Approximating Partial Likelihood Estimators via Optimal Subsampling

Haixiang Zhang[1*], Lulu Zuo[1], HaiYing Wang[2] and Liuquan Sun[3]

[1] *Center for Applied Mathematics, Tianjin University, Tianjin 300072, China*

[2] *Department of Statistics, University of Connecticut, Storrs, Mansfield, CT 06269, USA*

[3] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

Abstract

With the growing availability of large-scale biomedical data, it is often time-consuming or infeasible to directly perform traditional statistical analysis with relatively limited computing resources at hand. We propose a fast subsampling method to effectively approximate the full data maximum partial likelihood estimator in Cox's model, which largely reduces the computational burden when analyzing massive survival data. We establish consistency and asymptotic normality of a general subsample-based estimator. The optimal subsampling probabilities with explicit expressions are determined via minimizing the trace of the asymptotic variance-covariance matrix for a linearly transformed parameter estimator. We propose a two-step subsampling algorithm for practical implementation, which has a significant reduction in computing time compared to the full data method. The asymptotic properties of the resulting two-step subsample-based estimator is also established. Extensive numerical experiments and a real-world example are provided to assess our subsampling strategy. Supplemental materials for this article are available online.

Keywords: Asymptotic normality; Empirical process; L-optimality criterion; Massive data; Survival analysis.

*Corresponding author. Email: haixiang.zhang@tju.edu.cn (Haixiang Zhang)

1 Introduction

With the development of science and technology, the amounts of available data are rapidly increasing in recent years. A major bottleneck to analyze huge datasets is that the data volume exceeds the capacity of available computational resources. It is not always possible to meet the demands for computational speed and storage memory if we directly perform traditional analysis for large datasets with a single computer at hand. To cope with big data, there are many statistical methods in the literature dealing with the heavy calculation and storage burden. Basically, we could classify these methods into three categories. (i) *divide-and-conquer approach* (Zhao *et al.*, 2016; Battey *et al.*, 2018; Shi *et al.*, 2018; Jordan *et al.*, 2019; Volgushev *et al.*, 2019; Chen *et al.*, 2022; Fan *et al.*, 2021). (ii) *online updating approach* (Schifano *et al.*, 2016; Luo and Song, 2020; Lin *et al.*, 2020; Luo *et al.*, 2022; Wang *et al.*, 2022b). (iii) *subsampling-based approach*. The subsampling is an emerging field for big data. Many papers have been published during recent years. For example, Wang *et al.* (2018) and Wang (2019) studied the optimal subsampling for massive logistic regression. Wang *et al.* (2019) presented an information-based subdata selection approach for linear regression with big datasets. Zhang *et al.* (2020) studied an effective sketching method for massive datasets via A-optimal subsampling. Yao and Wang (2019), Han *et al.* (2020) and Yao *et al.* (2021) proposed several subsampling methods for large-scale multiclass logistic regression. Yu *et al.* (2022) considered optimal Poisson subsampling for maximum quasi-likelihood estimator with massive data. Zhang *et al.* (2021) proposed a response-free optimal sampling procedure for generalized linear models under measurement constraints. Wang and Ma (2021) studied the optimal subsampling for quantile regression in big data. Liu *et al.* (2021) proposed an optimal subsampling method for the functional linear model via L-optimality criterion. Zhang and Wang (2021) and Zuo *et al.* (2021b) considered optimal distributed subsampling methods for big data in the context of linear and logistic models, respectively. Ai *et al.* (2021) studied the optimal subsampling method for generalized linear models under the A-optimality criterion. Wang and Zhang (2022) proposed an optimal subsampling procedure for multiplicative regression with massive data. For more related results on massive data analysis, we refer to several review papers by Wang *et al.* (2016),

Lee and Ng (2020), Yao and Wang (2021), Chen *et al.* (2021b), Li and Meng (2021) and Yu *et al.* (2023).

