Optimal Downsampling for Imbalanced Classification with Generalized Linear Models

Yan Chen
Duke University

Jose Blanchet
Stanford University

Laura Fee Nern
Yahoo Research

Krzysztof Dembczynski
Yahoo Research &
Poznan University of Technology

Aaron Flores
Yahoo Research

Abstract

Downsampling or under-sampling is a technique that is utilized in the context of large and highly imbalanced classification models. We study optimal downsampling for imbalanced classification using generalized linear models (GLMs). We propose a pseudo maximum likelihood estimator and study its asymptotic normality in the context of increasingly imbalanced populations relative to an increasingly large sample size. We provide theoretical guarantees for the introduced estimator. Additionally, we compute the optimal downsampling rate using a criterion that balances statistical accuracy and computational efficiency. Our numerical experiments, conducted on both synthetic and empirical data, further validate our theoretical results, and demonstrate that the introduced estimator outperforms commonly available alternatives.

1 INTRODUCTION

The problem of training a machine learning classification model with imbalanced populations (or, equivalently, predicting rare events given contextual data) arises in a wide range of applications such online advertisement, healthcare, insurance or fraud detection(Basha et al., 2022; Haixiang et al., 2017; Krawczyk, 2016; Sisodia and Sisodia, 2022, 2023; Razzaghi et al., 2015; Huda et al., 2016; Li et al., 2017b; Hassan and Abraham, 2016; More and Rana, 2017; Yan et al., 2015). A common method in these imbalanced settings is downsampling or under-sampling (Lee and Seo, 2022; Zhang et al., 2014; Wu et al., 2009; Elad and Feuer, 1997; Taha et al., 2021; Li et al., 2017a; Basha et al., 2022; Haixiang et al., 2017; Krawczyk, 2016; Ksieniewicz, 2018). For example, in online advertising the conversion rate is usually less than 0.1% out of hundreds of millions impressions, making downsampling a very attractive option for training to reduce the costs. The downsampling process involves sampling a proportion of the majority population according to a suitable sampling rate, called the downsampling rate. After that, a model is directly trained on the down-sampled set of instances. If the model is used directly on test data, it will be expected to predict according to the original distribution, so a correction is needed either during training or during the prediction phase. There are multiple ways in which one can correct for downsampling, these methods are well documented in the literature, one can either apply a proper rescaling of the estimated conditional probabilities (Elkan, 2001), one can reweight the samples in training procedure, or one can apply an additional correction step using, for example, isotonic regression (Niculescu-Mizil and Caruana, 2005). All of the above techniques will correct biases induced by downsampling, but obviously their performance might be significantly different.

Perhaps surprisingly, given how prevalent downsampling is in practice, to our knowledge, the “optimal" estimator (at least in theory) and its practical implications for applying downsampling have not been widely studied. The most recent paper by Wang (2020) partially addresses these issues by exploring the asymptotic property of the downsample estimator under imbalanced classification. The analysis, however, has been constrained to the logistic regression model and have not included any guidelines for selecting downsample rate. This motivates our study. While most of our dis-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

cussion focuses on generalized linear models, our results also offer insights into the application of downsampling in classification tasks using neural networks.

Our goal is to offer new estimators that are not only consistent and asymptotically normal, but are guided by optimal design in terms of variance. It is well known that the maximum likelihood estimator (MLE) attains the Cramer-Rao lower bound and therefore it is optimal in this sense. We then use this minimum variance insight to help users select an optimal downsampling rate.

Contributions and Overview of Results. Our contributions are as follows:

1) We provide an explicit characterization of the maximum likelihood estimator in highly imbalanced data.
2) We introduce new MLE-informed estimators that offer significant improvements relative to the available alternatives.
3) We provide guidance for the choice of optimal downsampling rates.

In order to formally capture highly imbalanced data, we consider an asymptotic regime of the form $\mathbb{P}(Y = 1|X = x) = 1 - F(\tau_n + \theta_*^T x) \equiv \bar{F}(\tau_n + \theta_*^T x)$, where F is the *cumulative distribution function* (c.d.f.) of an underlying latent variable. The highly imbalanced data setup is captured by the location parameter $\tau_n \to \infty$ so that $n(1 - F(\tau_n)) \to \infty$. By doing this, we capture a wide range of situations for which data is imbalanced relative to the sample size. We do not think that the data is becoming increasingly imbalanced as the sample size grows, rather, the asymptotic statistics provide tools for approximate inference that are applied for a fixed sample size and, in the same spirit, our scaling introduces approximations that are valid within a wide range of imbalanced proportions. This creates a regime of data imbalance in which the minority proportion converges to zero as the location parameter τ_n becomes extreme. This setup is useful in practice. For example, if $n = 10^6$, $\tau_n = 0.6\log(n) \approx 8$, and $\theta_*^T x \geq 1$. If F follows a logistic regression model such that $F(\tau_n + \theta_*^T x) = e^{\tau_n + \theta_*^T x}/(1 + e^{\tau_n + \theta_*^T x})$, then $\mathbb{P}(Y = 1|X = x) = 1/(1 + e^{\tau_n + \theta_*^T x}) < 0.1\%$. So by appropriately controlling the growth rate of τ_n, we can accurately recover the ratio of the rare events that are the focus of our analysis.

Firstly, we show that the prediction score from downsample still follows a GLM structure, such that $\mathbb{P}(\tilde{Y} = 1|\tilde{X} = x) = 1 - G(\tau_0 + \theta_*^T x) \equiv \bar{G}(\tau_0 + \theta_*^T x)$, where $(\tilde{Y}, \tilde{X})$ is the random variable induced from downsample data, α is the downsample rate, and $G(z) = \frac{\alpha F(z)}{1-(1-\alpha)F(z)}$ is also a c.d.f. Based on this discovery, we propose a *pseudo maximum likelihood estimator* (pseudo MLE) which can be computed directly on the downsample from imbalanced data by (6).

Secondly, we obtain the asymptotic normality of the proposed pseudo MLE (see Theorem 1 and Theorem 2). The estimator is unbiased and has generally a larger variance than the full-sample estimator (with zero downsampling rate). But for some small values of α the asymptotic variance stays the same as that of the full-sampling estimator, meaning that downsampling with these α results in no efficiency loss at all. We conduct numerical experiments and apply our estimator to logistic regression for different values of τ_n. We use the mean squared error metric and compare the pseudo MLE with two commonly used alternative estimators: the inverse-weighting estimator (8) employed by Wang (2020), and the conditional MLE (9). The findings are that our estimator outperforms both of them when τ_n is large, but as τ_n decreases, it loses its advantage gradually, which is consistent with our theoretical conclusions.

Additionally, we introduce a notion that trades the statistical error induced by the downsampling mechanism with the computational benefits derived by downsampling. We adapt the framework of Glynn and Whitt (1992) to introduce a concept of *efficiency cost* that reflects the trade-off between statistical error and the computational cost. Our reasoning is as follows.

Suppose that there is a given computational routine to be applied in the training process. We assume that the routine produces an estimator with an ϵ error with complexity $O(\log(1/\epsilon)) \times n \times (\alpha(1 - p_1) + p_1)$, where n is the size of the original training set, α is the downsampling rate and p_1 is the minority population rate. The contribution of $\log(1/\epsilon)$ arises when training via gradient descent of a smooth convex loss (Rakhlin et al., 2011; Foster et al., 2019; Bubeck et al., 2015; Johnson and Zhang, 2013) and is used as an illustration of our reasoning (and is applicable to logistic regression, for example); the discussion that follows can be adapted to the cost incurred using other methods with different assumptions. In the instance we adopted here, consequently, with a computational budget of size $b = c \times (\alpha(1 - p_1) + p_1)n\log(n)\log(\log(n))$ for some $c > 0$, we obtain a mean squared error $\sigma(\alpha)^2/n + o(1/n)$, where $\sigma(\alpha)^2/n$ is the mean squared error of the asymptotic downsampled estimator. We can write n in terms of the budget, obtaining, for large n, $n = b(1+o(1))/(c\times(\alpha(1-p_1)+p_1)\log(b)\log(\log(b)))$. Therefore, minimizing the total error subject to a given budget constraint is asymptotically equivalent to just minimizing $\sigma(\alpha)^2 \times (\alpha(1 - p_1) + p_1)$ as a function of α. We provide expressions to guarantee this optimal choice of α in practice in Section 5 and we apply our theoretical findings to logistic regression in Section 6.

Finally, we apply our estimator to real-world imbalanced datasets using both logistic regression and neural network models. The numerical results indicate that our estimator performs well in this experiment. More details can be found in Section 7.2 and Appendix A.

Related Work The concept of Generalized Linear Model (GLM) has a long history dating back to as early as its introduction by Nelder and Wedderburn (1972) in the 1970s. Specifically, GLM plays an important role in classification tasks both in theory and applications (Gorriz et al., 2021; Dobson and Barnett, 2018; Deng et al., 2022). For example, a statistical test was developed by Gorriz et al. (2021) based on GLM for pattern classification in machine learning applications. A high-dimensional GLM classification problem was explored by Hsu et al. (2021) via support vector machine classifiers. GLM classifiers were also applied by Ding and Gentleman (2005) and Arnold et al. (2020) in healthcare and computational biology.

Classification for imbalanced data has generated the interest of many researchers (Japkowicz, 2000; King and Zeng, 2001; Chawla et al., 2004; Chawla, 2010; Douzas and Bacao, 2017; LemaÃŽtre et al., 2017; Estabrooks et al., 2004; Fithian and Hastie, 2014; Han et al., 2005; Mathew et al., 2017; Rahman and Davis, 2013; Drummond et al., 2003; Owen, 2007). In this context, also referred to as a rare event setting, one class within a population occurs much less frequently. One example is online advertising data, where typically only 1 conversion is observed out of 1,000 to 1,000,000 impressions (Lee et al., 2012, 2021; Shah and Nasnodkar, 2021; Jiang et al., 2021). An extensive survey of the area is given by Chawla et al. (2004), where significant attention was given to downsampling methods, in which some data from the positive examples are removed for a more balanced subsample. Another approach is oversampling (Douzas and Bacao, 2017; Mathew et al., 2017; Wang, 2020; Shelke et al., 2017; Yan et al., 2019) in which additional instances are generated as positive examples. Imbalanced classification is also closely related to optimization of complex performance metrics such as the F-measure or Balanced Accuracy (Menon et al., 2013; Dembczyński et al., 2017; Xu et al., 2020; Aghbalou et al., 2024).

Our work focuses on exploring the effects of downsampling (Fithian and Hastie, 2014; Wang, 2020; Chawla et al., 2004), in which we maintain all the positive examples and uniformly sample an equal number of negative examples to make a more balanced subsample dataset. This is also referred to as “case-control sampling” by previous literature (Fithian and Hastie, 2014). A method of subsampling for logistic regression proposed by Fithian and Hastie (2014) was to adjust the class balance locally in feature space via an accept-reject scheme.

The findings of a recent paper by Wang (2020) suggest that the available information in rare event data for binary logistic regression is at the scale of the number of positive examples, and downsampling a small proportion of the controls may induce an estimator with identical asymptotic distribution to the full-data maximum likelihood estimator (MLE). Besides, a different perspective was provided by Drummond et al. (2003) for the performance analysis of sampling technique by considering cost sensitivity of misclassification. Motivated by the previous work, we study the imbalanced binary classification problem with a generalized linear model structure. Similar to (Wang, 2020), we derive the asymptotic distribution of the downsampling estimator, and come to a similar conclusion that downsampling a small proportion of the controls by selecting downsampling rate within a certain range results in identical statistical efficiency as the full-sample MLE.

However, our paper is more general than Wang (2020) in the following aspects. Firstly, we consider the generalized linear model as a more general and commonly seen family of models. Secondly, we use a new covariate-adapted estimator based on the maximum likelihood estimator for the downsample distribution, which turns out to be more efficient than the estimator used by Wang (2020) under rare-event setup. Thirdly, we also provide clear guidance on the selection of downsample rate.

Notation Given a thrice differentiable function f, we use $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$ to denote the first-order, second-order or the third-order derivatives. We use $\bar{f}(\cdot)$ to denote $1 - f(\cdot)$. For a sequence of random variables $Z_1, Z_2, \dots$ in a metric space $\mathcal{Z}$, we say $Z_n \overset{d}{\to} Z$ if Z_n converges in distribution to Z, and $Z_n \overset{p}{\to} Z$ if Z_n converges in probability to Z. Given matrix A, B, $A \succ B$ indicates $A - B$ is positive definite. Given target parameter $\theta \in \mathbb{R}^d$ and its estimator $\hat{\theta} \in \mathbb{R}^d$, we use the mean squared error or the mean squared estimation error to refer to $\mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right]$. We denote $\mu(\cdot)$ as the density of X and $\mathcal{X}$ as the covariate space.

2 PROBLEM SETUP

We observe an i.i.d. generated dataset $\{(Y_i, X_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d, d \geq 1$, $Y_i \in \{0, 1\}$. We use $\mathbb{P}$ to denote the joint probability distribution of (X_i, Y_i), and $\mathbb{E}[\cdot]$ to denote the expectation induced by $\mathbb{P}$. The data generating process follows *Generalized Lin-*

ear Model (GLM), such that given a latent random variable Z, the observed binary label Y given covariate X is defined as $Y = \mathbf{1}\left(Z > \tau_n + \theta_*^T X\right)$, with $\bar{F}_Z(\tau_n + \theta_*^T x) = \mathbb{P}\left(Y = 1 \middle| X = x\right)$, where $\bar{F}(\cdot) := \bar{F}_Z(\cdot) = 1 - F_Z(\cdot)$ and $F_Z(\cdot)$ is the *cumulative distribution function* (c.d.f.) of Z. We are interested in studying the impact of downsampling for the unbalanced dataset with $\mathbb{P}(Y = 1) \ll \mathbb{P}(Y = 0)$. To do this we keep all the positive samples (i.e., those with $Y_i = 1$), and uniformly sample a proportion $\alpha \in (0, 1]$ of the samples with $Y_i = 0$, so that we get a more balanced downsample $\{(\tilde{Y}_i, \tilde{X}_i)\}_{i=1}^N$ of size $N = np_1 + (1 - p_1)n\alpha$, where p_1 is the ratio of positive samples. We use $\tilde{\mu}$ to denote the joint density of downsample random pairs $(\tilde{Y}, \tilde{X})$.

Specifically, we consider the case where τ_n is a known sequence such that $\tau_n \to \infty$ as $n \to \infty$. This corresponds to the rare-event setup where $\bar{F}(\tau_n + \theta_*^T x) \to 0$ for any $x \in \mathcal{X}$, where $\mathcal{X}$ is bounded in $\mathbb{R}^d$. The assumption that $\tau_n \to \infty$ is without loss of generality by noting that if $\tau_n \to -\infty$, then $\mathbb{P}(Y = 0) \approx 0$ and $\mathbb{P}(Y = 1) \approx 1$, so the analysis is similar to the case $\tau_n \to \infty$ by switching the role of $\mathbb{P}(Y = 0)$ and $\mathbb{P}(Y = 1)$. We consider the following data generation process: Given sample size n and τ_n, we observe i.i.d. data $\{X_i, Y_i\}_{i=1}^n$ with

$$\mathbb{P}(Y_i = 1|X_i) = \bar{F}(\tau_n + \theta_*^T X_i) = 1 - F(\tau_n + \theta_*^T X_i),$$

where θ_* is the estimation target.

3 PROPOSED ESTIMATOR

In this section, we first demonstrate that under the Generalized Linear Model (GLM) setup, the induced variables by downsampling still follow a GLM model. Based on this finding, we propose a *pseudo maximum likelihood estimator* which can be computed directly from downsample.

Given joint law $\mathbb{P}$ of the full-sample random variable (Y, X), and denote the induced joint law of downsample random variable $(\tilde{Y}, \tilde{X})$ as $\tilde{P}$, there exists a monotone function m such that $\mathbb{P}(Y = 1|X; \theta) = m(\tilde{P}(Y = 1|X; \theta))$, where the probability distributions are parametrized by $\theta \in \Theta$. Note that this justifies the correction method for downsampling by imposing a monotone transformation on the predicted scores based on the model trained from the downsample. Besides, we explicitly formulate the downsample maximum likelihood estimator (MLE), where the target parameter for the downsample MLE coincides with the one from the full-sample model.

One common practice for estimation on imbalanced dataset is that people usually fit the parameter as if the downsample model still belongs to the same class of models as $F(\cdot)$, after which they obtain $F(\tau_n + \hat{\theta}_1^T X)$ as the score from the model trained on the downsample, and then use isotonic regression, i.e. utilizing a monotone mapping $g(\cdot)$ so that $g(F(\tau_n + \hat{\theta}_1^T X))$ is obtained as the final output of their prediction score given any covariate X_i and τ_n, $\hat{\theta}_1$ is achieved by solving the following problem:

$$\begin{aligned}\hat{\theta}_1 = \operatorname{argmax}_{\theta_1} \tfrac{1}{N} \textstyle\sum_{i=1}^N \tilde{Y}_i \log\Big(1 - F(\tau_n + \theta_1^T \tilde{X}_i)\Big) \\ + (1 - \tilde{Y}_i) \log F(\tau_n + \theta_1^T \tilde{X}_i).\end{aligned} \tag{1}$$

Lemma 1 in Appendix B shows a counterexample to illustrate why this method can lead to biased predictions for some covariates — meaning that there may be no monotone mapping which can fully recover the true predictions at all covariates, with $\hat{\theta}_1$ obtained from (1).

Specifically, Lemma 1 constructs a counterexample where $\hat{\theta}_1$ leads to biasedness. If $h(x) = \tilde{\theta}_1^T x$ maps two distinct x, x' to the same value while $\theta_*^T x \neq \theta_*^T x'$, and the c.d.f. $F(\cdot)$ is strictly increasing, then the true model has different probabilities for x, x' whereas the model trained on the downsample maps them to the same prediction score, thus both the parameter estimator and the induced prediction score model are biased.

Proposition 1 (Downsample Prediction Score)**.** *The prediction score of the downsampled random variables follows* $\mathbb{P}(\tilde{Y}_i = 1|\tilde{X}_i) = \frac{\mathbb{P}(Y_i=1|X_i=1)}{\alpha+(1-\alpha)\mathbb{P}(Y_i=1|X_i=1)}$*. Hence* $\{(\tilde{Y}_i, \tilde{X}_i)\}_{i=1}^N$ *still follows GLM, with the conditional probability* $\mathbb{P}(\tilde{Y}_i = 1|\tilde{X}_i) = \bar{G}(\tau_n + \theta_1^T \tilde{X}_i)$*, and* $\bar{G}(z) = 1 - G(z)$*, where*

$$G(z) := \frac{\alpha F(z)}{1 - (1 - \alpha)F(z)} \tag{2}$$

is the c.d.f. of some random variable.

Proposition 1 illustrates that the downsample random variables generated from GLM still follow a GLM, where $F(z) = \frac{G(z)}{\alpha+(1-\alpha)G(z)}$, indicating that the true class probabilities are the monotonic transformations of the predicted class probabilities conditional on the downsample. Based upon this, the joint likelihood of the downsample random variables is as follows.

