# Maximum sampled conditional likelihood for informative subsampling

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

**Jae Kwang Kim**                                         JKIM@IASTATE.EDU
*Department of Statistics*
*Iowa State University*
*Ames, IA 50011, USA*

**Editor:** Tong Zhang

## Abstract

Subsampling is a computationally effective approach to extract information from massive data sets when computing resources are limited. After a subsample is taken from the full data, most available methods use an inverse probability weighted (IPW) objective function to estimate the model parameters. The IPW estimator does not fully utilize the information in the selected subsample. In this paper, we propose to use the maximum sampled conditional likelihood estimator (MSCLE) based on the sampled data. We established the asymptotic normality of the MSCLE and prove that its asymptotic variance covariance matrix is the smallest among a class of asymptotically unbiased estimators, including the IPW estimator. We further discuss the asymptotic results with the L-optimal subsampling probabilities and illustrate the estimation procedure with generalized linear models. Numerical experiments are provided to evaluate the practical performance of the proposed method.

**Keywords:** Asymptotic Distribution; Bias Correction; Estimation Efficiency; Lower Bound of Variance; Informative Subsampling

## 1. Introduction

In the era of big data, many data sets have huge volumes. If the data sets are too voluminous then traditional data processing software products are not capable of processing the data within a reasonable amount of time. In this case, a subsample or coreset of the full data is often used to alleviate the computational burden. Subsampling is an emerging area of research that balances the trade-off between computational efficiency and statistical efficiency by developing an efficient subsampling design and estimation strategy. For this purpose, existing research focuses more on designing the subsampling probabilities and less on improving the estimator based on the selected subsample, e.g., Drineas et al. (2006); Yang et al. (2015); Wang et al. (2018), among others.

In the linear regression model setup, optimal subsampling designs are well studied in the literature. Specifically, statistical leverage scores or their variants are often recommended to construct subsampling probabilities, see Drineas et al. (2006); Dhillon et al. (2013);

McWilliams et al. (2014); Ma et al. (2015); Yang et al. (2015); Nie et al. (2018), and the references therein. Instead of calculating exact leverage scores directly on the full data, Drineas et al. (2012) proposed fast algorithms to approximate them. The aforementioned sampling probabilities are not dependent on the response variable, and this type of sampling schemes is referred to as non-informative subsampling. For this scenario, Wang et al. (2019) proposed a deterministic selection algorithm that has high estimation efficiency.

Beyond linear regressions, Wang et al. (2018) proposed an optimal subsampling method under the A-optimality criterion for logistic regression, which defines subsampling probabilities that minimize the asymptotic mean squared error of the resulting subsample estimator. They further considered the L-optimality to further improve the computational efficiency. This method has been extended to include other models such as generalized linear models (Ai et al., 2021), quantile regressions (Wang and Ma, 2021), quasi-likelihood models (Yu et al., 2022), and etc. Ting and Brochu (2018) suggested using the influence function to define optimal probabilities. Shen et al. (2021) proposed the surprise sampling method that gives optimal forms to a variety of objectives. Wang et al. (2022) systematically compared with-replacement sampling with Poisson sampling and recommended Poisson sampling for its higher estimation efficiency and computational feasibility. Readers are refered to Yao and Wang (2021) for a systematic review on this topic.

Unlike non-informative subsampling, the optimal subsampling probabilities depend on the response variable as well as the covariates. If the sampling probabilities depend on the response variable in addition to the auxiliary (covariate) variable, the sampling mechanism is called informative (Pfeffermann et al., 1998). The selection probabilities in the informative sampling utilize information in both the covariates and the responses, so the resulting subsample often contain more relevant information compared with non-informative sampling. Under informative sampling, the selection probabilities are often inversely applied to obtain the IPW estimator (Chambers and Skinner, 2003). However, the inverse probability weighting scheme may not achieve efficient estimation.

Generally speaking, the informative subsampling can be viewed as a biased sampling problem in statistics, as discussed in Cox (1969) and Qin (2017). To understand the biased sampling problem, it is useful to consider selection bias in the context of two-phases of sampling. In the first phase, we have a random sample of size $N$. In the second phase, we select a subset of the original sample with known selection probability $\pi(\boldsymbol{x}, y)$ which is a function of the observations in the first-phase sample. If the selection probability depends on the outcome variable $y$, it is also called outcome-dependent (two-phase) sampling. The two-phase sampling design is commonly used in many disciplines. Kim et al. (2006) and Saegusa and Wellner (2013) developed some theory for two-phase sampling. Some examples of two-phase sampling can be found in Hsieh et al. (1985), Kalbfleisch and Lawless (1988), Wild (1991), Scott and Wild (1991), Hu and Lawless (1996), Scott and Wild (1997), Hu and Lawless (1997), Breslow and Holubkov (1997), and Whittemore (1997). The case-control study is a popular example of the outcome-dependent two-phase sampling. If the outcome is binary and the case with $y = 1$ is rare, it is sensible to oversample the cases with $y = 1$ in the final sample. Such outcome-dependent two-phase sampling is used because it is either more efficient or cost effective. Removing or reducing the selection bias in such sampling is a crucial part of the estimation problem.

For logistic regressions with case-control sampling, Scott and Wild (1986) showed that the bias for the unweighted estimator only appears in the intercept estimator and they provided the expression of the bias term. For extremely imbalanced data, Wang (2020) proved that with a case-control subsample, the IPW estimator and the unweighted estimator with bias correction have the same convergence rate and asymptotic distribution as the full data estimator if sufficient number of controls are selected; otherwise the latter is more efficient than the former. The case-control sampling probabilities depend on the responses only and do not utilize the information in the covariates. For binary logistic regression, Fithian and Hastie (2014) proposed the local case-control subsampling probabilities which depend on both the responses and the covariates. They showed that the bias in the regression coefficient estimator is corrected by the pilot estimator used to calculated the sampling probabilities. Wang (2019) extended this bias correction idea to the optimal subsampling probabilities under the A- and L-optimality, and proved that an unweighted estimator with bias correction has a higher estimation efficiency. For multi-class logistic regression, Han et al. (2020) developed the local uncertainty sampling (LUS) that focuses on correcting the unweighted objective function instead of correcting the bias in the resulting estimator. This approach is using the sampled data conditional likelihood and is not restricted by a specific form of the subsampling probabilities. Wang et al. (2021) adopted the idea and investigated optimal negative subsampling for the case of extremely imbalanced binary data.

The aforementioned investigations exclusively focus on logistic regression. In this paper, we will show that there is a general approach to extract more information from an informative subsample without using inverse probability weighting. The basic strategy is to treat the subsample estimation as a missing data problem and obtain the conditional likelihood of the sampled data, which is based on the conditional density function of the study variable given the covariate variable for the sampled data. The sampled conditional likelihood has been used in the context of biased sampling problem, but to our best knowledge, it has not been addressed in the subsampling area. The sampled conditional likelihood approach is applicable for general parametric models and sampling probabilities. The investigations of Fithian and Hastie (2014); Wang (2019); Han et al. (2020) are all specific cases of this general approach. We first establish the consistency and the asymptotic normality of the maximum sampled conditional likelihood estimator under some regularity conditions. Thus, statistical inference such as normal-based confidence intervals can be developed. We also show that the resulting estimator has the highest estimation efficiency among a class of asymptotically unbiased estimators, and it is more efficient than the IPW estimator. The maximum sampled conditional likelihood estimator can be computed by applying the Fisher-scoring algorithm with the closed-form formula for the Hessian matrix. Thus, the computation is relatively simple and fast. As illustrated in the simulation study in Section 6, the efficiency gains of using the sampled conditional likelihood over the IPW estimator are substantial.

If the subsampling probabilities are unknown, we may use an independent pilot sample to estimate the parameters in the subsampling probabilities. Because the subsampling probabilities are under our control, unlike the missing data problem, we can always use the correct subsampling probabilities in constructing the sampled conditional likelihood. Thus, even if the pilot samples are systematically different from the original sample, the statistical properties of the maximum sampled conditional likelihood estimator remain valid.

The rest of the paper is organized as follows. We present the proposed idea in Section 2 and discuss its asymptotic properties in Section 3. Section 4 considers the practical situation that informative subsampling probabilities depend on unknowns. Section 5 illustrates structural results using a version of widely used informative subsampling probabilities and Section 6 demonstrates the proposed method in the context of generalized linear models. Section 8 provides numerical results based on simulated and real data sets. Some concluding remarks are made in Section 8. Proofs, technical details, and additional numerical results are given in the appendix.

## 2. Sampled data conditional likelihood estimator

Let $(\boldsymbol{x}_i, y_i)$, $i = 1, ..., N$, be an independent sample from the distribution of $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is the covariate variable and $y$ is the main response variable. Denote the density function of $y_i$ given $\boldsymbol{x}_i$ as $f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter of interest, and it is often estimated by the maximum likelihood estimator (MLE),

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i), \tag{1}$$

where $\ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) = \log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$ is the log-likelihood function. For massive data, the computational cost in the maximization for $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ can be high, especially when there is no closed-form solution and an iterative algorithm has to be used. To solve this computational issue, approximation using a small subsample or a coreset of the data is regarded as an effective solution.

To emphasize the fact that informative subsampling probabilities depend on the responses, we denote them as $\pi(\boldsymbol{x}_i, y_i) = \mathbb{P}(\delta_i = 1 \mid \boldsymbol{x}_i, y_i) \in (0, 1]$, $i = 1, ..., N$. Here, $\delta_i$ is the indicator variable signifying if the $i$-th data point is selected, i.e., $\delta_i = 1$ if $(\boldsymbol{x}_i, y_i)$ is in the subsample and $\delta_i = 0$ otherwise. We assume that the distribution of $\delta_i$ is Bernoulli with parameter

$$\mathbb{P}(\delta_i = 1 \mid \boldsymbol{x}_i, y_i) = \pi(\boldsymbol{x}_i, y_i), \quad \text{for} \quad i = 1, ..., N. \tag{2}$$

A commonly used subsample estimator is based on a IPW objective function,

$$\hat{\boldsymbol{\theta}}_W = \arg \max_{\boldsymbol{\theta}} \ell_W(\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \frac{\delta_i \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)}{\pi(\boldsymbol{x}_i, y_i)}. \tag{3}$$

Here, the inverse probability weighting is necessary for $\hat{\boldsymbol{\theta}}_W$ to be consistent. An unweighted estimator is biased and inconsistent to neither the full data MLE nor the true parameter. However, the weighting scheme in (3) does not fully extract the information in the subsample as it down-weights more informative data points. Intuitively, one wants to assign a larger $\pi(\boldsymbol{x}_i, y_i)$ to a data point if it contains more information about $\boldsymbol{\theta}$ so that we sample it with a higher probability. If $\pi_i$'s are non-informative such as in the leverage-based sampling, then an unweighted estimator can still be consistent to the true parameter (e.g., Ma et al., 2015). Actually, for non-informative sampling, an unweighted estimator is the best estimator with the smallest asymptotic variance. If the subsampling mechanism is non-informative in the

sense that the subsampling probability satisfies $\pi(\boldsymbol{x}, y) = \pi(\boldsymbol{x})$, according to Rubin (1976), the MLE of $\boldsymbol{\theta}$ is obtained by maximizing the complete-case (CC) log-likelihood

$$\hat{\boldsymbol{\theta}}_{CC} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \delta_i \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i).$$

Thus, the IPW estimator in (3) is inefficient. However, non-informative subsampling probabilities may not be as effective as informative subsampling probabilities in identifying informative data points in the first place. Thus we focus on informative subsampling only.

To avoid the inverse probability weighting, we propose to use the sampled data conditional likelihood to obtain the subsample estimator. By Bayes' theorem the conditional density function of $y_i$ given $\boldsymbol{x}_i$ for sampled data is

$$f(y_i \mid \boldsymbol{x}_i, \delta_i = 1; \boldsymbol{\theta}) = \frac{f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{x}_i, y_i)}{\int f(y \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{x}_i, y)\mathrm{d}y}, \tag{4}$$

where $\mathrm{d}y$ is the Lebesgue measure for continuous responses and it is the counting measure for discrete responses.

The sampled data conditional log-likelihood function from (4) has a general form of

$$\ell_S(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) = \sum_{i=1}^{N} \delta_i \big[ \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \log\{\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})\} \big] + C, \tag{5}$$

where

$$\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbb{E}\{\pi(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i\} = \int f(y \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi(\boldsymbol{x}_i, y)\mathrm{d}y,$$

and $C = \sum_{i=1}^{N} \delta_i \log\{\pi(\boldsymbol{x}_i, y_i)\}$ does not contain $\boldsymbol{\theta}$. The proposed estimator $\hat{\boldsymbol{\theta}}_S$ is the maximizer of (5), namely,

$$\hat{\boldsymbol{\theta}}_S = \arg\max_{\boldsymbol{\theta}} \ell_S(\boldsymbol{\theta}). \tag{6}$$

Note that the density in (4) is not the joint density of $(\boldsymbol{x}_i, y_i)$ for sampled data; it is the conditional density of $y_i$ given $\boldsymbol{x}_i$ for sampled data. Therefore we call the corresponding likelihood function the conditional likelihood, and call our estimator the maximum sampled conditional likelihood estimator (MSCLE). If the subsampling probability $\pi(\boldsymbol{x}_i, y_i)$ depends on $\boldsymbol{x}_i$ only, then $\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta}) = \pi(\boldsymbol{x}_i)$ does not contain $\boldsymbol{\theta}$ and thus the MSCLE reduces to the MLE that maximizes the complete-case likelihood function.

Note that $\hat{\boldsymbol{\theta}}_W$ uses $\pi^{-1}(\boldsymbol{x}_i, y_i)$'s as weights while $\hat{\boldsymbol{\theta}}_S$ uses $\log\{\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})\}$ to correct the log-likelihood function. In computing $\hat{\boldsymbol{\theta}}_W$, if a data point with a very small value of $\pi(\boldsymbol{x}_i, y_i)$ is selected, then the objective function may be dominated by this data point. Although this scenario happens with a very small probability, the asymptotic variance of $\hat{\boldsymbol{\theta}}_W$ will be greatly inflated (Hesterberg, 1995; Owen and Zhou, 2000; Ma et al., 2015). However, if we use $\hat{\boldsymbol{\theta}}_S$, this problem will be ameliorated for the following two reasons. 1) Since $\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})$ is an weighted average of $\pi(\boldsymbol{x}_i, y_i)$ across the conditional distribution of $y_i$ given $\boldsymbol{x}_i$, we know that $\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})$'s are less variable than $\pi(\boldsymbol{x}_i, y_i)$'s. As a result, even when $\pi(\boldsymbol{x}_i, y_i)$ is very close to zero, $\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})$ can be bounded away from zero. 2) Even if $\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})$ and $\pi(\boldsymbol{x}_i, y_i)$ approach to zero at the same rate, $-\log\{\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})\}$ approaches to infinity much slower than

$\pi^{-1}(\boldsymbol{x}_i, y_i)$ does. Compared with $\hat{\boldsymbol{\theta}}_W$, $\hat{\boldsymbol{\theta}}_S$ is based on the conditional likelihood of the sampled data, and thus it has a higher estimation efficiency. The price to pay for using $\hat{\boldsymbol{\theta}}_S$ is that the integration in $\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})$ depends on the model structure, and $\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})$ may have complicated or no explicit expressions for some models. On the other hand, $\hat{\boldsymbol{\theta}}_W$ directly uses $\pi^{-1}(\boldsymbol{x}_i, y_i)$ and does not requires any extra model assumption and thus it is easier to implement. We will show in Section 6 that $\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})$'s have a closed-form solution for many popular generalized linear models, and Newton's method is applicable with closed-form expressions for the score functions and Hessian matrices.

Assuming that derivatives can pass integration, the score function associated with $\ell_S(\boldsymbol{\theta})$ is

$$\dot{\ell}_S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_S(\boldsymbol{\theta}) = \sum_{i=1}^N \delta_i \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) - \sum_{i=1}^N \delta_i \frac{\partial \bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})/\partial \boldsymbol{\theta}}{\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})},$$

where $\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) = \partial \log f(y \mid \boldsymbol{x}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ and

$$\frac{\partial}{\partial \boldsymbol{\theta}} \bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta}) = \int \frac{\partial}{\partial \boldsymbol{\theta}} f(y \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \pi(\boldsymbol{x}_i, y) dy = \mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\pi(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i\}. \tag{7}$$

Thus, we can express

$$\dot{\ell}_S(\boldsymbol{\theta}) = \sum_{i=1}^N \delta_i \left[ \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) - \mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i, \delta_i = 1\} \right], \tag{8}$$

where

$$\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i, \delta_i = 1\} = \frac{\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\pi(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i\}}{\bar{\pi}(\boldsymbol{x}_i; \boldsymbol{\theta})}. \tag{9}$$

The second term in (8) can be called the bias-adjustment term for the score function.

## 3. Asymptotic Results

We now presents some asymptotic results of the MSCLE of $\boldsymbol{\theta}$ proposed in (6). To do this, we need the following regularity assumptions. As a convention in this paper, we use $\dot{\ell}$ and $\ddot{\ell}$ to denote gradient vector (of the first derivatives) and Hessian matrix (of the second derivatives) of a function $\ell$ with respect to $\boldsymbol{\theta}$, respectively.

**Assumption 1** *Assume that $\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^2$ and $\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|$ are integrable, where $\|A\| = \mathrm{tr}^{1/2}(A^{\mathrm{T}}A)$ is the norm for a vector or matrix $A$. Here, $\ell(\boldsymbol{\theta}; \boldsymbol{x}, y) = \log f(y \mid \boldsymbol{x}, \boldsymbol{\theta})$ is the log-likelihood function of the original data distribution.*

**Assumption 2** *The parameter space $\boldsymbol{\Theta}$ is compact and the third order partial derivative of $\ell(\boldsymbol{\theta}; \boldsymbol{x}, y)$ and $\log\{\bar{\pi}(\boldsymbol{x}; \boldsymbol{\theta})\}$ with respect to any components of $\boldsymbol{\theta}$ is bounded in absolute value by an integrable function $B(\boldsymbol{x}, y)$ that does not depend on $\boldsymbol{\theta}$.*

**Assumption 3** *The matrix*

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \mathbb{E}\left[ \dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\pi(\boldsymbol{x}, y) - \frac{\mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\pi(\boldsymbol{x}, y) \mid \boldsymbol{x}\}}{\bar{\pi}(\boldsymbol{x}; \boldsymbol{\theta})} \right] \tag{10}$$

*is finite and positive definite, where $A^{\otimes 2} = AA^{\mathrm{T}}$ for a vector or matrix $A$.*

Compared with commonly used assumptions in maximum likelihood theory, the above Assumptions 2 and 3 impose additional constraints on the subsampling probability $\pi(\boldsymbol{x}, y)$. Assumption 2 imposes an integrable bound on $\log\{\bar{\pi}(\boldsymbol{x}; \boldsymbol{\theta})\}$ and its derivatives. This is to prevent the distribution of $\pi(\boldsymbol{x}, y)$ from having a large probability around zero. Since $\bar{\pi}(\boldsymbol{x}; \boldsymbol{\theta})$ is an weighted average of $\pi(\boldsymbol{x}, y)$ across the conditional distribution of $y$, it put less probability around the boundary zero, so the required condition here is less restrictive compared with those required by the IPW estimators for which $\pi(\boldsymbol{x}, y)$ is in the denominator. For the IPW estimator, it is often assumed that $\pi(\boldsymbol{x}, y)$ is bounded away from zero (e.g. Ai et al., 2021; Wang and Ma, 2021; Yu et al., 2022), which is a much stronger condition. If the subsampling probability is non-informative, i.e. $\pi(\boldsymbol{x}, y) = \pi(\boldsymbol{x})$, then $\bar{\pi}(\boldsymbol{x}; \boldsymbol{\theta})$ does not dependent on $\boldsymbol{\theta}$, and thus Assumption 2 reduces to the common condition that the third derivative of the log-likelihood is bounded in absolute value by an integrable random variable. Assumption 3 ensures that the variance covariance matrix for the sampled data conditional score function is positive definite. The key restriction here is the integrability of $\bar{\pi}^{-1}(\boldsymbol{x}; \boldsymbol{\theta})\mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\pi(\boldsymbol{x}, y) \mid \boldsymbol{x}\}$. Again, this is less restrictive compared with the conditions required by the IPW estimator, and a specific example will be provided in Section 5. For non-informative subsampling, a sufficient condition for Assumption 3 is that $\mathbb{E}\{\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\pi(\boldsymbol{x})\}$ is finite and positive definite.

We can also see that our conditions on the sampling probability in Assumptions 2 and 3 are less restrictive from another angle. These conditions imposes restrictions on the distribution of $\boldsymbol{x}$ only, while for IPW estimator with informative sampling probability the required conditions impose restrictions on the distribution of $y$ as well (e.g. Wang et al., 2018; Wang and Ma, 2021; Wang et al., 2022; Yu et al., 2022, etc).

**Theorem 1** *Let $\{(\boldsymbol{x}_i, y_i), i = 1, \cdots, N\}$ be $N$ independent realizations of $(X, Y)$ with joint density $f(y \mid \boldsymbol{x}; \boldsymbol{\theta})f_X(\boldsymbol{x})$ for some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $f_X(\boldsymbol{x})$ is completely unspecified. Let $\hat{\boldsymbol{\theta}}_S$ be the MSCLE of $\boldsymbol{\theta}$ defined in (6). Under Assumptions 1-3, as $N$ goes to infinity,*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}) \xrightarrow{D} \mathbb{N}(\boldsymbol{0}, \ \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}), \tag{11}$$

*where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is defined in (10), and $\xrightarrow{D}$ means convergence in distribution.*

In Theorem 1, the subsample size $n^* = \sum_{i=1}^{N} \delta_i$ is random, and the average subsample size $n = \mathbb{E}(n^*) = N\mathbb{E}\{\pi(\boldsymbol{x}, y)\}$ goes to infinity as $N \to \infty$.

The estimator $\hat{\boldsymbol{\theta}}_S$ is based on the conditional likelihood of the sampled data, so it is expected to be more efficient than the IPW estimator. Actually, it is the most efficient estimator in a class of asymptotically unbiased estimators. To see this, consider the following class of estimating equations

$$\sum_{i=1}^{N} \delta_i \boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) = \boldsymbol{0}, \tag{12}$$

where $\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ satisfies $\mathbb{E}\{\delta \boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}\} = \boldsymbol{0}$. Let $\hat{\boldsymbol{\theta}}_u$ be the class of estimators obtained through solving the class of estimating equations in (12). The class of estimators defined via solving (12) includes the IPW estimator and the MSCLE as special cases. If $\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y) = \pi^{-1}(\boldsymbol{x}, y)\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)$, then $\hat{\boldsymbol{\theta}}_u$ becomes the IPW estimator defined in (3); if $\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y) =$

$\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y) = \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1\}$, then $\hat{\boldsymbol{\theta}}_u$ becomes the MSCLE defined in (6).

The following theorem shows that the MSCLE is the most efficient among the class of estimators defined through solving (12).

