner 1993, Cheng & Huang 2010).

The rest of the paper is organized as follows. We present the model setup and asymptotic distributions in Section 2. In Section 3, we derive optimal subsampling probabilities and propose practical algorithms. We will also obtain theoretical properties for the practical algorithms. In Section 4, we perform numerical experiments demonstrating the performance of the proposed methods. Proofs of our theoretical results are provided in the appendix.

Here are some notation conventions to be used in the paper. We use * to indicate subsample quantities; use $\hat{}$ to indicate full data estimator; use $\tilde{}$ to indicate subsample estimator; use $_R$ and $_P$ to indicate subsampling with replacement and Poisson subsampling, respectively; use $\dot{m}$ and $\ddot{m}$ to denote the gradient and Hessian matrix of a function m with respect to the parameter $\boldsymbol{\theta}$; use $o_P(1)$ or $O_P(1)$ to denote a sequence that converges to zero in probability or is bounded in probability, respectively; use $\rightsquigarrow$ to denote convergence in distribution; use $\|\boldsymbol{v}\|$ to denote the Euclidean norm of a vector $\boldsymbol{v}$; and use $\|\boldsymbol{A}\|$ to denote the Frobenius norm of a matrix $\boldsymbol{A}$.

2 Problem setup and asymptotic distributions

Suppose that a set of training data $\mathcal{D}_n = \{Z_i\}_{i=1}^n$ consists of independent observations from the distribution that generates Z. To estimate some parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ about the data distribution, we want to calculate $\hat{\boldsymbol{\theta}}_n$, the maximizer of

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \boldsymbol{\theta}).$$

Here the dimension of Z_i does not have to be the same as $\boldsymbol{\theta}$, e.g., in softmax regression. Usually, there is no closed-form solution to $\hat{\boldsymbol{\theta}}_n$, and an iterative algorithm is required to find the solution numerically. For massive data, iterative calculations on the full data of size n are often too expensive, so subsampling is adopted to produce a subsampling estimator $\tilde{\boldsymbol{\theta}}$ to approximate $\hat{\boldsymbol{\theta}}_n$. Nonuniform subsampling probabilities are often used to improve the estimation efficiency.

Let $\boldsymbol{\pi} = \{\pi_{n,i}\}_{i=1}^n$ be a subsampling distribution such that $\pi_{n,i} > 0$ and $\sum_{i=1}^n \pi_{n,i} = 1$. For Poisson subsampling, we further assume that $\pi_{n,i} \leq s_n^{-1}$, where s_n is the expected subsample size. As stated early, we use * to indicate quantities with randomness due to subsampling. For instance, let $Z_1^*, ..., Z_{s_n}^*$ denote the resampled sample and let $\pi_{n,1}^*, ..., \pi_{n,s_n}^*$ be the corresponding resampled subsampling probabilities.

We present the general subsampling estimators $\tilde{\boldsymbol{\theta}}_{s_n,R}$ based on subsampling with replacement and $\tilde{\boldsymbol{\theta}}_{s_n,P}$ based on Poisson subsampling, comparatively, in the following Algorithm 1.

Algorithm 1 Subsampling with replacement vs Poisson subsampling

Sampling with replacement

- Calculate $\boldsymbol{\pi} = \{\pi_{n,i}\}_{i=1}^{n}$ based on $\mathcal{D}_n$;
- generate s_n independent random numbers from multinomial distribution with $\boldsymbol{\pi}$ to determine a subsample $\mathcal{D}_{s_n}^* = \{Z_1^*, Z_2^*, ..., Z_{s_n}^*\}$;
- record $\{\pi_{n,1}^*, \pi_{n,2}^*, ..., \pi_{n,s_n}^*\}$ in the subsample;
- obtain the subsample estimator

$$\tilde{\boldsymbol{\theta}}_{s_n,R} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{s_n} \frac{m(Z_i^*, \boldsymbol{\theta})}{n s_n \pi_{n,i}^*}. \quad (1)$$

Poisson Sampling:

- For each $i = 1, ..., n$, calculate an individual $\pi_{n,i}$ such that $\pi_{n,i} \leq s_n^{-1}$ based on Z_i;
- generate $u_i \sim U(0,1)$;
- if $u_i \leq s_n \pi_{n,i}$, include Z_i in the subsample and record $\pi_{n,i}$;
- obtain the subsample estimator

$$\tilde{\boldsymbol{\theta}}_{s_n,P} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{s_n^*} \frac{m(Z_i^*, \boldsymbol{\theta})}{n s_n^* \pi_{n,i}^*}. \quad (2)$$

Remark 1. In Algorithm 1, we see that subsampling with replacement requires to access the whole sampling distribution $\boldsymbol{\pi} = \{\pi_{n,i}\}_{i=1}^{n}$, i.e., all $\pi_{n,i}$'s, because they are the parameters in the multinomial distribution from which random numbers are generated. On the other hand, Poisson subsampling only needs to access one $\pi_{n,i}$ in each sampling consideration. This makes the Poisson subsampling more convenient to implement, especially when the available memory cannot hold all $\pi_{n,i}$'s or in distributed computing platforms. For subsampling with replacement, the subsample size is equal to s_n and there may be replicates in the subsample. Here $\pi_{n,i}$ is the probability that observation Z_i is selected when only one data point is selected, and the probability to include Z_i in the subsample of size s_n is $1 - (1 - \pi_{n,i})^{s_n}$, which is smaller than $s_n \pi_{n,i}$. For Poisson subsampling, the subsample size s_n^* is random with $\mathbb{E}(s_n^*) = s_n$; there is no replicates in the subsample; and $s_n \pi_{n,i}$ is the probability of including Z_i in the subsample of expected size s_n.

Remark 2. Another way of implementing Poisson subsampling is to remove the condition of $\pi_{n,i} \leq s_n^{-1}$ and replace $\pi_{n,i}$ with $\min(s_n \pi_{n,i}, 1)$. The expected subsample size from this approach would be difficult to determine as $\pi_{n,i}$'s are often calculated on the go as scanning through the full data. We only know that the expected subsample size would be smaller than s_n. In this paper, we focus on the Poisson subsampling procedure described in Algorithm 1.

We now derive asymptotic properties of $\tilde{\boldsymbol{\theta}}_{s_n,R}$ in (1) and $\tilde{\boldsymbol{\theta}}_{s_n,P}$ in (2), respectively, to compare their estimation efficiency theoretically. We need some regularity assumptions listed below.

Assumption 1. *The parameter $\boldsymbol{\theta}$ belongs to a compact set.*

Assumption 2. *The function $m(Z, \boldsymbol{\theta})$ is a concave function of $\boldsymbol{\theta}$ with a unique and finite maximum, and it satisfies that $\mathbb{E}\{m^2(Z, \boldsymbol{\theta})\} < \infty$ for any $\boldsymbol{\theta}$.*

Assumption 3. *The matrix $-\mathbb{E}\{\ddot{m}(Z, \boldsymbol{\theta})\}$ is positive-definite, $\mathbb{E}\{\ddot{m}^2_{k,l}(Z, \boldsymbol{\theta})\} < \infty$, and $\ddot{m}(Z, \boldsymbol{\theta})$ is Lipschitz continuous in $\boldsymbol{\theta}$ so that there exists a function $\psi(z)$ with $\mathbb{E}\{\psi^2(Z)\} < \infty$ and for every $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, $|\ddot{m}_{k,l}(z, \boldsymbol{\theta}_1) - \ddot{m}_{k,l}(z, \boldsymbol{\theta}_2)| \leq \psi(z)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, $k, l = 1, 2, ..., d$.*

Assumption 4. *The matrix $\Lambda(\boldsymbol{\theta}) = \mathbb{E}\{\dot{m}(Z, \boldsymbol{\theta})\dot{m}^{\mathrm{T}}(Z, \boldsymbol{\theta})\}$ is positive-definite, and for $\boldsymbol{\theta}$ in the neighborhood of $\hat{\boldsymbol{\theta}}_n$, $\frac{1}{n}\sum_{i=1}^{n} \|\dot{m}(Z_i, \boldsymbol{\theta})\|^4 = O_P(1)$.*

Assumption 5. *The sampling distribution $\boldsymbol{\pi}$ satisfies that $\max_{i=1,...,n}(n\pi_{n,i})^{-1} = O_P(1)$.*

Assumptions 1 and 2 are very mild, and they assure that the target function has a finite and unique maximum. Assumptions 3 and 4 impose some constraints on the Hessian matrix and gradient of $m(Z, \boldsymbol{\theta})$; Assumption 3 is used to prove the consistency of subsample estimators and Assumption 4 is used to establish the asymptotic normality of subsample estimators. Assumption 5 essentially requires that the minimum subsampling probability is at the same order of $\frac{1}{n}$ in probability. Here, $\pi_{n,i}$ can be random as it is allowed to depend on the data, so the notation $O_P(1)$ is used. This assumption is required so that the objective function based on a subsample would not be dominated by data points with very small $\pi_{n,i}$'s. Very small $\pi_{n,i}$'s may not matter when characterizing the worst-case bound, e.g., Drineas et al. (2012), but they do impact the statistical properties of subsampling algorithms. Due to this, Ma et al. (2015) proposed the "shrinkage" leverage scores to prevent the statistical performance of algorithmic leveraging algorithm from being deteriorated by very small leverage scores.

Let $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} \mathbb{E}\{m(Z, \boldsymbol{\theta})\}$ be the true parameter that generates the data. The following proposition is a known result (see, e.g., Chapter 5 of van der Vaart 1998).

Proposition 1. *Under Assumptions 1 and 3, if $\Lambda(\boldsymbol{\theta})$ is positive-definite (the first part of Assumption 4), then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \rightsquigarrow \mathbb{N}\{\mathbf{0}, V(\boldsymbol{\theta}_0)\},$$

where $V(\boldsymbol{\theta}) = \ddot{M}^{-1}(\boldsymbol{\theta})\Lambda(\boldsymbol{\theta})\ddot{M}^{-1}(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta}) = \mathbb{E}\{m(\boldsymbol{\theta}, Z)\}$.

To assess the distributional properties of subsample estimators, we need to derive the distribution asymptotically, i.e., to assume that $s_n \to \infty$ and $n \to \infty$. We assume that $s_n < n$, because a primary goal of subsampling is to reduce the subsample size, but we do not require $s_n = o(n)$.

We define some notations for convergence in conditional distribution and probability before presenting our results. Let Δ_{n,s_n} be a vector function of a subsample of size s_n from the full data $\mathcal{D}_n$, e.g., a subsample estimator. We say that Δ_{n,s_n} converges in conditional probability given $\mathcal{D}_n$ in probability and write it as $\Delta_{n,s_n} = o_{P|\mathcal{D}_n}(1)$, if $\mathbb{P}(\|\Delta_{n,s_n}\| > \delta|\mathcal{D}_n) = o_P(1)$ for any $\delta > 0$; this can be equivalently stated as for any $\delta > 0$ and $\epsilon > 0$, as $s_n \to \infty$ and $n \to \infty$,

$$\mathbb{P}\Big\{\mathbb{P}(\|\Delta_{n,s_n}\| > \delta|\mathcal{D}_n) \le \epsilon\Big\} \to 1.$$

We say that Δ_{n,s_n} is bounded in conditional probability given $\mathcal{D}_n$ in probability and write it as $\Delta_{n,s_n} = O_{P|\mathcal{D}_n}(1)$, if for any $\epsilon > 0$ there exists a $0 < K_\epsilon < \infty$ such that as $s_n \to \infty$ and $n \to \infty$,

$$\mathbb{P}\big\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_\epsilon|\mathcal{D}_n) \le \epsilon\big\} \to 1.$$

We say that Δ_{n,s_n} (of dimension d) converges in conditional distribution to a continuous random vector U given $\mathcal{D}_n$ in probability and denote this as $\Delta_{n,s_n} \overset{|\mathcal{D}_n}{\rightsquigarrow} U$, if $\mathbb{P}(\Delta_{n,s_n} \le \boldsymbol{x}|\mathcal{D}_n) - \mathbb{P}(U \le \boldsymbol{x}) = o_P(1)$ for every $\boldsymbol{x} \in \mathbb{R}^d$; this can also be stated as that for any $\epsilon > 0$ and every $\boldsymbol{x} \in \mathbb{R}^d$, as $s_n \to \infty$ and $n \to \infty$,

$$\mathbb{P}\Big\{\Big|\mathbb{P}(\Delta_{n,s_n} \le \boldsymbol{x}|\mathcal{D}_n) - \mathbb{P}(U \le \boldsymbol{x})\Big| \le \epsilon\Big\} \to 1.$$

Proposition 2. *The following results hold for conditional convergence.*

(a) If $\Delta_{n,s_n} = o_{P|\mathcal{D}_n}(1)$ then $\Delta_{n,s_n} = o_P(1)$, and vice versa.

(b) If $\Delta_{n,s_n} = O_{P|\mathcal{D}_n}(1)$ then $\Delta_{n,s_n} = O_P(1)$, and vice versa.

(c) If $\Delta_{n,s_n} \overset{|\mathcal{D}_n}{\rightsquigarrow} U$ then $\Delta_{n,s_n} \rightsquigarrow U$, and vice versa.

The following Theorems 1 and 2 present conditional asymptotic distributions of $\tilde{\boldsymbol{\theta}}_{s_n,R}$ in (1) and $\tilde{\boldsymbol{\theta}}_{s_n,P}$ in (2), respectively, when approximating the full data estimator $\hat{\boldsymbol{\theta}}_n$.

Theorem 1. *Under Assumptions 1-5, as $s_n \to \infty$ and $n \to \infty$, the estimator $\tilde{\boldsymbol{\theta}}_{s_n,R}$ in (1) satisfies that*

$$\sqrt{s_n}\{V_{n,R}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n) \overset{|\mathcal{D}_n}{\rightsquigarrow} \mathbb{N}(\mathbf{0}, \boldsymbol{I}_d), \tag{3}$$

where $\mathbb{N}(\mathbf{0}, \boldsymbol{I}_d)$ is a multivariate Gaussian distribution with mean $\mathbf{0}$ and variance $\boldsymbol{I}_d$ (the identity matrix of dimension d), $V_{n,R}(\boldsymbol{\theta}) = \ddot{M}_n^{-1}(\boldsymbol{\theta})\Lambda_{n,R}(\boldsymbol{\theta})\ddot{M}_n^{-1}(\boldsymbol{\theta})$,

$$\ddot{M}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ddot{m}(Z_i, \boldsymbol{\theta}), \quad \text{and} \quad \Lambda_{n,R}(\boldsymbol{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n} \frac{\dot{m}(Z_i, \boldsymbol{\theta})\dot{m}^{\mathrm{T}}(Z_i, \boldsymbol{\theta})}{\pi_{n,i}}. \tag{4}$$

Theorem 2. *Under Assumptions 1-5, as $s_n \to \infty$ and $n \to \infty$, the estimator $\tilde{\boldsymbol{\theta}}_{s_n,P}$ in (2) satisfies that,*

$$\sqrt{s_n}\{V_{n,P}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}(\tilde{\boldsymbol{\theta}}_{s_n,P} - \hat{\boldsymbol{\theta}}_n) \overset{|\mathcal{D}_n}{\rightsquigarrow} \mathbb{N}(\mathbf{0}, \boldsymbol{I}_d), \tag{5}$$

where $V_{n,P}(\boldsymbol{\theta}) = \ddot{M}_n^{-1}(\boldsymbol{\theta})\Lambda_{n,P}(\boldsymbol{\theta})\ddot{M}_n^{-1}(\boldsymbol{\theta})$, $\ddot{M}_n(\boldsymbol{\theta})$ is the same as in (4), and

$$\Lambda_{n,P}(\boldsymbol{\theta}) = \Lambda_{n,R}(\boldsymbol{\theta}) - \frac{s_n}{n^2}\sum_{i=1}^{n}\dot{m}(Z_i,\boldsymbol{\theta})\dot{m}^{\mathrm{T}}(Z_i,\boldsymbol{\theta}). \tag{6}$$

Remark 3. The asymptotic distributions in (3) and (5) mean that given a full data set for any $\delta > 0$, the probability that $\|\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n\| > \delta$ is accurately approximated by $\mathbb{P}(\|U_R\| > \delta)$ where $U_R \sim \mathbb{N}\{\mathbf{0}, V_{n,R}(\hat{\boldsymbol{\theta}}_n)\}$, and the probability that $\|\tilde{\boldsymbol{\theta}}_{s_n,P} - \hat{\boldsymbol{\theta}}_n\| > \delta$ is accurately approximated by $\mathbb{P}(\|U_P\| > \delta)$ where $U_P \sim \mathbb{N}\{\mathbf{0}, V_{n,P}(\hat{\boldsymbol{\theta}}_n)\}$. Thus, a smaller variance means a smaller probability of excess error at the same error bound, or a smaller error bound for the same excess probability.

Remark 4. Both $\tilde{\boldsymbol{\theta}}_{s_n,R}$ and $\tilde{\boldsymbol{\theta}}_{s_n,P}$ have Gaussian asymptotic distributions, but they have different asymptotic variances $V_{n,R}(\hat{\boldsymbol{\theta}}_n)$ and $V_{n,P}(\hat{\boldsymbol{\theta}}_n)$, respectively. Under Assumption 4, the second term on the right-hand-side of (6) goes to zero in probability if $s_n/n \to 0$, and it converges to a positive-definite matrix in probability if $s_n/n \to c > 0$. Thus, the difference $V_{n,R}(\hat{\boldsymbol{\theta}}_n) - V_{n,P}(\hat{\boldsymbol{\theta}}_n) \to \mathbf{0}$ in probability if $s_n/n \to 0$, and it converges to a positive-definite matrix in probability if s_n/n converges to a positive constant. This means that subsampling with replacement and Poisson subsampling have the same asymptotic estimation efficiency only if the subsampling ratio s_n/n goes to zero; otherwise, Poisson subsampling has a higher estimation efficiency. Thus, to obtain more accurate estimates in practice, Poisson subsampling is recommended unless the subsampling ratio s_n/n is very small.

Remark 5. If the sampling distribution $\boldsymbol{\pi}$ is constructed so that $\Lambda_{n,R}(\boldsymbol{\theta}) \to \Lambda(\boldsymbol{\theta})$ in probability uniformly in a neighborhood of $\boldsymbol{\theta}_0$, then $V_{n,R}(\hat{\boldsymbol{\theta}}_n)$ and $(1-c)^{-1}V_{n,P}(\hat{\boldsymbol{\theta}}_n)$ both converge in probability to $V(\boldsymbol{\theta}_0)$, the scaled asymptotic variance of $\hat{\boldsymbol{\theta}}$. This means both subsample estimators have the bootstrap consistency in this scenario. A class of sampling distributions satisfies this situation if $\boldsymbol{\pi}$ does not dependent on the data, such as the class of exchangeable bootstrap weights which includes the uniform sampling distribution (see Præstgaard & Wellner 1993, Cheng & Huang 2010). However if $\boldsymbol{\pi}$ depends on the data, then $\Lambda_{n,R}(\boldsymbol{\theta})$ may not converge to $\Lambda(\boldsymbol{\theta})$ [2], and in this case the subsample estimators do not have the bootstrap consistency. The goal of this paper is different from the line of research about bootstrap

[2]This is still possible in some special cases such as the local case control subsampling for logistic regression (Fithian & Hastie 2014, Wang 2019).

that focuses on constructing conference region nor approximating complicated distributions (see Bickel et al. 1997, Politis et al. 1999), so bootstrap inconsistency is not a concern. Nevertheless, if multiple subsamples are taken, then the average of the subsample estimates is recommended and the variance can be estimated from these subsample estimates using the approach proposed in Wang & Ma (2021).

Although the convergence in conditional distribution $\overset{|\mathcal{D}_n}{\rightsquigarrow}$ can be replaced by $\rightsquigarrow$ because of Proposition 2 (c), Theorems 1 and 2 are about approximating the full data estimator and they are conditional results in nature. In the following, we derive the unconditional asymptotic distribution when the true parameter is of interest to further compare the two subsampling approaches.

Theorem 1'. *Under Assumptions 1-5, if $\Lambda_{n,R}(\boldsymbol{\theta}_0)$ converges to a positive-definite matrix $\Lambda_\pi(\boldsymbol{\theta}_0)$ as $s_n \to \infty$ and $n \to \infty$, then the estimator $\tilde{\boldsymbol{\theta}}_{s_n,R}$ in (1) satisfies that*

$$\sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,R} - \boldsymbol{\theta}_0) \rightsquigarrow \mathbb{N}\big\{\mathbf{0},\ V_R^U(\boldsymbol{\theta}_0)\big\},$$

where $V_R^U(\boldsymbol{\theta}) = \ddot{M}^{-1}(\boldsymbol{\theta})\Lambda_R^U(\boldsymbol{\theta})\ddot{M}^{-1}(\boldsymbol{\theta})$, $\Lambda_R^U(\boldsymbol{\theta}) = \Lambda_\pi(\boldsymbol{\theta}) + c\Lambda(\boldsymbol{\theta})$, and $c = \lim \frac{s_n}{n}$.

Theorem 2'. *Under Assumptions 1-5, if $\Lambda_{n,R}(\boldsymbol{\theta}_0)$ converges to a positive-definite matrix $\Lambda_\pi(\boldsymbol{\theta}_0)$ as $s_n \to \infty$ and $n \to \infty$, then the estimator $\tilde{\boldsymbol{\theta}}_{s_n,P}$ in (2) satisfies that*

$$\sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,P} - \boldsymbol{\theta}_0) \rightsquigarrow \mathbb{N}\big\{\mathbf{0},\ V_P^U(\boldsymbol{\theta}_0)\big\},$$

where $V_P^U(\boldsymbol{\theta}) = \ddot{M}^{-1}(\boldsymbol{\theta})\Lambda_\pi(\boldsymbol{\theta})\ddot{M}^{-1}(\boldsymbol{\theta})$.

Remark 6. In Theorems 1' and 2', the unconditional asymptotic distributions of $\tilde{\boldsymbol{\theta}}_{s_n,R}$ and $\tilde{\boldsymbol{\theta}}_{s_n,P}$ for estimating the true parameter are also Gaussian with (scaled) variances $V_R^U(\boldsymbol{\theta}_0)$ and $V_P^U(\boldsymbol{\theta}_0)$, respectively. From the two theorems, we see that $V_R^U(\boldsymbol{\theta}_0) = V_P^U(\boldsymbol{\theta}_0) + cV(\boldsymbol{\theta}_0)$, where $V(\boldsymbol{\theta}_0)$ is the scaled asymptotic variance for the full data estimator in Propositio 1. Here, $V_P^U(\boldsymbol{\theta}_0)$ can be interpreted as the variation due to subsampling and $cV(\boldsymbol{\theta}_0)$ can be interpreted as the variation due to the randomness of the full data. It is interesting to note that the asymptotic variance components due to the two sources are additive for the subsampling with replacement estimator $\tilde{\boldsymbol{\theta}}_{s_n,R}$, while $cV(\boldsymbol{\theta}_0)$ does not contribute to the asymptotic variance of the Poisson subsampling estimator $\tilde{\boldsymbol{\theta}}_{s_n,P}$.