Proposition 2 (Downsample Joint Likelihood of $(\tilde{Y}, \tilde{X})$)**.** *Given any* τ_n*, the maximum likelihood estimator for the target estimand* θ_* *(i.e., the true value of the unknown parameter) is*

$$\hat{\theta} := \operatorname*{argmax}_{\theta_1} \frac{1}{N} \sum_{i=1}^N \ell(\tilde{X}_i, \tilde{Y}_i, \theta_1; \tau_n), \tag{3}$$

where

$$\begin{aligned}&\ell(\tilde{x}, \tilde{y}, \theta_1; \tau_n) \\ &= \tilde{y} \log \bar{F}(\tau_n + \theta_1^T \tilde{x}) + (1 - \tilde{y}) \log \alpha F(\tau_n + \theta_1^T \tilde{x}) \\ &\quad - \log \int_{\mathcal{X}} \left[1 - (1 - \alpha)F(\tau_n + \theta_1^T x)\right] \mu(x) dx.\end{aligned} \tag{4}$$

Proposition 2 exhibits the empirical maximum likelihood estimator of downsample GLM with respect to the joint distribution of $(\tilde{Y}, \tilde{X})$. However, we note that (4) utilizes the marginal density of full-sample X, which is expensive to estimate for large-scale imbalanced data. To tackle this, we first note that the density of down-sample $\tilde{X}$ can be written as $\tilde{\mu}(x) = \frac{[1-(1-\alpha)F(\tau_n+\theta_*^T x)]\mu(x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}$ (see Lemma 4 in Appendix B), where $\mu(\cdot)$ is the density of full-sample X. Thus the last term in (4) can be rewritten as

$$\log \int_{\mathcal{X}} \left[1-(1-\alpha)F(\tau_n+\theta_1^T x)\right]\mu(x)dx$$

$$= \log \int_{\mathcal{X}} [1-(1-\alpha)F(\tau_n+\theta_1^T x)] \times \frac{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}{1-(1-\alpha)F(\tau_n+\theta_*^T x)}\tilde{\mu}(x)dx.$$

Note that when $\tau_n \to \infty$, $1-(1-\alpha)F(\tau_n+\theta_*^T x) \to \alpha$, and $\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0) = \mathbb{E}[1-(1-\alpha)F(\tau_n+\theta_*^T X)] \to \alpha$, thus as $n \to \infty$, we have

$$\frac{\log \int_{\mathcal{X}} [1-(1-\alpha)F(\tau_n+\theta_1^T x)] \frac{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}{1-(1-\alpha)F(\tau_n+\theta_*^T x)}\tilde{\mu}(x)dx}{\log \int_{\mathcal{X}} [1-(1-\alpha)F(\tau_n+\theta_1^T x)]\tilde{\mu}(x)dx} \to 1. \tag{5}$$

Furthermore, by the law of large numbers, the sample average $\frac{1}{N}\sum_{i=1}^N [1-(1-\alpha)F(\tau_n+\theta_1^T \tilde{X}_i)]$ converges in probability to $\int_{\mathcal{X}} [1-(1-\alpha)F(\tau_n+\theta_1^T x)]\tilde{\mu}(x)dx$. Based upon the discovery above, we propose the following pseudo maximum likelihood estimator computed from downsample data:

$$\hat{\theta}_* := \operatorname{argmax}_{\theta_1} \tfrac{1}{N}\textstyle\sum_{i=1}^N \tilde{\ell}(\tilde{X}_i, \tilde{Y}_i, \theta_1; \tau_n), \tag{6}$$

where

$$\begin{aligned} &\tilde{\ell}(\tilde{X}_i, \tilde{Y}_i, \theta_1; \tau_n) \\ = &\tilde{Y}_i \log\Big(1 - F(\tau_n + \theta_1^T \tilde{X}_i)\Big) \\ &+(1-\tilde{Y}_i)\log \alpha F(\tau_n+\theta_1^T \tilde{X}_i) \\ &-\log\left\{\tfrac{1}{N}\textstyle\sum_{i=1}^N \Big[1-(1-\alpha)F(\tau_n+\theta_1^T \tilde{X}_i)\Big]\right\}. \end{aligned} \tag{7}$$

In the following, we use $\hat{\theta}_*$ to denote the pseudo MLE computed from (6).

Remark 1. Note that (5) holds only when $\tau_n \to \infty$, but the estimator $\hat{\theta}_*$ may be not consistent when $\tau_n < \infty$. So in practice the pseudo MLE should perform well for large values of τ_n, and its benefit could disappear as τ_n decays. We verify this claim through numerical experiment on logistic regression models by comparing the performance of $\hat{\theta}_*$ under different values of τ_n in Section 7.1.

Additionally, there are two alternative natural estimators for θ_*, which are *inverse-weighting estimator* and *conditional maximum likelihood estimator* (a.k.a. *conditional MLE*). The inverse weighting estimator $\hat{\theta}_I$ is defined as (Wang, 2020):

$$\begin{aligned} \hat{\theta}^I := \operatorname{argmax}_{\theta_1} \tfrac{1}{N}\textstyle\sum_{i=1}^N \tilde{Y}_i \log\Big(1-F(\tau_n+\theta_1^T\tilde{X}_i)\Big) \\ +\tfrac{1-\tilde{Y}_i}{\alpha}\log F(\tau_n+\theta_1^T\tilde{X}_i), \end{aligned} \tag{8}$$

and the conditional MLE is defined as:

$$\begin{aligned} \hat{\theta}^C := \operatorname{argmax}_{\theta_1} \tfrac{1}{N}\textstyle\sum_{i=1}^N \tilde{Y}_i \log\Big(1-G(\tau_n+\theta_1^T\tilde{X}_i)\Big) \\ +(1-\tilde{Y}_i)\log G(\tau_n+\theta_1^T\tilde{X}_i), \end{aligned} \tag{9}$$

where $G(\cdot)$ is defined as in (2). By the theory of M-estimators (Van der Vaart, 2000), both of the estimators are consistent, i.e. $\hat{\theta}^I \xrightarrow{p} \theta_*$ and $\hat{\theta}^C \xrightarrow{p} \theta_*$. We compare the performance of our proposed pseudo MLE $\hat{\theta}_*$ as (6) with both $\hat{\theta}^I$ and $\hat{\theta}^C$ through numerical experiments on synthetic data (see Section 7.1) and empirical data (see Section 7.2 and Appendix A). We find that $\hat{\theta}_*$ outperforms $\hat{\theta}^I$ and $\hat{\theta}^C$ for large values of τ_n, but it gradually under-performs as we decrease the value τ_n, which verifies our conjecture as in remark 1, implying its value under rare-event setup.

4 ASYMPTOTIC NORMALITY

In this section, we discover that for the rare event setup, there is a certain range of downsample rate $\alpha \ll 1$, as long as it doesn't go to zero too fast, downsampling with rate α maintains the statistical efficiency as that of the full-sampling estimator. To demonstrate our statement, we consider the case where $\mathbb{P}(Y=1|X=x) = 1-F(\tau_n+\theta_*^T X)$ with $\tau_n \to \infty$.

Failure of Classical Maximum Likelihood Estimator Analysis Given the definition of $\tilde{\ell}(\cdot,\cdot,\cdot;\tau_n)$ above, we note that $\mathbb{E}\left[\tilde{\ell}(\tilde{X},\tilde{Y},\theta_*;\tau_n)\right] \to 0$ regardless of θ_* if $\alpha > 0$. Thus as $n \to \infty$, the criterion function doesn't rely on the value of θ_1 in the varying-rate regime, hence the classical MLE theory cannot be applied here. A more detailed discussion is presented in the Appendix C. We now use an alternative way to derive the asymptotic normality of $\hat{\theta}_*$.

Assumption 1. $F \in C^1(\mathbb{R})$ is strictly increasing, and the covariate space $\mathcal{X} \subset \mathbb{R}^d$ is compact.

Assumption 2. The matrix $\mathbb{E}[XX^T]$ is nonsingular.

Assumption 3. F is thrice differentiable and there exists a thrice differentiable function $h : \mathbb{R} \to \mathbb{R}_+$ such that the function $\bar{F}(\cdot+\tau_n)/\bar{F}(\tau_n)$ together with its first three derivatives converges uniformly over compact sets to $h(\cdot)$ and its three first derivatives denoted as $h^{(1)}, h^{(2)}, h^{(3)}$, respectively.

The distributions that satisfy Assumption 3 include among others exponential distribution and logistic regression model. We now provide our first main result.

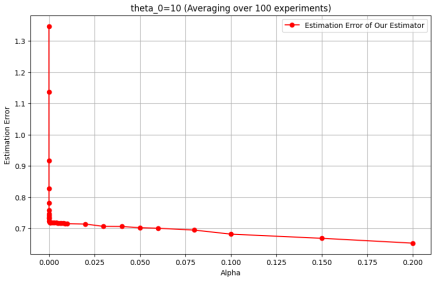


Figure 1: Estimation error for different α.

Theorem 1 (Asymptotic Normality of MLE as $\tau_n \to \infty$)**.** *Suppose Assumptions 1, 2, 3 hold, and for some $\Theta \subset \mathbb{R}^d$ as a neighborhood of θ_*, $\{F(\tau_n + \theta_1^T x) : \theta_1 \in \Theta\}$ is differentiable in quadratic mean at θ_*. Assume that $n(1-F(\tau_n)) \to \infty$, $\lim_{n\to\infty} \frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha} = c$, and*

$$\begin{aligned}&\mathbb{E}\left[\frac{h^{(1)}(\theta_*^T X)^2}{h(\theta_*^T X)} X X^T\right] \\ &\succ c\mathbb{E}\left[h^{(1)}(\theta_*^T X)X\right]\mathbb{E}\left[h^{(1)}(\theta_*^T X)X^T\right],\end{aligned} \tag{10}$$

then as $\tau_n \to \infty$, we have $\sqrt{n(1-F(\tau_n))}(\hat\theta_* - \theta_*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{-1}\right)$, where

$$\begin{aligned}\mathbf{V} \;&= \mathbb{E}\left[\frac{h^{(1)}(\theta_*^T X)^2}{h(\theta_*^T X)} X X^T\right] \\ &-c\mathbb{E}\left[h^{(1)}(\theta_*^T X)X\right]\mathbb{E}\left[h^{(1)}(\theta_*^T X)X^T\right].\end{aligned}$$

Remark 2. Theorem 1 suggests that if

$$\lim_{n\to\infty} \frac{(1-\alpha^2)(1-F(\tau_n))}{\alpha} = 0,$$

then the resulting estimator is as efficient as the full-sample estimator (i.e., when $\alpha = 1$). This happens when either α is bounded from below by a positive constant, or $\alpha \to 0$ but $\frac{1-F(\tau_n)}{\alpha} \to 0$, which is consistent with the discovery of Remark 2 in (Wang, 2020) for logistic regression in the rare event case.

Remark 3. We demonstrate the necessity of condition (10) through a numerical illustration. Imagine that when α is too close to zero, the right hand side of (10) can blow up so the condition will be violated. We look at a case of logistic regression where $\tau_n = 10$, $\theta_* = 0.5$, and $F(\tau_n + \theta_*^T x) = \frac{e^{\tau_n + \theta_*^T x}}{1+e^{\tau_n+\theta_*^T x}}$ and $X \sim \text{Unif}[0,1]$. We generate 10^5 random samples and $\mathbb{P}(Y=1) \approx 5.89 \times 10^{-5}$. The estimation error is computed as the average value of $|\theta_* - \hat\theta_*|$ from 500 random experiments. The x-axis corresponds to the α as the downsample rate from negative samples. Figure 1 shows that when α is too close to 0, the estimation error is large. But when α falls in a proper range such that condition (10) holds, the estimation error for $\hat\theta_*$ stabilizes and is kept as a small value, consistent with the theoretical findings that the mean squared error should be small and $\mathbb{E}[\|\hat\theta_* - \theta_*\|^2] \approx \text{tr}[\mathbf{V}^{-1}]/(n(1-F(\tau_n)))$.

Generalized Scaled Asymptotic Normality Now we generalize Assumption 3 to Assumption 4, which is satisfied by more distributions in addition to those satisfying Assumption 3, including both heavy-tailed distributions (e.g. Pareto) and very light tailed distributions (e.g. Gaussian tails).

Assumption 4. Suppose that F is thrice differentiable and that there exists a thrice differentiable function $g : \mathbb{R} \to \mathbb{R}_+$ such that the function $\bar F(r(\tau_n)\cdot + \tau_n)/\bar F(\tau_n)$ together with its first three derivatives converges uniformly over compact sets to $g(\cdot)$ and its three first derivatives denoted as $g^{(1)}, g^{(2)}, g^{(3)}$, respectively.

Assumption 4 includes more common distributions except for those satisfying Assumption 3. For example, for the standard normal distribution, we can take $r(\tau) = \frac{1}{\tau}$ and $g(z) = e^{-z}$. For the Pareto distribution, we can take $r(\tau) = \tau$ and $g(z) = \left(\frac{1}{1+z}\right)^\gamma$ with $\gamma > 1$ to satisfy Assumption 4, and we assume $\theta_*^T x > -1$ for any $x \in \mathcal{X}$, so $r(\tau_n)\{\theta_*^T x\} + \tau_n = \tau_n(1 + \theta_*^T x) > 0$ for any $x \in \mathcal{X}$, and the limit is well defined. Now we can extend our result of the asymptotic distribution to more general cases.

Theorem 2 (Generalized Scaled Asymptotic Normality)**.** *Suppose the underlying binary classification model is defined as $Y = \mathbf{1}(\tau_n + r(\tau_n)\theta_*^T X > 0)$ with $r(\cdot)$ defined as in Assumption 4. Assume that $n(1-F(\tau_n)) \to \infty$, $\lim_{n\to\infty} \frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha} = c$, and*

$$\begin{aligned}&\mathbb{E}\left[\frac{g^{(1)}(\theta_*^T X)^2}{g(\theta_*^T X)} X X^T\right] \\ &\succ c\mathbb{E}\left[g^{(1)}(\theta_*^T X)X\right]\mathbb{E}\left[g^{(1)}(\theta_*^T X)X^T\right].\end{aligned} \tag{11}$$

Then as $\tau_n \to \infty$, we have

$$\sqrt{n(1-F(\tau_n))}r(\tau_n)(\hat\theta_* - \theta_*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{-1}\right),$$

where

$$\begin{aligned}\mathbf{V} \;&= \mathbb{E}\left[\frac{g^{(1)}(\theta_*^T X)^2}{g(\theta_*^T X)} X X^T\right] \\ &-c\mathbb{E}\left[g^{(1)}(\theta_*^T X)X\right]\mathbb{E}\left[g^{(1)}(\theta_*^T X)X^T\right].\end{aligned}$$

Based on the asymptotic covariances, we are now ready to formulate the optimization problem by considering both statistical efficiency and computational gains. We now introduce the efficiency concept with a computational budget constraint.

5 EFFICIENCY WITH A BUDGET CONSTRAINT

Theorem 1 and Theorem 2 indicate that full-sampling (i.e. $\alpha = 1$) and downsampling with a rate α such that $\alpha \ll 1$ and $\frac{1-F(\tau_n)}{\alpha} = o(1)$ lead to the same rate of convergence and asymptotic MSE. At the same time, if we choose $\alpha \approx \mathbb{P}(Y=1)$, we still have the same convergence rate while the trace norm of the asymptotic covariance is kept as O(1). Intuitively, downsampling with a much smaller α can help to reduce computational cost significantly while maintaining statistical efficiency under the current rare event setup. This motivates us to formulate the choice of optimal downsampling rate by considering this tradeoff between statistical efficiency and computational cost with a budget constraint.

We adapt the framework of Glynn and Whitt (1992) for the definition of algorithm efficiency. Let the computational cost be some strictly increasing function of the downsample size, i.e., $C(\alpha; n, p_1) = f(np_1 + \alpha(1-p_1)n) = f(n[p_1 + \alpha(1-p_1)])$, where $f(\cdot)$ is strictly increasing, $p_1 = \mathbb{P}(Y = 1)$ and $C(\alpha; n, p_1)$ is the cost function of sampling α proportion of the negative samples (i.e., observations with label $Y = 0$) with the original sample size n and positive ratio p_1. According to the asymptotic efficiency principle in the canonical case of Glynn and Whitt (1992), we define the *asymptotic algorithm efficiency cost value* as the product of the sampling variance and the cost rate, i.e., $v(\alpha; p_1) = \lim_{n\to\infty} C(\alpha; n, p_1)\mathrm{Var}\left(\hat{\theta}_*\right)$. Specifically, when $f(\cdot)$ is a linear function, i.e., with some constant c_0, we have

$$v(\alpha; p_1) = \lim_{n\to\infty} c_0 n \left(p_1 + \alpha(1-p_1)\right) \mathrm{Var}\left(\hat{\theta}_*\right). \quad (12)$$

Recall from Theorem 1 that under regular conditions, when $\tau_n \to \infty$, we have $\sqrt{n(1-F(\tau_n))}(\hat{\theta}_* - \theta_*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{-1}\right)$. So $\frac{n(1-F(\tau_n))\mathrm{tr}\left[\mathrm{Cov}(\hat{\theta}_* - \theta_*)\right]}{\mathrm{tr}\left(\mathbf{V}^{-1}\right)} \to 1$ as $n \to \infty$. Under the definition of the efficiency cost (12), we have $\frac{c_0(p_1+\alpha(1-p_1))d^2}{\mathrm{tr}((1-F(\tau_n))\mathbf{V})} \le v(\alpha; p_1) \le \frac{c_0(p_1+\alpha(1-p_1))\kappa d^2}{\mathrm{tr}((1-F(\tau_n))\mathbf{V})}$, where κ is the condition number of $\mathbf{V}$. A natural objective for efficiency optimization is $\min_{\alpha\in[0,1]} \lim_{n\to\infty} \frac{p_1+\alpha(1-p_1)}{\mathrm{tr}(\mathbf{V})}$.

Theorem 3 (Optimal Downsampling Rate for Imbalanced Classification)**.** *Suppose Assumptions 1, 2, 3, and $\tau_n \to \infty$, $n(1-F(\tau_n)) \to \infty$, then the optimal choice of downsampling rate is*

$$\alpha^* = \frac{2(1-F(\tau_n))\mathrm{tr}\left\{\mathbb{E}\left[h^{(1)}(\theta_*^T X)X\right]\mathbb{E}\left[h^{(1)}(\theta_*^T X)X^T\right]\right\}}{\mathrm{tr}\left\{\mathbb{E}\left[\left\{h^{(1)}(\theta_*^T X)^2/h(\theta_*^T X)\right\}XX^T\right]\right\}}.$$

Henceforth, we have obtained explicit formulations of the optimization problems for selecting the optimal downsample rate. Theorem 3 has provided guidelines for downsampling schemes in practice. The result can easily be extended to generalized scaled result from Theorem 2.