**Theorem 2** *Assume that the partial derivatives of $\mathbb{E}\{\delta\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}\}$ with respect to $\boldsymbol{\theta}$ can be passed under the integration sign, and that the regularity conditions for the following standard asymptotic expansion holds.*

$$\hat{\boldsymbol{\theta}}_u = \boldsymbol{\theta} - \mathbf{M}_{\boldsymbol{\theta}}^{-1} N^{-1} \sum_{i=1}^{N} \delta_i \boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) + o_P(N^{-1/2}), \tag{13}$$

*where $\mathbf{M}_{\boldsymbol{\theta}} = \mathbb{E}\{\delta\dot{\boldsymbol{U}}(\boldsymbol{\theta}; \boldsymbol{x}, y)\}$ is full rank and $\dot{\boldsymbol{U}}(\boldsymbol{\theta}; \boldsymbol{x}, y) = \partial\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y)/\partial\boldsymbol{\theta}^{\mathrm{T}}$. Assume that $\mathbb{E}\{\delta\boldsymbol{U}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\}$ exists. The asymptotic variance covariance matrix of $\hat{\boldsymbol{\theta}}_u$ scaled by $N$ is $\mathbf{M}_{\boldsymbol{\theta}}^{-1}\mathbb{E}\{\delta\boldsymbol{U}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\}(\mathbf{M}_{\boldsymbol{\theta}}^{-1})^{\mathrm{T}}$, and it satisfies that*

$$\mathbf{M}_{\boldsymbol{\theta}}^{-1}\mathbb{E}\{\delta\boldsymbol{U}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\}(\mathbf{M}_{\boldsymbol{\theta}}^{-1})^{\mathrm{T}} \geq \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}, \tag{14}$$

*in the Loewner ordering for any $\boldsymbol{\theta}$, where the equality holds if $\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ is a linear function of the subsampled data conditional score function, namely $\boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y) = -\mathbf{M}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y)$.*

**Remark 3** *For the IPW estimator $\hat{\boldsymbol{\theta}}_W$ in (3), let $\mathbf{V}_W$ denote the asymptotic variance scaled by $N$, which typically has a form of (see Yu et al., 2022)*

$$\mathbf{V}_W = \mathbf{F}^{-1}\mathbb{E}\left\{\frac{\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)}{\pi(\boldsymbol{x}, y)}\right\}\mathbf{F}^{-1}, \tag{15}$$

*where $\mathbf{F} = \mathbb{E}\{\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\}$ is the Fisher information matrix of the original data distribution. From Theorem 2, $\mathbf{V}_W \geq \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$. Additionally, for $\mathbf{V}_W$ in (15), it requires $\pi^{-1}(\boldsymbol{x}, y)\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ to be integrable, which may be violated if $\pi(\boldsymbol{x}, y)$ has a high density in the neighborhood of zero. On the other hand, for $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$ in (10), it requires that $\bar{\pi}^{-1}(\boldsymbol{x}; \boldsymbol{\theta})\mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\pi(\boldsymbol{x}, y) \mid \boldsymbol{x}\}$ is integrable. It is the average probability $\bar{\pi}(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{E}\{\pi(\boldsymbol{x}, y) \mid \boldsymbol{x}\}$ that is in the denominator. The MSCLE is less restrictive compared with the IPW estimator because even if $\pi(\boldsymbol{x}, y)$ has a high density in the neighborhood of zero, $\bar{\pi}(\boldsymbol{x}; \boldsymbol{\theta})$ may not be small.*

## 4. Estimated subsampling probabilities

In practice, informative subsampling probabilities may depend on unknown parameters, say $\boldsymbol{\vartheta}$, and a pilot subsample is often used to estimate it. Here $\boldsymbol{\vartheta}$ may be the same as $\boldsymbol{\theta}$ or contain $\boldsymbol{\theta}$ as its components. We denote the pilot estimator of $\boldsymbol{\vartheta}$ as $\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}$, and assume that $\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}$ is independent of the data to sample and converges to a limit, i.e., $\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}} \xrightarrow{P} \boldsymbol{\vartheta}_p$. Here $\xrightarrow{P}$ means convergence in probability.

The assumption that a pilot estimator is independent of the data is commonly used in the literature (e.g., Fithian and Hastie, 2014; Han et al., 2020), and it is reasonable in the

context of subsampling. If one uses simple random sampling to obtain a pilot subsample or simply use the first certain number of observations as a pilot sample, then the pilot subsample is independent of the rest of the data. Thus, taking the rest of the data as the full data and performing informative subsampling, the pilot estimator is independent of the full data. Since one is likely to assign a larger pilot subsample size for a larger original data sample size, it is reasonable to assume that the pilot estimator converges to some limit. Here we do not have to assume that the pilot estimator converges to the true parameter, namely, the pilot estimator can be misspecified. As discussed in Section 1, because the subsampling probabilities are under our control, we can always use the realized subsampling probabilities in constructing the sampled likelihood. Thus, even if the pilot samples are systematically different from the original sample, we can still use the realized subsampling probabilities and the statistical properties of the maximum sampled conditional likelihood estimator remain valid. If the pilot sample is systematically different from the original sample, the pilot estimator may not be consistent to the true parameter and the resulting subsampling probabilities may not be optimal anymore. We will show numerically in Section 8 that the proposed MSCLE is more robust to pilot misspecification compared with the IPW estimator.

Another practical consideration is the subsample size. In Section 2, the average subsample size is $n = N\mathbb{E}\{\pi(\boldsymbol{x}, y)\}$ which is the same order of $N$. Since the intended average subsample size may be much smaller than the full data sample size in practice, in this section we use $n$ to denote the average subsample size, namely $\mathbb{E}\{\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})\} = n/N$, and allow $n = o(N)$. This means the subsampling probabilities are dependent on $N$ and may go to zero.

Given an estimated pilot $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$, the sampled data conditional log-likelihood function is written as

$$\ell_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = \sum_{i=1}^{N} \delta_i \big[ \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \log\{\bar{\pi}_N(\boldsymbol{x}_i; \boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\} \big] + C, \tag{16}$$

where

$$\bar{\pi}_N(\boldsymbol{x}_i; \boldsymbol{\theta} \mid \boldsymbol{\vartheta}) = \int f(y \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \pi_N(\boldsymbol{x}_i, y; \boldsymbol{\vartheta}) \mathrm{d}y, \tag{17}$$

and $C = \sum_{i=1}^{N} \delta_i \log\{\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\}$ does not contain $\boldsymbol{\theta}$. Here, we use notation conditional on $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$ to emphasize its dependence on $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$.

Denote the sampled data estimator through maximizing $\ell_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$ as $\hat{\boldsymbol{\theta}}_{s,\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}$. We need the following regularity assumptions to investigate the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{s,\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}$.

**Assumption 1'** *For $\boldsymbol{\vartheta}$ in an neighborhood of $\boldsymbol{\vartheta}_p$*

$$\limsup_{n,N\to\infty} \frac{N}{n}\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta})\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4\} < \infty \quad and \tag{18}$$

$$\limsup_{n,N\to\infty} \frac{N}{n}\mathbb{E}\{\pi_N(\boldsymbol{x}, y, \boldsymbol{\vartheta})\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^2\} < \infty. \tag{19}$$

**Assumption 2'** *The parameter space $\boldsymbol{\Theta}$ is compact and there exist a function $B_{\boldsymbol{\vartheta}}(\boldsymbol{x}, y)$ such that for any component of $\boldsymbol{\theta}$, say $\theta_{j_1}$, $\theta_{j_2}$, and $\theta_{j_3}$,*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial^3 \ell(\boldsymbol{\theta}; \boldsymbol{x}, y)}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3}} \right| \leq B_{\boldsymbol{\vartheta}}(\boldsymbol{x}, y), \tag{20}$$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{\partial^3 \log\{\bar{\pi}_N(\boldsymbol{x}; \boldsymbol{\theta} \mid \boldsymbol{\vartheta})\}}{\partial \theta_{j_1} \partial \theta_{j_2} \partial \theta_{j_3}} \right| \leq B_{\boldsymbol{\vartheta}}(\boldsymbol{x}, y), \tag{21}$$

*where $\bar{\pi}_N(\boldsymbol{x}; \boldsymbol{\theta} \mid \boldsymbol{\vartheta})$ is defined in (17) and $B_{\boldsymbol{\vartheta}}(\boldsymbol{x}, y)$ satisfies that for $\boldsymbol{\vartheta}$ in a neighborhood of $\boldsymbol{\vartheta}_p$*

$$\limsup_{n, N \to \infty} \frac{N}{n} \mathbb{E}\{\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) B_{\boldsymbol{\vartheta}}(\boldsymbol{x}, y)\} < \infty. \tag{22}$$

**Assumption 3'** *As $n$ and $N$ goes to infinity, for $\boldsymbol{\vartheta}$ in an neighborhood of $\boldsymbol{\vartheta}_p$, the matrix*

$$\boldsymbol{\Sigma}_{N, \boldsymbol{\theta}, \boldsymbol{\vartheta}} = \frac{N}{n} \mathbb{E}\left[ \dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y) \pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) - \frac{\mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}}{\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}} \right] \to \boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}}, \tag{23}$$

*where $\boldsymbol{\Sigma}_{N, \boldsymbol{\theta}, \boldsymbol{\vartheta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}}$ are finite, positive definite, and continuous with respective to $\boldsymbol{\vartheta}$.*

**Remark 4** *Assumption 1' is essentially moment conditions on the first and second derivatives of the log-likelihood. If $n^{-1} N \pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ are bounded, then the integrability of $\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4$ and $\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^2$ is sufficient for Assumption 1'. We impose stronger moment conditions here compared with the independent and identically distributed (i.i.d.) case in Section 2, because we allow $\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ to depend on $N$ and $\boldsymbol{\vartheta}$ has to be estimated. Assumptions 2'-3' are the counterparts corresponding to the Assumptions 2-3 in Section 2. Note that $\boldsymbol{\Sigma}_{N, \boldsymbol{\theta}, \boldsymbol{\vartheta}}$ is always semi-positive definite because it can be written as*

$$\boldsymbol{\Sigma}_{N, \boldsymbol{\theta}, \boldsymbol{\vartheta}} = \frac{N}{n} \mathbb{E}\left( \pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \left[ \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \frac{\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}}{\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}} \right]^{\otimes 2} \right). \tag{24}$$

*If $n^{-1} N \pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ are bounded away from zero, then with Assumption 1', a sufficient condition for Assumption 3' is that the sequence of random variables*

$$\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \frac{\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}}{\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}}$$

*is full rank, for which a sufficient condition is that $\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ and $\boldsymbol{c}^{\mathrm{T}} \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ has nonzero correlation for any nonzero and nonrandom vector $\boldsymbol{c}$. If $\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ is obtained by rescaling a function of $(\boldsymbol{x}, y)$ i.e., $\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}) = n/N \pi(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ (e.g., the case in Section 5), where $\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ does not depend on $N$, then Assumption 3' reduces to the same requirement as in Assumption 3 except that $\pi(\boldsymbol{x}, y)$ is replaced by $\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta})$. If $\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ is bounded away from zero, then a sufficient condition is the integrability of*

$$\mathbb{E}^{-1}\{\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\} \mathbb{E}^2\{\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\| \pi(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}$$

*together with a nonzero correlation between $\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ and $\boldsymbol{c}^{\mathrm{T}} \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ for any nonzero constant vector $\boldsymbol{c}$.*

**Theorem 5** *Under Assumptions 1'-3', as n and N goes to infinity,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{s,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} - \boldsymbol{\theta}) \xrightarrow{D} \mathbb{N}(\mathbf{0}, \ \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}), \tag{25}$$

*where $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}$ is the limit of $\boldsymbol{\Sigma}_{N,\boldsymbol{\theta},\boldsymbol{\vartheta}_p}$.*

Let $\hat{\boldsymbol{\theta}}_{u,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}$ be the estimator obtained through solving the following class of estimating equations

$$\sum_{i=1}^{N} \delta_i \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) = \mathbf{0}, \tag{26}$$

where $\boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ satisfies that $\mathbb{E}\{\delta \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\} = \mathbf{0}$. The class of estimators defined in (26) includes the IPW estimator with $\boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) = \pi_N^{-1}(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ and the MSCLE with $\boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) = \dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$.

Similar to Theorem 2, the following result shows that with estimated parameters in subsampling probabilities, the proposed estimator is the most efficient among the class of estimators defined through solving (26).

**Theorem 6** *Assume that the partial derivatives of $\mathbb{E}\{\delta \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}$ with respect to $\boldsymbol{\theta}$ can be passed under the integration sign, and the regularity conditions for the following standard asymptotic expansion holds.*

$$\hat{\boldsymbol{\theta}}_{u,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} = \boldsymbol{\theta} - n^{-1}\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1} \sum_{i=1}^{N} \delta_i \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) + o_P(n^{-1/2}), \tag{27}$$

*where $\mathbf{M}_{N,\boldsymbol{\theta},\boldsymbol{\vartheta}} = n^{-1}N\mathbb{E}\{\delta \dot{\boldsymbol{U}}_{\boldsymbol{\vartheta}}(\boldsymbol{\theta}; \boldsymbol{x}, y)\}$ and $\dot{\boldsymbol{U}}(\boldsymbol{\theta}; \boldsymbol{x}, y) = \partial \boldsymbol{U}(\boldsymbol{\theta}; \boldsymbol{x}, y)/\partial \boldsymbol{\theta}^{\mathrm{T}}$. Assume that $\mathbf{M}_{N,\boldsymbol{\theta},\boldsymbol{\vartheta}}$ and $\mathbf{V}_{N,\boldsymbol{\theta},\boldsymbol{\vartheta}} = n^{-1}N\mathbb{E}\{\delta \boldsymbol{U}_{\boldsymbol{\vartheta}}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\}$ are continuous in $\boldsymbol{\vartheta}$ and they converge to finite and full rank matrices $\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}}$ and $\mathbf{V}_{\boldsymbol{\theta},\boldsymbol{\vartheta}}$, respectively. The asymptotic variance covariance matrix of $\hat{\boldsymbol{\theta}}_u$ (multiplied by n) is $\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbf{V}_{\boldsymbol{\vartheta}_p}\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$, and it satisfies that*

$$\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbf{V}_{\boldsymbol{\vartheta}_p}\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1} \geq \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}, \tag{28}$$

*where the equality holds if $\boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ is asymptotically a linear function of $\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$ in the sense that $n^{-1}N\mathbb{E}(\delta\|T_{N\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}\|^2 \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = o_P(1)$, where $T_{N\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} = \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) + \mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$.*

## 5. Subsampling probabilities based on gradient norms (GN)

We use a specific form of subsampling probabilities to illustrate our results. One way to specify subsampling probabilities is to use the norm of the per-observation score function, the gradient of the log-likelihood, by letting

$$\pi(\boldsymbol{x}_i, y_i) \propto \|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\|, \quad \text{for} \quad i = 1, ..., N, \tag{29}$$

where $\propto$ means "proportional to". This option leads to the L-optimal subsampling probabilities that minimize the trace of the asymptotic variance covariance matrix for a linearly transformed IPW estimator (Wang et al., 2018; Ai et al., 2021; Yu et al., 2022). While the optimal probabilities in (29) are for the IPW estimator in (3), it can be used to obtain more efficient subsample for our MSCLE.

Since (29) implies that the subsampling probabilities are dependent on the unknown $\boldsymbol{\theta}$, a pilot estimate is required. In addition, to control the average subsample size, the subsampling probabilities need to be re-scaled and the scaling value may also be unknown. Let $\boldsymbol{\vartheta}$ be the vector consisting of parameters that are required to calculate the subsampling probabilities. The subsampling probabilities are presented as

$$\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta}) = \frac{n\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\|}{N\Psi}, \quad i = 1, ..., N, \tag{30}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\theta}^{\mathrm{T}}, \Psi)^{\mathrm{T}}$ and $\Psi = \mathbb{E}\{\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|\}$. In practice, the unknown $\boldsymbol{\vartheta}$ is replaced by a pilot estimate $\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}$.

If the subsampling ratio $n/N$ is large (far from zero), then it is possible that some $\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta}_p) > 1$ in (30), and therefore the resulting average subsample size may be smaller than $n$. However, in subsampling, the subsample size is typically much smaller than the full data sample size, so it is reasonable to assume that $n = o(N)$. In this case, the truncation can be ignored, and the average subsample size is $n$ asymptotically, namely,

$$\frac{N}{n}\mathbb{E}\left\{\frac{n\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\|}{N\Psi} \wedge 1\right\} \to 1,$$

where $a \wedge b$ means the smaller value of $a$ and $b$. For the rest of the paper, we assume that $n = o(N)$, and we call $\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})$ in (30) subsampling probabilities since the number of cases that $\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta}) > 1$ is negligible in this scenario.

With the specific form of $\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})$ in (30), the sampled data conditional log-likelihood function given pilot estimate $\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}} = (\tilde{\boldsymbol{\theta}}_{\mathrm{plt}}^{\mathrm{T}}, \tilde{\Psi}_{\mathrm{plt}})^{\mathrm{T}}$ can be written specifically as

$$\ell_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \sum_{i=1}^N \delta_i \big[\log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) - \log \mathbb{E}\{\|\dot{\ell}(\tilde{\boldsymbol{\theta}}_{\mathrm{plt}}; \boldsymbol{x}_i, y_i)\| \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}\big] + C, \tag{31}$$

where $C$ does not contain $\boldsymbol{\theta}$. Note that the second term contains $\boldsymbol{\theta}$ through the expectation. Accordingly, a sufficient condition for Assumption 1' is that $\|\dot{\ell}(\dot{\boldsymbol{\theta}}; \boldsymbol{x}, y)\|\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^2$ and $\|\dot{\ell}(\dot{\boldsymbol{\theta}}; \boldsymbol{x}, y)\|\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4$ are integrable for $\dot{\boldsymbol{\theta}}$ in the neighborhood of $\boldsymbol{\theta}_p$, and a sufficient condition for (22) in Assumption 2' is that $\|\dot{\ell}(\dot{\boldsymbol{\theta}}; \boldsymbol{x}, y)\|B_{\boldsymbol{\theta}}(\boldsymbol{x}, y)$ is integrable for $\dot{\boldsymbol{\theta}}$ in the neighborhood of $\boldsymbol{\theta}_p$. These sufficient conditions for Assumptions 1' and 2' with the specific subsampling probabilities $\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})$ are essentially moments conditions on the log-likelihood of the original data distribution and its derivatives. Assumption 3' requires that

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p} = \frac{1}{\Psi_p}\mathbb{E}\left[\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|\dot{\ell}(\boldsymbol{\theta}_p; \boldsymbol{x}, y)\| - \frac{\mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|\dot{\ell}(\boldsymbol{\theta}_p; \boldsymbol{x}, y)\| \mid \boldsymbol{x}\}}{\mathbb{E}\{\|\dot{\ell}(\boldsymbol{\theta}_p; \boldsymbol{x}, y)\| \mid \boldsymbol{x}\}}\right] \tag{32}$$

is finite and positive definite, and a sufficient condition for $\boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p}$ in (32) to be positive definite is that the quantity $\boldsymbol{l}^{\mathrm{T}}\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ is dependent on $y$ for any constant vector $\boldsymbol{l} \neq \boldsymbol{0}$.

For the IPW estimator, under some regularity conditions, the asymptotic variance-covariance matrix (multiplied by $n$) is

$$\mathbf{V}_{W,\boldsymbol{\vartheta}_p} = \Psi_p \mathbf{F}^{-1} \mathbb{E}\left\{ \frac{\dot{\ell}^{\otimes 2}(\boldsymbol{\theta};\boldsymbol{x},y)}{\|\dot{\ell}(\boldsymbol{\theta}_p;\boldsymbol{x},y)\|} \right\} \mathbf{F}^{-1}. \tag{33}$$

Comparing the expressions of $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$ in (32) and $\mathbf{V}_{W,\boldsymbol{\vartheta}_p}$ in (33), we see that the MSCLE requires weaker assumptions than the IPW estimator. For example, for a normal linear regression model with a known error variance, say $\sigma^2 = 1$, $\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x},y) = -0.5(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\boldsymbol{x}$. If $\boldsymbol{\theta}_p \neq \boldsymbol{\theta}$, i.e., the pilot estimate is not consistent, then given $\boldsymbol{x}$,

$$\frac{\dot{\ell}_j^2(\boldsymbol{\theta};\boldsymbol{x},y)}{\|\dot{\ell}(\boldsymbol{\theta}_p;\boldsymbol{x},y)\|} = \frac{(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})^2 x_j^2}{2|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_p|\|\boldsymbol{x}\|} \tag{34}$$

is not integrable; its expectation exists and equals $+\infty$. Thus, any element of $\mathbf{V}_{W,\boldsymbol{\vartheta}_p}$ (if the integral exists) will be $\pm\infty$. This means that for a subsample taken according to (30), the IPW estimator is not applicable for linear regression. On the other hand for the term in the denominator of (32), by direct calculation we know that

$$\mathbb{E}\{\|\dot{\ell}(\boldsymbol{\theta}_p;\boldsymbol{x},y)\| \mid \boldsymbol{x}\} = \mathbb{E}(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_p| \mid \boldsymbol{x})\|\boldsymbol{x}\| = \{2(\mu_d)\Phi(\mu_d) + 2\phi(\mu_d) - \mu_d\}\|\boldsymbol{x}\|, \tag{35}$$

where $\mu_d = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\theta}_p - \boldsymbol{\theta})$, and $\Phi$ and $\phi$ are the cumulative distribution function and the probability density function, respectively, of the standard normal distribution. Thus under some integrability requirement on the covariate $\boldsymbol{x}$, the MSCLE is applicable.

From Theorem 6, we know that $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1} \leq \mathbf{V}_{W,\boldsymbol{\vartheta}_p}$. Actually, for the IPW estimator, we have

$$\mathbf{V}_{W,\boldsymbol{\vartheta}_p} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1} = \Psi_p \mathbb{E}\{\|\dot{\ell}(\boldsymbol{\theta}_p;\boldsymbol{x},y)\|\boldsymbol{\xi}^{\otimes 2}\}, \tag{36}$$

where

$$\boldsymbol{\xi} = \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\left[\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x},y) - \frac{\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x},y)\|\dot{\ell}(\boldsymbol{\theta}_p;\boldsymbol{x},y)\| \mid \boldsymbol{x}\}}{\mathbb{E}\{\|\dot{\ell}(\boldsymbol{\theta}_p;\boldsymbol{x},y)\| \mid \boldsymbol{x}\}}\right] - \mathbf{F}^{-1}\frac{\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x},y)}{\|\dot{\ell}(\boldsymbol{\theta}_p;\boldsymbol{x},y)\|}. \tag{37}$$

This implies that $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1} = \mathbf{V}_{W,\boldsymbol{\vartheta}_p}$ if and only if $\boldsymbol{\xi} = \mathbf{0}$ almost surely, which occurs if and only if $\dot{\ell}(\boldsymbol{\theta}_p;\boldsymbol{x},y) = \mathbf{0}$ almost surely and this is not possible. Thus with the subsampling probabilities $\pi_N(\boldsymbol{x}_i,y_i;\boldsymbol{\vartheta})$'s in (30) $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1} < \mathbf{V}_{W,\boldsymbol{\vartheta}_p}$.