The aforementioned investigations focused on developing statistical methods for large datasets with uncensored observations. In recent years, huge biomedical datasets become increasingly common, and they are often subject to censoring (Kleinbaum and Klein, 2005). There have been several recent papers on statistical analysis of massive censored survival data. For example, Xue *et al.* (2019) and Wu *et al.* (2021) studied the online updating approach for streams of survival data. Keret and Gorfine (2020) presented an optimal Cox regression subsampling procedure with rare events. Tarkhan and Simon (2020) and Xu *et al.* (2020) used the stochastic gradient descent algorithms to analyse large-scale survival datasets with Cox's model and the accelerated failure time models, respectively. Li *et al.* (2020) proposed a batch screening iterative Lasso method for large-scale and ultrahigh-dimensional Cox model. Zuo *et al.* (2021a) proposed a sampling-based method for massive survival data with additive hazards model. Wang *et al.* (2021) studied an efficient divide-and-conquer algorithm to fit high-dimensional Cox regression for massive datasets. Yang *et al.* (2022) studied the optimal subsampling algorithms for parametric accelerate failure time models with massive survival data. In spite of the aforementioned papers, existing research on massive survival data is relatively limited, and it is meaningful to further investigate the statistical theories in the area of large-scale survival analysis.

It is worthy mentioning that subsampling is an emerging area of research, which has attracted great attentions in both statistics and computer science (Ma *et al.*, 2015; Bai *et al.*, 2021). Subsampling methods focus on selecting a small proportion of the full data as a surrogate to perform statistical computations. A key to success of subsampling is to design nonuniform sampling probabilities so that those influential or informative data points are sampled with high probabilities. Although significant progress has been made towards developing optimal subsampling theory for uncensored observations, to the best of our knowledge, the research on optimal subsampling for large-scale survival data lags behind. In consideration of the important role of Cox's model in the field of survival analysis (Cox, 1972; Fleming and Harrington, 1991), it is desirable to develop effective subsampling methods in the context of Cox's model for massive survival data. This paper aims to close this

gap by developing a subsample-based estimator to fast and effectively approximate the full data maximum partial likelihood estimator. Our aim is to design an efficient subsampling and estimation strategy to better balance the trade-off between computational efficiency and statistical efficiency. Here are some key differences between our proposed subsampling approach and some recently developed approaches on Cox's model with large-scale data: (i) Keret and Gorfine (2020) proposed a subsampling-based estimation for Cox's model with rare events by including all observed failures, while our optimal subsampling method is developed for Cox's model under the regular setting that observed failure times are not rare compared with the observed censoring times. (ii) Tarkhan and Simon (2020) presented a stochastic gradient descent (SGD) procedure for Cox's model. This method primarily intends to resolve the problems that the whole dataset cannot be easily loaded in memory; the main aim is to deal with the out-of-memory issue rather than speeding up the calculation. (iii) Li *et al.* (2020) and Wang *et al.* (2021) studied the variable selection problem for ultrahigh-dimensional Cox's model, which is different from the focus of our paper on dealing with very large sample sizes.

The main contributions of our proposed subsampling method include three aspects: First, the computation of our subsample-based estimator is much faster than that of the full data estimator calculated by the standard R function `coxph`. Therefore, it effectively reduces the computational burden when analysing massive survival data with Cox's model. Second, we provide an explicit expression for the optimal subsampling distribution, which has much better performance than the uniform subsampling distribution in terms of statistical efficiency. Third, we establish consistency and asymptotic normality of the proposed subsample estimator, which is useful for performing statistical inference (e.g. constructing confidence intervals and testing hypotheses).

The remainder of this paper is organized as follows. In Section 2, we review the setup and notations for Cox's model. A general subsample-based estimator is proposed to approximate the full data maximum partial likelihood estimator. In Section 3, we establish consistency and asymptotic normality of a general subsample-based estimator. The optimal subsampling probabilities are explicitly specified in the context of L-optimality criterion. In Section 4, we give a two-step subsampling algorithm together with the asymptotic properties of the

resulting estimator. In Section 5, extensive simulations together with an application are conducted to verify the validity of the proposed subsampling procedure. Some concluding remarks are presented in Section 6. Technical proofs are given in the supplement.