6 APPLICATION TO LOGISTIC REGRESSION

Equipped with all the findings from the previous sections, we now focus on the logistic regression model as an application. Given a sequence of $\tau_n \to \infty$, define $F(\tau_n + \theta_1^T x) := \frac{e^{\tau_n+\theta_1^T x}}{1+e^{\tau_n+\theta_1^T x}}$. A direct application of Theorem 1 indicates:

Proposition 3 (Asymptotic Normality for Logistic Regression)**.** *Assume $\frac{n}{1+e^{\tau_n}} \to \infty$ and $\lim_{n\to\infty} \frac{(1-\alpha)^2}{\alpha(1+e^{\tau_n})} = c$ where c is a constant. Then as $n \to \infty$, we have*

$$\sqrt{n\mathbb{P}(Y=1)}\left(\hat{\theta}_* - \theta_*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbb{E}_X\left[e^{-\theta_*^T X}\right]\mathbf{V}^{-1}\right),$$

$$\textit{where } \mathbf{V} = \mathbb{E}\left[e^{-\theta_*^T X} XX^T\right] - c\mathbb{E}\left[e^{-\theta_*^T X} X\right]\mathbb{E}\left[e^{-\theta_*^T X} X\right].$$

Remark 4. If either α is bounded away from zero, i.e., $\exists$ some absolute constant $\underline{\alpha} \in (0,1)$ such that $0 < \underline{\alpha} \le \alpha \le 1$, or $\alpha \to 0$ and $\frac{1}{\alpha(1+e^{\tau_n})} \to 0$, then $c = 0$, and the asymptotic distribution is the same as that of the full-sampling case, so the estimator is as efficient as the full-sampling one. This is also consistent with the findings of (Wang, 2020).

Remark 5. Our estimator is different from Wang (2020) which uses the inverse weighting in the estimator (8). We illustrate the differences in the performance of our estimator and that of the estimator considered by Wang (2020) further through numerical experiments in Section 7.1.

Proposition 4 (Optimal Downsampling Rate for Logistic Regression)**.** *Suppose $\tau_n \to \infty$, $\frac{n}{1+e^{\tau_n}} \to \infty$, the optimal choice of downsampling rate is*

$$\alpha^* = \frac{2(1+e^{\tau_n})^{-1}\mathrm{tr}\left\{\mathbb{E}[e^{-\theta_*^T X}X]\mathbb{E}[e^{-\theta_*^T X}X]\right\}}{\mathrm{tr}\{\mathbb{E}[e^{-\theta_*^T X}XX^T]\}}.$$

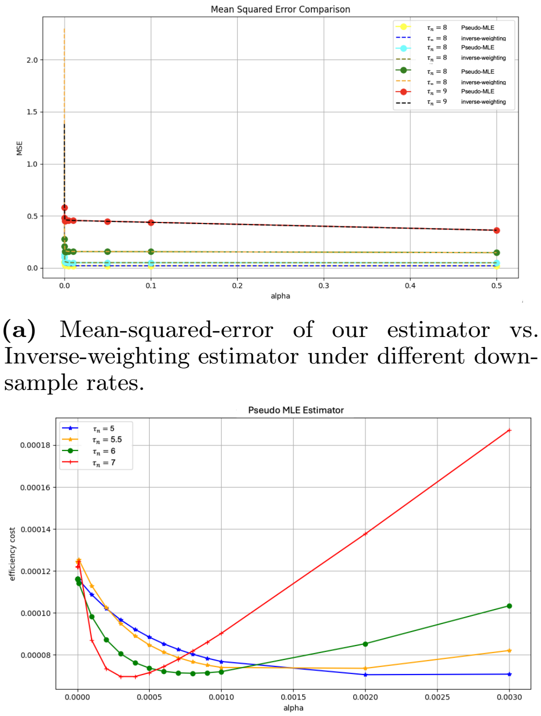
7 EXPERIMENTAL STUDIES

In this section we illustrate the above results empirically in two types of studies. We first run numerical experiments using logistic regression on synthetic data. Next, we use benchmark data sets for which we do not know the underlining model.

7.1 Experiments on Synthetic Data

We focus on a setting where $\theta_* = 0.5$, and the covariates X are drawn i.i.d. from a uniform distribution $[0,1]$. The sample size is $n = 10^5$. We fix different values of τ_n and estimate θ_* for these different τ_n. Under each α, we compute the solutions to (3), i.e., maximum likelihood estimators, under 500 random environments, and we also compute the mean squared error (MSE) for θ_* with respect to each α by averaging over the 500 environments.

Firstly, we compare the mean squared estimation error of our estimator and the inverse weighting estimator



(a) Mean-squared-error of our estimator vs. Inverse-weighting estimator under different downsample rates.

(b) Efficiency cost combining both statistical efficiency and computational cost for $\tau_n = 5, 5.5, 6, 7$.

Figure 2: Mean squared error and Efficiency costs

(Wang, 2020) for $\tau_n = 6, 7, 8, 9$ for some of the $\alpha \in [0.00005, 0.5]$. These values τ_n correspond to $\mathbb{P}(Y = 1)$ approximately equal to $0.002, 0.0007, 0.0002, 0.000097$. The numerical results shown in Figure 2a are consistent with our findings: When α is small and falls into the proper range satisfying the conditions of Theorem 1, the mean square error is close to that generated by $\alpha = 0.5$.

Secondly, we focus on the range of α very close to $\mathbb{P}(Y = 1)$ for $\tau_n = 10.0, 9.8, 6.0, 5.0$, we compute the mean squared error for logistic regression, which corresponds to the cases where $\mathbb{P}(Y = 1)$ is approximately equal to 3.57×10^{-5}, 4.36×10^{-5}, 0.0019, 0.0053 respectively. We replicate our simulations 500 times for each τ_n to compare the inverse weighting, the conditional maximum likelihood, and the pseudo-MLE estimator.

From Figure 3 we see that for $\tau_n = 10, 9.8$ (large), our proposed estimator outperforms both the inverse weighting estimator and the conditional maximum likelihood estimator. This verifies our statement in Remark 1. However, for $\tau_n = 6, 5$ (small), our proposed estimator is worse than the other two. Note that our estimator is not consistent when τ_n is small by Remark 1, so its underperformance is not surprising in this scenario.

Finally, we plot the efficiency cost in Figure 2b with the computational budget constraint according to our definition. The tradeoff is depicted numerically for the downsample rate selection as we discussed in Section 5. We have observed that when τ_n increases, the efficiency cost function becomes sharper, indicating that it is more sensitive to the choice of α. Although our theoretical finding shows that under the rare event case, a small choice of α such that $(1 - F(\tau_n))/\alpha = o(1)$ does not bring information loss while reduces computational cost massively, the efficiency cost can be very sensitive to α as the positive ratio goes to 0 according to Figure 2b, suggesting the necessity of a more prudent method of downsampling rate selection.

Dataset	Sample Size	Neg:Pos Ratio	Feature Number
abalone_19	4,177	130:1	10
mammography	11,183	42:1	6
yeast_me2	1,484	28:1	8
abalone	4,177	9.7:1	10
ecoli	336	8.6:1	7

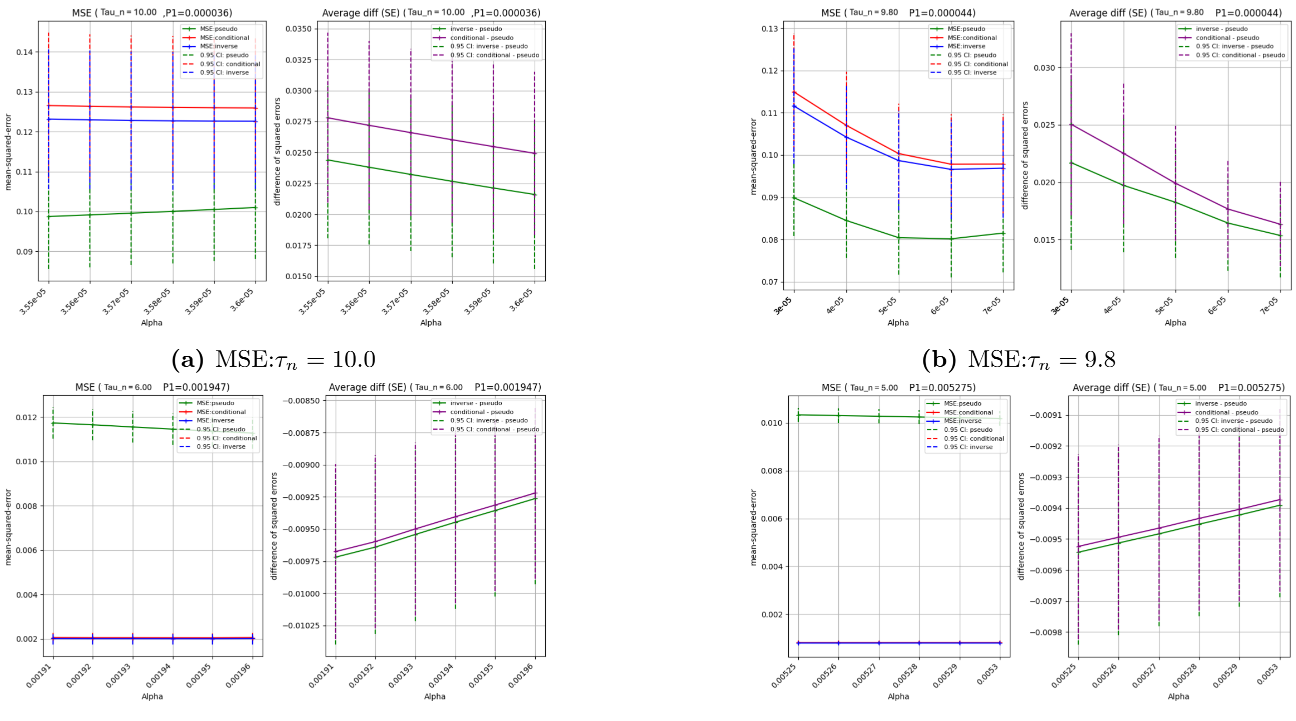
Table 1: Datasets from the UCI repository along of their size, negative/positive ratio, and the number of features.

7.2 Experiments on Benchmark Datasets

To verify the performance of the proposed pseudo-MLE on real imbalanced data, we compare the performance of our estimator with the inverse weighting estimator on UCI imbalanced datasets summarized in Table 1.

In order to adapt to our setup where we consider the regime with $\tau_n \to \infty$, we set τ_n such that $1/(1+e^{\tau_n}) = p_1$, where the values of p_1 are the positive ratio in the imbalanced dataset. Then we use the inverse weighting estimator and our pseudo-MLE estimator to fit θ_* (the coefficient for the features), and we compute the log loss for both estimators on the test set. We replicate the experiment 500 times, and during each round we randomly split the dataset into 80% for training and 20% for testing. We then plot the average log-losses and the confidence intervals for the log-losses for each downsampling rate α in Figure 4, where these α's are chosen close to p_1 (i.e., positive ratio) of each dataset. We refer the reader to Appendix A.1 for performance with additional moderate and small values of τ_n. We also apply our method to neural networks on some of these datasets. The experiment details and insights are presented in Appendix A.2.

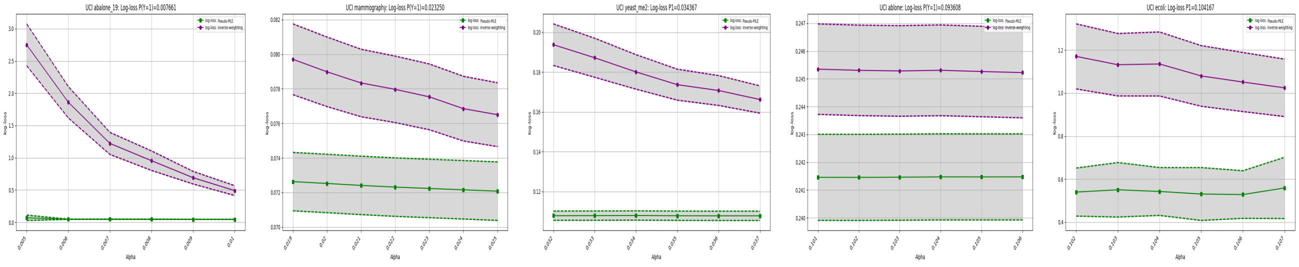
The numerical results suggest that the application of the pseudo-MLE gives lower test loss than the one of the commonly used inverse weighting estimators.



(a) MSE:$\tau_n = 10.0$ **(b)** MSE:$\tau_n = 9.8$

(c) MSE:$\tau_n = 6.0$ **(d)** MSE:$\tau_n = 5.0$

Figure 3: On the left panel of each figure, we plot MSE of inverse-weighting (blue) vs. pseudo-MLE (green) vs. conditional MLE (red) for α chosen around $\mathbb{P}(Y = 1)$ for $\tau_n = 10.0, 9.8, 6.0, 5.0$ with Logistic Regression. The dashed lines correspond to the 95% confidence intervals for the squared errors. The upper and lower ends are computed by $\pm 1.96 * \frac{\hat{\sigma}}{\sqrt{500}}$ and $\hat{\sigma}$ is the standard deviation of the squared errors at each α computed over 500 random environments. On the right panel of each figure the solid lines correspond to the difference of squared errors between inverse weighting and pseudo MLE (green) and conditional MLE and pseudo MLE (purple). The dashed lines are the 95% confidence intervals.



(a) UCI `abalone_19` **(b)** UCI `mammmography` **(c)** UCI `yeast_me2` **(d)** UCI `abalone` **(e)** UCI `ecoli`

Figure 4: Log loss of inverse-weighting (purple) vs. log-loss of pseudo-MLE (green) for α chosen around $\mathbb{P}(Y = 1)$ for each data set by applying Logistic Regression. We randomly split the original dataset into 80% for training and 20% for testing during each replication. The log-losses are all computed on test datasets. The dashed lines correspond to the 95% confidence intervals of the log loss. The upper and lower ends are computed by $\pm 1.96 * \frac{\hat{\sigma}}{\sqrt{500}}$ and $\hat{\sigma}$ is the standard deviation of log loss values at each α computed over 500 random environments.

8 DISCUSSION

We propose a pseudo maximum likelihood estimator for a Generalized Linear Model binary classifier under downsampling, with theoretical convergence guarantees for imbalanced data. We propose an efficiency cost notion to guide downsampling rate selection. For future work we are interested in investigating the performance under other performance metrics (e.g., AM-risk), exploring the oversampling strategy, and analyzing the optimization complexity of the log-likelihood function.

References

Anass Aghbalou, Anne Sabourin, and François Portier. Sharp error bounds for imbalanced classification: how many examples in the minority class? In *International Conference on Artificial Intelligence and Statistics*, pages 838–846. PMLR, 2024.

Kellyn F Arnold, Vinny Davies, Marc de Kamps, Peter WG Tennant, John Mbotwa, and Mark S Gilthorpe. Reflection on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *International journal of epidemiology*, 49 (6):2074–2082, 2020.

Shaik Johny Basha, Srinivasa Rao Madala, Kolla Vivek, Eedupalli Sai Kumar, and Tamminina Ammannamma. A review on imbalanced data classification techniques. In *2022 International conference on advanced computing technologies and applications (ICACTA)*, pages 1–6. IEEE, 2022.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Nitesh V Chawla. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pages 875–886, 2010.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6 (1):1–6, 2004.

Krzysztof Dembczyński, Wojciech Kotłowski, Oluwasanmi Koyejo, and Nagarajan Natarajan. Consistency analysis for binary classification revisited. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML '17, page 961–969. JMLR.org, 2017.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

Beiying Ding and Robert Gentleman. Classification using generalized partial least squares. *Journal of Computational and Graphical Statistics*, 14(2):280–298, 2005.

Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

Georgios Douzas and Fernando Bacao. Self-organizing map oversampling (somo) for imbalanced data set learning. *Expert systems with Applications*, 82:40–52, 2017.

Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, 2003.

Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997.

Charles Elkan. The foundations of cost-sensitive learning. JCAI'01, page 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608125.

Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20 (1):18–36, 2004.

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

Dylan J Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345. PMLR, 2019.

Peter W Glynn and Ward Whitt. The asymptotic efficiency of simulation estimators. *Operations research*, 40(3):505–520, 1992.

JM Gorriz, Carmen Jimenez-Mesa, Fermín Segovia, Javier Ramírez, SiPBA Group, and J Suckling. A connection between pattern classification by machine learning and statistical inference with the general linear model. *IEEE Journal of Biomedical and Health Informatics*, 26(11):5332–5343, 2021.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239, 2017.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

Amira Kamil Ibrahim Hassan and Ajith Abraham. Modeling insurance fraud detection using imbalanced data classification. In *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015*, pages 117–127. Springer, 2016.

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

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99. PMLR, 2021.

Shamsul Huda, John Yearwood, Herbert F Jelinek, Mohammad Mehedi Hassan, Giancarlo Fortino, and Michael Buckland. A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. *IEEE access*, 4:9145–9154, 2016.

Nathalie Japkowicz. *Learning from Inbalanced Data Sets: Papers from the AAAI Workshop*. AAAI Press, 2000.

Dan Jiang, Rongbin Xu, Xin Xu, and Ying Xie. Multi-view feature transfer for click-through rate prediction. *Information Sciences*, 546:961–976, 2021.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

Pawel Ksieniewicz. Undersampled majority class ensemble for highly imbalanced binary classification. In *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 82–94. PMLR, 2018.

Jungwon Lee, Okkyung Jung, Yunhye Lee, Ohsung Kim, and Cheol Park. A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(5): 1472–1491, 2021.

Kuang-chih Lee, Burkay Orten, Ali Dasdan, and Wentong Li. Estimating conversion rate in display advertising from past erformance data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 768–776, 2012.

Wonjae Lee and Kangwon Seo. Downsampling for binary classification with a highly imbalanced dataset using active learning. *Big Data Research*, 28:100314, 2022.

Guillaume LemaÃŽtre, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, 18 (17):1–5, 2017.

Jinyan Li, Simon Fong, Shimin Hu, Victor W Chu, Raymond K Wong, Sabah Mohammed, and Nilanjan Dey. Rare event prediction using similarity majority under-sampling technique. In *Soft Computing in Data Science: Third International Conference, SCDS 2017, Yogyakarta, Indonesia, November 27–28, 2017, Proceedings 3*, pages 23–39. Springer, 2017a.

Jinyan Li, Lian-sheng Liu, Simon Fong, Raymond K Wong, Sabah Mohammed, Jinan Fiaidhi, Yunsick Sung, and Kelvin KL Wong. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PloS one*, 12(7):e0180830, 2017b.

Josey Mathew, Chee Khiang Pang, Ming Luo, and Weng Hoe Leong. 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, 2017.

Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611. PMLR, 2013.

AS More and Dipti P Rana. Review of random forest classification techniques to resolve data imbalance. In *2017 1st International conference on intelligent systems and information management (ICISIM)*, pages 72–78. IEEE, 2017.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135 (3):370–384, 1972.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL `https://doi.org/10.1145/1102351.1102430`.

Art B Owen. Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, 8(4), 2007.

M Mostafizur Rahman and Darryl N Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224, 2013.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Talayeh Razzaghi, Oleg Roderick, Ilya Safro, and Nick Marko. Fast imbalanced classification of healthcare data with missing values. In *2015 18th International Conference on Information Fusion (Fusion)*, pages 774–781. IEEE, 2015.

Akshay Shah and Siddhesh Nasnodkar. The impacts of user experience metrics on click-through rate (ctr) in digital advertising: A machine learning approach. *Sage Science Review of Applied Machine Learning*, 4 (1):27–44, 2021.