The $\pi_N(\boldsymbol{x}_i,y_i;\boldsymbol{\vartheta})$'s in (30) are a version of the L-optimal subsampling probabilities. If the pilot $\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}$ is consistent, then $\pi_N(\boldsymbol{x}_i,y_i;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$'s minimize the trace of the asymptotic variance-covariance matrix of the IPW estimator of $\mathbf{F}\boldsymbol{\theta}$ among all subsampling probabilities with the same average subsample size. If the dimension of $\boldsymbol{\theta}$ is one, then $\pi_N(\boldsymbol{x}_i,y_i;\boldsymbol{\vartheta})$'s minimize the asymptotic variance of the IPW estimator of $\boldsymbol{\theta}$. Our results show that the estimation efficiency can be further improved by using the proposed MSCLE.

## 6. Generalized linear models

We provide more detailed discussions to illustrate the proposed estimator in the context of generalized linear models (GLMs). As before, we use $\dot{\ell}$ and $\ddot{\ell}$, respectively, to denote gradient vector and Hessian matrix of $\ell$ with respect to a vector variable; and we use $b'$ and $b''$, respectively, to denote the first derivative and the second derivative of a function $b$ with respect to a scalar variable.

Let $y_i$ be the response and $\boldsymbol{x}_i$ be the corresponding covariate. A GLM assumes that the conditional mean of the response $y_i$ give the covariate $\boldsymbol{x}_i$, $\mu_i = \mathbb{E}(y_i \mid \boldsymbol{x}_i)$, satisfies

$$g(\mu_i) = g\{\mathbb{E}(y_i \mid \boldsymbol{x}_i)\} = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta},$$

where $g$ is the link function, $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}$ is the linear predictor, and $\boldsymbol{\theta}$ is the regression coefficient. For 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,

$$f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}, \phi) = a(y_i, \phi) \exp\left\{\frac{y_i b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) - c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})}{\phi}\right\}, \tag{38}$$

where $a$, $b$ and $c$ are known scalar functions, and $\phi$ is the dispersion parameter. In the framework of GLMs, if the link function $g$ is selected such that $b$ is the identity function, i.e., $b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}$, then the link function is called the canonical link. With a canonical link function, $g\{\mathbb{E}(y_i \mid \boldsymbol{x}_i)\} = c'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})$ where $c'$ is the derivative function of $c$. For example, binary logistic regression is a special case of GLMs with the canonical link and the response variable follows the Bernoulli distribution. Specifically, in logistic regression, $a(y_i, \phi) = 1$, $b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}$, $c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) = \log\{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\}$, and $\phi = 1$. Multi-class logistic regression can also be formulated into the family of GLMs with multivariate responses. We will provide more details in Section 6.2.2.

For a GLM from the exponential family, the MLE of $\boldsymbol{\theta}$ is the maximizer of

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) = \frac{1}{\phi} \sum_{i=1}^{N} \{y_i b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) - c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\} + C,$$

where $C$ does not contain $\boldsymbol{\theta}$, and the maximizer is the solution to the score equation

$$\dot{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) = \frac{1}{\phi} \sum_{i=1}^{N} (y_i - \mu_i) b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) \boldsymbol{x}_i = \boldsymbol{0}, \tag{39}$$

$\mu_i = c'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})/b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})$, and $b'$ and $c'$ are the first derivative functions of $b$ and $c$, respectively.

### 6.1 Informative subsampling estimation

Multiple informative subsampling designs are available such as the local case-control (Fithian and Hastie, 2014), the A- and L-optimal subsampling (Ai et al., 2021; Wang, 2019), and the local uncertainty subsampling (Han et al., 2020). In GLMs with univariate responses, these probabilities have a unified expression and they satisfy that

$$\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta}) \propto |y_i - \mu_i| |b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})| h(\boldsymbol{x}_i), \quad i = 1, ..., N, \tag{40}$$

where $h(\boldsymbol{x}_i) > 0$ is a criterion function that may or may not depend on $\boldsymbol{\theta}$. For example, in logistic regression, if $h(\boldsymbol{x}_i) = 1$, then $\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})$ corresponds to the local case-control subsampling; if $h(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|$, then $\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})$ corresponds to the L-optimal subsampling discussed in Section 5; if $h(\boldsymbol{x}_i) = \|\ddot{\ell}^{-1}(\boldsymbol{\theta})\boldsymbol{x}_i\|$ with $\ddot{\ell}(\boldsymbol{\theta})$ being the Hessian matrix of $\ell(\boldsymbol{\theta})$, then $\pi_N(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta})$ corresponds to the A-optimal subsampling and $h(\boldsymbol{x}_i)$ depends on $\boldsymbol{\theta}$ in general for this case.

In practical implementation, a pilot estimate of $\boldsymbol{\vartheta}$ is used, and to control the average subsample size as $n$ the subsampling probabilities are often taken as

$$\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{n}{N \widetilde{\Psi}_{\mathrm{plt}}} |y_i - \tilde{\mu}_i| \big| b'(\boldsymbol{x}_i^{\mathrm{T}} \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) \big| h(\boldsymbol{x}_i), \quad i = 1, ..., N, \tag{41}$$

where $\tilde{\mu}_i = c'(\boldsymbol{x}_i^{\mathrm{T}} \tilde{\boldsymbol{\theta}}_{\mathrm{plt}})/b'(\boldsymbol{x}_i^{\mathrm{T}} \tilde{\boldsymbol{\theta}}_{\mathrm{plt}})$, and $\tilde{\boldsymbol{\theta}}_{\mathrm{plt}}$ and $\widetilde{\Psi}_{\mathrm{plt}}$ are pilot estimates of parameters $\boldsymbol{\theta}$ and $\mathbb{E}\{|(y - \mu)b'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})|h(\boldsymbol{x})\}$, respectively.

For a subsample taken according to $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$'s, by direct calculation, the sampled data log-likelihood function for $\boldsymbol{\theta}$ simplifies to

$$\ell_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{1}{\phi} \sum_{i=1}^{N} \delta_i \big\{ y_i b(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) - c(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) - \phi \log \mathbb{E}(|y_i - \tilde{\mu}_i| \mid \boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) \big\} + C, \tag{42}$$

where $C$ do not contain $\boldsymbol{\theta}$. By direct calculation, we know that the MSCLE can be obtained by solving the score equation,

$$\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{1}{\phi} \sum_{i=1}^{N} \delta_i \Big( y_i - \frac{\tilde{\kappa}_{1,i}}{\tilde{\kappa}_{0,i}} \Big) b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) \boldsymbol{x}_i = \boldsymbol{0}, \tag{43}$$

where $\tilde{\kappa}_{0,i} = \mathbb{E}(|y_i - \tilde{\mu}_i| \mid \boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}_{\mathrm{plt}})$ and $\tilde{\kappa}_{1,i} = \mathbb{E}(y_i|y_i - \tilde{\mu}_i| \mid \boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}_{\mathrm{plt}})$. The Hessian matrix is

$$\ddot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{1}{\phi} \sum_{i=1}^{N} \delta_i \Big( y_i - \frac{\tilde{\kappa}_{1,i}}{\tilde{\kappa}_{0,i}} \Big) b''(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$$
$$- \frac{1}{\phi^2} \sum_{i=1}^{N} \delta_i \Big( \frac{\tilde{\kappa}_{2,i}}{\tilde{\kappa}_{0,i}} - \frac{\tilde{\kappa}_{1,i}^2}{\tilde{\kappa}_{0,i}^2} \Big) \{b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\}^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}. \tag{44}$$

where $\tilde{\kappa}_{2,i} = \mathbb{E}(y_i^2|y_i - \tilde{\mu}_i| \mid \boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}_{\mathrm{plt}})$. Note that on the right-hand-side of (44), the second term is the dominating term, and the first term is often ignored in numerical optimization, namely, using $\ddot{\ell}^F_{S,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta})$ instead of $\ddot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$, where

$$\ddot{\ell}^F_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = -\frac{1}{\phi^2} \sum_{i=1}^{N} \delta_i \Big( \frac{\tilde{\kappa}_{2,i}}{\tilde{\kappa}_{0,i}} - \frac{\tilde{\kappa}_{1,i}^2}{\tilde{\kappa}_{0,i}^2} \Big) \{b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\}^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}. \tag{45}$$

Here we use the superscript $^F$ because the resulting form of Newton's method is often called the Fisher scoring algorithm. If the canonical link is used as mostly implemented in practice, then $b'(\cdot) = 1$ and $b''(\cdot) = 0$, and thus $\ddot{\ell}^F_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \ddot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$. The MSCLE can be calculated from the Fisher scoring algorithm by iteratively applying

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \{\ddot{\ell}^F_S(\boldsymbol{\theta}^{(t)} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})\}^{-1} \dot{\ell}_S(\boldsymbol{\theta}^{(t)} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}). \tag{46}$$

We will show the explicit expressions of $\tilde{\kappa}_{0,i}$, $\tilde{\kappa}_{1,i}$, and $\tilde{\kappa}_{2,i}$ for the examples in the following subsections.

For GLMs with the subsampling probabilities defined in (41), the $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}$ has the following specific form.

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p} = \frac{1}{\phi^2 \Psi_p} \mathbb{E}\left( b'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_p)\{b'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\}^2 h(\boldsymbol{x}) \left[ |y - \mu_p|(y-\mu)^2 - \frac{\mathbb{E}^2\{(y-\mu)|y-\mu_p| \mid \boldsymbol{x}\}}{\mathbb{E}\{|y-\mu_p| \mid \boldsymbol{x}\}} \right] \boldsymbol{x}^{\otimes 2} \right),$$

where $\mu = c'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})/b'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})$ and $\mu_p = c'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_p)/b'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_p)$. With a canonical link and a consistent pilot estimate, $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}$ simplifies to

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p} = \frac{1}{\phi^2 \mathbb{E}\{|y-\mu|h(\boldsymbol{x})\}} \mathbb{E}\left( \left[ |y-\mu|^3 - \frac{\mathbb{E}^2\{(y-\mu)|y-\mu| \mid \boldsymbol{x}\}}{\mathbb{E}\{|y-\mu| \mid \boldsymbol{x}\}} \right] h(\boldsymbol{x})\boldsymbol{x}^{\otimes 2} \right).$$

### 6.2 Examples

6.2.1 EXAMPLE: BINARY RESPONSE MODELS

Binary data are ubiquitous in case-control studies and classifications. Let $y \in \{0,1\}$ be the binary response variable and $\boldsymbol{x}$ be the covariate vector. Assume that the probability for $y = 1$ given $\boldsymbol{x}$ is

$$\mathbb{P}(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) = p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}), \tag{47}$$

where $p(\cdot)$ is a smooth and monotonic function. The model in (47) is a GLM with $a(y,\phi) = 1$, $b(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}) = \log\{p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\} - \log\{1 - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\}$, $c(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}) = -\log\{1 - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\}$, and $\phi = 1$. It is easy to obtain that $\mu = c'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})/b'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}) = p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})$.

For independent full data from the model in (47), $(\boldsymbol{x}_i, y_i)$, $i = 1,...,N$, the subsampling probabilities in (41) reduce to

$$\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{n\big|\{y_i - p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}})\}b'(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}})\big|h(\boldsymbol{x}_i)}{N\tilde{\Psi}_{\mathrm{plt}}}, \quad i = 1,...,N. \tag{48}$$

For a subsample taken according to $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$ in (48), $\tilde{\kappa}_{0,i}$, $\tilde{\kappa}_{1,i}$, and $\tilde{\kappa}_{3,i}$ have the following specific expressions

$$\tilde{\kappa}_{0,i} = p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) + p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) - 2p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}}), \tag{49}$$

$$\tilde{\kappa}_{1,i} = \tilde{\kappa}_{2,i} = p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) - p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}}). \tag{50}$$

We can use these expressions in (43) and (45), and use the Fisher scoring algorithm to find the MSCLE $\hat{\boldsymbol{\theta}}_{s,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}$, which is the solution to

$$\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \sum_{i=1}^{n} \delta_i \left[ y_i - p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) - \frac{\{1 - 2p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}})\}p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\{1 - p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\}}{p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta}) + p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) - 2p(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})p(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_{\mathrm{plt}})} \right] b'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\theta})\boldsymbol{x}_i$$

$$= \boldsymbol{0}. \tag{51}$$

According to Theorem 5, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}_{s,\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}$ (multiplied by $n$) has an expression of

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p} = \frac{1}{\Psi_p}\mathbb{E}\left(\frac{p'(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}_p)p'(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})b'(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})h(\boldsymbol{x})\boldsymbol{x}^{\otimes 2}}{\left[p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}_p)\{1 - p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\} + p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\{1 - p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}_p)\}\right]}\right), \tag{52}$$

where

$$\Psi_p = \mathbb{E}\big[\big|\{y - p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}_p)\}b'(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}_p)\big|h(\boldsymbol{x})\big]. \tag{53}$$

If the pilot estimates are consistent, then the expression in (52) simplifies to

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p} = \frac{\mathbb{E}\big[p'(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\{b'(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\}^2 h(\boldsymbol{x})\boldsymbol{x}^{\otimes 2}\big]}{4\mathbb{E}\{|p'(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})|h(\boldsymbol{x})\}}. \tag{54}$$

A special case of the model in (47) is the widely used binary logistic regression model,

$$\mathbb{P}(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) = p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})}{1 + \exp(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})}, \tag{55}$$

for which $b(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}) = \boldsymbol{x}^{\text{T}}\boldsymbol{\theta}$, $c(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}) = \log\{1 + \exp(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\}$, and $b'(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}) = 1$. For this model, if $h(\boldsymbol{x}_i) = 1$ and $\widetilde{\Psi}_{\text{plt}} = n^{-1}N$, then $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\theta}}_{\text{plt}})$ in (48) become the local case-control subsampling probabilities (Fithian and Hastie, 2014); letting $\tilde{\mathbf{F}}$ be a pilot estimate of the Fisher information matrix $\mathbb{E}[p(\boldsymbol{x}, \boldsymbol{\theta})\{1 - p(\boldsymbol{x}, \boldsymbol{\theta})\}\boldsymbol{x}\boldsymbol{x}^{\text{T}}]$, if $h(\boldsymbol{x}_i) = \|\boldsymbol{x}_i\|$ or $\|\tilde{\mathbf{F}}^{-1}\boldsymbol{x}_i\|$, then $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\theta}}_{\text{plt}})$ in (48) become the L-optimal or A-optimal subsampling probabilities for the IPW estimator (Wang et al., 2018; Wang, 2019).

By direct calculation, the sampled data conditional log-likelihood score equation in (51) for logistic regression is simplified to

$$\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = \sum_{i=1}^{N} \delta_i\big[y_i - p\{\boldsymbol{x}_i^{\text{T}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_{\text{plt}})\}\big]\boldsymbol{x}_i = \mathbf{0}. \tag{56}$$

Fithian and Hastie (2014) and Wang (2019) proposed to solve

$$\sum_{i=1}^{N} \delta_i\{y_i - p(\boldsymbol{x}_i^{\text{T}}\boldsymbol{\theta})\}\boldsymbol{x}_i = \mathbf{0}, \tag{57}$$

and then add $\tilde{\boldsymbol{\theta}}_{\text{plt}}$ to the resulting estimator to correct the bias, which is identical to the solution to (56). This indicates that the method in Fithian and Hastie (2014) and Wang (2019) is actually the MSCLE, and the score equation in (56) explains where the magic bias correction term in their method comes from.

Specific for logistic regression, the $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}$ in (52) simplifies to

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p} = \frac{\mathbb{E}\big[p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}_p)\{1 - p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\}p\{\boldsymbol{x}^{\text{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}_p)\}h(\boldsymbol{x})\boldsymbol{x}^{\otimes 2}\big]}{\mathbb{E}\big[\{y - p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta}_p)\}h(\boldsymbol{x})\big]}, \tag{58}$$

which is the same as the result in Theorem 21 of Wang (2019). If the pilot estimate is consistent so that $\boldsymbol{\theta}_p = \boldsymbol{\theta}$, then the above expression simplifies to

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p} = \frac{\mathbb{E}\big[p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\{1 - p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\}h(\boldsymbol{x})\boldsymbol{x}^{\otimes 2}\big]}{4\mathbb{E}\big[p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\{1 - p(\boldsymbol{x}^{\text{T}}\boldsymbol{\theta})\}h(\boldsymbol{x})\big]}. \tag{59}$$

When $h(\boldsymbol{x}) = 1$, we see that

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p} = \frac{\mathbf{F}}{4\mathbb{E}\big[p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\{1 - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\}\big]} > \mathbf{F}, \tag{60}$$

if $\boldsymbol{\theta} \neq \mathbf{0}$, indicating that the per-observation information matrix of a local case-control subsample is larger than that of a uniform subsample.

There is no general relationship between the efficiency of the MSCLE and that of the full data MLE, except that the full data MLE has a higher estimation efficiency than a subsample estimator. The interesting result that the asymptotic variance of the local case-control subsampling estimator is twice of that of the full data MLE (Fithian and Hastie, 2014) lies in the very unique structure of the logistic regression model and the special form of the local case-control inclusion probability. The MSCLE includes the local case-control subsampling as a special case, so we can obtain some insights on this interesting result. Since a larger sample size results in a smaller variance in general, it is important to pay attention to the average sample size for a subsampling method. For the local case-control subsampling probability with a consistent pilot under correctly specified logistic regression model, the average sample size is asymptotically $n = N\mathbb{E}\{|y - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})|\} = 2N\mathbb{E}[p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\{1 - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\}]$ and the per-observation information matrix is presented as $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}$ in (60). Thus the asymptotic variance of the local case-control subsampling is $n^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}}^{-1} = 2N^{-1}\mathbf{F}$, while the asymptotic variance of the full data MLE is $N^{-1}\mathbf{F}$.

By separating the average sample size and the per-observation information matrix, we see that the local case-control subsampling is more appealing for more imbalanced data. For the extreme case that $p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}) = 0.5$ almost surely, the local case-control subsampling estimator uses half of the full data on average to achieve the twice variance and it is just the uniform subsampling. However for highly imbalanced data, $n$ can be much smaller than $0.5N$ and the per-observation matrix is much larger than $\mathbf{F}$.

One key for the interesting result in logistic regression with local case-control subsampling is that elements of $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}}$ and $\mathbf{F}$ are proportional to each other. Unfortunately, if the model is not the logistic regression, if the sampling probability is not the local case-control subsampling probability, or if the pilot is not consistent, then this proportional relationship does not hold in general, as seen in (52), (54), (58), and (59) for different cases with binary response models.

The local case-control subsampling sampling is also a quite unique example that the IPW estimator is asymptotically as efficient as the MSCLE as noticed in Shen et al. (2021) and Wang (2019), but again if the model or the sampling probability is changed, then the result does not hold anymore.

### 6.2.2 EXAMPLE: MULTI-CLASS LOGISTIC REGRESSIONS

Now we discuss the multi-class logistic regression, which assume that an experiment has $K$ possible outcomes. This is a GLM with a multinomial distribution, a distribution in the multivariate exponential family. For $k = 1, ..., K$, let $y_{i,k} = 1$ if the $k$-th outcomes occurs in the $i$-th experiment and $y_{i,k} = 0$ otherwise. Thus $\sum_{k=1}^{K} y_{i,k} = 1$. The multinomial logistic

regression assumes that

$$\mathbb{P}(y_{i,k} = 1 \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = p_k(\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}_k)}{\sum_{l=1}^{K} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}_l)}, \quad k = 1, 2, ..., K, \tag{61}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, ..., \boldsymbol{\theta}_K^{\mathrm{T}})^{\mathrm{T}}$ is the regression coefficient vector. Let $\boldsymbol{y}_i = (y_{i,1}, ..., y_{i,K})^{\mathrm{T}}$. For an observation $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, the density of the multivariate response $\boldsymbol{y}_i$ at $\boldsymbol{x}_i$ is

$$f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{\exp(\sum_{k=1}^{K} y_{i,k} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}_k)}{\sum_{l=1}^{K} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}_l)} = \frac{\exp\{\boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{y}_i \otimes \boldsymbol{x}_i)\}}{\sum_{l=1}^{K} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}_l)}, \quad y_{i,k} = 0, 1, \tag{62}$$

where $\otimes$ is the kronecker product. The full data log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[ (\boldsymbol{y}_i \otimes \boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\theta} - \log\left\{ \sum_{l=1}^{K} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}_l) \right\} \right], \tag{63}$$

with the corresponding score function as

$$\dot{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \{ \boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \boldsymbol{\theta}) \} \otimes \boldsymbol{x}_i, \tag{64}$$

where $\boldsymbol{p}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \{ p_1(\boldsymbol{x}_i, \boldsymbol{\theta}), ..., p_K(\boldsymbol{x}_i, \boldsymbol{\theta}) \}^{\mathrm{T}}$. Note that for this model $\sum_{k=1}^{K} p_k(\boldsymbol{x}_i, \boldsymbol{\theta}) = 1$, so not all $\boldsymbol{\theta}_k$'s are estimable and a common constrain is to assume that the regression coefficient corresponding to a baseline class is $\boldsymbol{0}$, e.g., $\boldsymbol{\theta}_K = \boldsymbol{0}$. Thus the full vector of unknown regression parameter is $\boldsymbol{\theta}_{-K} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, ..., \boldsymbol{\theta}_{K-1}^{\mathrm{T}})^{\mathrm{T}}$, whose dimension is $(K-1)d$.

The approximate L-optimal subsampling probabilities for the IPW estimator (Yao and Wang, 2019) are

$$\pi_N(\boldsymbol{x}_i, \boldsymbol{y}_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{n \| \boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) \| \| \boldsymbol{x}_i \|}{N \widetilde{\Psi}_{\mathrm{plt}}}, \quad i = 1, ..., N, \tag{65}$$

where $\widetilde{\Psi}_{\mathrm{plt}}$ is a pilot estimate of $\mathbb{E}\{ \| \boldsymbol{y} - \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{\theta}) \| \| \boldsymbol{x} \| \}$. For the sampled data, the conditional log-likelihood score equation is (detailed derivations in Section A.5.1).

$$\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \sum_{i=1}^{N} \delta_i \{ \boldsymbol{y}_i - \boldsymbol{p}_i^g(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) \} \otimes \boldsymbol{x}_i = \boldsymbol{0}, \tag{66}$$

where

$$\boldsymbol{p}_i^g(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) = \begin{pmatrix} \tilde{p}_{i,1}^g \\ \vdots \\ \tilde{p}_{i,K}^g \end{pmatrix}, \quad \tilde{p}_{i,k}^g = \frac{\exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}_k + \tilde{g}_{i,k})}{\sum_{l=1}^{K} \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}_l + \tilde{g}_{i,l})}, \quad \tilde{g}_{i,k} = \log\{ \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_k; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) \}, \tag{67}$$

and $\boldsymbol{1}_k$ is the $K$-dimensional unit vector with the $k$-th element being one and other elements being zero. Here, the expression of $\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$ in (66) holds in general and it is not restricted to the sampling probability in (65). For the sampling probability in (65), the $\tilde{g}_{i,k}$ in (66) has a specific form of $\tilde{g}_{i,k} = 0.5 \log\left\{ \sum_{l=1}^{K} p_l^2(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) + 1 - 2 p_k(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) \right\}$.