3 Optimal subsampling probabilities

From the results in Theorems 1 and 2, the asymptotic variances $V_{n,R}(\hat{\boldsymbol{\theta}}_n)$ and $V_{n,P}(\hat{\boldsymbol{\theta}}_n)$ depend on $\boldsymbol{\pi} = \{\pi_{n,i}\}_{i=1}^n$. To improve the estimation efficiency, we want to choose optimal $\boldsymbol{\pi}$

to minimize $V_{n,R}(\hat{\boldsymbol{\theta}}_n)$ or $V_{n,P}(\hat{\boldsymbol{\theta}}_n)$. Specifically, we consider the L-optimality criterion (Section 10.5 of Atkinson et al. 2007). The L-optimality minimizes the trace of the asymptotic variance matrix for some **l**inear transformation, say L, of the parameter estimator. For our case, this is to minimize $\text{tr}\{LV_{n,R}(\hat{\boldsymbol{\theta}}_n)L^{\text{T}}\}$ or $\text{tr}\{LV_{n,P}(\hat{\boldsymbol{\theta}}_n)L^{\text{T}}\}$ for some matrix L, because $LV_{n,R}(\hat{\boldsymbol{\theta}}_n)L^{\text{T}}$ and $LV_{n,P}(\hat{\boldsymbol{\theta}}_n)L^{\text{T}}$ are the asymptotic variances of $L\tilde{\boldsymbol{\theta}}_{s_n,R}$ and $L\tilde{\boldsymbol{\theta}}_{s_n,R}$, respectively. If we take $L = \boldsymbol{I}$, then the resulting criterion is also called the A-optimality; this is to minimize the **a**verage of the variances for all parameter components by minimizing the trace of the variance matrix, i.e., minimizing $\text{tr}\{V_{n,R}(\hat{\boldsymbol{\theta}}_n)\}$ or $\text{tr}\{V_{n,P}(\hat{\boldsymbol{\theta}}_n)\}$. If we take $L = \ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$, then the resultant criterion is to minimize $\text{tr}\{\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)\}$ or $\text{tr}\{\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n)\}$. This has a computational advantage compared with other choices, so we focus more on this choice in this paper. The following Theorems 3 and 4 present the optimal subsampling probabilities for subsampling with replacement and Poisson subsampling, respectively.

Theorem 3. *For the subsampling with replacement estimator in (1), the L-optimal subsampling probabilities with $L = \ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ that minimize $\text{tr}\{\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)\}$ are*

$$\pi_{n,Ri}^{\text{opt}} = \frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|}{\sum_{j=1}^{n} \|\dot{m}(Z_j, \hat{\boldsymbol{\theta}}_n)\|}, \quad i = 1, ..., n. \tag{7}$$

Theorem 4. *For the Poisson subsampling estimator in (2), the L-optimal subsampling probabilities with $L = \ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ that minimize $\text{tr}\{\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n)\}$ are*

$$\pi_{n,Pi}^{\text{opt}} = \frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \wedge H}{\sum_{j=1}^{n} \{\|\dot{m}(Z_j, \hat{\boldsymbol{\theta}}_n)\| \wedge H\}}, \quad i = 1, ..., n, \tag{8}$$

where $a \wedge b = \min(a, b)$,

$$H = \frac{\sum_{i=1}^{n-g} \|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(i)}}{s_n - g}, \tag{9}$$

$\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(1)} \leq ... \leq \|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n)}$ are the order statistics of $\|\dot{m}(Z_1, \hat{\boldsymbol{\theta}}_n)\|, ..., \|\dot{m}(Z_n, \hat{\boldsymbol{\theta}}_n)\|$, and g is an integer such that

$$\frac{\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n-g)}}{\sum_{i=1}^{n-g} \|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(i)}} < \frac{1}{s_n - g} \quad \textit{and} \quad \frac{\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n-g+1)}}{\sum_{i=1}^{n-g+1} \|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(i)}} \geq \frac{1}{s_n - g + 1}, \tag{10}$$

in which we define $\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n+1)} = \infty$.

Remark 7. For a general choice of L, we can obtain optimal subsampling probabilities by replacing $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|$ with $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|_L = \|L\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|$. However, these quantities require $O(nd^2)$ time to compute when $\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)$ and $\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)$ are available, where n is the full data sample size and d is dimension of $\hat{\boldsymbol{\theta}}_n$. On the other hand, it only takes $O(nd)$ time to compute all $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|$'s. Thus the choice of $L = \ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ has a significant computational advantage.

Remark 8. In Theorems 3 and 4, $\pi_{n,Ri}^{\text{opt}}$ in (7) and $\pi_{n,Pi}^{\text{opt}}$ in (8) have both similarities and differences. Assuming that $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| > 0$ for all i, then $0 < \pi_{n,Ri}^{\text{opt}} < 1$ while $0 < \pi_{n,Pi}^{\text{opt}} \leq \frac{1}{s_n}$. This means that the inclusion of any data point through optimal subsampling with replacement is random, while the inclusion of data points with $\pi_{n,Pi}^{\text{opt}} = \frac{1}{s_n}$ is deterministic through optimal Poisson subsampling. The order statistics constraint in (10) indicates that if there are data points such that $\frac{s_n}{n}\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| > \frac{1}{n}\sum_{j=1}^{n}\|\dot{m}(Z_j, \hat{\boldsymbol{\theta}}_n)\|$, then $\pi_{n,Ri}^{\text{opt}}$ and $\pi_{n,Pi}^{\text{opt}}$ are different. This means that if the subsampling ratio $\frac{s_n}{n}$ is larger or if the tail of the distribution of $\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|$ is heavier, then optimal probabilities for Poisson subsampling and subsampling with replacement are more likely to be different. If $s_n\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n)} < \sum_{i=1}^{n}\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|$, then $\pi_{n,Ri}^{\text{opt}}$ and $\pi_{n,Pi}^{\text{opt}}$ are identical. This condition is true with probability approaching one under some conditions, e.g., when $s_n \log n = o(n)$ and the distribution of $\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|$ has a sub-Gaussian tail because in this case $\frac{s_n}{n}\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n)} = o_P(1)$ and $\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|$ goes to a positive constant in probability.

Remark 9. In Theorem 4, H is the threshold so that all $\pi_{n,Pi}^{\text{opt}}$ are no larger than $\frac{1}{s_n}$, and it satisfies that

$$\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n-g)} < H \leq \|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n-g+1)}. \tag{11}$$

Here g is the number of cases that $\pi_{n,Pi}^{\text{opt}} = \frac{1}{s_n}$, i.e., the number of data points that will be included in the subsample for sure.

Now we discuss an example to illustrate the optimal structural results. Additional examples are available in Section A.2 of the Appendix.

Example 1 (Binary response models)**.** Consider a binary classification model such that

$$\mathbb{P}(y_i = 1) = p(\boldsymbol{x}_i, \boldsymbol{\theta}), \quad i = 1, ..., n,$$

where $y_i \in \{0, 1\}$ is the binary class label, $\boldsymbol{x}_i$ is the covariate, and $\boldsymbol{\theta}$ is the unknown parameter. To estimate $\boldsymbol{\theta}$ using the maximum likelihood estimator (MLE), let $Z_i = (\boldsymbol{x}_i, y_i)$ and

$$m(Z_i, \boldsymbol{\theta}) = y_i \log\{p(\boldsymbol{x}_i, \boldsymbol{\theta})\} + (1 - y_i)\log\{1 - p(\boldsymbol{x}_i, \boldsymbol{\theta})\}.$$

Direct calculations yield that

$$\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = \frac{y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}\dot{\hat{p}}_i, \quad \text{and} \quad \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| = \frac{|y_i - \hat{p}_i|}{\hat{p}_i(1 - \hat{p}_i)}\|\dot{\hat{p}}_i\|, \tag{12}$$

where $\hat{p}_i = p(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_n)$, and $\dot{\hat{p}}_i = \dot{p}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_n)$ is the gradient of $p(\boldsymbol{x}_i, \boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}_n$. We can obtain optimal sampling probabilities by inserting the expression in (12) into Theorems 3 and 4.

To obtain the general L-optimal subsampling probabilities with any L, the Hessian matrix $\ddot{m}(Z_i, \hat{\boldsymbol{\theta}})$ of $m(Z_i, \hat{\boldsymbol{\theta}})$ is

$$\ddot{m}(Z_i, \hat{\boldsymbol{\theta}}) = \frac{y_i - \hat{p}_i}{\hat{p}_i(1-\hat{p}_i)}\hat{\ddot{p}}_i - \Big\{\frac{y_i}{\hat{p}_i^2} + \frac{1-y_i}{(1-\hat{p}_i)^2}\Big\}\hat{\dot{p}}_i\hat{\dot{p}}_i^{\mathrm{T}}, \tag{13}$$

where $\hat{\ddot{p}}_i = \ddot{p}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_n)$ is the Hessian matrix of $p(\boldsymbol{x}_i, \boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}}_n$. Thus, we obtain the general L-optimal sampling probabilities by using

$$\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|_L = \frac{|y_i - \hat{p}_i|}{\hat{p}_i(1-\hat{p}_i)}\|L\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\hat{\dot{p}}_i\|, \tag{14}$$

to replace $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|$ in Theorems 3 and 4, for any L, where

$$\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = \sum_{i=1}^{n}\frac{y_i - \hat{p}_i}{\hat{p}_i(1-\hat{p}_i)}\hat{\ddot{p}}_i - \sum_{i=1}^{n}\Big\{\frac{y_i}{\hat{p}_i^2} + \frac{1-y_i}{(1-\hat{p}_i)^2}\Big\}\hat{\dot{p}}_i\hat{\dot{p}}_i^{\mathrm{T}}. \tag{15}$$

Under some regularity conditions, $\frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\hat{p}_i}{\hat{p}_i(1-\hat{p}_i)}\hat{\ddot{p}}_i$ is a small term in (15), and therefore $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ in (14) can be replaced by

$$\ddot{M}_n^a(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{y_i}{\hat{p}_i^2} + \frac{1-y_i}{(1-\hat{p}_i)^2}\Big\}\hat{\dot{p}}_i\hat{\dot{p}}_i^{\mathrm{T}}. \tag{16}$$

Thus, there is no need to calculate the Hessian matrix $\hat{\ddot{p}}_i$.

From (12) or (14), the optimal subsampling probabilities are proportional to $|y_i - \hat{p}_i|$. Thus if $y_i = 1$, data points with smaller values of $\hat{p}_i$ are sampled with higher probabilities; if $y_i = 0$, data points with larger values of $\hat{p}_i$ are sampled with higher probabilities. The optimal subsampling probabilities give higher preference to data points that are closer to the class boundary. This increases the classification accuracy because if these data points can be classified correctly, then other data points are easier to classify.

Specifically for Logistic regression in which

$$p(\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}}}{(1 + e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}})},$$

we have $\hat{\dot{p}}_i = \hat{p}_i(1-\hat{p}_i)\boldsymbol{x}_i$ and $\hat{\ddot{p}}_i = \hat{p}_i(1-\hat{p}_i)(1-2\hat{p}_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$. Thus, for this case

$$\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| = |y_i - \hat{p}_i|\|\boldsymbol{x}_i\|, \quad \text{and} \tag{17}$$

$$\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|_L = |y_i - \hat{p}_i|\|L\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\boldsymbol{x}_i\|, \quad \text{with} \quad \ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i(1-\hat{p}_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}. \tag{18}$$

If the expression in (17), or the expression in (18) with $L = \boldsymbol{I}$, is used in Theorems 3, the structural results for optimal probabilities of subsampling with replacement are identical to

those in Wang et al. (2018). If (16) is used, then the expression of $\ddot{M}_n^a(\hat{\boldsymbol{\theta}}_n)$ is $\ddot{M}_n^a(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{i=1}^n (y_i - \hat{p}_i)^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$, which has the same limit as $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ in (18).

From Theorem 4 we see that if there are data points such that $\frac{s_n}{n}|y_i - \hat{p}_i|\|\boldsymbol{x}_i\| > \frac{1}{n}\sum_{j=1}^n |y_j - \hat{p}_j|\|\boldsymbol{x}_j\|$, then optimal probabilities for Poisson subsampling are different from that for subsampling with replacement.

3.1 Practical algorithms

The optimal subsampling probabilities depend on the full data estimator $\hat{\boldsymbol{\theta}}_n$, so the structural results in the previous section do not translate into useful algorithms directly. We need a pilot estimator to approximate the optimal subsampling probabilities in order to obtain practically implementable algorithms. This can be done by taking a pilot subsample of size s_0 through a subsampling distribution that does not depend on $\hat{\boldsymbol{\theta}}_n$. For simplicity, we use the uniform subsampling distribution $\boldsymbol{\pi}^{\mathrm{uni}} = \{\pi_{n,i} = \frac{1}{n}\}_{i=1}^n$, and present the approximated optimal subsampling with replacement procedure in Algorithm 2.

Compared with the exact $\pi_{n,Ri}^{\mathrm{opt}}$, the approximated $\tilde{\pi}_{n,Ri}^{\mathrm{opt}}$ in (20) are subject to additional disturbance due to the randomness of $\tilde{\boldsymbol{\theta}}_{s_0,R}^{0*}$, the maximizer of (19). From Theorem 1, the subsampling probabilities are in the denominators of $\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)$. Thus the additional disturbance may be amplified for data points with $\pi_{n,Ri}^{\mathrm{opt}}$ being close to zero, and this may inflate the asymptotic variance of the subsample estimator. To protect the estimator from these data points, we adopt the idea of defensive importance sampling (Hesterberg 1995, Owen & Zhou 2000) and mix the approximated optimal subsampling distribution with the uniform subsampling distribution. Specifically, we use $\tilde{\pi}_{n,R\alpha i}^{\mathrm{opt}}$ instead of $\tilde{\pi}_{n,Ri}^{\mathrm{opt}}$ in (20) to perform the subsampling. The same idea was also adopted in Ma et al. (2015).

In $\tilde{\boldsymbol{\pi}}_{R\alpha i} = \{\tilde{\pi}_{n,R\alpha i}^{\mathrm{opt}}\}_{i=1}^n$, α controls the proportion of mixture, and $\tilde{\boldsymbol{\pi}}_{R\alpha i}$ is close to the optimal subsampling distribution if α is close to 0 while it is close to the uniform subsampling distribution if α is close to 1. If $\alpha > 0$, then $n\pi_{n,R\alpha i}^{\mathrm{opt}}$ are bounded away from zero, which add to robustness of the subsampling estimator.

Algorithm 2 Practical algorithm based on optimal subsampling with replacement

• *Pilot subsampling*: use sampling with replacement with $\boldsymbol{\pi}^{\text{uni}}$ to obtain $\{Z_1^{0*}, ..., Z_{s_0}^{0*}\}$; obtain $\tilde{\boldsymbol{\theta}}_{s_0,R}^{0*}$ through maximizing

$$M_R^{0*}(\boldsymbol{\theta}) = \sum_{i=1}^{s_0} \frac{m(Z_i^{0*}, \boldsymbol{\theta})}{s_0}. \tag{19}$$

• *Approximated optimal subsampling:*

calculate the whole subsampling distribution $\tilde{\boldsymbol{\pi}}_{R\alpha i} = \{\tilde{\pi}_{n,R\alpha i}^{\text{opt}}\}_{i=1}^n$, where $\alpha \in (0,1)$,

$$\tilde{\pi}_{n,Ri}^{\text{opt}} = \frac{\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}_{s_0,R}^{0*})\|}{\sum_{j=1}^n \|\dot{m}(Z_j, \tilde{\boldsymbol{\theta}}_{s_0,R}^{0*})\|} \quad \text{and} \quad \tilde{\pi}_{n,R\alpha i}^{\text{opt}} = (1-\alpha)\tilde{\pi}_{n,Ri}^{\text{opt}} + \alpha\frac{1}{n}; \tag{20}$$

use $\tilde{\boldsymbol{\pi}}_{R\alpha i}$ to take a subsample $\{Z_1^*, ..., Z_{s_n}^*\}$, and record the corresponding probabilities $\{\tilde{\pi}_{R\alpha 1}^{\text{opt}*}, ..., \tilde{\pi}_{R\alpha s}^{\text{opt}*}\}$.

• *Estimation:* obtain $\tilde{\boldsymbol{\theta}}_{s_n,R}^{\alpha}$ through maximizing

$$M_{R\alpha}^*(\boldsymbol{\theta}) = \sum_{i=1}^{s_n} \frac{m(Z_i^*, \boldsymbol{\theta})}{n s_n \tilde{\pi}_{n,R\alpha i}^{\text{opt}*}}. \tag{21}$$

Algorithm 3 Practical algorithm based on optimal Poisson subsampling

• *Pilot subsampling*: use Poisson sampling with $\boldsymbol{\pi}^{\text{uni}}$ to obtain $\{Z_1^{0*}, ..., Z_{s_0^*}^{0*}\}$; obtain $\tilde{\boldsymbol{\theta}}_{s_0,P}^{0*}$ through maximizing

$$M_P^{0*}(\boldsymbol{\theta}) = \sum_{i=1}^{s_0^*} \frac{m(Z_i^{0*}, \boldsymbol{\theta})}{s_0^*}; \tag{22}$$

calculate

$$H^{0*} = \|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\|_{\frac{s}{bn}}, \quad \text{and} \quad \Psi^{0*} = \sum_{i=1}^{s_0^*} \frac{\{\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\| \wedge H^{0*}\}}{s_0^*}. \tag{23}$$

• *Approximated optimal subsampling:* For each i of $i = 1, ..., n$,

calculate

$$\tilde{\pi}_{n,Pi}^{\text{opt}} = \frac{\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\| \wedge H^{0*}}{n\Psi^{0*}}, \quad \text{and} \quad \tilde{\pi}_{n,P\alpha i}^{\text{opt}} = (1-\alpha)\tilde{\pi}_{n,Pi}^{\text{opt}} + \alpha\frac{1}{n}; \tag{24}$$

generate $u_i \sim U(0,1)$;

if $u_i \leq s_n \tilde{\pi}_{n,P\alpha i}^{\text{opt}}$, include Z_i in the subsample and record $\tilde{\pi}_{n,P\alpha i}^{\text{opt}}$.

• *Estimation:* obtain $\tilde{\boldsymbol{\theta}}_{s_n,P}^{\alpha}$ through maximizing

$$M_{P\alpha}^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{s_n^*} \frac{m(Z_i^*, \boldsymbol{\theta})}{(s_n \tilde{\pi}_{n,P\alpha i}^{\text{opt}*}) \wedge 1}. \tag{25}$$

For the optimal Poisson subsampling probability $\pi_{n,Pi}^{\text{opt}}$, we also need to use the pilot subsample to approximate H and $\Psi = \frac{1}{n}\sum_{i=1}^{n}\{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \wedge H\}$ in order to determine the inclusion probability based on each data point itself, as described in Algorithm 3. From (11), H is between the $(n-g)$-th and the $(n-g+1)$-th order statistics of $\{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|\}_{i=1}^{n}$, and g is between 0 and s_n, so we can roughly approximate H with $\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\|_{\frac{s_n}{bn}}$, the upper $\frac{s_n}{bn}$-th sample quantile of $\{\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\|\}_{i=1}^{s_0}$, where $b \geq 1$ is a tuning parameter. Since g is typically closer to 0 and farther from s_n, taking $b = 1$ underestimates H and the resulting subsampling probabilities lean towards the uniform subsampling probability (if $H \leq \|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(1)}$, then $\pi_{n,Pi}^{\text{opt}}$ would be all equal to $\frac{1}{n}$). When subsampling from massive data, s_n is often much smaller than n and the number of cases for $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|$ to be larger than H is small. For this scenario, one may simply ignore H and use ∞ to replace H. This simple option in general overestimates H, but it may perform reasonably well for small subsampling ratios. For Ψ, it can be approximated by Ψ^{0*} defined in (23).

When we use Ψ^{0*} and H^{0*} to replace Ψ and H in (24), it is possible that some $\tilde{\pi}_{n,Pi}^{\text{opt}}$ in (24) are larger than $\frac{1}{s_n}$ and thus $s_n\tilde{\pi}_{n,Pi}^{\text{opt}}$ are larger than one. Thus, we use one as a threshold in the denominator of (25).

Remark 10. In Algorithm 2, $\tilde{\boldsymbol{\theta}}_{s_0,R}^{0*}$ and $\tilde{\boldsymbol{\theta}}_{s_n,R}^{\alpha}$ can be combined to obtain an aggregated estimator,

$$\check{\boldsymbol{\theta}}_R = \big(s_0\ddot{M}_R^{0*} + s\ddot{M}_R^{*}\big)^{-1} \times \big(s_0\ddot{M}_R^{0*} \times \tilde{\boldsymbol{\theta}}_{s_0,R}^{0*} + s\ddot{M}_R^{*} \times \tilde{\boldsymbol{\theta}}_{s_n,R}^{\alpha}\big),$$

where $\ddot{M}_R^{0*}$ is the Hessian matrix of $M_R^{0*}(\boldsymbol{\theta})$ in (19) evaluated at $\tilde{\boldsymbol{\theta}}_{s_0,R}^{0*}$ and $\ddot{M}_R^{*}$ is the Hessian matrix of $M_R^{*}(\boldsymbol{\theta})$ in (21) evaluated at $\tilde{\boldsymbol{\theta}}_{s_n,R}^{\alpha}$. Similarly, in Algorithm 3, $\tilde{\boldsymbol{\theta}}_{s_0,P}^{0*}$ and $\tilde{\boldsymbol{\theta}}_{s_n,P}^{\alpha}$ can be combined to obtain an aggregated estimator,

$$\check{\boldsymbol{\theta}}_P = \big(s_0^*\ddot{M}_P^{0*} + s_n\ddot{M}_P^{*}\big)^{-1} \times \big(s_0^*\ddot{M}_P^{0*} \times \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*} + s_n\ddot{M}_P^{*} \times \tilde{\boldsymbol{\theta}}_{s_n,P}^{\alpha}\big),$$

where $\ddot{M}_P^{0*}$ is the Hessian matrix of $M_P^{0*}(\boldsymbol{\theta})$ in (22) evaluated at $\tilde{\boldsymbol{\theta}}_{s_0,P}^{0*}$ and $\ddot{M}_P^{*}$ is the Hessian matrix of $M_P^{*}(\boldsymbol{\theta})$ in (25) evaluated at $\tilde{\boldsymbol{\theta}}_{s_n,P}^{\alpha}$. Here, $\check{\boldsymbol{\theta}}_R$ is obtained as a linear combination of $\tilde{\boldsymbol{\theta}}_{s_0,R}^{0*}$ and $\tilde{\boldsymbol{\theta}}_{s_n,R}^{\alpha}$, and $\check{\boldsymbol{\theta}}_P$ is obtained as a linear combination of $\tilde{\boldsymbol{\theta}}_{s_0,P}^{0*}$ and $\tilde{\boldsymbol{\theta}}_{s_n,P}^{\alpha}$ in a way similar to the aggregation step in the divide-and-conquer method (Lin & Xie 2011, Schifano et al. 2016). This further improves the estimation efficiency.

3.2 Theoretical analysis of practical algorithms

We obtain the following distributional results in Theorems 5 and 6 for Algorithms 2 and 3, respectively.