2 Model and Subsample-Based Estimation

In many biomedical applications, the outcome of interest is measured as a "time-to-event", such as death and onset of cancer. The time to occurrence of an event is referred to as a failure time (Kalbfleisch and Prentice, 2002), and its typical characteristic is subject to possible right censoring. For $i = 1, \cdots, n$, let T_i be the failure time, C_i be the censoring time, and $\mathbf{X}_i$ be the p-dimensional vector of time-independent covariates (e.g., treatment indicator, blood pressure, age, and gender). We assume that T_i and C_i are conditionally independent given $\mathbf{X}_i$. The observed failure time is $Y_i = \min(T_i, C_i)$, and the failure indicator is $\Delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. For convenience, we denote the full data of independent and identically distributed observations from the population as $\mathcal{D}_n = \{(\mathbf{X}_i, \Delta_i, Y_i), i = 1, \cdots, n\}$. The Cox's proportional hazard regression model (Cox, 1972) is commonly used to describe the relationship between covariates of an individual and the risk of experiencing an event. This model assumes that the conditional hazard rate function of T_i given $\mathbf{X}_i$ is

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{X}_i), \tag{1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is a p-dimensional vector of regression parameters, and its true value belongs to a compact set $\Theta \subset \mathbb{R}^p$. To estimate $\boldsymbol{\beta}$, Cox (1975) proposed a novel *partial likelihood* method. The negative log-partial likelihood function is

$$\ell(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \left[\boldsymbol{\beta}'\mathbf{X}_i - \log \left\{ \sum_{j=1}^{n} I(Y_j \geq t) \exp(\boldsymbol{\beta}'\mathbf{X}_j) \right\} \right] dN_i(t), \tag{2}$$

where $N_i(t) = I(\Delta_i = 1, Y_i \leq t)$ is a counting process and τ is a prespecified positive constant. One advantage of Cox's *partial likelihood* method is that the criterion function given in (2) does not involve the nonparametric baseline hazard function $\lambda_0(t)$, and the

resulting estimator of $\boldsymbol{\beta}$ is asymptotically equivalent to the parameter estimator obtained by maximizing the full likelihood function (Cox, 1975).

For convenience, we introduce the following notations to ease the presentation:

$$S^{(k)}(t,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq t)\mathbf{X}_i^{\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}_i),\ k = 0, 1 \text{ and } 2,$$

where the notation $\mathbf{u}^{\otimes k}$ means $\mathbf{u}^{\otimes 0} = 1$, $\mathbf{u}^{\otimes 1} = \mathbf{u}$ and $\mathbf{u}^{\otimes 2} = \mathbf{u}\mathbf{u}'$ for a vector $\mathbf{u}$. Throughout this paper, $\|\mathbf{A}\| = (\sum_{1\leq i,j\leq p} a_{ij}^2)^{1/2}$ for a matrix $\mathbf{A} = (a_{ij})$.

The gradient of $\ell(\boldsymbol{\beta})$ is

$$\begin{aligned}\dot{\ell}(\boldsymbol{\beta}) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{X}_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dN_i(t)\\ &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{X}_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta}),\end{aligned} \tag{3}$$

where $M_i(t,\boldsymbol{\beta}) = N_i(t) - \int_0^t I(Y_i \geq u)\exp(\boldsymbol{\beta}'\mathbf{X}_i)\lambda_0(u)du$, and

$$\bar{\mathbf{X}}(t,\boldsymbol{\beta}) = \frac{S^{(1)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})}. \tag{4}$$

The Hessian matrix of $\ell(\boldsymbol{\beta})$ is given by

$$\ddot{\ell}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left[\frac{S^{(2)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})} - \left\{\frac{S^{(1)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})}\right\}^{\otimes 2}\right]dN_i(t). \tag{5}$$

According to Cox (1975), the full data maximum partial likelihood (MPL) estimator $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ is the solution to $\dot{\ell}(\boldsymbol{\beta}) = 0$, and the asymptotic properties of $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ have been investigated by Andersen and Gill (1982). There is no closed-form to $\hat{\boldsymbol{\beta}}_{\text{MPL}}$, and it is numerically calculated by Newton's method through iteratively applying

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \{\ddot{\ell}(\boldsymbol{\beta}^{(m)})\}^{-1}\dot{\ell}(\boldsymbol{\beta}^{(m)}). \tag{6}$$

For small datasets with hundreds of observations or even fewer, the iterative algorithm given in (6) is able to converge in a reasonable time. For moderate datasets, it is common to use the gold standard `coxph` function in the R package of Therneau (2021), where a smart updating procedure is adopted to speed up the computation (Simon *et al.*, 2011). The computational efficiency of `coxph` will be presented in the simulation section.

