

Nonuniform Negative Sampling and Log Odds Correction with Rare Events Data

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

Aonan Zhang
ByteDance Inc.
aonan.zhang@bytedance.com

Chong Wang
ByteDance Inc.
chong.wang@bytedance.com

Abstract

We investigate the issue of parameter estimation with nonuniform negative sampling for imbalanced data. We first prove that, with imbalanced data, the available information about unknown parameters is only tied to the relatively small number of positive instances, which justifies the usage of negative sampling. However, if the negative instances are subsampled to the same level of the positive cases, there is information loss. To maintain more information, we derive the asymptotic distribution of a general inverse probability weighted (IPW) estimator and obtain the optimal sampling probability that minimizes its variance. To further improve the estimation efficiency over the IPW method, we propose a likelihood-based estimator by correcting log odds for the sampled data and prove that the improved estimator has the smallest asymptotic variance among a large class of estimators. It is also more robust to pilot misspecification. We validate our approach on simulated data as well as a real click-through rate dataset with more than 0.3 trillion instances, collected over a period of a month. Both theoretical and empirical results demonstrate the effectiveness of our method.

1 Introduction

Imbalanced data are ubiquitous in scientific fields and applications with binary response outputs, where the number of instances in the positive class is much smaller than that in the negative class. For example, in online search or recommendation systems where billions of impressions appear each day, non-click impressions usually dominate. Using these non-click impressions as negative instances is prevalent and proved to be useful especially in modern machine learning systems (McMahan *et al.*, 2013; He *et al.*, 2014; Chen *et al.*, 2016; Guo *et al.*, 2017; Huang *et al.*, 2020). A common approach to address imbalanced data is to balance it by subsampling the negative class (Drummond *et al.*, 2003; Liu *et al.*, 2009) and/or up-sampling the positive class (Chawla *et al.*, 2002; Han *et al.*, 2005; Mathew *et al.*, 2017; Douzas and Bacao, 2017), and a great deal of attentions have been drawn to this problem, see Japkowicz (2000); King and Zeng (2001); Chawla *et al.* (2004); Estabrooks *et al.* (2004); Owen (2007); Sun *et al.* (2007); Chawla (2009); Rahman and Davis (2013); Fithian and Hastie (2014); Lemaître *et al.* (2017); Wang (2020), and the references therein.

In this paper, we focus on subsampling the negative class, i.e., negative sampling, while keeping all rare positive instances. This is beneficial for modern online learning systems, where negative sampling significantly reduces data volume in distributed storage while saving training time for multiple downstream models. Rare events data and negative sampling are studied in Wang (2020),

but it focused on linear logistic regression and only considered uniform sampling. We approach this problem under general binary response models from the perspective of optimal subsampling, which aims to minimize the asymptotic variance of the IPW estimator (Wang *et al.*, 2018; Ting and Brochu, 2018; Wang and Ma, 2021; Yu *et al.*, 2020). This topic was not well investigated neither for imbalanced data nor for negative sampling. In addition, existing optimal probability formulation minimizes the conditional asymptotic variance and the variations due to the data randomness are ignored. We fill these gaps by deriving the asymptotic distributions for both full data and subsample estimators under general binary response models, and we find the unconditional optimal probability for negative sampling. In addition, since IPW may inflate the variance (Hesterberg, 1995; Owen and Zhou, 2000), we develop a more efficient likelihood-based estimator through nonuniform log odds correction to avoid IPW. Our main contributions are summarized as follows.

- Under a general binary response model with rare events data, we find that the difference between the full data estimator and the true parameter converges to a normal distribution at a rate that is tied to the number of rare positive instances only. This indicates the possibility to throw away the majority of negative instances without information loss, i.e, it justifies the usage of negative sampling.
- We show that there is no asymptotic information loss with aggressive negative sampling if the negative instances kept dominates the positive instances. However there is information loss expressed as a variance inflation if the negative instances are brought down to the same level of the positive instances. For this case, we obtain optimal subsampling probabilities by minimizing the unconditional asymptotic variance of a general IPW estimator.
- We develop a likelihood estimator through nonuniform log odds correction for sampled data, and prove that it has the smallest asymptotic variance among a large class of estimators.
- We apply the proposed method to a real online streaming click-through rate (CTR) dataset with more than 0.3 trillion instances and demonstrate its effectiveness.

The rest of paper is organized as follows. Section 2 defines the problem this paper focuses on and shows the full data results. Section 3 presents the optimal probability for the IPW estimator. Section 4 proposes the likelihood-based estimator and establishes its asymptotic optimality. Section 5 discusses practical implementation and the theoretical properties of the resulting estimators. Section 6 presents numerical experiments with simulated data, and applies the proposed method to a real online streaming CTR dataset with more than 0.3 trillion instances. Section 7 concludes the paper and points out limitations of our investigation. Proofs and additional experiments with misspecifications are in the supplementary materials.

2 Problem setup and full data results

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be training data of size N that satisfies the binary response model,

$$\Pr(y = 1 \mid \mathbf{x}) = p(\mathbf{x}; \boldsymbol{\theta}), \quad (1)$$

where $y \in \{0, 1\}$ is the binary class label ($y = 1$ for the positive and $y = 0$ for the negative), \mathbf{x} is the vector of features, and $\boldsymbol{\theta}$ is the unknown d -dimensional parameter.

Let N_1 be the number of positive instances ($N_1 = \sum_{i=1}^N y_i$) and N_0 be the number of negative instances ($N_0 = N - N_1$). We consider the scenario of massive imbalanced data, i.e., $N_1 \ll N_0$. For asymptotic investigations when N_1 is much smaller than N_0 , it is more appropriate to assume that N_1 increases in a slower rate compared with N_0 , i.e., $N_1/N_0 \rightarrow 0$ in probability as $N \rightarrow \infty$ (Wang, 2020). This requires $\Pr(y = 1) = \mathbb{E}\{p(\mathbf{x}; \boldsymbol{\theta})\} \rightarrow 0$ as $N \rightarrow \infty$ on the model side. We assume that the parameter $\boldsymbol{\theta}$ contains two components $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$ and the log odds can be written as

$$g(\mathbf{x}; \boldsymbol{\theta}) := \log \left\{ \frac{p(\mathbf{x}; \boldsymbol{\theta})}{1 - p(\mathbf{x}; \boldsymbol{\theta})} \right\} = \alpha + f(\mathbf{x}; \boldsymbol{\beta}),$$

where the true parameter, denoted as $\boldsymbol{\theta}^* = (\alpha^*, \boldsymbol{\beta}^{*T})^T$, satisfies that $\alpha^* \rightarrow -\infty$ as $N \rightarrow \infty$ and $\boldsymbol{\beta}^*$ is fixed. Here $f(\mathbf{x}; \boldsymbol{\beta})$ is a smooth function of $\boldsymbol{\beta}$, such as a linear function or a nonlinear neural network model. If it is linear, the model reduces to the logistic regression model. A diverging α^* and a fixed $\boldsymbol{\beta}^*$ indicate that both the marginal and conditional probabilities for a positive instance are

small. Heuristically, this implies that a change in the feature value does make a rare event become a large probability event. We can also allow β^* to change with N , but as long as β^* has a finite limit, the situation is essentially the same as a fixed β^* . So here we assume β^* to be fixed to simplify the presentation.

Throughout this paper, we use $\dot{g}(\mathbf{x}; \boldsymbol{\theta})$ and $\ddot{g}(\mathbf{x}; \boldsymbol{\theta})$ to denote the gradient and Hessian of $g(\mathbf{x}; \boldsymbol{\theta})$ with respect to (w.r.t.) $\boldsymbol{\theta}$, respectively. Let $\|\mathbf{v}\|$ be the Frobenius norm and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$ for a vector or a matrix \mathbf{v} . We denote $\pi(\mathbf{x}, y)$ as the sampling probability for an instance (\mathbf{x}, y) . For negative cases, we sometimes use a shorthand notation $\pi(\mathbf{x})$ to denote its sampling probability. We use $o_P(1)$ to represent a random quantity that converges to 0 in probability as $N \rightarrow \infty$.

2.1 It is the number of positive instances that matters

We show that for rare events data, the estimation error rate for $\boldsymbol{\theta}$ is related to N_1 instead of N . Using the full data, a commonly used estimator of $\boldsymbol{\theta}$ is the maximum likelihood estimator (MLE)

$$\hat{\boldsymbol{\theta}}_f := \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^N [y_i g(\mathbf{x}_i; \boldsymbol{\theta}) - \log \{1 + e^{g(\mathbf{x}_i; \boldsymbol{\theta})}\}].$$

To investigate the theoretical properties of the estimator $\hat{\boldsymbol{\theta}}_f$, we make the following assumptions.

Assumption 1. The first, second, and third derivatives of $f(\mathbf{x}; \boldsymbol{\beta})$ and $e^{f(\mathbf{x}; \boldsymbol{\beta})} f(\mathbf{x}; \boldsymbol{\beta})$ w.r.t. any components of $\boldsymbol{\beta}$ are bounded by a square integrable random variable $B(\mathbf{x})$.

Assumption 2. The matrix $\mathbb{E}\{\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\}$ is finite and positive definite.

Assumption 1 imposes constraints on the smoothness of $f(\mathbf{x}; \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ and on the tightness of the distribution of \mathbf{x} . Assumption 2 ensures that all components of $\boldsymbol{\theta}$ are estimable.

Theorem 1. Let $\mathbf{V}_f = \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\} \mathbf{M}_f^{-1}$ where $\mathbf{M}_f = \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\}$. Under Assumptions 1 and 2, as $N \rightarrow \infty$,

$$\sqrt{N_1}(\hat{\boldsymbol{\theta}}_f - \boldsymbol{\theta}^*) \longrightarrow \mathbb{N}(\mathbf{0}, \mathbf{V}_f), \quad \text{in distribution.}$$

This result shows that even if all the N instances are used for model training, the resulting estimator $\hat{\boldsymbol{\theta}}_f$ converges to the true parameter $\boldsymbol{\theta}^*$ at the rate of $N_1^{-1/2}$, which is much slower than the usual rate of $N^{-1/2}$ for regular cases. This indicates that with rare events data, the available information about unknown parameters is at the scale of N_1 . Although $N \rightarrow \infty$ much faster than $N_1 \rightarrow \infty$ (in terms of $N_1/N \rightarrow 0$), N does not explicitly show in the asymptotic normality result of Theorem 1. For the specific case of linear logistic regression with $g(\mathbf{x}; \boldsymbol{\theta}) = \alpha + \mathbf{x}^T \boldsymbol{\beta}$, our Theorem 1 reduces to Theorem 1 of Wang (2020).

2.2 Negative sampling

Since the available information ties to the number of positive instances instead of the full data size, one can keep all the positive instances and significantly subsample the negative instances to reduce the computational cost. For better estimation efficiency, we consider nonuniform sampling. Let $\varphi(\mathbf{x}) > 0$ be an integrable function and ρ be the sampling rate on the negative class. Without loss of generality, assume $\mathbb{E}\{\varphi(\mathbf{x})\} = 1$ so that $\pi(\mathbf{x}_i) := \rho\varphi(\mathbf{x}_i)$ is the sampling probability for the i -th data point if $y_i = 1$. We present the nonuniform negative sampling procedure in Algorithm 1.

Algorithm 1 Negative sampling

For $i = 1, \dots, N$: if $y_i = 1$, include $\{\mathbf{x}_i, y_i, \pi(\mathbf{x}_i, y_i) = 1\}$ in the sample;
if $y_i = 0$, calculate $\varphi(\mathbf{x}_i)$ and generate $u_i \sim \mathbb{U}(0, 1)$;
if $u_i \leq \rho\varphi(\mathbf{x}_i)$, include $\{\mathbf{x}_i, y_i, \pi(\mathbf{x}_i, y_i) = \rho\varphi(\mathbf{x}_i)\}$ in the sample.

3 Weighted estimation and its optimal negative sampling probability

Let $\delta_i = 1$ if the i -th data point is selected and $\delta_i = 0$ otherwise. Given a subsample taken according to $\pi(\mathbf{x}_i, y_i) = y_i + (1 - y_i)\rho\varphi(\mathbf{x}_i)$, $i = 1, \dots, N$, the IPW estimator of θ is

$$\hat{\theta}_w = \arg \max_{\theta} \sum_{i=1}^N \delta_i \pi^{-1}(\mathbf{x}_i, y_i) [y_i g(\mathbf{x}_i; \theta) - \log \{1 + e^{g(\mathbf{x}_i; \theta)}\}]. \quad (2)$$

We need an assumption on the subsampling rate to investigate the subsample estimators.

Assumption 3. The subsampling rate ρ satisfies that $c_N := e^{\alpha^*} / \rho \rightarrow c$ for a constant c such that $0 \leq c < \infty$.

Note that c_N cannot be zero but its limit c can be exactly zero. It can be shown that $c_N \mathbb{E}\{e^{f(\mathbf{x}; \beta^*)}\} = N_1(N_0\rho)^{-1}\{1 + o_P(1)\}$, so $c\mathbb{E}\{e^{f(\mathbf{x}; \beta^*)}\}$ is the asymptotic positive/negative ratio for the sample. The theorem below shows the asymptotic distribution of $\hat{\theta}_w$.

Theorem 2. Under Assumptions 1-3, if $\mathbb{E}[\{\varphi(\mathbf{x}) + \varphi^{-1}(\mathbf{x})\}B^2(\mathbf{x})] < \infty$ then as $N \rightarrow \infty$,

$$\sqrt{N_1}(\hat{\theta}_w - \theta^*) \longrightarrow \mathbb{N}(\mathbf{0}, \mathbf{V}_w), \quad \text{in distribution,}$$

where $\mathbf{V}_w = \mathbf{V}_f + \mathbf{V}_{\text{sub}}$ and $\mathbf{V}_{\text{sub}} = c\mathbb{E}\{e^{f(\mathbf{x}; \beta^*)}\}\mathbf{M}_f^{-1}\mathbb{E}\{\varphi^{-1}(\mathbf{x})e^{2f(\mathbf{x}; \beta^*)}\dot{g}^{\otimes 2}(\mathbf{x}; \theta^*)\}\mathbf{M}_f^{-1}$.

The subsampled estimator $\hat{\theta}_w$ have the same convergence rate as the full data estimator. However, the asymptotic variance \mathbf{V}_w may be inflated by the term \mathbf{V}_{sub} from subsampling the negative cases. Here \mathbf{V}_{sub} is zero if $c = 0$. If we keep much more negative instances than the positive instances in the subsample ($c = 0$), then there is no asymptotic information loss due to subsampling ($\mathbf{V}_w = \mathbf{V}_f$). In this scenario, the number of negative instances can still be aggressively reduced ($\rho \rightarrow 0$) so that the computational efficiency is significantly improved, and the sampling function $\varphi(\mathbf{x})$ does not play an significant role. If we have to reduce the negative instances to the same level of the positive instances ($0 < c < \infty$), then the variance inflation \mathbf{V}_{sub} is not negligible. In this scenario, a well designed sampling function $\varphi(\mathbf{x})$ it is more relevant and critical.

Theorem 2 shows that the distribution of the estimation error $\text{err} := \sqrt{N_1}(\hat{\theta}_w - \theta^*)$ is approximated by a normal distribution $\mathbb{N}(\mathbf{0}, \mathbf{V}_w)$. Thus for any $\epsilon > 0$ the probability of excess error $\Pr(\text{err} > \epsilon)$ is approximated by $\Pr\{\|\mathbb{N}(\mathbf{0}, \mathbf{V}_w)\| > \epsilon\} = \Pr(\sum_{j=1}^d \lambda_j z_j^2 > \epsilon)$, where λ_j 's are eigenvalues of \mathbf{V}_w and z_j 's are independent standard normal random variables. Therefore, a smaller trace of \mathbf{V}_w means a smaller probability of excess error of the same level since $\text{tr}(\mathbf{V}_w) = \sum_{j=1}^d \lambda_j$.

The following theorem gives the optimal negative sampling function for the IPW estimator.

Theorem 3. For a given sampling rate ρ , the asymptotically optimal $\varphi(\mathbf{x})$ that minimizes $\text{tr}(\mathbf{V}_w)$ is

$$\varphi_{\text{os}}(\mathbf{x}) = \frac{\min\{t(\mathbf{x}; \theta^*), T\}}{\mathbb{E}[\min\{t(\mathbf{x}; \theta^*), T\}]}, \quad (3)$$

where $t(\mathbf{x}; \theta) = p(\mathbf{x}; \theta)\|\mathbf{M}_f^{-1}\dot{g}(\mathbf{x}; \theta)\|$ and T is the maximum number so that $\rho[\min\{t(\mathbf{x}; \theta^*), T\}] \leq \mathbb{E}[\min\{t(\mathbf{x}; \theta^*), T\}]$ with probability one.

Remark 1. If $\rho \rightarrow 0$ then $\rho t(\mathbf{x}; \theta^*) \leq \mathbb{E}\{t(\mathbf{x}; \theta^*)\}$ almost surely, so T can be dropped (i.e., $T = \infty$). If $\lim_{N \rightarrow \infty} \rho > 0$ then $c = 0$, so the variance inflation due to subsampling is negligible. Thus, the truncation term T can be ignored in practice with imbalanced data; it only plays an role for not very imbalanced data when the sampling ratio is not very small.

In $\varphi_{\text{os}}(\mathbf{x})$, $t(\mathbf{x}; \theta^*)$ consists of two components, $p(\mathbf{x}; \theta^*)$ and $\|\mathbf{M}_f^{-1}\dot{g}(\mathbf{x}; \theta^*)\|$. The term $p(\mathbf{x}; \theta^*)$ is from the binary response model structure, and it gives a higher preference for a data point with larger $p(\mathbf{x}; \theta^*)$ in the negative class. The term $\|\mathbf{M}_f^{-1}\dot{g}(\mathbf{x}; \theta^*)\|$ corresponds to the A optimality criterion (Pukelsheim, 2006). For computational benefit in optimal sampling, the A optimality is often replaced by the L optimality (e.g., Wang *et al.*, 2018; Ting and Brochu, 2018). Under the L optimality, $\|\mathbf{M}_f^{-1}\dot{g}(\mathbf{x}; \theta^*)\|$ is replaced by $\|\dot{g}(\mathbf{x}; \theta^*)\|$, and this minimizes the asymptotic mean squared error (MSE) of $\mathbf{M}_f\hat{\theta}_w$. In case the gradient is difficult to obtain, one could ignore the gradient term and simply use $p(\mathbf{x}; \theta^*)$ to replace $t(\mathbf{x}; \theta^*)$. Although this does not give an optimal sampling probability, it outperforms the uniform subsampling because it takes into account the information of the model structure like the local case-control (LCC) sampling.