Theorem 5. *For $\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}$ obtained from Algorithm 2, under Assumptions 1-4, as s_0, s_n, and n get large, the following result holds. Given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$ in probability,*

$$\sqrt{s_n}\{V^{\alpha}_{n,R}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n)\to\mathbb{N}\left(\mathbf{0},\boldsymbol{I}_d\right),$$

in conditional distribution, where $V^{\alpha}_{n,R}(\hat{\boldsymbol{\theta}}_n)=\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\Lambda^{\alpha}_R(\hat{\boldsymbol{\theta}}_n)\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)$,

$$\Lambda^{\alpha}_R(\hat{\boldsymbol{\theta}}_n)=\frac{1}{n^2}\sum_{i=1}^{n}\frac{\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{\pi^{\mathrm{opt}}_{n,R\alpha i}(\hat{\boldsymbol{\theta}}_n)},\quad \textit{and}\quad \pi^{\mathrm{opt}}_{n,R\alpha i}(\hat{\boldsymbol{\theta}}_n)=(1-\alpha)\pi^{\mathrm{opt}}_{n,Ri}(\hat{\boldsymbol{\theta}}_n)+\alpha\frac{1}{n}.$$

Theorem 6. *For $\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,P}$ obtained from Algorithm 3, under Assumptions 1-4, as s_0, s_n, and n get large, if $s_0=o(n)$, $\varrho_n=s_n/(bn)\to\varrho\in[0,1)$, and the distribution of Z is continuous, the following result hold. If $\varrho=0$, then given $\mathcal{D}_n$ and the pilot estimates in probability,*

$$\sqrt{s_n}\{V^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,P}-\hat{\boldsymbol{\theta}}_n)\to\mathbb{N}\left(\mathbf{0},\boldsymbol{I}_d\right),$$

in conditional distribution, where $V^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)=\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\Lambda^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)$,

$$\Lambda^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)=\frac{1}{n^2}\sum_{i=1}^{n}\frac{\{1-s_n\pi^{\mathrm{opt}}_{n,P\alpha i}(\hat{\boldsymbol{\theta}}_n)\}\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{\pi^{\mathrm{opt}}_{n,P\alpha i}(\hat{\boldsymbol{\theta}}_n)}\tag{26}$$

and $\pi^{\mathrm{opt}}_{n,P\alpha i}(\hat{\boldsymbol{\theta}}_n)=(1-\alpha)\pi^{\mathrm{opt}}_{n,Pi}(\hat{\boldsymbol{\theta}}_n)+\alpha\frac{1}{n}$. If $\varrho>0$, then $\pi^{\mathrm{opt}}_{n,Pi}(\hat{\boldsymbol{\theta}}_n)$ in (26) is replaced by $\pi^{\mathrm{opt}}_{n,Pi}=\frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\wedge H_{\varrho_n}}{\sum_{j=1}^{n}\{\|\dot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)\|\wedge H_{\varrho_n}\}}$, where H_{ϱ_n} is the ϱ_n-th upper sample quantile of $\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|$'s for $i=1,...,n$.

Remark 11. Denote $\Lambda^{\mathrm{opt}}_R(\hat{\boldsymbol{\theta}}_n)$ and $\Lambda^{\mathrm{opt}}_{n,P}(\hat{\boldsymbol{\theta}}_n)$ as $\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)$ and $\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n)$ with optimal subsampling probabilities that produce the minimum trace values, respectively. In Theorems 5 and 6, $\Lambda^{\alpha}_R(\hat{\boldsymbol{\theta}}_n)$ and $\Lambda^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)$ are different from $\Lambda^{\mathrm{opt}}_R(\hat{\boldsymbol{\theta}}_n)$ and $\Lambda^{\mathrm{opt}}_{n,P}(\hat{\boldsymbol{\theta}}_n)$, respectively. However, it can be shown that

$$\mathrm{tr}\{\Lambda^{\mathrm{opt}}_R(\hat{\boldsymbol{\theta}}_n)\}<\mathrm{tr}\{\Lambda^{\alpha}_R(\hat{\boldsymbol{\theta}}_n)\}<\frac{\mathrm{tr}\{\Lambda^{\mathrm{opt}}_R(\hat{\boldsymbol{\theta}}_n)\}}{1-\alpha},\ \text{ and }\ \mathrm{tr}\{\Lambda^{\mathrm{opt}}_{n,P}(\hat{\boldsymbol{\theta}}_n)\}<\mathrm{tr}\{\Lambda^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)\}<\frac{\mathrm{tr}\{\Lambda^{\mathrm{opt}}_{n,P}(\hat{\boldsymbol{\theta}}_n)\}}{1-\alpha}.$$

Thus, if α is small enough, $\mathrm{tr}\{\Lambda^{\alpha}_R(\hat{\boldsymbol{\theta}}_n)\}$ and $\mathrm{tr}\{\Lambda^{\mathrm{opt}}_R(\hat{\boldsymbol{\theta}}_n)\}$ can be arbitrarily close, and $\mathrm{tr}\{\Lambda^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)\}$ and $\mathrm{tr}\{\Lambda^{\mathrm{opt}}_{n,P}(\hat{\boldsymbol{\theta}}_n)\}$ can be arbitrarily close.

Remark 12. If the pilot subsample size is much smaller than the approximated optimal subsample size, i.e., $s_0=o(s_n)$, then the aggragated estimator $\check{\boldsymbol{\theta}}_R$ and $\check{\boldsymbol{\theta}}_P$ have the same asymptotic distributions as those for $\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}$ and $\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,P}$, respectively.

4 Numerical experiments

In this section, we use numerical examples to compare the optimal subsampling probabilities under the two sampling procedures considered in this paper. We will also use numerical experiments to evaluate the performance of the practical algorithms proposed in Section 3.1.

4.1 Comparisons of optimal subsampling probabilities

In this section, we use numerical examples to compare the optimal probabilities for subsampling with replacement presented in Theorem 3 with the optimal Poisson subsampling probabilities presented in Theorem 4.

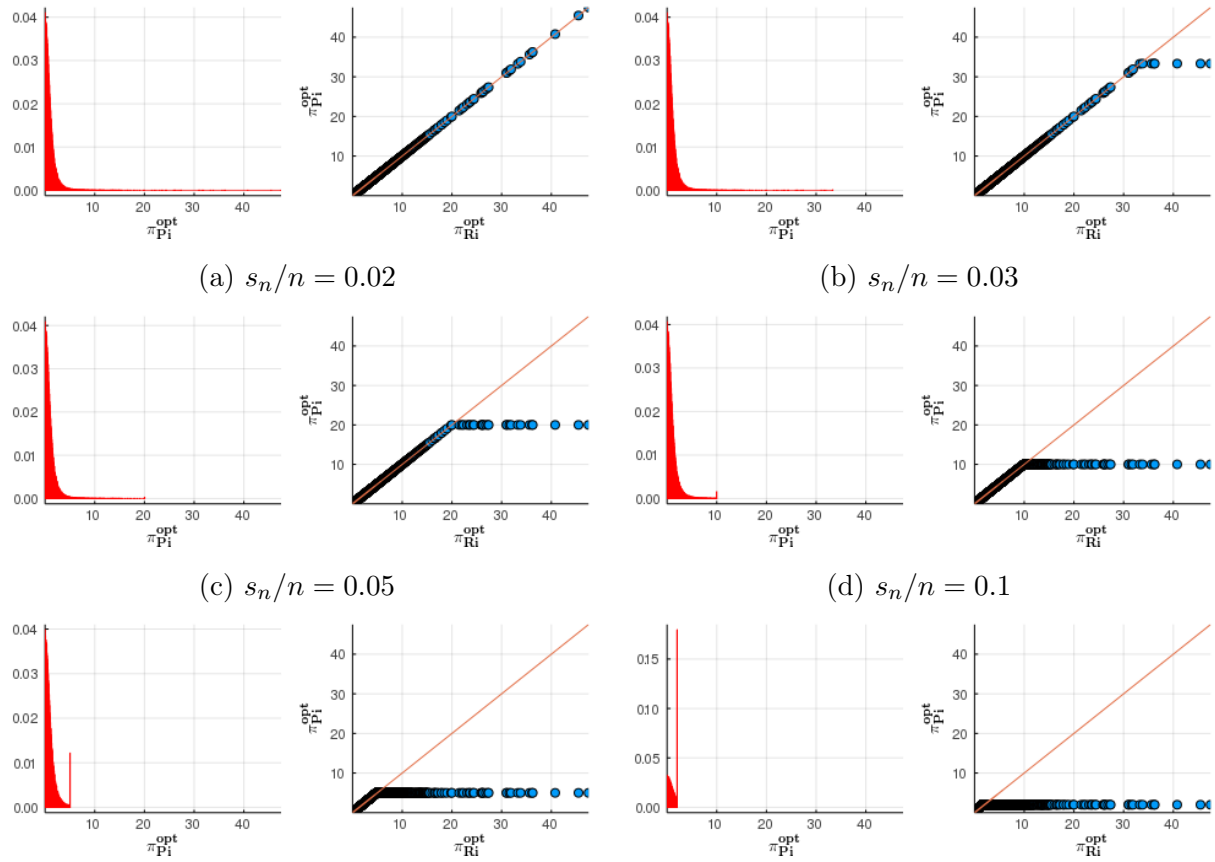
Example 2 (Linear regression). Consider solving the OLS for a linear regression model $y_i = \theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_1 + \varepsilon_i$, $i = 1, ..., n$, with $n = 10^5$, $\theta_0 = 1$, $\boldsymbol{\theta}_1$ being a 50 dimensional vector of ones, and ε_i being i.i.d. $\mathbb{N}(0, 1)$. For the expected subsample sizes, we consider $s_n = 2 \times 10^3, 3 \times 10^3, 5 \times 10^3, 10^4, 2 \times 10^4$, and 5×10^4, so that the subsampling ratios are $s_n/n = 0.02, 0.03, 0.05, 0.1, 0.2$, and 0.5. In this example, we use the L-optimality criterion with $L = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{1/2}$ so that the optimal subsampling probabilities are closely related to the statistical leverage scores. Specially, $\pi_{n,Ri}^{\mathrm{opt}} \propto |\hat{\varepsilon}_i|\sqrt{h_i}$ and $\pi_{n,Pi}^{\mathrm{opt}} \propto (|\hat{\varepsilon}_i|\sqrt{h_i}) \wedge H$ for the two subsampling procedures, respectively. To generate $\boldsymbol{x}_i$'s, we used normal distribution $\mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and multivariate t distributions $t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$ with degrees of freedom $v = 5, 4, 3, 2$, and 1, where $\boldsymbol{\Sigma}$ is a matrix with the (i, j)-th element being $0.5^{I(i \neq j)}$ and $I()$ being the indicator function. For this sequence of covariate distributions, the statistical leverage scores become more and more nonuniform.

Table 1 gives the values of g in the expression of the optimal Poisson subsampling probabilities in Theorem 4 for different combinations of the subsampling ratio s_n/n and covariate distribution. Note that g is the number of cases that $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \propto |\hat{\varepsilon}_i|\sqrt{h_i}$ are truncated by H. Thus, $\pi_{n,Ri}^{\mathrm{opt}}$ and $\pi_{n,Pi}^{\mathrm{opt}}$ are more different for larger values of g, and they are identical if $g = 0$. It is clear that g increases as s_n/n increases, indicating that $\pi_{n,Ri}^{\mathrm{opt}}$ and $\pi_{n,Pi}^{\mathrm{opt}}$ are more different as the subsampling ratio s_n/n gets larger. We also see that as the tail of the covariate distribution get heavier, g gets larger. This tells us that the difference between $\pi_{n,Ri}^{\mathrm{opt}}$ and $\pi_{n,Pi}^{\mathrm{opt}}$ is more significant if the statistical leverage scores are more nonuniform, as a heavier-tailed covariate distribution leads to more nonuniform leverage scores.

Table 1: The values of g in the optimal Poisson subsampling probabilities for OLS with different expected subsample sizes s_n and different distributions of $\boldsymbol{x}_i$'s. The full data sample size is $n = 10^5$.

	Distribution of $\boldsymbol{x}_i$'s					
s_n/n	Normal	t_5	t_4	t_3	t_2	t_1
0.02	0	0	0	0	16	120
0.03	0	0	0	7	39	203
0.05	0	0	1	28	113	342
0.1	0	23	58	154	492	756
0.2	15	584	762	1216	2242	1734
0.5	14364	16569	17191	17954	19038	15481

Figure 1 presents histograms and scatter plots of optimal probabilities for the two subsampling procedures to show more details on the distributions of $\pi_{n,Ri}^{\text{opt}}$'s and $\pi_{n,Pi}^{\text{opt}}$'s when $\boldsymbol{x}_i$'s are from the t_3 distribution. In each sub-figure, the left panel is the histogram for $\pi_{n,Pi}^{\text{opt}}$'s and the right panel is the scatter plot of $\pi_{n,Pi}^{\text{opt}}$'s against $\pi_{n,Ri}^{\text{opt}}$'s. We multiply all probabilities by n for better presentations. Note that this does not change the shapes of the figures. We only create the histogram for $\pi_{n,Pi}^{\text{opt}}$'s, because the distribution of $\pi_{n,Ri}^{\text{opt}}$'s does not depend on s_n and remains the same for all values of s_n. In addition, since $g = 0$ for the case with $s_n/n = 0.02$, the histogram in Figure 1(a) is the same to the histogram for $\pi_{n,Ri}^{\text{opt}}$ and we can compare it with other histograms to see the difference between the distributions of $\pi_{n,Ri}^{\text{opt}}$'s and $\pi_{n,Pi}^{\text{opt}}$'s. From Figure 1 (a)-(f), we see that as s_n/n increases the optimal probabilities for Poisson sampling and sampling with replacement are more different, because more larger $\pi_{n,Pi}^{\text{opt}}$'s are truncated to $1/s$.



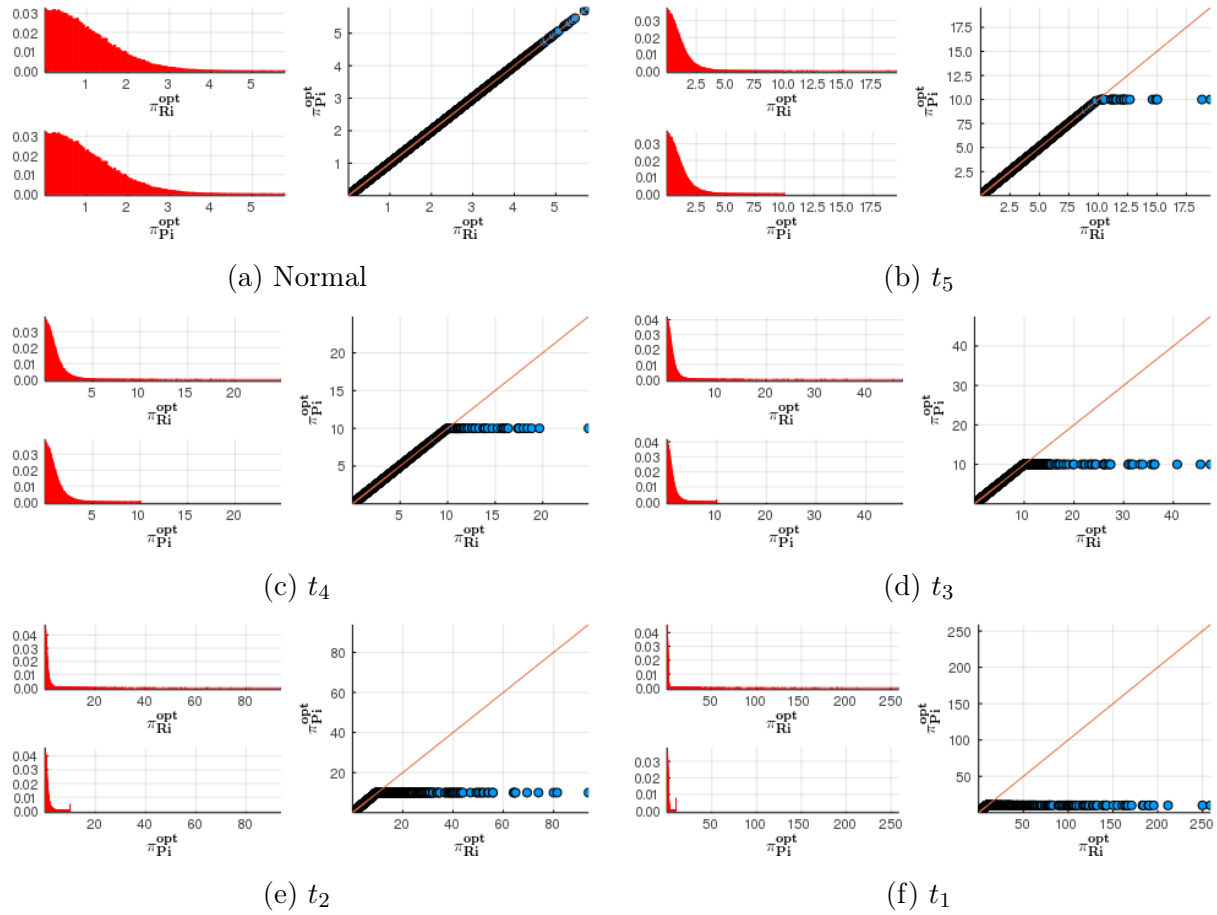
(a) $s_n/n = 0.02$ (b) $s_n/n = 0.03$

(c) $s_n/n = 0.05$ (d) $s_n/n = 0.1$

(e) $s_n/n = 0.2$ (f) $s_n/n = 0.5$

Figure 1: Histograms and scatter plots of optimal probabilities for subsampling with replacement and Poisson subsampling for different subsampling ratio s_n/n. Here $\boldsymbol{x}_i$'s are from the t_3 distribution.

Figure 2 presents histograms and scatter plots of optimal probabilities $\pi_{n,Ri}^{\text{opt}}$'s and $\pi_{n,Pi}^{\text{opt}}$'s for different distributions of $\boldsymbol{x}_i$'s when $s_n/n = 0.1$. In each sub-figure, the upper and lower plots in the left panel are the histograms for $\pi_{n,Ri}^{\text{opt}}$'s and $\pi_{n,Pi}^{\text{opt}}$'s, respectively, and the right panel is the scatter plot of $\pi_{n,Pi}^{\text{opt}}$'s against $\pi_{n,Ri}^{\text{opt}}$'s. Again, we multiply all probabilities by n for better presentations. We see that for a fixed subsampling ratio s_n/n, $\pi_{n,Ri}^{\text{opt}}$'s and $\pi_{n,Pi}^{\text{opt}}$'s become more different as the leverage scores become more nonuniform (the tail of the covariate distribution becomes heavier), because more large values of $\pi_{n,Pi}^{\text{opt}}$'s are truncated.



(a) Normal (b) t_5

(c) t_4 (d) t_3

(e) t_2 (f) t_1

Figure 2: Histograms and scatter plots of optimal probabilities for subsampling with replacement and Poisson subsampling for different covariate distributions. Here the subsampling ratio $s_n/n = 0.1$.

4.2 Comparison of estimation efficiency of the practical algorithms

We compare the estimation efficiency for the two subsampling procedures using both synthetic and real data sets.

Example 3 (Logistic regression). Form model $\mathbb{P}(y_i = 1|\boldsymbol{x}_i) = e^{\theta_0+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_1}/(1+e^{\theta_0+\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_1})$, $i = 1, ..., n$, we generate synthetic data sets by setting $n = 10^5$, $\theta_0 = 0.5$, and $\boldsymbol{\theta}_1$ to be a 9 dimensional vector of 0.5. We consider the following three cases to generate $\boldsymbol{x}_i$. In Cases 1 and 3, the responses y_i are balanced, while in Case 2 about 98% of the data points are with $y_i = 1$.

Case 1: **Normal**. Generate $\boldsymbol{x}_i$ from a multivariate normal distribution, $\mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where the (i, j)-th element of $\boldsymbol{\Sigma}$ is $\Sigma_{ij} = 0.5^{I(i\neq j)}$ and $I()$ is the indicator function. This distribution is symmetric with light tails.

Case 2: **LogNormal**. Generate $\boldsymbol{v}_i$ from $\mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as defined in Case 1 and then set $\boldsymbol{x}_i = e^{\boldsymbol{v}_i}$, where the exponentiation is element-wise. This distribution is asymmetric and positively skewed.

Case 3: $\boldsymbol{T}_3$. We generate $\boldsymbol{x}_i$ from a multivariate t distribution with three degrees of freedom $t_3(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ defined in Case 1. This distribution is symmetric with heavy tails.

We also consider two real data sets: the covtype data from the LIBSVM data website (`https://www.csie.ntu.edu.tw/~cjlin/libsvm/`) and the SUSY data (Baldi et al. 2014). Both data sets are also available from the UCI data repository (Dheeru & Karra Taniskidou 2017). We present them as Cases 4 and 5 below.

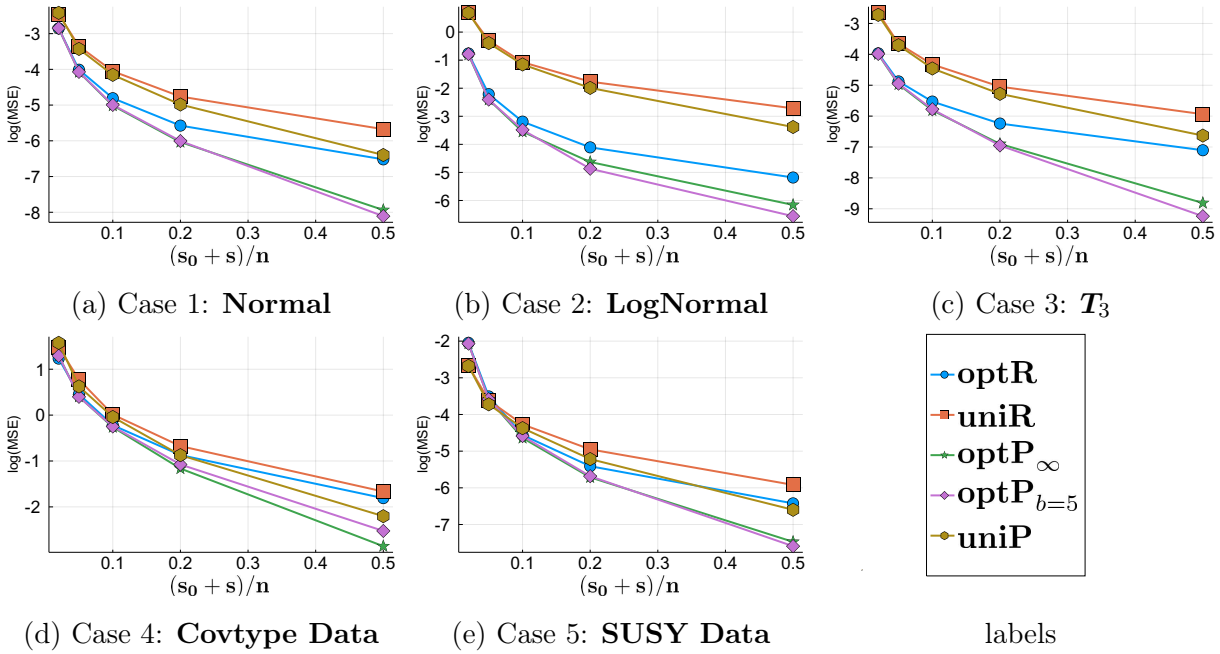
Case 4: **Covtype Data**. It has $n = 581,012$ observations with about 48.76% of the responses are $y_i = 1$. We use the ten quantitative covariate variables as $\boldsymbol{x}_i$'s.

Case 5: **SUSY Data**. It has $n = 5,000,000$ observations with about 54.24% of the responses are $y_i = 1$. We use the 18 kinematic features to classify whether new SUSY particles are produced.

To implement Algorithms 2 and 3, we set $\alpha = 0.1$, and choose $s_0 = 0.01n$ and different values for s_n so that the sampling ratio $(s_0 + s_n)/n = 0.02$, 0.05, 0.1, 0.2, and 0.5. Two different options of H^{0*} are considered: $H^{0*} = \|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\|_{\frac{s_n}{5n}}$ and $H^{0*} = \infty$. We aggregate the pilot estimator with the approximated optimal subsampling estimator using the procedure described in Remark 10. For comparison, we also implement the uniform subsampling method with expected subsample sizes $s_0 + s_n$. Newton's method is used for optimization on all subsamples. We repeat the simulation for $T = 1000$ times to calculate the empirical mean squared error (MSE), defined as MSE$=\frac{1}{T}\sum_{t=1}^{T}\|\check{\boldsymbol{\theta}}^{(t)} - \hat{\boldsymbol{\theta}}_n\|^2$, where $\check{\boldsymbol{\theta}}^{(t)}$ is the subsampling estimate at the t-th repetition and $\hat{\boldsymbol{\theta}}_n$ is the full data estimate.