Mayuri S Shelke, Prashant R Deshmukh, and Vijaya K Shandilya. A review on imbalanced data handling using undersampling and oversampling technique. *Int. J. Recent Trends Eng. Res*, 3(4):444–449, 2017.

Deepti Sisodia and Dilip Singh Sisodia. Data sampling strategies for click fraud detection using imbalanced user click data of online advertising: an empirical review. *IETE Technical Review*, 39(4):789–798, 2022.

Deepti Sisodia and Dilip Singh Sisodia. A hybrid data-level sampling approach in learning from skewed user-click data for click fraud detection in online advertising. *Expert Systems*, 40(2):e13147, 2023.

Adil Yaseen Taha, Sabrina Tiun, Abdul Hadi Abd Rahman, and Ali Sabah. Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification. *Journal of Information and Communication Technology*, 20(3):423–456, 2021.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

HaiYing Wang. Logistic regression for massive data with rare events. In *International Conference on Machine Learning*, pages 9829–9836. PMLR, 2020.

Xiaolin Wu, Xiangjun Zhang, and Xiaohan Wang. Low bit-rate image compression via adaptive down-sampling and constrained least squares upconversion. *IEEE Transactions on Image Processing*, 18(3):552–561, 2009.

Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International conference on machine learning*, pages 10544–10554. PMLR, 2020.

Yilin Yan, Min Chen, Mei-Ling Shyu, and Shu-Ching Chen. Deep learning for imbalanced multimedia data classification. In *2015 IEEE international symposium on multimedia (ISM)*, pages 483–488. IEEE, 2015.

Yuguang Yan, Mingkui Tan, Yanwu Xu, Jiezhang Cao, Michael Ng, Huaqing Min, and Qingyao Wu. Oversampling for imbalanced data via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5605–5612, 2019.

Gang Zhang, Lixin Wang, Alistair P Duffy, Hugh Sasse, Danilo Di Febo, Antonio Orlandi, and Karol Aniserowicz. Downsampled and undersampled datasets in feature selective validation (fsv). *IEEE transactions on electromagnetic compatibility*, 56(4):817–824, 2014.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] We provide the mathematical setting in the section of problem setup, assumptions and model in Section 3, etc.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We analyze the asymptotic normality of our estimator in Section 4,

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] We upload the code to a github link.

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes] We have stated the assumptions in section 4 or provide them as the conditions of the theorems.

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The code is uploaded to a Github repository and the hyperlink is available at the beginning of Appendix A. We also provide a hyperlink for the UCI dataset in Section 7.2.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] The information can be found in the subsection of numerical experiments and section 7.2.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] We provide that information in the beginning of Appendix A.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Yes]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] We have properly cited the sources of the lemmas in existing literature in supplementary.

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

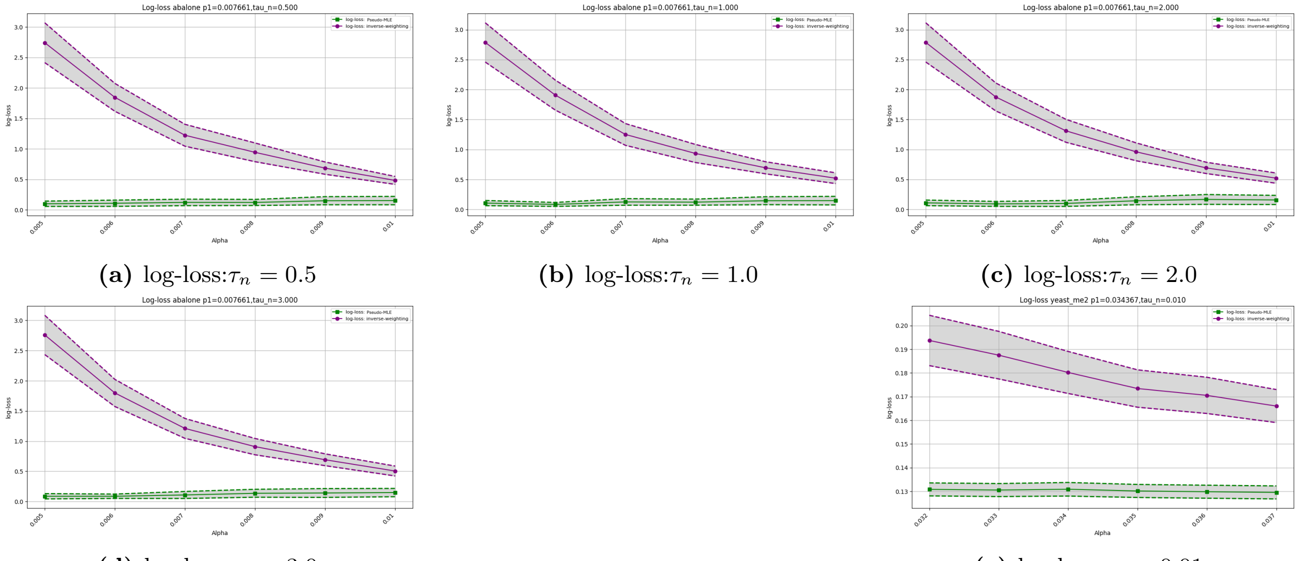
Supplementary Material

A ADDITIONAL SIMULATION DETAILS

The replication code for the simulations in the paper is provided at `https://github.com/Yan-Chen-Stats-ORer/Downsampling-GLM/tree/main`. We use a MacBook Pro with Apple M2 chip for all simulations.

A.1 Additional Results for Applying Logistic Regression

Varying values of τ_n for Logistic Regression on Empirical Data We run additional simulations on `abalone_19` and `yeast_me2` data for varying values of τ_n. The previous setting of `abalone_19` data is $\tau_n = \log(1/p_1 - 1) = 4.86$ and we plot the results with $\tau_n = 0.01, 0.5, 1.0, 2.0, 3.0$ in Figure 5. The previous setting of `yeast_me2` is $\tau_n = \log(1/p_1 - 1) = 3.34$ and we plot the results with $\tau_n = 0.01, 0.5, 1.0, 2.0, 2.5$ in Figure 6. The plots show that our pseudo MLE estimator outperforms the inverse-weighting estimator even for those moderate and small values of τ_n.



(a) log-loss:$\tau_n = 0.5$ **(b)** log-loss:$\tau_n = 1.0$ **(c)** log-loss:$\tau_n = 2.0$

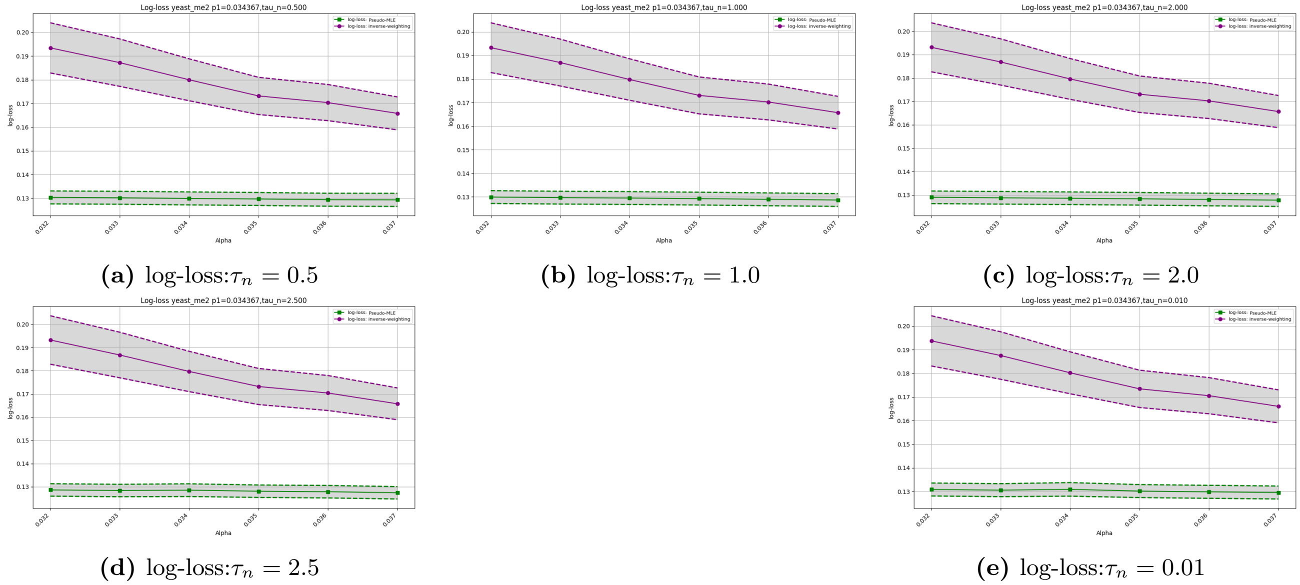
(d) log-loss:$\tau_n = 3.0$ **(e)** log-loss:$\tau_n = 0.01$

Figure 5: Additional results for `abalone_19` dataset for small and moderate values of τ_n.

A.2 Insights for Neural Networks

Neural Networks We plot the log losses of neural networks applied to imbalanced UCI real data (`yeat_me2`, `abalone_19`, `ecoli`) for different downsampling rates α in Figures 7a, 7b, 7c. Dashed lines correspond to the cross-entropy loss, solid lines correspond to the customized loss function implied by our estimator. The results are average log losses for each training epoch averaging across 500 random train/test splits of the real data. The solid lines are below the dashed lines within the same color, indicating the customized loss leading to better performance under the fixed small downsample rate α given here. The neural networks are trained with 3 dense layers with the ReLU activation and one outer layer with sigmoid output.

Though a maximum likelihood analysis of a neural network model is significantly more challenging than generalized linear model, the insight about downsampling alone is intuitive and should carry over to the case of neural networks in addition to GLMs. For example, we could use a small dataset to train a neural network, and exploit the fact that the output activation function is often a GLM. We can then apply our results to the GLM portion and use this to adjust the sample size according to our optimal sample selection.



(a) log-loss:$\tau_n = 0.5$

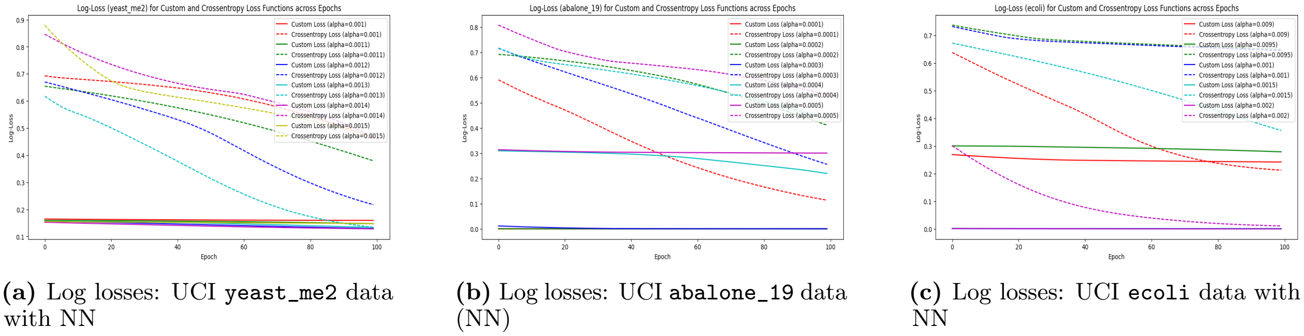
(b) log-loss:$\tau_n = 1.0$

(c) log-loss:$\tau_n = 2.0$

(d) log-loss:$\tau_n = 2.5$

(e) log-loss:$\tau_n = 0.01$

Figure 6: Additional results for `yeast_me2` dataset for small and moderate values of τ_n.



(a) Log losses: UCI `yeast_me2` data with NN

(b) Log losses: UCI `abalone_19` data (NN)

(c) Log losses: UCI `ecoli` data with NN

Figure 7: Log losses of Neural Networks trained on UCI datasets.

B ADDITIONAL LEMMAS AND PROOFS FOR THE PROPOSED ESTIMATOR

For the rest of the paper, we use $\tilde{P}$ to denote the joint distibution of downsample variables $(\tilde{X}, \tilde{Y})$, use $\tilde{E}$ to denote the expectation with respect to $\tilde{P}$, use $\tilde{P}_N$ for the empirical measure induced by $\{\tilde{X}_i, \tilde{Y}_i\}_{i=1}^N$, and $\tilde{E}_N$ for the expectation taken with respect to $\tilde{P}_N$.

Lemma 1 (Counterexample)**.** *Suppose $F(z)$ is strictly increasing, and suppose*

1) $\mathbb{E}_X\left[F^{(1)}(\tau_n + \theta_*^T X)X\right] \neq \mathbf{0}$*,*

2) there exists a unique $\tilde{\theta}_1 \in \Theta$ *such that* $\mathbb{E}_X\left[\frac{[1-(1-\alpha)F(\tau_n+\theta_*^T X)]F^{(1)}(\tau_n+\tilde{\theta}_1^T X)X}{1-(1-\alpha)F(\tau_n+\tilde{\theta}_1^T X)}\right] = \mathbf{0}$*,*

Then (1) leads to an inconsistent estimator for θ_*.

Furthermore, if $\{x \in \mathcal{X} | \tilde{\theta}_1^T x = 0\} \neq \emptyset$ and $\{x \in \mathcal{X} | \theta_*^T x = 0\} \cap \{x \in \mathcal{X} | \tilde{\theta}_1^T x = 0\} \notin \{\emptyset, \mathcal{X}\}$, then the prediction score obtained by the procedure described above (i.e. solving (1) and then applying isotonic regression) is also biased.

Proof of Lemma 1. The first-order condition for (1) is

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{Y}_i\frac{-F^{(1)}(\tau_n+\hat{\theta}_1^T\tilde{X}_i)\tilde{X}_i^T}{1-F(\tau_n+\hat{\theta}_1^T\tilde{X}_i)}+(1-\tilde{Y}_i)\frac{F^{(1)}(\tau_n+\hat{\theta}_1^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\hat{\theta}_1^T\tilde{X}_i)}=\mathbf{0},$$

i.e., $\hat{\theta}_1$ satisfies

$$\frac{1}{N}\sum_{i=1}^{N}\frac{(1-\tilde{Y}_i-F(\tau_n+\hat{\theta}_1^T\tilde{X}_i))F^{(1)}(\tau_n+\hat{\theta}_1^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\hat{\theta}_1^T\tilde{X}_i)(1-F(\tau_n+\hat{\theta}_1^T\tilde{X}_i))}=\mathbf{0}.$$

By Lemma 3 and the strong law of large numbers, for any $\theta_1\in\Theta$,

$$\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\tfrac{(1-\tilde{Y}_i-F(\tau_n+\theta_1^T\tilde{X}_i))F^{(1)}(\tau_n+\theta_1^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\theta_1^T\tilde{X}_i)(1-F(\tau_n+\theta_1^T\tilde{X}_i))}\\
&\xrightarrow{a.s.}\tilde{E}\left[\tfrac{(1-\tilde{Y}_i-F(\tau_n+\theta_1^T\tilde{X}_i))F^{(1)}(\tau_n+\theta_1^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\theta_1^T\tilde{X}_i)(1-F(\tau_n+\theta_1^T\tilde{X}_i))}\right]\\
&=\tilde{E}\left[\left[\tfrac{(1-\tilde{Y}_i-F(\tau_n+\theta_1^T\tilde{X}_i))F^{(1)}(\tau_n+\theta_1^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\theta_1^T\tilde{X}_i)(1-F(\tau_n+\theta_1^T\tilde{X}_i))}\middle|\tilde{X}_i\right]\right]\\
&=_{(1)}\tilde{E}\left[\tfrac{[G(\tau_n+\theta_1^T\tilde{X}_i)-F(\tau_n+\theta_1^T\tilde{X}_i)]F^{(1)}(\tau_n+\theta_1^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\theta_1^T\tilde{X}_i)(1-F(\tau_n+\theta_1^T\tilde{X}_i))}\right]\\
&=-(1-\alpha)\tilde{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_1^T\tilde{X}_i)\tilde{X}_i^T}{1-(1-\alpha)F(\tau_n+\theta_1^T\tilde{X}_i)}\right]\\
&=_{(2)}-(1-\alpha)\tfrac{\mathbb{E}_X\left[\frac{[1-(1-\alpha)F(\tau_n+\theta_*^TX)]F^{(1)}(\tau_n+\theta_1^TX)X}{1-(1-\alpha)F(\tau_n+\theta_1^TX)}\right]}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)},
\end{aligned}$$

where (1) follows from Proposition 2 and (2) uses Lemma 4. Thus by Lemma 7, $\hat{\theta}_1\xrightarrow{p}\tilde{\theta}_1$ such that $\mathbb{E}_X\left[\frac{[1-(1-\alpha)F(\tau_n+\theta_*^TX)]F^{(1)}(\tau_n+\tilde{\theta}_1^TX)X^T}{1-(1-\alpha)F(\tau_n+\tilde{\theta}_1^TX)}\right]=\mathbf{0}$.

Furthermore, taking in true parameter θ_*, the expected value under probability measure $\tilde{P}$ (i.e. the joint distribution of $(\tilde{Y},\tilde{X})$) is equal to

$$\begin{aligned}
&\tilde{E}\left[\tfrac{(1-\tilde{Y}_i-F(\tau_n+\theta_*^T\tilde{X}_i))F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\theta_*^T\tilde{X}_i)(1-F(\tau_n+\theta_*^T\tilde{X}_i))}\right]\\
&=\tilde{E}\left[\tilde{E}\left[\tfrac{(1-\tilde{Y}_i-F(\tau_n+\theta_*^T\tilde{X}_i))F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\theta_*^T\tilde{X}_i)(1-F(\tau_n+\theta_*^T\tilde{X}_i))}\middle|\tilde{X}_i\right]\right]\\
&=_{(1)}\tilde{E}\left[-\tfrac{F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)\bar{G}(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i^T}{1-F(\tau_n+\theta_*^T\tilde{X}_i)}+\tfrac{F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)G(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i^T}{F(\tau_n+\theta_*^T\tilde{X}_i)}\right]\\
&=_{(2)}\tfrac{-(1-\alpha)\mathbb{E}_X\left[F^{(1)}(\tau_n+\theta_*^TX)X^T\right]}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)},
\end{aligned}$$

where (1) uses Proposition 2 and (2) uses Lemma 4, and $\mathbb{E}_X$ is the expectation taken with respect to the distribution of X from the full data. Thus under the given conditions, $\mathbb{E}_X\left[F^{(1)}(\tau_n+\theta_*^TX)X^T\right]\neq\mathbf{0}$, while there exists some $\tilde{\theta}_1$,

$$\mathbb{E}_X\left[\tfrac{[1-(1-\alpha)F(\tau_n+\theta_*^TX)]F^{(1)}(\tau_n+\tilde{\theta}_1^TX)X^T}{1-(1-\alpha)F(\tau_n+\tilde{\theta}_1^TX)}\right]=\mathbb{E}_X\left[\tfrac{[1-(1-\alpha)F(\tau_n+\theta_*^TX)]\varphi(\tau_n+\tilde{\theta}_1^TX)X^T}{1-(1-\alpha)F(\tau_n+\tilde{\theta}_1^TX)}\right]=\mathbf{0}.$$

Thus $\hat{\theta}_1\xrightarrow{p}\tilde{\theta}_1$ but $\tilde{\theta}_1\neq\theta_*$.