It is desirable to develop an effective and computationally stable method when handling massive survival datasets with Cox's model. Recently, Keret and Gorfine (2020) introduced a novel subsampling procedure for Cox regression with rare events, while our aim is to propose a subsampling procedure for large-scale Cox model under non-rare-events setting. To be specific, we assign subsampling probabilities $\{\pi_i\}_{i=1}^n$ to the full data $\mathcal{D}_n$, where $\sum_{i=1}^n \pi_i = 1$ and $\pi_i > 0$ for $i = 1, \cdots, n$. Draw a random subsample of size r with replacement based on $\{\pi_i\}_{i=1}^n$ from the full data $\mathcal{D}_n$, where r is typically much smaller than n. Let $\mathcal{D}_r^* = \{(\mathbf{X}_i^*, \Delta_i^*, Y_i^*, \pi_i^*)\}_{i=1}^r$ be a selected subsample with size r from the full data $\mathcal{D}_n$, where $\mathbf{X}_i^*$, Δ_i^*, Y_i^*, and π_i^* are the covariate, the failure indicator, the observed failure times, and the subsampling probability, respectively, in the subsample. We propose a weighted pseudo log partial likelihood using the subsample $\mathcal{D}_r^*$:

$$\ell^*(\boldsymbol{\beta}) = -\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\int_0^\tau \left\{\boldsymbol{\beta}'\mathbf{X}_i^* - \log\left[r^{-1}\sum_{j=1}^{r}\pi_j^{*-1} I(Y_j^* \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_j^*)\right]\right\} dN_i^*(t), \quad (7)$$

where $N_i^*(t) = I(\Delta_i^* = 1, Y_i^* \leq t)$. The inverse probability weighting in (7) is to ensure the consistency of the resulting subsample estimator towards $\hat{\boldsymbol{\beta}}_{\text{MPL}}$, which will be carefully investigated in Section 3. The corresponding weighted subsample score function is

$$\begin{aligned}\dot{\ell}^*(\boldsymbol{\beta}) &= -\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}^*(t,\boldsymbol{\beta})\} dN_i^*(t) \quad (8)\\ &= -\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}^*(t,\boldsymbol{\beta})\} dM_i^*(t,\boldsymbol{\beta}),\end{aligned}$$

where $M_i^*(t,\boldsymbol{\beta}) = N_i^*(t) - \int_0^t I(Y_i^* \geq u)\exp(\boldsymbol{\beta}'\mathbf{X}_i^*)\lambda_0(u)du$, and

$$\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) = \frac{S^{*(1)}(t,\boldsymbol{\beta})}{S^{*(0)}(t,\boldsymbol{\beta})} \quad (9)$$

with

$$S^{*(k)}(t,\boldsymbol{\beta}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*} I(Y_i^* \geq t)(\mathbf{X}_i^*)^{\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}_i^*), \quad k = 0, 1 \text{ and } 2. \quad (10)$$

The subsample-based estimator $\tilde{\boldsymbol{\beta}}$ is the solution to $\dot{\ell}^*(\boldsymbol{\beta}) = 0$, which is computationally easier to solved by Newton's method due to the smaller subsample size. Here $\tilde{\boldsymbol{\beta}}$ can be viewed as a subsample approximation to the full data $\hat{\boldsymbol{\beta}}_{\text{MPL}}$. A natural question is how to select the subsample so that $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ are close. We will derive the asymptotic distribution of $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}$ and then find the probabilities that minimize a function of the asymptotic variance.

3 Asymptotic Properties and Subsampling Strategy

In this section, we establish the asymptotic properties of a general subsample-based estimator $\tilde{\boldsymbol{\beta}}$ obtained via solving $\dot{\ell}^*(\tilde{\boldsymbol{\beta}}) = 0$, where $\dot{\ell}^*(\boldsymbol{\beta})$ is given in (8). A strategy on how to specify optimal subsampling probabilities for our method is presented. We need the following assumptions for theoretical derivation. Throughout this paper we allow π_i's to depend on the data, so they may be random.

Assumption 1 *The baseline hazard satisfies that* $\int_0^\tau \lambda_0(t)dt < \infty$, *and* $P(T_i \geq \tau) > 0$.