Figure 3 plots the empirical MSE (natural logarithm is taken for better presentation) against the subsampling ratio $(s_0 + s_n)/n$. When the subsampling ratio $(s_0 + s_n)/n$ is close to zero, subsampling with replacement and Poisson subsampling have similar performance for both approximated optimal subsampling and uniform subsampling. However, when $(s_0 + s_n)/n$ gets larger, Poisson subsampling outperforms subsampling with replacement, and the improvement from subsampling with replacement to Poisson subsampling is more significant for approximated optimal subsampling than for uniform subsampling. For both subsampling with replacement and Poisson subsampling, approximated optimal subsampling methods outperform the uniform subsampling method. Their performances are closer for smaller $(s_0+s_n)/n$ because the proportions of uniform subsamples are higher for smaller $(s_0+s_n)/n$. For Poisson subsampling, the results for the two choices of H^{0*}, $H^{0*} = \infty$ and $H^{0*} =$

$\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\|_{\frac{s_n}{5n}}$, are similar when $(s_0 + s_n)/n$ is small, but they start to differ for larger $(s_0 + s_n)/n$.



(a) Case 1: **Normal** (b) Case 2: **LogNormal** (c) Case 3: $\boldsymbol{T}_3$

(d) Case 4: **Covtype Data** (e) Case 5: **SUSY Data**

Figure 3: Log empirical MSEs (y-axis) against subsampling ratio $(s_0 + s_n)/n$ (x-axis) for logistic regression. Here, "optR" means optimal subsampling with replacement; "uniR" means uniform subsampling with replacement; "optP$_\infty$" means approximated optimal Poisson subsampling with $H^{0*} = \infty$; "optP$_{b=5}$" means approximated optimal Poisson subsampling with $H^{0*} = \|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\|_{\frac{s_n}{5n}}$; and"uniP" means uniform Poisson subsampling.

Example 4 (Linear regression). We consider a linear model $y_i = \theta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}_1 + \varepsilon_i$, $i = 1, ..., n$, with $n = 10^5$, $\theta_0 = 1$, $\boldsymbol{\theta}_1$ being a 50 dimensional vector of ones, and ε_i being i.i.d. $\mathbb{N}(0, 1)$. We use the same distributions in Cases 1-3 to generate $\boldsymbol{x}_i$ and refer them as Cases 1'-3'. We also consider a gas sensor data Fonollosa et al. (2015) from the UCI data repository (Dheeru & Karra Taniskidou 2017). We present it as Case 6 below.

Case 6: **Gas Sensor Data**. After cleaning, the data contain $n = 4,188,261$ readings on 15 sensors. We use log of readings from the last sensor as responses and log of other readings as covariates.

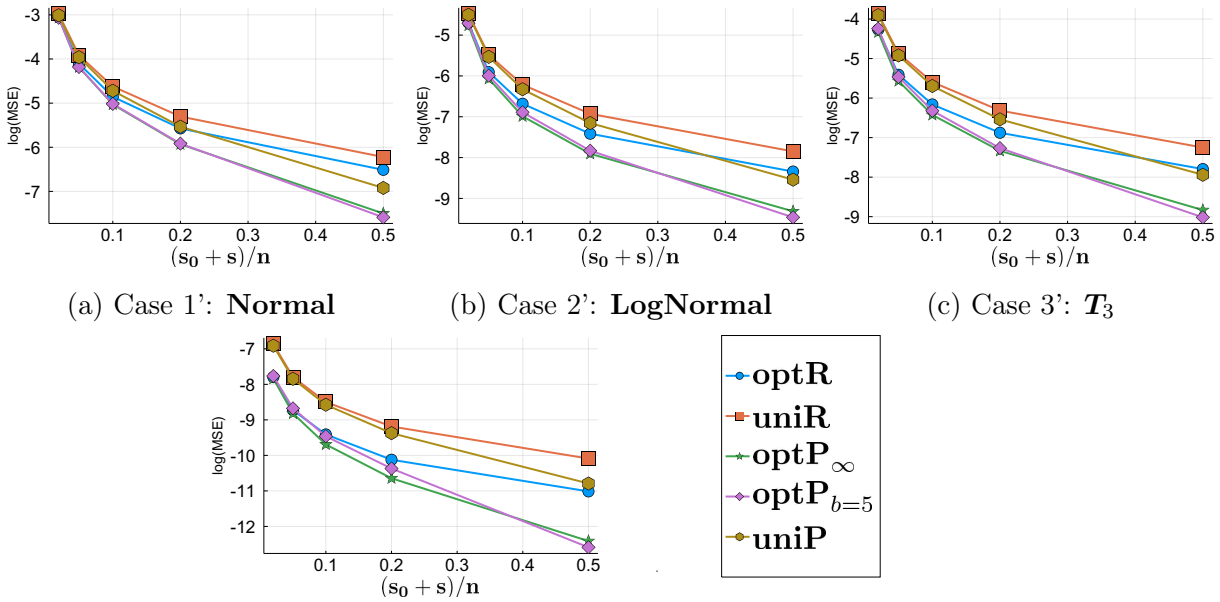
To implement Algorithms 2 and 3, we use the same setup for α, s_0, s_n, and H^{0*}, as used in logistic regression. Specifically, $\alpha = 0.1$, $s_0 = 0.01n$ and different values for s_n so that $(s_0 + s_n)/n = 0.02$, 0.05, 0.1, 0.2, and 0.5. We also consider both $H^{0*} = \|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\|_{\frac{s_n}{5n}}$ and $H^{0*} = \infty$, and aggregate the pilot estimator with the approximated optimal subsampling

estimator using the procedure described in Remark 10. We repeat the simulation for $T = 1000$ times to calculate the empirical MSE.

Figure 4 gives results for empirical MSE from least-squares in linear regression model. The overall pattern in Figure 4 is similar to that in Figure 3. We see that subsampling with replacement and Poisson subsampling have similar performance if the subsampling ratio $(s_0 + s_n)/n$ is close to zero, while Poisson subsampling outperforms subsampling with replacement as $(s_0 + s_n)/n$ gets larger. This trend is true for both approximated optimal subsampling and uniform subsampling, and we observe that the advantage of Poisson subsmapling over subsampling with replacement is more significant for approximated optimal subsampling. Furthermore, for linear regression, the advantage of Poisson subsampling compared with subsampling with replacement is more significant. For example, in Case 4', the synthetic data sets with $\boldsymbol{x}_i$'s from the t_3 distribution, the uniform Poisson subsampling can even outperform the approximated optimal subsampling with replacement when $(s_0+s_n)/n = 0.5$. We also observe that approximated optimal subsampling methods outperform the uniform subsampling methods, and the gap between their performance in terms of estimation efficiency is larger for larger $(s_0+s_n)/n$. This is because the proportions of more informative observations in the subsample are higher for larger $(s_0+s_n)/n$. Another pattern is that when the approximated optimal subsampling probabilities are more nonuniform, their advantage over uniform subsampling is more significant. For example, from the gas sensor data set, approximated optimal subsampling methods have significantly higher estimation efficiency than the uniform subsampling methods even when $s_0 = s_n = 1000$. For Poisson subsampling, the performance with $H^{0*} = \infty$ and that with $H^{0*} = \|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|_{\frac{s_n}{5n}}$ are similar for small $(s_0 + s_n)/n$, but the choice with $H^{0*} = \|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|_{\frac{s_n}{5n}}$ starts to show its advantage for larger $(s_0 + s_n)/n$.

5 Conclusion and Discussion

In this paper, we derived optimal subsampling probabilities in the context of maximizing an additive target function for both subsampling with replacement and Poisson subsampling. Theoretical and empirical results show that the two different subsampling procedure have similar performance when the subsampling ratio is small. However, when subsampling ratio does not converge to zero, Poisson subsampling has a higher estimation efficiency. One problem warrants for further investigation is how to choose the tuning parameter b in Algorithm 3 so that the approximated optimal subsampling probabilities produce an estimator with an asymptotic variance-covariance matrix that is near optimal even when the subsampling ratio does not converge to zero.



(a) Case 1': **Normal** (b) Case 2': **LogNormal** (c) Case 3': $\boldsymbol{T}_3$

(d) Case 6: **Gas Sensor** labels

Figure 4: Log Empirical MSEs (y-axis) against subsampling ratio $(s_0 + s_n)/n$ (x-axis) for linear regression. Here, "optR" means optimal subsampling with replacement; "uniR" means uniform subsampling with replacement; "optP$_\infty$" means approximated optimal Poisson subsampling with $H^{0*} = \infty$; "optP$_{b=5}$" means approximated optimal Poisson subsampling with $H^{0*} = \|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}_{s_0,P}^{0*})\|_{\frac{s_n}{5n}}$; and"uniP" means uniform Poisson subsampling.

Acknowledgments

The authors are deeply grateful to Professor Michael Mahoney, the Associate Editor Professor Stephane Boucheron, and two anonymous reviewers for their insightful comments, question, and suggestions that significantly improved the manuscript. This work was partially supported by the NSF grant CCF-2105571.

An early short version of the paper is presented in the AISTATS 2021 conference (Wang & Zou 2021). This version contains substantially more technical results such as the relationships between conditional and unconditional convergences, and the unconditional asymptotic distributions about the true parameter. It also contains additional examples and numerical comparison results.

Appendix

A.1 Proofs

In this section, we prove all the theoretical results in the paper.

A.1.1 Proof of Proposition 2

Proof. For Proposition 2 (a) since $\mathbb{P}(\|\Delta_{n,s_n}\| > \delta|\mathcal{D}_n)$ is a nonnegative and bounded random variable, from Theorem 1.3.6 of Serfling (1980), $\mathbb{P}(\|\Delta_{n,s_n}\| > \delta|\mathcal{D}_n) = o_P(1)$ if and only if $\mathbb{E}\{\mathbb{P}(\|\Delta_{n,s_n}\| > \delta|\mathcal{D}_n)\} \to 0$. Note that

$$\mathbb{E}\{\mathbb{P}(\|\Delta_{n,s_n}\| > \delta|\mathcal{D}_n)\} = \mathbb{E}[\mathbb{E}\{I(\|\Delta_{n,s_n}\| > \delta)|\mathcal{D}_n\}] = \mathbb{E}\{I(\|\Delta_{n,s_n}\| > \delta)\} = \mathbb{P}(\|\Delta_{n,s_n}\| > \delta).$$

Thus $\mathbb{E}\{\mathbb{P}(\|\Delta_{n,s_n}\| > \delta|\mathcal{D}_n)\} \to 0$ if and only if $\mathbb{P}(\|\Delta_{n,s_n}\| > \delta) \to 0$, which is true if and only if $\Delta_{n,s_n} = o_P(1)$.

Now we prove Proposition 2 (b). Note that $\Delta_{n,s_n} = O_{P|\mathcal{D}_n}(1)$ means that for any $\epsilon > 0$ and any $\delta > 0$, there exist a finite $K_\epsilon > 0$ and a finite $N_{\epsilon,\delta} > 0$ such that $\mathbb{P}\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_\epsilon|\mathcal{D}_n) > \epsilon\} < \delta$ for all $s_n > N_{\epsilon,\delta}$ and $n > N_{\epsilon,\delta}$. Thus if $\Delta_{n,s_n} = O_{P|\mathcal{D}_n}(1)$, then for any $\epsilon > 0$ and $\delta = \epsilon$, there exist a finite $K_{0.5\epsilon} > 0$ and a finite $N_{0.5\epsilon,0.5\epsilon} > 0$ such that for all $s_n > N_{0.5\epsilon,0.5\epsilon}$ and $n > N_{0.5\epsilon,0.5\epsilon}$,

$$\mathbb{P}\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_{0.5\epsilon}|\mathcal{D}_n) > 0.5\epsilon\} < 0.5\epsilon.$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\|\Delta_{n,s_n}\| > K_{0.5\epsilon}) &= \mathbb{E}\big\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_{0.5\epsilon}|\mathcal{D}_n)\big\} \\
&\le \mathbb{E}\big[\mathbb{P}(\|\Delta_{n,s_n}\| > K_{0.5\epsilon}|\mathcal{D}_n) I\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_{0.5\epsilon}|\mathcal{D}_n) > 0.5\epsilon\}\big] + 0.5\epsilon \\
&\le \mathbb{E}\big[I\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_{0.5\epsilon}|\mathcal{D}_n) > 0.5\epsilon\}\big] + 0.5\epsilon \\
&\le \mathbb{P}\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_{0.5\epsilon}|\mathcal{D}_n) > 0.5\epsilon\} + 0.5\epsilon \\
&\le \epsilon,
\end{aligned}$$

meaning that $\Delta_{n,s_n} = O_P(1)$.

On the other hand, if $\Delta_{n,s_n} = O_P(1)$, then for any $\epsilon > 0$ and $\delta > 0$, there exist a finite $K_{\delta\epsilon}$ and a finite $N_{\delta\epsilon}$ such that $\mathbb{P}(\|\Delta_{n,s_n}\| > K_{\delta\epsilon}) \le \delta\epsilon$ for all $s_n > N_{\delta\epsilon}$ and $n > N_{\delta\epsilon}$. Thus

$$\mathbb{P}\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_{\delta\epsilon}|\mathcal{D}_n) > \epsilon\} \le \epsilon^{-1}\mathbb{E}\big\{\mathbb{P}(\|\Delta_{n,s_n}\| > K_{\delta\epsilon}|\mathcal{D}_n)\big\} = \epsilon^{-1}\mathbb{P}(\|\Delta_{n,s_n}\| > K_{\delta\epsilon}) < \delta,$$

which means that $\Delta_{n,s_n} = O_{P|\mathcal{D}_n}(1)$.

Now we prove Proposition 2 (c). Because $\mathbb{P}(\Delta_{n,s_n} \leq \boldsymbol{x}|\mathcal{D}_n)$ is bounded, $\mathbb{P}(\Delta_{n,s_n} \leq \boldsymbol{x}|\mathcal{D}_n) - \mathbb{P}(U \leq \boldsymbol{x}) = o_P(1)$ if and only if $\mathbb{E}\{\mathbb{P}(\Delta_{n,s_n} \leq \boldsymbol{x}|\mathcal{D}_n)\} - \mathbb{P}(U \leq \boldsymbol{x}) = \mathbb{P}(\Delta_{n,s_n} \leq \boldsymbol{x}) - \mathbb{P}(U \leq \boldsymbol{x}) = o(1)$. Thus $\Delta_{n,s_n} \overset{|\mathcal{D}_n}{\rightsquigarrow} U$ if and only if $\Delta_{n,s_n} \rightsquigarrow U$. □

A.1.2 Proof for Theorem 1

Recall that

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \boldsymbol{\theta}).$$

For the sampling with replacement estimator in (2), let

$$M^*_{s_n}(\boldsymbol{\theta}) = \frac{1}{s_n}\sum_{i=1}^{s_n} \frac{m(Z^*_i, \boldsymbol{\theta})}{n\pi^*_{n,i}}.$$

To prove Theorem 1, we first establish Lemma 1 and Lemma 2 in the following.

Lemma 1. *Under Assumptions 3 and 5, if* $\|\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n\| = o_P(1)$*, then conditional on* $\mathcal{D}_n$*,*

$$B_{s_n} - \ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = o_P(1),$$

where

$$\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} \ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n),$$
$$B_{s_n} = \int_0^1 \frac{1}{s_n}\sum_{i=1}^{s_n} \frac{\ddot{m}\{Z^*_i, \hat{\boldsymbol{\theta}}_n + \lambda(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n)\}}{n\pi^*_{n,i}}\,\mathrm{d}\lambda.$$

In Lemma 1, the notation $o_P(1)$ means convergence to 0 in probability. Here the probability is conditional probability. From Xiong & Li (2008), Cheng & Huang (2010), a sequence converges to 0 in conditional probability is equivalent to the fact that it converges to 0 in unconditional probability. Thus we use $o_P(1)$ to indicate convergence to 0 either in unconditional or conditional probability.

Proof. Firstly, note that

$$\mathbb{E}\left(\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\psi(Z^*_i)}{n\pi^*_{n,i}}\Big|\mathcal{D}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\psi(Z_i) = \mathbb{E}\{\psi(Z)\} + o_P(1), \quad \text{and}$$
$$\mathbb{V}\left(\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\psi(Z^*_i)}{n\pi^*_{n,i}}\Big|\mathcal{D}_n\right) = \frac{1}{s_n}\sum_{i=1}^{n}\frac{\psi^2(Z_i)}{n^2\pi_{n,i}} \leq \max_{i=1,\ldots,n}\left(\frac{1}{n\pi_{n,i}}\right)\frac{1}{s_n n}\sum_{i=1}^{n}\psi^2(Z_i) = O_P(s_n^{-1}).$$

Thus,

$$\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\psi(Z_i^*)}{n\pi_{n,i}^*} = O_{P|\mathcal{D}_n}(1).$$

For every $k, l = 1, 2, ..., d$, from Lipschitz continuity, for $\lambda \in (0, 1)$, we have

$$\begin{aligned}
&\left|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}\{Z_i^*, \hat{\boldsymbol{\theta}}_n + \lambda(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n)\}}{n\pi_{n,i}^*} - \frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*, \hat{\boldsymbol{\theta}}_n)}{n\pi_{n,i}^*}\right| \\
&= \lambda\|\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n\|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\psi(Z_i^*)}{n\pi_{n,i}^*} = o_P(1), \qquad \text{(A.1)}
\end{aligned}$$

and for any fixed $\boldsymbol{\theta}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}^2(Z_i, \hat{\boldsymbol{\theta}}_n) \le \frac{2}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}^2(Z_i, \boldsymbol{\theta}) + \frac{2\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|^2}{n}\sum_{i=1}^{n}\psi^2(Z_i) = O_P(1). \qquad \text{(A.2)}$$

In addition, according to (A.2),

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*, \hat{\boldsymbol{\theta}}_n)}{n\pi_{n,i}^*}\middle|\mathcal{D}_n\right\} &= \frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n), \\
\mathbb{V}\left\{\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*, \hat{\boldsymbol{\theta}}_n)}{n\pi_{n,i}^*}\middle|\mathcal{D}_n\right\} &\le \frac{1}{s_n}\sum_{i=1}^{n}\frac{\ddot{m}_{k,l}^2(Z_i, \hat{\boldsymbol{\theta}}_n)}{n^2\pi_{n,i}} \\
&\le \max_i\left(\frac{1}{n\pi_{n,i}}\right)\frac{1}{s_n n}\sum_{i=1}^{n}\ddot{m}_{k,l}^2(Z_i, \hat{\boldsymbol{\theta}}_n) = O_P(s_n^{-1}).
\end{aligned}$$

Thus, by Chebyshev's inequality, we have

$$\left|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*, \hat{\boldsymbol{\theta}}_n)}{n\pi_{n,i}^*} - \frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)\right| = O_{P|\mathcal{D}_n}(s_n^{-1/2}) = o_{P|\mathcal{D}_n}(1). \qquad \text{(A.3)}$$

Combining (A.1) and (A.3), we have

$$\begin{aligned}
&\left|B_{s_n,k,l} - \frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)\right| \\
&\le \int_0^1\left|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}\{Z_i^*, \hat{\boldsymbol{\theta}}_n + \lambda(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n)\}}{n\pi_{n,i}^*} - \frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)\right|\mathrm{d}\lambda \\
&\le \int_0^1\left[\left|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}\{Z_i^*, \hat{\boldsymbol{\theta}}_n + \lambda(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n)\}}{n\pi_{n,i}^*} - \frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*, \hat{\boldsymbol{\theta}}_n)}{n\pi_{n,i}^*}\right|\right]\mathrm{d}\lambda \\
&\quad + \left|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*, \hat{\boldsymbol{\theta}}_n)}{n\pi_{n,i}^*} - \frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)\right| \\
&= o_{P|\mathcal{D}_n}(1).
\end{aligned}$$

□

Lemma 2. *Under Assumptions 4-5, given $\mathcal{D}_n$ in probability,*

$$\sqrt{s_n}\{\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}\dot{M}^*_{s_n}(\hat{\boldsymbol{\theta}}_n) \overset{|\mathcal{D}_n}{\rightsquigarrow} \mathbb{N}(\mathbf{0}, \boldsymbol{I}), \tag{A.4}$$

in conditional distribution.

Proof. Note that

$$\sqrt{s_n}\dot{M}^*_{s_n}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{\sqrt{s_n}}\sum_{i=1}^{s_n}\frac{\dot{m}(Z_i^*, \hat{\boldsymbol{\theta}}_n)}{n\pi_i^*} \equiv \frac{1}{\sqrt{s_n}}\sum_{i=1}^{s_n}\boldsymbol{\eta}_i \tag{A.5}$$

Given $\mathcal{D}_n, \boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_{s_n}$ are i.i.d, with

$$\mathbb{E}(\boldsymbol{\eta}|\mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n}\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = \mathbf{0}, \text{ and} \tag{A.6}$$

$$\begin{aligned}\mathbb{V}(\boldsymbol{\eta}_i|\mathcal{D}_n) = \Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n) &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}_n)}{\pi_{n,i}} \\ &\leq \max_{i=1,...,n}\left(\frac{1}{n\pi_{n,i}}\right)\frac{1}{n}\sum_{i=1}^{n}\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}_n) = O_P(1),\end{aligned} \tag{A.7}$$

where the inequality in (A.7) is in the Loewner ordering, i.e., $\boldsymbol{A}_1 \leq \boldsymbol{A}_2$ means $\boldsymbol{A}_1 - \boldsymbol{A}_2$ is a negative semi-definite matrix.