Consequently, under the third condition in the lemma, there exists $x_1,x_2\in\mathcal{X}$ such that $\tilde{\theta}_1^T(x_1-x_2)=0$ while $\theta_*^T(x_1-x_2)\neq0$. This is because there exists $\Delta\in\mathcal{X}$ such that Δ is in the subspace defined by the hyperplane induced by $\tilde{\theta}_1$: $\Delta\in\{x\in\mathcal{X}|\tilde{\theta}_1^Tx=0\}$ while $\theta_*^T\Delta\neq0$. Then for any $x_1\in\mathcal{X}$, there exists λ sufficiently small such that $x_2=x_1+\lambda\Delta\in\mathcal{X}$, with $\tilde{\theta}_1^T(x_1-x_2)=0$ while $\theta_*^T(x_1-x_2)\neq0$. Therefore, $\tau_n+\tilde{\theta}_1^Tx_1=\tau_n+\tilde{\theta}_1^Tx_2$ while $\tau_n+\theta_*^Tx_1\neq\tau_n+\theta_*^Tx_2$. Thus suppose there exists a monotone transformation g such that $F(\tau_n+\theta_*^Tx)=g\circ F(\tau_n+\tilde{\theta}_1^Tx)$ for all $x\in\mathcal{X}$, then $F(\tau_n+\theta_*^Tx_1)=g\circ F(\tau_n+\tilde{\theta}_1^Tx_1)$. Also note that $\tau_n+\tilde{\theta}_1^Tx_1=\tau_n+\tilde{\theta}_1^Tx_2$ thus $g\circ F(\tau_n+\tilde{\theta}_1^Tx_1)=g\circ F(\tau_n+\tilde{\theta}_1^Tx_2)=F(\tau_n+\theta_*^Tx_2)$, thus $F(\tau_n+\theta_*^Tx_1)=F(\tau_n+\theta_*^Tx_2)$, which leads to contradiction because F is strictly increasing. □

Lemma 2. *Let $\mathbb{P}$ be the joint distribution of (Y,X), for any $y\in\{0,1\}$ and $x\in\mathcal{X}$, define*

$$\tilde{P}(\tilde{Y}=y,\tilde{X}=x):=\frac{y\mathbb{P}\left(Y=1,X=x\right)+(1-y)\alpha\mathbb{P}(Y=0,X=x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)},\tag{13}$$

then (13) defines a probability distribution $\tilde{P}$ with respect to downsample random variables $(\tilde{Y}_i, \tilde{X}_i)$.

Proof of Lemma 2. Note that the downsampling procedure is equivalent to

$$\left(\tilde{Y}, \tilde{X}\right) = \mathbf{1}\left(Y = 1\right)\left(Y, X\right) + \mathbf{1}(Y = 0)\mathbf{1}(U \leq \alpha)(Y, X), \tag{14}$$

where $U \sim$ Uniform $[0, 1]$and $U \perp\!\!\!\perp (Y, X)$. Thus the joint distribution (density) of $(\tilde{Y}, \tilde{X})$ with respect to $\mathbb{P}_{(Y,X)}$ as the joint law of (Y, X) can be written as

$$\begin{aligned}\mathbb{P}_{(Y,X)}\left(\left(\tilde{Y}, \tilde{X}\right) = (y, x)\right) &= \mathbf{1}(y = 1)\mathbb{P}\left(Y = 1, X = x\right) + \mathbf{1}(y = 0)\alpha\mathbb{P}(Y = 0, X = x)\\ &= y\mathbb{P}\left(Y = 1, X = x\right) + (1 - y)\alpha\mathbb{P}(Y = 0, X = x).\end{aligned} \tag{15}$$

So integrating with respect to (y, x) we have

$$\int_{\mathcal{X}} \mathbb{P}\left[(Y = 1, X = x) + \alpha\mathbb{P}(Y = 0, X = x)\right] dx = \mathbb{P}(Y = 1) + \alpha\mathbb{P}(Y = 0).$$

So by definition

$$\int_{\mathcal{X}} \sum_{y \in \{0,1\}} \tilde{P}(\tilde{Y} = y, \tilde{X} = x) dx = 1.$$

Further note that $\tilde{P}(\tilde{Y} = y, \tilde{X} = x) \in [0, 1]$ for any $x \in \mathcal{X}$ and $y \in \{0, 1\}$. Thus $\tilde{P}$ is indeed a valid probability distribution defined for $(\tilde{Y}, \tilde{X})$. □

Lemma 3 (i.i.d. Property)**.** *The downsampled data $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^N$ are i.i.d. generated with respect to $\tilde{P}$ defined in (13) of Lemma 2.*

Proof of Lemma 3. Recall that the *Generalized Linear Model* is defined as follows: For some latent variable Z, the label is defined as $Y = \mathbf{1}\left(Z > \tau_n + \theta_*^T X\right)$, and $\mathbb{P}\left(Y = 1 \middle| X = x\right) = \bar{F}_Z(\tau_n + \theta_*^T x)$. For $\forall i \in [N]$, let (Y_i, X_i) denote the full-sample random variable. Thus for $i, j \in [N]$, $i \neq j$, given any pairs of event (y, x), (y', x'), with $\mathbb{P}$ denoting the joint distribution of the full-sample random variables (Y_i, X_i), and U denote a uniform random variable on $[0, 1]$ such that $U \perp\!\!\!\perp \{(X_i, Y_i)\}_{i=1}^n$, by (15) we have

$$\begin{aligned}&\mathbb{P}\left(\left(\tilde{Y}_i, \tilde{X}_i\right) = (y, x), \left(\tilde{Y}_j, \tilde{X}_j\right) = (y', x')\right)\\ =_{(a)}\ &\mathbb{P}\left((Y_i, X_i) = (1, x), \left(\tilde{Y}_j, \tilde{X}_j\right) = (y', x')\right)\\ &+\mathbb{P}\left((Y_i, X_i) = (0, x), U \leq \alpha, \left(\tilde{Y}_j, \tilde{X}_j\right) = (y', x')\right)\\ =_{(b)}\ &\mathbb{P}\left((Y_i, X_i) = (1, x)\right)\mathbb{P}\left(\left(\tilde{Y}_j, \tilde{X}_j\right) = (y', x')\right)\\ &+\mathbb{P}\left((Y_i, X_i) = (0, x), U \leq \alpha\right)\mathbb{P}\left(\left(\tilde{Y}_j, \tilde{X}_j\right) = (y', x')\right)\\ =_{(c)}\ &\mathbb{P}\left(\left(\tilde{Y}_i, \tilde{X}_i\right) = (y, x)\right)\mathbb{P}\left(\left(\tilde{Y}_j, \tilde{X}_j\right) = (y', x')\right),\end{aligned}$$

where (a) uses the fact that $\mathbf{1}((Y_i, X_i) = (y, x), Y_i = 1)$ and $\mathbf{1}((Y_i, X_i) = (y, x), Y_i = 0, U \leq \alpha)$ are disjoint events, and (b) uses the fact that $(\tilde{Y}_j, \tilde{X}_j) = \mathcal{L}(Y_j, X_j, U)$ as some joint law of (Y_j, X_j, U), which is independent of (Y_i, X_i). Moreover, (c) uses the fact that $\mathbb{P}\left(\left(\tilde{Y}_i, \tilde{X}_i\right) = (y, x)\right) = \mathbb{P}\left((Y_i, X_i) = (1, x)\right) + \mathbb{P}\left((Y_i, X_i) = (0, x), U \leq \alpha\right)$ according to (14). Thus $(\tilde{Y}_i, \tilde{X}_i)$ and $(\tilde{Y}_j, \tilde{X}_j)$ are independent with respect to $\mathbb{P}$.

Note that $\tilde{P} = \mathbb{P}/(\mathbb{P}(Y = 1) + \alpha\mathbb{P}(Y = 0))$, where given τ_n, θ_*, $\mathbb{P}(Y = 1) + \alpha\mathbb{P}(Y = 0)$ is a constant, thus $(\tilde{Y}_i, \tilde{X}_i)$ and $(\tilde{Y}_j, \tilde{X}_j)$ are independent with respect to $\tilde{P}$. Obviously $(\tilde{Y}_j, \tilde{X}_j)$ are identically generated. So the result follows. □

Proof of Proposition 1. From (14) we know that for $y \in \{0, 1\}$, $x \in \mathcal{X}$, and $\mathbb{P}$ as the joint law of (Y, X), we have

$$\begin{aligned}\mathbb{P}\left(\left(\tilde{Y}, \tilde{X}\right) \in (y, x)\right) &= \mathbb{P}\left((Y, X) \in (y, x), Y = 1\right) + \mathbb{P}\left((Y, X) = (y, x), Y = 0, U \leq \alpha\right)\\ &= \mathbf{1}(y = 1)\mathbb{P}(Y = 1, X = x) + \mathbf{1}(y = 0)\alpha\mathbb{P}(Y = 0, X = x)\end{aligned}$$

Note that

$$\begin{aligned}
\tilde{P}\left(\tilde{Y}=1\middle|\tilde{X}=x\right) &= \tfrac{\tilde{P}(\tilde{Y}=1,\tilde{X}=x)}{\tilde{P}(\tilde{X}=x)} \\
&= \tfrac{\mathbb{P}(\tilde{Y}=1,\tilde{X}=x)/(\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0))}{\mathbb{P}(\tilde{Y}=1,\tilde{X}=x)/(\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0))+\mathbb{P}(\tilde{Y}=0,\tilde{X}=x)/(\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0))} \\
&= \tfrac{\mathbb{P}(\tilde{Y}=1,\tilde{X}=x)}{\mathbb{P}(\tilde{Y}=1,\tilde{X}=x)+\mathbb{P}(\tilde{Y}=0,\tilde{X}=x)} = \mathbb{P}(\tilde{Y}=1|\tilde{X}=x) \\
&= \tfrac{\mathbb{P}(Y=1,X=x)}{\mathbb{P}(Y=1,X=x)+\alpha\mathbb{P}(Y=0,X=x)} \\
&= \tfrac{\mathbb{P}(Y=1|X=x)\mathbb{P}(X=x)}{\mathbb{P}(Y=1|X=x)\mathbb{P}(X=x)+\mathbb{P}(Y=0|X=x)\mathbb{P}(X=x)} \\
&= \tfrac{\mathbb{P}(Y=1\middle|X=x)}{\mathbb{P}(Y=1\middle|X=x)+\alpha\mathbb{P}(Y=0\middle|X=x)}.
\end{aligned}$$

When $\alpha=1$, $\mathbb{P}\left(\tilde{Y}=1\middle|\tilde{X}=x\right)=\mathbb{P}(Y=1\middle|X=x)$. Note that $\mathbb{P}\left(Y=1\middle|X=x\right)=\bar{F}_Z(\tau_n+\theta_*^Tx)=1-F_Z(\tau_n+\theta_*^Tx)$, then $\mathbb{P}\left(\tilde{Y}=1\middle|\tilde{X}=x\right)=\frac{\bar{F}_Z(\tau_n+\theta_*^Tx)}{\bar{F}_Z(\tau_n+\theta_*^Tx)+(1-\bar{F}_Z(\tau_n+\theta_*^Tx))\alpha}=\frac{\bar{F}_Z(\tau_n+\theta_*^Tx)}{(1-\alpha)\bar{F}_Z(\tau_n+\theta_*^Tx)+\alpha}$. Let $\bar{G}(z)=\frac{\bar{F}_Z(z)}{\bar{F}_Z(z)(1-\alpha)+\alpha}$, then $\mathbb{P}(\tilde{Y}=1|\tilde{X}=x)=\bar{G}(\tau_n+\theta_*^Tx)$. Note $\bar{G}(\infty)=0$, $\bar{G}^{(1)}(z)<0$, $\bar{G}(-\infty)=1$, thus there exists some W, such that $\bar{G}(z)=\bar{F}_W(z)$, where $\bar{F}_W(z)=1-F_W(z)$, and $F_W(\cdot)$ is the c.d.f. of W, and $\bar{G}(z)=\frac{1}{(1-\alpha)+\alpha/\bar{F}_Z}\iff\bar{F}_Z=\frac{\alpha\bar{G}}{1-(1-\alpha)\bar{G}}$. □

Lemma 4 (Distribution of $\tilde{X}_i$ with respect to $\tilde{P}$)**.** *The density function of $\tilde{X}_i$ at $\tilde{X}_i=x$ is $\tilde{\mu}(x)=\frac{[1-(1-\alpha)F(\tau_n+\theta_*^Tx)]\mu(x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}$.*

Proof of Lemma 4. From (14) and previous proofs, with $\mathbb{P}$ denoting the joint law of (Y,X), we have

$$\begin{aligned}
&\tfrac{\bar{F}(\tau_n+\theta_*^Tx)\mu(x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)} = \tfrac{\mathbb{P}(Y=1,X=x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)} \\
&= \tfrac{\mathbb{P}\left(\tilde{Y}=1,\tilde{X}=x\right)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)} = \tilde{P}\left(\tilde{Y}=1,\tilde{X}=x\right) = \bar{G}(\tau_n+\theta_*^Tx)\tilde{\mu}(x),
\end{aligned}$$

$$\begin{aligned}
&\tfrac{\alpha F(\tau_n+\theta_*^Tx)\mu(x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)} = \tfrac{\alpha\mathbb{P}(Y=0,X=x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)} \\
&= \tfrac{\mathbb{P}\left(\tilde{Y}=0,\tilde{X}=x\right)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)} = \tilde{P}\left(\tilde{Y}=0,\tilde{X}=x\right) = G(\tau_n+\theta_*^Tx)\tilde{\mu}(x).
\end{aligned}$$

so the density function of $\tilde{X}_i$ wtih respect to $\tilde{P}$ at $\tilde{X}_i=x$ is equal to $\tilde{\mu}(x)=\frac{[1-(1-\alpha)F(\tau_n+\theta_*^Tx)]\mu(x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}$.

□

Proof of Proposition 2. Note that by definition the joint distribution (density) of $(\tilde{Y},\tilde{X})$ with respect to $\tilde{P}$ can be written as

$$\begin{aligned}
&\tilde{P}\left(\tilde{Y}=y,\tilde{X}=x\right) \\
&= y\tilde{P}\left(\tilde{Y}=1|\tilde{X}=x\right)\tilde{\mu}(x)+(1-y)\tilde{P}\left(\tilde{Y}=0|\tilde{X}=x\right)\tilde{\mu}(x) \\
&= y\bar{G}(\tau_n+\theta_*^Tx)\tilde{\mu}(x)+(1-y)G(\tau_n+\theta_*^Tx)\tilde{\mu}(x) \\
&= \left(\bar{F}(\tau_n+\theta_*^Tx)\tfrac{\mu(x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}\right)^y\left(\alpha F(\tau_n+\theta_*^Tx)\tfrac{\mu(x)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}\right)^{1-y},
\end{aligned}$$

where the third equality in the above uses Lemma 4. Let $\tilde{E}[\cdot]$ denote the expectation taken with respect to $\tilde{P}$, then we have

$$\begin{aligned}
&M(\theta_1;\tau_n) := \tilde{E}\left[\log\tilde{P}\left(\tilde{Y}_i,\tilde{X}_i\right)\right] \\
&= \tilde{E}\Bigg[\tilde{Y}_i\left(\log\left(\bar{F}(\tau_n+\theta_1^T\tilde{X}_i)\right)+\log\tfrac{\mu(\tilde{X}_i)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}\right) \\
&\qquad +(1-\tilde{Y}_i)\left(\log\left(\alpha F(\tau_n+\theta_1^T\tilde{X}_i)\right)+\log\tfrac{\mu(\tilde{X}_i)}{\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)}\right)\Bigg] \\
&= \tilde{E}\left[\tilde{Y}_i\log\bar{F}(\tau_n+\theta_1^T\tilde{X}_i)+(1-\tilde{Y}_i)\log\alpha F(\tau_n+\theta_1^T\tilde{X}_i)+\log\mu(\tilde{X}_i)\right] \\
&\qquad -\log\left[\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)\right] \\
&= \tilde{E}\left[\tilde{Y}_i\log\bar{F}(\tau_n+\theta_1^T\tilde{X}_i)+(1-\tilde{Y}_i)\log\alpha F(\tau_n+\theta_1^T\tilde{X}_i)+\log\mu(\tilde{X}_i)\right] \\
&\qquad -\log\int_{\mathcal{X}}\left[1-(1-\alpha)F(\tau_n+\theta_1^Tx)\right]\mu(x)dx.
\end{aligned}$$

We then have $\theta_* \in \operatorname{argmax}_{\theta_1 \in \Theta} M(\theta_1; \tau_n)$, and note from Lemma 3 $(\tilde{X}_i, \tilde{Y}_i)$ are i.i.d. with respect to $\tilde{P}$, so by strong law of large numbers,

$$\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{i=1}^{N} \tilde{Y}_i \log \bar{F}(\tau_n + \theta_1^T \tilde{X}_i) + (1-\tilde{Y}_i)\log \alpha F(\tau_n + \theta_1^T \tilde{X}_i) + \log \mu(\tilde{X}_i)\\
&\quad - \log \textstyle\int_{\mathcal{X}} \left[1 - (1-\alpha)F(\tau_n + \theta_1^T x)\right] \mu(x) dx\\
&\xrightarrow{p} \tilde{E}\left[\tilde{Y}_i \log \bar{F}(\tau_n + \theta_1^T \tilde{X}_i) + (1-\tilde{Y}_i)\log \alpha F(\tau_n + \theta_1^T \tilde{X}_i) + \log \mu(\tilde{X}_i)\right]\\
&\quad - \log \textstyle\int_{\mathcal{X}} \left[1 - (1-\alpha)F(\tau_n + \theta_1^T x)\right] \mu(x) dx.
\end{aligned}$$

Further note that given the down-sample $(\tilde{X}_i, \tilde{Y}_i)$, $\frac{1}{N}\sum_{i=1}^{N} \log \mu(\tilde{X}_i)$ doesn't depend on τ_n or θ_1, thus the maximum likelihood estimator can be defined as

$$\hat{\theta}_* = \operatorname{argmax}_{\theta_1} \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \ell\left(\tilde{X}_i, \tilde{Y}_i, \theta_1; \tau_n\right),$$

where

$$\ell(\tilde{x}, \tilde{y}, \theta_1; \tau_n) = \tilde{y} \log \bar{F}(\tau_n + \theta_1^T \tilde{x}) + (1-\tilde{y})\log \alpha F(\tau_n + \theta_1^T \tilde{x}) - \log \int_{\mathcal{X}} \left[1 - (1-\alpha)F(\tau_n + \theta_1^T x)\mu(x)\right] dx.$$

Thus the result follows. □

C PROOFS FOR ASYMPTOTIC ANALYSIS

Discussion on the failure of classical MLE analysis Note that

$$\begin{aligned}
\mathbb{E}\left[\tilde{\ell}(\tilde{X}, \tilde{Y}, \theta_*; \tau_n)\right] &= \mathbb{E}_{(\tilde{Y},\tilde{X})}\left[\tilde{Y} \log \bar{F}(\tau_n + \theta_*^T \tilde{X}) + (1-\tilde{Y})\log \alpha F(\tau_n + \theta_*^T \tilde{X})\right]\\
&\quad - \mathbb{E}_{(\tilde{Y},\tilde{X})}\{\log \tfrac{1}{N}\textstyle\sum_{i=1}^{N}[1-(1-\alpha)F(\tau_n + \theta_1^T \tilde{X}_i)]\}
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{\tilde{X}}\left[\bar{G}(\tau_n + \theta_*^T \tilde{X}) \log \bar{F}(\tau_n + \theta_*^T \tilde{X}) + G(\tau_n + \theta_*^T \tilde{X})\log \alpha F(\tau_n + \theta_*^T \tilde{X})\right]\\
&\quad - \mathbb{E}_{(\tilde{Y},\tilde{X})}\{\log \tfrac{1}{N}\textstyle\sum_{i=1}^{N}[1-(1-\alpha)F(\tau_n + \theta_1^T \tilde{X}_i)]\}
\end{aligned}$$

$$\begin{aligned}
&= \tfrac{\mathbb{E}_X\left[(1-F(\tau_n+\theta_*^T X))\log\left(1-F(\tau_n+\theta_1^T X)\right)+\alpha F(\tau_n+\theta_*^T X)\log \alpha F(\tau_n+\theta_1^T X)\right]}{\mathbb{P}(Y=1)+\alpha \mathbb{P}(Y=0)}\\
&\quad - \mathbb{E}_{(\tilde{Y},\tilde{X})}\{\log \tfrac{1}{N}\textstyle\sum_{i=1}^{N}[1-(1-\alpha)F(\tau_n + \theta_1^T \tilde{X}_i)]\}
\end{aligned}$$

Note that $F(\tau_n + \theta_1^T x) \to 1$ as $n \to \infty$ for any $x \in \mathcal{X}$ and $\theta_1 \in \Theta$. Since $x \log x \to 0$ as $x \to 0$, so $\mathbb{E}\left[\tilde{\ell}(\tilde{X}, \tilde{Y}, \theta_*; \tau_n)\right] \to \log(\alpha) - \log(\alpha) = 0$ regardless of the value of θ_* if $\alpha > 0$. Thus as $n \to \infty$, the criterion function doesn't rely on the value of θ_1 in the varying-rate regime, thus the classical MLE theory cannot be applied here.