4 More efficient estimation based on the likelihood with log odds correction

4.1 Nonuniform log odds correction

The optimal function $\varphi_{\text{os}}(\mathbf{x})$ assigns larger probabilities to more informative instances. However, the IPW estimator in (2) assigns smaller weights to more informative instances in estimation, so the resulting estimator can be improved to have higher estimation efficiency. A naive unweighted estimator is biased, and the bias correction approach in Fithian and Hastie (2014); Wang (2019) does not work for the sampling procedure in Algorithm 1 even when the underlying model is the logistic regression. We adopted the idea of Han *et al.* (2020) and seek to define a more efficient estimator through finding the corrected likelihood of the sampled data based on any negative sampling probability $\pi(\mathbf{x})$. For data included in the subsample (where $\delta = 1$), the conditional probability of $y = 1$ is

$$p_{\pi}(\mathbf{x}; \boldsymbol{\theta}) := \Pr(y = 1 \mid \mathbf{x}, \delta = 1) = \frac{1}{1 + e^{-g(\mathbf{x}; \boldsymbol{\theta}) - l}}, \quad (4)$$

where $l = -\log\{\pi(\mathbf{x})\}$. Please see the detailed derivation of (4) in the supplement. To avoid the IPW, we propose the estimator based on the log odds corrected likelihood in (4), namely,

$$\hat{\boldsymbol{\theta}}_{\text{lik}} = \arg \max_{\boldsymbol{\theta}} \ell_{\text{lik}}(\boldsymbol{\theta}), \quad \text{where} \quad \ell_{\text{lik}}(\boldsymbol{\theta}) = \sum_{i=1}^N \delta_i [y_i g(\mathbf{x}_i; \boldsymbol{\theta}) - \log \{1 + e^{g(\mathbf{x}_i; \boldsymbol{\theta}) + l_i}\}]. \quad (5)$$

With $\hat{\boldsymbol{\theta}}_w$ in (2), $\pi(\mathbf{x}_i)$ is in the inverse. If an instance with much smaller $\pi(\mathbf{x}_i)$ is selected in the sample, it dominates the objective function, making the resulting estimator unstable. With $\hat{\boldsymbol{\theta}}_{\text{lik}}$, this problem is ameliorated because $\pi(\mathbf{x}_i)$ is in the logarithm in the log-likelihood of (5) and $\log(v)$ goes to infinity much slower than v^{-1} as $v \downarrow 0$.

4.2 Theoretical analysis of the likelihood-based estimator

The following shows the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{\text{lik}}$.

Theorem 4. *Under Assumptions 1-3, if $\mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \varphi^{-1}(\mathbf{x}) B(\mathbf{x})\} < \infty$, then as $N \rightarrow \infty$,*

$$\sqrt{N_1}(\hat{\boldsymbol{\theta}}_{\text{lik}} - \boldsymbol{\theta}) \longrightarrow \mathbb{N}(\mathbf{0}, \mathbf{V}_{\text{lik}}) \quad \text{in distribution,}$$

where $\mathbf{V}_{\text{lik}} = \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\} \boldsymbol{\Lambda}_{\text{lik}}^{-1}$ and $\boldsymbol{\Lambda}_{\text{lik}} = \mathbb{E}\left[\frac{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)}{1 + c\varphi^{-1}(\mathbf{x})e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}\right]$.

Theorem 3 shows that the proposed estimator $\hat{\boldsymbol{\theta}}_{\text{lik}}$ is asymptotic normal with variance \mathbf{V}_{lik} . We compare \mathbf{V}_{lik} with \mathbf{V}_w to show that $\hat{\boldsymbol{\theta}}_{\text{lik}}$ has a higher estimation efficiency than $\hat{\boldsymbol{\theta}}_w$.

Theorem 5. *If \mathbf{V}_w and \mathbf{V}_{lik} are finite and positive definite, then $\mathbf{V}_{\text{lik}} \leq \mathbf{V}_w$, i.e., $\mathbf{V}_w - \mathbf{V}_{\text{lik}}$ is non-negative definite. The equality holds when $c = 0$ and in this case $\mathbf{V}_{\text{lik}} = \mathbf{V}_w = \mathbf{V}_f$.*

Since the estimator $\hat{\boldsymbol{\theta}}_{\text{lik}}$ in Eg. (5) is based on the conditional log-likelihood of the sampled data, it actually has the highest estimation efficiency among a class of asymptotically unbiased estimators.

Theorem 6. *Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$ be the feature matrix and \mathcal{D}_{δ} denote the sampled data. Let $\hat{\boldsymbol{\theta}}_U$ be a subsample estimator with the following asymptotic representation:*

$$\hat{\boldsymbol{\theta}}_U = \mathbf{U}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta}) + N_1^{-1/2} o_P(1), \quad (6)$$

where the variance $\mathbb{V}\{\mathbf{U}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta})\}$ exists, and $\mathbf{U}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta})$ satisfies that $\mathbb{E}\{\mathbf{U}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta}) \mid \mathbf{X}\} = \boldsymbol{\theta}^*$ and $\mathbb{E}\{\dot{\mathbf{U}}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta}) \mid \mathbf{X}\} = \mathbf{0}$ with $\dot{\mathbf{U}}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta}) = \partial \mathbf{U}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta}) / \partial \boldsymbol{\theta}^{*T}$. If $N_1 \mathbb{V}\{\mathbf{U}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta})\} \rightarrow \mathbf{V}_U$ in probability, then $\mathbf{V}_U \geq \mathbf{V}_{\text{lik}}$.

Remark 2. Theorem 6 tells us that $\hat{\boldsymbol{\theta}}_{\text{lik}}$ is statistically the most efficient among a class of estimators, which includes both $\hat{\boldsymbol{\theta}}_w$ and $\hat{\boldsymbol{\theta}}_{\text{lik}}$ as special cases. The condition $\mathbb{E}\{\mathbf{U}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta}) \mid \mathbf{X}\} = \boldsymbol{\theta}^*$ ensures that the estimator is asymptotically unbiased. The condition $\mathbb{E}\{\dot{\mathbf{U}}(\boldsymbol{\theta}^*; \mathcal{D}_{\delta}) \mid \mathbf{X}\} = \mathbf{0}$ can be intuitively interpreted as the requirement that the derivative of the $o_P(1)$ term in (6) is also negligible. Clearly, this is satisfied by all unbiased estimators for which the $o_P(1)$ term in (6) is $\mathbf{0}$.

5 Practical considerations

In this section, we first show how we design a more practical estimator based on the previous results and then present the theoretical analysis behind these improvements.

5.1 Making estimators more practical

Like the LCC sampling (Fithian and Hastie, 2014) and the A-optimal sampling (Wang, 2019), the $\varphi_{\text{os}}(\mathbf{x})$ in (3) depends on θ , and thus a pilot value of θ , say $\tilde{\theta}$, is required in practice. Here, $\tilde{\theta}$ can be constructed from a pilot sample, say $(\tilde{\mathbf{x}}_i, \tilde{y}_i)$, $i = 1, \dots, \tilde{n}$, taken by uniform sampling from both classes with equal expected sample sizes. With $\tilde{\theta}$ obtained, calculate $\tilde{t}_i = p(\tilde{\mathbf{x}}_i; \tilde{\theta}) \|\tilde{\mathbf{M}}_f^{-1} \dot{g}(\tilde{\mathbf{x}}_i; \tilde{\theta})\|$ for $i = 1, \dots, \tilde{n}$, where $\tilde{\mathbf{M}}_f$ is the Hessian matrix of the pilot sample objective function. In case that the Hessian matrix and gradients are hard to record, the gradient term can be dropped, i.e., use $\tilde{t}_i = p(\tilde{\mathbf{x}}_i; \tilde{\theta})$. As mentioned in Remark 1, T can be ignored for very imbalanced data. The expectation in the denominator of (3) can be approximated by mimicking the method of moment estimator, namely by

$$\tilde{\omega} = 2N_1(\tilde{n}N)^{-1} \sum_{\tilde{y}_i=1} \tilde{t}_i + 2N_0(\tilde{n}N)^{-1} \sum_{\tilde{y}_i=0} \tilde{t}_i. \quad (7)$$

In practice, it is common to set a lower threshold, say ϱ , on sampling probabilities, to ensure that all instances have positive probabilities to be selected. This is more critical for the IPW estimator while the likelihood estimator is not very sensitive to very small sampling probabilities.

Denote the pilot value as $\tilde{\theta} = (\tilde{\theta}, \tilde{\omega})$. The following probabilities can be practically implemented

$$\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\theta}) = \min[\max\{\rho \tilde{\varphi}_{\text{os}}(\mathbf{x}), \varrho\}, 1], \quad \text{where} \quad \tilde{\varphi}_{\text{os}}(\mathbf{x}) = \tilde{\omega}^{-1} t(\mathbf{x}; \tilde{\theta}), \quad (8)$$

and we use this in our experiments in Section 6. The $\tilde{\omega}$ in (7) is essentially a normalizing term so that the expected subsample size is ρ . For some practical systems such as online models, $\tilde{\omega}$ is often treated as a tuning parameter. We will illustrate this in our real data example in Section 6.2.

5.2 Theoretical analysis of practical estimators

Denote $\hat{\theta}_w^{\tilde{\theta}}$ and $\hat{\theta}_{\text{lik}}^{\tilde{\theta}}$ the IPW estimator and the likelihood estimator, respectively, with practically estimated optimal probability in (8). We have the following results.

Theorem 7. Assume that the pilot estimator $\tilde{\theta}$ is independent of the data, $\tilde{\varphi}_{\text{os}}(\mathbf{x}) \rightarrow \varphi_{\text{plt}}(\mathbf{x})$ in probability, and $\rho^{-1}\varrho \rightarrow c_l > 0$. Under Assumptions 1-3, as $N \rightarrow \infty$, the $\hat{\theta}_w^{\tilde{\theta}}$ satisfies that,

$$\sqrt{N_1}(\hat{\theta}_w^{\tilde{\theta}} - \theta) \rightarrow \mathbb{N}(\mathbf{0}, \mathbf{V}_w^{\text{plt}}) \quad \text{in distribution,}$$

where $\mathbf{V}_w^{\text{plt}}$ has the same expression of \mathbf{V}_w except that $\varphi(\mathbf{x})$ is replaced by $\max\{\varphi_{\text{plt}}(\mathbf{x}), c_l\}$. If $\Pr\{\varphi_{\text{plt}}(\mathbf{x}) \geq c_l\} = 1$ and the pilot estimator is consistent, then $\mathbf{V}_w^{\text{plt}}$ achieves the optimal variance.

Theorem 8. Assume that the pilot estimator $\tilde{\theta}$ is independent of the data, $\tilde{\varphi}_{\text{os}}(\mathbf{x}) \rightarrow \varphi_{\text{plt}}(\mathbf{x})$ in probability, and $\varrho = o(\rho)$. Under Assumptions 1-3, as $N \rightarrow \infty$, $\hat{\theta}_{\text{lik}}^{\tilde{\theta}}$ satisfies that,

$$\sqrt{N_1}(\hat{\theta}_{\text{lik}}^{\tilde{\theta}} - \theta) \rightarrow \mathbb{N}(\mathbf{0}, \mathbf{V}_{\text{lik}}^{\text{plt}}) \quad \text{in distribution,}$$

where $\mathbf{V}_{\text{lik}}^{\text{plt}}$ has the same expression of \mathbf{V}_{lik} except that $\varphi(\mathbf{x})$ is replaced by $\varphi_{\text{plt}}(\mathbf{x})$. If the pilot estimator is consistent, then $\mathbf{V}_{\text{lik}}^{\text{plt}}$ achieves the variance with the optimal probability.

Remark 3. Theorems 7 and 8 do not require $\tilde{\theta}$ to be consistent, i.e., the pilot estimator can be misspecified. However $\tilde{\theta}$ has to be consistent in order to achieve the asymptotic variances with the optimal probability. Compared with the likelihood estimator $\hat{\theta}_{\text{lik}}^{\tilde{\theta}}$, the IPW estimator $\hat{\theta}_w^{\tilde{\theta}}$ requires a stronger condition on ϱ to have asymptotic normality, because $\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\theta})$ is in the denominator of the objective function for $\hat{\theta}_w^{\tilde{\theta}}$ so the turbulence of $\tilde{\theta}$ is amplified. The requirement $c_l > 0$ means that ϱ cannot be too small in practice. The likelihood estimator does not have this constraint.

6 Numerical experiments

We present simulated results and report the performance on a CTR dataset with more than 0.3 trillion instances. More experiments with model and pilot misspecifications are available in the supplement.

6.1 Experiments on simulated data

We implemented the following negative sampling methods for logistic regression. 1) uniW: uniform sampling with the IPW estimator. 2) uniLik: uniform sampling with the likelihood estimator. 3) optW: optimal sampling with the IPW estimator. 4) optLik: optimal sampling with the likelihood estimator. We also implement the full data MLE (Full) and the LCC for comparisons.

In each repetition of the simulation, we generate full data of size $N = 5 \times 10^5$ from the logistic regression with $g(\mathbf{x}; \boldsymbol{\theta}) = \alpha + \mathbf{x}^T \boldsymbol{\beta}$. We set the true $\boldsymbol{\beta}^*$ to be a 6×1 vector of ones and set different values for α for different feature distributions so that the positive/negative ratio is close to 1:400. We consider the following feature distributions: (a) Normal distribution which is symmetric with light tails; the true intercept is $\alpha = -7.65$. (b) Log-normal distribution which is asymmetric and positively skewed; the true intercept is $\alpha = -0.5$. (c) t_3 distribution which is symmetric with heavier tails; the true intercept is $\alpha = -7$.

We set the sampling rate as $\rho = 0.002, 0.004, 0.006, 0.01, \text{ and } 0.02$ for all sampling methods. In each repetition, uniform samples of average size 100 are selected from each class to calculate the pilot estimates, so the uncertainty due to the pilot estimates are taken into account. We also consider pilot misspecification by adding a uniform random number from $\mathcal{U}(0, 1.5)$ to the pilot so that the pilot is systematically different from the true parameter. We repeat the simulation for $R = 1000$ times to calculate the MSE as $R^{-1} \sum_{r=1}^R \|\hat{\boldsymbol{\theta}}^{(r)} - \boldsymbol{\theta}\|^2$, where $\boldsymbol{\theta}^{(r)}$ is the estimate at the r -th repetition.

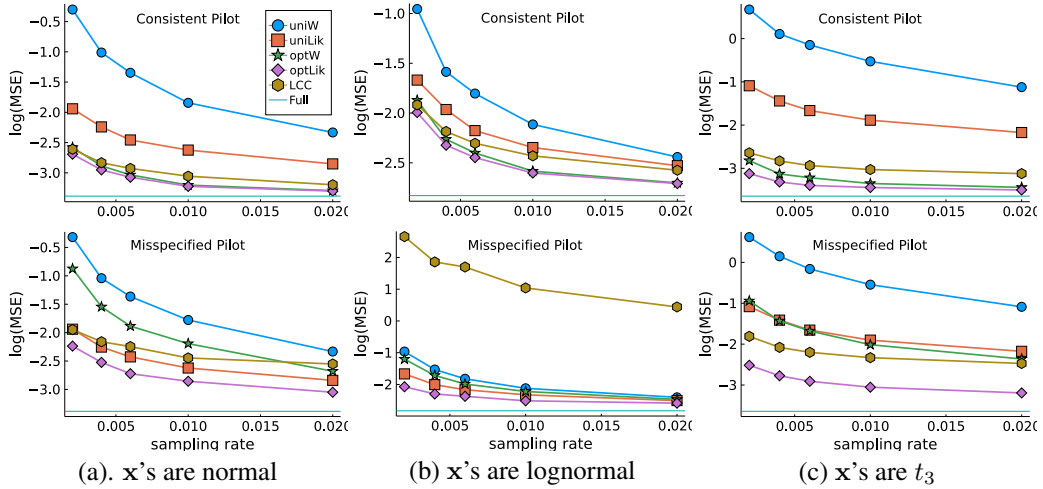


Figure 1: Log(MSEs) for different estimators (the smaller the better). The top row uses consistent pilot estimators; the bottom uses misspecified pilot estimators.

Results are presented in Figure 1. The MSEs of all estimators decrease as the sampling rate ρ increases. The optLik outperforms other sampling methods in general, and its performance is close to the full data MLE when $\rho = 0.02$, especially if the pilot is consistent. The advantage of optLik over optW is more evident for smaller ρ . If the pilot is misspecified, optLik is quite robust but optLik and LCC significantly deteriorate. Since uniLik and uniW do not require a pilot, they are not affected by pilot misspecification, but they have much larger MSE than optLik and optW when the pilot is consistent, especially uniW. Different feature distributions also affect the performances of different sampling methods.

To verify Theorem 1 numerically, we let the full data size N increase at a faster rate than the average number of positive instances $N_1^a = \mathbb{E}(N_1)$ so that the probability $\Pr(y = 1)$ decreases towards zero. As N increases, we fixed the value of $\boldsymbol{\beta}^*$ and set decreasing values to α^* to mimic our scaling regime. We considered two covariate distributions: 1) a multivariate normal distribution for which a logistic

regression model is a correct model; and 2) a multivariate log-normal distribution for which a logistic regression model is misspecified. We simulated for 100 times to calculate the empirical variance $\hat{\mathbf{V}}_f$ of the full data MLE, and report the results in Table 1. It is seen that $\text{tr}(\hat{\mathbf{V}}_f)$ is decreasing towards zero. According to Theorem 1, $\text{tr}(\hat{\mathbf{V}}_f)$ should converge in a rate of $1/N_1$, so $N_1^a \text{tr}(\hat{\mathbf{V}}_f)$ should be relatively stable as N increases. This is indeed the case as seen in the third and sixth columns of Table 1. On the other hand, $N \text{tr}(\hat{\mathbf{V}}_f)$ gets large dramatically as N increases. The aforementioned observations confirm the theoretical result in Theorem 1. Furthermore, Table 1 shows that the convergence rate in Theorem 1 may also be true for some misspecified models.