Assumption 2 *The quantity* $\frac{1}{n}\sum_{i=1}^n \int_0^\tau \left[\frac{S^{(2)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})} - \left\{\frac{S^{(1)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})}\right\}^{\otimes 2}\right] dN_i(t)$ *converges in probability to a positive definite matrix for all* $\boldsymbol{\beta} \in \Theta$, *where* Θ *is a compact set containing the true value of* $\boldsymbol{\beta}$.

Assumption 3 *The time-independent covariates* $\mathbf{X}_i$*'s are bounded.*

Assumption 4 *The subsampling probabilities satisfy* $\max_{1\leq i\leq n}(n\pi_i)^{-1} = O_P(1)$.

Assumptions 1 and 2 are two classical regularity conditions for Cox's model (Andersen and Gill, 1982); Assumption 3 is a bounded condition, which was commonly imposed in the literature about Cox's model, e.g., Huang *et al.* (2013) and Fang *et al.* (2017). This assumption is reasonable in most practical applications, because for biomedical survival data the covariates of an individual are often treatment indicator, blood pressure, age, and gender, etc. These biomedical related features are usually bounded (Keret and Gorfine, 2020). Assumption 4 is required to protect the weighted subsample pseudo-score function given in (8) from being dominated by those data points with extremely small subsampling probabilities. i.e., Assumption 4 requires that the minimum subsampling probability is at the same order of $1/n$ in probability. This assumption was also imposed by Wang *et al.* (2022a).

We establish the consistency and asymptotic normality of the subsample-based estimator $\tilde{\boldsymbol{\beta}}$ conditional on the full data $\mathcal{D}_n$ in the following. This result plays an important role in performing statistical inference. In addition, the asymptotic distribution is a key foundation to design optimal subsampling probabilities for our method. Throughout this paper, the

notation $b = O_{P|\mathcal{D}_n}(1)$ denotes that b is bounded in conditional probability, i.e., for any $\epsilon > 0$, there exists a finite $b_\epsilon > 0$ such that $P\{P(|b| \geq b_\epsilon|\mathcal{D}_n) < \epsilon\} \to 1$.

Theorem 1 *Under assumptions 1-4, if $r = o(n)$ as $n \to \infty$ and $r \to \infty$, then the subsample-based estimator $\tilde{\boldsymbol{\beta}}$ is consistent to $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ with a convergence rate $O_{P|\mathcal{D}_n}(r^{-1/2})$. In addition, conditional on $\mathcal{D}_n$ in probability, we have*

$$\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}), \tag{11}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\boldsymbol{\Sigma} = \boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}$ with

$$\boldsymbol{\Psi} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\frac{S^{(2)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{(1)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}\right] dN_i(t), \tag{12}$$

and

$$\boldsymbol{\Gamma} = \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_i}\left[\int_0^\tau \left\{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\right\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\right]^{\otimes 2}. \tag{13}$$

Remark 1 *The convergence rate indicates that $\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| = O_{P|\mathcal{D}_n}(r^{-1/2})$. Since a random sequence that converges to zero in conditional probability also converges to zero in unconditional probability (Xiong and Li, 2008), we know that $\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| = o_{P|\mathcal{D}_n}(1) = o_P(1)$ as $r \to \infty$. Therefore, the subsample-based estimator $\tilde{\boldsymbol{\beta}}$ is close to $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ as long as r is large enough. It is reasonable to use $\tilde{\boldsymbol{\beta}}$ as a surrogate for $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ in order to reduce computational burden when handling large-scale Cox's model in practice.*

Remark 2 *The asymptotic normality in condition distribution indicates that the distribution of the error term $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}$ conditional on $\mathcal{D}_n$ can be approximated by that of a normal random variable, say $\mathbf{Z}$, with conditional distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. This means that for any $\epsilon > 0$, the probability $P(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| \geq \epsilon|\mathcal{D}_n)$ is accurately approximated by $P(\|\mathbf{Z}\| \geq \epsilon|\mathcal{D}_n)$. Hence, a smaller variance ensures a smaller excess error bound. This sheds light on how to design optimal subsampling probabilities for our proposed sampling method.*