Note that $\tilde{g}_{i,k}$'s do not contain $\boldsymbol{\theta}$; they depend only on the sample data and the pilot $\tilde{\boldsymbol{\theta}}_{\text{plt}}$. Thus, solving (66) is as easy as solving the score function without correction. When we use the constrain that $\boldsymbol{\theta}_K = \mathbf{0}$, we only need to solve the first $(K-1)d$ components of (66) and the last $d$ equations are automatically satisfied. For the case of $K = 2$, the first $d$ equations are the same as the last $d$ equations. In this special case, $\tilde{g}_{i,k} = -\boldsymbol{x}_i^{\text{T}}\tilde{\boldsymbol{\theta}}_{\text{plt}}$, and the score equation in (66) reduce to that in (56).

The asymptotic variance-covariance matrix (multiplied by $n$) for the MSCLE of $\boldsymbol{\theta}_{-K}$ with data taken according to (65) is

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p} = \frac{\mathbb{E}\big[\|\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{x},\boldsymbol{\theta}_p)\|\|\boldsymbol{x}\|\{\boldsymbol{y}_{-K} - \boldsymbol{p}^g_{-K}(\boldsymbol{\theta},\boldsymbol{\theta}_p)\}^{\otimes 2} \otimes \boldsymbol{x}^{\otimes 2}\big]}{\mathbb{E}\{\|\boldsymbol{y} - \boldsymbol{p}(\boldsymbol{x},\boldsymbol{\theta}_p)\|\|\boldsymbol{x}\|\}}, \tag{68}$$

where $(\boldsymbol{x}, \boldsymbol{y})$ is an observation from the data distribution, $\boldsymbol{y}_{-K}$ is $\boldsymbol{y}$ with the $K$-th element removed, $\boldsymbol{p}^g_{-K}(\boldsymbol{\theta},\boldsymbol{\theta}_p)$ is $\boldsymbol{p}^g(\boldsymbol{\theta},\boldsymbol{\theta}_p)$ with the $K$-th element removed, and $\boldsymbol{p}^g(\boldsymbol{\theta},\boldsymbol{\theta}_p)$ has the same expression as $\boldsymbol{p}^g(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}_{\text{plt}})$ except that $\tilde{\boldsymbol{\theta}}_{\text{plt}}$ is replaced by $\boldsymbol{\theta}_p$.

Since the expression of $\tilde{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$ in (66) holds in general, it holds for the LUS probability as well, so the LUS estimator is a specific case of the MSCLE. The LUS probability is designed for imbalanced multi-class response models, and it has an expression of

$$\pi_N^{\text{LUS}}(\boldsymbol{x},\boldsymbol{y};\tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = 2I\{\gamma \geq 2q(\tilde{\boldsymbol{\theta}}_{\text{plt}})\}\frac{q(\tilde{\boldsymbol{\theta}}_{\text{plt}}) + \{1 - 2q(\tilde{\boldsymbol{\theta}}_{\text{plt}})\}\eta(\tilde{\boldsymbol{\theta}}_{\text{plt}})}{\gamma}$$
$$+ I\{\gamma < 2q(\tilde{\boldsymbol{\theta}}_{\text{plt}})\}\frac{\{\gamma - q(\tilde{\boldsymbol{\theta}}_{\text{plt}})\} + (1-\gamma)\eta(\tilde{\boldsymbol{\theta}}_{\text{plt}})}{\gamma - q(\tilde{\boldsymbol{\theta}}_{\text{plt}})},$$

where $q(\tilde{\boldsymbol{\theta}}_{\text{plt}}) = \max\{0.5, p_1(\boldsymbol{x},\tilde{\boldsymbol{\theta}}_{\text{plt}}), ..., p_K(\boldsymbol{x},\tilde{\boldsymbol{\theta}}_{\text{plt}})\}$, $\eta(\tilde{\boldsymbol{\theta}}_{\text{plt}}) = I\{\boldsymbol{y}^{\text{T}}\boldsymbol{p}(\boldsymbol{x},\tilde{\boldsymbol{\theta}}_{\text{plt}}) = q(\tilde{\boldsymbol{\theta}}_{\text{plt}})\}$ with $I()$ being the indicator function, and $\gamma$ is a tuning parameter that corresponds to an upper bound of the average subsample size, namely, $n \geq N/\gamma$.

Our Theorem 5 indicates that the inverse of the asymptotic variance of the LUS estimator is proportional to

$$\sum_{k=1}^{K}\mathbb{E}\Big[p_k(\boldsymbol{x},\boldsymbol{\theta})\pi^{\text{LUS}}(\boldsymbol{x},\mathbf{1}_k;\boldsymbol{\theta}_p)\{\mathbf{1}_k - \boldsymbol{p}^g(\boldsymbol{\theta},\boldsymbol{\theta}_p)\}^{\otimes 2} \otimes \boldsymbol{x}^{\otimes 2}\Big], \tag{69}$$

where $\boldsymbol{p}^g(\boldsymbol{\theta},\boldsymbol{\theta}_p)$ has the same expression as $\boldsymbol{p}^g_i(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}_{\text{plt}})$ in (66) except that $\tilde{\boldsymbol{\theta}}_{\text{plt}}$ and $\boldsymbol{x}_i$ are replaced by $\boldsymbol{\theta}_p$ and $\boldsymbol{x}$, respectively. Our results allow $n = o(N)$ which corresponds to the scenario that $\gamma \to \infty$. In this case, the LUS probability reduces to satisfy that $\pi_N^{\text{LUS}}(\boldsymbol{x},\boldsymbol{y};\tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \propto q(\tilde{\boldsymbol{\theta}}_{\text{plt}}) + \{1 - 2q(\tilde{\boldsymbol{\theta}}_{\text{plt}})\}\eta(\tilde{\boldsymbol{\theta}}_{\text{plt}})$, and (69) still holds with $\pi^{\text{LUS}}(\boldsymbol{x},\mathbf{1}_k;\boldsymbol{\theta}_p)$ replaced by $q(\boldsymbol{\theta}_p) + \{1 - 2q(\boldsymbol{\theta}_p)\}I\{\mathbf{1}_k^{\text{T}}\boldsymbol{p}(\boldsymbol{x},\boldsymbol{\theta}_p) = q(\boldsymbol{\theta}_p)\}$. If the pilot is consistent, i.e., $\boldsymbol{\theta}_p = \boldsymbol{\theta}$, and $\gamma$ is fixed and finite, then the expression in (69) reduces to be proportional to the inverse of $\mathcal{V}_{\text{LUS}}$ in Corollary 4.1 of Han et al. (2020).

Note that the parameter $\gamma$ only controls an upper bound of the average subsample size, so the expressions in (68) and (69) are not comparable because they are scaled by different average subsample sizes. We will provide numerical comparisons by using the save average subsample sizes in Section 8.1. We will also implement the IPW estimator to show that the MSCLE has a higher estimation efficiency than the IPW estimator for the LUS probability as well.

### 6.2.3 Example: Poisson regression

Poisson regression models are commonly used for modeling count data. It assumes that given the covariate $\boldsymbol{x}_i$, the response $y_i$ follows a Poisson distribution with density

$$f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \frac{1}{y_i!} \exp(y_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta} - e^{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}}), \quad y_i = 0, 1, ..., \tag{70}$$

where $\mu_i = \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta})$. This is a specific case of (38) with $\phi = 1$, $a(y_i, \phi) = (y_i!)^{-1}$, $b(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}$ and $c(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta}) = \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\theta})$. The link function in Poisson regression is $g(\mu_i) = \log(\mu_i)$ and it is the canonical link.

For this model, the subsampling probabilities in (41) reduce to

$$\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{n|y_i - \tilde{\mu}_i| h(\boldsymbol{x}_i)}{N \tilde{\Psi}_{\mathrm{plt}}}, \quad i = 1, ..., N, \tag{71}$$

where $\tilde{\mu}_i = \exp(\boldsymbol{x}_i^{\mathrm{T}} \tilde{\boldsymbol{\theta}}_{\mathrm{plt}})$. For a subsample taken according to $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$, the sampled data conditional log-likelihood has the score equation and the Hessian matrix as

$$\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \sum_{i=1}^{N} \delta_i \left( y_i - \frac{\tilde{\kappa}_{1,i}}{\tilde{\kappa}_{0,i}} \right) \boldsymbol{x}_i, \tag{72}$$

and

$$\ddot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = -\sum_{i=1}^{N} \delta_i \left( \frac{\tilde{\kappa}_{2,i}}{\tilde{\kappa}_{0,i}} - \frac{\tilde{\kappa}_{1,i}^2}{\tilde{\kappa}_{0,i}^2} \right) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}, \tag{73}$$

respectively.

Now we show the closed-form expressions of $\tilde{\kappa}_{0,i}$, $\tilde{\kappa}_{1,i}$, and $\tilde{\kappa}_{2,i}$. Let $m_i = \lfloor \tilde{\mu}_i \rfloor$ be the largest integer that is smaller than or equal to $\tilde{\mu}_i$, and let $F(\cdot; \mu_i)$ be the cumulative distribution function for a Poisson distribution with mean $\mu_i$. Here we show the results and give the detailed derivations in Section A.5.2. We have that

$$\tilde{\kappa}_{0,i} = 2\tilde{\mu}_i F(m_i; \mu_i) - 2\mu_i F(m_i - 1; \mu_i) + \mu_i - \tilde{\mu}_i, \tag{74}$$

$$\tilde{\kappa}_{1,i} = 2\mu_i(\tilde{\mu}_i - 1)F(m_i - 1; \mu_i) - 2\mu_i^2 F(m_i - 2; \mu_i) + \mu_i + \mu_i^2 - \mu_i \tilde{\mu}_i, \tag{75}$$

$$\tilde{\kappa}_{2,i} = \tilde{\kappa}_{1,i} + \mu_i^2(\tilde{\mu}_i - 2)\{2F(m_i - 2; \mu_i) - 1\} - 2\mu_i^3 F(m_i - 3; \mu_i) + \mu_i^3. \tag{76}$$

With the above expressions, we can find the MSCLE by applying the Newton's algorithm. We will demonstrate the performance of the resulting estimator in Section 8

## 7. Model misspecification

When the assumed working model is misspecified, the meaning of consistency of an estimator needs to be carefully defined. One definition of the consistency of a subsample estimator is that it has the same probability limit as the full data estimator (Fithian and Hastie, 2014; Han et al., 2020; Shen et al., 2021). In this definition, the IPW estimator $\hat{\boldsymbol{\theta}}_W$ defined in (3) is consistent, because the conditional expectation of the selection indicator $\delta$ give the data equals the inclusion probability and therefore the expectation of the objective function for

the IPW estimator is the same as that of the full data MLE. Specifically, the full data MLE under a misspecified model estimates the $\boldsymbol{\theta}_l$ that solves the following population estimation equation

$$\mathbb{E}^t\{\dot{\ell}(\boldsymbol{\theta}_l; \boldsymbol{x}, y)\} = \mathbf{0}, \tag{77}$$

where we use $\mathbb{E}^t$ to emphasize that the expectation is taken with respect to the true data distribution. Since $\mathbb{E}\{\delta\pi^{-1}(\boldsymbol{x}, y)\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)] = \mathbb{E}^t\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)]$, the IPW estimator $\hat{\boldsymbol{\theta}}_W$ defined in (3) always estimates the same $\boldsymbol{\theta}_l$ as the full data MLE. Actually, the subsample IPW estimator $\hat{\boldsymbol{\theta}}_W$ is often investigated as an estimator of the full data estimator in the literature (e.g. Wang et al., 2018; Yu et al., 2022).

For the MSCLE, we use the assumed model structure to derive the subsampled conditional likelihood, so if the subsampling probabilities also depend on the data then the MSCLE is not consistent to the same limit as the full data MLE when the assumed model structure is misspecified. The MSCLE is consistent to the $\boldsymbol{\theta}_{ls}$ that solves the following population estimation equation if the assumed model is misspecified,

$$\mathbb{E}^t\left[\mathbb{E}^t\{\dot{\ell}(\boldsymbol{\theta}_{ls}; \boldsymbol{x}, y)\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\} - \frac{\mathbb{E}^t\{\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}}{\mathbb{E}^w\{\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}}\mathbb{E}^w\{\dot{\ell}(\boldsymbol{\theta}_{ls}; \boldsymbol{x}, y)\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta}) \mid \boldsymbol{x}\}\right] = \mathbf{0}, \tag{78}$$

where $\mathbb{E}^w$ is the expectation under the assumed working model and $\boldsymbol{\vartheta}$ the fixed parameter in calculating the sampling probabilities. Clearly, $\boldsymbol{\theta}_l$ and $\boldsymbol{\theta}_{ls}$ are different in general, although they can be the same under the special case of the local case-control subsampling.

For an illustration, consider the problem under the more specific class of binary response models discussed in Section 6.2.1. For this class of models, (77) and (78) are simplified as

$$\mathbb{E}_{\boldsymbol{x}}\left[\{p_t(\boldsymbol{x}) - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_l)\}b'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_l)\boldsymbol{x}\right] = \mathbf{0}, \tag{79}$$

and

$$\mathbb{E}_{\boldsymbol{x}}\left[\{p_t(\boldsymbol{x}) - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_{ls})\}\frac{\pi(\boldsymbol{x}, 0; \boldsymbol{\vartheta})\pi(\boldsymbol{x}, 1; \boldsymbol{\vartheta})}{p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_{ls})\pi(\boldsymbol{x}, 1; \boldsymbol{\vartheta}) + \{1 - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_{ls})\}\pi(\boldsymbol{x}, 0; \boldsymbol{\vartheta})}b'(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_{ls})\boldsymbol{x}\right] = \mathbf{0}, \tag{80}$$

respectively, where $\mathbb{E}_{\boldsymbol{x}}$ is the expectation respect to the distribution of $\boldsymbol{x}$ and $p_t(\boldsymbol{x})$ is the true probability of $y = 1$ given $\boldsymbol{x}$. If the subsampling probabilities do not depend on the data, i.e., $\pi(\boldsymbol{x}, y; \boldsymbol{\vartheta})$ is a consistent across the data distribution, then (79) and (80) are identical, meaning that the MSCLE is consistent to the full data MLE. In general for a correctly specified model, $p_t(\boldsymbol{x}) = p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_t)$ at the true parameter $\boldsymbol{\theta}_t$ and thus (79) and (80) are both true when $\boldsymbol{\theta}_l = \boldsymbol{\theta}_{ls} = \boldsymbol{\theta}_t$. For a misspecified model, heuristically, the $\boldsymbol{\theta}_l$ in (79) tries to make $p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_l)$ to be close to $p_t(\boldsymbol{x})$ and thus it is a reasonable value to consider. On the other hand, the $\boldsymbol{\theta}_{ls}$ in (80) also tries to make $p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_{ls})$ to be close to $p_t(\boldsymbol{x})$, and thus it is also a reasonable value for the parameter when the model is misspecified. It is difficult to argue which of $\boldsymbol{\theta}_l$ and $\boldsymbol{\theta}_{ls}$ is better, because the true $p_t(\boldsymbol{x})$ is unknown and the definition of the closeness of $p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})$ to $p_t(\boldsymbol{x})$ varies. For example, the $\boldsymbol{\theta}$'s that minimize $\mathbb{E}_{\boldsymbol{x}}[\{p_t(\boldsymbol{x}) - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\}^2]$ and $\mathbb{E}_{\boldsymbol{x}}\{|p_t(\boldsymbol{x}) - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})|\}$ are typically different although they are both reasonable objectives, and there is no easy way to tell whether $\boldsymbol{\theta}_l$ or $\boldsymbol{\theta}_{ls}$ is closer to any of them.

Specifically, if the sampling probabilities in (48) are used for logistic regression ($b'(\cdot) = 1$) and assume that the pilot $\boldsymbol{\vartheta} = \boldsymbol{\theta}_{ls}$, then (80) simplifies to

$$\mathbb{E}_{\boldsymbol{x}}\left[\{p_t(\boldsymbol{x}) - p(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}_{ls})\}h(\boldsymbol{x})\boldsymbol{x}\right] = \boldsymbol{0}. \tag{81}$$

We see that $\boldsymbol{\theta}_{ls}$ in (81) is in general different from $\boldsymbol{\theta}_l$ unless $h(\boldsymbol{x}) = 1$. Again, the $\boldsymbol{\theta}_{ls}$ in the above population estimation equation also seems to be a reasonable target if the model is misspecified. Existing investigations on optimal subsampling focus on defining the $h(\boldsymbol{x})$ to improve the estimation efficiency under a correctly specified model. Finding $h(\boldsymbol{x})$ to improve the subsample performance under model misspecifications is also interesting and it requires future investigations.

## 8. Numerical experiments

### 8.1 Multi-class logistic regression

We demonstrate the performance of the proposed MSCLE estimator using the multi-class logistic regression model discussed in Section 6.2.2. We set the full data sample size $N = 10^6$, and let the subsample sizes be $n = 500, 1000, 1500$; and 2000. We assume that the responses have three possible categories ($K = 3$), and let the dimension of the covariates $\boldsymbol{x}_i = (1, \boldsymbol{x}_{-1,i}^{\mathrm{T}})^{\mathrm{T}}$'s be $d = 4$ where the first element of one is for the intercept parameters. For this setup, the dimension of the unknown regression coefficient vector is eight, and we set the true parameter to be $\boldsymbol{\theta}_{-K} = (0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4)^{\mathrm{T}}$ when generating the data. To generate the covariates corresponding to the slope parameters, $\boldsymbol{x}_{-1,i}$'s, we consider the following distributions.

(a) Multivariate normal distribution $\boldsymbol{x}_{-1,i} \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, where the $(i,j)$-th element of $\boldsymbol{\Omega}$ is $\Sigma_{ij} = 0.5^{|i-j|}$ and $I()$ is the indicator function. This distribution is symmetric with light tails. The resulting probabilities of the responses for the three possible categories are 0.3, 0.39, and 0.31, respectively.

(b) Multivariate log-normal distribution $\boldsymbol{x}_{-1,i} \sim \mathbb{LN}(\boldsymbol{0}, \boldsymbol{\Omega})$, where $\boldsymbol{x}_{-1,i} = \exp(\boldsymbol{z}_i)$, $\boldsymbol{z}_i \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, and $\boldsymbol{\Omega}$ is the same as defined in the above case a). This distribution is asymmetric and positively skewed. The resulting probabilities of the responses for the three possible categories are 0.22, 0.65, and 0.13, respectively.

(c) Multivariate $t$ distribution, $\boldsymbol{x}_{-1,i} \sim \mathbb{T}_3(\boldsymbol{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is defined in case a). This distribution is symmetric, and response composition is similar to case a). However, this distribution have heavier tails.

(d) Independent exponential distribution, $\boldsymbol{x}_{-1,i} \sim \mathbb{EXP}(1)$, where components of $\boldsymbol{x}_{-1,i}$ independently follow the standard exponential distribution. This distribution is asymmetric and positively skewed. The response composition is similar to case b).

We repeat the simulation for $R = 1000$ times and calculate the empirical mean squared error (MSE) as $R^{-1}\sum_{r=1}^{R}\|\hat{\boldsymbol{\theta}}_{-K}^{(r)} - \boldsymbol{\theta}_{-K}\|^2$, empirical variance as $R^{-1}\sum_{r=1}^{R}\|\hat{\boldsymbol{\theta}}_{-K}^{(r)} - \bar{\boldsymbol{\theta}}_{-K}\|^2$ and empirical squared bias as $\|\bar{\boldsymbol{\theta}}_{-K} - \boldsymbol{\theta}_{-K}\|^2$, where $\hat{\boldsymbol{\theta}}_{-K}^{(r)}$ is the estimate at the $r$-th repetition and $\bar{\boldsymbol{\theta}}_{-K}$ is the average of $\hat{\boldsymbol{\theta}}_{-K}^{(r)}$'s. In each repetition, we generate the full data set and

use a uniform sample of size 400 to calculate the pilot estimates. Results are presented in Figure 1. Since the empirical MSE is the sum of the asymptotic variance and empirical squared bias, we only plot the empirical variances and squared biases. For comparison, we also implement the commonly used IPW estimator (IPW) and a naive estimator (naive) that is obtained by maximizing $\sum_{i=1}^{N} \delta_i \ell(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$. We implement both the gradient normal (GN) based sampling probability and the LUS probability. For the LUS probability, with $\gamma = N/n$, the average subsample size would be smaller than $n$ so the comparison would not be fair. Thus we have to set $\gamma$ to be smaller than $N/n$ for different distributions of $\boldsymbol{x}$ in order to have a fair comparison. Since the sampling ratios $n/N$ considered here are small, this is equivalent to scale the LUS probability by numbers that are larger than one.

The simulation results in Figure 1 can be summarized as follows: 1) The variance is the dominating term in the MSE and the squared bias is a small term for the MSCLE and the IPW estimators. 2) The bias of the Naive estimator does not decrease to zero with the increase of the subsample size, which suggests that the naive estimator is subject to non-negligible biases. 3) Both the IPW estimator and our proposed MSCLE show that the variance decreases as the sample size increases. 3) In all scenarios for both the GN probability and the LUS probability, our proposed MSCLE has smaller MSEs than the IPW estimator, which confirms our theory in Theorem 6. We see that the GN probability outperforms the LUS probability especially if the responses are balanced or if the covariate distributions have heavy tails. This is because the LUS probability is designed specifically to address imbalanced data and it does not take into account the structure information in the variance matrix represented by $\boldsymbol{x}$.

To examine the impact of misspecification in the pilot estimator, we run the afore-mentioned simulation with the same setting except that we use a wrong pilot estimator. Specifically, for each component of the pilot estimator obtained from the pilot sample, we add to it a random number generated from a uniform distribution between 1 and 2, and then we use this misspecified pilot estimator to implement the all the estimators in comparison. Results are presented in Figure 2.

From Figure 2, we see that the MSE gets larger for all the methods, especially for the naive estimator and the IPW estimator. This is because with this misspecified pilot estimator the resulting informative subsamples are more different from the original data distribution. The inflation on the MSE due to the misspecified pilot estimator is much smaller for the MSCLE than that for the IPW estimator. The advantage of the MSCLE over the IPW estimator become more significant with the misspecified pilot estimator. One reason for this behavior is that adding a random bias to the pilot estimator $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$ brings in additional variation to the estimated probabilities $\pi_N(\boldsymbol{x}_i, \boldsymbol{y}_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$'s. These probabilities are in the denominators for the IPW estimator while they are in the expectation and then log-transformed for the MSCLE, so the additional variation expresses more significantly for the IPW estimator, resulting in a larger variance of the estimator. Another reason is that the GN sampling probability is a version of the L-optimal probabilities for the IPW estimator with consistent pilot (Wang et al., 2018; Ai et al., 2021; Yu et al., 2022), while it is not optimal for the MSCLE, so a systematic bias on the pilot has a larger negative impact on the IPW estimator. We have additional numerical experiments to investigate the effect of different levels of pilot misspecification on the final estimator. Please see the details in Section B.1 of the Appendix.