Table 1: Empirical variances of the full data MLE for different full data size and average number of positive instances combinations.

| (N, N_1^a) | Correct model | | | Mis-specified model | | |
|---------------|---------------------------------|---------------------------------------|-----------------------------------|---------------------------------|---------------------------------------|-----------------------------------|
| | $\text{tr}(\hat{\mathbf{V}}_f)$ | $N_1^a \text{tr}(\hat{\mathbf{V}}_f)$ | $N \text{tr}(\hat{\mathbf{V}}_f)$ | $\text{tr}(\hat{\mathbf{V}}_f)$ | $N_1^a \text{tr}(\hat{\mathbf{V}}_f)$ | $N \text{tr}(\hat{\mathbf{V}}_f)$ |
| $(10^3, 32)$ | 0.169 | 5.41 | 169.17 | 0.969 | 30.99 | 968.70 |
| $(10^4, 64)$ | 0.097 | 6.20 | 969.29 | 0.322 | 20.59 | 3217.12 |
| $(10^5, 128)$ | 0.045 | 5.76 | 4497.24 | 0.135 | 17.32 | 13527.60 |
| $(10^6, 256)$ | 0.018 | 4.62 | 18048.40 | 0.046 | 11.74 | 45847.40 |

6.2 Experiments on real data

We conduct experiments on an internal CTR dataset from our products with more than 0.3 trillion instances. There are 299 input features, including user and ad profiles together with rich context features. We concatenate all features as a single high-dimensional vector \mathbf{x} , whose length is greater than 5,000. We use a 3-layer fully connected deep neural network as a feature extractor that maps \mathbf{x} into a 256-dimensional dense features $h(\mathbf{x})$. The downstream model is a linear logistic regression model taking $h(\mathbf{x})$ as input to predict the binary output whether the user clicks the ad or not.

The positive/negative ratio is around 1:80. Due to limited storage, we first uniformly drop 80% negative instances and try various negative sampling algorithms on the rest 60 billion instances. On this pre-processed dataset, we then split out 1 percent of the data for testing, and do negative sampling on the rest data. The entire training procedure scans the subsampled data set from begin to end in one-pass according to the timestamp. The testing is done all the way through training. We calculate the area under the ROC curve (AUC) for testing instances. The entire model is trained on a distributed machine learning platform using 1,200 CPU cores and can finish within two days.

We adopt the sampling probability $\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})$ as demonstrated in (8) and use the empirical sampling rate to calibrate the normalizing term $\tilde{\omega}$. As mentioned before, we remove the gradient term for implementation tractability so $t(\mathbf{x}; \tilde{\boldsymbol{\theta}}) = p(\mathbf{x}; \tilde{\boldsymbol{\theta}})$. As the result will show, this approximate score produces consistent results with the simulation experiments. We compare four methods: 1) uniW and 2) uniLik are the same as in simulated experiments; and 3) optW and 4) optLik represent nonuniform sampling probability using $\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})$ with the IPW estimator and with the likelihood estimator. We use “opt” to refer to both optW and optLik when focusing on the sampling.

We study two sets of negative sample rates (w.r.t. the original 0.3 trillion data): (i) [0.01, 0.05]. In this case, negative instances still outnumber positive instances. (ii) [0.001, 0.008]. In this case, negative subsampling is too aggressive that positive instances dominates negative instances. We demonstrate the result in Figure 2. When the sample rate is moderate (Figure 2.(b)), opt is nearly optimal and can significantly outperform uniform subsampling. There is still a small gap between the IPW estimator and the likelihood estimator. When the sample rate is extremely small (Figure 2.(a)), opt becomes sub-optimal and its gap w.r.t. uniform sampling is closer. Still we see a clear difference between the IPW estimator and the likelihood estimator. This is due to a larger asymptotic variance for uniform sampling when the sampling rate goes to zero. Note that on this huge data set, a small relative AUC gain (e.g., 0.0005) usually could bring about significant revenue gain in products.

6.3 Sensitivity analysis

In practice, we find it necessary to use cross-validation to find the optimal ϱ in (8) given a fixed sample rate. As demonstrated in Figure 3, for each sampling rate we tune ϱ from a very small value (in this

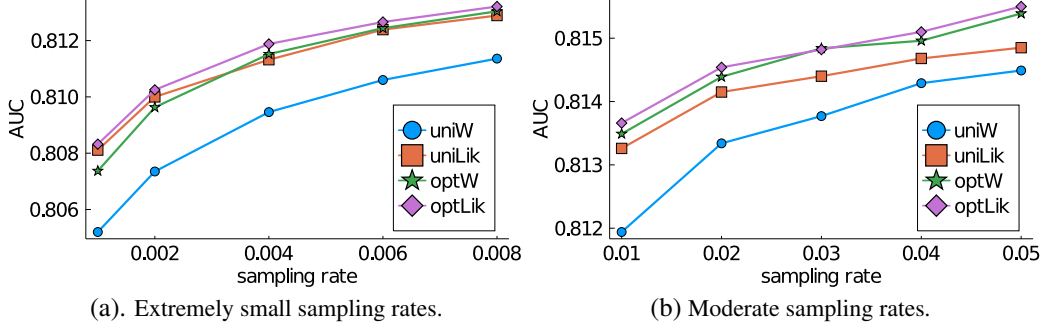


Figure 2: Empirical testing AUC of subsample estimators for different sample sizes (the larger the better).

case, 10^{-6}) all the way to its largest value. When ϱ achieves its largest value, $\rho\tilde{\omega}^{-1} = 0$ and $\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})$ reduces to uniform negative sampling. We observe that $\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})$ with a moderate ϱ achieves the best results for various sampling rates. Note that ϱ is the lower truncation level for negative sampling probabilities. Thus, a larger ϱ means more “simple” negative instances ($\rho\tilde{\omega}^{-1}p(\mathbf{x}_i; \tilde{\boldsymbol{\theta}})$ is smaller than ϱ) will have better chances to be selected. This coincides with recent empirical results for negative subsampling in large search systems (Huang *et al.*, 2020), where they mix “simple” (smaller $p(\mathbf{x}_i; \tilde{\boldsymbol{\theta}})$) and “hard” (larger $p(\mathbf{x}_i; \tilde{\boldsymbol{\theta}})$) negative instances. We also observe an empirical variance of around 0.0001 for each setup, demonstrating that the relative improvement is consistent.

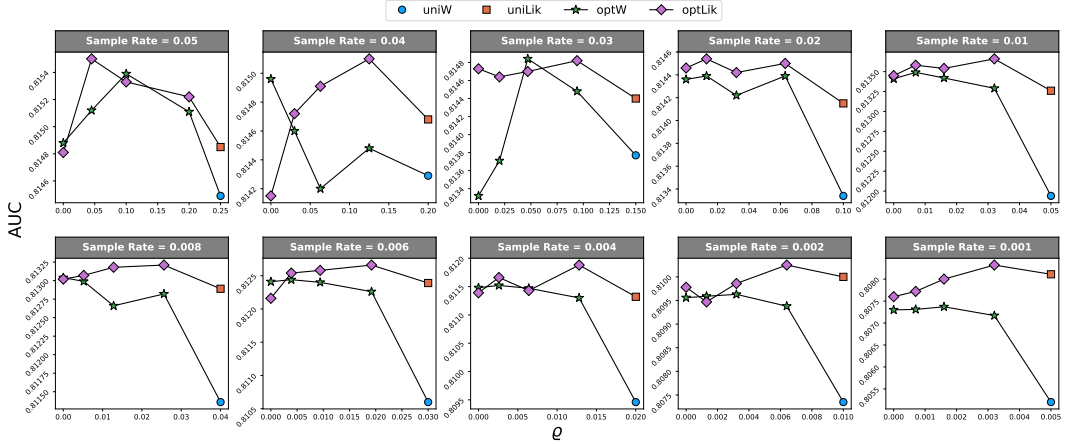


Figure 3: AUC for IPW and likelihood estimators when tuning the truncation lower bound ϱ .

7 Conclusions, discussion, and limitations

In this paper, we have derived asymptotic distributions for full data and subsample estimators with rare events data. We have also found the optimal sampling probability that minimizes the unconditional asymptotic variance of the IPW estimator, and proposed the likelihood estimator through nonuniform log odds correction to further improve the estimation efficiency. The optimal probability depends on unknown parameters so we have discussed practical implementations and investigated the convergences of the resultant estimators. Experiments on both simulated and real big online streaming data confirm our theoretical findings. Heuristically, our nonuniform negative sampling method gives preference to data points that are harder to observe, so it would give preference to sub-populations that are difficult to observe. We do not think this would bring any negative societal impacts. We have also examined our real data application and did not notice discrimination against any sub-populations.

This paper has the following limitations: 1) We assume that the model is correctly specified. Theoretical properties with model misspecification is not considered, and this important question requires

future studies. 2) Oversampling the rare positive instances is another common practice to deal with imbalanced data. Its application together with negative sampling is not considered. 3) The assumption of a fixed β^* means that both the marginal and conditional probabilities for a positive instance to occur are small. If the proportion of the positive instances for a given feature in a dataset is large, this assumption may not be appropriate. Further investigation is need to see if our results is still applicable. In the following, we present three scenarios that the scaling regime may or may not fit in:

Scenario 1 (phone call while driving) Car crashes occur with small probabilities, and making phone calls while driving significantly increases the probability of a car crash. However, the probability of car crashes among people making phone calls while driving is still small, so for these types of features, our scaling regime is appropriate to model the rare events.

Scenario 2 (anti-causal learning) Anti-causal learning Schölkopf *et al.* (2012); Kilbertus *et al.* (2018) assumes that label (y) causes observations (x). Thus $\Pr(x|y)$ represents the causal mechanism and is fixed. One standard example is that diseases (y) cause symptoms (x). Our scaling regime fits the framework of anti-causal learning. To see this, using Bayes Theorem we write the log odds as

$$\alpha + f(\mathbf{x}; \beta) = \log \frac{\Pr(y = 1)}{\Pr(y = 0)} + \log \frac{\Pr(\mathbf{x} | y = 1)}{\Pr(\mathbf{x} | y = 0)}.$$

In anti-causal learning, the marginal distribution of y changes while the conditional distribution of \mathbf{x} given y is fixed. Thus only the scale factor α changes, and $f(\mathbf{x}; \beta)$ is fixed.

Scenario 3 (local transmission of the COVID-19) Our scaling regime has some limitations and does not apply to all types of rare events data. For example, although the COVID-19 infection rate for the whole population is low, the infection rate for people whose family members are in the same house with positive test results is high. This means the change of a family member’s test result converts a small-probability-event to a large-probability-event, and our scaling regime would not be appropriate.

Acknowledgments and Disclosure of Funding

HaiYing Wang’s research was partially supported by NSF grant CCF 2105571.

Proofs and Additional Numerical Experiments

Before presenting the proof, we point out that Assumption 2 implies that $\mathbb{E}\{v(\mathbf{x})\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta})\}$ is positive definite for a positive function $v(\mathbf{x}) > 0$ almost surely. This is because Assumption 2 means that $\Pr\{l^T \dot{g}(\mathbf{x}; \boldsymbol{\theta}) \neq 0\} > 0$ for any $l \neq \mathbf{0}$, and therefore $\Pr\{v^{1/2}(\mathbf{x})l^T \dot{g}(\mathbf{x}; \boldsymbol{\theta}) \neq 0\} > 0$ for any $l \neq \mathbf{0}$, implying that $\mathbb{E}\{v(\mathbf{x})\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta})\}$ is positive definite. We also point out that if a sequence is $o_P(1)$ conditionally then it is also $o_P(1)$ unconditionally and vice versa (Xiong and Li, 2008; Cheng *et al.*, 2010).

To facilitate the presentation, define $a_N = \sqrt{Ne^{\alpha^*}}$. We notice that

$$N_1 = N\mathbb{E}\left\{\frac{e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}}{1 + e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}}\right\}\{1 + o_P(1)\} = a_N^2 \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\}\{1 + o_P(1)\}, \quad (\text{S.1})$$

from the dominated convergence theorem. Thus the normalizing term $\sqrt{N_1}$ for the asymptotic normality in the paper can be replaced by $a_N \mathbb{E}^{1/2}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\}$.

S.1 Proof of Theorem 1

The estimator $\hat{\boldsymbol{\theta}}_f$ is the maximizer of

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N [y_i g(\mathbf{x}_i; \boldsymbol{\theta}) - \log\{1 + e^{g(\mathbf{x}_i; \boldsymbol{\theta})}\}],$$

so $a_N(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is the maximizer of

$$\gamma(\mathbf{u}) = \ell(\boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) - \ell(\boldsymbol{\theta}^*).$$

By Taylor's expansion,

$$\gamma(\mathbf{u}) = a_N^{-1}\mathbf{u}^T \dot{\ell}(\boldsymbol{\theta}^*) + 0.5a_N^{-2} \sum_{i=1}^N \phi(\mathbf{x}_i; \boldsymbol{\theta}^*) \{\mathbf{u}^T \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\}^2 + \Delta_\ell + R,$$

where

$$\dot{\ell}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^N \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta})\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}),$$

$$\Delta_\ell = \frac{1}{2a_N^2} \sum_{i=1}^N \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \mathbf{u}^T \ddot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mathbf{u},$$

and R is the remainder term. By direct calculation,

$$\begin{aligned} R &= \frac{1}{6a_N^3} \sum_{i=1}^N \phi(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) \{1 - 2p(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u})\} \{\mathbf{u}^T \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u})\}^3 \\ &\quad + \frac{1}{6a_N^3} \sum_{i=1}^N \phi(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) \{\mathbf{u}^T \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u})\} \{\mathbf{u}^T \ddot{g}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) \mathbf{u}\} \\ &\quad + \frac{1}{6a_N^3} \sum_{i=1}^N \left[\{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u})\} \sum_{j_1, j_2=1}^d u_{j_1} u_{j_2} \ddot{g}_{j_1 j_2}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) \mathbf{u} \right], \end{aligned}$$

where $\ddot{g}_{j_1 j_2}(\mathbf{x}; \boldsymbol{\theta})$ is the gradient of $\ddot{g}_{j_1 j_2}(\mathbf{x}, \boldsymbol{\theta})$ and \mathbf{u} lies between $\mathbf{0}$ and \mathbf{u} . We see that

$$\begin{aligned} |R| &\leq \frac{\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N p(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) C(\mathbf{x}_i, \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N y_i B(\mathbf{x}_i) \\ &\leq \frac{\|\mathbf{u}\|^3 e^{a_N^{-1}\|\mathbf{u}\|}}{6Na_N} \sum_{i=1}^N \exp\{f(\mathbf{x}_i; \boldsymbol{\beta} + a_N^{-1}\mathbf{u}_{(-1)})\} C(\mathbf{x}_i, \boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N y_i B(\mathbf{x}_i) \end{aligned}$$

$$\leq \frac{d^3 \|\mathbf{u}\|^3 e^{a_N^{-1} \|\mathbf{u}\|}}{6N a_N} \sum_{i=1}^N B(\mathbf{x}_i) + \frac{d \|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N y_i B(\mathbf{x}_i) \equiv \Delta_{R1} + \Delta_{R2}, \quad (\text{S.2})$$

where $C(\mathbf{x}, \boldsymbol{\theta}) = \|\dot{g}(\mathbf{x}, \boldsymbol{\theta})\|^3 + \|\dot{g}(\mathbf{x}, \boldsymbol{\theta})\| \|\ddot{g}(\mathbf{x}, \boldsymbol{\theta})\| + \sum_{j_1 j_2}^d \|\ddot{g}_{j_1 j_2}(\mathbf{x}, \boldsymbol{\theta})\|$ and $\dot{\mathbf{u}}_{(-1)}$ is $\dot{\mathbf{u}}$ with the first element removed. From Assumption 1, $\mathbb{E}(\Delta_{R1}) \rightarrow 0$ and

$$\mathbb{E}(\Delta_{R2}) \leq \frac{d \|\mathbf{u}\|^3}{6N a_N} \sum_{i=1}^N \mathbb{E}\{e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)} B(\mathbf{x}_i)\} = \frac{d \|\mathbf{u}\|^3}{6a_N} \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} B(\mathbf{x})\} \rightarrow 0.$$

Since Δ_{R1} and Δ_{R2} are both positive, Markov's inequality shows that they are both $o_P(1)$ so $R = o_P(1)$. For Δ_ℓ , the mean $\mathbb{E}(\Delta_\ell) = \mathbf{0}$ and the variance satisfies that

$$\begin{aligned} \mathbb{V}(\Delta_\ell) &\leq \frac{\|\mathbf{u}\|^4}{4a_N^4} \sum_{i=1}^N \mathbb{E}\{p(\mathbf{x}_i; \boldsymbol{\theta}^*) \|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^2\} \leq \frac{\|\mathbf{u}\|^4}{4N a_N^2} \sum_{i=1}^N \mathbb{E}[e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^2] \\ &= \frac{\|\mathbf{u}\|^4}{4a_N^2} \mathbb{E}[e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}, \boldsymbol{\theta}^*)\|^2] \rightarrow 0, \end{aligned}$$

so $\Delta_\ell = o_P(1)$.

If we can show that

$$a_N^{-1} \dot{\ell}(\boldsymbol{\theta}^*) \longrightarrow \mathbb{N}(\mathbf{0}, \mathbf{M}_f), \quad (\text{S.3})$$

in distribution, and

$$a_N^{-2} \sum_{i=1}^N \phi(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \longrightarrow \mathbf{M}_f, \quad (\text{S.4})$$

in probability, then from the Basic Corollary in page 2 of Hjort and Pollard (2011), we know that $a_N(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, the maximizer of $\gamma(\mathbf{u})$, satisfies that

$$a_N(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \mathbf{M}_f^{-1} \times a_N^{-1} \dot{\ell}(\boldsymbol{\theta}^*) + o_P(1). \quad (\text{S.5})$$

Slutsky's theorem together with (S.3) and (S.5) implies the result in Theorem 1. We prove (S.3) and (S.4) in the following.

Note that

$$\dot{\ell}(\boldsymbol{\theta}^*) = \sum_{i=1}^N \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)$$

is a summation of i.i.d. quantities. Since the distribution of $\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)$ depends on N because $\alpha^* \rightarrow -\infty$ as $N \rightarrow \infty$, we need to use the Lindeberg-Feller central limit theorem for triangular arrays (see, Section *2.8 of van der Vaart, 1998).

We examine the mean and variance of $a_N^{-1} \dot{\ell}(\boldsymbol{\theta}^*)$. For the mean, from the fact that

$$\mathbb{E}[\{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)] = \mathbb{E}\left[\mathbb{E}\{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*) \mid \mathbf{x}_i\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mathbf{x}_i\right] = \mathbf{0},$$

we know that $\mathbb{E}\{a_N^{-1} \dot{\ell}(\boldsymbol{\theta}^*)\} = \mathbf{0}$.