For practical application of the proposed sampling strategy, we need to specify the subsampling distribution $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$. A simple choice is the uniform subsampling with $\{\pi_i = n^{-1}\}_{i=1}^n$. However, this is not optimal, because it does not distinguish the importances

among different data points. It is desirable to design nonuniform subsampling probabilities such that more informative data points are more likely to be selected into a subsample (Wang *et al.*, 2018; Wang and Ma, 2021). In view of Remark 2, we propose to determine nonuniform subsampling probabilities by minimizing the asymptotic variance-covariance matrix $\boldsymbol{\Sigma}$ given in (11). However, the meaning of "minimizing" a matrix needs to be carefully defined. Here we adopt the idea from design of experiments (Kiefer, 1959), and determine the optimal subsampling probabilities by minimizing a convex function of $\boldsymbol{\Sigma}$. We follow the idea of Wang *et al.* (2018) and focus on minimizing $\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi})$=$\mathrm{tr}(\boldsymbol{\Gamma})$, where

$$\begin{aligned}\mathrm{tr}(\boldsymbol{\Gamma}) &= \mathrm{tr}\left(\frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_i}\left[\int_0^\tau \left\{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\right\} dM_i(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\right]^{\otimes 2}\right) \\ &= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\left\|\int_0^\tau \left\{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\right\} dM_i(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\right\|^2.\end{aligned}$$

As a matter of fact, this optimality criterion of minimizing $\mathrm{tr}(\boldsymbol{\Gamma})$ is a version of L-optimality criterion (Atkinson *et al.*, 2007), because $\mathrm{tr}(\boldsymbol{\Gamma})$ is trace of the asymptotic variance-covariance matrix of $\boldsymbol{\Psi}\tilde{\boldsymbol{\beta}}$, which is a linearly transformed subsample estimator. The following theorem provides an explicit expression for the optimal subsampling distribution $\boldsymbol{\pi}^{\mathrm{Lopt}} = \{\pi_i^{\mathrm{Lopt}}\}_{i=1}^n$ in the context of L-optimality criterion.

Theorem 2 *If the subsampling probabilities are chosen as*

$$\pi_i^{\mathrm{Lopt}} = \frac{\| \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\} dM_i(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\|}{\sum_{j=1}^n \| \int_0^\tau \{\mathbf{X}_j - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\} dM_j(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\|}, \quad i = 1, \cdots, n, \tag{14}$$

then $tr(\boldsymbol{\Gamma})$ *attains its minimum.*

Remark 3 *The numerator of* π_i^{Lopt} *has a term* $\bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})$*, which contains all individuals of the full data* $\mathcal{D}_n$*. This is different from existing results on parametric models without censoring, for which numerators of optimal subsampling probabilities involve only individual observations' information (except the dependency of the full data estimator). Practical adjustments are required to implement the optimal subsampling probabilities to tackle the additional computational challenge due to censored survival data with Cox's model. We will discuss this in Section 4.*

Remark 4 *With the A-optimality criterion (Wang* et al.*, 2018), we can derive the corresponding optimal subsampling probabilities that minimize tr($\boldsymbol{\Sigma}$). They are*

$$\pi_i^{\text{Aopt}} = \frac{\|\boldsymbol{\Psi}^{-1}\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\|}{\sum_{j=1}^n \|\boldsymbol{\Psi}^{-1}\int_0^\tau \{\mathbf{X}_j - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_j(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\|}, \quad i = 1, \cdots, n, \tag{15}$$

where $\boldsymbol{\Psi}$ is given in (12). Due to the term $\boldsymbol{\Psi}$ in (15), the computational burden of π_i^{Aopt} is much heavier than that of π_i^{Lopt}. Therefore, we focus on π_i^{Lopt} in the presentation of our subsampling procedure. We provide numerical comparisons between the A-optimality criterion and the L-optimality criterion in Section 5.1.

We provide more insights on the optimal subsampling probabilities $\{\pi_i^{\text{Lopt}}\}_{i=1}^n$ from two aspects: First, the numerator $\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})$ is actually the ith sample's score given in (3), which is also referred to as the residual (Therneau *et al.*, 1990). The subsampling probabilities in Keret and Gorfine (2020) for censored individuals also share a similar spirit, but the subsampling probabilities for observed events are one in Keret and Gorfine (2020)'s approach. Second, since the failure times are observed for uncensored observations, they contain more information than censored observations. The optimal subsampling probabilities give higher preferences to uncensored observations compared with censored observations. This will be demonstrated numerically in Section 5.1.

4 Practical Implementation

4.1 Two-Step Subsampling Algorithm

In this section, we discuss some issues on practical implementation and provide strategies to resolve them.