(a) $\boldsymbol{x}_{-1,i}$'s are normal

(b) $\boldsymbol{x}_{-1,i}$'s are lognormal

(c) $\boldsymbol{x}_{-1,i}$'s are $\mathbb{T}_3$

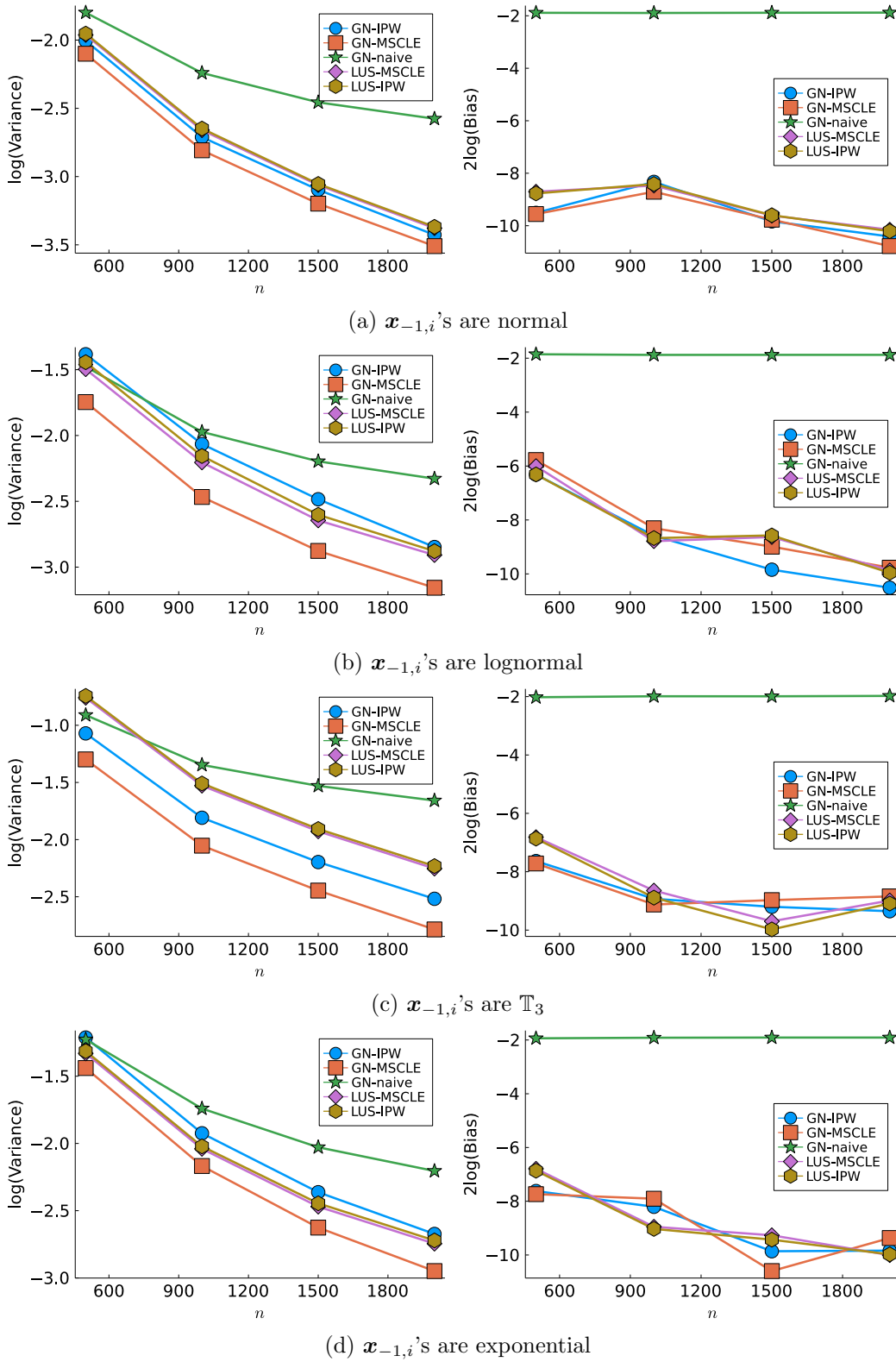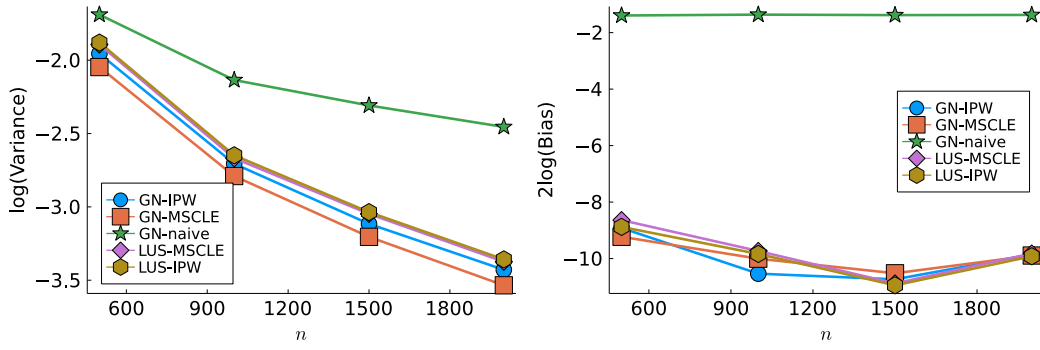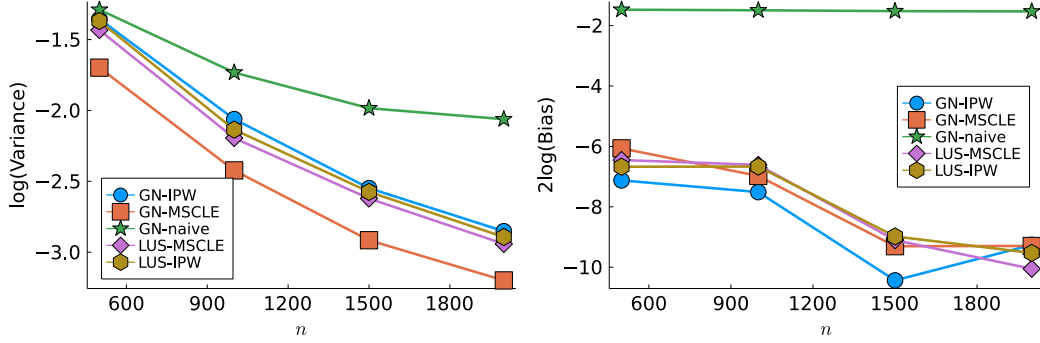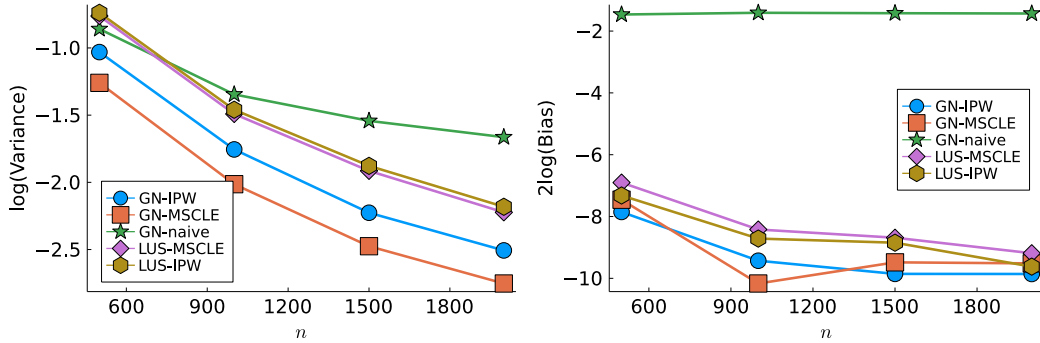(d) $\boldsymbol{x}_{-1,i}$'s are exponential

Figure 1: Log of empirical variances and squared biases (the smaller the better) of subsample estimators for different sample sizes in multi-class logistic regression.
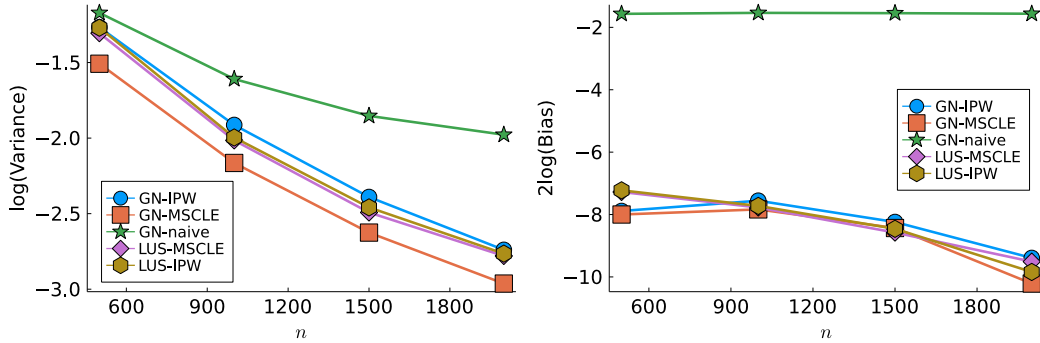
(a) $\boldsymbol{x}_{-1,i}$'s are normal

(b) $\boldsymbol{x}_{-1,i}$'s are lognormal

(c) $\boldsymbol{x}_{-1,i}$'s are $\mathbb{T}_3$

(d) $\boldsymbol{x}_{-1,i}$'s are exponential

Figure 2: Log of empirical variances and squared biases (the smaller the better) of subsample estimators for different sample sizes in multi-class logistic regression when the pilot estimator is misspecified.

To evaluate the performance of the asymptotic variance in Theorem 5, we also calculated the estimated variance by plug-in estimation. The results are reported in Figure 3. The empirical variances and the estimated variances are quite close, and this is the case for both the GN sampling probability and the LUS probability, conforming our theoretical result in Theorem 5.
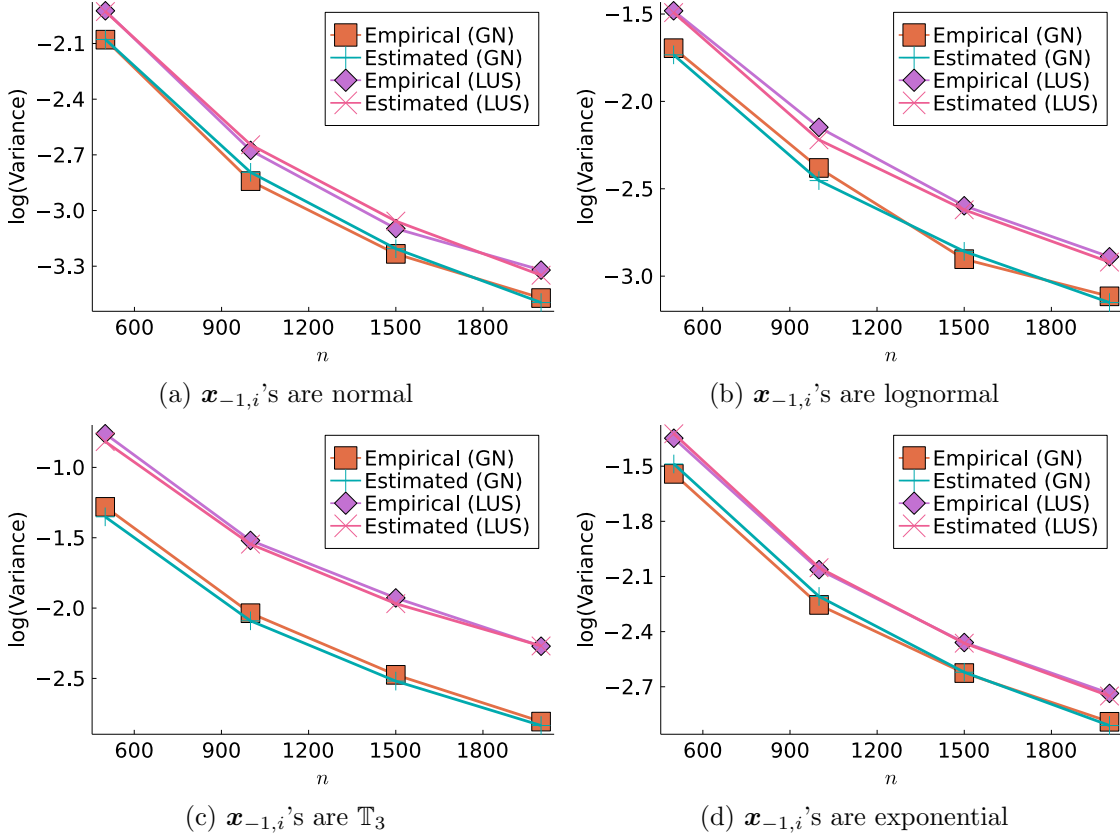


Figure 3: Log of empirical variances and average estimated variances of subsample MSCLE for different sample sizes in multi-class logistic regression. Logarithm is taken for better presentation.

The MSCLE needs to calculate $\bar{\pi}_N(\boldsymbol{x}_i; \boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$'s and may require more computational resources. However, this is done only on the selected subsample, so the actual difference for computational cost will not be large. To check this, we also recorded the computational costs in terms of CPU times and memory allocations for the MSCLE and the IPW estimator. As expected, the time differences would not be noticeable if we present the total time of calculating a subsample estimator and calculating $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$'s. To see the differences more clearly, we separated the times for calculating subsample estimators from those of calculating $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$'s. We implemented all the algorithm in Julia (Bezanson et al., 2017) on a Desktop running Ubuntu 20.04. We restricted all the calculations to use one thread of the CPU with a base frequency of 2,200 megahertz and a maximum boosted frequency of 4,549 megahertz. We repeated the simulation for 100 times, and calculated

the average CPU times and memory allocations. We used a smaller number of iterations here because the variations of the computational costs across different repetitions are much smaller than that of the estimators.

Results for case (a) with multivariate normal covariates are reported in Table 1. Indeed, the MSCLE takes more time than the IPW, but the difference is very small and is negligible compared with the major time of calculating $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$'s that both methods require. In terms of the memory allocations, the major cost is also on calculating $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$'s. Here IPW uses a little more memory because we created an additional weighted covariate matrix in the implementation to save some CPU time. The difference on memory allocations is not significant either. Results for other covariate distributions are similar and thus we omit them.

Table 1: Average CPU times and Memory allocations for the multi-class logistic regression example. Here "calPI" is for the step of calculating sampling probabilities $\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$'s and taking subsamples.

| | CPU time (millisecond) | | | | Memory allocation (megabyte) | | | |
|---|---|---|---|---|---|---|---|---|
| | $n$=500 | 1000 | 1500 | 2000 | $n$ =500 | 1000 | 1500 | 2000 |
| MSCLE | 0.28 | 0.45 | 0.62 | 0.77 | 0.88 | 1.72 | 2.56 | 3.40 |
| IPW | 0.27 | 0.41 | 0.54 | 0.71 | 0.89 | 1.78 | 2.64 | 3.53 |
| calPI | 41.17 | 39.18 | 38.99 | 38.19 | 160.38 | 160.41 | 160.44 | 160.47 |
| Full | 660.40 | | | | 1,777.66 | | | |

To exam the performance of the proposed method with a higher dimension, we also performed experiments using a setting considered in Han et al. (2020). Specifically, the conditional distribution of $\boldsymbol{x}_{-1,i}$ given $\boldsymbol{y}_i = \mathbf{1}_k$ is $\mathbb{N}(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$, for $k = 1, 2$, and 3. Here, $\boldsymbol{\mu}_1$ is a 20 dimensional vector with the first ten elements being one and the last ten elements being zero; $\boldsymbol{\mu}_2$ is a 20 dimensional vector with the first ten elements being zero and the last ten elements being one; $\boldsymbol{\mu}_2$ is a 20 dimensional vector of zeros, and $\boldsymbol{\Omega}_s$'s are all equal to the identity matrix. The marginal distribution of $\boldsymbol{y}_i$ is $\mathbb{P}(\boldsymbol{y}_i = \mathbf{1}_1) = \mathbb{P}(\boldsymbol{y}_i = \mathbf{1}_3) = 0.1$ and $\mathbb{P}(\boldsymbol{y}_i = \mathbf{1}_2) = 0.8$. We consider two settings of sampling rates. The first setting is with $N = 10^6$ and $n = 2000, 4000, 6000$ and 10000, so the sampling rate $n/N$ is low. For this low sampling rate, we can control the average sample size by scaling the LUS probability without affecting its form. The second setting is the same as that in Han et al. (2020) with $N = 50000$ and $\gamma = 3, 2$, and 1.1. For this setting the sampling rate is high and we cannot control the average sample size for the LUS probability without changing its form. We implemented the LUS probability with the given values of $\gamma$, and then implemented the GN sampling probability with the actual average sample size acquired by the LUS probability. Figure 4 presents the empirical variances and squared biases for the two settings. The overall pattern here is similar to that with the low dimensional covariates. We omit the results for the naive estimator here because it is biased as seen in previous examples.

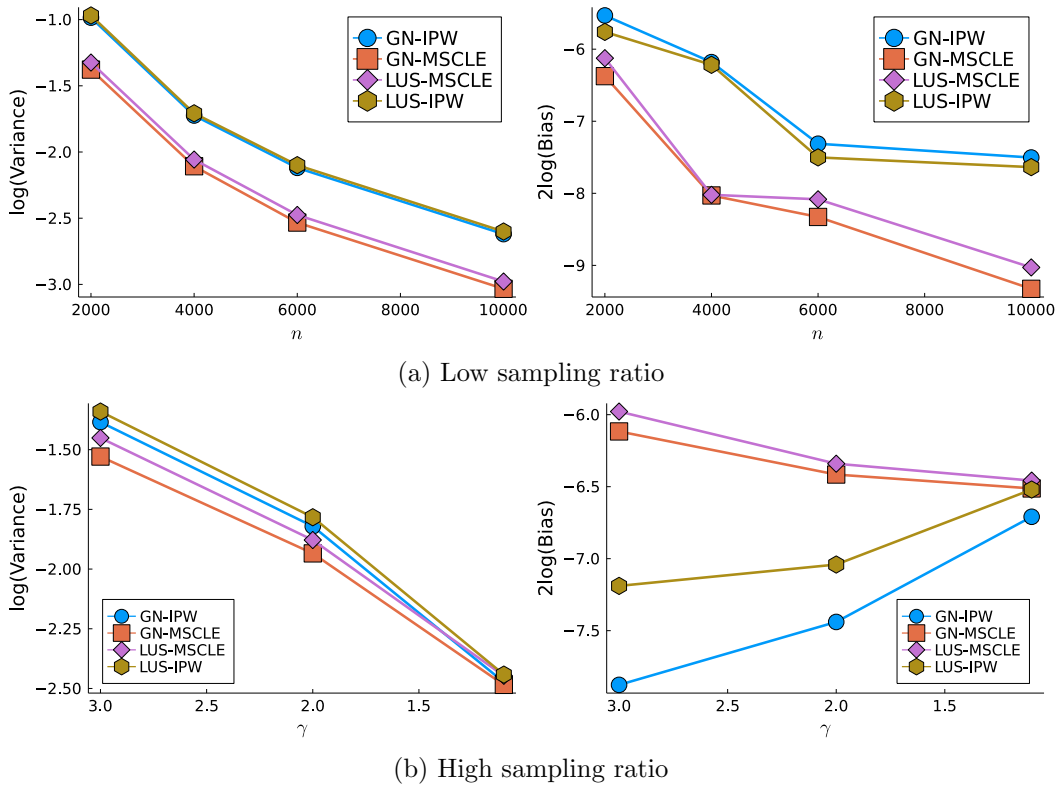(a) Low sampling ratio



(b) High sampling ratio

Figure 4: Log of empirical variances and squared biases (the smaller the better) of subsample estimators for different sample sizes in multi-class logistic regression with the setting of the LUS paper (Han et al., 2020).

## 8.2 Poisson regression

In this section, we consider the Poisson regression model discussed in Section 6.2.3. As in the previous example, we set the full data sample size $N = 10^6$, and let the subsample sizes be $n = 500; 1000; 1500;$ and 2000. We let the dimension of the covariates $\boldsymbol{x}_i = (1, \boldsymbol{x}_{-1,i}^{\mathrm{T}})^{\mathrm{T}}$'s be $d = 7$ with the first element of one for the slope parameters. We set the true value of $\boldsymbol{\theta}$ as a vector of 0.25's to generate the data, and consider the following distributions for the covariates corresponding to the slope parameters, $\boldsymbol{x}_{-1,i}$'s.

(a) Independent uniform distribution $\boldsymbol{x}_{-1,i} \sim \mathbb{U}(\boldsymbol{0}, \boldsymbol{1})$, where components of $\boldsymbol{x}_{-1,i}$ independently follow the standard uniform distribution. This distribution has a bounded support and it is symmetric.

(b) Independent Beta distribution $\boldsymbol{x}_{-1,i} \sim \mathbb{B}(\boldsymbol{2}, \boldsymbol{5})$, where components of $\boldsymbol{x}_{-1,i}$ independently follow the Beta distribution with parameters 2 and 5. This distribution has a bounded support and it is skewed to the right.

(c) Multivariate normal distribution $\boldsymbol{x}_{-1,i} \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is defined similarly to that defined in Section 8.1. This distribution has an unbounded support and it is symmetric.

(d) Independent exponential distribution, $\boldsymbol{x}_{-1,i} \sim \mathbb{EXP}(2)$, where components of $\boldsymbol{x}_{-1,i}$ independently follow the exponential distribution with rate parameter 2. This distribution has an unbounded support, and it is asymmetric and positively skewed.

Here, the distribution in case (a) is the Case 1 used in Yu et al. (2022). We also considered other cases of distributions used in their paper based on uniform distributions. The results are omitted because the relative performance of the three estimators are similar to that of case (a).

Again, we repeat the simulation for $R = 1000$ to calculate the empirical MSEs, variances, and squared biases, for the three estimators: the proposed MSCLE estimator, the IPW estimator, and the naive estimator (naive). Results on the empirical variances and squared biases are presented in Figure 5.

The simulation results in Figure 5 shows similar patterns of the simulation study in Section 8.1. The variance is the dominating term in the MSE for both the IPW estimator and the MSCLE. The bias of the naive estimator does not decrease with the subsample size. Both the IPW estimator and the MSCLE have smaller variances for larger sample sizes. The MSCLE is uniformly more efficient than the other estimators in terms of the variances.

For the computational costs, the relative pattern is very similar to that for the multiclass logistic regression and thus we omit the results.

We also have additional numerical results on the effect of pilot misspecification for Poisson regression. Please see them in Section B.2 of the Appendix.