For the variance,

$$\begin{aligned} \mathbb{V}\{a_N^{-1} \dot{\ell}(\boldsymbol{\theta}^*)\} &= a_N^{-2} \sum_{i=1}^N \mathbb{V}[\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)] = a_N^{-2} N \mathbb{E}\{\phi(\boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\} \\ &= a_N^{-2} N \mathbb{E}\left[\frac{e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}}{\{1 + e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\}^2} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right] = \mathbb{E}\left[\frac{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}{\{1 + e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\}^2} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right]. \end{aligned}$$

We have

$$\frac{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}{\{1 + e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\}^2} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*) \longrightarrow e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*) \quad \text{almost surely,}$$

and

$$\frac{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}{\{1 + e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\}^2} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 \leq e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 \quad \text{with} \quad \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2\} \leq \infty.$$

Note that $\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*) = \{1, \dot{f}^T(\mathbf{x}; \boldsymbol{\beta}^*)\}^T$ does not depend on N . Thus, from the dominated convergence theorem,

$$\mathbb{V}\{a_N^{-1} \dot{\ell}(\boldsymbol{\theta}^*)\} \longrightarrow \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\}.$$

Now we check the Lindeberg-Feller condition. For any $\epsilon > 0$,

$$\begin{aligned} & \sum_{i=1}^N \mathbb{E} \left[\|\{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^2 I(\|\{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\| > a_N \epsilon) \right] \\ &= N \mathbb{E} \left[\|\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon) \right] \\ &= N \mathbb{E} \left[p(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon) \right] \\ &\quad + N \mathbb{E} \left[\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \{p(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|p(\mathbf{x}; \boldsymbol{\theta}^*) \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon) \right] \\ &\leq N \mathbb{E} \left[p(\mathbf{x}; \boldsymbol{\theta}^*) \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon) \right] \\ &\quad + N \mathbb{E} \left[\{p(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|p(\mathbf{x}; \boldsymbol{\theta}^*) \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon) \right] \\ &\leq a_N^2 \mathbb{E} \{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon)\} \\ &\quad + a_N^2 \mathbb{E} \{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|p(\mathbf{x}; \boldsymbol{\theta}^*) \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon)\} \\ &= o(a_N^2), \end{aligned}$$

where the last step is from the dominated convergence theorem. Thus, applying the Lindeberg-Feller central limit theorem (Section 2.8 of van der Vaart, 1998), we finish the proof of (S.3).

Now we prove (S.4). This is done by noting that

$$\begin{aligned} & a_N^{-2} \sum_{i=1}^N \phi(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \\ &= \frac{1}{N e^{\alpha^*}} \sum_{i=1}^N \frac{e^{\alpha^* + f(\mathbf{x}_i; \boldsymbol{\beta}^*)}}{(1 + e^{\alpha^* + f(\mathbf{x}_i; \boldsymbol{\beta}^*)})^2} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \\ &= \frac{1}{N} \sum_{i=1}^N \frac{e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)}}{(1 + e^{\alpha^* + f(\mathbf{x}_i; \boldsymbol{\beta}^*)})^2} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) = \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\} + o_P(1), \end{aligned}$$

where the last step is from Lemma 28 of Wang (2019).

S.2 Proof of Theorem 2

First, note that \mathbf{V}_w can be written as

$$\mathbf{V}_w = \mathbb{E}(e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}) \mathbf{M}_f^{-1} \mathbf{M}_w \mathbf{M}_f^{-1},$$

where

$$\mathbf{M}_w = \mathbb{E} \left[\left\{ 1 + \frac{c e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}{\varphi(\mathbf{x})} \right\} e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*) \right].$$

We then point out that under assumptions 1 and 2 the condition $\mathbb{E}[\{\varphi(\mathbf{x}) + \varphi^{-1}(\mathbf{x})\} B^2(\mathbf{x})] < \infty$ implies the following:

$$\mathbb{E}[\{\varphi(\mathbf{x}) e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2] < \infty, \quad (\text{S.6})$$

$$\mathbb{E}[\{1 + \varphi^{-1}(\mathbf{x})\} e^{2f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^4] < \infty, \quad (\text{S.7})$$

$$\mathbb{E}\{\varphi^{-1}(\mathbf{x})e^{2f(\mathbf{x};\boldsymbol{\beta}^*)}\|\ddot{g}(\mathbf{x};\boldsymbol{\theta}^*)\|^2\} < \infty. \quad (\text{S.8})$$

The estimator $\hat{\boldsymbol{\theta}}_w$ is the maximizer of

$$\ell_w(\boldsymbol{\theta}) = \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} [y_i g(\mathbf{x}_i; \boldsymbol{\theta}) - \log\{1 + e^{g(\mathbf{x}_i; \boldsymbol{\theta})}\}],$$

so $a_N(\hat{\boldsymbol{\theta}}_w - \boldsymbol{\theta}^*)$ is the maximizer of $\gamma_w(\mathbf{u}) = \ell_w(\boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) - \ell_w(\boldsymbol{\theta}^*)$. By Taylor's expansion,

$$\gamma_w(\mathbf{u}) = \frac{1}{a_N} \mathbf{u}^T \dot{\ell}_w(\boldsymbol{\theta}^*) + \frac{1}{2a_N^2} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \phi(\mathbf{x}_i; \boldsymbol{\theta}^*) (\mathbf{z}_i^T \mathbf{u})^2 + \Delta_{\ell_w} + R_w,$$

where

$$\begin{aligned} \dot{\ell}_w(\boldsymbol{\theta}) &= \frac{\partial \ell_w(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta})\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}), \\ \Delta_{\ell_w} &= \frac{1}{2a_N^2} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \mathbf{u}^T \ddot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mathbf{u}, \end{aligned}$$

and R_w is the remainder. Similarly to the proof of Theorem 1, we only need to show that

$$a_N^{-1} \dot{\ell}_w(\boldsymbol{\theta}^*) \longrightarrow \mathbb{N}(\mathbf{0}, \mathbf{M}_w), \quad (\text{S.9})$$

in distribution,

$$a_N^{-2} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \phi(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \longrightarrow \mathbf{M}_f, \quad (\text{S.10})$$

in probability for any \mathbf{u} , and $\Delta_{\ell_w} = o_P(1)$ and $R_w = o_P(1)$.

We prove (S.9) first. Let $\eta_i = \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)$, we know that $\eta_i, i = 1, \dots, N$, are i.i.d., with the underlying distribution of η_i being dependent on N . From direct calculation, we have that $\mathbb{E}(\eta_i \mid \mathbf{x}_i) = \mathbf{0}$ and

$$\begin{aligned} \mathbb{V}(\eta_i \mid \mathbf{x}_i) &= \mathbb{E} \left[\frac{\{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\}^2}{\{y_i + (1 - y_i)\rho\varphi(\mathbf{x}_i)\}} \mid \mathbf{x}_i \right] \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \\ &= p(\mathbf{x}_i; \boldsymbol{\theta}^*) \{1 - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\}^2 \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) + \{1 - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \frac{p^2(\mathbf{x}_i; \boldsymbol{\theta}^*)}{\rho\varphi(\mathbf{x}_i)} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \\ &\leq e^{\alpha^*} e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) + \rho^{-1} e^{2\alpha^*} \frac{e^{2f(\mathbf{x}_i; \boldsymbol{\beta}^*)}}{\varphi(\mathbf{x}_i)} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*). \end{aligned}$$

Thus, by the dominated convergence theorem, we obtain that

$$\mathbb{V}(\eta_i) = \mathbb{E}\{\mathbb{V}(\eta_i \mid \mathbf{x}_i)\} = e^{\alpha^*} \mathbb{E} \left[\left\{ 1 + \frac{ce^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)}}{\varphi(\mathbf{x}_i)} \right\} e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \right] \{1 + o(1)\}.$$

Now we check the Lindeberg-Feller condition (Section *2.8 of van der Vaart, 1998). Denote $\delta = y + (1 - y)I\{u \leq \rho\varphi(\mathbf{x})\}$ where $u \sim \mathbb{U}(0, 1)$. For any $\epsilon > 0$,

$$\begin{aligned} &\sum_{i=1}^N \mathbb{E}\{\|\eta_i\|^2 I(\|\eta_i\| > a_N \epsilon)\} \\ &= N \mathbb{E}[\|\pi^{-1}(\mathbf{x}, y) \delta \{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\pi^{-1}(\mathbf{x}, y) \delta \{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon)] \\ &= \rho N \mathbb{E} \left[\varphi(\mathbf{x}) \|\pi^{-1}(\mathbf{x}, y) \{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\pi^{-1}(\mathbf{x}, y) \{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N \epsilon) \right] \\ &\quad + N \mathbb{E} \left[\{1 - \rho\varphi(\mathbf{x})\} \|\pi^{-1}(\mathbf{x}, y) y \{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 \right] \end{aligned}$$

$$\begin{aligned}
& \times I(\|\pi^{-1}(\mathbf{x}, y)y\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\Big] \\
& = \rho N \mathbb{E}\left[\varphi(\mathbf{x})p(\mathbf{x}; \boldsymbol{\theta}^*)\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\
& + N \mathbb{E}\left[\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\rho^{-1}\varphi^{-1}(\mathbf{x})\|p(\mathbf{x}; \boldsymbol{\theta}^*)\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2\right. \\
& \quad \left.\times I(\|\rho^{-1}\varphi^{-1}(\mathbf{x})\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\
& + N \mathbb{E}\left[\{1 - \rho\varphi(\mathbf{x})\}p(\mathbf{x}; \boldsymbol{\theta}^*)\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\
& \leq \rho N e^{\alpha^*} \mathbb{E}\left[\varphi(\mathbf{x})e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\
& + N e^{2\alpha^*} \rho^{-1} \mathbb{E}\left[\varphi^{-1}(\mathbf{x})\|e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\rho^{-1}\varphi^{-1}(\mathbf{x})\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\
& + N e^{\alpha^*} \mathbb{E}\left[e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\
& = o(Ne^{\alpha^*}),
\end{aligned}$$

where the last step is due to Assumptions 1, (S.6)-(S.8), the dominated convergence theorem, and the facts that $a_N \rightarrow \infty$ and $\lim_{N \rightarrow \infty} e^{\alpha^*}/\rho = c < \infty$. Thus, applying the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998) finishes the proof of (S.9).

Now we prove (S.10). Let

$$\begin{aligned}
H_w & \equiv a_N^{-2} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \phi(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \\
& = \frac{1}{n} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \frac{e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)}}{(1 + e^{\alpha^* + f(\mathbf{x}_i; \boldsymbol{\beta}^*)})^2} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \\
& = \frac{1}{n} \sum_{i=1}^N \frac{y_i + (1 - y_i)I\{u_i \leq \rho\varphi(\mathbf{x}_i)\}}{y_i + (1 - y_i)\rho\varphi(\mathbf{x}_i)} \frac{e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)}}{(1 + e^{\alpha^* + f(\mathbf{x}_i; \boldsymbol{\beta}^*)})^2} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*).
\end{aligned}$$

We notice that

$$\mathbb{E}(H_w) = \mathbb{E}\left\{\frac{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}{(1 + e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)})^2} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right\} = \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\} + o(1), \quad (\text{S.11})$$

where the last step is by the dominated convergence theorem. In addition, the variance of each component of H_w is bounded by

$$\begin{aligned}
& \frac{1}{N} \mathbb{E}\left\{\frac{\delta}{\pi^2(\mathbf{x}, y)} e^{2f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^4\right\} \\
& \leq \frac{1}{N} \mathbb{E}\left\{\frac{1}{\pi(\mathbf{x}, y)} e^{2f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^4\right\} \\
& \leq \frac{1}{N\rho} \mathbb{E}\left[\left\{\rho + \frac{1}{\varphi(\mathbf{x})}\right\} e^{2f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^4\right] = o(1). \quad (\text{S.12})
\end{aligned}$$

where the last step is because of (S.7) and the fact that $Ne^{\alpha^*} \rightarrow \infty$ and $e^{\alpha^*}/\rho \rightarrow c < \infty$ imply that $N\rho \rightarrow \infty$. From (S.11) and (S.12), Chebyshev's inequality implies that $H_w \rightarrow \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\}$ in probability.

In the following we finish the proof by showing that $\Delta_{\ell_w} = o_P(1)$ and $R_w = o_P(1)$. For Δ_{ℓ_w} , the mean $\mathbb{E}(\Delta_{\ell_w}) = \mathbf{0}$ and the variance satisfies that

$$\begin{aligned}
\mathbb{V}(\Delta_{\ell_w}) & \leq \frac{\|\mathbf{u}\|^4 N}{4a_N^4} \mathbb{E}\left[\frac{\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2}{y + (1 - y)\rho\varphi(\mathbf{x})}\right] \\
& \leq \frac{\|\mathbf{u}\|^4 N}{4a_N^4} \mathbb{E}\left[\left\{1 + \frac{p(\mathbf{x}; \boldsymbol{\theta}^*)}{\rho\varphi(\mathbf{x})}\right\} p(\mathbf{x}; \boldsymbol{\theta}^*) \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2\right]
\end{aligned}$$

$$\leq \frac{\|\mathbf{u}\|^4}{4a_N^2} \mathbb{E} \left[\left\{ 1 + \frac{e^{\alpha^*} e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}{\rho \varphi(\mathbf{x})} \right\} e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{\mathbf{g}}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 \right] = o_P(1),$$

so $\Delta_{\ell_w} = o_P(1)$. For the remainder term R_w . By direct calculation,

$$\begin{aligned} R_w &= \frac{1}{6a_N^3} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \phi(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}}) \{1 - 2p(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}})\} \{\mathbf{u}^T \dot{\mathbf{g}}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}})\}^3 \\ &\quad + \frac{1}{6a_N^3} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \phi(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}}) \{\mathbf{u}^T \dot{\mathbf{g}}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}})\} \{\mathbf{u}^T \ddot{\mathbf{g}}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}}) \mathbf{u}\} \\ &\quad + \frac{1}{6a_N^3} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \left[\{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}})\} \sum_{j_1 j_2}^d u_{j_1} u_{j_2} \ddot{g}_{j_1 j_2}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}}) \mathbf{u} \right], \end{aligned}$$

where $\dot{\mathbf{u}}$ lies between $\mathbf{0}$ and \mathbf{u} . We see that

$$\begin{aligned} |R_w| &\leq \frac{\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} p(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}}) C(\mathbf{x}_i, \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}}) + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} y_i B(\mathbf{x}_i) \\ &\leq \frac{\|\mathbf{u}\|^3 e^{a_N^{-1} \|\mathbf{u}\|}}{6Na_N} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} \exp\{f(\mathbf{x}_i; \boldsymbol{\beta} + a_N^{-1} \dot{\mathbf{u}}_{(-1)})\} C(\mathbf{x}_i, \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}}) \\ &\quad + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} y_i B(\mathbf{x}_i) \\ &\leq \frac{d^3 \|\mathbf{u}\|^3 e^{a_N^{-1} \|\mathbf{u}\|}}{6Na_N} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} B(\mathbf{x}_i) + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} y_i B(\mathbf{x}_i) \\ &\equiv \Delta_{R_w 1} + \Delta_{R_w 2}, \end{aligned}$$

where $C(\mathbf{x}, \boldsymbol{\theta})$ is defined in (S.2) and $\dot{\mathbf{u}}_{(-1)}$ is $\dot{\mathbf{u}}$ with the first element removed. From Assumption 1, $\mathbb{E}(\Delta_{R_w 1}) \rightarrow 0$ and

$$\mathbb{E}(\Delta_{R_w 2}) \leq \frac{d\|\mathbf{u}\|^3}{6a_N} \mathbb{E}[e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} B(\mathbf{x})] \rightarrow 0.$$

Since $\Delta_{R_w 1}$ and $\Delta_{R_w 2}$ are both positive, Markov's inequality shows that they are both $o_P(1)$ so $R_w = o_P(1)$.

S.3 Proof of Theorem 3

Since only \mathbf{V}_{sub} in \mathbf{V}_w is affected by subsampling, the probability that minimizes $\text{tr}(\mathbf{V}_{\text{sub}})$ also minimizes $\text{tr}(\mathbf{V}_w)$. Let $\Lambda_{\text{sub}} = \mathbb{E}\{\varphi^{-1}(\mathbf{x}) e^{2f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{\mathbf{g}}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\}$. We notice that

$$\begin{aligned} \mathbf{V}_{\text{sub}} &= c \mathbb{E}(e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}) \mathbf{M}_f^{-1} \Lambda_{\text{sub}} \mathbf{M}_f^{-1} \\ &= c \mathbb{E}(e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}) e^{-2\alpha^*} \{1 + o_P(1)\} \mathbf{M}_f^{-1} \mathbb{E}\{\varphi^{-1}(\mathbf{x}) p^2(\mathbf{x}; \boldsymbol{\theta}^*) \dot{\mathbf{g}}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\} \mathbf{M}_f^{-1}, \end{aligned}$$

where the term $c \mathbb{E}(e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}) e^{-2\alpha^*}$ does not depend on $\varphi(\mathbf{x})$. Thus we can focus on minimizing the term $\mathbf{M}_f^{-1} \mathbb{E}\{\varphi^{-1}(\mathbf{x}) p^2(\mathbf{x}; \boldsymbol{\theta}^*) \dot{\mathbf{g}}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\} \mathbf{M}_f^{-1}$.