C.1 Proof of Theorem 1

Proof of Theorem 1. Given τ_n, note that

$$\begin{aligned}
\hat{\theta}_* &= \operatorname{argmax}_{\theta_1 \in \Theta} \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \tilde{Y}_i \log \bar{F}(\tau_n + \theta_1^T \tilde{X}_i) + (1-\tilde{Y}_i)\log \alpha F(\tau_n + \theta_1^T \tilde{X}_i)\\
&\qquad - \log\left\{\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\left[1-(1-\alpha)F(\tau_n + \theta_1^T \tilde{X}_i)\right]\right\}
\end{aligned}$$

Denote

$$\begin{aligned}
&L_n(\theta_1) \triangleq L(\theta_1; \tau_n)\\
&\triangleq \textstyle\sum_{i=1}^{N}\left[\tilde{Y}_i \log \bar{F}(\tau_n + \theta_1^T \tilde{X}_i) + (1-\tilde{Y}_i)\log \alpha F(\tau_n + \theta_1^T \tilde{X}_i)\right]\\
&\quad - \log\{\tfrac{1}{N}\textstyle\sum_{i=1}^{N}[1-(1-\alpha)F(\tau_n + \theta_1^T \tilde{X}_i)]\}\\
&= \textstyle\sum_{i=1}^{N}\left[\tilde{Y}_i \log \frac{\bar{F}(\tau_n+\theta_1^T \tilde{X}_i)}{\{\frac{1}{N}\sum_{i=1}^{N}[1-(1-\alpha)F(\tau_n+\theta_1^T \tilde{X}_i)]\}} + (1-\tilde{Y}_i)\log \frac{\alpha F(\tau_n+\theta_1^T \tilde{X}_i)}{\{\frac{1}{N}\sum_{i=1}^{N}[1-(1-\alpha)F(\tau_n+\theta_1^T \tilde{X}_i)]\}}\right]\\
&= \textstyle\sum_{i=1}^{n}\left[Y_i \log \frac{\bar{F}(\tau_n+\theta_1^T X_i)}{\{\frac{1}{N}\sum_{i=1}^{N}[1-(1-\alpha)F(\tau_n+\theta_1^T \tilde{X}_i)]\}}\right.\\
&\qquad \left. + (1-Y_i)\mathbf{1}(U_i \le \alpha)\log \frac{\alpha F(\tau_n+\theta_1^T X_i)}{\{\frac{1}{N}\sum_{i=1}^{N}[1-(1-\alpha)F(\tau_n+\theta_1^T \tilde{X}_i)]\}}\right],
\end{aligned}$$

where $\{U_i\}_{i=1}^n$ is i.i.d. uniform random variable on $[0,1]$, and $U_i \perp\!\!\!\perp \{Y_i, X_i\}_{i=1}^n$. Let $a_n = \sqrt{n(1-F(\tau_n))}$. So $w_n = a_n(\hat{\theta}_* - \theta_*)$ is the maximizer of

$$H(w) := L_n(\theta_* + a_n^{-1} w) - L_n(\theta_*).$$

For $w = a_n(\hat{\theta} - \theta_*)$, define $g(t) = L_n(\theta_* + t(\hat{\theta} - \theta_*))$ for $t \in [0,1]$. By Taylor expansion, for some $\gamma \in (0,1)$, we have $g(1) - g(0) = g^{(1)}(0) + \frac{1}{2} g^{(2)}(\gamma)$, i.e.

$$\begin{aligned} H(w) &= \nabla_{\theta_1} L_n(\theta_*)^T (a_n^{-1} w) + \tfrac{1}{2}(\hat{\theta} - \theta_*)^T \nabla_{\theta_1}^2 L_n(\theta_* + \gamma(\hat{\theta} - \theta_*))(\hat{\theta} - \theta_*) \\ &= (a_n^{-1} w^T) \nabla_{\theta_1} L_n(\theta_*) + \tfrac{1}{2} a_n^{-2} w^T \nabla_{\theta_1}^2 L_n(\theta_* + \gamma(\hat{\theta} - \theta_*)) w. \end{aligned}$$

Denote $J_N(\theta_1) := \frac{1}{N}\sum_{i=1}^N [1 - (1-\alpha)F(\tau_n + \theta_1^T \tilde{X}_i)]$, and

$$\begin{aligned} \ell(x, y, u, \theta_1; \tau_n) &= y \log \tfrac{\bar{F}(\tau_n + \theta_1^T x)}{\{\frac{1}{N}\sum_{i=1}^N [1-(1-\alpha)F(\tau_n + \theta_1^T \tilde{X}_i)]\}} \\ &\quad + (1-y)\mathbf{1}(u \le \alpha) \log \tfrac{\alpha F(\tau_n + \theta_1^T x)}{\{\frac{1}{N}\sum_{i=1}^N [1-(1-\alpha)F(\tau_n + \theta_1^T \tilde{X}_i)]\}} \\ &= y \log \tfrac{\bar{F}(\tau_n + \theta_1^T x)}{J_N(\theta_1)} + (1-y)\mathbf{1}(u \le \alpha) \log \tfrac{\alpha F(\tau_n + \theta_1^T x)}{J_N(\theta_1)}. \end{aligned}$$

Then $\nabla_{\theta_1} J_N(\theta_1) = -(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i^T$, and

$$\begin{aligned} &\nabla_{\theta_1} \ell(x, y, u, \theta_1; \tau_n) \\ &= y\left[-\tfrac{F^{(1)}(\tau_n + \theta_1^T x) x^T}{1 - F(\tau_n + \theta_1^T x)} + \tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_1)}\right] \\ &\quad + (1-y)\mathbf{1}(u \le \alpha)\left[\tfrac{F^{(1)}(\tau_n + \theta_1^T x) x^T}{F(\tau_n + \theta_1^T x)} + \tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_1)}\right] \\ &= F^{(1)}(\tau_n + \theta_1^T x) x^T \left[-\tfrac{y}{1 - F(\tau_n + \theta_1^T x)} + \mathbf{1}(u \le \alpha)\tfrac{1-y}{F(\tau_n + \theta_1^T x)}\right] \\ &\quad + (y + (1-y)\mathbf{1}(u \le \alpha)) \tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_1)} \end{aligned}$$

and

$$\begin{aligned} &\nabla_{\theta_1} L_n(\theta_1) \\ &= \textstyle\sum_{i=1}^n F^{(1)}(\tau_n + \theta_1^T X_i) X_i \left[-\tfrac{Y_i}{1 - F(\tau_n + \theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n + \theta_1^T X_i)}\right] \\ &\quad + (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha)) \tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_1)}. \end{aligned}$$

Hence

$$\begin{aligned} &\nabla_{\theta_1}^2 L_n(\theta_1) = \textstyle\sum_{i=1}^n \nabla_{\theta_1}^2 \ell(X_i, Y_i, U_i, \theta_1; \tau_n) \\ &= \textstyle\sum_{i=1}^n F^{(2)}(\tau_n + \theta_1^T X_i) X_i X_i^T \left[-\tfrac{Y_i}{1 - F(\tau_n + \theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n + \theta_1^T X_i)}\right] \\ &\quad + \textstyle\sum_{i=1}^n F^{(1)}(\tau_n + \theta_1^T X_i)^2 X_i X_i^T \left[\tfrac{-Y_i}{(1 - F(\tau_n + \theta_1^T X_i))^2} - \tfrac{\mathbf{1}(U_i \le \alpha)(1-Y_i)}{F(\tau_n + \theta_1^T X_i)^2}\right] \\ &\quad + \textstyle\sum_{i=1}^n (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha)) \left[\tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(2)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i \tilde{X}_i^T}{J_N(\theta_1)}\right] \\ &\quad + \textstyle\sum_{i=1}^n (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha)) \left[\tfrac{(1-\alpha)^2 \{\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i\}\{\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i^T\}}{J_N(\theta_1)^2}\right], \end{aligned}$$

and $\quad H(w) = (a_n^{-1} w^T)\nabla_{\theta_1} L_n(\theta_*) + \frac{1}{2} a_n^{-2} \sum_{i=1}^n \Phi_i(\theta_* + \gamma a_n^{-1} w), \quad$ where

$$\begin{aligned} &\Phi_i(\theta_1) \\ &= F^{(2)}(\tau_n + \theta_1^T X_i) X_i X_i^T \left[-\tfrac{Y_i}{1 - F(\tau_n + \theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n + \theta_1^T X_i)}\right] \\ &\quad + F^{(1)}(\tau_n + \theta_1^T X_i)^2 X_i X_i^T \left[-\tfrac{Y_i}{(1 - F(\tau_n + \theta_1^T X_i))^2} - \tfrac{\mathbf{1}(U_i \le \alpha)(1-Y_i)}{F(\tau_n + \theta_1^T X_i)^2}\right] \\ &\quad + (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha)) \left[\tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(2)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i \tilde{X}_i^T}{J_N(\theta_1)}\right] \\ &\quad + (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha)) \left[\tfrac{(1-\alpha)^2 \{\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i\}\{\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n + \theta_1^T \tilde{X}_i)\tilde{X}_i^T\}}{J_N(\theta_1)^2}\right] \end{aligned}$$

In the following, we want to show that for some matrices $\mathbf{V}$ and $\tilde{\mathbf{V}}_\Phi$, we have

$$a_n^{-1} \nabla_{\theta_1} L_n(\theta_*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}),$$

and for any u and $\gamma \in [0,1]$, $a_n^{-2}\sum_{i=1}^n \Phi_i(\theta_* + \gamma a_n^{-1} w) \xrightarrow{p} \tilde{\mathbf{V}}_\Phi$. Note that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\Big[F^{(1)}(\tau_n + \theta_*^T X_i)X_i \left[-\tfrac{Y_i}{1-F(\tau_n+\theta_*^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_*^T X_i)}\right] \\
&\quad + (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha))\,\tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n+\theta_*^T \tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_*)}\Big] \\
&= \lim_{n\to\infty} \mathbb{E}\left[F^{(1)}(\tau_n + \theta_*^T X_i)X_i \left(-\tfrac{Y_i}{1-F(\tau_n+\theta_*^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_*^T X_i)}\right)\right] \\
&\quad +(1-\alpha)[\mathbb{P}(Y=1) + \alpha\mathbb{P}(Y=0)]\,\tfrac{\tilde{E}[F^{(1)}(\tau_n+\theta_*^T \tilde{X})\tilde{X}]}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]} \\
&=_{(1)} \lim_{n\to\infty} (\alpha-1)\,\mathbb{E}\left[F^{(1)}(\tau_n+\theta_*^T X_i)X_i\right] \\
&\qquad + \tfrac{(1-\alpha)[\mathbb{P}(Y=1)+\alpha\mathbb{P}(Y=0)]\mathbb{E}[F^{(1)}(\tau_n+\theta_*^T X_i)(1-(1-\alpha)F(\tau_n+\theta_*^T X_i))X_i]}{\mathbb{E}[(1-(1-\alpha)F(\tau_n+\theta_*^T X))^2]} \\
&= \mathbf{0},
\end{aligned}$$

where we use Lemma 4 in (1), so $\mathbb{E}\left[a_n^{-1}\nabla_{\theta_1} L_n(\theta_*)\right] \to \mathbf{0}$. Furthermore, by letting $\theta_1 = \theta_*$ in the below, we have

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{Cov}\left[a_n^{-1}\nabla_{\theta_1} L_n(\theta_*)\right] \\
&= \lim_{n\to\infty} a_n^{-2}\sum_{i=1}^n \mathrm{Cov}\Big[F^{(1)}(\tau_n + \theta_1^T X_i)X_i \left[-\tfrac{Y_i}{1-F(\tau_n+\theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_1^T X_i)}\right] \\
&\quad + (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha))\,\tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n+\theta_*^T \tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_*)}\Big] \\
&= \lim_{n\to\infty} \sum_{i=1}^n \mathrm{Cov}\Big[\tfrac{F^{(1)}(\tau_n+\theta_1^T X_i)}{n(1-F(\tau_n))}X_i \left[-\tfrac{Y_i}{1-F(\tau_n+\theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_1^T X_i)}\right] \\
&\quad + (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha))\,\tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n+\theta_*^T \tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_*)}\Big] \\
&= \lim_{n\to\infty} \mathrm{Cov}\Big[\tfrac{F^{(1)}(\tau_n+\theta_1^T X_i)}{\sqrt{1-F(\tau_n)}}X_i \left[-\tfrac{Y_i}{1-F(\tau_n+\theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_1^T X_i)}\right] \\
&\quad + (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha))\,\tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n+\theta_*^T \tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_*)\sqrt{1-F(\tau_n)}}\Big] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_1^T X_i)^2}{1-F(\tau_n)} \left[-\tfrac{Y_i}{1-F(\tau_n+\theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_1^T X_i)}\right]^2 X_i X_i^T\right] \\
&\quad +\mathbb{E}\left[(Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha))^2\right]\,\tfrac{(1-\alpha)^2\tilde{E}[F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i]\tilde{E}[F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i^T]}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]^2(1-F(\tau_n))} \\
&\quad +2\mathbb{E}\Big[\tfrac{F^{(1)}(\tau_n+\theta_1^T X_i)X_i}{\sqrt{1-F(\tau_n)}} \left[-\tfrac{Y_i}{1-F(\tau_n+\theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_1^T X_i)}\right] \\
&\qquad \times \tfrac{(Y_i+(1-Y_i)\mathbf{1}(U_i\le\alpha))}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]}(1-\alpha)\tilde{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_1^T\tilde{X})\tilde{X}^T}{\sqrt{1-F(\tau_n)}}\right]\Big] \\
&=_{(g)} \lim_{n\to\infty} \mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^T X_i)^2}{1-F(\tau_n)} \left[\tfrac{1}{1-F(\tau_n+\theta_*^T X_i)} + \tfrac{\alpha}{F(\tau_n+\theta_*^T X_i)}\right] X_i X_i^T\right] \\
&\quad + \tfrac{(1-\alpha)^2\tilde{E}\left[\frac{F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)}{\sqrt{1-F(\tau_n)}}\tilde{X}_i\right]\tilde{E}\left[\frac{F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)}{\sqrt{1-F(\tau_n)}}\tilde{X}_i^T\right](\mathbb{P}(Y_i=1)+\alpha\mathbb{P}(Y_i=0))}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]^2} \\
&\quad +2\mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^T X_i)}{\sqrt{1-F(\tau_n)}}\,\tfrac{\mathbf{1}(U_i\le\alpha)-1}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]}(1-\alpha)\tilde{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i}{\sqrt{1-F(\tau_n)}}\right]X_i^T\right] \\
&= \lim_{n\to\infty} \mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^T X_i)^2}{1-F(\tau_n)} \left[\tfrac{1}{1-F(\tau_n+\theta_*^T X_i)} + \tfrac{\alpha}{F(\tau_n+\theta_*^T X_i)}\right] X_i X_i^T\right] \\
&\quad + \tfrac{(1-\alpha)^2\tilde{E}\left[\frac{F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)}{\sqrt{1-F(\tau_n)}}\tilde{X}_i\right]\tilde{E}\left[\frac{F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)}{\sqrt{1-F(\tau_n)}}\tilde{X}_i^T\right](\mathbb{P}(Y_i=1)+\alpha\mathbb{P}(Y_i=0))}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]^2} \\
&\quad -2\tfrac{(1-\alpha)^2}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]}\mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^T X_i)}{\sqrt{1-F(\tau_n)}}X_i\right]\tilde{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i^T}{\sqrt{1-F(\tau_n)}}\right] \\
&= \mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^T X_i)^2}{1-F(\tau_n)} \left[\tfrac{1}{1-F(\tau_n+\theta_*^T X_i)} + \tfrac{\alpha}{F(\tau_n+\theta_*^T X_i)}\right] X_i X_i^T\right] \\
&\quad -\tfrac{(1-\alpha)^2}{\alpha}\mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^T X_i)}{1-F(\tau_n)}X_i\right]\mathbb{E}\left[F^{(1)}(\tau_n+\theta_*^T X_i)X_i^T\right],
\end{aligned}$$

where equation (g) uses the fact that $\mathbb{E}[(Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha))^2] = \mathbb{P}(Y=1) + \alpha\mathbb{P}(Y=0)$. Thus by dominated convergence theorem and Assumption 3,

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{Cov}\left[a_n^{-1}\nabla_{\theta_1} L_n(\theta_*)\right] \\
&= \mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X) X X^T\right] - \mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\mathbb{E}\left[\lim_{n\to\infty}\tfrac{(1-\alpha)^2 F^{(1)}(\tau_n+\theta_*^T X)X^T}{\alpha}\right].
\end{aligned}$$