(a) $\boldsymbol{x}_{-1,i}$'s are uniform



(b) $\boldsymbol{x}_{-1,i}$'s are Beta



(c) $\boldsymbol{x}_{-1,i}$'s are normal



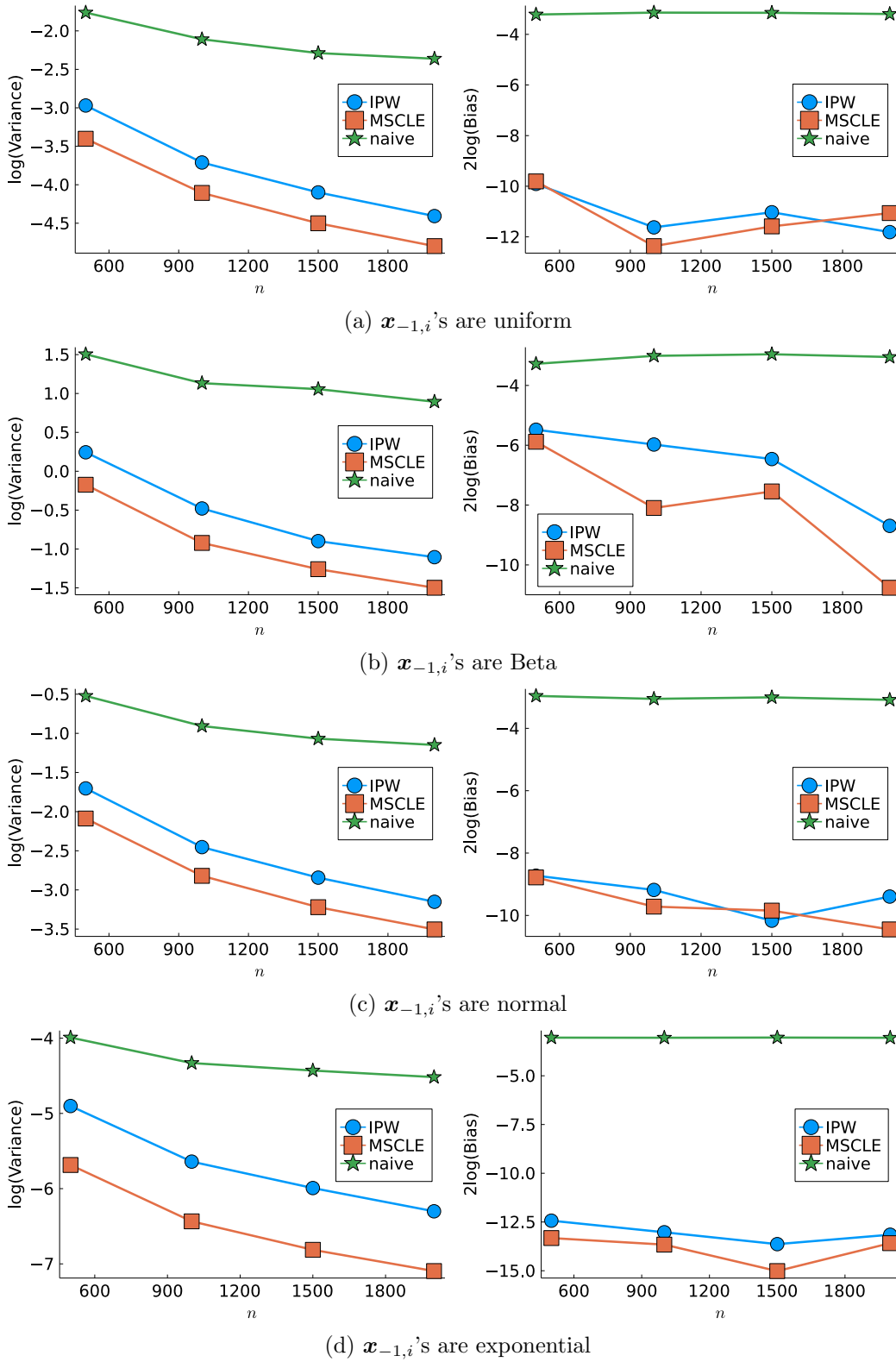(d) $\boldsymbol{x}_{-1,i}$'s are exponential

Figure 5: Log of empirical variances and squared biases (the smaller the better) of subsample estimators for different sample sizes in Poisson regression.

Figure 6 plots the empirical variances and estimated variances to evaluate the performance of the asymptotic variance in Theorem 5 for Poisson regression. We see that the empirical variances and the estimated variances are very close, showing that the approximation based on the asymptotic distribution in Theorem 5 is accurate.



(a) $\boldsymbol{x}_{-1,i}$'s are uniform

(b) $\boldsymbol{x}_{-1,i}$'s are Beta

(c) $\boldsymbol{x}_{-1,i}$'s are normal

(d) $\boldsymbol{x}_{-1,i}$'s are exponential
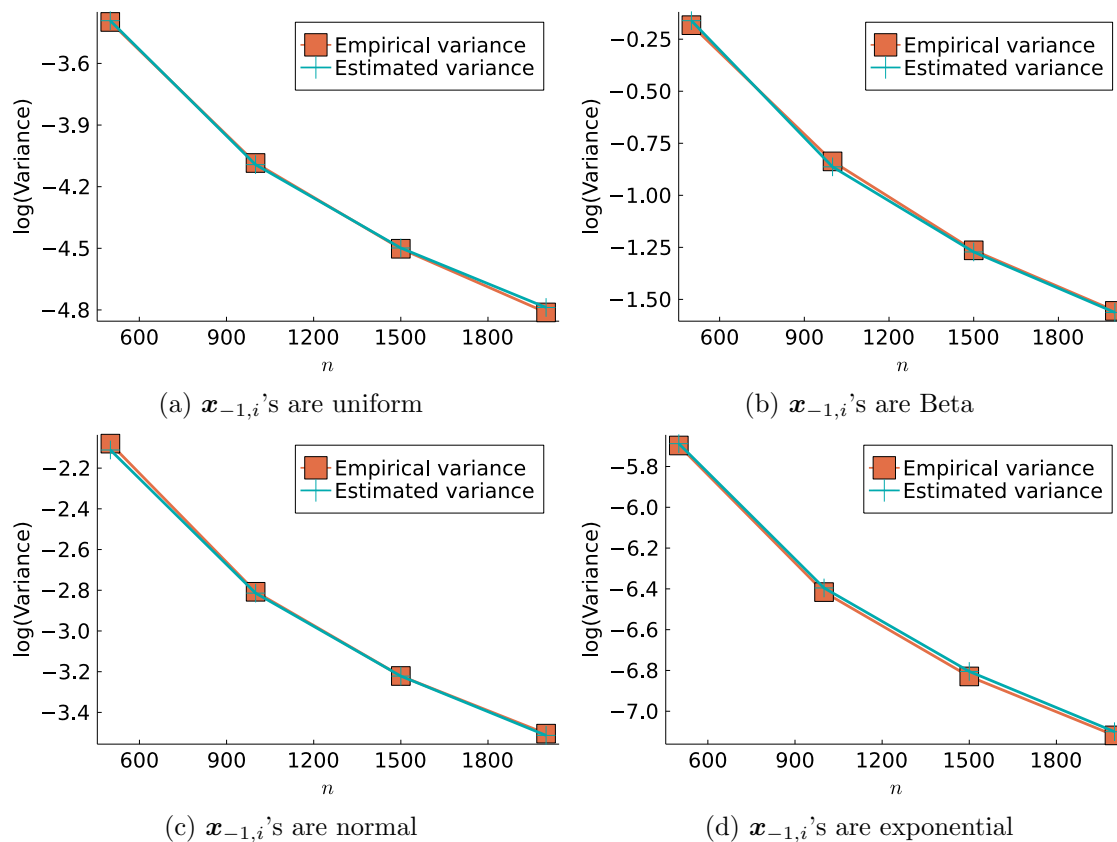
Figure 6: Log of empirical variances and estimated variance of subsample estimators for different sample sizes in Poisson regression. Logarithm is taken for better presentation.

## 8.3 Real data example: cover type data

To demonstrate the performance of the MSCLE, we applied it to the cover type data (Blackard and Dean, 1999). This dataset contains $N = 581,012$ observations on ten quantitative variables that measure geographical features and lighting conditions. The interest is to use these variables to build a model to predict the forest type. There are $K = 7$ forest types: Spruce/Fir (36.46%), Lodgepole Pine (48.76%), Ponderosa Pine (6.15%), Aspen (1.63%), Douglas-fir (2.99%), Krummholz (3.53%) and Cottonwood/Willow (0.427%), where the number in the parentheses is the percentage of the forest type in the full data set. We include an intercept in the model so $d = 11$ and the dimension of unknown parameter is $(K-1)d = 66$. To fit a multi-class logistic regression model as in (61), we use the probabilities defined in (65) to take subsamples of sizes $n$ from the full data, and use the subsamples to train the model. Pilot estimates are obtained from subsamples of average

size 2000 taken according to proportional subsampling probabilities. Note that for real data the true data generating model is unknown and any parametric model may be subject to a certain level of model misspecification. We use the full data estimator as the "true parameter" and repeat the subsampling estimator for 1000 times to calculated empirical biases, variances, and MSEs.

Results are presented in Table 2. We present the squared biases so that they are directly comparable to the variances. For comparison, we obtained the IPW estimator (IPW) and the naive estimator (Naive), and also implement the uniform subsampling method (Uniform). In terms of MSE, MSCLE outperforms the IPW estimator, which outperforms the Uniform method. The Naive method has small variances, but its biases are large and do not seem to converge to zero.

Table 2: Empirical bias (squared), Variance (Var.) and the mean squared error (MSE) of the four subsample estimators for cover type data

|  | $n = 4000$ | | | $n = 6000$ | | | $n = 8000$ | | |
|---|---|---|---|---|---|---|---|---|---|
|  | $\text{Bias}^2$ | Var. | MSE | $\text{BIAS}^2$ | Var. | MSE | $\text{Bias}^2$ | Var. | MSE |
| MSCLE | 25.73 | 248.27 | 273.74 | 14.53 | 120.17 | 134.58 | 12.56 | 76.87 | 89.35 |
| IPW | 22.97 | 325.70 | 348.35 | 6.76 | 163.67 | 170.27 | 3.16 | 104.57 | 107.62 |
| Naive | 146.04 | 231.76 | 377.57 | 146.73 | 118.05 | 264.66 | 150.49 | 80.78 | 231.49 |

## 8.4 Real data example: the MNIST data

In this section, we illustrate the advantage of the MSCLE over the IPW estimator using the famous MNIST data that is available at `http://yann.lecun.com/exdb/mnist/`. The data has a training set with 60,000 instances and a testing set with 10,000 instances. Each instance is an image of a handwritten digit with 28 by 28 greyscale pixels. The goal is to train a model using the training set to predict the handwritten digits of $\{0, 1, 2, ..., 9\}$ in the testing set. We implement the convolutional neural network LeNet-5 (LeCun et al., 1998) with Flux.jl (Innes, 2018), and use a subsample of average size $n = 5,000$ out of $N = 60,000$ (about 8.3% of the training data) to train the model.

To apply the sampling probabilities presented in (65), we use the norms of the of the 28 by 28 greyscale pixels to replace $\|\boldsymbol{x}_i\|$'s. Note that this does not give the GN sampling probability for this model. We adopt this simplified version of the sampling probability because the complicated model structure and high dimension (with 44,426 parameters) makes it difficult to calculate the real gradient norms. The pilot probabilities $\boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})$'s are obtained from a pilot model trained with 2,000 uniformly selected instances from the training set.

Due to the complexity of the model the parameters do not have explicit interpretations, so we report the test accuracy which is the percentage of correct classification in the testing set. Figure 7 plots the test accuracy against the epoch for the MSCLE and the IPW estimator. The epoch value is the number of passes throughout the entire training set that the mini batch gradient descent algorithm has completed. It is seen that the MSCLE outperforms the IPW estimator uniformly across epochs.

Again, no model can be the exact data generating model for real data, so although being a very rich model, the LeNet-5 model here may be subject to a certain level of

misspecification as well. As a matter of fact, since the number of parameters is much larger than the subsample size, some of the regularity assumptions in Section 4 may not hold. The promising performance in Figure 7 indicates that the proposed MSCLE may be applicable in more general scenarios than that restricted by the regularity assumptions in Section 4.
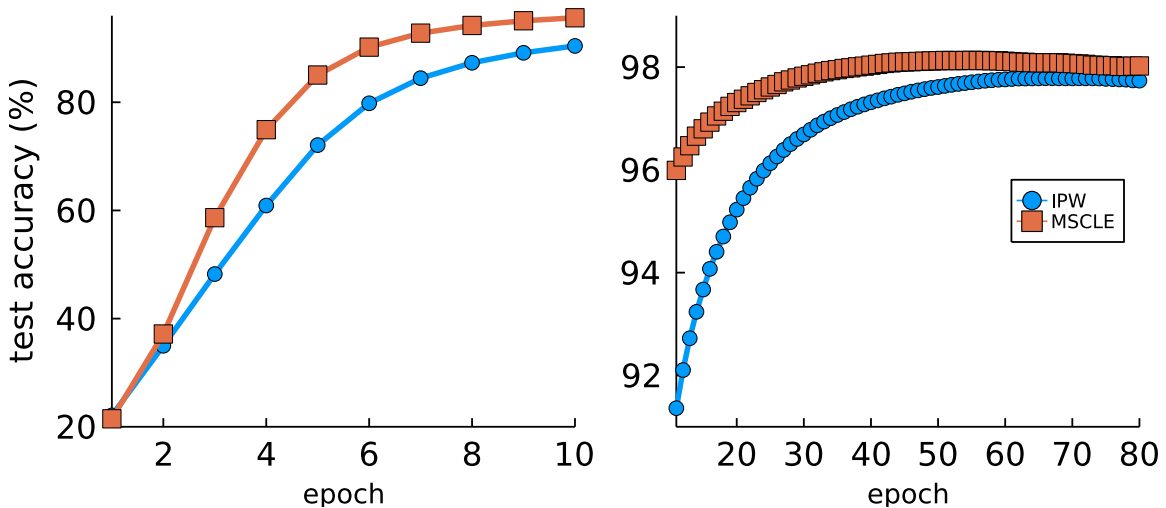


Figure 7: Classification accuracy (in percentage) on the test data against epoch in the training using subsamples of size $n = 5,000$ from the MNIST data.

## 9. Discussion

Subsampling is a useful technique for handling big data. To estimate the parameters with the subsample data, the inverse probability IPW estimator have been used as a gold standard method. In this paper, we consider an alternative method of parameter estimation using the sampled conditional likelihood function. The resulting MSCLE is consistent and is more efficient than the IPW estimator. The computation for obtaining MSCLE requires computing the bias-adjustment term in the naive (complete-case) method. Explicit closed-form formula for the bias-adjustment terms are given in Section 6.

To obtain efficient subsampling probabilities, we need consistent estimates of the model parameters. In this paper, we assume that an independent pilot subsample is available outside of the current sample. In this case, the theory can be developed as presented in Section 4. Otherwise, one can consider an adaptive estimation method which simultaneously updates the parameter estimates and the corresponding selection probabilities. Such adaptive methods introduce additional computational burden and also theoretical challenges. In addition, the subsampling idea is closely related with reservior sampling (Efraimidis and Spirakis, 2006) which is a useful tool for handling streaming data. Thus, the proposed MSCLE can be used to handle the reservior sample with unequal selection probabilities. Such extensions will be the topics for future research.

## Acknowledgments

## Appendix A. Proofs and technical details

### A.1 Proof of Theorem 1

For any functions $h_1(\boldsymbol{x}, y)$ and $h_2(\boldsymbol{x})$, assuming integrability in the following, we have

$$
\begin{aligned}
\mathbb{E}\{\delta h_1(\boldsymbol{x}, y) h_2(\boldsymbol{x})\} &= \mathbb{E}\big[\delta h_2(\boldsymbol{x}) \mathbb{E}\{h_1(\boldsymbol{x}, y) \mid \delta, \boldsymbol{x}\}\big] \\
&= \mathbb{E}\Big[\pi(\boldsymbol{x}, y) h_2(\boldsymbol{x}) \mathbb{E}\{h_1(\boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1\} \,\Big|\, \boldsymbol{x}, y\Big] \\
&= \mathbb{E}\Big[\delta h_2(\boldsymbol{x}) \mathbb{E}\{h_1(\boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1\} \,\Big|\, \boldsymbol{x}, y\Big] \\
&= \mathbb{E}\big[\delta h_2(\boldsymbol{x}) \mathbb{E}\{h_1(\boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1\}\big].
\end{aligned}
\tag{A.1}
$$

The estimator $\hat{\boldsymbol{\theta}}_S$ is the maximizer of $\ell_S(\boldsymbol{\theta})$ in (5), so $\sqrt{N}(\hat{\boldsymbol{\theta}}_S - \boldsymbol{\theta})$ is the maximizer of $\gamma(\boldsymbol{\eta}) = \ell_S(\boldsymbol{\theta} + \boldsymbol{\eta}/\sqrt{N}) - \ell_S(\boldsymbol{\theta})$. By Taylor's expansion,

$$
\gamma(\boldsymbol{\eta}) = \frac{1}{\sqrt{N}} \boldsymbol{\eta}^{\mathrm{T}} \dot{\ell}_S(\boldsymbol{\theta}) + \frac{1}{2N} \boldsymbol{\eta}^{\mathrm{T}} \ddot{\ell}_S(\boldsymbol{\theta}) \boldsymbol{\eta} + R,
\tag{A.2}
$$

where $R = o_P(1)$ because Assumption 2 indicates that

$$
|R| \leq \frac{d^3 \|\boldsymbol{\eta}\|^3}{3N^{1/2}} \times \frac{1}{N} \sum_{i=1}^{N} B(\boldsymbol{x}_i, y_i) = o_P(1).
\tag{A.3}
$$

For $\dot{\ell}_S(\boldsymbol{\theta}) = \sum_{i=1}^{N} \dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$, $\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$ are i.i.d. random vectors. Let $h_1(\boldsymbol{x}_i, y_i) = \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)$ and $h_2(\boldsymbol{x}_i) = 1$ in (A.1), we know that $\mathbb{E}\{\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\} = \boldsymbol{0}$, and using (8) we have

$$
\mathbb{V}\{\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\} = \mathbb{E}\Big(\delta[\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1)\}]^{\otimes 2}\Big)
\tag{A.4}
$$

$$
= \mathbb{E}\Big(\delta[\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1)\}]\Big)
\tag{A.5}
$$

$$
= \mathbb{E}\left(\delta\left[\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \frac{\mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \pi(\boldsymbol{x}, y) \mid \boldsymbol{x}\}}{\bar{\pi}^2(\boldsymbol{x}; \boldsymbol{\theta})}\right]\right) = \boldsymbol{\Sigma_\theta}.
\tag{A.6}
$$

Thus, from the central limit theorem,

$$
\frac{1}{\sqrt{N}} \dot{\ell}_S(\boldsymbol{\theta}) \xrightarrow{D} \mathbb{N}(\boldsymbol{0}, \boldsymbol{\Sigma_\theta}).
\tag{A.7}
$$

Now we investigate the Hessian matrix

$$
\frac{1}{N} \ddot{\ell}_S(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \delta_i[\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) - \mathbb{E}\{\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i, \delta_i = 1)\}]
$$

$$-\frac{1}{N}\sum_{i=1}^{N}\delta_i\left[\int\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)\dot{f}^{\mathrm{T}}(y_i\mid\boldsymbol{x}_i,\delta_i=1;\boldsymbol{\theta})\mathrm{d}y\right]$$

$$\equiv\Delta_1-\mathbf{H}_N. \tag{A.8}$$

Note that $\Delta_1$ is an average of i.i.d terms, and from Assumption 1 and (A.1), we know that $\mathbb{E}(\Delta_1)=\mathbf{0}$. Thus, from the strong law of large numbers, $\Delta_1=o(1)$ almost surely.

Under Assumptions 1 and 2, and the fact that $\pi(\boldsymbol{x},y)$ is bounded above by one, derivatives can pass the integration sign in $\int f(y\mid\boldsymbol{x};\boldsymbol{\theta})\pi(\boldsymbol{x},y)\mathrm{d}y$ by the dominated convergence theorem. Thus, using (7) we have

$$\dot{f}(y_i\mid\boldsymbol{x}_i,\delta_i=1;\boldsymbol{\theta})$$

$$=\frac{\dot{f}(y_i\mid\boldsymbol{x}_i;\boldsymbol{\theta})\pi(\boldsymbol{x}_i,y_i)}{\bar{\pi}(\boldsymbol{x}_i;\boldsymbol{\theta})}-\frac{f(y_i\mid\boldsymbol{x}_i;\boldsymbol{\theta})\pi(\boldsymbol{x}_i,y_i)\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)\pi(\boldsymbol{x}_i,y)\mid\boldsymbol{x}_i\}}{\bar{\pi}^2(\boldsymbol{x}_i;\boldsymbol{\theta})}$$

$$=\frac{\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)\pi(\boldsymbol{x}_i,y_i)f(y_i\mid\boldsymbol{x}_i;\boldsymbol{\theta})}{\bar{\pi}(\boldsymbol{x}_i;\boldsymbol{\theta})}-\frac{f(y_i\mid\boldsymbol{x}_i;\boldsymbol{\theta})\pi(\boldsymbol{x}_i,y_i)\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)\pi(\boldsymbol{x}_i,y)\mid\boldsymbol{x}_i\}}{\bar{\pi}^2(\boldsymbol{x}_i;\boldsymbol{\theta})}. \tag{A.9}$$

Thus, we have that

$$\mathbf{H}_N=\frac{1}{N}\sum_{i=1}^{N}\delta_i\left[\int\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)\dot{f}^{\mathrm{T}}(y\mid\boldsymbol{x}_i,\delta_i=1;\boldsymbol{\theta})\mathrm{d}y\right] \tag{A.10}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\delta_i\left[\frac{\mathbb{E}\{\dot{\ell}^{\otimes2}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)\pi(\boldsymbol{x}_i,y_i)\mid\boldsymbol{x}_i\}}{\bar{\pi}(\boldsymbol{x}_i;\boldsymbol{\theta})}-\frac{\mathbb{E}^{\otimes2}\{\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i)\pi(\boldsymbol{x}_i,y)\mid\boldsymbol{x}_i\}}{\bar{\pi}^2(\boldsymbol{x}_i;\boldsymbol{\theta})}\right]. \tag{A.11}$$

Since $\mathbf{H}_N$ is an average of i.i.d terms, and (8) and (A.1) tells us that its expectation is $\mathbb{E}(\mathbf{H}_N)=\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, from the strong law of large numbers, $\mathbf{H}_N\to\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ almost surely. Therefore,

$$-\frac{1}{N}\ddot{\ell}_S(\boldsymbol{\theta})=\boldsymbol{\Sigma}_{\boldsymbol{\theta}}+o_P(1). \tag{A.12}$$

From (A.2), (A.7) and (A.12), applying the Basic Corollary in page 2 of Hjort and Pollard (2011) finishes the proof.