Note that

$$\text{tr}[p^2(\mathbf{x}; \boldsymbol{\theta}^*) \mathbf{M}_f^{-1} \dot{\mathbf{g}}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*) \mathbf{M}_f^{-1}] = \|p(\mathbf{x}; \boldsymbol{\theta}) \mathbf{M}_f^{-1} \dot{\mathbf{g}}(\mathbf{x}; \boldsymbol{\theta})\|^2 = t^2(\mathbf{x}; \boldsymbol{\theta}^*).$$

Thus the problem is to find $\varphi(\mathbf{x})$ that minimizes

$$\mathbb{E} \left\{ \frac{t^2(\mathbf{x}; \boldsymbol{\theta}^*)}{\varphi(\mathbf{x})} \right\}, \quad \text{subject to } 0 < \varphi(\mathbf{x}) \leq \rho^{-1} \text{ and } \mathbb{E}\{\varphi(\mathbf{x})\} = 1.$$

To facilitate the presentation, denote

$$\zeta = \frac{1}{\mathbb{E}[\min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}]},$$

which is non-random. We notice that

$$\begin{aligned}
& \mathbb{E} \left\{ \frac{\zeta^2 t^2(\mathbf{x}; \boldsymbol{\theta}^*)}{\varphi(\mathbf{x})} \right\} \\
&= \mathbb{E} \left(\frac{\zeta^2 [\min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}]^2}{\varphi(\mathbf{x})} \right) + \mathbb{E} \left[\frac{\zeta^2 \{t^2(\mathbf{x}; \boldsymbol{\theta}^*) - T^2\}}{\varphi(\mathbf{x})} I\{t(\mathbf{x}; \boldsymbol{\theta}^*) > T\} \right] \\
&= \mathbb{E} \left(\left[\sqrt{\varphi(\mathbf{x})} - \frac{\zeta \min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}}{\sqrt{\varphi(\mathbf{x})}} \right]^2 \right) - \mathbb{E}\{\varphi(\mathbf{x})\} + 2\mathbb{E}[\zeta \min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}] \\
&\quad + \mathbb{E} \left[\frac{\zeta^2 \{t^2(\mathbf{x}; \boldsymbol{\theta}^*) - T^2\}}{\varphi(\mathbf{x})} I\{t(\mathbf{x}; \boldsymbol{\theta}^*) > T\} \right] \\
&= \mathbb{E} \left(\left[\sqrt{\varphi(\mathbf{x})} - \frac{\zeta \min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}}{\sqrt{\varphi(\mathbf{x})}} \right]^2 \right) + \mathbb{E} \left[\frac{\zeta^2 \{t^2(\mathbf{x}; \boldsymbol{\theta}^*) - T^2\}}{\varphi(\mathbf{x})} I\{t(\mathbf{x}; \boldsymbol{\theta}^*) > T\} \right] + 1 \\
&\equiv E_1 + E_2 + 1,
\end{aligned}$$

where $I(\cdot)$ is the indicator function. Note that E_i 's are non-negative and $E_1 = 0$ if and only if

$$\varphi(\mathbf{x}) = \zeta \min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\} = \frac{\min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}}{\mathbb{E}[\min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}]},$$

which is attainable because

$$\rho \min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\} \leq \mathbb{E}[\min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}], \quad \text{almost surely.} \quad (\text{S.13})$$

If $t(\mathbf{x}; \boldsymbol{\theta}^*) \leq T$ almost surely, then $E_2 = 0$ and the proof finishes. If $\Pr\{t(\mathbf{x}; \boldsymbol{\theta}^*) > T\} > 0$ then

$$\frac{\rho T}{\mathbb{E}[\min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}]} = 1,$$

because if this is not true then we can find a larger T that satisfies (S.13). This means that if $t(\mathbf{x}; \boldsymbol{\theta}^*) > T$, then

$$\varphi(\mathbf{x}) = \frac{\min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}}{\mathbb{E}[\min\{t(\mathbf{x}; \boldsymbol{\theta}^*), T\}]} = \rho^{-1},$$

which minimizes E_2 as well since $\varphi(\mathbf{x})$ is in the denominator. This finishes the proof.

S.4 Derivation of corrected model (4)

Note that $\pi(\mathbf{x}, 1) = 1$ and $\pi(\mathbf{x}, 0) = \pi(\mathbf{x})$. By direct calculation, we have

$$\begin{aligned}
& \Pr(y = 1 \mid \mathbf{x}, \delta = 1) \\
&= \frac{\Pr(y = 1, \delta = 1 \mid \mathbf{x})}{\Pr(\delta = 1 \mid \mathbf{x})} \\
&= \frac{\Pr(y = 1, \delta = 1 \mid \mathbf{x})}{\Pr(y = 1, \delta = 1 \mid \mathbf{x}) + \Pr(y = 0, \delta = 1 \mid \mathbf{x})} \\
&= \frac{\Pr(y = 1 \mid \mathbf{x}) \Pr(\delta = 1 \mid \mathbf{x}, y = 1)}{\Pr(y = 1 \mid \mathbf{x}) \Pr(\delta = 1 \mid \mathbf{x}, y = 1) + \Pr(y = 0 \mid \mathbf{x}) \Pr(\delta = 1 \mid \mathbf{x}, y = 0)} \\
&= \frac{p(\mathbf{x}; \boldsymbol{\theta}^*) \pi(\mathbf{x}, 1)}{p(\mathbf{x}; \boldsymbol{\theta}^*) \pi(\mathbf{x}, 1) + \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \pi(\mathbf{x}, 0)} \\
&= \frac{e^{g(\mathbf{x}; \boldsymbol{\theta}^*)}}{e^{g(\mathbf{x}; \boldsymbol{\theta}^*)} + \pi(\mathbf{x})} \\
&= \frac{e^{g(\mathbf{x}; \boldsymbol{\theta}^*) + l}}{1 + e^{g(\mathbf{x}; \boldsymbol{\theta}^*) + l}}.
\end{aligned}$$

S.5 Proof of Theorem 4

The estimator $\hat{\boldsymbol{\theta}}_{\text{lik}}$ is the maximizer of $\ell_{\text{lik}}(\boldsymbol{\theta})$, so $\mathbf{u}_N = a_N(\hat{\boldsymbol{\theta}}_{\text{lik}} - \boldsymbol{\theta})$ is the maximizer of

$$\gamma_{\text{lik}}(\mathbf{u}) = \ell_{\text{lik}}(\boldsymbol{\theta} + a_N^{-1}\mathbf{u}) - \ell_{\text{lik}}(\boldsymbol{\theta}).$$

By Taylor's expansion,

$$\gamma_{\text{lik}}(\mathbf{u}) = a_N^{-1}\mathbf{u}^T \dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*) + 0.5a_N^{-2} \sum_{i=1}^N \phi_{\pi}(\mathbf{x}_i; \boldsymbol{\theta}^*) \{\mathbf{u}^T \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\}^2 + \Delta_{\ell_{\text{lik}}} + R_{\text{lik}},$$

where $\phi_{\pi}(\mathbf{x}; \boldsymbol{\theta}) = p_{\pi}(\mathbf{x}; \boldsymbol{\theta})\{1 - p_{\pi}(\mathbf{x}; \boldsymbol{\theta})\}$,

$$p_{\pi}(\mathbf{x}; \boldsymbol{\theta}) = \frac{e^{g(\mathbf{x}; \boldsymbol{\theta})+l}}{1 + e^{g(\mathbf{x}; \boldsymbol{\theta})+l}} \quad \text{with} \quad l = -\log\{\rho\varphi(\mathbf{x})\}, \quad (\text{S.14})$$

$$\dot{\ell}_{\text{lik}}(\boldsymbol{\theta}) = \frac{\partial \ell_{\text{lik}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^N \delta_i \{y_i - p_{\pi}(\mathbf{x}_i; \boldsymbol{\theta})\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}),$$

$$\Delta_{\ell_{\text{lik}}} = \frac{1}{2a_N^2} \sum_{i=1}^N \delta_i \{y_i - p_{\pi}(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \mathbf{u}^T \ddot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mathbf{u},$$

and R_{lik} is the remainder term. By direct calculation,

$$\begin{aligned} R_{\text{lik}} &= \frac{1}{6a_N^3} \sum_{i=1}^N \delta_i \phi_{\pi}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}}) \{1 - 2p_{\pi}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}})\} \{\mathbf{u}^T \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}})\}^3 \\ &\quad + \frac{1}{6a_N^3} \sum_{i=1}^N \delta_i \phi_{\pi}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}}) \{\mathbf{u}^T \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}})\} \{\mathbf{u}^T \ddot{g}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}}) \mathbf{u}\} \\ &\quad + \frac{1}{6a_N^3} \sum_{i=1}^N \delta_i \left[\{y_i - p_{\pi}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}})\} \sum_{j_1, j_2=1}^d u_{j_1} u_{j_2} \ddot{g}_{j_1 j_2}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}}) \mathbf{u} \right], \end{aligned}$$

where $\dot{\mathbf{u}}$ lies between $\mathbf{0}$ and \mathbf{u} . We see that

$$\begin{aligned} |R_{\text{lik}}| &\leq \frac{\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \delta_i p_{\pi}(\mathbf{x}_i; \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}}) C(\mathbf{x}_i, \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}}) + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \delta_i y_i B(\mathbf{x}_i) \\ &\leq \frac{\|\mathbf{u}\|^3 e^{a_N^{-1}\|\mathbf{u}\|}}{6Na_N} \sum_{i=1}^N \delta_i \exp[f(\mathbf{x}_i; \boldsymbol{\beta} + a_N^{-1}\dot{\mathbf{u}}_{(-1)}) - \log\{\rho\varphi(\mathbf{x}_i)\}] C(\mathbf{x}_i, \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}}) \\ &\quad + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \delta_i y_i B(\mathbf{x}_i) \\ &= \frac{\|\mathbf{u}\|^3 e^{a_N^{-1}\|\mathbf{u}\|}}{6Na_N \rho} \sum_{i=1}^N \delta_i \varphi^{-1}(\mathbf{x}_i) \exp[f(\mathbf{x}_i; \boldsymbol{\beta} + a_N^{-1}\dot{\mathbf{u}}_{(-1)})] C(\mathbf{x}_i, \boldsymbol{\theta}^* + a_N^{-1}\dot{\mathbf{u}}) \\ &\quad + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \delta_i y_i B(\mathbf{x}_i) \\ &\leq \frac{d^3\|\mathbf{u}\|^3 e^{a_N^{-1}\|\mathbf{u}\|}}{6Na_N \rho} \sum_{i=1}^N \delta_i \varphi^{-1}(\mathbf{x}_i) B(\mathbf{x}_i) + \frac{d\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N y_i B(\mathbf{x}_i) \\ &\equiv \Delta_{R_{\text{lik}1}} + \Delta_{R2} = \Delta_{R_{\text{lik}1}} + o_P(1), \end{aligned}$$

where $\dot{\mathbf{u}}_{(-1)}$ is $\dot{\mathbf{u}}$ with the first element removed, $C(\mathbf{x}, \boldsymbol{\theta})$ and $\Delta_{R2} = o_P(1)$ are defined in (S.2). For $\Delta_{R_{\text{lik}1}}$, we see that

$$\mathbb{E}\left\{(N\rho)^{-1} \sum_{i=1}^N \delta_i \varphi^{-1}(\mathbf{x}_i) B(\mathbf{x}_i)\right\} = \rho^{-1} \mathbb{E}\left([\{p(\mathbf{x}_i; \boldsymbol{\theta}^*)\varphi^{-1}(\mathbf{x}) + \rho\{1 - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\}\} B(\mathbf{x})\right]$$

$$\leq e^{\alpha^*} \rho^{-1} \mathbb{E}\{e^{f(\mathbf{x}; \beta^*)} \varphi^{-1}(\mathbf{x}) B(\mathbf{x})\} + \mathbb{E}\{B(\mathbf{x})\}.$$

Therefore, $\mathbb{E}(\Delta_{R_{\text{lik}1}}) \rightarrow 0$. Since $\Delta_{R_{\text{lik}1}}$ is positive, Markov's inequality shows that $\Delta_{R_{\text{lik}1}} = o_P(1)$ and thus $R_{\text{lik}} = o_P(1)$.

For $\Delta_{\ell_{\text{lik}}}$, the mean $\mathbb{E}(\Delta_{\ell_{\text{lik}}}) = \mathbf{0}$ and the variance satisfies that

$$\mathbb{V}(\Delta_{\ell_{\text{lik}}}) \leq \frac{\|\mathbf{u}\|^4}{4a_N^4} \sum_{i=1}^N \mathbb{E}\{\delta_i p_\pi(\mathbf{x}_i; \boldsymbol{\theta}^*) \|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^2\} \leq \frac{\|\mathbf{u}\|^4}{4a_N^2} \mathbb{E}[e^{f(\mathbf{x}; \beta^*)} \|\dot{g}(\mathbf{x}, \boldsymbol{\theta}^*)\|^2] \rightarrow 0,$$

where the last step is because Assumption 1 implies that $e^{f(\mathbf{x}; \beta^*)} \|\dot{g}(\mathbf{x}, \boldsymbol{\theta}^*)\|^2$ is integrable, so $\Delta_{\ell_{\text{lik}}} = o_P(1)$.

If we can show that

$$a_N^{-1} \dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*) \longrightarrow \mathbb{N}(\mathbf{0}, \boldsymbol{\Lambda}_{\text{lik}}), \quad (\text{S.15})$$

in distribution, and

$$a_N^{-2} \sum_{i=1}^N \delta_i \phi_\pi(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \longrightarrow \boldsymbol{\Lambda}_{\text{lik}}, \quad (\text{S.16})$$

in probability, then from the Basic Corollary in page 2 of Hjort and Pollard (2011), we know that $a_N(\hat{\boldsymbol{\theta}}_{\text{lik}} - \boldsymbol{\theta}^*)$, the maximizer of $\gamma_{\text{lik}}(\mathbf{u})$, satisfies that

$$a_N(\hat{\boldsymbol{\theta}}_{\text{lik}} - \boldsymbol{\theta}^*) = \boldsymbol{\Lambda}_{\text{lik}}^{-1} \times a_N^{-1} \dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*) + o_P(1). \quad (\text{S.17})$$

Slutsky's theorem together with (S.15) and (S.17) implies the result in Theorem 1. We prove (S.15) and (S.16) in the following.

Note that $\dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*)$ is a summation of i.i.d. quantities, $\delta_i \{y_i - p_\pi(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)$'s, whose distribution depends on N . Note that

$$\begin{aligned} & \mathbb{E}[\delta \{y - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &= \mathbb{E}[\pi(\mathbf{x}, y) \{y - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &= \mathbb{E}\left([p(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\} - \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \rho \varphi(\mathbf{x}) p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)] \dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\right) = \mathbf{0}, \end{aligned} \quad (\text{S.18})$$

where the last step is by a direct calculation from (S.14) to obtain the following equality

$$p(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\} = \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \rho \varphi(\mathbf{x}) p_\pi(\mathbf{x}; \boldsymbol{\theta}^*). \quad (\text{S.19})$$

Thus we know that $\mathbb{E}\{a_N^{-1} \dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*)\} = \mathbf{0}$. The variance of $a_N^{-1} \dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*)$ satisfies that

$$\begin{aligned} & \mathbb{V}\{a_N^{-1} \dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*)\} \\ &= a_N^{-2} N \mathbb{E}[\delta \{y - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &= e^{-\alpha^*} \mathbb{E}[\{y + (1 - y) \rho \varphi(\mathbf{x})\} \{y - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &= e^{-\alpha^*} \mathbb{E}\left([p(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} \rho \varphi(\mathbf{x}) p_\pi^2(\mathbf{x}; \boldsymbol{\theta}^*)] \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right) \end{aligned} \quad (\text{S.20})$$

$$= e^{-\alpha^*} \mathbb{E}\left([p(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + p(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\} p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)] \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right) \quad (\text{S.21})$$

$$= e^{-\alpha^*} \mathbb{E}\left([\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + \{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\} p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)] p(\mathbf{x}; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right) \quad (\text{S.22})$$

$$= e^{-\alpha^*} \mathbb{E}\left([\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + \phi_\pi(\mathbf{x}; \boldsymbol{\theta}^*)] p(\mathbf{x}; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right) \quad (\text{S.23})$$

$$= \mathbb{E}\left([\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + \phi_\pi(\mathbf{x}; \boldsymbol{\theta}^*)] \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} e^{f(\mathbf{x}; \beta^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right) \quad (\text{S.24})$$

where the forth equality uses (S.19). Note that elements of the term in the expectation of (S.24) are all bounded by integrable random variable $e^{f(\mathbf{x}; \beta^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2$. Thus, from the dominated convergence theorem and the fact that

$$[\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + \phi_\pi(\mathbf{x}; \boldsymbol{\theta}^*)] \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}$$

$$= 1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*) + o_P(1) = \frac{1}{1 + c\varphi^{-1}(\mathbf{x})e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}} + o_P(1),$$

we have

$$\mathbb{V}\{a_N^{-1}\dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*)\} \longrightarrow \mathbb{E}\left[\frac{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)}{1 + c\varphi^{-1}(\mathbf{x})e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}\right] = \boldsymbol{\Lambda}_{\text{lik}}. \quad (\text{S.25})$$

Now we check the Lindeberg-Feller condition. For any $\epsilon > 0$,

$$\begin{aligned} & \sum_{i=1}^N \mathbb{E}\left[\|\delta_i\{y_i - p_\pi(\mathbf{x}_i; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^2 I(\|\delta_i\{y_i - p_\pi(\mathbf{x}_i; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &= N\mathbb{E}\left[\delta\|y - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\delta\|y - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &= N\mathbb{E}\left[p(\mathbf{x}; \boldsymbol{\theta}^*)\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\quad + N\mathbb{E}\left[\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\rho\varphi(\mathbf{x})\{p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\rho\varphi(\mathbf{x})p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\leq N\mathbb{E}\left[e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\quad + N\mathbb{E}\left[e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\leq a_N^2\mathbb{E}\left[e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\quad + a_N^2\mathbb{E}\left[e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &= o(a_N^2), \end{aligned}$$

where the inequality in the third step uses the following fact derived from (S.14)

$$\rho\varphi(\mathbf{x})p_\pi(\mathbf{x}; \boldsymbol{\theta}^*) = e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\} \leq e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}, \quad (\text{S.26})$$

and the last step is from the dominated convergence theorem. Thus, applying the Lindeberg-Feller central limit theorem (Section 2.8 of van der Vaart, 1998), we finish the proof of (S.15).

Now we prove (S.16). Denote

$$H_{\text{lik}} = a_N^{-2} \sum_{i=1}^N \delta_i \phi_\pi(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*).$$

We notice that

$$\begin{aligned} \mathbb{E}(H_{\text{lik}}) &= Na_N^{-2}\mathbb{E}[\{y + (1 - y)\rho\varphi(\mathbf{x})\}\phi_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &= e^{-\alpha^*}\mathbb{E}[\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\rho\varphi(\mathbf{x})p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &\quad + e^{-\alpha^*}\mathbb{E}[p(\mathbf{x}; \boldsymbol{\theta}^*)p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &= e^{-\alpha^*}\mathbb{E}[p(\mathbf{x}; \boldsymbol{\theta}^*)\{1 - p_\pi(\mathbf{x}; \boldsymbol{\theta}^*)\}^2\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &\quad + e^{-\alpha^*}\mathbb{E}[\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\rho\varphi(\mathbf{x})p_\pi^2(\mathbf{x}; \boldsymbol{\theta}^*)\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)] \\ &= \mathbb{V}\{a_N^{-1}\dot{\ell}_{\text{lik}}(\boldsymbol{\theta}^*)\} = \boldsymbol{\Lambda}_{\text{lik}} + o(1), \end{aligned} \quad (\text{S.27})$$

where the third equality is obtained by applying (S.19) and the forth equality is from (S.20). In addition, the variance of each component of H_{lik} is bounded by

$$\begin{aligned} & a_N^{-4}N\mathbb{E}\{\delta\phi_\pi^2(\mathbf{x}_i; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4\} \\ & \leq a_N^{-4}N\mathbb{E}[\{y + (1 - y)\rho\varphi(\mathbf{x})\}p_\pi(\mathbf{x}_i; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4] \\ & \leq a_N^{-4}N\mathbb{E}[\{p(\mathbf{x}; \boldsymbol{\theta}^*) + \rho\varphi(\mathbf{x})\}p_\pi(\mathbf{x}_i; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4] \\ & \leq 2a_N^{-4}Ne^{\alpha^*}\mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4\} = o(1), \end{aligned} \quad (\text{S.28})$$

where the last inequality is because of (S.26) and the last step is because Assumption 1 implies that $e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4$ is integrable. From (S.27) and (S.28), Chebyshev's inequality implies that $H_{\text{lik}} \rightarrow \boldsymbol{\Lambda}_{\text{lik}}$ in probability.