Furthermore, we check the Lindeberg-Feller CLT condition for the asymptotic normality result. Recall that

$$\begin{aligned}
&\nabla_{\theta_1}\ell(x,y,u,\theta_*;\tau_n) \\
&= F^{(1)}(\tau_n+\theta_1^T x)x\left[-\tfrac{y}{1-F(\tau_n+\theta_1^T x)} + \mathbf{1}(u\le\alpha)\tfrac{1-y}{F(\tau_n+\theta_1^T x)}\right] \\
&\quad + (y + (1-y)\mathbf{1}(u\le\alpha))\,\tfrac{(1-\alpha)\frac{1}{N}\sum_{i=1}^N F^{(1)}(\tau_n+\theta_1^T\tilde{X}_i)\tilde{X}_i^T}{J_N(\theta_1)}.
\end{aligned}$$

For $\epsilon > 0$, let $\mathcal{A}_i$ denote the event that $\|\nabla_{\theta_1}\ell(X_i, Y_i, U_i, \theta_*; \tau_n)\| > a_n\epsilon$, we then have

$$\begin{aligned}
&\textstyle\sum_{i=1}^n \mathbb{E}\left[\|\nabla_{\theta_1}\ell(X_i, Y_i, U_i, \theta_*; \tau_n)\|^2 \mathbf{1}\left(\|\nabla_{\theta_1}\ell(X_i, Y_i, U_i, \theta_*; \tau_n)\| > a_n\epsilon\right)\right] \\
&= n\mathbb{E}\left[\|\nabla_{\theta_1}\ell(X_i, Y_i, U_i, \theta_*; \tau_n)\|^2 \mathbf{1}\left(\|\nabla_{\theta_1}\ell(X_i, Y_i, U_i, \theta_*; \tau_n)\| > a_n\epsilon\right)\right] \\
&\leq_{(1)} 2Cn\mathbb{E}\left[\left(\tfrac{1-Y_i}{F(\tau_n+\theta_*^T X_i)}\mathbf{1}(U_i \leq \alpha) - \tfrac{Y_i}{1-F(\tau_n+\theta_*^T X_i)}\right)^2 F^{(1)}(\tau_n + \theta_*^T X_i)^2 \|X_i\|^2 \mathbf{1}_{\mathcal{A}_i}\right] \\
&\quad + 2Cn\mathbb{E}\left[\tfrac{(Y_i+(1-Y_i)\mathbf{1}(U_i\leq\alpha))^2(1-\alpha)^2}{J_N(\theta_*)^2}\left\|\tilde{E}[F^{(1)}(\tau_n + \theta_*^T \tilde{X})\tilde{X}]\right\|^2 \mathbf{1}_{\mathcal{A}_i}\right] \\
&= 2n\mathbb{E}\left[\mathbb{E}\left[\left(\tfrac{1-Y_i}{F(\tau_n+\theta_*^T X_i)}\mathbf{1}(U_i \leq \alpha) - \tfrac{Y_i}{1-F(\tau_n+\theta_*^T X_i)}\right)^2 F^{(1)}(\tau_n + \theta_*^T X_i)^2 \|X_i\|^2 \mathbf{1}_{\mathcal{A}_i} \middle| X_i\right]\right] \\
&\quad + 2Cn\mathbb{E}\left[\mathbb{E}\left[\tfrac{(Y_i+(1-Y_i)\mathbf{1}(U_i\leq\alpha))^2(1-\alpha)^2}{J_N(\theta_*)^2}\left\|\tilde{E}[F^{(1)}(\tau_n + \theta_*^T \tilde{X})\tilde{X}]\right\|^2 \mathbf{1}_{\mathcal{A}_i} \middle| X_i\right]\right] \\
&= 2n\mathbb{E}\left[\left(\tfrac{\mathbf{1}(U_i\leq\alpha)}{F(\tau_n+\theta_*^T X_i)} + \tfrac{1}{1-F(\tau_n+\theta_*^T X_i)}\right) F^{(1)}(\tau_n + \theta_*^T X_i)^2 \|X_i\|^2 \mathbf{1}_{\mathcal{A}_i}\right] \\
&\quad + 2Cn\tfrac{(1-\alpha)^2\left\|\tilde{E}[F^{(1)}(\tau_n+\theta_*^T \tilde{X})\tilde{X}]\right\|^2}{J_N(\theta_*)^2}\mathbb{E}\left[(1 - F(\tau_n + \theta_*^T X_i) + \mathbf{1}(U_i \leq \alpha)F(\tau_n + \theta_*^T X_i))\mathbf{1}_{\mathcal{A}_i}\right] \\
&\leq 2n(1 - F(\tau_n)) \\
&\quad \times\mathbb{E}\Bigg[\left(\tfrac{\mathbf{1}(U_i\leq\alpha)}{F(\tau_n+\theta_*^T X_i)} + \tfrac{1}{1-F(\tau_n+\theta_*^T X_i)}\right) \tfrac{F^{(1)}(\tau_n+\theta_*^T X_i)^2}{1-F(\tau_n)} \|X_i\|^2 \\
&\qquad \times\mathbf{1}\left\{\tfrac{(1-Y_i)\mathbf{1}(U_i\leq\alpha)F^{(1)}(\tau_n+\theta_*^T X_i)\|X_i\|}{F(\tau_n+\theta_*^T X_i)(1-F(\tau_n))} > \tfrac{n\epsilon}{2}\right\}\Bigg] \\
&\quad + 2Cn(1 - F(\tau_n))\frac{(1-\alpha)^2\left\|\tilde{E}\left[\frac{F^{(1)}(\tau_n+\theta_*^T \tilde{X})}{1-F(\tau_n)}\tilde{X}\right]\right\|^2}{J_N(\theta_*)^2} \\
&\qquad \times\mathbb{E}\left[(1 - F(\tau_n + \theta_*^T X_i) + \mathbf{1}(U_i \leq \alpha)F(\tau_n + \theta_*^T X_i))\mathbf{1}_{\mathcal{A}_i}\right] \\
&\overset{(*)}{=} \mathrm{o}(n(1 - F(\tau_n))) = \mathrm{o}(a_n^2),
\end{aligned}$$

where (1) is because $\mathbb{E}[\|A + B\|^2] \leq 2\mathbb{E}[\|A\|^2 + \|B\|^2]$, and (*) uses dominated convergence theorem and Assumption 3, and C is some absolute constant.

For the rest of the proof, we denote

$$g_1(\cdot) := \frac{F^{(1)}}{1 - F(\tau_n)}\frac{1 - F(\tau_n)}{1 - F} = \frac{-h^{(1)}(\cdot)}{h},$$

$$g_2(\cdot) := \frac{F^{(2)}}{1 - F(\tau_n)}\frac{1 - F(\tau_n)}{1 - F} = \frac{-h^{(2)}}{h},$$

$$g_3(\cdot) := \frac{F^{(3)}}{1 - F(\tau_n)}\frac{1 - F(\tau_n)}{1 - F} = \frac{-h^{(3)}}{h}.$$

Thus applying the Lindeberg-Feller central limit theorem as Proposition 2.27 from (Van der Vaart, 2000) (i.e. Lemma 5), we have

$$a_n^{-1}\nabla_{\theta_1}L_n(\theta_*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}\right),$$

where

$$\begin{aligned}
\mathbf{V} \quad &= \mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X)XX^T\right] - \mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\mathbb{E}\left[\lim_{n\to\infty}\tfrac{(1-\alpha)^2 F^{(1)}(\tau_n+\theta_*^T X)X^T}{\alpha}\right] \\
&= \mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X)XX^T\right] \\
&\quad - \left(\lim_{n\to\infty}\tfrac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\right)\mathbb{E}[h(\theta_*^T X)g_1(\theta_*^T X)X]\mathbb{E}\left[h(\theta_*^T X)g_1(\theta_*^T X)X^T\right]
\end{aligned}$$

Now we want to show that for any u and any $\gamma \in [0, 1]$, for some matrix $\tilde{\mathbf{V}}_\Phi$ we have

$$a_n^{-2}\textstyle\sum_{i=1}^n \Phi_i(\theta_* + \gamma a_n^{-1}w) \xrightarrow{p} \tilde{\mathbf{V}}_\Phi.$$

Note that

$$\begin{aligned}
&\lim_{n\to\infty} a_n^{-2}\textstyle\sum_{i=1}^n \Phi_i(\theta_*) \\
&= \lim_{n\to\infty}\tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{F^{(2)}(\tau_n+\theta_*^T X_i)}{1-F(\tau_n)}X_i X_i^T\left[-\tfrac{Y_i}{1-F(\tau_n+\theta_*^T X_i)} + \mathbf{1}(U_i \leq \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_*^T X_i)}\right] \\
&\quad + \tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{F^{(1)}(\tau_n+\theta_*^T X_i)^2}{1-F(\tau_n)}X_i X_i^T\left[-\tfrac{Y_i}{(1-F(\tau_n+\theta_*^T X_i))^2} - \tfrac{\mathbf{1}(U_i\leq\alpha)(1-Y_i)}{F(\tau_n+\theta_*^T X_i)^2}\right] \\
&\quad + \tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{(Y_i+(1-Y_i)\mathbf{1}(U_i\leq\alpha))}{1-F(\tau_n)}\left[\tfrac{(1-\alpha)\tilde{E}[F^{(2)}(\tau_n+\theta_*^T \tilde{X}_i)\tilde{X}_i\tilde{X}_i^T]}{J_N(\theta_*)}\right] \\
&\quad + \tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{(1-\alpha)^2\tilde{E}[F^{(1)}(\tau_n+\theta_*^T \tilde{X})\tilde{X}]\tilde{E}[F^{(1)}(\tau_n+\theta_*^T \tilde{X})\tilde{X}^T]}{J_N(\theta_*)^2},
\end{aligned}$$

and note that by dominated convergence theorem and Assumption 3,

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{E}\left[\tfrac{F^{(2)}(\tau_n+\theta_*^TX_i)}{1-F(\tau_n)}X_iX_i^T\left(-\tfrac{Y_i}{1-F(\tau_n+\theta_*^TX_i)}+\mathbf{1}(U_i\le\alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_*^TX_i)}\right)\right]\\ &=(\alpha-1)\mathbb{E}\left[h(\theta_*^TX)g_2(\theta_*^TX)XX^T\right],\end{aligned}$$

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+\theta_*^TX_i)^2}{1-F(\tau_n)}X_iX_i^T\left[-\tfrac{Y_i}{(1-F(\tau_n+\theta_*^TX_i))^2}-\tfrac{\mathbf{1}(U_i\le\alpha)(1-Y_i)}{F(\tau_n+\theta_*^TX_i)^2}\right]\right]\\ &=-\mathbb{E}\left[g_1(\theta_*^TX)^2h(\theta_*^TX)XX^T\right],\end{aligned}$$

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{E}\Bigg[\tfrac{(Y_i+(1-Y_i)\mathbf{1}(U_i\le\alpha))}{1-F(\tau_n)}\Bigg[\tfrac{(1-\alpha)\tilde{E}[F^{(2)}(\tau_n+\theta_*^T\tilde{X}_i)\tilde{X}_i\tilde{X}_i^T]}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]}\\ &\qquad+\tfrac{(1-\alpha)^2\tilde{E}[F^{(1)}(\tau_n+\theta_*^T\tilde{X})\tilde{X}]\tilde{E}[F^{(1)}(\tau_n+\theta_*^T\tilde{X})\tilde{X}^T]}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_*^T\tilde{X})]^2}\Bigg]\Bigg]\\ &=(1-\alpha)\mathbb{E}[g_2(\theta_*^TX)h(\theta_*^TX)XX^T]\\ &\quad+\mathbb{E}[h(\theta_*^TX)g_1(\theta_*^TX)X]\mathbb{E}\left[\lim_{n\to\infty}\tfrac{(1-\alpha)^2F^{(1)}(\tau_n+\theta_*^TX)}{\alpha}X^T\right]\\ &=(1-\alpha)\mathbb{E}[g_2(\theta_*^TX)h(\theta_*^TX)XX^T]\\ &\quad+\left(\lim_{n\to\infty}\tfrac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\right)\mathbb{E}[h(\theta_*^TX)g_1(\theta_*^TX)X]\mathbb{E}\left[h(\theta_*^TX)g_1(\theta_*^TX)X^T\right]\end{aligned}$$

Thus

$$\begin{aligned}&a_n^{-2}\textstyle\sum_{i=1}^n\Phi_i(\theta_*)\\ &\xrightarrow{p}-\mathbb{E}\left[g_1(\theta_*^TX)^2h(\theta_*^TX)XX^T\right]\\ &\quad+\left(\lim_{n\to\infty}\tfrac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\right)\mathbb{E}[h(\theta_*^TX)g_1(\theta_*^TX)X]\mathbb{E}\left[h(\theta_*^TX)g_1(\theta_*^TX)X^T\right].\end{aligned}$$

For the last step of the proof, we want to show that indeed

$$\begin{aligned}\tilde{\mathbf{V}}_\Phi&=-\mathbb{E}\left[g_1(\theta_*^TX)^2h(\theta_*^TX)XX^T\right]\\ &\quad+\left(\lim_{n\to\infty}\tfrac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\right)\mathbb{E}[h(\theta_*^TX)g_1(\theta_*^TX)X]\mathbb{E}\left[h(\theta_*^TX)g_1(\theta_*^TX)X^T\right].\end{aligned}$$

Note that

$$\begin{aligned}&\nabla_{\theta_1}\Phi_i(\theta_1)\\ &=F^{(3)}(\tau_n+\theta_1^TX_i)X_i^TX_iX_i^T\left[-\tfrac{Y_i}{1-F(\tau_n+\theta_1^TX_i)}+\mathbf{1}(U_i\le\alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_1^TX_i)}\right]\\ &\quad+F^{(2)}(\tau_n+\theta_1^TX_i)F^{(1)}(\tau_n+\theta_1^TX_i)X_i^TX_iX_i^T\left[-\tfrac{Y_i}{(1-F(\tau_n+\theta_1^TX_i))^2}-\tfrac{\mathbf{1}(U_i\le\alpha)(1-Y_i)}{F(\tau_n+\theta_1^TX_i)^2}\right]\\ &\quad+2F^{(1)}(\tau_n+\theta_1^TX_i)F^{(2)}(\tau_n+\theta_1^TX_i)X_i^TX_iX_i^T\left[-\tfrac{Y_i}{(1-F(\tau_n+\theta_1^TX_i))^2}-\tfrac{\mathbf{1}(U_i\le\alpha)(1-Y_i)}{F(\tau_n+\theta_1^TX_i)^2}\right]\\ &\quad+F^{(1)}(\tau_n+\theta_1^TX_i)^2X_i^TX_iX_i^T\Big[-2Y_i(1-F(\tau_n+\theta_1^TX_i))^{-3}F^{(1)}(\tau_n+\theta_1^TX_i)\\ &\qquad\qquad+2\cdot\mathbf{1}(U_i\le\alpha)(1-Y_i)F(\tau_n+\theta_1^TX_i)^{-3}F^{(1)}(\tau_n+\theta_1^TX_i)\Big]\\ &\quad+(Y_i+(1-Y_i)\mathbf{1}(U_i\le\alpha))\left[\tfrac{(1-\alpha)\tilde{E}_N\left[F^{(3)}(\tau_n+\theta_1^T\tilde{X}_i)\tilde{X}_i^T\tilde{X}_i\tilde{X}_i^T\right]}{\tilde{E}_N[1-(1-\alpha)F(\tau_n+\theta_1^T\tilde{X}_i)]}\right]\\ &\quad+(Y_i+(1-Y_i)\mathbf{1}(U_i\le\alpha))\left[2(1-\alpha)^2\tfrac{\tilde{E}[F^{(2)}(\tau_n+\theta_1^T\tilde{X})\tilde{X}^T\tilde{X}]\tilde{E}[F^{(1)}(\tau_n+\theta_1^T\tilde{X})\tilde{X}^T]}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_1^T\tilde{X})]^2}\right],\end{aligned}$$

hence

$$\begin{aligned}&\left|a_n^{-2}\textstyle\sum_{i=1}^n\left\|\Phi_i(\theta_*+\gamma a_n^{-1}w)\right\|-a_n^{-2}\textstyle\sum_{i=1}^n\left\|\Phi_i(\theta_*)\right\|\right|\\ &\le_{(d)}a_n^{-2}\left\|a_n^{-1}w\right\|\textstyle\sum_{i=1}^n\left|\Delta_i(\theta_*+\tilde{\gamma}a_n^{-1}w)\right|\left\|X_i\right\|^3+a_n^{-2}\left\|a_n^{-1}w\right\|\textstyle\sum_{i=1}^n\left|\tilde{\Delta}_i(\theta_*+\tilde{\gamma}a_n^{-1}w)\right|\\ &=\tfrac{\left\|a_n^{-1}w\right\|}{n}\textstyle\sum_{i=1}^n\tfrac{\left|\Delta_i(\theta_*+\tilde{\gamma}a_n^{-1}w)\right|}{1-F(\theta_n)}\left\|X_i\right\|^3+\tfrac{\left\|a_n^{-1}w\right\|}{n}\textstyle\sum_{i=1}^n\tfrac{\left|\tilde{\Delta}_i(\theta_*+\tilde{\gamma}a_n^{-1}w)\right|}{1-F(\theta_n)},\end{aligned}$$

where $\tilde{\gamma}$ is some constant such that $\tilde{\gamma}\in(0,1)$, inequality (d) uses mean value theorem and Cauchy-Schwarz inequality and the fact that $\gamma \in (0,1)$, and

$$\begin{aligned}&\Delta_i(\theta_* + \tilde{\gamma} a_n^{-1} w)\\ &= F^{(3)}(\tau_n + \theta_1^T X_i)\left[-\tfrac{Y_i}{1-F(\tau_n+\theta_1^T X_i)} + \mathbf{1}(U_i \le \alpha)\tfrac{1-Y_i}{F(\tau_n+\theta_1^T X_i)}\right]\\ &\quad +F^{(2)}(\tau_n + \theta_1^T X_i)F^{(1)}(\tau_n + \theta_1^T X_i)\left[-\tfrac{Y_i}{(1-F(\tau_n+\theta_1^T X_i))^2} - \tfrac{\mathbf{1}(U_i\le\alpha)(1-Y_i)}{F(\tau_n+\theta_1^T X_i)^2}\right]\\ &\quad +2F^{(1)}(\tau_n + \theta_1^T X_i)F^{(2)}(\tau_n + \theta_1^T X_i)\left[-\tfrac{Y_i}{(1-F(\tau_n+\theta_1^T X_i))^2} - \tfrac{\mathbf{1}(U_i\le\alpha)(1-Y_i)}{F(\tau_n+\theta_1^T X_i)^2}\right]\\ &\quad +F^{(1)}(\tau_n + \theta_1^T X_i)^2\Big[-2Y_i(1-F(\tau_n+\theta_1^T X_i))^{-3}F^{(1)}(\tau_n+\theta_1^T X_i)\\ &\qquad\qquad +2\cdot\mathbf{1}(U_i \le \alpha)(1-Y_i)F(\tau_n+\theta_1^T X_i)^{-3}F^{(1)}(\tau_n+\theta_1^T X_i)\Big],\end{aligned}$$

$$\begin{aligned}&\tilde{\Delta}_i(\theta_* + \tilde{\gamma} a_n^{-1} w)\\ &= (Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha))\left[\tfrac{(1-\alpha)\tilde{E}_N\left[F^{(3)}(\tau_n+\theta_1^T\tilde{X}_i)\tilde{X}_i^T\tilde{X}_i\tilde{X}_i^T\right]}{\tilde{E}_N[1-(1-\alpha)F(\tau_n+\theta_1^T\tilde{X}_i)]}\right]\\ &\quad +(Y_i + (1-Y_i)\mathbf{1}(U_i \le \alpha))\left[2(1-\alpha)^2\tfrac{\tilde{E}[F^{(2)}(\tau_n+\theta_1^T\tilde{X})\tilde{X}^T\tilde{X}]\tilde{E}[F^{(1)}(\tau_n+\theta_1^T\tilde{X})\tilde{X}^T]}{\tilde{E}[1-(1-\alpha)F(\tau_n+\theta_1^T\tilde{X})]^2}\right].\end{aligned}$$