Meanwhile, for every $\varepsilon > 0$ and some $\delta \in (0, 2]$,

$$\begin{aligned}&\frac{1}{s_n}\sum_{i=1}^{s_n}\mathbb{E}\left\{\|\boldsymbol{\eta}_i\|^2 I(\|\boldsymbol{\eta}_i\| > s_n^{1/2}\varepsilon)|\mathcal{D}_n\right\} \leq \frac{1}{s_n^{1+\delta/2}\varepsilon^\delta}\sum_{i=1}^{s_n}\mathbb{E}\left\{\|\boldsymbol{\eta}_i\|^{2+\delta} I(\|\boldsymbol{\eta}_i\| > s_n^{1/2}\varepsilon)|\mathcal{D}_n\right\} \\ &\leq \frac{1}{s_n^{1+\delta/2}\varepsilon^\delta}\sum_{i=1}^{s_n}\mathbb{E}\left(\|\boldsymbol{\eta}_i\|^{2+\delta}|\mathcal{D}_n\right) \leq \frac{1}{s_n^{\delta/2}n^{2+\delta}\varepsilon^\delta}\sum_{i=1}^{n}\frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^{2+\delta}}{\pi_{n,i}^{1+\delta}} \\ &= \max_{i=1,...,n}\left(\frac{1}{n\pi_{n,i}}\right)^{1+\delta}\frac{1}{ns_n^{\delta/2}\varepsilon^\delta}\sum_{i=1}^{n}\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^{2+\delta} = O_P(s_n^{-\delta/2}).\end{aligned}$$

This shows that Lindeberg's condition is satisfied in probability. From (A.5), (A.6) and (A.7), by the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart (1998)), conditionally on $\mathcal{D}_n$, (A.4) follows. □

Proof of Theorem 1. Based on Lemma 1 and Lemma 2, now we are ready to prove Theorem 1. By direct calculation, we have that for any $\boldsymbol{\theta}$,

$$\mathbb{E}\left(M^*_{s_n}(\boldsymbol{\theta})|\mathcal{D}_n\right) = M_n(\boldsymbol{\theta}).$$

By Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\left\{\left|M^*_{s_n}(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta})\right| \geq \varepsilon|\mathcal{D}_n\right\} \leq \frac{\mathbb{V}\{M^*_{s_n}(\boldsymbol{\theta})|\mathcal{D}_n\}}{\varepsilon^2} = \frac{1}{\varepsilon^2 s_n n^2}\sum_{i=1}^{n}\frac{m^2(Z_i, \boldsymbol{\theta})}{\pi_{n,i}}$$

$$\leq \frac{1}{\varepsilon^2 s_n} \max_{i=1,\ldots,n} \left(\frac{1}{n\pi_{n,i}} \right) \frac{1}{n} \sum_{i=1}^{n} m^2(Z_i, \boldsymbol{\theta}) = O_P\left(s_n^{-1}\right).$$

Thus, for every $\boldsymbol{\theta}$,

$$M^*_{s_n}(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}) = o_{P|\mathcal{D}_n}(1). \tag{A.8}$$

Note that under Assumptions 1, 2, the parameter space is compact and $\hat{\boldsymbol{\theta}}_n$ is the unique global maximum of the continuous concave function $M_n(\boldsymbol{\theta})$. Thus from Theorem 5.9 and its remark of van der Vaart (1998), conditionally on $\mathcal{D}_n$,

$$\|\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n\| = o_{P|\mathcal{D}_n}(1) = o_P(1). \tag{A.9}$$

The consistency ensures that $\tilde{\boldsymbol{\theta}}_{s_n,R}$ is close to $\hat{\boldsymbol{\theta}}_n$ as long as s_n is large. By Taylor expansion,

$$0 = \dot{M}^*_{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,R}) = \dot{M}^*_{s_n}(\hat{\boldsymbol{\theta}}_n) + B_{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n), \tag{A.10}$$

where

$$B_{s_n} = \int_0^1 \frac{1}{s_n} \sum_{i=1}^{s_n} \frac{\ddot{m}\{Z^*_i, \hat{\boldsymbol{\theta}}_n + \lambda(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n)\}}{n\pi^*_{n,i}} \, \mathrm{d}\lambda.$$

From (A.10) and Lemma 1,

$$0 = \dot{M}^*_{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,R}) = \dot{M}^*_{s_n}(\hat{\boldsymbol{\theta}}_n) + \{\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) + o_P(1)\}(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n), \tag{A.11}$$

which shows that

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n &= -\{\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) + o_P(1)\}^{-1} \dot{M}^*_{s_n}(\hat{\boldsymbol{\theta}}_n) \\ &= -\frac{1}{\sqrt{s_n}} \{\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) + o_P(1)\}^{-1} \{\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)\}^{1/2} \sqrt{s_n} \{\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2} \dot{M}^*_{s_n}(\hat{\boldsymbol{\theta}}_n). \end{aligned} \tag{A.12}$$

By Lemma 2 and Slutsky's theorem, we obtain that, given full data $\mathcal{D}_n$ in probability,

$$\sqrt{s_n}\{V_{n,R}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}(\tilde{\boldsymbol{\theta}}_{s_n,R} - \hat{\boldsymbol{\theta}}_n) \to \mathbb{N}(\mathbf{0}, \boldsymbol{I}), \tag{A.13}$$

in conditional distribution, and this finishes the proof. □

A.1.3 Proof for Theorem 2

Let $\nu_i = 1$ if the i-th data point is selected in the subsample and $\nu_i = 0$ otherwise. The estimator in (2) is the same as the maximizer of

$$M^*_P(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{s^*_n} \frac{m(Z^*_i, \boldsymbol{\theta})}{s_n \pi^*_{n,i}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\nu_i m(Z_i, \boldsymbol{\theta})}{s_n \pi_{n,i}}.$$

Here, we use s_n to replace s^*_n in (2) for convenience, and the resulting estimator is identical to $\tilde{\boldsymbol{\theta}}_{s_n,P}$.

To prove Theorem 2, we first establish the following Lemmas 3 and 4.

Lemma 3. *If Assumptions 4-5 hold, then, given $\mathcal{D}_n$,*

$$\sqrt{s_n}\{\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}\dot{M}_P^*(\hat{\boldsymbol{\theta}}_n) \overset{|\mathcal{D}_n}{\rightsquigarrow} \mathbb{N}(\mathbf{0}, \boldsymbol{I}),$$

in conditional distribution, where

$$\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{(1-s_n\pi_{n,i})\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{\pi_{n,i}}.$$

Proof. Write

$$\sqrt{s_n}\dot{M}_P^*(\hat{\boldsymbol{\theta}}_n) = \sum_{i=1}^{n}\frac{\nu_i\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)}{n\sqrt{s_n}\pi_{n,i}} \equiv \sum_{i=1}^{n}\boldsymbol{\eta}_{Pi}.$$

By direct calculation and according to the definition of $\hat{\boldsymbol{\theta}}_n$,

$$\mathbb{E}\left(\sum_{i=1}^{n}\boldsymbol{\eta}_{Pi}\Big|\mathcal{D}_n\right) = \sqrt{s_n}\sum_{i=1}^{n}\frac{\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)}{n} = \mathbf{0},$$

and

$$\begin{aligned}
\mathbb{V}\left(\sum_{i=1}^{n}\boldsymbol{\eta}_{Pi}\Big|\mathcal{D}_n\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\mathbb{V}(\nu_i|\mathcal{D}_n)\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{r\pi_i^2} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\frac{(1-s_n\pi_{n,i})\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{\pi_{n,i}} = \Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n) \\
&\leq \Big(\max_i\frac{1}{n\pi_{n,i}}\Big)\frac{1}{n}\sum_{i=1}^{n}\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n) = O_P(1).
\end{aligned}$$

Next, we check Lindeberg's condition in conditional distribution. Note that for $\rho \in (0,2]$ and any $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}\left\{\|\boldsymbol{\eta}_{Pi}\|I(\|\boldsymbol{\eta}_{Pi}\| > \varepsilon)\Big|\mathcal{D}_n\right\} \leq \frac{1}{\varepsilon^\rho}\sum_{i=1}^{n}\mathbb{E}\left\{\|\boldsymbol{\eta}_{Pi}\|^{2+\rho}I(\|\boldsymbol{\eta}_{Pi}\| > \varepsilon)\Big|\mathcal{D}_n\right\} \\
\leq\ &\frac{1}{\varepsilon^\rho}\sum_{i=1}^{n}\mathbb{E}\left(\|\boldsymbol{\eta}_{Pi}\|^{2+\rho}\Big|\mathcal{D}_n\right) = \frac{1}{\varepsilon^\rho}\mathbb{E}\left\{\sum_{i=1}^{n}\frac{\nu_i^{2+\rho}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^{2+\rho}}{n^{2+\rho}s_n^{1+\rho/2}\pi_{n,i}^{2+\rho}}\Big|\mathcal{D}_n\right\} \\
=\ &\frac{1}{\varepsilon^\rho}\sum_{i=1}^{n}\frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^{2+\rho}}{n^{2+\rho}s_n^{\rho/2}\pi_{n,i}^{1+\rho}} \\
\leq\ &\max_i\left(\frac{1}{n\pi_{n,i}}\right)^{1+\rho}\frac{1}{s_n^{\rho/2}\varepsilon^\rho n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^{2+\rho} = O_P(s_n^{-\rho/2}) = o_P(1).
\end{aligned}$$

According to the Lindeberg-Feller Central Limit Theorem (van der Vaart 1998, cf.), given $\mathcal{D}_n$,

$$\sqrt{s_n}\{\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}\dot{M}_P^*(\hat{\boldsymbol{\theta}}_n) \to \mathbb{N}(\mathbf{0}, \boldsymbol{I}),$$

in conditional distribution. □

Lemma 4. *Under Assumptions 3 and 5, for any $\boldsymbol{u}_{s_n} = o_P(1)$, conditional on $\mathcal{D}_n$,*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n + \boldsymbol{u}_{s_n})}{s_n \pi_{n,i}} - \frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = o_P(1).$$

Proof. First, note that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \psi(Z_i)}{s_n \pi_{n,i}} = O_{P|\mathcal{D}_n}(1), \tag{A.14}$$

by Chebyshev's inequality and the fact that

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \psi(Z_i)}{s_n \pi_{n,i}}\middle|\mathcal{D}_n\right) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\psi(Z_i)\mathbb{E}(\nu_i|\mathcal{D}_n)}{s_n \pi_{n,i}} = \frac{1}{n}\sum_{i=1}^{n}\psi(Z_i) = \mathbb{E}\{\psi(Z_i)\} + o_P(1), \\
\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \psi(Z_i)}{s_n \pi_{n,i}}\middle|\mathcal{D}_n\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi^2(Z_i)\mathbb{V}(\nu_i|\mathcal{D}_n)}{s_n^2 \pi_{n,i}^2} \leq \frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi^2(Z_i)\mathbb{E}(\nu_i^2)}{s_n^2 \pi_{n,i}^2} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi^2(Z_i)}{s_n \pi_{n,i}} \leq \frac{1}{s_n n}\sum_{i=1}^{n}\psi^2(Z_i)\max_i\frac{1}{n\pi_{n,i}} = O_P(s_n^{-1}).
\end{aligned}$$

Thus, for every $k, l = 1, 2, ..., d$, from Assumption 3, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n + \boldsymbol{u}_{s_n})}{s_n \pi_{n,i}} - \frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)}{s_n \pi_{n,i}}\right| \leq \frac{\|\boldsymbol{u}_{s_n}\|}{n}\sum_{i=1}^{n}\frac{\nu_i \psi(Z_i)}{s_n \pi_{n,i}} = o_P(1).$$

which shows that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n + \boldsymbol{u}_{s_n})}{s_n \pi_{n,i}} - \frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)}{s_n \pi_{n,i}} = o_P(1). \tag{A.15}$$

According to (A.2), for every $k, l = 1, 2, ..., d$

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)}{s_n \pi_{n,i}}\middle|\mathcal{D}_n\right) &= \frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n), \\
\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)}{s_n \pi_{n,i}}\middle|\mathcal{D}_n\right) &= \frac{1}{s_n n^2}\sum_{i=1}^{n}\frac{(1 - s_n \pi_{n,i})\ddot{m}_{k,l}^2(Z_i, \hat{\boldsymbol{\theta}}_n)}{\pi_{n,i}} \leq \frac{1}{s_n n^2}\sum_{i=1}^{n}\frac{\ddot{m}_{k,l}^2(Z_i, \hat{\boldsymbol{\theta}}_n)}{\pi_{n,i}} \\
&\leq \max_i\left(\frac{1}{n\pi_{n,i}}\right)\frac{1}{s_n n}\sum_{i=1}^{n}\ddot{m}_{k,l}^2(Z_i, \hat{\boldsymbol{\theta}}_n) = O_P(s_n^{-1}).
\end{aligned}$$

Thus, Chebyshev's inequality tells us that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)}{s_n \pi_{n,i}} - \frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = O_{P|\mathcal{D}_n}(s_n^{-1/2}) = o_P(1). \tag{A.16}$$

Therefore, combining (A.15) and (A.16), we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i\ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n+\boldsymbol{u}_{s_n})}{s_n\pi_{n,i}}-\frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)=o_P(1).$$

□

Proof of Theorem 2. Denote

$$\gamma_P(\boldsymbol{u})=s_nM_P^*(\hat{\boldsymbol{\theta}}_n+\boldsymbol{u}/\sqrt{s_n})-s_nM_P^*(\hat{\boldsymbol{\theta}}_n).$$

Under Assumption 2, $\sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,P}-\hat{\boldsymbol{\theta}}_n)$ is the unique maximizer of $\gamma_P(\boldsymbol{u})$ as $\tilde{\boldsymbol{\theta}}_{s_n,P}$ is the unique maximizer of $M_P^*(\boldsymbol{u})$. By Taylor's expansion,

$$\gamma_P(\boldsymbol{u})=\sqrt{s_n}\boldsymbol{u}^T\dot{M}_P^*(\hat{\boldsymbol{\theta}}_n)+\frac{1}{2}\boldsymbol{u}^T\ddot{M}_P^*(\hat{\boldsymbol{\theta}}_n+\acute{\boldsymbol{u}}/\sqrt{s_n})\boldsymbol{u}$$

where $\acute{\boldsymbol{u}}$ lies between $\mathbf{0}$ and $\boldsymbol{u}$. From Lemma 3 $\sqrt{s_n}\dot{M}_P^*(\hat{\boldsymbol{\theta}}_n)$ is stochastically bounded in conditional probability given $\mathcal{D}_n$. From Lemma 4, conditional on $\mathcal{D}_n$, $\ddot{M}_P^*(\hat{\boldsymbol{\theta}}_n+\acute{\boldsymbol{u}}/\sqrt{s_n})-\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)=o_P(1)$ and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ converges to a positive-definite matrix.

Thus from the Basic Corollary in page 2 of Hjort & Pollard (2011), the maximizer of $s_n\gamma_P(\boldsymbol{u})$, $\sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,P}-\hat{\boldsymbol{\theta}}_n)$, satisfies that

$$\sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,P}-\hat{\boldsymbol{\theta}}_n)=\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\sqrt{s_n}\dot{M}_P^*(\hat{\boldsymbol{\theta}}_n)+o_P(1),\tag{A.17}$$

which implies that

$$\sqrt{s_n}\{V_{n,P}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}(\tilde{\boldsymbol{\theta}}_{s_n,P}-\hat{\boldsymbol{\theta}}_n)\to\mathbb{N}(\mathbf{0},\boldsymbol{I}),\tag{A.18}$$

in conditional distribution, given $\mathcal{D}_n$ in probability. This finishes the proof. □

A.1.4 Proof of Theorem 1'

Proof of Theorem 1'. Letting $S_{n,R}=\sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,R}-\hat{\boldsymbol{\theta}}_n)$ and $Y_n=\sqrt{s_n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$, we have

$$\sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,R}-\boldsymbol{\theta}_0)=S_{n,R}+Y_n.$$

According to Theorem 1, we know that under Assumptions 1-5 the characteristic function of $S_{n,R}$ given $\mathcal{D}_n$ satisfies that

$$\mathbb{E}\big(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}S_{n,R}}\big|\mathcal{D}_n\big)=e^{-0.5\boldsymbol{t}^{\mathrm{T}}V_{n,R}(\hat{\boldsymbol{\theta}}_n)\boldsymbol{t}}+o_{P|\mathcal{D}_n}(1)=e^{-0.5\boldsymbol{t}^{\mathrm{T}}V_{n,R}(\hat{\boldsymbol{\theta}}_n)\boldsymbol{t}}+o_P(1),\tag{A.19}$$

where $\mathbb{i}$ is the imaginary unit. For every $k,l=1,2,...,d$, from Lipschitz continuity, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i,\hat{\boldsymbol{\theta}}_n)-\frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i,\boldsymbol{\theta}_0)\right|\le\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|\frac{1}{n}\sum_{i=1}^{n}\psi(Z_i)=o_P(1).$$

Thus, applying the law of large numbers, we know that $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = \ddot{M}_n(\boldsymbol{\theta}_0) + o_P(1) = \ddot{M}(\boldsymbol{\theta}_0) + o_P(1)$.

Next we prove that $\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n) = \Lambda_\pi(\boldsymbol{\theta}_0) + o_P(1)$. We have

$$\begin{aligned}
&\|\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n) - \Lambda_{n,R}(\boldsymbol{\theta}_0)\| \\
&= \frac{1}{n}\left\| \sum_{i=1}^n \frac{\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}_n) - \dot{m}(Z_i, \boldsymbol{\theta}_0)\dot{m}^{\mathrm{T}}(Z_i, \boldsymbol{\theta}_0)}{n\pi_{n,i}} \right\| \\
&\leq \max_i \left(\frac{1}{n\pi_{n,i}} \right) \frac{1}{n} \sum_{i=1}^n \left\| \dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}_n) - \dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i, \boldsymbol{\theta}_0) \right\| \\
&\quad + \max_i \left(\frac{1}{n\pi_{n,i}} \right) \frac{1}{n} \sum_{i=1}^n \left\| \dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i, \boldsymbol{\theta}_0) - \dot{m}(Z_i, \boldsymbol{\theta}_0)\dot{m}^{\mathrm{T}}(Z_i, \boldsymbol{\theta}_0) \right\| \\
&= \max_i \left(\frac{1}{n\pi_{n,i}} \right) \frac{1}{n} \sum_{i=1}^n \left\{ \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| + \|\dot{m}(Z_i, \boldsymbol{\theta}_0)\| \right\} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) - \dot{m}(Z_i, \boldsymbol{\theta}_0)\|.
\end{aligned}$$

Using Taylor's expansion, we obtain

$$\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = \dot{m}(Z_i, \boldsymbol{\theta}_0) + B_{n,i}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0),$$

where $B_{n,i} = \int_0^1 \ddot{m}\{Z_i, \boldsymbol{\theta}_0 + \lambda(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\}\mathrm{d}\lambda$ satisfies that

$$\|B_{n,i} - \ddot{m}(Z_i, \boldsymbol{\theta}_0)\| \leq d \int_0^1 \lambda\psi(Z_i)\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|\mathrm{d}\lambda = 0.5d\psi(Z_i)\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|.$$

due to the Lipschitz continuity in Assumption 3. This shows that

$$\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) - \dot{m}(Z_i, \boldsymbol{\theta}_0)\| \leq 0.5d\psi(Z_i)\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 + \|\ddot{m}(Z_i, \boldsymbol{\theta}_0)\|\|(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\|.$$

Thus,

$$\begin{aligned}
&\|\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n) - \Lambda_{n,R}(\boldsymbol{\theta}_0)\| \\
&\leq \max_i \left(\frac{1}{n\pi_{n,i}} \right) \left[0.5d\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 \left\{ \frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|\psi(Z_i) + \frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i, \boldsymbol{\theta}_0)\|\psi(Z_i) \right\} \right. \\
&\quad \left. + \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \left\{ \frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|\|\ddot{m}(Z_i, \boldsymbol{\theta}_0)\| + \frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i, \boldsymbol{\theta}_0)\|\|\ddot{m}(Z_i, \boldsymbol{\theta}_0)\| \right\} \right].
\end{aligned}$$

From Hölder's inequality

$$\frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|\psi(Z_i) \leq \left\{ \frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^4 \right\}^{\frac{1}{4}} \left\{ \frac{1}{n}\sum_{i=1}^n \psi(Z_i)^{\frac{4}{3}} \right\}^{\frac{3}{4}} = O_P(1).$$

Similarly, we can show that $\frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i,\boldsymbol{\theta}_0)\|\psi(Z_i)$, $\frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\|\ddot{m}(Z_i,\boldsymbol{\theta}_0)\|$, and $\frac{1}{n}\sum_{i=1}^n \|\dot{m}(Z_i,\boldsymbol{\theta}_0)\|\|\ddot{m}(Z_i,\boldsymbol{\theta}_0)\|$ are all $O_P(1)$. Therefore, $\|\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)-\Lambda_{n,R}(\boldsymbol{\theta}_0)\| = o_P(1)$, and thus (A.19) implies that

$$\mathbb{E}\big(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}S_{n,R}}\big|\mathcal{D}_n\big) = e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\Lambda_\pi(\boldsymbol{\theta}_0)\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}} + o_P(1),$$

where the $o_P(1)$ is bounded.

Note that $Y_n = \sqrt{\frac{s_n}{n}}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$. Using Proposition 1, we have

$$\mathbb{E}(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}Y_n}) \to e^{-0.5\boldsymbol{t}^{\mathrm{T}}c\ddot{M}^{-1}(\boldsymbol{\theta}_0)\Lambda(\boldsymbol{\theta}_0)\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}}.$$

Since Y_n is $\mathcal{D}_n$ measurable, we have

$$\begin{aligned}
&\left|\mathbb{E}\left\{e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}(Y_n+S_{n,R})} - e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}Y_n}e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\Lambda_\pi(\boldsymbol{\theta}_0)\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}}\right\}\right| \\
&= \left|\mathbb{E}\left[\mathbb{E}\left\{e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}(Y_n+S_{n,R})} - e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}Y_n}e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\Lambda_\pi(\boldsymbol{\theta}_0)\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}}\Big|\mathcal{D}_n\right\}\right]\right| \\
&= \left|\mathbb{E}\left[e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}Y_n}\left\{\mathbb{E}\left(e^{\mathbb{i}\boldsymbol{t}S_{n,R}}\big|\mathcal{D}_n\right) - e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\Lambda_\pi(\boldsymbol{\theta}_0)\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}}\right\}\right]\right| \\
&\le \mathbb{E}\left\{\left|\mathbb{E}\left(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}S_{n,R}}|\mathcal{D}_n\right) - e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\Lambda_\pi(\boldsymbol{\theta}_0)\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}}\right|\right\} \\
&= o(1),
\end{aligned}$$

where the last step is from the dominated convergence theorem. Therefore,

$$\mathbb{E}\big\{e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}(Y_n+S_{n,R})}\big\} = \mathbb{E}(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}Y_n})e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\Lambda_\pi(\boldsymbol{\theta}_0)\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}} + o(1) \to \mathbb{E}\big\{e^{-0.5\boldsymbol{t}^{\mathrm{T}}V_R^U(\boldsymbol{\theta}_0)\boldsymbol{t}}\big\}.$$

Hence, we obtain that

$$\sqrt{s_n}\{V_R^U(\boldsymbol{\theta}_0)\}^{-1/2}(\tilde{\boldsymbol{\theta}}_{s_n,R} - \boldsymbol{\theta}_0) \rightsquigarrow \mathbb{N}(\mathbf{0}, \boldsymbol{I}_d).$$

□

A.1.5 Proof of Theorem 2'

Proof of Theorem 2'. The technique of proving Theorem 2' is similar to that of proving Theorem 1'. Denoting $S_{n,P} = \sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,P} - \hat{\boldsymbol{\theta}}_n)$, we write $\sqrt{s_n}(\tilde{\boldsymbol{\theta}}_{s_n,P} - \boldsymbol{\theta}_0) = S_{n,P} + Y_n$. From Theorem 2, we know that under Assumptions 1-5,

$$\mathbb{E}\big(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}S_{n,P}}\big|\mathcal{D}_n\big) = e^{-0.5\boldsymbol{t}^{\mathrm{T}}V_{n,P}(\hat{\boldsymbol{\theta}}_n)\boldsymbol{t}} + o_{P|\mathcal{D}_n} = e^{-0.5\boldsymbol{t}^{\mathrm{T}}V_{n,P}(\hat{\boldsymbol{\theta}}_n)\boldsymbol{t}} + o_P(1).$$

In the proof of Theorem 1', we have proved that $\ddot{M}(\hat{\boldsymbol{\theta}}_n) = \ddot{M}(\boldsymbol{\theta}_0) + o_P(1)$ and $\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n) = \Lambda_\pi + o_P(1)$. Using a similar approach, we can show that

$$\frac{s_n}{n^2}\sum_{i=1}^n \dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n) = c\Lambda(\boldsymbol{\theta}_0) + o_P(1).$$

Therefore,

$$\mathbb{E}\big(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}S_{n,P}}\big|\mathcal{D}_n\big) = e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\{\Lambda_\pi(\boldsymbol{\theta}_0)-c\Lambda(\boldsymbol{\theta}_0)\}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}} + o_P(1).$$

Now we use the same technique used in the proof of Theorem 1'. Since Y_n is $\mathcal{D}_n$ measurable, we have

$$\begin{aligned}
&\left|\mathbb{E}\left\{e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}(Y_n+S_{n,P})} - e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}Y_n}e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\{\Lambda_\pi(\boldsymbol{\theta}_0)-c\Lambda(\boldsymbol{\theta}_0)\}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}}\right\}\right| \\
&= \left|\mathbb{E}\left[e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}Y_n}\left\{\mathbb{E}\left(e^{\mathbb{i}\boldsymbol{t}S_{n,P}}|\mathcal{D}_n\right) - e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\{\Lambda_\pi(\boldsymbol{\theta}_0)-c\Lambda(\boldsymbol{\theta}_0)\}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}}\right\}\right]\right| \\
&\leq \mathbb{E}\left\{\left|\mathbb{E}\left(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}S_{n,P}}|\mathcal{D}_n\right) - e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\{\Lambda_\pi(\boldsymbol{\theta}_0)-c\Lambda(\boldsymbol{\theta}_0)\}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}}\right|\right\} \to 0,
\end{aligned}$$

where the last step is from the dominated convergence theorem. Hence,

$$\mathbb{E}\big\{e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}(Y_n+S_{n,P})}\big\} = \mathbb{E}(e^{\mathbb{i}\boldsymbol{t}^{\mathrm{T}}Y_n})e^{-0.5\boldsymbol{t}^{\mathrm{T}}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\{\Lambda_\pi(\boldsymbol{\theta}_0)-c\Lambda(\boldsymbol{\theta}_0)\}\ddot{M}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{t}} + o(1) \to \mathbb{E}\big\{e^{-0.5\boldsymbol{t}^{\mathrm{T}}V_P^U(\boldsymbol{\theta}_0)\boldsymbol{t}}\big\},$$

and this finishes the proof. □

A.1.6 Proof of Theorem 3

Proof of Theorem 3. For the result in (7),

$$\mathrm{tr}\{\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)\} = \frac{1}{n^2}\sum_{i=1}^n \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{\pi_{n,i}} = \frac{1}{n^2}\sum_{i=1}^n \pi_{n,i}\sum_{i=1}^n \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{\pi_{n,i}} \geq \frac{1}{n^2}\left\{\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\right\}^2.$$

Here, the last step is from the Cauchy-Schwarz inequality and the equality holds if and only if $\pi_{n,i} \propto \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|$.