## A.2 Proofs of Theorem 2

Let $\ell_S(\boldsymbol{\theta};\boldsymbol{x},y)=\log\{f(y\mid\boldsymbol{x},\delta=1;\boldsymbol{\theta})\}$. Note that $\mathbb{E}\{\delta\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y)\mid\boldsymbol{x}\}=\mathbf{0}$. Using

$$\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\mathbb{E}\{\delta\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y)\mid\boldsymbol{x}\}=\mathbb{E}\{\delta\dot{\boldsymbol{U}}(\boldsymbol{\theta};\boldsymbol{x},y)\mid\boldsymbol{x}\}+\mathbb{E}\{\delta\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y)\dot{\ell}_S^{\mathrm{T}}(\boldsymbol{\theta};\boldsymbol{x},y)\mid\boldsymbol{x}\},$$

we have

$$\mathbb{E}\{\delta\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y)\dot{\ell}_S^{\mathrm{T}}(\boldsymbol{\theta};\boldsymbol{x},y)\}=-\mathbb{E}\{\delta\dot{\boldsymbol{U}}(\boldsymbol{\theta};\boldsymbol{x},y)\}=-\mathbf{M}_{\boldsymbol{\theta}}.$$

Hence, by calculating the variance covariance matrix

$$\mathbb{V}\{\delta\mathbf{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y)+\delta\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\dot{\ell}_S(\boldsymbol{\theta};\boldsymbol{x},y)\}$$

$$=\mathbb{V}\{\delta\mathbf{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y)\}+\mathbf{M}_{\boldsymbol{\theta}}^{-1}\mathbb{E}\{\delta\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y)\dot{\ell}_S^{\mathrm{T}}(\boldsymbol{\theta};\boldsymbol{x},y)\}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$$

$$+ \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbb{E}\{\delta\dot{\ell}_S(\boldsymbol{\theta};\boldsymbol{x},y)\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{\theta};\boldsymbol{x},y)\}(\mathbf{M}_{\boldsymbol{\theta}}^{-1})^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbb{V}\{\delta\dot{\ell}_S(\boldsymbol{\theta};\boldsymbol{x},y)\}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$$
$$= \mathbf{M}_{\boldsymbol{\theta}}^{-1}\mathbb{E}\{\delta\boldsymbol{U}^{\otimes 2}(\boldsymbol{\theta};\boldsymbol{x},y)\}(\mathbf{M}_{\boldsymbol{\theta}}^{-1})^{\mathrm{T}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \geq \mathbf{0},$$

we have proved (14). Now, we know that the equality holds if $T(\boldsymbol{\theta}) := \delta\mathbf{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y) + \delta\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\dot{\ell}_S(\boldsymbol{\theta};\boldsymbol{x},y)$ is a constant vector. Since $\mathbb{E}\{T(\boldsymbol{\theta})\} = \mathbf{0}$, we have $T(\boldsymbol{\theta}) = \mathbf{0}$, or $\boldsymbol{U}(\boldsymbol{\theta};\boldsymbol{x},y) = -\mathbf{M}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\dot{\ell}_S(\boldsymbol{\theta};\boldsymbol{x},y)$. This completes the proof.

### A.3 Proof of Theorem 5

Under Assumptions 1' and 2', and the fact that $\pi_N(\boldsymbol{x},y;\boldsymbol{\vartheta})$ is bounded above by one, we know that derivatives can pass the integration sign in $\int f(y \mid \boldsymbol{x};\boldsymbol{\theta})\pi_N(\boldsymbol{x},y;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})\mathrm{d}y$ by the dominated convergence theorem. The score function for $\ell_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$ can be written as

$$\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \sum_{i=1}^{N} \delta_i[\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i) - \int \dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y)f(y \mid \boldsymbol{x}_i,\delta=1;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})\mathrm{d}y], \quad (A.13)$$

$$= \sum_{i=1}^{N} \delta_i[\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i) - \mathbb{E}\{\dot{\ell}(\boldsymbol{\theta};\boldsymbol{x}_i,y_i) \mid \boldsymbol{x}_i,\delta=1;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}], \quad (A.14)$$

where

$$f(y_i \mid \boldsymbol{x}_i,\delta_i=1,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{f(y_i \mid \boldsymbol{x}_i;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})\pi(\boldsymbol{x}_i,y_i)}{\int f(y \mid \boldsymbol{x}_i;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})\pi(\boldsymbol{x}_i,y)\mathrm{d}y}. \quad (A.15)$$

Similarly to (A.1), for any functions $h_1(\boldsymbol{x},y)$ and $h_2(\boldsymbol{x})$, we have

$$\mathbb{E}\{\delta h_1(\boldsymbol{x},y)h_2(\boldsymbol{x}) \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\} = \mathbb{E}\left[\delta h_2(\boldsymbol{x})\mathbb{E}\{h_1(\boldsymbol{x},y) \mid \delta,\boldsymbol{x}\} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\right]$$
$$= \mathbb{E}\left[\pi_N(\boldsymbol{x},y;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})h_2(\boldsymbol{x})\mathbb{E}\{h_1(\boldsymbol{x},y) \mid \boldsymbol{x},\delta=1\} \mid \boldsymbol{x},y;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\right]$$
$$= \mathbb{E}\left[\delta h_2(\boldsymbol{x})\mathbb{E}\{h_1(\boldsymbol{x},y) \mid \boldsymbol{x},\delta=1\} \mid \boldsymbol{x},y;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\right]$$
$$= \mathbb{E}\left[\delta h_2(\boldsymbol{x})\mathbb{E}\{h_1(\boldsymbol{x},y) \mid \boldsymbol{x},\delta=1\} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\right] \quad (A.16)$$

Now we prove Theorem 5. Since $\hat{\boldsymbol{\theta}}_{s,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}$ is the maximizer of $\ell_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{s,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} - \boldsymbol{\theta})$ is the maximizer of $\gamma(\boldsymbol{\eta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \ell_{S,\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{\theta} + \boldsymbol{\eta}/\sqrt{n}) - \ell_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$. By Taylor's expansion,

$$\gamma(\boldsymbol{\eta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \frac{1}{\sqrt{n}}\boldsymbol{\eta}^{\mathrm{T}}\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) + \frac{1}{2n}\boldsymbol{\eta}^{\mathrm{T}}\ddot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})\boldsymbol{\eta} + R, \quad (A.17)$$

where $R = o_P(1)$. We show why $R$ is a small term in probability in the following. By (20) and (21) of Assumption 2', $R$ satisfies that

$$|R| \leq \frac{d^3\|\boldsymbol{\eta}\|^3}{3} \times \frac{1}{n^{3/2}}\sum_{i=1}^{N} \delta_i B_{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}(\boldsymbol{x}_i,y) \equiv \frac{d^3\|\boldsymbol{\eta}\|^3}{3} \times \Delta. \quad (A.18)$$

For any constant $\epsilon > 0$, from Markov's inequality,

$$\mathbb{P}(\Delta > \epsilon \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) \leq \frac{N\int \pi_N(\boldsymbol{x},y;\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})B(\boldsymbol{x},y)f(y \mid \boldsymbol{x};\boldsymbol{\theta})\mathrm{d}y}{n^{3/2}\epsilon} = o_P(1), \quad (A.19)$$

37

where the last step above is because (22) and the fact that $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$ is independent of the data imply that $Nn^{-1}\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})B_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{x}, y) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} = O_P(1)$. Because a conditional probability is a bounded random variable, $\mathbb{P}(\Delta > \epsilon) = \mathbb{E}\{\mathbb{P}(\Delta > \epsilon \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\} \to 0$, indicating that $\Delta = o_P(1)$, and therefore $R = o_P(1)$.

For $\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$, it is a sum of i.i.d terms conditionally on $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$. Let $h_1(\boldsymbol{x}, y) = \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)$ and $h_2(\boldsymbol{x}) = 1$ in (A.16), we know that $\mathbb{E}\{\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} = \boldsymbol{0}$; and using a similar procedure to obtain (A.6), we have

$$\frac{1}{n}\mathbb{V}\{\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} = \frac{N}{n}\mathbb{E}\Big(\delta[\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}]^{\otimes 2} \ \Big| \ \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\Big)$$

$$= \frac{N}{n}\mathbb{E}\Big(\delta[\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}] \ \Big| \ \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\Big)$$

$$= \boldsymbol{\Sigma}_{N\tilde{\boldsymbol{\vartheta}}_{\text{plt}}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p}, \tag{A.20}$$

where the convergence in probability in the last step is due to Assumption 3. Note that $\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = \delta_i[\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) - \mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i, \delta_i = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}]$. We check the Lindeberg-Feller condition given $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$. For any $\epsilon > 0$,

$$\frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\{\|\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|^2 I(\|\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|^2 > n\epsilon) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}$$

$$= \frac{N}{n}\mathbb{E}\{\|\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|^2 I(\|\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|^2 > n\epsilon) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}$$

$$\leq \frac{N}{n^2\epsilon}\mathbb{E}\{\|\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|^4 \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}$$

$$= \frac{N}{n^2\epsilon}\mathbb{E}\big[\delta\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) - \mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}_i, \delta_i = 1; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}\|^4 \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\big]$$

$$\leq \frac{8N}{n^2\epsilon}\mathbb{E}\big[\delta\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4 + \mathbb{E}\{\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4 \mid \boldsymbol{x}_i, \delta_i = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\big]$$

$$= \frac{16N}{n^2\epsilon}\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4 \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}$$

$$= o_P(1),$$

where the last step is because (18) in Assumption 1 and the fact that $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$ is independent of the data imply that

$$\frac{N}{n}\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4 \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} = O_P(1). \tag{A.21}$$

Thus, from the Lindeberg-Feller central limit theorem, conditionally on $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$,

$$\frac{1}{\sqrt{n}}\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \xrightarrow{D} \mathbb{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p}). \tag{A.22}$$

Now we investigate the Hessian matrix. Note that

$$\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}] = \int \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)f(y \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\mathrm{d}y, \tag{A.23}$$

where $f(y \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$ is the density of $y$ conditional on $\boldsymbol{x}$, $\delta = 1$, and $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$. We have

$$
\frac{1}{n}\ddot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = \frac{1}{n}\sum_{i=1}^{N}\delta_i\big[\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) - \mathbb{E}\{\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i, \delta_i = 1; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}\big]
$$

$$
- \frac{1}{n}\sum_{i=1}^{N}\delta_i\bigg[\int \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\dot{f}^{\text{T}}(y_i \mid \boldsymbol{x}_i, \delta_i = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\mathrm{d}y\bigg]
$$

$$
\equiv \Delta_2 - \mathbf{H}_N^{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}. \tag{A.24}
$$

From (A.16), we know that $\mathbb{E}(\Delta_2) = \mathbf{0}$. For the $j_1, j_2$-th element of $\Delta_2$, $\Delta_{2,j_1j_2}$,

$$
\mathbb{E}\Big(\Delta_{2,j_1,j_2}^2 \;\Big|\; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\Big)
$$

$$
= \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}\Big(\delta_i\big[\ddot{\ell}_{j_1j_2}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) - \mathbb{E}\{\ddot{\ell}_{j_1j_2}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i, \delta_i = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\}\big]^2 \;\Big|\; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\Big)
$$

$$
\leq \frac{2}{n^2}\sum_{i=1}^{N}\mathbb{E}\Big(\delta_i\big[\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\|^2 + \|\mathbb{E}\{\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i, \delta_i = 1; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|^2\big] \;\Big|\; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\Big)
$$

$$
\leq \frac{2}{n^2}\sum_{i=1}^{N}\mathbb{E}\Big(\delta_i[\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\|^2 + \mathbb{E}\{\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\|^2 \mid \boldsymbol{x}_i, \delta_i = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}})] \;\Big|\; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\Big)
$$

$$
= \frac{4N}{n^2}\mathbb{E}\Big\{\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^2 \;\Big|\; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\Big\} = o_P(1),
$$

where the last step is because (19) of Assumption 1' and the fact that the $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$ is independent of the data imply that $Nn^{-1}\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|\ddot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^2 \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} = O_P(1)$. Thus, $\Delta_2 \xrightarrow{P} \mathbf{0}$.

The partial derivative of $f(y_i \mid \boldsymbol{x}_i, \delta_i = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$ with respect to $\boldsymbol{\theta}$ is

$$
\dot{f}(y_i \mid \boldsymbol{x}_i, \delta_i = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}})
$$

$$
= \frac{\dot{f}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})}{\mathbb{E}\{\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}}
$$

$$
- \frac{f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\pi(\boldsymbol{x}_i, y; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}}{\mathbb{E}^2\{\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}}
$$

$$
= \frac{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})}{\mathbb{E}\{\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}}
$$

$$
- \frac{f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\mathbb{E}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\pi(x_i, y; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}}{\mathbb{E}^2\{\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}}.
$$

Thus, we have that

$$
\mathbf{H}_N^{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}} = \frac{1}{n}\sum_{i=1}^{N}\delta_i\bigg[\int \dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\dot{f}^{\text{T}}(y \mid \boldsymbol{x}_i, \delta_i = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\mathrm{d}y\bigg]
$$

$$
= \frac{1}{n}\sum_{i=1}^{N}\delta_i\bigg[\frac{\mathbb{E}\{\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}}{\mathbb{E}\{\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}}
$$

$$- \frac{\mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)\pi(x_i, y; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}}{\mathbb{E}^2\{\pi_N(\boldsymbol{x}_i, y_i; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}}\bigg].$$

Now we exam the limit of $\mathbf{H}_N^{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}$. First, the expectation satisfies that

$$\mathbb{E}(\mathbf{H}_N^{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$$
$$= \frac{N}{n}\mathbb{E}\left[\dot{\ell}^{\otimes 2}(\boldsymbol{\theta}; \boldsymbol{x}, y)\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) - \frac{\mathbb{E}^{\otimes 2}\{\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) \mid \boldsymbol{x}, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}}{\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) \mid \boldsymbol{x}, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}} \,\bigg|\, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\right] = \boldsymbol{\Sigma}_{N\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}}.$$

By the continuous mapping theorem, Assumption 3', and the fact that $\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}$ is independent of the data, we know that $\mathbb{E}(\mathbf{H}_N^{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \boldsymbol{\Sigma}_{N\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p}$.

For the $j$-th diagonal element of $\mathbf{H}_{N, jj}$, using (A.16), we have

$$\mathbb{V}(\mathbf{H}_{N, jj}^{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})$$
$$= \frac{N}{n^2}\mathbb{V}\left[\delta\mathbb{E}\{\dot{\ell}_j^2(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\} - \delta\mathbb{E}^2\{\dot{\ell}_j(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\} \,\Big|\, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\right]$$
$$\leq \frac{4N}{n^2}\mathbb{E}\left[\delta\mathbb{E}^2\{\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^2 \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}\right]$$
$$\leq \frac{4N}{n^2}\mathbb{E}\left[\delta\mathbb{E}\{\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4 \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}\right]$$
$$= \frac{4N}{n^2}\mathbb{E}\{\delta\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4 \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\}$$
$$= \frac{4N}{n^2}\mathbb{E}\{\pi_N(\boldsymbol{x}, y; \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}})\|\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, y)\|^4 \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\} = o_P(1),$$

where the last step is from (A.21). Thus, noting that $\mathbf{H}_N^{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} \geq \mathbf{0}$ and $\boldsymbol{\Sigma}_{N\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p} + o_P(1)$, we know $\mathbf{H}_N^{\tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p} + o_P(1)$. Therefore,

$$-\frac{1}{n}\ddot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p} + o_P(1) \tag{A.25}$$

From (A.17), (A.22) and (A.25), applying the Basic Corollary in page 2 of Hjort and Pollard (2011) gives that for any constant vector $c$,

$$\mathbb{P}\big\{\sqrt{n}\boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p}(\hat{\boldsymbol{\theta}}_{s, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} - \boldsymbol{\theta}) \leq c \,\big|\, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}\big\} \xrightarrow{P} \Phi(c),$$

where $\Phi$ is the multivariate standard normal distribution function. Since a probability is bounded, this implies that

$$\mathbb{P}\big\{\sqrt{n}\boldsymbol{\Sigma}_{\boldsymbol{\theta}, \boldsymbol{\vartheta}_p}(\hat{\boldsymbol{\theta}}_{s, \tilde{\boldsymbol{\vartheta}}_{\mathrm{plt}}} - \boldsymbol{\theta}) \leq c\big\} \xrightarrow{P} \Phi(c),$$

and this finishes the proof.

## A.4 Proof of Theorem 6

Note that $\ell_S(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = \log\{f(y \mid \boldsymbol{x}, \delta = 1, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}; \boldsymbol{\theta})\}$. Taking partial derivatives of $\mathbb{E}\{\delta \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} = \boldsymbol{0}$, we have

$$\boldsymbol{0} = \mathbb{E}\{\delta \dot{\boldsymbol{U}}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \boldsymbol{x}, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} + \mathbb{E}\{\delta \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y)\dot{\ell}_S^{\text{T}}(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \boldsymbol{x}, \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\},$$

which implies

$$\mathbb{E}\{\delta \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y)\dot{\ell}_S^{\text{T}}(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\} = -\mathbb{E}\{\delta \dot{\boldsymbol{U}}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}.$$

Hence, by calculating the variance covariance matrix conditional on $\tilde{\boldsymbol{\vartheta}}_{\text{plt}}$,

$$\mathbb{V}\big\{\delta \mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) + \delta \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\big\}$$

$$= \mathbb{V}\big\{\delta \mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\big\} + \mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbb{E}\{\delta \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y)\dot{\ell}_S^{\text{T}}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$$

$$+ \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbb{E}\{\delta \dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}^{\text{T}}(\boldsymbol{\theta}; \boldsymbol{x}, y) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}(\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1})^{\text{T}}$$

$$+ \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbb{V}\{\delta \dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}\}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$$

$$= nN^{-1}\Big\{\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbf{V}_{N,\boldsymbol{\theta},\boldsymbol{\vartheta}}(\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1})^{\text{T}} - \mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbf{M}_{N,\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$$

$$- \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbf{M}_{N,\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}^{\text{T}}(\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1})^{\text{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\boldsymbol{\Sigma}_{N\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\Big\}\{1 + o_P(1)\}$$

$$= nN^{-1}\big\{\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbf{V}_{\boldsymbol{\theta},\boldsymbol{\vartheta}}(\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1})^{\text{T}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\big\}\{1 + o_P(1)\},$$

where the last step is from the continuous mapping theorem. Thus,

$$\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbf{V}_{\boldsymbol{\theta},\boldsymbol{\vartheta}}(\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1})^{\text{T}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1} = n^{-1}N\mathbb{V}(\delta \mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}T_{N\tilde{\boldsymbol{\vartheta}}_{\text{plt}}} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\{1 + o_P(1)\},$$

where

$$T_{N\tilde{\boldsymbol{\vartheta}}_{\text{plt}}} = \boldsymbol{U}_{\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}(\boldsymbol{\theta}; \boldsymbol{x}, y) + \mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\dot{\ell}_S(\boldsymbol{\theta}; \boldsymbol{x}, y \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}).$$

Letting $N \to \infty$, we know that $\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\mathbf{V}_{\boldsymbol{\theta},\boldsymbol{\vartheta}}(\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1})^{\text{T}} \geq \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$ in the Loewner ordering, and the equality holds if $n^{-1}N\mathbb{E}(\delta\|T_{N\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}\|^2 \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = o_P(1)$ because $\|\mathbb{V}(\delta \mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}T_{N\tilde{\boldsymbol{\vartheta}}_{\text{plt}}} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\| \leq \|\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\|^2\|\mathbb{V}(\delta T_{N\tilde{\boldsymbol{\vartheta}}_{\text{plt}}} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\| \leq \|\mathbf{M}_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}\|^2\|\mathbb{E}(\delta\|T_{N\tilde{\boldsymbol{\vartheta}}_{\text{plt}}}\|^2 \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\|$.

## A.5 Technical details for examples

### A.5.1 DERIVATION OF EQUATION (66) FOR MULTI-CLASS LOGISTIC REGRESSION

**Derivation for the specific $\pi_N(\boldsymbol{x}_i, \boldsymbol{y}_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$ in (66):**
The sampled data log-likelihood function for $\boldsymbol{\theta}$ (up to an additive constant) is

$$\ell_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = \sum_{i=1}^{N}\delta_i\left[(\boldsymbol{y}_i\otimes\boldsymbol{x}_i)^{\text{T}}\boldsymbol{\theta} - \log\left\{\sum_{l=1}^{K}\exp(\boldsymbol{x}_i^{\text{T}}\boldsymbol{\theta}_l)\right\} - \log\mathbb{E}\{\|\boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\| \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}\}\right].$$

By direct calculation, we know that the score function is

$$\dot{\ell}_S(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = \sum_{i=1}^{N}\delta_i\left[\boldsymbol{y}_i - \frac{\mathbb{E}\{\boldsymbol{y}_i\|\boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\| \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}\}}{\mathbb{E}\{\|\boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\| \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}\}}\right] \otimes \boldsymbol{x}_i.$$

41

For $k = 1, ..., K$, when $y_{i,k} = 1$,

$$\|\boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\|^2 = \{1 - p_k(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\}^2 + \sum_{l \neq k} p_l^2(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})$$

$$= 1 - 2p_k(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}) + \sum_{l=1}^K p_l^2(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}) = \exp(2\tilde{g}_{i,k}) \left\{ \sum_{l=1}^K p_l^2(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}) \right\}.$$

Using the above expression, we obtain the following expectations.

$$\mathbb{E}\{y_{i,k}\|\boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\| \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}\} = p_k(\boldsymbol{x}_i, \boldsymbol{\theta}) \exp(\tilde{g}_{i,k}) \left\{ \sum_{l=1}^K p_l^2(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}) \right\}^{1/2}, \quad \text{and}$$

$$\mathbb{E}\{\|\boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\| \mid \boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}\} = \left\{ \sum_{k=1}^K p_k(\boldsymbol{x}_i, \boldsymbol{\theta}) \exp(\tilde{g}_{i,k}) \right\} \left\{ \sum_{l=1}^K p_l^2(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}}) \right\}^{1/2}.$$

Thus, the ratio is

$$\frac{\mathbb{E}\{y_{i,k}\|\boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\| \mid \boldsymbol{x}_i\}}{\mathbb{E}\{\|\boldsymbol{y}_i - \boldsymbol{p}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}}_{\text{plt}})\| \mid \boldsymbol{x}_i\}} = \frac{p_k(\boldsymbol{x}_i, \boldsymbol{\theta}) \exp(\tilde{g}_{i,k})}{\sum_{k=1}^K p_k(\boldsymbol{x}_i, \boldsymbol{\theta}) \exp(\tilde{g}_{i,k})} = \frac{\exp(\boldsymbol{x}_i^{\text{T}} \boldsymbol{\theta}_k + \tilde{g}_{i,k})}{\sum_{l=1}^K \exp(\boldsymbol{x}_i^{\text{T}} \boldsymbol{\theta}_l + g_{i,l})} = \tilde{p}_{i,k}^g.$$

**Derivation for general $\pi_N(\boldsymbol{x}_i, \boldsymbol{y}_i; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})$:**
From the facts of $\bar{\pi}_N(\boldsymbol{x}_i; \boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) = \sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta})$ and $\partial p_l(\boldsymbol{x}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta} = p_l(\boldsymbol{x}_i, \boldsymbol{\theta})\{\boldsymbol{1} - \boldsymbol{p}(\boldsymbol{x}_i, \boldsymbol{\theta})) \otimes \boldsymbol{x}_i$, we know that

$$\frac{\partial \log \bar{\pi}_N(\boldsymbol{x}_i; \boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})}{\partial \boldsymbol{\theta}} = \frac{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) \partial p_l(\boldsymbol{x}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}}{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta})} \tag{A.26}$$

$$= \frac{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta})\{\boldsymbol{1}_l - \boldsymbol{p}(\boldsymbol{x}_i, \boldsymbol{\theta})) \otimes \boldsymbol{x}_i}{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta})} \tag{A.27}$$

$$= \frac{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{1}_l}{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta})} \otimes \boldsymbol{x}_i - \boldsymbol{p}(\boldsymbol{x}_i, \boldsymbol{\theta}) \otimes \boldsymbol{x}_i. \tag{A.28}$$

Note that with $\tilde{g}_{i,k} = \log\{\pi_N(\boldsymbol{x}_i, \boldsymbol{1}_k; \tilde{\boldsymbol{\vartheta}}_{\text{plt}})\}$, the $k$-th element of

$$\frac{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{1}_l}{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta})} \tag{A.29}$$

is

$$\frac{\pi_N(\boldsymbol{x}_i, \boldsymbol{1}_k; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_k(\boldsymbol{x}_i, \boldsymbol{\theta})}{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) p_l(\boldsymbol{x}_i, \boldsymbol{\theta})} = \frac{\pi_N(\boldsymbol{x}_i, \boldsymbol{1}_k; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) e^{\boldsymbol{x}_i^{\text{T}} \boldsymbol{\theta}_k}}{\sum_{l=1}^K \pi_N(\boldsymbol{x}_i, \boldsymbol{1}_l; \tilde{\boldsymbol{\vartheta}}_{\text{plt}}) e^{\boldsymbol{x}_i^{\text{T}} \boldsymbol{\theta}_l}} = \frac{e^{\boldsymbol{x}_i^{\text{T}} \boldsymbol{\theta}_k + \tilde{g}_{i,k}}}{\sum_{l=1}^K e^{\boldsymbol{x}_i^{\text{T}} \boldsymbol{\theta}_l + \tilde{g}_{i,l}}} = \tilde{p}_{i,k}^g. \tag{A.30}$$

Thus the specific expression of (62) gives

$$\dot{\ell}(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i) - \frac{\partial \log \bar{\pi}_N(\boldsymbol{x}_i; \boldsymbol{\theta} \mid \tilde{\boldsymbol{\vartheta}}_{\text{plt}})}{\partial \boldsymbol{\theta}} = \{\boldsymbol{y}_i - \boldsymbol{p}_i^g(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}_{\text{plt}})\} \otimes \boldsymbol{x}_i, \tag{A.31}$$

which finishes the proof.