S.6 Proof of Theorem 5

Let $\kappa(\mathbf{x}) = 1 + c\varphi^{-1}(\mathbf{x})e^{f(\mathbf{x};\beta^*)}$. We first notice that

$$\begin{aligned} [\mathbb{E}\{e^{f(\mathbf{x};\beta^*)}\}]^{-1}\mathbf{V}_w &= \mathbf{M}_f^{-1}\mathbf{M}_f\mathbf{M}_f^{-1} + \mathbf{M}_f^{-1}c\Lambda_{\text{sub}}\mathbf{M}_f^{-1} \\ &= \mathbf{M}_f^{-1}\mathbb{E}[\kappa(\mathbf{x})e^{f(\mathbf{x};\beta^*)}\dot{g}^{\otimes 2}(\mathbf{x};\theta^*)]\mathbf{M}_f^{-1}, \end{aligned}$$

and

$$[\mathbb{E}\{e^{f(\mathbf{x};\beta^*)}\}]^{-1}\mathbf{V}_{\text{lik}} = \Lambda_{\text{lik}}^{-1} = [\mathbb{E}\{\kappa^{-1}(\mathbf{x})e^{f(\mathbf{x};\beta^*)}\dot{g}^{\otimes 2}(\mathbf{x};\theta^*)\}]^{-1}$$

We only need to show that

$$[\mathbb{E}\{e^{f(\mathbf{x};\beta^*)}\}]^{-1}\mathbf{V}_w \geq \Lambda_{\text{lik}}^{-1}.$$

This is proved by the following calculation

$$\begin{aligned} \mathbf{0} &\leq \mathbb{E}\left([\kappa^{1/2}(\mathbf{x})\mathbf{M}_f^{-1} - \kappa^{-1/2}(\mathbf{x})\Lambda_{\text{lik}}^{-1}]e^{f(\mathbf{x};\beta^*)/2}\dot{g}(\mathbf{x};\theta^*)\right]^{\otimes 2} \\ &= \mathbb{E}\left\{\mathbf{M}_f^{-1}\kappa(\mathbf{x})e^{f(\mathbf{x};\beta^*)}\dot{g}^{\otimes 2}(\mathbf{x};\theta^*)\mathbf{M}_f^{-1}\right\} + \mathbb{E}\left\{\Lambda_{\text{lik}}^{-1}\kappa^{-1}(\mathbf{x})e^{f(\mathbf{x};\beta^*)}\dot{g}^{\otimes 2}(\mathbf{x};\theta^*)\Lambda_{\text{lik}}^{-1}\right\} \\ &\quad - 2\mathbb{E}\left\{\mathbf{M}_f^{-1}e^{f(\mathbf{x};\beta^*)}\dot{g}^{\otimes 2}(\mathbf{x};\theta^*)\Lambda_{\text{lik}}^{-1}\right\} \\ &= [\mathbb{E}\{e^{f(\mathbf{x};\beta^*)}\}]^{-1}\mathbf{V}_w + \Lambda_{\text{lik}}^{-1} - 2\Lambda_{\text{lik}}^{-1} = [\mathbb{E}\{e^{f(\mathbf{x};\beta^*)}\}]^{-1}\mathbf{V}_w - \Lambda_{\text{lik}}^{-1}. \end{aligned}$$

If $c = 0$ then $\kappa(\mathbf{x}) = 1$, and we can directly verify that

$$[\mathbb{E}\{e^{f(\mathbf{x};\beta^*)}\}]^{-1}\mathbf{V}_w = \Lambda_{\text{lik}}^{-1} = \mathbf{M}_f^{-1}. \quad (\text{S.29})$$

S.7 Proof of Theorem 6

For sampled data, (5) tell us that the joint density w.r.t. the product counting measure of the responses given the features is $\exp\{\ell_{\text{lik}}(\theta) + \sum_{i=1}^N \delta_i y_i l_i\}$, whose support has a finite number of possible values. Thus the expectation $\mathbb{E}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \mid \mathbf{X}\}$ is a sum of finite number of elements, and therefore the partial derivatives w.r.t. θ can be passed under the integration sign in $\mathbb{E}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \mid \mathbf{X}\}$. Therefore we have

$$\begin{aligned} \mathbf{I} &= \frac{\partial}{\partial \theta^T} \mathbb{E}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \mid \mathbf{X}\} \\ &= \frac{\partial}{\partial \theta^T} \int \mathbf{U}(\theta; \mathcal{D}_\delta) \exp\left\{\ell_{\text{lik}}(\theta) + \sum_{i=1}^N \delta_i y_i l_i\right\} d\mathbf{y} \\ &= \int \dot{\mathbf{U}}(\theta; \mathcal{D}_\delta) \exp\left\{\ell_{\text{lik}}(\theta) + \sum_{i=1}^N \delta_i y_i l_i\right\} d\mathbf{y} \\ &\quad + \int \mathbf{U}(\theta; \mathcal{D}_\delta) \exp\left\{\ell_{\text{lik}}(\theta) + \sum_{i=1}^N \delta_i y_i l_i\right\} \dot{\ell}_{\text{lik}}^T(\theta) d\mathbf{y} \\ &= \mathbb{E}\{\dot{\mathbf{U}}(\theta; \mathcal{D}_\delta) \mid \mathbf{X}\} + \mathbb{E}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \dot{\ell}_{\text{lik}}^T(\theta) \mid \mathbf{X}\}, \end{aligned}$$

where $d\mathbf{y} = dy \times dy \times \dots dy$ is a product measure and dy is the counting measure. This gives us that

$$\mathbb{E}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \dot{\ell}_{\text{lik}}^T(\theta) \mid \mathbf{X}\} = \mathbf{I} \quad (\text{S.30})$$

Hence we have

$$\begin{aligned} \mathbb{V}\{\mathbf{U}(\theta; \mathcal{D}_\delta) - a_N^{-2} \Lambda_{\text{lik}}^{-1} \dot{\ell}_{\text{lik}}(\theta) \mid \mathbf{X}\} &\geq \mathbf{0} \\ &= \mathbb{V}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \mid \mathbf{X}\} + \mathbb{V}\{a_N^{-2} \Lambda_{\text{lik}}^{-1} \dot{\ell}_{\text{lik}}(\theta) \mid \mathbf{X}\} - 2a_N^{-2} \mathbb{E}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \dot{\ell}_{\text{lik}}^T(\theta) \mid \mathbf{X}\} \Lambda_{\text{lik}}^{-1} \\ &= \mathbb{V}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \mid \mathbf{X}\} + a_N^{-2} \Lambda_{\text{lik}}^{-1} \mathbb{V}\{a_N^{-1} \dot{\ell}_{\text{lik}}(\theta) \mid \mathbf{X}\} \Lambda_{\text{lik}}^{-1} - 2a_N^{-2} \Lambda_{\text{lik}}^{-1}. \end{aligned}$$

Taking expectation of the above conditional variance and using (S.25), we have

$$\mathbb{E}[\mathbb{V}\{\mathbf{U}(\theta; \mathcal{D}_\delta) \mid \mathbf{X}\}] + a_N^{-2} \Lambda_{\text{lik}}^{-1} \mathbb{E}[a_N^{-1} \mathbb{V}\{\dot{\ell}_{\text{lik}}(\theta) \mid \mathbf{X}\}] \Lambda_{\text{lik}}^{-1} - 2a_N^{-2} \Lambda_{\text{lik}}^{-1}$$

$$\begin{aligned}
&= \mathbb{V}\{\mathbf{U}(\boldsymbol{\theta}; \mathcal{D}_\delta)\} + a_N^{-2} \boldsymbol{\Lambda}_{\text{lik}}^{-1} \{1 + o_P(1)\} - 2a_N^{-2} \boldsymbol{\Lambda}_{\text{lik}}^{-1} \\
&= \mathbb{V}\{\mathbf{U}(\boldsymbol{\theta}; \mathcal{D}_\delta)\} - a_N^{-2} \boldsymbol{\Lambda}_{\text{lik}}^{-1} \{1 + o_P(1)\} \\
&= \mathbb{V}\{\mathbf{U}(\boldsymbol{\theta}; \mathcal{D}_\delta)\} - N_1^{-1} \mathbf{V}_{\text{lik}} \{1 + o_P(1)\} \geq \mathbf{0},
\end{aligned}$$

where the last equality is because (S.1) implies that $a_N^{-2} = N_1^{-1} \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\} \{1 + o_P(1)\}$. This finishes the proof.

S.8 Proof of Theorem 7

The outline of the proof is similar to that of the proof of Theorem 2. Write $\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\vartheta}}) = y + (1 - y)\pi_\varrho^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\vartheta}})$ and $\delta^{\tilde{\boldsymbol{\vartheta}}} = y + (1 - y)I\{u \leq \pi_\varrho^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\vartheta}})\}$ where $u \sim \mathbb{U}(0, 1)$. The estimator $\hat{\boldsymbol{\theta}}_w^{\tilde{\boldsymbol{\vartheta}}}$ is the maximizer of

$$\ell_w^{\tilde{\boldsymbol{\vartheta}}}(\boldsymbol{\theta}) = \sum_{i=1}^N \frac{\delta_i^{\tilde{\boldsymbol{\vartheta}}}}{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\vartheta}})} [y_i g(\mathbf{x}_i; \boldsymbol{\theta}) - \log\{1 + e^{g(\mathbf{x}_i; \boldsymbol{\theta})}\}],$$

so $a_N(\hat{\boldsymbol{\theta}}_w^{\tilde{\boldsymbol{\vartheta}}} - \boldsymbol{\theta}^*)$ is the maximizer of $\gamma_w^{\tilde{\boldsymbol{\vartheta}}}(\mathbf{u}) = \ell_w^{\tilde{\boldsymbol{\vartheta}}}(\boldsymbol{\theta}^* + a_N^{-1}\mathbf{u}) - \ell_w^{\tilde{\boldsymbol{\vartheta}}}(\boldsymbol{\theta}^*)$. By Taylor's expansion,

$$\gamma_w^{\tilde{\boldsymbol{\vartheta}}}(\mathbf{u}) = \frac{1}{a_N} \mathbf{u}^T \dot{\ell}_w^{\tilde{\boldsymbol{\vartheta}}}(\boldsymbol{\theta}^*) + \frac{1}{2a_N^2} \sum_{i=1}^N \frac{\delta_i^{\tilde{\boldsymbol{\vartheta}}}}{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\vartheta}})} \phi(\mathbf{x}_i; \boldsymbol{\theta}^*) (\mathbf{z}_i^T \mathbf{u})^2 + \Delta_{\ell_w}^{\tilde{\boldsymbol{\vartheta}}} + R_w^{\tilde{\boldsymbol{\vartheta}}},$$

where

$$\begin{aligned}
\dot{\ell}_w^{\tilde{\boldsymbol{\vartheta}}}(\boldsymbol{\theta}) &= \sum_{i=1}^N \frac{\delta_i^{\tilde{\boldsymbol{\vartheta}}}}{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\vartheta}})} \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta})\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}), \\
\Delta_{\ell_w}^{\tilde{\boldsymbol{\vartheta}}} &= \frac{1}{2a_N^2} \sum_{i=1}^N \frac{\delta_i^{\tilde{\boldsymbol{\vartheta}}}}{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\vartheta}})} \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \mathbf{u}^T \ddot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mathbf{u},
\end{aligned}$$

and $R_w^{\tilde{\boldsymbol{\vartheta}}}$ is the remainder. Similarly to the proof of Theorem 2, we only need to show that

$$a_N^{-1} \dot{\ell}_w^{\tilde{\boldsymbol{\vartheta}}}(\boldsymbol{\theta}^*) \longrightarrow \mathbb{N}(\mathbf{0}, \boldsymbol{\Lambda}_w^{\text{plt}}), \quad (\text{S.31})$$

in distribution with

$$\boldsymbol{\Lambda}_w^{\text{plt}} = \mathbb{E} \left[e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) + \frac{ce^{2f(\mathbf{x}_i; \boldsymbol{\beta}^*)}}{\max\{\varphi_{\text{plt}}(\mathbf{x}), c_l\}} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \right],$$

and for any \mathbf{u} ,

$$H_w^{\tilde{\boldsymbol{\vartheta}}} := a_N^{-2} \sum_{i=1}^N \frac{\delta_i^{\tilde{\boldsymbol{\vartheta}}}}{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\vartheta}})} \phi(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \longrightarrow \mathbf{M}_f, \quad (\text{S.32})$$

in probability, and $\Delta_{\ell_w}^{\tilde{\boldsymbol{\vartheta}}} = o_P(1)$ and $R_w^{\tilde{\boldsymbol{\vartheta}}} = o_P(1)$.

We prove (S.31) first. Let $\eta_i^{\tilde{\boldsymbol{\vartheta}}} = \frac{\delta_i^{\tilde{\boldsymbol{\vartheta}}}}{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\vartheta}})} \{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)$. Given $\tilde{\boldsymbol{\vartheta}}$, $\eta_i^{\tilde{\boldsymbol{\vartheta}}}$, $i = 1, \dots, N$, are i.i.d. with mean zero, and the variance satisfies that

$$\begin{aligned}
&\mathbb{V}(\eta_i^{\tilde{\boldsymbol{\vartheta}}} \mid \tilde{\boldsymbol{\vartheta}}) \\
&= \mathbb{E} \left[\frac{\{y_i - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\}^2}{\{y_i + (1 - y_i)\pi_\varrho^{\text{os}}(\mathbf{x}_i; \tilde{\boldsymbol{\vartheta}})\}} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\vartheta}} \right] \\
&= \mathbb{E} \left[p(\mathbf{x}_i; \boldsymbol{\theta}^*) \{1 - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\}^2 \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) + \{1 - p(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \frac{p^2(\mathbf{x}_i; \boldsymbol{\theta}^*)}{\pi_\varrho^{\text{os}}(\mathbf{x}_i; \tilde{\boldsymbol{\vartheta}})} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\vartheta}} \right] \\
&\leq \mathbb{E} \{ p(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) + \varrho^{-1} p^2(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\vartheta}} \} \\
&\leq e^{\alpha^*} \mathbb{E} \{ e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) + (\varrho^{-1} e^{\alpha^*}) e^{2f(\mathbf{x}_i; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \}
\end{aligned}$$

Since $\rho^{-1}\varrho \rightarrow c_l > 0$, for large enough N ,

$$\mathbb{V}\{a_N^{-1}\dot{\ell}_w^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\theta}}\} \leq \mathbb{E}\{e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)}\dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) + (1 + c_l^{-1}c)e^{2f(\mathbf{x}_i; \boldsymbol{\beta}^*)}\dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*)\}.$$

Thus, we know that

$$\mathbb{V}\{a_N^{-1}\dot{\ell}_w^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*)\} \rightarrow \Lambda_w^{\text{plt}}$$

Now we check the Lindeberg-Feller condition (Section *2.8 of van der Vaart, 1998) given $\tilde{\boldsymbol{\theta}}$. For any $\epsilon > 0$,

$$\begin{aligned} & \sum_{i=1}^N \mathbb{E}\{\|\eta_i^{\tilde{\boldsymbol{\theta}}}\|^2 I(\|\eta_i^{\tilde{\boldsymbol{\theta}}}\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}\} \\ &= N\mathbb{E}\left[\|\{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})\}^{-1}\delta^{\tilde{\boldsymbol{\theta}}}\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 \right. \\ &\quad \left. \times I(\|\{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})\}^{-1}\delta^{\tilde{\boldsymbol{\theta}}}\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}\right] \\ &= N\mathbb{E}\left[\|\{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})\}^{-1}\| \{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 \right. \\ &\quad \left. \times I(\|\{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})\}^{-1}\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}\right] \\ &+ N\mathbb{E}\left[\{1 - \pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})\} \|\{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})\}^{-1}y\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 \right. \\ &\quad \left. \times I(\|\{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})\}^{-1}y\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}\right] \\ &\leq N\mathbb{E}\left[p(\mathbf{x}; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\quad + N\mathbb{E}\left[\varrho^{-1}\|p(\mathbf{x}; \boldsymbol{\theta}^*)\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\varrho^{-1}p(\mathbf{x}; \boldsymbol{\theta}^*)\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\quad + N\mathbb{E}\left[p(\mathbf{x}; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\leq 2a_N^2\mathbb{E}\left[e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &\quad + a_N^2e^{\alpha^*}\varrho^{-1}\mathbb{E}\left[e^{2f(\mathbf{x}_i; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\varrho^{-1}e^{\alpha^*}\|e^{f(\mathbf{x}_i; \boldsymbol{\beta}^*)}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\right] \\ &= o(a_N^2), \end{aligned}$$

where the last step is due to Assumptions 1, the dominated convergence theorem, and the facts that $a_N \rightarrow \infty$ and $\lim_{N \rightarrow \infty} e^{\alpha^*}/\varrho = c/c_l < \infty$. Thus, applying the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998) finishes the proof of (S.31).

Now we prove (S.32). We notice that

$$\mathbb{E}(H_w^{\tilde{\boldsymbol{\theta}}} \mid \tilde{\boldsymbol{\theta}}) = \mathbb{E}\left\{\frac{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}}{(1 + e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)})^2}\dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)\right\} = \mathbf{M}_f + o(1), \quad (\text{S.33})$$

where the last step is by the dominated convergence theorem. In addition, the conditional variance of each component of $H_w^{\tilde{\boldsymbol{\theta}}}$ is bounded by

$$\frac{1}{N}\mathbb{E}\left\{\frac{e^{2f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^4}{\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})} \mid \tilde{\boldsymbol{\theta}}\right\} \leq \frac{1}{N\varrho}\mathbb{E}\{e^{2f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^4\} = o(1). \quad (\text{S.34})$$

where the last step is because of Assumption 1 and the fact that $Ne^{\alpha^*} \rightarrow \infty$, $e^{\alpha^*}/\varrho \rightarrow c < \infty$, and $\rho^{-1}\varrho \rightarrow c_l > 0$ imply that $N\varrho \rightarrow \infty$. From (S.33) and (S.34), applying Chebyshev's inequality finishes the proof of (S.32).