□

A.1.7 Proof of Theorem 4

Proof. Note that

$$\begin{aligned}
\mathrm{tr}\{\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n)\} &= \mathrm{tr}\left\{\frac{1}{n^2}\sum_{i=1}^n \frac{(1-s_n\pi_{n,i})\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{\pi_{n,i}}\right\} \\
&= \frac{1}{n^2}\left[\sum_{i=1}^n \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{\pi_{n,i}} - s_n\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\right].
\end{aligned}$$

Thus, minimizing $\mathrm{tr}\{\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n)\}$ is equal to minimizing $\sum_{i=1}^n \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{\pi_{n,i}}$. For $i = 1,...,n$, let $t_i = \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|$ and let $t_{(i)}$ denote the order statistics of $\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|$, i.e., $t_{(i)} =$

$\|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(i)}$. The optimization problem of minimizing $\text{tr}\{\Lambda_{n,P}(\hat{\boldsymbol{\theta}}_n)\}$ subject to the constraints on $\pi_{n,i}$ can be presented as minimizing

$$T(\pi_1, \pi_2, ..., \pi_n) = \sum_{i=1}^{n} \frac{t_{(i)}^2}{\pi_{n,i}}, \tag{A.20}$$

$$\text{subject to } \sum_{i=1}^{n} \pi_{n,i} = 1 \quad \text{and} \quad 0 \leq \pi_{n,i} \leq \frac{1}{s_n}, i = 1, 2, ..., n.$$

Defining slack variables $\omega_1^2, \omega_2^2, ..., \omega_n^2$, to use Lagrangian multiplier method, we can construct

$$H(\pi_1, ..., \pi_n, \tau, \mu_1, ..., \mu_n, \omega_1, ..., \omega_n) = \sum_{i=1}^{n} \frac{t_{(i)}^2}{\pi_{n,i}} + \tau \left(\sum_{i=1}^{n} \pi_{n,i} - 1 \right) + \sum_{i=1}^{n} \mu_i \left(\pi_{n,i} + \omega_i^2 - \frac{1}{s_n} \right).$$

By taking the derivatives, the Karush–Kuhn–Tucker (KKT) conditions (Nocedal & Wright 1999) are

$$\frac{\partial H}{\partial \pi_{n,i}} = -\frac{t_{(i)}^2}{\pi_{n,i}^2} + \tau + \mu_i = 0, \qquad i = 1, 2, ..., n. \tag{A.21}$$

$$\frac{\partial H}{\partial \tau} = \sum_{i=1}^{n} \pi_{n,i} - 1 = 0, \tag{A.22}$$

$$\frac{\partial H}{\partial \mu_i} = \pi_{n,i} + \omega_i^2 = \frac{1}{s_n}, \qquad i = 1, 2, ..., n. \tag{A.23}$$

$$\frac{\partial H}{\partial \omega_i} = 2\mu_i \omega_i = 0, \qquad i = 1, 2, ..., n. \tag{A.24}$$

$$\mu_i \geq 0, \qquad i = 1, 2, ..., n. \tag{A.25}$$

From (A.21), we have

$$\pi_{n,i} = \frac{t_{(i)}}{\sqrt{\tau + \mu_i}}, \quad i = 1, 2, ..., n. \tag{A.26}$$

Combining it with (A.23), we have

$$\frac{t_{(i)}}{\sqrt{\tau + \mu_i}} + \omega_i^2 = \frac{1}{s_n}, \quad i = 1, 2, ..., n. \tag{A.27}$$

According to (A.24), at least one of μ_i and ω_i must be 0. From (A.26) and (A.27),

$$\text{if } t_{(i)} < \frac{\sqrt{\tau}}{s}, \qquad \mu = 0 \text{ and } \pi_{n,i} = \frac{t_{(i)}}{\sqrt{\tau}} < \frac{1}{s_n}; \tag{A.28}$$

$$\text{if } t_{(i)} \geq \frac{\sqrt{\tau}}{s}, \qquad \omega_i = 0 \text{ and } \pi_{n,i} = \frac{t_{(i)}}{\sqrt{\tau + \mu_i}} = \frac{1}{s_n}. \tag{A.29}$$

Thus, letting g be the number of cases that $t_{(i)} \geq \frac{\sqrt{\tau}}{s}$, from (A.22) and the fact that $t_{(i)}$ is non-decreasing in i,

$$1 = \sum_{i=1}^{n} \pi_{n,i} = \sum_{i=1}^{n-g} \frac{t_{(i)}}{\sqrt{\tau}} + \sum_{i=n-g+1}^{n} \frac{1}{s_n} = \frac{\sum_{i=1}^{n-g} t_{(i)}}{\sqrt{\tau}} + \frac{g}{s}, \tag{A.30}$$

which shows that

$$\sqrt{\tau} = \frac{s}{s-g} \sum_{i=1}^{n-g} t_{(i)}. \tag{A.31}$$

Combining(A.28), (A.29), and (A.31),

$$\pi_{n,i} = \begin{cases} \dfrac{t_{(i)}(s_n - g)}{s_n \sum_{i=1}^{n-g} t_{(i)}}, & \text{for } i = 1, 2, ..., n-g; \quad \text{(A.32)} \\ \dfrac{1}{s_n}, & \text{for } i = n-g+1, ..., n. \quad \text{(A.33)} \end{cases}$$

From (A.31),

$$H = \frac{\sum_{i=1}^{n-g} t_{(i)}}{s_n - g} = \frac{\sqrt{\tau}}{s_n}, \tag{A.34}$$

Thus, from (A.28) and (A.29), we know $t_{(i)} < H$ for $i = 1, 2, ..., n-g$, and $t_{(i)} \geq H$, for $i = n-g+1, ..., n$. Therefore

$$\sum_{i=1}^{n} (t_{(i)} \wedge H) = \sum_{i=1}^{n-g} t_{(i)} + \sum_{i=n-g+1}^{n} H = s_n H \tag{A.35}$$

Thus, from (A.32), for $i = 1, 2, ..., n-g$,

$$\pi_{n,i} = \frac{t_{(i)}}{s_n H} = \frac{t_{(i)} \wedge H}{\sum_{i=1}^{n} (t_{(i)} \wedge H)}; \tag{A.36}$$

from (A.33), for $i = n-g+1, ..., n$,

$$\pi_{n,i} = \frac{H}{s_n H} = \frac{t_{(i)} \wedge H}{\sum_{i=1}^{n} (t_{(i)} \wedge H)}. \tag{A.37}$$

For the result under the A-optimality, define $t_{(i)} = \|\ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(i)}$ and the proof is the same as the used for the L-optimality. □

A.1.8 Proof of Theorem 5

The proof of Theorem 5 relies on Lemmas 5 and 6 below.

Lemma 5. *Under Assumption 3, if* $\|\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n\|=o_P(1)$*, then conditional on* $\mathcal{D}_n$ *and* $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$*,*

$$B_{s_n}^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}-\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)=o_P(1), \tag{A.38}$$

where

$$B_{s_n}^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}=\int_0^1\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}\left\{Z_i^*,\hat{\boldsymbol{\theta}}_n+\lambda(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n)\right\}}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}\,\mathrm{d}\lambda.$$

Proof. For every $k,l=1,2,...,d$, from Lipschitz continuity, we have

$$\begin{aligned}
&\left|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}\{Z_i^*,\hat{\boldsymbol{\theta}}_n+\lambda(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n)\}}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}-\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*,\hat{\boldsymbol{\theta}}_n)}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}\right|\\
&\leq\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\psi(Z_i^*)\|\lambda(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n)\|}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}\\
&\leq\lambda\|\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n\|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\psi(Z_i^*)}{\alpha}=\|\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n\|O_P(1)=o_P(1).
\end{aligned}\tag{A.39}$$

According to (A.2), we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*,\hat{\boldsymbol{\theta}}_n)}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}\middle|\mathcal{D}_n,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\right)=\frac{1}{n}\sum_{i=1}^{n}\ddot{m}_{k,l}(Z_i,\hat{\boldsymbol{\theta}}_n),\\
&\mathbb{V}\left(\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}_{k,l}(Z_i^*,\hat{\boldsymbol{\theta}}_n)}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}\middle|\mathcal{D}_n,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\right)\leq\frac{1}{s_n}\sum_{i=1}^{n}\frac{\ddot{m}^2_{k,l}(Z_i,\hat{\boldsymbol{\theta}}_n)}{n^2\pi^{\text{opt}}_{n,R\alpha i}}\leq\frac{1}{\alpha s_n n}\sum_{i=1}^{n}\ddot{m}^2_{k,l}(Z_i,\hat{\boldsymbol{\theta}}_n)=O_P(s_n^{-1}).
\end{aligned}$$

Thus, by Chebyshev's inequality, similar to (A.3), we have

$$\left\|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}(Z_i^*,\hat{\boldsymbol{\theta}}_n)}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}-\frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\right\|=o_{P|\mathcal{D}_n,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}(1).\tag{A.40}$$

Combining (A.39) and (A.40), we have

$$\begin{aligned}
\left\|B_{s_n}-\frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\right\|&\leq\int_0^1\left\|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}\{Z_i^*,\hat{\boldsymbol{\theta}}_n+\lambda(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n)\}}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}-\frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\right\|\mathrm{d}\lambda\\
&\leq\int_0^1\left[\left\|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}\{Z_i^*,\hat{\boldsymbol{\theta}}_n+\lambda(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}-\hat{\boldsymbol{\theta}}_n)\}}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}-\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}(Z_i^*,\hat{\boldsymbol{\theta}}_n)}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}\right\|\right.\\
&\quad\left.+\left\|\frac{1}{s_n}\sum_{i=1}^{s_n}\frac{\ddot{m}(Z_i^*,\hat{\boldsymbol{\theta}}_n)}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}-\frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\right\|\right]\mathrm{d}\lambda=o_P(1),
\end{aligned}$$

which finishes the proof. □

Lemma 6. *If Assumption 4 hold, then given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$ in probability,*

$$\sqrt{s_n}\{\Lambda^{\alpha}_{R}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\}^{-1/2}\dot{M}^{*}_{R\alpha}(\hat{\boldsymbol{\theta}}_n) \to \mathbb{N}\left(\mathbf{0}, \boldsymbol{I}\right), \tag{A.41}$$

in conditional distribution, where

$$\dot{M}^{*}_{R\alpha}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{ns_n}\sum_{i=1}^{s_n}\frac{\dot{m}(Z^{*}_i, \hat{\boldsymbol{\theta}}_n)}{\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}}, \quad \textit{and} \quad \Lambda^{\alpha}_{R}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}_n)}{\tilde{\pi}^{\text{opt}}_{n,R\alpha i}}.$$

Proof. Note that

$$\sqrt{s_n}\dot{M}^{*}_{R\alpha}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{\sqrt{s_n}}\sum_{i=1}^{s_n}\frac{\dot{m}(Z^{*}_i, \hat{\boldsymbol{\theta}}_n)}{n\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}} \equiv \frac{1}{\sqrt{s_n}}\sum_{i=1}^{s_n}\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}. \tag{A.42}$$

Given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$, $\boldsymbol{\eta}_1^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}, ..., \boldsymbol{\eta}_{s_n}^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}$ are i.i.d, with

$$\mathbb{E}(\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}) = \frac{1}{n}\sum_{i=1}^{n}\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = \mathbf{0}, \quad \text{and} \tag{A.43}$$

$$\mathbb{V}(\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}) = \mathbb{E}\left\{\frac{\dot{m}(Z^{*}_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z^{*}_i, \hat{\boldsymbol{\theta}}_n)}{n^2(\tilde{\pi}^{\text{opt}*}_{n,R\alpha i})^2}\middle|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\right\} \tag{A.44}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\frac{\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}_n)}{\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}} = \Lambda^{\alpha}_{R}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}). \tag{A.45}$$

Meanwhile, for every $\varepsilon > 0$ and some $\delta \in (0, 2]$,

$$\begin{aligned}
&\frac{1}{s_n}\sum_{i=1}^{s_n}\mathbb{E}\left\{\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}\|^2 I(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}\| > s_n^{1/2}\varepsilon)\middle|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\right\} \\
&\leq \frac{1}{s_n^{1+\delta/2}\varepsilon^{\delta}}\sum_{i=1}^{s_n}\mathbb{E}\left\{\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}\|^{2+\delta} I(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}\| > s_n^{1/2}\varepsilon)\middle|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\right\} \\
&\leq \frac{1}{s_n^{1+\delta/2}\varepsilon^{\delta}}\sum_{i=1}^{s_n}\mathbb{E}\left(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}\|^{2+\delta}\middle|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\right) \\
&\leq \frac{1}{s_n^{\delta/2}n^{2+\delta}\varepsilon^{\delta}}\sum_{i=1}^{n}\frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^{2+\delta}}{(\tilde{\pi}^{\text{opt}*}_{n,R\alpha i})^{1+\delta}} \\
&\leq \frac{1}{s_n^{\delta/2}\alpha^{1+\delta}\varepsilon^{\delta}}\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^{2+\delta} = O_P(s_n^{-\delta/2}) = o_P(1).
\end{aligned}$$

where the second last equality is from Assumption 4. This show that Lindeberg's condition is satisfied in probability. From (A.42), (A.43) and (A.45), by the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart (1998)), conditional on $\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$, we obtain (A.41). □

Proof of Theorem 5. By direct calculation, we have

$$\mathbb{E}\left\{M^*_{R\alpha}(\boldsymbol{\theta})\middle|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\right\} = M_n(\boldsymbol{\theta}),$$
$$\mathbb{V}\left\{M^*_{R\alpha}(\boldsymbol{\theta})\middle|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\right\} \leq \frac{1}{s_n n^2}\sum_{i=1}^n \frac{m^2(Z_i,\boldsymbol{\theta})}{\tilde{\pi}^{\text{opt}*}_{n,R\alpha i}} \leq \frac{1}{s_n n}\sum_{i=1}^n \frac{m^2(Z_i,\boldsymbol{\theta})}{\alpha} = O_P(s_n^{-1}).$$

By Chebyshev's inequality, for each $\boldsymbol{\theta}$, we have

$$M^*_{R\alpha}(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}) = o_{P|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}(1).$$

Under Assumptions 1 and 2, the parameter space is compact and $\hat{\boldsymbol{\theta}}_n$ is the unique global maximum of the continuous concave function $M_n(\boldsymbol{\theta})$. Thus from Theorem 5.9 and its remark of van der Vaart (1998), conditionally on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$,

$$\|\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R} - \hat{\boldsymbol{\theta}}_n\| = o_P(1).$$

By Taylor expansion

$$0 = \dot{M}^*_{R\alpha}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R}) = \dot{M}^*_{R\alpha}(\hat{\boldsymbol{\theta}}_n) + B_{s_n}^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R} - \hat{\boldsymbol{\theta}}_n),$$

so

$$\begin{aligned}\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R} - \hat{\boldsymbol{\theta}}_n &= -\left(B_{s_n}^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}\right)^{-1}\dot{M}^*_{R\alpha}(\hat{\boldsymbol{\theta}}_n) \\ &= -\frac{1}{\sqrt{s_n}}\left(B_{s_n}^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}}\right)^{-1}\{\Lambda^{\alpha}_R(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\}^{1/2}\sqrt{s_n}\{\Lambda^{\alpha}_R(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\}^{-1/2}\dot{M}^*_{R\alpha}(\hat{\boldsymbol{\theta}}_n).\end{aligned}$$

Therefore, from Lemma 5 and Lemma 6, conditional on $\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$, by Slutsky's theorem

$$\sqrt{s_n}\{\Lambda^{\alpha}_R(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\}^{-1/2}\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R} - \hat{\boldsymbol{\theta}}_n) \to \mathbb{N}\left(\mathbf{0}, \boldsymbol{I}\right), \tag{A.46}$$

in conditional distribution.

Next, we check the distance bewtween $\Lambda^{\alpha}_R(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})$ and $\Lambda^{\alpha}_R(\hat{\boldsymbol{\theta}}_n)$.

$$\begin{aligned}&\|\Lambda^{\alpha}_R(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}) - \Lambda^{\alpha}_R(\hat{\boldsymbol{\theta}}_n)\| \\ &= \left\|\frac{1}{n^2}\sum_{i=1}^n \frac{\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{(1-\alpha)\pi^{\text{opt}}_{Ri}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}) + \alpha\frac{1}{n}} - \frac{1}{n^2}\sum_{i=1}^n \frac{\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{(1-\alpha)\pi^{\text{opt}}_{Ri}(\hat{\boldsymbol{\theta}}_n) + \alpha\frac{1}{n}}\right\| \\ &\leq \frac{1}{n^2}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\left|\frac{1}{(1-\alpha)\pi^{\text{opt}}_{Ri}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}) + \alpha\frac{1}{n}} - \frac{1}{(1-\alpha)\pi^{\text{opt}}_{Ri}(\hat{\boldsymbol{\theta}}_n) + \alpha\frac{1}{n}}\right| \\ &< \frac{1}{\alpha^2}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\left|\pi^{\text{opt}}_{Ri}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}) - \pi^{\text{opt}}_{Ri}(\hat{\boldsymbol{\theta}}_n)\right|\end{aligned}$$

$$\leq \frac{1}{\alpha^2}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\Bigg\{\frac{\left|\|\dot{m}(Z_i,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\|-\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\right|}{\sum_{j=1}^{n}\|\dot{m}(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\|} + \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\frac{\sum_{j=1}^{n}\left|\|\dot{m}(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\|-\|\dot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)\|\right|}{\sum_{j=1}^{n}\|\dot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)\|\sum_{j=1}^{n}\|\dot{m}(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\|}\Bigg\}$$

$$\equiv \frac{1}{\alpha^2}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\left(\Delta_{1i}+\Delta_{2i}\right). \tag{A.47}$$

Under Assumption 3, for any $j=1,2,...,n$

$$\begin{aligned}
&\left|\|\dot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)\|-\|\dot{m}(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\|\right| \leq \|\dot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)-\dot{m}(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\| \\
&\leq \sqrt{\sum_{k=1}^{d}\{\dot{m}_k(Z_j,\hat{\boldsymbol{\theta}}_n)-\dot{m}_k(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\}^2} \leq \sum_{k=1}^{d}\left|\dot{m}_k(Z_j,\hat{\boldsymbol{\theta}}_n)-\dot{m}_k(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\right| \\
&\leq \sum_{k=1}^{d}\left|\ddot{m}_k^T(Z_j,\xi_k)(\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\right| \leq \|\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\|\sum_{k=1}^{d}\|\ddot{m}_k(Z_j,\xi_k)\| \equiv \|\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\|h(Z_j),
\end{aligned} \tag{A.48}$$

where $\dot{m}_k(Z_j,\hat{\boldsymbol{\theta}}_n)$ is the kth element of $\dot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)$, $\ddot{m}_k(Z_j,\hat{\boldsymbol{\theta}}_n)$ is the kth column of $\ddot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)$, and all ξ_k are between $\hat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$. Thus,

$$\Delta_{1i} \leq \frac{\|\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\|h(Z_i)}{\sum_{j=1}^{n}\|\dot{m}(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\|}, \tag{A.49}$$

and

$$\Delta_{2i} \leq \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\|\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\|\sum_{j=1}^{n}h(Z_j)}{\sum_{j=1}^{n}\|\dot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)\|\sum_{j=1}^{n}\|\dot{m}(Z_j,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R})\|} \tag{A.50}$$

From (A.2) and Assumption 3

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}h^2(Z_j) \leq d\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\|\ddot{m}_k(Z_j,\xi_k)\|^2 = d\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}\ddot{m}^2_{k,l}(Z_j,\xi_k) \\
&\leq d\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\sum_{l=1}^{d}\left(2\ddot{m}^2_{k,l}(Z_j,\hat{\boldsymbol{\theta}}_n)+2\psi^2(Z_j)\|\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}-\hat{\boldsymbol{\theta}}_n\|^2\right) = O_P(1)
\end{aligned} \tag{A.51}$$

which also implies that $\frac{1}{n}\sum_{j=1}^{n}h(Z_j)=O_P(1)$. Thus,

$$\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\Delta_{1i} \leq \frac{O_P(\|\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\|)}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2h(Z_i)$$

$$\leq O_P(\|\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\|)\left\{\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^4\right\}^{\frac{1}{2}}\left\{\frac{1}{n}\sum_{i=1}^{n}h^2(Z_i)\right\}^{\frac{1}{2}}, \quad \text{(A.52)}$$

and

$$\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\Delta_{2i} = O_P(\|\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\|)\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2 \quad \text{(A.53)}$$

Combining (A.47), (A.52), and (A.53), we obtain that for large s_0, s_n and n,

$$\|\Lambda^{\alpha}_R(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}) - \Lambda^{\alpha}_R(\hat{\boldsymbol{\theta}}_n)\| = \|\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}\|O_P(1) = o_P(1).$$

Thus, Slutsky's theorem and (A.46) indicate that given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$, as s_0, s_n and $n \to \infty$

$$\sqrt{s_n}\{V^{\alpha}_{n,R}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,R} - \hat{\boldsymbol{\theta}}_n) \to \mathbb{N}\left(\mathbf{0}, \boldsymbol{I}\right),$$

in conditional distribution. □

A.1.9 Proof of Theorem 6

The proof of Theorem 6 relies on Lemmas 7, 8 and 9.