A.5.2 Derivations of equations (74), (75), and (76) for Poisson regression

For non-negative integers $m$ and $k$, denote

$$q(m, k) = \sum_{y=0}^{m} \frac{y^k e^{-\mu} \mu^y}{y!}.$$

We have

$$
\begin{aligned}
q(m, k) &:= \sum_{y=0}^{m} \frac{e^{-\mu} \mu^y}{y!} y^k = \mu \sum_{y=1}^{m} \frac{e^{-\mu} \mu^{y-1}}{(y-1)!} y^{k-1} \\
&= \mu \sum_{y=0}^{m-1} \frac{e^{-\mu} \mu^y}{y!} (y+1)^{k-1} = \sum_{l=0}^{k-1} \binom{k-1}{l} \mu \sum_{y=0}^{m-1} \frac{e^{-\mu} \mu^y}{y!} y^l \\
&= \mu \sum_{l=0}^{k-1} \binom{k-1}{l} q(m-1, l).
\end{aligned}
$$

Thus, we know that

$$
\begin{aligned}
q(m, 0) &= F(m; \mu), \\
q(m, 1) &= \mu F(m-1; \mu), \\
q(m, 2) &= \mu F(m-1; \mu) + \mu^2 F(m-2; \mu), \\
q(m, 3) &= \mu F(m-1; \mu) + 3\mu^2 F(m-2; \mu) + \mu^3 F(m-3; \mu).
\end{aligned}
$$

Now, we derive the expectations. First,

$$
\begin{aligned}
\mathbb{E}(|y_i - \tilde{\mu}_i| \mid \boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}_{\text{plt}}) &= \sum_{y_i=0}^{\infty} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} |y_i - \tilde{\mu}_i| = 2 \sum_{y_i=0}^{m_i} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} (\tilde{\mu}_i - y_i) + \mu_i - \tilde{\mu}_i \\
&= 2\tilde{\mu}_i q(m_i, 0) - 2q(m_i, 1) + \mu_i - \tilde{\mu}_i \\
&= 2\tilde{\mu}_i F(m_i; \mu_i) - 2\mu_i F(m_i - 1; \mu_i) + \mu_i - \tilde{\mu}_i \\
&= 2(\tilde{\mu}_i - \mu_i) F(m_i - 1; \mu_i) + 2\tilde{\mu}_i f(m_i; \mu_i) + \mu_i - \tilde{\mu}_i.
\end{aligned}
$$

Second,

$$
\mathbb{E}(y_i|y_i - \tilde{\mu}_i| \mid \boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}_{\text{plt}}) = \sum_{y_i=0}^{\infty} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} (y_i|y_i - \tilde{\mu}_i|)
$$

$$
\begin{aligned}
&= 2 \sum_{y_i=0}^{m_i} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} (\tilde{\mu}_i y_i - y_i^2) + \mathbb{E}(y_i^2 - y_i \tilde{\mu}_i \mid \boldsymbol{x}_i) \\
&= 2\tilde{\mu}_i q(m_i, 1) - 2q(m_i, 2) + \mu_i + \mu_i^2 - \mu_i \tilde{\mu}_i \\
&= 2\tilde{\mu}_i \mu_i F(m_i - 1; \mu_i) - 2\{\mu_i F(m_i - 1; \mu_i) + \mu_i^2 F(m_i - 2; \mu_i)\} + \mu_i + \mu_i^2 - \mu_i \tilde{\mu}_i \\
&= 2\mu_i (\tilde{\mu}_i - 1) F(m_i - 1; \mu_i) - 2\mu_i^2 F(m_i - 2; \mu_i) + \mu_i + \mu_i^2 - \mu_i \tilde{\mu}_i.
\end{aligned}
$$

Third,

$$
\begin{aligned}
\mathbb{E}(y_i^2|y_i - \tilde{\mu}_i| \mid \boldsymbol{x}_i; \tilde{\boldsymbol{\theta}}_{\mathrm{plt}}) &= \sum_{y_i=0}^{\infty} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}(y_i^2|y_i - \tilde{\mu}_i|) \\
&= 2\sum_{y_i=0}^{m_i} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}(\tilde{\mu}_i y_i^2 - y_i^3) + \mathbb{E}(y_i^3 - y_i^2\tilde{\mu}_i \mid \boldsymbol{x}_i) \\
&= 2\tilde{\mu}_i q(m_i, 2) - 2q(m_i, 3) + \mu_i + 3\mu_i^2 + \mu_i^3 - \tilde{\mu}_i(\mu_i + \mu_i^2) \\
&= 2\tilde{\mu}_i\{\mu_i F(m_i - 1; \mu_i) + \mu_i^2 F(m_i - 2; \mu_i)\} \\
&\quad - 2\{\mu_i F(m_i - 1; \mu_i) + 3\mu_i^2 F(m_i - 2; \mu_i) + \mu_i^3 F(m_i - 3; \mu_i)\} \\
&\quad + \mu_i + 3\mu_i^2 + \mu_i^3 - \tilde{\mu}_i(\mu_i + \mu_i^2) \\
&= 2\mu_i(\tilde{\mu}_i - 1)F(m_i - 1; \mu_i) + 2\mu_i^2(\tilde{\mu}_i - 3)F(m_i - 2; \mu_i) - 2\mu_i^3 F(m_i - 3; \mu_i) \\
&\quad + \mu_i + 3\mu_i^2 + \mu_i^3 - \tilde{\mu}_i(\mu_i + \mu_i^2).
\end{aligned}
$$

## Appendix B. Additional numerical experiments on pilot misspecifications

We carried out additional numerical experiments to investigate the effect of the magnitude of pilot misspecification on different subsample estimators.

### B.1 multi-class logistic regression

We used exactly the same setup as in Section 8.1 with $n_0 = 400$ and $n = 2,000$ in this experiment. The only difference is that the pilot is set to be the true parameter plus $\lambda$ times a vector of ones. With this setting, $\lambda$ controls the level of pilot misspecification with $\lambda = 0$ corresponding to the consistent pilot. We used the same approach used in Section 8.1 to calculate the empirical MSEs, variances, and squared biases. The relative pattern between the variances and the squared biases are the same as in Section 8.1, so we report the empirical MSEs here only. We omit the results for the native estimator due to its high bias.

Figure A.1 reports the results. The MSCLE dominates the corresponding IPW estimator for both GN and LUS probabilities uniformly in $\lambda$. For MSCLE, it has good performance with a consistent pilot ($\lambda = 0$), but its best performance may not always be achieved in this case. The reason is that $\Sigma_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$ in Theorem 5 may not necessarily be minimized at the true parameter. Actually, for the MSCLE, there is no general solution to $\boldsymbol{\vartheta}_p$ so that $\Sigma_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$ is minimized.

Note that $\Sigma_{\boldsymbol{\theta},\boldsymbol{\vartheta}_p}^{-1}$ depends on $\boldsymbol{\vartheta}_p$ through $\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}_p)$ so the problem of find the optimal $\boldsymbol{\vartheta}_p$ is essentially finding the optimal sampling probability $\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}_p)$ for the MSCLE. The problem is complicated even with a much simplified scenario of noninformative subsampling so that $\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}_p)$ does not dependent on $y$, i.e., $\pi_N(\boldsymbol{x}, y; \boldsymbol{\vartheta}_p) = \pi_N(\boldsymbol{x}; \boldsymbol{\vartheta}_p)$. In this scenario, the problem of determining the optimal $\pi_N(\boldsymbol{x}; \boldsymbol{\vartheta}_p)$ is called optimal design of experiments (see, e.g., Kiefer, 1959; Pukelsheim, 2006), and the optimal $\pi_N(\boldsymbol{x}; \boldsymbol{\vartheta}_p)$ is binary with possible values of zero and one (Pronzato and Wang, 2021). This topic is beyond the scope our investigation because the primary focus of this study is to propose an improved estimator for informative subsamples.

For the IPW estimator, we see that the optimal performance is not achieved with $\lambda = 0$ for cases (b) and (d). This is because the GM sampling probabilities in (65) minimize the variance of a certain linear function of the IPW estimator which is different from the MSE. Nevertheless, for the IPW estimator, the optimal sampling probabilities exist under some optimality criteria such as the A- and L- optimality, and it requires the pilot to be consistent to achieve the optimal variance. The proposed MSCLE has a smaller variance matrix than the weighted estimator. Thus, we can only conclude that the asymptotic variance of the MSCLE has a smaller upper bound with a consistent pilot than with a misspecified pilot.
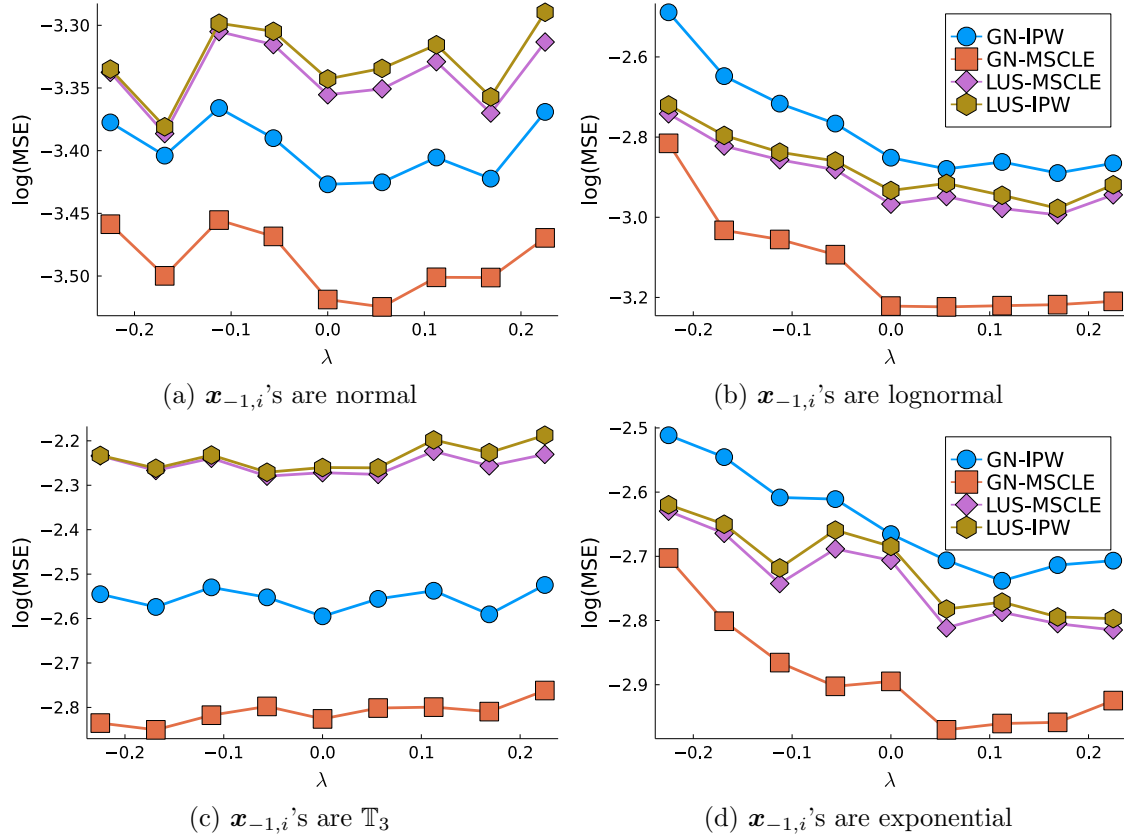


Figure A.1: Log of empirical MSEs of subsample estimators with $n_0 = 400$ and $n = 2,000$ in multi-class logistic regression when the misspecified pilot estimator is set to be the true parameter plus a vector of $\lambda$'s.

## B.2 Poisson regression

We used exactly the same setup and procedure as in Section 8.2 to generate data and calculate the empirical MSEs, variances, and squared biases for $n_0 = 400$ and $n = 2,000$. The only difference is that the pilot is set to be the true parameter plus $\lambda$ times a vector of ones, so that $\lambda$ controls the level of pilot misspecification with $\lambda = 0$ corresponding to the consistent pilot. Again, we report the empirical MSEs here only and omit the results for the native estimator.

Figure A.2 reports the results. The MSCLE dominates the corresponding IPW estimator uniformly, and both methods achieve the best performance for all the four cases with a consistent pilot ($\lambda = 0$) in this example.

We also consider the case of misspecified pilot estimator using the same procedure considered in Section 8.1. The resulting impacts are similar to those observed in Section 8.1 so we omit the results.



(a) $\boldsymbol{x}_{-1,i}$'s are uniform

(b) $\boldsymbol{x}_{-1,i}$'s are Beta

(c) $\boldsymbol{x}_{-1,i}$'s are normal

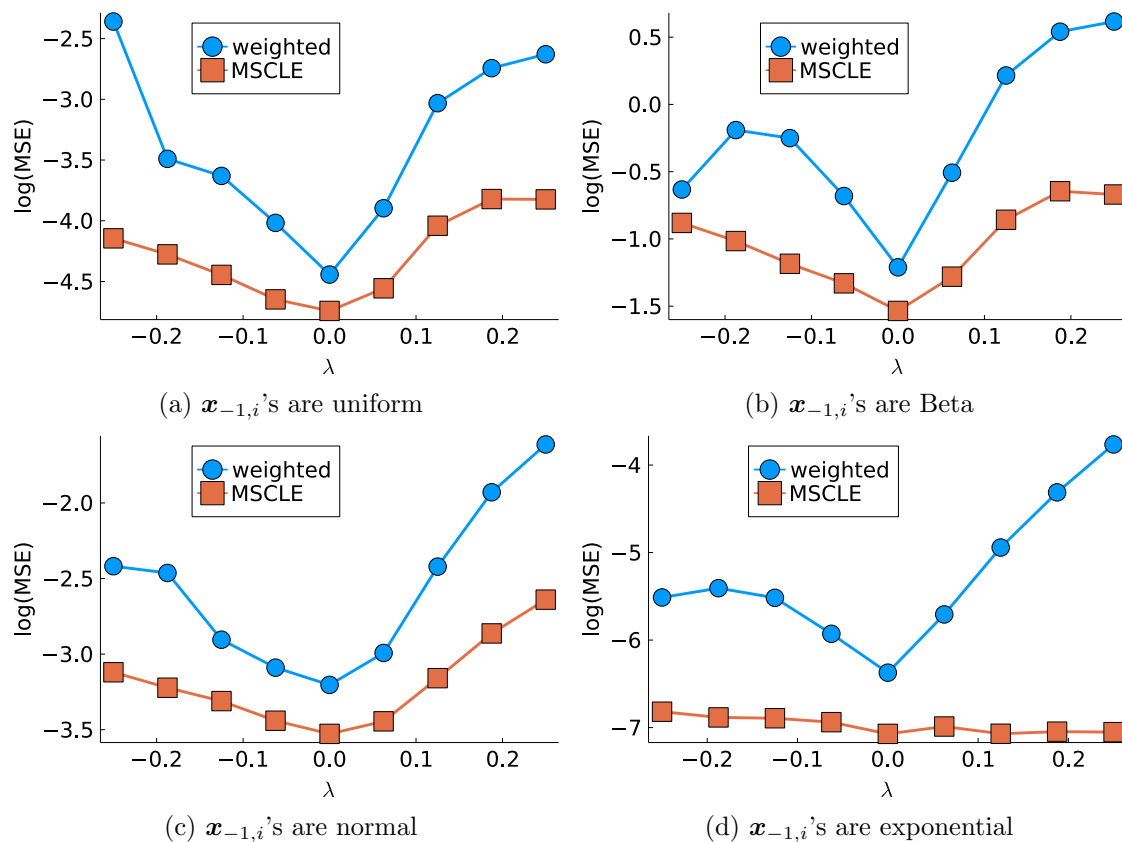(d) $\boldsymbol{x}_{-1,i}$'s are exponential

Figure A.2: Log of empirical MSEs of subsample estimators with $n_0 = 400$ and $n = 2,000$ in Poisson regression when the misspecified pilot estimator is set to be the true parameter plus a vector of $\lambda$'s.

# References

Mingyao Ai, Jun Yu, Huiming Zhang, and HaiYing Wang. Optimal subsampling algorithms for big data regressions. *Statistica Sinica*, 31(2):749–772, 2021. doi: 10.5705/ss.202018. 0439.

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

Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.

N.E. Breslow and R. Holubkov. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B*, 59:447–461, 1997.

R. L. Chambers and C. J. Skinner, editors. *Analysis of Survey Data*. John Wiley & Sons, Chichester, U. K., 2003.

D. R. Cox. Some sampling problems in technology. In U.L. Johnson and H. Smith, editors, *New Developments in Survey Sampling*, pages 506–527. Wiley Interscience, 1969.

Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar. New subsampling algorithms for fast least squares regression. In *Advances in Neural Information Processing Systems*, pages 360–368, 2013.

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

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

Pavlos S. Efraimidis and Paul G. Spirakis. Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5):181–185, 2006.

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

Lei Han, Kean Ming Tan, Ting Yang, and Tong Zhang. Local uncertainty sampling for large-scale multiclass logistic regression. *The Annals of Statistics*, 48(3):1770–1788, 2020.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

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

D.A. Hsieh, C.F. Manski, and D. McFadden. Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association*, 80:651–662, 1985.

X.J. Hu and J.F. Lawless. Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika*, 83:651–662, 1996.

X.J. Hu and J.F. Lawless. Pseudo likelihood estimation in a class of problems with response-related missing covariates. *Canadian Journal of Statistics*, 25:125–142, 1997.

Mike Innes. Flux: Elegant machine learning with julia. *Journal of Open Source Software*, 2018. doi: 10.21105/joss.00602.

J.D. Kalbfleisch and J.F. Lawless. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7:149–160, 1988.

J. Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society. Series B*, 21(2):272–319, 1959.

J.K. Kim, A. Navarro, and W.A. Fuller. Replicate variance estimation after multi-phase stratified sampling. *Journal of the American Statistical Association*, 101:312–320, 2006.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

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

Rui Nie, Douglas P Wiens, and Zhichun Zhai. Minimax robust active learning for approximately specified regression models. *Canadian Journal of Statistics*, 46(1):104–122, 2018.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

D. Pfeffermann, A. M. Krieger, and Y. Rinnot. Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8:1087–1114, 1998.

Luc Pronzato and HaiYing Wang. Sequential online subsampling for thinning experimental designs. *Journal of Statistical Planning and Inference*, 212:169 – 193, 2021. ISSN 0378-3758. doi: 10.1016/j.jspi.2020.08.001. URL `http://www.sciencedirect.com/science/article/pii/S0378375820300999`.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Jing Qin. *Biased sampling, over-identified parameter problems and beyond*. Springer, 2017.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

T. Saegusa and J. A. Wellner. Weighted likelihood estimation under two-phase sampling. *The Annals of Statistics*, 41:269–295, 2013.

A. J. Scott and C. J. Wild. Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society. Series B*, 48(2):170–182, 1986.

A.J. Scott and C.J. Wild. Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47:497–510, 1991.

A.J. Scott and C.J. Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84:57–71, 1997.

Xinwei Shen, Kani Chen, and Wen Yu. Surprise sampling: Improving and extending the local case-control sampling. *Electronic Journal of Statistics*, 15(1):2454–2482, 2021.

Daniel Ting and Eric Brochu. Optimal subsampling with influence functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/57c0531e13f40b91b3b0f1a30b529a1d-Paper.pdf`.

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

HaiYing Wang. Logistic regression for massive data with rare events. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9829–9836. PMLR, 13–18 Jul 2020. URL `http://proceedings.mlr.press/v119/wang20a.html`.

Haiying Wang and Yanyuan Ma. Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112, 2021.

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

HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525): 393–405, 2019. doi: 10.1080/01621459.2017.1408468.

HaiYing Wang, Aonan Zhang, and Chong Wang. Nonuniform negative sampling and log odds correction with rare events data. In *Proceedings of The 35 Conference on Neural Information Processing Systems (NeurIPS 2021).*, Proceedings of Machine Learning Research. PMLR, 2021.

Jing Wang, Jiahui Zou, and HaiYing Wang. Sampling with replacement vs poisson sampling: A comparative study in optimal subsampling. *IEEE Transactions on Information Theory*, 68(10):6605–6630, 2022. doi: 10.1109/TIT.2022.3176955.

A.S. Whittemore. Multistage sampling designs and estimating equations. *Journal of the Royal Statistical Society. Series B*, 59:589–602, 1997.

C.J. Wild. Fitting prospective regression models to case-control data. *Biometrika*, 78: 705–717, 1991.

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

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

Yaqiong Yao and HaiYing Wang. A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, 19(1):151–172, 2021.

Jun Yu, HaiYing Wang, Mingyao Ai, and Huiming Zhang. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 117(537):265–276, 2022. doi: 10.1080/01621459.2020.1773832. URL https://doi.org/10.1080/01621459.2020.1773832.