In the following we finish the proof by showing that given $\tilde{\boldsymbol{\theta}}$, $\Delta_{\ell_w}^{\tilde{\boldsymbol{\theta}}} = o_P(1)$ and $R_w^{\tilde{\boldsymbol{\theta}}} = o_P(1)$.

For $\Delta_{\ell_w}^{\tilde{\boldsymbol{\theta}}}$, the conditional mean $\mathbb{E}(\Delta_{\ell_w}^{\tilde{\boldsymbol{\theta}}} \mid \tilde{\boldsymbol{\theta}}) = \mathbf{0}$ and the conditional variance satisfies that

$$\mathbb{V}(\Delta_{\ell_w}^{\tilde{\boldsymbol{\theta}}} \mid \tilde{\boldsymbol{\theta}}) \leq \frac{\|\mathbf{u}\|^4 N}{4a_N^4} \mathbb{E}\left[\frac{\{y - p(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \|\ddot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2}{y + (1 - y)\pi_\varrho^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})} \mid \tilde{\boldsymbol{\theta}}\right]$$

$$\begin{aligned}
&\leq \frac{\|\mathbf{u}\|^4 N}{4a_N^4} \mathbb{E}[\{1 + \varrho^{-1} p(\mathbf{x}; \boldsymbol{\theta}^*)\} p(\mathbf{x}; \boldsymbol{\theta}^*) \|\ddot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2] \\
&\leq \frac{\|\mathbf{u}\|^4}{4a_N^2} \mathbb{E}[\{1 + \varrho^{-1} e^{\alpha^*} e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\} e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\ddot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2] = o(1),
\end{aligned}$$

so $\Delta_{\ell_w}^{\tilde{\boldsymbol{\theta}}} = o_P(1)$.

For the remainder term $R_w^{\tilde{\boldsymbol{\theta}}}$. By direct calculations similar to those used for the proof of Theorem 2, we know that

$$\begin{aligned}
|R_w^{\tilde{\boldsymbol{\theta}}}| &\leq \frac{d^3 \|\mathbf{u}\|^3 e^{a_N^{-1} \|\mathbf{u}\|}}{6Na_N} \sum_{i=1}^N \frac{\delta_i^{\tilde{\boldsymbol{\theta}}}}{\pi_{\varrho}^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})} B(\mathbf{x}_i) + \frac{d \|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \frac{\delta_i^{\tilde{\boldsymbol{\theta}}}}{\pi_{\varrho}^{\text{os}}(\mathbf{x}, y; \tilde{\boldsymbol{\theta}})} y_i B(\mathbf{x}_i) \\
&\equiv \Delta_{R_w1}^{\tilde{\boldsymbol{\theta}}} + \Delta_{R_w2}^{\tilde{\boldsymbol{\theta}}}.
\end{aligned}$$

From Assumption 1, $\mathbb{E}(\Delta_{R_w1}^{\tilde{\boldsymbol{\theta}}}) \rightarrow 0$ and $\mathbb{E}(\Delta_{R_w2}^{\tilde{\boldsymbol{\theta}}}) \rightarrow 0$. Since $\Delta_{R_w1}^{\tilde{\boldsymbol{\theta}}}$ and $\Delta_{R_w2}^{\tilde{\boldsymbol{\theta}}}$ are both positive, Markov's inequality shows that they are both $o_P(1)$ so $R_w^{\tilde{\boldsymbol{\theta}}} = o_P(1)$.

S.9 Proof of Theorem 8

The outline of the proof is similar to that of the proof of Theorem 4. The estimator $\hat{\boldsymbol{\theta}}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}$ is the maximizer of $\ell_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})$, so $\mathbf{u}^{\tilde{\boldsymbol{\theta}}} = a_N^{-1}(\hat{\boldsymbol{\theta}}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta})$ is the maximizer of

$$\gamma_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\mathbf{u}) = \ell_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta} + a_N^{-1} \mathbf{u}) - \ell_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}).$$

By Taylor's expansion,

$$\gamma_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\mathbf{u}) = a_N^{-1} \mathbf{u}^T \dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) + 0.5 a_N^{-2} \sum_{i=1}^N \phi_{\pi}(\mathbf{x}_i; \boldsymbol{\theta}^*) \{\mathbf{u}^T \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\}^2 + \Delta_{\ell_{\text{lik}}}^{\tilde{\boldsymbol{\theta}}} + R_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}, \quad (\text{S.35})$$

where $\phi_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}) = p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}) \{1 - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta})\}$,

$$p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}) = \frac{e^{g(\mathbf{x}; \boldsymbol{\theta}) + \tilde{l}}}{1 + e^{g(\mathbf{x}; \boldsymbol{\theta}) + \tilde{l}}}, \quad \text{with} \quad \tilde{l}_i = -\log\{\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})\}, \quad (\text{S.36})$$

$$\dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}) = \sum_{i=1}^N \delta_i^{\tilde{\boldsymbol{\theta}}} \{y_i - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta})\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}),$$

$$\Delta_{\ell_{\text{lik}}}^{\tilde{\boldsymbol{\theta}}} = \frac{1}{2a_N^2} \sum_{i=1}^N \delta_i^{\tilde{\boldsymbol{\theta}}} \{y_i - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \mathbf{u}^T \ddot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*) \mathbf{u}, \quad (\text{S.37})$$

and $R_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}$ is the remainder term. By direct calculations similar to those used for the proof of Theorem 4, we know that

$$|R_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}| \leq \frac{\|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \delta_i^{\tilde{\boldsymbol{\theta}}} C(\mathbf{x}_i, \boldsymbol{\theta}^* + a_N^{-1} \dot{\mathbf{u}}) + \frac{d \|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \delta_i^{\tilde{\boldsymbol{\theta}}} y_i B(\mathbf{x}_i) \leq \frac{(d+1) \|\mathbf{u}\|^3}{6a_N^3} \sum_{i=1}^N \delta_i^{\tilde{\boldsymbol{\theta}}} B(\mathbf{x}_i),$$

where $\dot{\mathbf{u}}$ lies between $\mathbf{0}$ and \mathbf{u} .

Since $\varrho = o(\rho)$, we know that for large enough N ,

$$\begin{aligned}
&\mathbb{E}\left\{a_N^{-3} \sum_{i=1}^N \delta_i^{\tilde{\boldsymbol{\theta}}} B(\mathbf{x}_i) \mid \tilde{\boldsymbol{\theta}}\right\} \\
&= Na_N^{-3} \mathbb{E}\left\{[y + (1-y)\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})] B(\mathbf{x}) \mid \tilde{\boldsymbol{\theta}}\right\} \\
&\leq Na_N^{-3} \mathbb{E}\left\{[e^{\alpha^*} e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} + \tilde{\omega}^{-1} \rho t(\mathbf{x}; \tilde{\boldsymbol{\theta}})] B(\mathbf{x}) \mid \tilde{\boldsymbol{\theta}}\right\}
\end{aligned}$$

$$\begin{aligned}
&\leq a_N^{-1} \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} B(\mathbf{x})\} + a_N^{-1} \tilde{\lambda}_{\min}^{-1} (e^{-\alpha^*} \rho) (e^{\tilde{\alpha} \tilde{\omega}})^{-1} \mathbb{E}\{e^{f(\mathbf{x}; \tilde{\boldsymbol{\theta}})} \|\dot{g}(\mathbf{x}; \tilde{\boldsymbol{\theta}})\| B(\mathbf{x}) \mid \tilde{\boldsymbol{\theta}}\} \\
&\leq a_N^{-1} \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} B(\mathbf{x})\} + a_N^{-1} \tilde{\lambda}_{\min}^{-1} (e^{-\alpha^*} \rho) (e^{\tilde{\alpha} \tilde{\omega}})^{-1} \mathbb{E}\{B^2(\mathbf{x})\} = o_P(1),
\end{aligned}$$

where $\tilde{\lambda}_{\min}$ is the minimum eigenvalue of $\tilde{\mathbf{M}}_{\mathbf{f}}$. Thus, Markov's inequality shows that $R_{\text{lik}}^{\tilde{\boldsymbol{\theta}}} = o_P(1)$. For $\Delta_{\ell_{\text{lik}}}^{\tilde{\boldsymbol{\theta}}}$, the conditional mean $\mathbb{E}(\Delta_{\ell_{\text{lik}}}^{\tilde{\boldsymbol{\theta}}} \mid \tilde{\boldsymbol{\theta}}) = \mathbf{0}$ and the conditional variance satisfies that

$$\begin{aligned}
\mathbb{V}(\Delta_{\ell_{\text{lik}}}^{\tilde{\boldsymbol{\theta}}} \mid \tilde{\boldsymbol{\theta}}) &\leq \frac{\|\mathbf{u}\|^4}{4a_N^4} \sum_{i=1}^N \mathbb{E}\{\delta_i^{\tilde{\boldsymbol{\theta}}} p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*) \|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^2 \mid \tilde{\boldsymbol{\theta}}\} \\
&\leq \frac{\|\mathbf{u}\|^4}{4a_N^2} \mathbb{E}[e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}, \boldsymbol{\theta}^*)\|^2] \rightarrow 0,
\end{aligned}$$

so $\Delta_{\ell_{\text{lik}}}^{\tilde{\boldsymbol{\theta}}} = o_P(1)$.

Now we need to show that given $\tilde{\boldsymbol{\theta}}$

$$a_N^{-1} \dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) \longrightarrow \mathbb{N}(\mathbf{0}, \boldsymbol{\Lambda}_{\text{lik}}^{\text{plt}}), \quad (\text{S.38})$$

in distribution, with

$$\boldsymbol{\Lambda}_{\text{lik}}^{\text{plt}} = \mathbb{E} \left[\frac{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*)}{1 + c\varphi_{\text{plt}}^{-1}(\mathbf{x}) e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}} \right],$$

and

$$H_{\text{lik}}^{\tilde{\boldsymbol{\theta}}} := a_N^{-2} \sum_{i=1}^N \delta_i^{\tilde{\boldsymbol{\theta}}} \phi_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}_i; \boldsymbol{\theta}^*) \longrightarrow \boldsymbol{\Lambda}_{\text{lik}}^{\text{plt}}, \quad (\text{S.39})$$

in probability, then from the Basic Corollary in page 2 of Hjort and Pollard (2011), we know that $a_N(\hat{\boldsymbol{\theta}}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}^*)$, the maximizer of $\gamma_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\mathbf{u})$, satisfies that

$$a_N(\hat{\boldsymbol{\theta}}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}} - \boldsymbol{\theta}^*) = (\boldsymbol{\Lambda}_{\text{lik}}^{\text{plt}})^{-1} \times a_N^{-1} \dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) + o_P(1). \quad (\text{S.40})$$

Slutsky's theorem together with (S.38) and (S.40) implies the result in Theorem 1. We prove (S.38) and (S.39) in the following.

Given $\tilde{\boldsymbol{\theta}}$, $\dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*)$ is a summation of i.i.d. quantities, $\delta_i^{\tilde{\boldsymbol{\theta}}} \{y_i - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)$'s. Using a similar approach to obtain (S.18) we know that $\mathbb{E}\{a_N^{-1} \dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\theta}}\} = \mathbf{0}$. The conditional variance of $a_N^{-1} \dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*)$ satisfies that

$$\begin{aligned}
&\mathbb{V}\{a_N^{-1} \dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\theta}}\} \\
&= e^{-\alpha^*} \mathbb{E}[\{y + (1 - y)\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})\} \{y - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\theta}}] \\
&= e^{-\alpha^*} \mathbb{E}([p(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + \pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}}) \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} p_{\pi}^2(\mathbf{x}; \boldsymbol{\theta}^*)] \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\theta}}) \\
&= e^{-\alpha^*} \mathbb{E}([\{1 - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + \phi_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}^2] p(\mathbf{x}; \boldsymbol{\theta}^*) \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\theta}}) \\
&= \mathbb{E}([\{1 - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 + \phi_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*) \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}^2] \{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\} e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \dot{g}^{\otimes 2}(\mathbf{x}; \boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\theta}}).
\end{aligned}$$

The elements of the term on the right hand side of the above equation are bounded by an integrable random variable $e^{f(\mathbf{x}; \boldsymbol{\beta}^*)} \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2$, so if $\varrho = o(\rho)$ then

$$\mathbb{V}\{a_N^{-1} \dot{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*)\} \longrightarrow \boldsymbol{\Lambda}_{\text{lik}}^{\text{plt}}. \quad (\text{S.41})$$

Using an similar approach as in checking the Lindeberg-Feller condition in the proof of Theorem 4, we obtain that for any $\epsilon > 0$,

$$\sum_{i=1}^N \mathbb{E} \left[\|\delta_i^{\tilde{\boldsymbol{\theta}}} \{y_i - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^2 I(\|\delta_i^{\tilde{\boldsymbol{\theta}}} \{y_i - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*)\} \dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\| > a_N \epsilon) \mid \tilde{\boldsymbol{\theta}} \right]$$

$$\begin{aligned}
&= N\mathbb{E}\left[\delta^{\tilde{\boldsymbol{\theta}}}\|\{y - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\delta^{\tilde{\boldsymbol{\theta}}}\|\{y - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}\right] \\
&= N\mathbb{E}[p(\mathbf{x}; \boldsymbol{\theta}^*)\{1 - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\{1 - p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}] \\
&\quad + N\mathbb{E}\left[\{1 - p(\mathbf{x}; \boldsymbol{\theta}^*)\}\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})\{p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\}^2 \|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 \right. \\
&\quad \quad \left. \times I(\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}\right] \\
&\leq N\mathbb{E}[p(\mathbf{x}; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}] \\
&\quad + N\mathbb{E}[e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon) \mid \tilde{\boldsymbol{\theta}}] \\
&\leq a_N^2 \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)\} \\
&\quad + a_N^2 \mathbb{E}[e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\|^2 I(e^{\alpha^* + f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}; \boldsymbol{\theta}^*)\| > a_N\epsilon)] \\
&= o(a_N^2),
\end{aligned}$$

where the last step is from the dominated convergence theorem. Thus, applying the Lindeberg-Feller central limit theorem (Section *2.8 of van der Vaart, 1998), we finish the proof of (S.38).

Now we prove (S.39). By similar derivations in (S.27), we know that

$$\mathbb{E}(H_{\text{lik}}^{\tilde{\boldsymbol{\theta}}} \mid \tilde{\boldsymbol{\theta}}) = \mathbb{V}\{a_N^{-1}\tilde{\ell}_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}^*) \mid \tilde{\boldsymbol{\theta}}\} = \boldsymbol{\Lambda}_{\text{lik}}^{\text{plt}} + o_P(1). \quad (\text{S.42})$$

where the second equality is from (S.41). In addition, the conditional variance of each component of $H_{\text{lik}}^{\tilde{\boldsymbol{\theta}}}$ is bounded by

$$\begin{aligned}
&a_N^{-4} N\mathbb{E}\{\delta^{\tilde{\boldsymbol{\theta}}}\{\phi_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*)\}^2 \|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4 \mid \tilde{\boldsymbol{\theta}}\} \\
&\leq a_N^{-4} N\mathbb{E}[\{y + (1 - y)\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})\}p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4 \mid \tilde{\boldsymbol{\theta}}] \\
&\leq a_N^{-4} N\mathbb{E}[\{p(\mathbf{x}; \boldsymbol{\theta}^*) + \pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}})\}p_{\pi}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4 \mid \tilde{\boldsymbol{\theta}}] \\
&\leq 2a_N^{-4} N e^{\alpha^*} \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4\} = 2a_N^{-2} \mathbb{E}\{e^{f(\mathbf{x}; \boldsymbol{\beta}^*)}\|\dot{g}(\mathbf{x}_i; \boldsymbol{\theta}^*)\|^4\} = o(1). \quad (\text{S.43})
\end{aligned}$$

From (S.42) and (S.43), Chebyshev's inequality implies that $H_{\text{lik}}^{\tilde{\boldsymbol{\theta}}} \rightarrow \boldsymbol{\Lambda}_{\text{lik}}^{\text{plt}}$ in probability.

S.10 Experiments on Pilot Mis-specification

This section complements Section 6.1 in the main text. We generate all data with a logistic regression model with $g(\mathbf{x}; \boldsymbol{\theta}) = \alpha + \mathbf{x}^T \boldsymbol{\beta}$. The total number of data is $N = 5 \times 10^5$ with N_0 negative samples and N_1 positive samples. When generating data, we tune the bias α to force $N_0/N_1 \approx 400$. In order to test the robustness of our methods, we add an artificial biased to the pilot model, which will produce skewed sampling probabilities through the equation

$$\pi_{\varrho}^{\text{os}}(\mathbf{x}; \tilde{\boldsymbol{\theta}}) = \min[\max\{\rho\tilde{\varphi}_{\text{os}}(\mathbf{x}), \varrho\}, 1], \quad \text{where} \quad \tilde{\varphi}_{\text{os}}(\mathbf{x}) = \tilde{\omega}^{-1}t(\mathbf{x}; \tilde{\boldsymbol{\theta}}). \quad (\text{S.44})$$

We try two types of biases: one on the intercept term α and the other on weights $\boldsymbol{\beta}$.

S.10.1 Perturb the pilot intercept $\tilde{\alpha}$.

We perturb $\tilde{\alpha} \leftarrow \tilde{\alpha} + \xi \times \mathbb{U}(0, 1) \times \log(N_0/N_1)$, where ξ is a hyperparameter controlling the degree of perturbation. We try $\xi = \{0.1, 0.5, 1.0\}$. As the result will show, perturbing the intercept will not affect the relative performance for various types of sampling algorithms.