First, the optimal subsampling probabilities $\{\pi_i^{\text{Lopt}}\}_{i=1}^n$ contain the full data estimator $\hat{\boldsymbol{\beta}}_{\text{MPL}}$. We take a pilot subsample from $\mathcal{D}_n$ by uniform subsampling with replacement, say $\mathcal{D}_{r_0}^* = \{(\mathbf{X}_i^{0*}, \Delta_i^{0*}, Y_i^{0*}), i = 1, \cdots, r_0\}$, obtain a pilot estimator $\tilde{\boldsymbol{\beta}}_0$ using $\mathcal{D}_{r_0}^*$, and use $\tilde{\boldsymbol{\beta}}_0$ to replace the $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ in (14) for practical implementation.

Second, the resultant probabilities still involve a term $\bar{\mathbf{X}}(t,\tilde{\boldsymbol{\beta}}_0)$ after replacing $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ with $\tilde{\boldsymbol{\beta}}_0$. This term involves the full data $\mathcal{D}_n$ so it requires heavy computation burden. To tackle

this problem, we recommend replacing $\bar{\mathbf{X}}(t, \tilde{\boldsymbol{\beta}}_0)$ with $\bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)$, where

$$\bar{\mathbf{X}}^{0*}(t, \boldsymbol{\beta}) = \frac{\sum_{j=1}^{r_0} I(Y_j^{0*} \geq t) \mathbf{X}_j^{0*} \exp(\boldsymbol{\beta}' \mathbf{X}_j^{0*})}{\sum_{j=1}^{r_0} I(Y_j^{0*} \geq t) \exp(\boldsymbol{\beta}' \mathbf{X}_j^{0*})}. \tag{16}$$

This is a reasonable choice because it can be shown that $\bar{\mathbf{X}}^{0*}(t, \boldsymbol{\beta}) = \bar{\mathbf{X}}(t, \boldsymbol{\beta}) + o_P(1)$, for any $t \in [0, \tau]$ and $\boldsymbol{\beta} \in \Theta$ (see Eq. (S.6) in the Appendix).

Third, the term $dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})$ involves the unknown baseline hazard function $\lambda_0(t)$. We propose a subsample Breslow-type estimator for $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ using $\mathcal{D}_{r_0}^*$ as follows:

$$\hat{\Lambda}_0^{\text{UNIF}}(t, \boldsymbol{\beta}) = \sum_{i=1}^{r_0} \left\{ \frac{\Delta_i^{0*} I(Y_i^{0*} \leq t)}{\sum_{j=1}^{r_0} I(Y_j^{0*} \geq Y_i^{0*}) \exp(\boldsymbol{\beta}' \mathbf{X}_j^{0*})} \right\}. \tag{17}$$

Taking into account previous discussions, the approximated optimal subsampling probabilities are

$$\pi_i^{\text{app}} = \frac{\| \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\} d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) \|}{\sum_{j=1}^n \| \int_0^\tau \{\mathbf{X}_j - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\} d\hat{M}_j(t, \tilde{\boldsymbol{\beta}}_0) \|}, \quad i = 1, \cdots, n, \tag{18}$$

where $d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) = dN_i(t) - I(Y_i \geq t) \exp(\tilde{\boldsymbol{\beta}}_0' \mathbf{X}_i) d\hat{\Lambda}_0^{\text{UNIF}}(t, \tilde{\boldsymbol{\beta}}_0)$.

Fourth, we see from (18) that π_i^{app} is proportional to $\| \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\} d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) \|$, which could be small for some data points. Since π_i^{app}'s are obtained by inserting the pilot $\tilde{\boldsymbol{\beta}}_0$, they are not the real optimal probabilities. The variation of $\tilde{\boldsymbol{\beta}}_0$ may be significantly amplified by data points with much smaller values of π_i^{app} than other data points, because the sampling probabilities appear in the denominator of the asymptotic variance as shown in (13). From another angle, if some data points with much smaller π_i^{app} are selected into a subsample, the weighted subsample pseudo-score function (8) may be dominated by these data points and thus the variance of the resulting estimator is inflated by them. Following the idea of defensive importance sampling (Hesterberg, 1995; Owen and Zhou, 2000), we mix the approximated optimal subsampling distribution with the uniform subsampling distribution. Specifically, we use $\pi_{\delta i}^{\text{app}} = (1 - \delta)\pi_i^{\text{app}} + \delta/n$ instead of π_i^{app} in (18) for practical implementation, where $\delta \in (0, 1)$ controls the proportion of mixture. A main advantage of this approach is that $n\pi_{\delta i}^{\text{app}}$ is lower-bounded by δ, so it ensures robustness of the subsampling estimator. The same idea was also adopted by other subsampling methods in the literature, such as Ma *et al.* (2015); Yu *et al.* (2022); Wang *et al.* (2022a). We use $\delta = 0.1$