Lemma 7. *Under Assumptions 4, conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,R}$, then*

$$\sqrt{s_n}\{\Lambda^{\alpha}_{n,P}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\}^{-1/2}\dot{M}^{*}_{P\alpha}(\hat{\boldsymbol{\theta}}_n) \to \mathbb{N}(\mathbf{0}, \boldsymbol{I}),$$

in conditional distribution, where

$$\Lambda^{\alpha}_{n,P}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}) = \frac{s_n}{n^2}\sum_{i=1}^{n}\frac{\{1-(s_n\tilde{\pi}^{\text{opt}}_{n,P\alpha i})\wedge 1\}\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{(s_n\tilde{\pi}^{\text{opt}}_{n,P\alpha i})\wedge 1}.$$

Proof. For the sake of readability, in the sequel, we redefine ν_i as $\nu_i = I(u_i \leq s_n\pi^{\text{opt}}_{n,P\alpha i})$ and let

$$\sqrt{s_n}\dot{M}^{*}_{P\alpha}(\hat{\boldsymbol{\theta}}_n) = \sum_{i=1}^{n}\frac{\nu_i\sqrt{s_n}\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)}{n\{(s_n\tilde{\pi}^{\text{opt}}_{n,P\alpha i})\wedge 1\}} \equiv \sum_{i=1}^{n}\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}}. \quad \text{(A.54)}$$

From direct calculation and the definition of $\hat{\boldsymbol{\theta}}_n$, we have

$$\mathbb{E}\Big(\sqrt{s_n}\dot{M}^{*}_{P\alpha}(\hat{\boldsymbol{\theta}}_n)\Big|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}\Big) = \frac{\sqrt{s_n}}{n}\sum_{i=1}^{n}\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n) = \mathbf{0},$$

and

$$\mathbb{V}\Big(\sqrt{s_n}\dot{M}^{*}_{P\alpha}(\hat{\boldsymbol{\theta}}_n)\Big|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}\Big) = \frac{s_n}{n^2}\sum_{i=1}^{n}\frac{\mathbb{V}(\nu_i|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\dot{m}^{\mathrm{T}}(Z_i,\hat{\boldsymbol{\theta}}_n)}{\{(s_n\tilde{\pi}^{\text{opt}}_{n,P\alpha i})\wedge 1\}^2}$$

$$= \frac{s_n}{n^2} \sum_{i=1}^{n} \frac{\{1 - (s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1\} \dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) \dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}_n)}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1}$$
$$\leq \frac{1}{\alpha n} \sum_{i=1}^{n} \dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) \dot{m}^{\mathrm{T}}(Z_i, \hat{\boldsymbol{\theta}}_n) = O_P(1).$$

Next, we check Lindeberg's condition. For any $\epsilon > 0$ and $\rho \in (0, 2]$,

$$\mathbb{E}\left\{ \sum_{i=1}^{n} \|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}}\| I(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}}\| > \varepsilon) \middle| \mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P} \right\}$$
$$\leq \frac{1}{\varepsilon^{\rho}} \sum_{i=1}^{n} \mathbb{E}\left\{ \|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}}\|^{2+\rho} I(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}}\| > \varepsilon) \middle| \mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P} \right\}$$
$$\leq \frac{1}{\varepsilon^{\rho}} \sum_{i=1}^{n} \mathbb{E}\left(\|\boldsymbol{\eta}_i^{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}}\|^{2+\rho} \middle| \mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P} \right) = \frac{s_n^{1+\rho/2}}{\varepsilon^{\rho} n^{2+\rho}} \sum_{i=1}^{n} \frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^{2+\rho}}{\{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1\}^{1+\rho}}$$
$$\leq \frac{s_n^{1+\rho/2}}{\varepsilon^{\rho} n^{2+\rho}} \sum_{i=1}^{n} \frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^{2+\rho}}{(s_n \alpha / n)^{1+\rho}} = \frac{1}{\alpha^{1+\rho} \varepsilon^{\rho} {s_n}^{\rho/2}} \frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^{2+\rho} = O_P({s_n}^{-\rho/2}).$$

Thus, from the Lindeberg-Feller Central Limit Theorem (cf. van der Vaart 1998), Lemma 7 follows. □

Lemma 8. *Under Assumption 3, for any $\boldsymbol{u}_{s_n} = o_P(1)$, conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}$,*

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\nu_i \ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n + \boldsymbol{u}_{s_n})}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1} - \frac{1}{n} \sum_{i=1}^{n} \ddot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = o_P(1).$$

Proof. First, using an approach similar to the one used to prove (A.14), we can show that given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}$,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\nu_i \psi(Z_i)}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1} = O_P(1). \tag{A.55}$$

For every $k, l = 1, 2, ..., d$, from Lipschitz continuity, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \frac{\nu_i \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n + \boldsymbol{u}_{s_n})}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1} - \frac{1}{n} \sum_{i=1}^{n} \frac{\nu_i \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1} \right| \leq \frac{1}{n} \sum_{i=1}^{n} \frac{\nu_i \psi(Z_i) \|\boldsymbol{u}_{s_n}\|}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1} = o_P(1). \tag{A.56}$$

For each $k, l = 1, 2, ..., d$, direct calculations show that

$$\mathbb{E}\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\nu_i \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1} \middle| \mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P} \right\} = \frac{1}{n} \sum_{i=1}^{n} \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n),$$
$$\mathbb{V}\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\nu_i \ddot{m}_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1} \middle| \mathcal{D}_n, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P} \right\} \leq \frac{1}{s_n n^2} \sum_{i=1}^{n} \frac{\ddot{m}^2_{k,l}(Z_i, \hat{\boldsymbol{\theta}}_n)}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1} \leq \frac{1}{\alpha s_n n} \sum_{i=1}^{n} h^2(Z_i) = O_P(s_n^{-1}).$$

According to Chebyshev's inequality, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)}{(s_n\tilde{\pi}^{\mathrm{opt}}_{n,P\alpha i})\wedge 1}-\frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)=O_P({s_n}^{-1/2}). \tag{A.57}$$

Therefore, combining (A.56) and (A.57), we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_i \ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n+\boldsymbol{u}_{s_n})}{s_n\pi^{\mathrm{opt}}_{n,P\alpha i}}-\frac{1}{n}\sum_{i=1}^{n}\ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)=o_P(1).$$

□

Lemma 9. *Under Assumptions 3 and 4,*

1) if $\varrho_n = s_n/(bn) \to \varrho \in (0,1)$, *then* $H^{0*} - H_{\varrho_n} = o_P(1)$;

2) $\Psi^{0*} - \Psi_{\varrho_n} = o_P(1)$, *where*

$$\Psi_{\varrho_n} = \frac{1}{n}\sum_{i=1}^{n}\{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n}\}; \tag{A.58}$$

3) if $s_n/(bn) \to \varrho = 0$, *then* $\Psi^{0*} - \Psi_\infty = o_P(1)$.

Proof. Note that H^{0*} is the $\lceil s_0^* - s_0^* s_n/b/n\rceil$-th order statistics of $\|\dot{m}(Z_i^{0*},\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|$, $i = 1,...,s_0^*$. For any $\rho > 0$, let $\tilde{H}_\rho$ be the $\lceil n(1-\rho)\rceil$-th order statistics of $\|\dot{m}(Z_i,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|$, $i = 1,...,n$. Let $\nu^0_{(i)} = 1$ if $\|\dot{m}(Z,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|_{(i)}$ is included in $\|\dot{m}(Z_1^{0*},\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|,...,\|\dot{m}(Z_{s_0^*}^{0*},\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|$, and $\nu^0_{(i)} = 0$ otherwize. For any $\varrho_+ > \varrho$,

$$\mathbb{P}(H^{0*} \le \tilde{H}_{\varrho_+}) = \mathbb{P}\left(\sum_{i=1}^{\lceil n(1-\varrho_+)\rceil}\nu^0_{(i)} \ge \lceil s_0^* - s_0^* s_n/b/n\rceil\right). \tag{A.59}$$

Note that

$$\frac{1}{s_0}\sum_{i=1}^{\lceil n(1-\varrho_+)\rceil}\nu^0_{(i)} = 1-\varrho_+ + o_P(1) \quad \text{and} \quad \frac{\lceil s_0^* - s_0^* s_n/b/n\rceil}{s_0} = 1-\varrho+o_P(1). \tag{A.60}$$

Thus,

$$\mathbb{P}(H^{0*} \le \tilde{H}_{\varrho_+}) \to 0. \tag{A.61}$$

Similarly, we obtain that for any $\varrho_- < \varrho$,

$$\mathbb{P}(H^{0*} \le \tilde{H}_{\varrho_-}) \to 1. \tag{A.62}$$

Note that $\tilde{H}_{\varrho_+}$ is between the $\lceil n(1-\varrho_+)\rceil - s_0^*$-th and the $\lceil n(1-\varrho_+)\rceil$-th order statistics of $\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|$'s that are not included in $\|\dot{m}(Z_1^{0*}, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|, ..., \|\dot{m}(Z_{s_0^*}^{0*}, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|$. The joint distribution of these $\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|$'s are exchangeable, and $s_0^*/n \to 0$ in probability. Therefore, both the $\lceil n(1-\varrho_+)\rceil - s_0^*$-th and the $\lceil n(1-\varrho_+)\rceil$-th order statistics of these $\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|$'s converge to the ϱ_+-quantile of the distribution of $\|\dot{m}(Z_i, \boldsymbol{\theta}_0)\|$ in probability (Chanda 1971), where $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} \mathbb{E}\{m(Z, \boldsymbol{\theta})\}$. As a result, $\tilde{H}_{\varrho_+}$ converge in probability to the ϱ_+-quantile of the distribution of $\|\dot{m}(Z, \boldsymbol{\theta}_0)\|$, say $\zeta_{\varrho+}$. Similarly, $\tilde{H}_{\varrho_-}$ converge in probability to the ϱ_--quantile of the distribution of $\|\dot{m}(Z, \boldsymbol{\theta}_0)\|$, say $\zeta_{\varrho-}$. Thus, (A.61) and (A.62) together imply that for any $\epsilon > 0$,

$$\mathbb{P}(\zeta_{\varrho_+} - \epsilon < H^{0*} < \zeta_{\varrho_-} + \epsilon) \to 1. \tag{A.63}$$

Since the distribution of Z is continuous and so is that of $\|\dot{m}(Z, \boldsymbol{\theta}_0)\|$, we can choose ϱ_+ and ϱ_- close enough to ϱ such that $\zeta_{\varrho_-} - \zeta_\varrho < \epsilon$ and $\zeta_\varrho - \zeta_{\varrho_+} < \epsilon$, which implies that

$$\mathbb{P}(\zeta_\varrho - 2\epsilon < H^{0*} < \zeta_\varrho + 2\epsilon) \to 1, \tag{A.64}$$

for any ϵ. Thus, $H^{0*} = \zeta_\varrho + o_P(1)$. Since $\|\dot{m}(Z_1, \hat{\boldsymbol{\theta}}_n)\|, ..., \|\dot{m}(Z_n, \hat{\boldsymbol{\theta}}_n)\|$ are exchangeable, $H_{\varrho_n} = \zeta_\varrho + o_P(1)$, where H_{ϱ_n} is the $\lceil n(1-\varrho_n)\rceil$-th order statistics of $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|$, $i = 1, ..., n$. Therefore, $H^{0*} - H_{\varrho_n} = o_P(1)$.

Now we prove 2) of Lemma 9. If $\varrho = 0$ and $\|\dot{m}(Z, \boldsymbol{\theta})\|$ is bounded, then

$$\begin{aligned}
\Psi^{0*} &= \sum_{i=1}^{\lceil s_0^* - s_0^* s_n/b/n\rceil} \frac{\|\dot{m}(Z^{0*}, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|_{(i)}}{s_0^*} + \frac{s_0^* - \lceil s_0^* - s_0^* s_n/b/n\rceil}{s_0^*} H^{0*} \\
&= \sum_{i=1}^{s_0^*} \frac{\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|}{s_0^*} + o_P(1),
\end{aligned}$$

and similarly,

$$\Psi_{\varrho_n} = \frac{1}{n}\sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| + o_P(1).$$

Thus the proof reduce to prove that

$$\sum_{i=1}^{s_0^*} \frac{\|\dot{m}(Z_i^{0*}, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|}{s_0^*} = \frac{1}{n}\sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| + o_P(1),$$

which can be proved by Taylor's expansion and Markov's inequality. To prove other cases, let $\nu_i^0 = 1$ if the i-th observation is included in the pilot subsample and $\nu_i^0 = 0$ otherwise; then Ψ^{0*} can be written as

$$\Psi^{0*} = \frac{1}{s_0^*}\sum_{i=1}^{n} \nu_i^0 \{\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| \wedge H^{0*}\}.$$

Define

$$\Psi^{0*}_{H_{\varrho n}} = \frac{1}{s_0^*}\sum_{i=1}^n \nu_i^0\{\|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| \wedge H_{\varrho n}\} \quad \text{and} \quad \Psi^{0*}_{\hat{\boldsymbol\theta}_n} = \frac{1}{s_0^*}\sum_{i=1}^n \nu_i^0\{\|\dot m(Z_i,\hat{\boldsymbol\theta}_n)\| \wedge H_{\varrho n}\}.$$

If $\varrho > 0$, then

$$\begin{aligned}|\Psi^{0*} - \Psi^{0*}_{H_{\varrho n}}| &= \frac{1}{s_0^*}\sum_{i=1}^n \nu_i^0\Big|\|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| \wedge H^{0*} - \|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| \wedge H_{\varrho n}\Big| \\ &\le \frac{|H^{0*} - H_{\varrho n}|}{s_0^*}\sum_{i=1}^n \nu_i^0 I\Big\{\|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| \ge H^{0*} \wedge H_{\varrho n}\Big\} \le |H^{0*} - H_{\varrho n}| = o_P(1).\end{aligned}$$

If $\varrho = 0$ and $\|\dot m(Z,\boldsymbol\theta)\|$ is unbounded, then $H^{0*} \wedge H_{\varrho n} \to \infty$ in probability. Under Assumptions 3 and 4, it can be shown that $\frac{1}{s_0^*}\sum_{i=1}^n \nu_i^0 \|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\|^2 = O_{P|\mathcal{D}_n}(1)$. Thus,

$$\begin{aligned}|\Psi^{0*} - \Psi^{0*}_{H_{\varrho n}}| &\le \frac{1}{s_0^*}\sum_{i=1}^n \nu_i^0 \|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| I\Big\{\|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| \ge H^{0*} \wedge H_{\varrho n}\Big\} \\ &\quad + \frac{H^{0*}}{s_0^*}\sum_{i=1}^n \nu_i^0 I\Big\{\|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| \ge H^{0*}\Big\} + \frac{H_{\varrho n}}{s_0^*}\sum_{i=1}^n \nu_i^0 I\Big\{\|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| \ge H_{\varrho n}\Big\} \\ &\le \left\{\frac{1}{H^{0*} \wedge H_{\varrho n}} + \frac{1}{H^{0*}} + \frac{1}{H_{\varrho n}}\right\}\frac{1}{s_0^*}\sum_{i=1}^n \nu_i^0 \|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\|^2 = o_P(1). \qquad \text{(A.65)}\end{aligned}$$

Furthermore, we can show that

$$\begin{aligned}|\Psi^{0*}_{H_{\varrho n}} - \Psi^{0*}_{\hat{\boldsymbol\theta}_n}| &\le \frac{1}{s_0^*}\sum_{i=1}^n \nu_i^0\{\|\dot m(Z_i,\tilde{\boldsymbol\theta}^{0*}_{s_0,P})\| - \|\dot m(Z_i,\hat{\boldsymbol\theta}_n)\|\} \\ &\le \frac{\|\hat{\boldsymbol\theta}_n - \tilde{\boldsymbol\theta}^{0*}_{s_0,P}\|}{s_0^*}\sum_{i=1}^n \nu_i^0 h(Z_i) = o_P(1)\end{aligned}$$

and

$$|\Psi^{0*}_{\hat{\boldsymbol\theta}_n} - \Psi_{\varrho_n}| = o_P(1),$$

where the last two $o_P(1)$ are obtained by mean and variance calculations under the conditional distribution of ν_i^0's. Thus, we have that

$$|\Psi^{0*} - \Psi_{\varrho_n}| = o_P(1). \qquad \text{(A.66)}$$

With 2) of Lemma 9 proved, in order to prove 3), we only need to show that $\Psi_\infty - \Psi_{\varrho_n} = o_P(1)$ if $s_n/(bn) \to \varrho = 0$. This is true because if $\|\dot m(Z,\boldsymbol\theta)\|$ is bounded, then

$$|\Psi_\infty - \Psi_{\varrho_n}| \le \frac{1}{n}\sum_{i=1}^n \Big|\|\dot m(Z_i,\hat{\boldsymbol\theta}_n)\| - \|\dot m(Z_i,\hat{\boldsymbol\theta}_n)\| \wedge H_{\varrho_n}\Big|$$

$$\leq \frac{n - \lceil n(1-\varrho_n)\rceil}{n} \|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|_{(n)} = o_P(1);$$

otherwise,

$$\begin{aligned}
|\Psi_\infty - \Psi_{\varrho_n}| &\leq \frac{1}{n}\sum_{i=1}^{n} \Big| \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| - \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n} \Big| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| I\Big\{ \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \geq H_{\varrho_n} \Big\} \\
&\leq \frac{1}{nH_{\varrho_n}} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^2 = o_P(1).
\end{aligned}$$

□

Proof of Theorem 6. For Algorithm 3, $M^*_{P\alpha}(\boldsymbol{\theta})$ can be written as

$$M^*_{P\alpha}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\nu_i m(Z_i, \boldsymbol{\theta})}{(s_n \tilde{\pi}^{\text{opt}}_{n,P\alpha i}) \wedge 1}.$$

Denote

$$\gamma_{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P} P}(\boldsymbol{u}) = s_n M^*_{P\alpha}(\hat{\boldsymbol{\theta}}_n + \boldsymbol{u}/\sqrt{s_n}) - s_n M^*_{P\alpha}(\hat{\boldsymbol{\theta}}_n).$$

Under Assumption 2, $\sqrt{s_n}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,P} - \hat{\boldsymbol{\theta}}_n)$ is the unique maximizer of $\gamma_{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P} P}(\boldsymbol{u})$. By Taylor's expansion,

$$\gamma_{\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P} P}(\boldsymbol{u}) = \sqrt{s_n}\boldsymbol{u}^T \dot{M}^*_{P\alpha}(\hat{\boldsymbol{\theta}}_n) + \frac{\boldsymbol{u}^T \ddot{M}^*_{P\alpha}(\hat{\boldsymbol{\theta}}_n + \acute{\boldsymbol{u}}/\sqrt{s_n})\boldsymbol{u}}{2}$$

where $\acute{\boldsymbol{u}}$ lies between $\mathbf{0}$ and $\boldsymbol{u}$. From Lemma 7, $\sqrt{s_n}\dot{M}^*_{P\alpha}(\hat{\boldsymbol{\theta}}_n)$ is stochastically bounded in conditional probability given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}$; from Lemma 8, conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}$, $\ddot{M}^*_{P\alpha}(\hat{\boldsymbol{\theta}}_n + \acute{\boldsymbol{u}}/\sqrt{s_n}) - \ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = o_P(1)$ and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ converges to a positive-definite matrix. Thus, from the Basic Corollary in page 2 of Hjort & Pollard (2011), the minimizer of $s_n\gamma(\boldsymbol{u})$, $\sqrt{s_n}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,P} - \hat{\boldsymbol{\theta}}_n)$, satisfies that

$$\sqrt{s_n}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,P} - \hat{\boldsymbol{\theta}}_n) = \ddot{M}_n^{-1}(\hat{\boldsymbol{\theta}}_n)\sqrt{s_n}\dot{M}^*_P(\hat{\boldsymbol{\theta}}_n) + o_P(1), \tag{A.67}$$

which implies that

$$\sqrt{s_n}\{\Lambda^{\alpha}_{n,P}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\}^{-1/2}\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,P} - \hat{\boldsymbol{\theta}}_n) \to \mathbb{N}(\mathbf{0}, \boldsymbol{I}), \tag{A.68}$$

in conditional distribution given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}$.

Next, we will check the distance between $\Lambda^{\alpha}_{n,P}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})$ and $\Lambda^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)$. Let $\Lambda^{\alpha}_{P\varrho_n}(\hat{\boldsymbol{\theta}}_n)$ have the same expression as $\Lambda^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)$ in (26) except that $\pi^{\rm opt}_{n,Pi}(\hat{\boldsymbol{\theta}}_n)$ in the denominator is replaced by

$$\pi^{\varrho_n}_{n,P\alpha i}(\hat{\boldsymbol{\theta}}_n) = (1-\alpha)\pi^{\varrho_n}_{n,Pi}(\hat{\boldsymbol{\theta}}_n) + \alpha\frac{1}{n} \quad \text{with} \quad \pi^{\varrho_n}_{n,Pi} = \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n}}{\sum_{j=1}^n \{\|\dot{m}(Z_j,\hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n}\}}.$$

We have that

$$\begin{aligned}
\|\Lambda^{\alpha}_{n,P}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}) - \Lambda^{\alpha}_{P\varrho_n}(\hat{\boldsymbol{\theta}}_n)\| &\le \frac{s_n}{n^2}\sum_{i=1}^n \left| \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{\{s_n\tilde{\pi}^{\rm opt}_{n,P\alpha i}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\} \wedge 1} - \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{\{s_n\pi^{\varrho_n}_{n,P\alpha i}(\hat{\boldsymbol{\theta}}_n)\} \wedge 1} \right| \\
&= \frac{s_n}{n^2}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2 \left| \frac{1}{\{s_n\tilde{\pi}^{\rm opt}_{n,P\alpha i}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\} \wedge 1} - \frac{1}{\{s_n\pi^{\varrho_n}_{n,P\alpha i}(\hat{\boldsymbol{\theta}}_n)\} \wedge 1} \right| \\
&= \frac{s_n}{n^2}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2 \left| \frac{\{s_n\tilde{\pi}^{\rm opt}_{n,P\alpha i}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\} \wedge 1 - \{s_n\pi^{\varrho_n}_{n,P\alpha i}(\hat{\boldsymbol{\theta}}_n)\} \wedge 1}{[\{s_n\tilde{\pi}^{\rm opt}_{n,P\alpha i}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\} \wedge 1][\{s_n\pi^{\varrho_n}_{n,P\alpha i}(\hat{\boldsymbol{\theta}}_n)\} \wedge 1]} \right| \\
&< \frac{1}{\alpha^2}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2 \left| \tilde{\pi}^{\rm opt}_{n,Pi}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}) - \pi^{\varrho_n}_{n,Pi}(\hat{\boldsymbol{\theta}}_n) \right| \qquad \text{(A.69)}
\end{aligned}$$

If $\varrho > 0$, then from

$$\begin{aligned}
& n\left| \tilde{\pi}^{\rm opt}_{n,Pi}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}) - \pi^{\varrho_n}_{n,Pi}(\hat{\boldsymbol{\theta}}_n) \right| \\
&= \left| \frac{\|\dot{m}(Z_i,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| \wedge H^{0*}}{\Psi^{0*}_{\varrho_n}} - \frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n}}{\Psi_{\varrho_n}} \right| \\
&\le \left| \frac{\|\dot{m}(Z_i,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| \wedge H^{0*} - \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n}}{\Psi^{0*}_{\varrho_n}} \right| + \{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n}\} \left| \frac{\Psi^{0*}_{\varrho_n} - \Psi_{\varrho_n}}{\Psi^{0*}_{\varrho_n}\Psi_{\varrho_n}} \right| \\
&\le \frac{\left| \|\dot{m}(Z_i,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| - \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| \right|}{\Psi^{0*}_{\varrho_n}} + \frac{\left|H^{0*} - H_{\varrho_n}\right|}{\Psi^{0*}_{\varrho_n}} + \{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n}\}\frac{\left|\Psi^{0*}_{\varrho_n} - \Psi_{\varrho_n}\right|}{\Psi^{0*}_{\varrho_n}\Psi_{\varrho_n}},
\end{aligned}$$

we have that

$$\begin{aligned}
& \|\Lambda^{\alpha}_{n,P}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}) - \Lambda^{\alpha}_{P\varrho_n}(\hat{\boldsymbol{\theta}}_n)\| \\
&< \frac{1}{\alpha^2}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2 \left| \tilde{\pi}^{\rm opt}_{n,Pi}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}) - \pi^{\varrho_n}_{n,Pi}(\hat{\boldsymbol{\theta}}_n) \right| \\
&\le \frac{1}{\alpha^2\Psi^{0*}_{\varrho_n}}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2 \left| \|\dot{m}(Z_i,\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| - \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| \right| \\
&\quad + \frac{\left|H^{0*} - H_{\varrho_n}\right|}{\alpha^2\Psi^{0*}_{\varrho_n}}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2 + \frac{\left|\Psi^{0*}_{\varrho_n} - \Psi_{\varrho_n}\right|}{\alpha^2\Psi^{0*}_{\varrho_n}\Psi_{\varrho_n}}\sum_{i=1}^n \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^3 = o_P(1),
\end{aligned}$$

by (A.52) and Lemma 9.