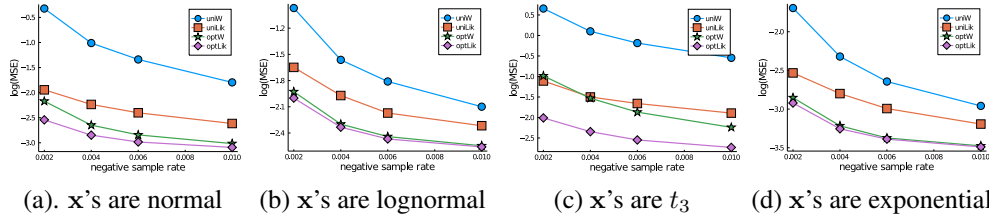


Figure S.4: Degree of perturbation $\xi = 0.1$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

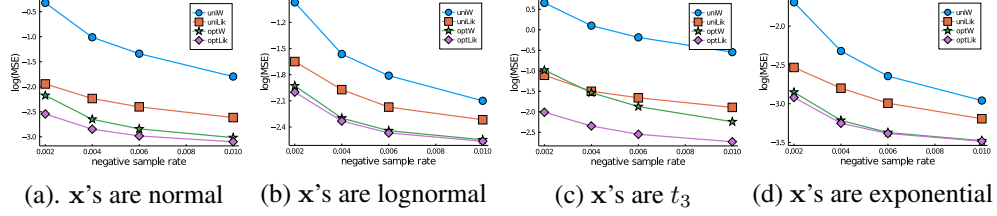


Figure S.5: Degree of perturbation $\xi = 0.5$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

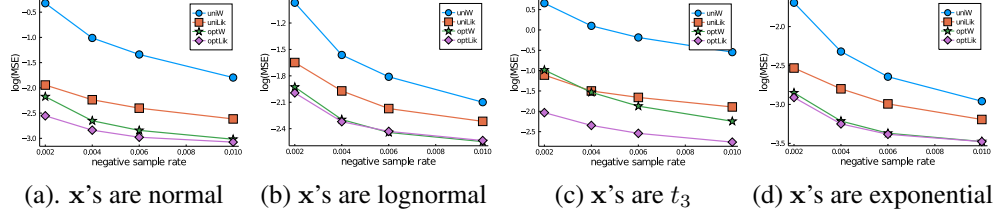


Figure S.6: Degree of perturbation $\xi = 1.0$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

S.10.2 Perturb the pilot weights β .

We perturb $\tilde{\beta} \leftarrow \tilde{\beta} + \xi \times \mathcal{N}(0, I)$, where ξ is a hyperparameter to control the degree of perturbation. We try $\xi = \{0.1, 0.5, 1.0\}$. As the result shows, the IPW estimator is pretty sensitive to the perturbation, and optW can be even worse than uniLik when the perturbation is significant (Figure S.9). uniLik may exceed optLik when the perturbation is too large such that the optimal sampling probability is significantly mis-calculated (Figure S.9(d)).

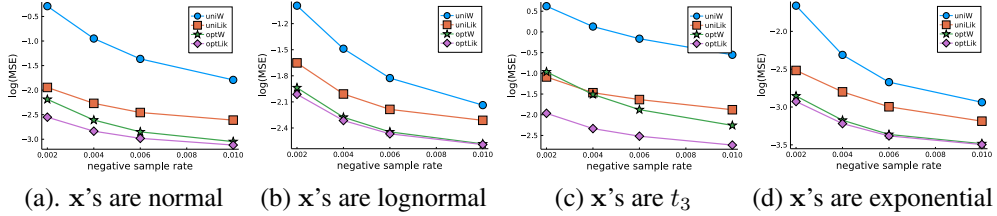


Figure S.7: Degree of perturbation $\xi = 0.1$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

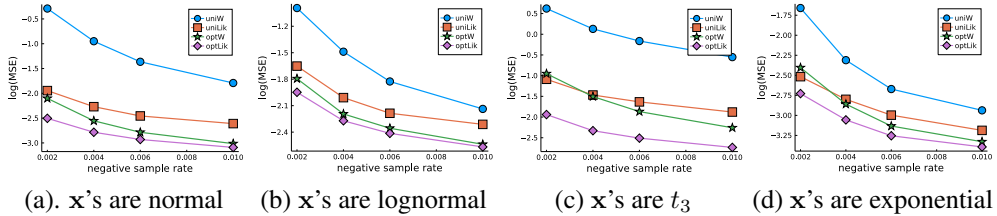


Figure S.8: Degree of perturbation $\xi = 0.5$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

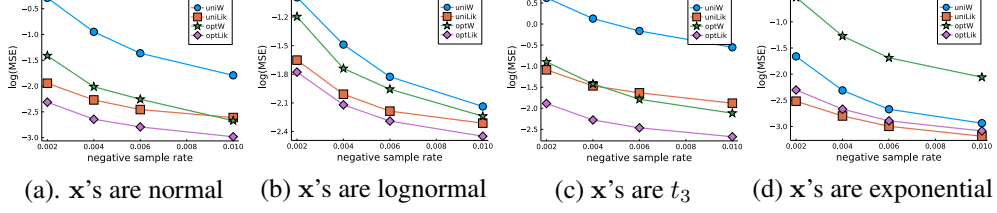


Figure S.9: Degree of perturbation $\xi = 1.0$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

S.11 Experiments on Model Mis-specification

This section studies the case when model is mis-specified. That is, the ground truth model fall out of the domain of linear logistic regression model. We design the ground truth model as a two-layer neural network with non-linear activations:

$$g(\mathbf{x}; \boldsymbol{\theta}) = \alpha + \frac{1}{\xi} \tanh(\xi \cdot \mathbf{x}^T \mathbf{W}) \beta, \quad (\text{S.45})$$

where ξ is a constant to control the “cut-off” threshold of the non-linear activation \tanh . Note that as $\xi \rightarrow 0$, the scaled \tanh function $\frac{1}{\xi} \cdot \tanh(\xi \cdot x) \approx x$ as shown in Figure S.10. Thus, by tuning ξ we have the control of the deviation of scaled \tanh function from a linear, identity transformation.

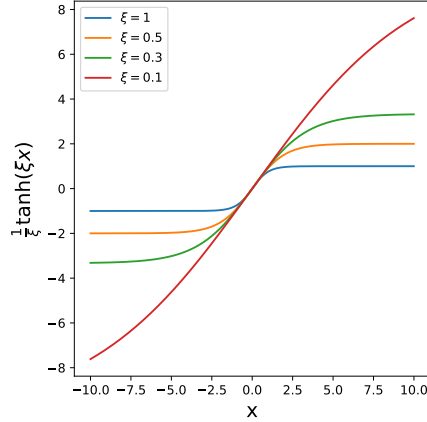


Figure S.10: Scaled \tanh function approximates identity transformation when $\xi \rightarrow 0$.

We also generate \mathbf{W} to be an approximately identity matrix, by setting the diagonal elements of \mathbf{W} to be 1, and draw off diagonal terms to be i.i.d. Normal distribution scaled by a factor ξ_w . Clearly, the operator norm $\|\mathbf{W} - I\|_{2 \rightarrow 2}$ is controlled by ξ_w . When $\xi_w = 0$, $\mathbf{W} = I$.

When perturbing the ground-truth model by setting ξ and ξ_w (thus \mathbf{W}), we need to reset the intercept α to match the pos/neg proportion to be $1/400$. For each experiment setup, we record the tuned α in the caption for reproducible research.

S.11.1 Ablation study.

When ξ and ξ_w are extremely small, $g(\mathbf{x}; \boldsymbol{\theta}) \approx \alpha + \mathbf{x}^T \beta$. Empirical results shown in Figure S.11 match results without model mis-specification.

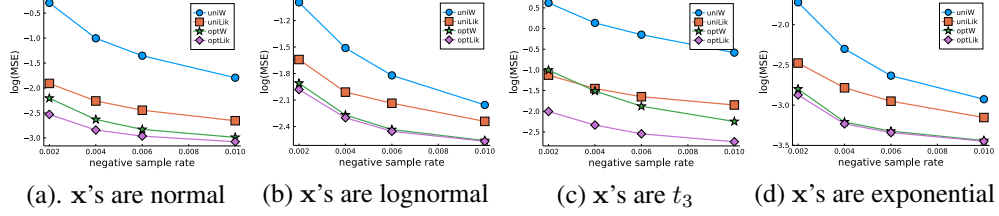


Figure S.11: Degree of non-linearity $\xi = 10^{-3}$, $\xi_w = 10^{-3}$. Ground-truth intercepts: (a) $\alpha = -7.65$, (b) $\alpha = -0.5$, (c) $\alpha = -7$, (d) $\alpha = -1.8$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

S.11.2 Tune the degree of activation function non-linearity by trying different ξ .

We fix $\xi_w = 10^{-3}$ and tune ξ to study the effect of non-linear activation. Note that when $\xi \rightarrow 0$, $\frac{1}{\xi} \tanh(\xi \mathbf{x}) \approx \mathbf{x}$ and when $\xi \rightarrow \infty$, $\frac{1}{\xi} \tanh(\xi \mathbf{x}) \approx 0$. As we shall see, the value ξ significantly affect the non-linearity of the model. In Figure S.12, we choose a relatively small $\xi = 0.1$ and observe an almost indistinguishable loss in $\log(\text{MSE})$. In Figure S.13 we study the case when $\xi = 0.5$ and find that it significantly affect the model estimation when \mathbf{x} has long tails (Figure S.13(b), (d)). When $\xi = 1$, all method crash since the true (non-linear) model deviate from the tractable domain (linear). In order to avoid this happen, we suggest to re-scale and centering the data before feeding them to the model. Note that when models are not crashed, optLk still achieves best performance among all methods.

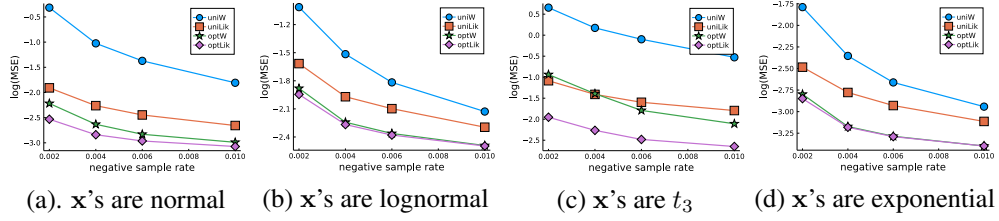


Figure S.12: Degree of non-linearity $\xi = 10^{-1}$, $\xi_w = 10^{-3}$. Ground-truth intercepts: (a) $\alpha = -7.7$, (b) $\alpha = -0.6$, (c) $\alpha = -6.9$, (d) $\alpha = -1.9$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

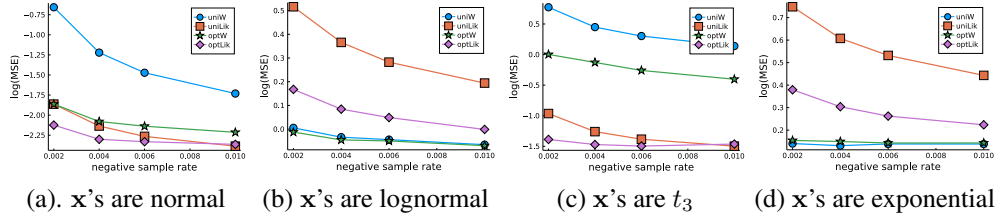


Figure S.13: Degree of non-linearity $\xi = 5 \cdot 10^{-1}$, $\xi_w = 10^{-3}$. Ground-truth intercepts: (a) $\alpha = -7.4$, (b) $\alpha = -0.95$, (c) $\alpha = -6.7$, (d) $\alpha = -2.15$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

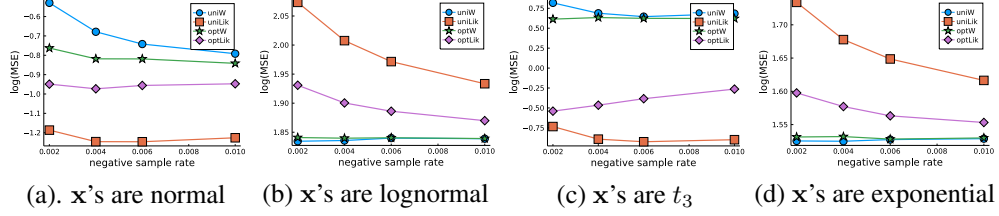


Figure S.14: Degree of non-linearity $\xi = 1.0, \xi_w = 10^{-3}$. Ground-truth intercepts: (a) $\alpha = -7.05$, (b) $\alpha = -1.75$, (c) $\alpha = -6.4$, (d) $\alpha = -2.85$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

S.11.3 Tune $\|\mathbf{W} - \mathbf{I}\|_{2 \rightarrow 2}$ by trying different ξ_w .

We fix $\xi = 10^{-3}$ and focus on studying the effect of generative process of \mathbf{W} to the estimation error. In Figure S.15, we set $\xi_w = 2 \cdot 10^{-2}$ to be a relatively small number such that $\|\mathbf{W} - \mathbf{I}\|_{2 \rightarrow 2} \approx 0.20 < 1$. In this case, each sampling method suffers from model mis-specification with higher $\log(\text{MSE})$. Still optLk is the best among all generative processes of \mathbf{x} . In Figure S.16, we set $\xi_w = 10^{-1}$ and $\|\mathbf{W} - \mathbf{I}\|_{2 \rightarrow 2} \approx 1$. In this case, the noise level is comparable to the ground-truth. All sampling models crash such that the estimated models deviate significantly from the ground-truth.

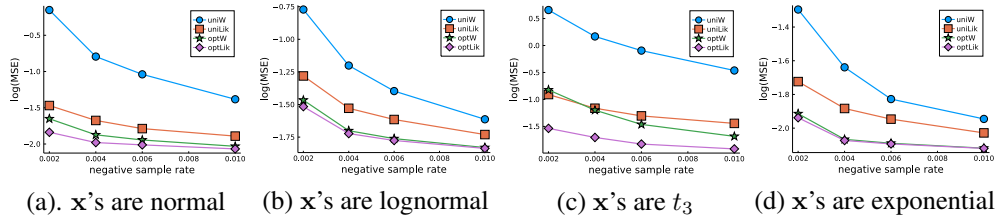


Figure S.15: Degree of non-linearity $\xi = 10^{-3}, \xi_w = 2 \cdot 10^{-2}$. Ground-truth intercepts: (a) $\alpha = -7.7$, (b) $\alpha = -0.5$, (c) $\alpha = -6.9$, (d) $\alpha = -1.9$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

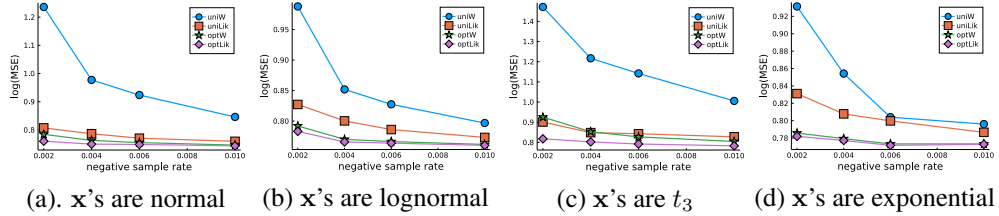


Figure S.16: Degree of non-linearity $\xi = 10^{-3}, \xi_w = 10^{-1}$. Ground-truth intercepts: (a) $\alpha = -8.25$, (b) $\alpha = -0.25$, (c) $\alpha = -6.6$, (d) $\alpha = -2.4$. Log (MSEs) of subsample estimators for different sample sizes (the smaller the better).

References

- Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*, 875–886. Springer.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357.
- Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter* **6**, 1, 1–6.
- Chen, J., Sun, B., Li, H., Lu, H., and Hua, X.-S. (2016). Deep ctr prediction in display advertising. In *Proceedings of the 24th ACM International Conference on Multimedia (MM)*, 811–820.

- Cheng, G., Huang, J. Z., *et al.* (2010). Bootstrap consistency for general semiparametric m-estimation. *The Annals of Statistics* **38**, 5, 2884–2915.
- Douzas, G. and Bacao, F. (2017). Self-organizing map oversampling (somo) for imbalanced data set learning. *Expert Systems with Applications* **82**, 40–52.
- Drummond, C., Holte, R. C., *et al.* (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on Learning from Imbalanced Datasets II*, vol. 11, 1–8. Citeseer.
- Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational intelligence* **20**, 1, 18–36.
- Fithian, W. and Hastie, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics* **42**, 5, 1693.
- Guo, H., Tang, R., Ye, Y., Li, Z., and He, X. (2017). Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*.
- Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, and G.-B. Huang, eds., *Advances in Intelligent Computing*, 878–887, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Han, L., Tan, K. M., Yang, T., Zhang, T., *et al.* (2020). Local uncertainty sampling for large-scale multiclass logistic regression. *Annals of Statistics* **48**, 3, 1770–1788.
- He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., Shi, Y., Atallah, A., Herbrich, R., Bowers, S., *et al.* (2014). Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 1–9.
- Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**, 2, 185–194.
- Hjort, N. L. and Pollard, D. (2011). Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*.
- Huang, J.-T., Sharma, A., Sun, S., Xia, L., Zhang, D., Pronin, P., Padmanabhan, J., Ottaviano, G., and Yang, L. (2020). Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2553–2561.
- Japkowicz, N. (2000). Learning from imbalanced data sets: Papers from the AAAI workshop, AAAI, 2000. Technical Report WS-00-05.
- Kilbertus, N., Parascandolo, G., and Schölkopf, B. (2018). Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*.
- King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis* **9**, 2, 137–163.
- Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research* **18**, 1, 559–563.
- Liu, X., Wu, J., and Zhou, Z. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39**, 2, 539–550.
- Mathew, J., Pang, C. K., Luo, M., and Leong, W. H. (2017). Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE Transactions on Neural Networks and Learning Systems* **29**, 9, 4065–4076.
- McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., *et al.* (2013). Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1222–1230.

- Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association* **95**, 449, 135–143.
- Owen, A. B. (2007). Infinitely imbalanced logistic regression. *The Journal of Machine Learning Research* **8**, 761–773.
- Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.
- Rahman, M. M. and Davis, D. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing* **3**, 2, 224.
- Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. In *ICML 2012: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, June 26-July 1, 2012*, 1255–1262.
- Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* **40**, 12, 3358–3378.
- Ting, D. and Brochu, E. (2018). Optimal subsampling with influence functions. In *Advances in neural information processing systems*, 3650–3659.
- van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, London.
- Wang, H. (2019). More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research* **20**, 132, 1–59.
- Wang, H. (2020). Logistic regression for massive data with rare events. In H. D. III and A. Singh, eds., *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 9829–9836. PMLR.
- Wang, H. and Ma, Y. (2021). Optimal subsampling for quantile regression in big data. *Biometrika* **108**, 1, 99–112.
- Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* **113**, 522, 829–844.
- Xiong, S. and Li, G. (2008). Some results on the convergence of conditional distributions. *Statistics & Probability Letters* **78**, 18, 3249–3253.
- Yu, J., Wang, H., Ai, M., and Zhang, H. (2020). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association* **0**, 0, 1–12.