Again, similar to the previous argument, by Assumption 3 and dominated convergence theorem, the terms $\lim_{n\to\infty}\mathbb{E}\left[\frac{|\Delta_i(\theta_*+\tilde{\gamma}a_n^{-1}w)|}{1-F(\theta_n)}\|X_i\|^3\right], \lim_{n\to\infty}\mathbb{E}\left[\frac{|\tilde{\Delta}_i(\theta_*+\tilde{\gamma}a_n^{-1}w)|}{1-F(\theta_n)}\right]$ are bounded, so

$$\left|a_n^{-2}\sum_{i=1}^n \left\|\Phi_i(\theta_* + \gamma a_n^{-1} w)\right\| - a_n^{-2}\sum_{i=1}^n \|\Phi_i(\theta_*)\|\right| = \mathrm{o}_P(1).$$

Combining with the previous arguments, we have proved that

$$\begin{aligned}\tilde{\mathbf{V}}_\Phi &= -\mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X) X X^T\right]\\ &\quad + \left(\lim_{n\to\infty}\tfrac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\right)\mathbb{E}[h(\theta_*^T X)g_1(\theta_*^T X)X]\mathbb{E}\left[h(\theta_*^T X)g_1(\theta_*^T X)X^T\right].\end{aligned}$$

Recall that $a_n(\hat{\theta}_* - \theta_*)$ is the maximizer of

$$H(w) = (a_n^{-1}w^T)\nabla_{\theta_1}L_n(\theta_*) + \frac{1}{2}a_n^{-2}w^T\nabla_{\theta_1}^2 L_n(\theta_* + \gamma(\hat{\theta}-\theta_*))w,$$

which is equivalently the minimizer of $-\frac{1}{2}a_n^{-2}w^T\nabla_{\theta_1}^2 L_n(\theta_* + \gamma(\hat{\theta}-\theta_*))u - (a_n^{-1}w^T)\nabla_{\theta_1}L_n(\theta_*)$. Then by the Basic Corollary of (Hjort and Pollard, 2011) (or Lemma 6), we have

$$a_n(\hat{\theta} - \theta_*) = -\tilde{\mathbf{V}}_\Phi^{-1} \times a_n^{-1}\nabla_{\theta_1}L_n(\theta_*) + \mathrm{o}_P(1) = \mathbf{V}_\Phi^{-1} \times a_n^{-1}\nabla_{\theta_1}L_n(\theta_*) + \mathrm{o}_P(1),$$

where

$$\begin{aligned}\mathbf{V}_\Phi &= \mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X) X X^T\right]\\ &\quad - \left(\lim_{n\to\infty}\tfrac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\right)\mathbb{E}[h(\theta_*^T X)g_1(\theta_*^T X)X]\mathbb{E}\left[h(\theta_*^T X)g_1(\theta_*^T X)X^T\right].\end{aligned}$$

By the given condition that $\lim_{n\to\infty}\frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha} = c$, we have

$$\begin{aligned}\mathbf{V} = \mathbf{V}_\Phi &= \mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X) X X^T\right] - c\mathbb{E}[h(\theta_*^T X)g_1(\theta_*^T X)X]\mathbb{E}\left[h(\theta_*^T X)g_1(\theta_*^T X)X^T\right]\\ &= \mathbb{E}\left[\tfrac{h^{(1)}(\theta_*^T X)^2}{h(\theta_*^T X)}XX^T\right] - c\mathbb{E}\left[h^{(1)}(\theta_*^T X)X\right]\mathbb{E}\left[h^{(1)}(\theta_*^T X)X^T\right]\end{aligned}$$

Thus we have $\sqrt{n(1-F(\tau_n))}(\hat{\theta}_* - \theta_*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}_\Phi^{-1}\mathbf{V}\mathbf{V}_\Phi^{-1}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{-1}\right)$. □

C.2 Proof for Theorem 2

Proof of Theorem 2. By definition, the (scaled) maximum likelihood estimator is now defined as

$$\begin{aligned}r(\tau_n)\hat{\theta}_* &= \operatorname{argmax}_{\theta_1\in\Theta}\tfrac{1}{N}\textstyle\sum_{i=1}^N \tilde{Y}_i\log\bar{F}(\tau_n + \theta_1^T\tilde{X}_i) + (1-\tilde{Y}_i)\log\alpha F(\tau_n + \theta_1^T\tilde{X}_i)\\ &\qquad\qquad - \log\{\tfrac{1}{N}\textstyle\sum_{i=1}^N[1-(1-\alpha)F(\tau_n + \theta_1^T\tilde{X}_i)]\},\end{aligned}$$

then following similar steps as in the proof of Theorem 1, we use the same definition of $L_n(\theta_1)$ and a_n, and we can get

$$a_n^{-1}\nabla_{\theta_1}L_n(r(\tau_n)\theta_*) \xrightarrow{d} \mathcal{N}(0,\mathbf{V}),$$

where

$$\begin{aligned}\mathbf{V} &= \lim_{n\to\infty}\mathbb{E}\left[\tfrac{F^{(1)}(\tau_n+r(\tau_n)\theta_*^TX_i)^2}{1-F(\tau_n)}\left[\tfrac{1}{1-F(\tau_n+r(\tau_n)\theta_*^TX_i)}+\tfrac{\alpha}{F(\tau_n+r(\tau_n)\theta_*^TX_i)}\right]X_iX_i^T\right]\\ &-\lim_{n\to\infty}\frac{(1-\alpha)^2\mathbb{E}\left[\frac{F^{(1)}(\tau_n+r(\tau_n)\theta_*^TX_i)}{1-F(\tau_n)}X_i\right]\mathbb{E}\left[F^{(1)}(\tau_n+r(\tau_n)\theta_*^TX_i)X_i^T\right]}{\tilde{E}[1-(1-\alpha)F(\tau_n+r(\tau_n)\theta_*^T\tilde{X}_i)]}.\end{aligned}$$

Note that by Assumption 4 for any $\theta_1\in\Theta$ and $x\in\mathcal{X}$, we have

$$g(\theta^Tx)=\frac{1-F(r(\tau_n)(\theta^Tx)+\tau_n)}{1-F(\tau_n)},$$

thus $\frac{F^{(1)}(\tau_n+r(\tau_n)\theta_*^Tx)}{1-F(\tau_n+r(\tau_n)\theta_*^Tx)}=\frac{\frac{F^{(1)}(\tau_n+r(\tau_n)\theta_*^Tx)}{1-F(\tau_n)}}{\frac{1-F(\tau_n+r(\tau_n)\theta_*^Tx)}{1-F(\tau_n)}}=-\frac{g^{(1)}(\theta_*^Tx)}{g(\theta_*^Tx)}$. So by dominated convergence theorem, we have

$$\mathbf{V}=\mathbb{E}\left[\frac{g^{(1)}(\theta_*^TX)^2}{g(\theta_*^TX)}XX^T\right]-\lim_{n\to\infty}\frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\mathbb{E}\left[g^{(1)}(\theta_*^TX)X\right]\mathbb{E}\left[g^{(1)}(\theta_*^TX)X^T\right].$$

Similarly, we can also get

$$a_n^{-2}\textstyle\sum_{i=1}^n\Phi_i(r(\tau_n)(\theta_*+\gamma a_n^{-1}w)) \xrightarrow{p} -\mathbf{V}_\Phi,$$

where

$$\mathbf{V}_\Phi=\mathbb{E}\left[\frac{g^{(1)}(\theta_*^TX)^2}{g(\theta_*^TX)}XX^T\right]-\lim_{n\to\infty}\frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\mathbb{E}\left[g^{(1)}(\theta_*^TX)X\right]\mathbb{E}\left[g^{(1)}(\theta_*^TX)X^T\right].$$

Then the rest of the proof follows by similar steps of checking regularity conditions, etc. as in Theorem 1. □

C.3 Technical Lemmas

Lemma 5 (Proposition 2.27 from (Van der Vaart, 2000))**.** *For each n let $Y_{n,1},\ldots,Y_{n,k_n}$ be independent random vectors with finite variances such that for every $\epsilon>0$, $\sum_{i=1}^{k_n}\mathbb{E}[\|Y_{n,i}\|^2]\mathbf{1}\{\|Y_{n,i}\|>\epsilon\}\to 0$, and $\sum_{i=1}^{k_n}Y_{n,i}\to\Sigma$, then the sequence $\sum_{i=1}^{k_n}(Y_{n,i}-\mathbb{E}[Y_{n,i}])$ converges in distribution to $\mathcal{N}(\mathbf{0},\Sigma)$.*

Lemma 6 (Basic Corollary of (Hjort and Pollard, 2011))**.** *Let $A_n(s)=\frac{1}{2}s^TVs+U_n^Ts+C_n+r_n(s)$, where V is symmetric and postive definite, U_n is stochastically bounded, C_n is arbitrary, and $r_n(s)$ goes to zero in probability for every s. Then $\alpha_n=\operatorname{argmin}A_n$ is $o_p(1)$ away from $-V^{-1}U_n$ as the argmin of $\frac{1}{2}s^TVs+U_n^Ts+C_n$. If also $U_n\xrightarrow{d}U$ then $\alpha_n\xrightarrow{d}-V^{-1}U$.*

Lemma 7 (Theorem 5.7 of (Van der Vaart, 2000))**.** *Let M_n be random functions and let M be a fixed function of θ such that for $\forall\epsilon>0$,*

$$\sup_{\theta\in\Theta}|M_n(\theta)-M(\theta)|\xrightarrow{p}0,\quad\textit{and}\quad\sup_{\theta:d(\theta,\theta_*)\geq\epsilon}M(\theta)<M(\theta_*).$$

Then any sequence of $\hat{\theta}_n$ with $M_n(\hat{\theta}_n)\geq M_n(\theta_*)-\mathrm{o}_{\mathbb{P}}(1)$ converges in probability to θ_*.

D PROOFS FOR EFFICIENCY WITH A BUDGET CONSTRAINT

Proof of Theorem 3. We use $g_1(\cdot)$ to denote $-\frac{h^{(1)}(\cdot)}{h(\cdot)}$. The objective function is equal to

$$\begin{aligned}&\lim_{n\to\infty}\tfrac{p_1+\alpha(1-p_1)}{\mathrm{tr}\{\mathbf{V}\}}\\ &=\lim_{n\to\infty}\tfrac{p_1+\alpha(1-p_1)}{\mathrm{tr}\{\mathbb{E}[g_1(\theta_*^TX)^2h(\theta_*^TX)XX^T]-c\mathbb{E}[g_1(\theta_*^TX)h(\theta_*^TX)X]\mathbb{E}[g_1(\theta_*^TX)h(\theta_*^TX)X]\}},\end{aligned}$$

where $\lim_{n\to\infty}\frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}=c$. Thus replacing c with $\frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}$ in the above equation, the objective becomes

$$J_n(\alpha):=\frac{[p_1+\alpha(1-p_1)]}{\mathrm{tr}\left\{\mathbb{E}\left[g_1(\theta_*^TX)^2h(\theta_*^TX)XX^T\right]-\frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\mathbb{E}\left[g_1(\theta_*^TX)h(\theta_*^TX)X\right]\mathbb{E}\left[g_1(\theta_*^TX)h(\theta_*^TX)X\right]\right\}}.$$

So

$$\begin{aligned}\log J_n(\alpha) \quad &= \log(p_1 + \alpha(1-p_1)) \\ &\quad - \log\big[\mathrm{tr}\big\{\mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X) X X^T\right] \\ &\qquad - \tfrac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\mathbb{E}\left[g_1(\theta_*^T X) h(\theta_*^T X) X\right]\mathbb{E}\left[g_1(\theta_*^T X) h(\theta_*^T X) X\right]\big\}\big],\end{aligned}$$

taking derivative with respect to α, we have

$$\begin{aligned}&\tfrac{\partial \log J_n(\alpha)}{\partial \alpha} \\ = &\tfrac{(1-p_1)}{p_1+\alpha(1-p_1)} \\ &- \tfrac{\left(1/\alpha^2-1\right)(1-F(\tau_n))\mathrm{tr}\left\{\mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\right\}}{\mathrm{tr}\left\{\mathbb{E}[g_1(\theta_*^T X)^2 h(\theta_*^T X)XX^T] - \frac{(1-\alpha)^2(1-F(\tau_n))}{\alpha}\mathbb{E}[g_1(\theta_*^T X)h(\theta_*^T X)X]\mathbb{E}[g_1(\theta_*^T X)h(\theta_*^T X)X]\right\}}.\end{aligned}$$

So

$$\begin{aligned}&\tfrac{\partial \log J_n(\alpha)}{\partial \alpha} = 0 \\ \Longleftrightarrow\ &\tfrac{\mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X)XX^T\right]}{\mathbb{E}[g_1(\theta_*^T X)h(\theta_*^T X)X]\mathbb{E}[g_1(\theta_*^T X)h(\theta_*^T X)X]} = \left[\tfrac{(1-\alpha)^2}{\alpha} + \left(\tfrac{p_1}{\alpha(1-p_1)}+1\right)\tfrac{1-\alpha^2}{\alpha}\right](1-F(\tau_n)),\end{aligned}$$

where we use abuse of notation with the equality sign holds above when $n \to \infty$ (i.e. $a_n = b_n$ implies $a_n - b_n \to 0$ as $n \to \infty$). Then we have $\alpha^* = \mathrm{o}(1)$, because if α is bounded away from zero, the right hand side of the above equation is $\mathrm{o}(1)$ while the left hand side is $\mathrm{O}(1)$. Let $\beta = \lim_{n\to\infty}\frac{1-F(\tau_n)}{\alpha}$, since $\alpha \to 0$ as $n \to \infty$, and note that $\frac{p_1}{1-p_1} \to 0$, then the right hand side converges to 2β, so we have

$$\beta = \frac{1}{2}\frac{\mathrm{tr}\left\{\mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X)XX^T\right]\right\}}{\mathrm{tr}\left\{\mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\right\}},$$

indicating that

$$\alpha^* = \frac{2(1-F(\tau_n))\mathrm{tr}\left\{\mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\right\}}{\mathrm{tr}\left\{\mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X)XX^T\right]\right\}}.$$

□

E PROOFS FOR APPLICATION TO LOGISTIC REGRESSION

Proof of Proposition 3. Note that

$$h(\theta_1^T x) = \lim_{n\to\infty}\frac{1-F(\tau_n+\theta_1^T x)}{1-F(\tau_n)} = e^{-\theta_1^T x},$$

$g_1(\theta_1^T x) = \lim_{n\to\infty}\frac{F^{(1)}(\tau_n+\theta_1^T x)}{1-F(\tau_n+\theta_1^T x)} = \lim_{n\to\infty} F(\tau_n+\theta_1^T x) = 1$, and also we have $g_2(\theta_1^T x) = \lim_{n\to\infty}\frac{F^{(2)}(\tau_n+\theta_1^T x)}{1-F(\tau_n+\theta_1^T x)} = -1$, $g_3(\theta_1^T x) = \lim_{n\to\infty}\frac{F^{(3)}(\tau_n+\theta_1^T x)}{1-F(\tau_n+\theta_1^T x)} = 1$, so by Theorem 1 we have

$$\sqrt{\frac{n}{1+e^{\tau_n}}}\left(\hat{\theta}_* - \theta_*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{-1}\right), \quad \text{where}$$

$$\begin{aligned}\mathbf{V} \quad &= \mathbb{E}\left[g_1(\theta_*^T X)^2 h(\theta_*^T X)XX^T\right] - c\mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right]\mathbb{E}\left[g_1(\theta_*^T X)h(\theta_*^T X)X\right] \\ &= \mathbb{E}\left[e^{-\theta_*^T X}XX^T\right] - c\mathbb{E}\left[e^{-\theta_*^T X}X\right]\mathbb{E}\left[e^{-\theta_*^T X}X\right]\end{aligned}$$

Thus

$$\sqrt{ne^{-\tau_n}}\left(\hat{\theta}_* - \theta_*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{-1}\right),$$

then by dominated convergence theorem and Slutsky's theorem, we have

$$\sqrt{n\mathbb{E}\left[\frac{1}{1+e^{\tau_n+\theta_*^T X}}\right]}\left(\hat{\theta}_* - \theta_*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbb{E}_X\left[e^{-\theta_*^T X}\right]\mathbf{V}^{-1}\right),$$

which gives

$$\sqrt{n\mathbb{P}(Y=1)}\left(\hat{\theta}_* - \theta_*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbb{E}_X\left[e^{-\theta_*^T X}\right]\mathbf{V}^{-1}\right).$$

□

Proof of Proposition 4. First note that Assumption 2 holds. We use $g_1(\cdot), g_2(\cdot), g_3(\cdot)$ to denote $-h^{(1)}(\cdot)/h(\cdot), -h^{(2)}(\cdot)/h(\cdot), -h^{(3)}(\cdot)/h(\cdot)$. By definition, for

$$F(\tau_n + \theta_1^T x) = \frac{e^{\tau_n + \theta_1^T x}}{1 + e^{\tau_n + \theta_1^T x}}, h(\theta_1^T x) = \lim_{n\to\infty} \frac{1 - F(\tau_n + \theta_1^T x)}{1 - F(\tau_n)} = \lim_{n\to\infty} \frac{1 + e^{\tau_n}}{1 + e^{\tau_n + \theta_1^T x}} = e^{-\theta_1^T x},$$

$$g_1(\theta_1^T x) = \lim_{n\to\infty} \frac{F^{(1)}(\tau_n + \theta_1^T x)}{1 - F(\tau_n + \theta_1^T x)} = 1,$$

$$g_2(\theta_1^T x) = \lim_{n\to\infty} \frac{F^{(2)}(\tau_n + \theta_1^T x)}{1 - F(\tau_n + \theta_1^T x)} = \lim_{n\to\infty} \frac{F^{(2)}(\tau_n + \theta_1^T x)}{1 - F(\tau_n + \theta_1^T x)} = -1,$$

$g_3(\theta_1^T x) = \lim_{n\to\infty} \frac{F^{(3)}(\tau_n + \theta_1^T x)}{1 - F(\tau_n + \theta_1^T x)} = 0$. Thus the conditions in Assumption 3 are satisfied. Further note that the conditions of Theorem 1 also hold. Then by Theorem 3,

$$\alpha^* = \frac{2(1 + e^{\tau_n})^{-1} \mathrm{tr} \left\{ \mathbb{E}[e^{-\theta_*^T X} X] \mathbb{E}[e^{-\theta_*^T X} X] \right\}}{\mathrm{tr}\{\mathbb{E}[e^{-\theta_*^T X} X X^T]\}}$$

□