If $\varrho = 0$, then,

$$
\begin{aligned}
& n \left| \tilde{\pi}^{\text{opt}}_{n,Pi}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}) - \pi^{\varrho_n}_{n,Pi}(\hat{\boldsymbol{\theta}}_n) \right| \\
& \leq \frac{\left| \|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| \wedge H^{0*} - \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \wedge H_{\varrho_n} \right|}{\Psi^{0*}_{\varrho_n}} + \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \left| \frac{\Psi^{0*}_{\varrho_n} - \Psi_{\varrho_n}}{\Psi^{0*}_{\varrho_n} \Psi_{\varrho_n}} \right| \\
& \leq \frac{\left| \|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| - \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \right|}{\Psi^{0*}_{\varrho_n}} + \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \frac{\left| \Psi^{0*}_{\varrho_n} - \Psi_{\varrho_n} \right|}{\Psi^{0*}_{\varrho_n} \Psi_{\varrho_n}} \\
& \quad + \frac{\|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|}{\Psi^{0*}_{\varrho_n}} I\Big\{ \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \geq H_{\varrho_n} \Big\} + \frac{H_{\varrho_n}}{\Psi^{0*}_{\varrho_n}} I\Big\{ \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \geq H_{\varrho_n} \Big\} \\
& \quad + \frac{\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|}{\Psi^{0*}_{\varrho_n}} I\Big\{ \|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| \geq H^{0*} \Big\} + \frac{H^{0*}}{\Psi^{0*}_{\varrho_n}} I\Big\{ \|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| \geq H^{0*} \Big\} \\
& \equiv \Delta_{3i} + \Delta_{4i} + \Delta_{5i} + \Delta_{6i} + \Delta_{7i} + \Delta_{8i}. \qquad \text{(A.70)}
\end{aligned}
$$

From (A.52) and Lemma 9, we know that

$$
\frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^2 \Delta_{3i} = o_P(1) \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^2 \Delta_{4i} = o_P(1). \qquad \text{(A.71)}
$$

Note that

$$
\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^2 \|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\| I\Big\{ \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \geq H_{\varrho_n} \Big\} \\
& \leq \left\{ \frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^4 \right\}^{\frac{1}{2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|^4 \right\}^{\frac{1}{4}} \left[\frac{1}{n} \sum_{i=1}^{n} I\Big\{ \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \geq H_{\varrho_n} \Big\} \right]^{\frac{1}{4}} \\
& = o_P(1),
\end{aligned}
$$

because

$$
\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{n} I\big\{ \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \geq H_{\varrho_n} \big\} = o_P(1), \quad \frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^4 = O_P(1), \\
& \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})\|^4 = O_P(1).
\end{aligned}
$$

Thus,

$$
\frac{1}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^2 \Delta_{5i} = o_P(1). \qquad \text{(A.72)}
$$

If $\|\dot{m}(Z, \boldsymbol{\theta})\|$ is bounded, then

$$
\frac{H_{\varrho_n}}{n} \sum_{i=1}^{n} \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|^2 I\Big\{ \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| \geq H_{\varrho_n} \Big\} \leq \frac{n - \lceil n(1-\varrho_n) \rceil}{n} \|\dot{m}(Z, \hat{\boldsymbol{\theta}}_n)\|^3_{(n)} = o_P(1);
$$

otherwise

$$\frac{H_{\varrho_n}}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2 I\Big\{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\geq H_{\varrho_n}\Big\}\leq\frac{1}{nH_{\varrho_n}}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^4=o_P(1).$$

Thus we know that

$$\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\Delta_{6i}=o_P(1). \tag{A.73}$$

Similarly, we can obtain that

$$\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\Delta_{7i}=o_P(1)\quad\text{and}\quad\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\Delta_{8i}=o_P(1). \tag{A.74}$$

Combining (A.69), (A.70), (A.71), (A.72), (A.73), and (A.74), we know that

$$\|\Lambda^{\alpha}_{n,P}(\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P})-\Lambda^{\alpha}_{P\varrho_n}(\hat{\boldsymbol{\theta}}_n)\|=o_P(1).$$

To finish the proof for the case of $\varrho=0$, we only need to show that $\|\Lambda^{\alpha}_{P\varrho_n}(\hat{\boldsymbol{\theta}}_n)-\Lambda^{\alpha}_{R}(\hat{\boldsymbol{\theta}}_n)\|=o_P(1)$. Let $\Psi_\infty=\frac{1}{n}\sum_{i=1}^{n}\{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\}$. We notice that

$$\begin{aligned}
&n\left|\pi^{\varrho_n}_{n,Pi}(\hat{\boldsymbol{\theta}}_n)-\pi^{\text{opt}}_{n,Ri}(\hat{\boldsymbol{\theta}}_n)\right|\\
&\leq\frac{\left|\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\wedge H_{\varrho_n}-\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\right|}{\Psi_{\varrho_n}}+\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\left|\frac{\Psi_{\varrho_n}-\Psi_\infty}{\Psi_{\varrho_n}\Psi_\infty}\right|\\
&\leq\frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|}{\Psi_{\varrho_n}}I\Big\{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\geq H_{\varrho_n}\Big\}+\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\frac{\left|\Psi_{\varrho_n}-\Psi_\infty\right|}{\Psi_{\varrho_n}\Psi_\infty}\\
&\leq\frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{\Psi_{\varrho_n}H_{\varrho_n}}+\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\frac{\left|\Psi_{\varrho_n}-\Psi_\infty\right|}{\Psi_{\varrho_n}\Psi_\infty}\equiv\Delta_{9i}+\Delta_{10i}.
\end{aligned} \tag{A.75}$$

With this result, it can be shown that

$$\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\Delta_{9i}=o_P(1)\quad\text{and}\qquad\frac{1}{n}\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2\Delta_{10i}=o_P(1),$$

which indicates that $\|\Lambda^{\alpha}_{P\varrho_n}(\hat{\boldsymbol{\theta}}_n)-\Lambda^{\alpha}_{R}(\hat{\boldsymbol{\theta}}_n)\|=o_P(1)$.

From Slutsky's theorem, we know that given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\theta}}^{0*}_{s_0,P}$, as s_0, s_n, and n go to infinity,

$$\sqrt{s_n}\{V^{\alpha}_{n,P}(\hat{\boldsymbol{\theta}}_n)\}^{-1/2}(\tilde{\boldsymbol{\theta}}^{\alpha}_{s_n,P}-\hat{\boldsymbol{\theta}}_n)\to\mathbb{N}\left(\mathbf{0},\boldsymbol{I}\right),$$

in conditional distribution. □

Proof of Remark 11. Since $\Lambda_R^{\text{opt}}(\hat{\boldsymbol{\theta}}_n)$ has the minimum trace among all choices of sampling probabilities, if $\alpha \neq 0$ then $\text{tr}\{\Lambda_R^{\text{opt}}(\hat{\boldsymbol{\theta}}_n)\} < \text{tr}\{\Lambda_R^{\alpha}(\hat{\boldsymbol{\theta}}_n)\}$. On the other hand,

$$\text{tr}\{\Lambda_R^{\alpha}(\hat{\boldsymbol{\theta}}_n)\} = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{(1-\alpha)\pi_{n,i}^{R\text{opt}}+\alpha\frac{1}{n}} < \frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|^2}{(1-\alpha)\pi_{n,i}^{R\text{opt}}}$$
$$= \frac{1}{(1-\alpha)n^2}\left\{\sum_{i=1}^{n}\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|\right\}^2 = \frac{\text{tr}_{opt}\{\Lambda_{n,R}(\hat{\boldsymbol{\theta}}_n)\}}{1-\alpha},$$

and this finishes the proof for $\Lambda_R^{\alpha}(\hat{\boldsymbol{\theta}}_n)$ from subsampling with replacement. For $\Lambda_{n,P}^{\alpha}(\hat{\boldsymbol{\theta}}_n)$ from Poisson subsampling, the proof is similar.

□

A.2 Additional examples on optimal structural results

Example 5 (Least-squares)**.** Consider least-squares estimator

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}}\sum_{i=1}^{n}\{y_i - g(\boldsymbol{x}_i,\boldsymbol{\theta})\}^2,$$

where y_i is the response, $\boldsymbol{x}_i$ is the covariate, and $g(\boldsymbol{x}_i,\boldsymbol{\theta})$ is a smooth function. The least-squares estimator of $\boldsymbol{\theta}$ can be presented in our framework by letting $Z_i = (\boldsymbol{x}_i, y_i)$ and defining

$$m(Z_i,\boldsymbol{\theta}) = -0.5\{y_i - g(\boldsymbol{x}_i,\boldsymbol{\theta})\}^2.$$

From direct calculation, we have

$$\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n) = \hat{\varepsilon}_i\dot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n), \quad \text{and} \quad \ddot{m}(Z_i,\hat{\boldsymbol{\theta}}_n) = \hat{\varepsilon}_i\ddot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n) - \dot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)\dot{g}^{\text{T}}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n), \quad \text{(A.76)}$$

where $\hat{\varepsilon}_i = y_i - g(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)$, $\dot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)$ and $\ddot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)$ are the gradient and Hessian matrix of $g(\boldsymbol{x}_i,\boldsymbol{\theta})$, respectively, evaluated at $\hat{\boldsymbol{\theta}}_n$. Note that $\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i\ddot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)$ is a small term, so there is no need to calculate the Hessian matrix $\ddot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)$, and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ can be replaced by

$$\ddot{M}_n^a(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{i=1}^{n}\dot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)\dot{g}^{\text{T}}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n). \quad \text{(A.77)}$$

From (A.76) and (A.77), we obtain optimal sampling probabilities by using

$$\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\| = |\hat{\varepsilon}_i|\|\dot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)\|, \quad \text{or} \quad \|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|_L = n|\hat{\varepsilon}_i|\left\|L\{\ddot{M}_n^a(\hat{\boldsymbol{\theta}}_n)\}^{-1}\dot{g}(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}_n)\right\|, \quad \text{(A.78)}$$

to replace $\|\dot{m}(Z_i,\hat{\boldsymbol{\theta}}_n)\|$ in Theorems 3 and 4 for different subsampling procedures.

Specifically for ordinary least-squares (OLS) in linear regression, $g(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}$, $\dot{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_n) = \boldsymbol{x}_i$, and $\ddot{g}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_n) = \mathbf{0}$. Therefore, the expression in (A.78) is simplified to

$$\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| = |\hat{\varepsilon}_i|\|\boldsymbol{x}_i\|, \quad \text{or} \quad \|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\|_L = n|\hat{\varepsilon}_i|\big\|L(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{x}_i\big\|, \tag{A.79}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^{\mathrm{T}}$.

With $\|\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)\| = |\hat{\varepsilon}_i|\|\boldsymbol{x}_i\|$ inserted into (7), the sampling probabilities reduce to gradient-based sampling probabilities (Zhu 2016). Furthermore, if we take $L = \{-n\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)\}^{1/2} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{1/2}$ in (A.79), the optimal probabilities for subsampling with replacement satisfy that

$$\pi_{n,Ri}^{\mathrm{opt}} \propto |\hat{\varepsilon}_i|\sqrt{h_i}, \quad i = 1, ..., n, \tag{A.80}$$

where h_i's are statistical leverage scores of $\boldsymbol{x}_i$'s, i.e., diagonal elements of $\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$. This clearly shows the connection between leverage scores and the L optimality.

Form (A.80) and Theorem 4, optimal probabilities for Poisson subsampling and subsampling with replacement differ if there are data points such that $\frac{s_n}{n}|\hat{\varepsilon}_i|\sqrt{h_i} > \frac{1}{n}\sum_{j=1}^{n}|\hat{\varepsilon}_j|\sqrt{h_j}$. This is more likely to happen if $|\hat{\varepsilon}_i|$'s or $\sqrt{h_i}$'s are more nonuniform. Yang et al. (2015) showed that if statistical leverage scores are very nonuniform, then using the square roots of statistical leverage scores to construct subsampling probabilities yields better approximation than using the original leverage scores. An intuitive explanation for their conclusion is that taking score roots on leverage scores has some shrinkage effect on the resulting probabilities toward the uniform subsampling probability. Our results echos their conclusion, and further indicates that for optimal Poisson subsampling it may be necessary to perform truncation for high leverage scores.

Example 6 (Generalized linear models)**.** Let y_i be the response and $\boldsymbol{x}_i$ be the corresponding covariate. A generalized linear model (GLM) assumes that the conditional mean of the response y_i given the covariate $\boldsymbol{x}_i$, $\mathbb{E}(y_i|\boldsymbol{x}_i)$, satisfies

$$g\{\mathbb{E}(y_i|\boldsymbol{x}_i)\} = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta},$$

where g is the link function, $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ is the linear predictor, and $\boldsymbol{\beta}$ is the regression coefficient. For most of the commonly used GLMs, it is assumed that the distribution of the response y_i given the covariate $\boldsymbol{x}_i$ belongs to the exponential family, namely,

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \phi) = a(y_i, \phi)\exp\Big[\frac{y_i b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\phi}\Big],$$

where a, b and c are known scalar functions, and ϕ is the dispersion parameter. In the framework of GLM. If the link function g is selected such that b is the identity function, i.e.,

$b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, then the link function is called the canonical link. With a canonical link function, $g\{\mathbb{E}(y_i|\boldsymbol{x}_i)\} = c'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ where c' is the derivative function of c.

Let $Z_i = (\boldsymbol{x}_i, y_i)$. If both the regression coefficient $\boldsymbol{\beta}$ and the dispersion parameter ϕ are of interest, then let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \phi)^{\mathrm{T}}$. The MLE of $\boldsymbol{\theta}$ corresponds to

$$m(Z_i, \boldsymbol{\theta}) = \frac{y_i b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\phi} + \log\{a(y_i, \phi)\}.$$

If $\boldsymbol{\beta}$ is the only parameter of interest, then $\boldsymbol{\theta} = \boldsymbol{\beta}$, and the MLE of $\boldsymbol{\theta}$ corresponds to

$$m(Z_i, \boldsymbol{\theta}) = y_i b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}).$$

For this case, direct calculations give us that

$$\dot{m}(Z_i, \boldsymbol{\theta}) = \{y_i b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - c'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}\boldsymbol{x}_i \text{ and } \ddot{m}(Z_i, \boldsymbol{\theta}) = \{y_i b''(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - c''(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \quad \text{(A.81)}$$

where b' and b'' are the first and second derivative functions of b, and and c'' is the second derivative function of c. Thus, optimal sampling probabilities under the L-optimality can be obtained by using the expressions in (A.81) for Theorems 3 and 4. If the canonical link is used, then the expressions in (A.81) simplify to

$$\dot{m}(Z_i, \boldsymbol{\theta}) = \{y_i - c'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\}\boldsymbol{x}_i \quad \text{and} \quad \ddot{m}(Z_i, \boldsymbol{\theta}) = -c''(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}.$$

The following list gives the forms of $m(Z_i, \boldsymbol{\theta})$, $\dot{m}(Z_i, \boldsymbol{\theta})$, and $\ddot{m}(Z_i, \boldsymbol{\theta})$ for commonly used GLMs with the canonical links.

- **Normal distribution**, $y_i|\boldsymbol{x}_i \sim \mathbb{N}(\mu_i, \sigma^2)$.
  - Canonical link: $g(\mu_i) = \mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$.
  - Parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$:
    - $m(Z_i, \boldsymbol{\theta}) = \frac{-(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2}{2\sigma^2} - \frac{\log(\sigma^2)}{2}$.
    - $\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \hat{\varepsilon}_i\boldsymbol{x}_i \\ \frac{\hat{\varepsilon}_i^2 - \hat{\sigma}^2}{2\hat{\sigma}^2} \end{bmatrix}$, and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = \frac{-1}{n\hat{\sigma}^2}\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\hat{\sigma}^2} \end{bmatrix}$, where $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^{\mathrm{T}}$, $\hat{\varepsilon}_i = y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2$.
  - Parameter $\boldsymbol{\theta} = \boldsymbol{\beta}^{\mathrm{T}}$ when σ^2 is not of interest:
    - $m(Z_i, \boldsymbol{\theta}) = -(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2$.
    - $\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n)$ and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n)$ are the same to case of OLS in Example 5.
- **Binomial distribution**, $y_i|\boldsymbol{x}_i \sim \mathbb{BIN}(k_i, p_i)$. The problem is often converted to model the ratio $y_i^r = y_i/k_i$.

– Canonical link: $g(p_i) = \log(\frac{p_i}{1-p_i}) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$.

– Parameter $\boldsymbol{\theta} = \boldsymbol{\beta}$:

∗ $m(Z_i, \boldsymbol{\theta}) = k_i\{y_i^r \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}})\}$.

∗ $\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = k_i(y_i^r - \hat{p}_i)\boldsymbol{x}_i$, and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{i=1}^n k_i\hat{p}_i(1-\hat{p}_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$, where $\hat{p}_i = e^{\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}/(1 + e^{\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}})$.

If $k_i = 1$ for all i, the results reduce to the case of logistic regression in Example 1.

- **Poisson distribution**, $y_i|\boldsymbol{x}_i \sim \mathbb{POI}(\mu_i)$.

– Canonical link: $g(\mu_i) = \log(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$.

– Parameter $\boldsymbol{\theta} = \boldsymbol{\beta}$:

∗ $m(Z_i, \boldsymbol{\theta}) = y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}$.

∗ $\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = (y_i - e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}})\boldsymbol{x}_i$, and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{i=1}^n e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$

- **Gamma distribution**, $y_i|\boldsymbol{x}_i \sim \mathbb{GAM}(\nu, \mu_i)$, with density function

$$f(y_i) = \frac{\nu^\nu}{\Gamma(\nu)\mu_i^\nu}\, y_i^{\nu-1}\, e^{-\frac{\nu y_i}{\mu_i}}, \quad y_i > 0, \tag{A.82}$$

where ν is the shape parameter and μ_i is the mean parameter.[3]

– Canonical link: $g(\mu_i) = \frac{-1}{\mu_i} = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$.

– Parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \nu)^{\mathrm{T}}$:

∗ $m(Z_i, \boldsymbol{\theta}) = \nu y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \nu\log(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \nu\log\nu + (\nu-1)\log(y_i) - \log\{\Gamma(\nu)\}$.

∗ $\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = \begin{bmatrix} \hat{\nu}\Big(y_i + \frac{1}{\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}\Big)\boldsymbol{x}_i \\ y_i\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \log(-\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + \log(\hat{\nu}) + 1 + \log(y_i) - \frac{\Gamma'(\hat{\nu})}{\Gamma(\hat{\nu})} \end{bmatrix}$, and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = \begin{bmatrix} -\frac{\hat{\nu}}{n}\sum_{i=1}^n \frac{1}{(\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\hat{\nu}} - \frac{\Gamma''(\hat{\nu})\Gamma(\hat{\nu}) - \{\Gamma''(\hat{\nu})\}^2}{\{\Gamma(\hat{\nu})\}^2} \end{bmatrix}$, where $\Gamma'(\hat{\nu})$ and $\Gamma''(\hat{\nu})$ are the first and second derivative of $\Gamma(\nu)$ evaluated at $\hat{\nu}$.

– Parameter $\boldsymbol{\theta} = \boldsymbol{\beta}$:

∗ $m(Z_i, \boldsymbol{\theta}) = y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \log(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$.

∗ $\dot{m}(Z_i, \hat{\boldsymbol{\theta}}_n) = \Big(y_i + \frac{1}{\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}\Big)\boldsymbol{x}_i$, and $\ddot{M}_n(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{i=1}^n \frac{1}{(\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$.

[3] A Gamma distribution is also often parameterized in terms of the shape and rate parameters or the shape and scale parameters. With our notations here, the shape and rate parameters are ν and ν/μ_i, respectively, and the shape and scale parameters are ν and μ_i/ν, respectively.

If $\nu = 1$ in (A.82), then the Gamma distribution reduces to an exponential, and thus the results reduce to the case of exponential regression. For inverse Gamma distribution, one can use the reciprocal transformation, i.e., $1/y_i$, to convert the problem to Gamma distribution.

References

Atkinson, A., Donev, A. & Tobias, R. (2007), *Optimum experimental designs, with SAS*, Vol. 34, Oxford University Press.

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

Bickel, P., Gotze, F. & van Zwet, W. (1997), 'Resampling fewer than n observations: gains, losses, and remedies for losses', *Statistica Sinica* **7**, 1–31.

Campbell, T. & Broderick, T. (2018), Bayesian coreset construction via greedy iterative geodesic ascent, *in* 'International Conference on Machine Learning', PMLR, pp. 698–706.

Campbell, T. & Broderick, T. (2019), 'Automated scalable bayesian inference via hilbert coresets', *The Journal of Machine Learning Research* **20**(1), 551–588.

Chanda, K. (1971), 'Asymptotic distribution of sample quantiles for exchangeable random variables', *Calcutta Statistical Association Bulletin* **20**(4), 135–142.

Cheng, G. & Huang, J. (2010), 'Bootstrap consistency for general semiparametric M-estimation', *The Annals of Statistics* **38**(5), 2884–2915.

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

Dheeru, D. & Karra Taniskidou, E. (2017), 'UCI machine learning repository'.
URL: *http://archive.ics.uci.edu/ml*

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

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

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

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

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

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

Fonollosa, J., Sheik, S., Huerta, R. & Marco, S. (2015), 'Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring', *Sensors and Actuators B: Chemical* **215**, 618–629.

Hesterberg, T. (1995), 'Weighted average importance sampling and defensive mixture distributions', *Technometrics* **37**(2), 185–194.

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

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

Lin, N. & Xie, R. (2011), 'Aggregated estimating equation estimation', *Statistics and Its Interface* **4**, 73–83.

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

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

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

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

Nocedal, J. & Wright, S. J. (1999), *Numerical Optimization*, Springer.

Owen, A. & Zhou, Y. (2000), 'Safe and effective importance sampling', *Journal of the American Statistical Association* **95**(449), 135–143.

Politis, D., Romano, J. & Wolf, M. (1999), *Subsampling*, Springer-Verlag, New York.

Præstgaard, J. & Wellner, J. A. (1993), 'Exchangeably weighted bootstraps of the general empirical process', *The Annals of Probability* pp. 2053–2086.

Särndal, C.-E., Swensson, B. & Wretman, J. (2003), *Model assisted survey sampling*, Springer Science & Business Media.

Schifano, E. D., Wu, J., Wang, C., Yan, J. & Chen, M.-H. (2016), 'Online updating of statistical inference in the big data setting', *Technometrics* **58**(3), 393–403.

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

Shao, J. & Tu, D. (1995), *The jackknife and bootstrap*, Springer-Verlag, New York.

Ting, D. & Brochu, E. (2018), Optimal subsampling with influence functions, *in* 'Advances in Neural Information Processing Systems 31', Curran Associates, Inc., pp. 3654–3663.

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

Wang, H. (2019), 'More efficient estimation for logistic regression with optimal subsamples', *Journal of Machine Learning Research* **20**(132), 1–59.

Wang, H. & Ma, Y. (2021), 'Optimal subsampling for quantile regression in big data', *Biometrika* **108**(1), 99–112.

Wang, H., Yang, M. & Stufken, J. (2019), 'Information-based optimal subdata selection for big data linear regression', *Journal of the American Statistical Association* **114**(525), 393–405.

Wang, H., Zhu, R. & Ma, P. (2018), 'Optimal subsampling for large sample logistic regression', *Journal of the American Statistical Association* **113**(522), 829–844.

Wang, H. & Zou, J. (2021), A comparative study on sampling with replacement vs poisson sampling in optimal subsampling, *in* A. Banerjee & K. Fukumizu, eds, 'Proceedings of The 24th International Conference on Artificial Intelligence and Statistics', Vol. 130 of *Proceedings of Machine Learning Research*, PMLR, pp. 289–297.
URL: *http://proceedings.mlr.press/v130/wang21a.html*

Woodruff, D. P. et al. (2014), 'Sketching as a tool for numerical linear algebra', *Foundations and Trends® in Theoretical Computer Science* **10**(1–2), 1–157.

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

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

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

Zhu, R. (2016), Gradient-based sampling: An adaptive importance sampling for least-squares, *in* D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon & R. Garnett, eds, 'Advances in Neural Information Processing Systems 29', Curran Associates, Inc., pp. 406–414.