in the numerical simulations and real-world application in Section 5, and this choice works well.

We present a practical two-step subsampling method for Cox's model in Algorithm 1.

Algorithm 1 Two-Step Subsampling Procedure

• **Step 1**. Take a pilot subsample of size r_0 $\mathcal{D}_{r_0}^* = \{(\mathbf{X}_i^{0*}, \Delta_i^{0*}, Y_i^{0*})\}_{i=1}^{r_0}$ using uniform subsampling with replacement from the full data $\mathcal{D}_n$. Here r_0 is typically much smaller than r. Compute a pilot estimator $\tilde{\boldsymbol{\beta}}_0$ by solving

$$\dot{\ell}^{0*}(\boldsymbol{\beta}) = -\frac{1}{r_0}\sum_{i=1}^{r_0}\Delta_i^{0*}\{\mathbf{X}_i^{0*} - \bar{\mathbf{X}}^{0*}(Y_i^{0*}, \boldsymbol{\beta})\} = 0,$$

where $\bar{\mathbf{X}}^{0*}(t, \boldsymbol{\beta})$ is given in (16). Calculate

$$\pi_{\delta i}^{\text{app}} = (1-\delta)\pi_i^{\text{app}} + \frac{1}{n}\delta, \quad , i = 1, ..., n, \tag{19}$$

where π_i^{app}'s are given in (18) and δ is often a small number. e.g., $\delta = 0.1$.

• **Step 2**. Draw r data points with replacement from the full data $\mathcal{D}_n$ using the subsampling probabilities $\{\pi_{\delta i}^{\text{app}}\}_{i=1}^n$ given in (19). Let $\mathcal{D}_r^* = \{(\mathbf{X}_i^*, \Delta_i^*, Y_i^*, \pi_{\delta i}^{\text{app}*}\}_{i=1}^r$ be the selected subsample. Obtain the two-step subsample-based estimator $\breve{\boldsymbol{\beta}}$ by solving

$$\dot{\ell}_{\tilde{\boldsymbol{\beta}}_0}^*(\boldsymbol{\beta}) = -\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi_{\delta i}^{\text{app}*}}\Delta_i^*\{\mathbf{X}_i^* - \bar{\mathbf{X}}_{\tilde{\boldsymbol{\beta}}_0}^*(Y_i^*, \boldsymbol{\beta})\} = 0, \tag{20}$$

where $\bar{\mathbf{X}}_{\tilde{\boldsymbol{\beta}}_0}^*(t, \boldsymbol{\beta})$ has the same expression as $\bar{\mathbf{X}}^*(t, \boldsymbol{\beta})$ given in (9) except that π_i^* is replaced with $\pi_{\delta i}^{\text{app}*}$.

Note that we do not recommend combining $\mathcal{D}_{r_0}^*$ and $\mathcal{D}_r^*$ together for Step 2 of Algorithm 1. If we are able to handle the calculation on the combined data from $\mathcal{D}_{r_0}^*$ and $\mathcal{D}_r^*$, then it is more efficient to increase the second step subsample size to $r_0 + r$. That is to say, the pilot subsample $\mathcal{D}_{r_0}^*$ does not come into the estimation step for $\breve{\boldsymbol{\beta}}$. Therefore, we do not need to allocate two subsample sizes r_0 and r when implementing our method. Some discussion and guidance on the selection of r_0 are provided in Section 5.1.

We established the asymptotic normality of the estimator $\breve{\boldsymbol{\beta}}$ from the practical Algorithm 1 in the following theorem.

Theorem 3 *Under assumptions 1-3, if $r = o(n)$, then as $r_0 \to \infty$, $r \to \infty$ and $n \to \infty$, conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, the two-step estimator $\breve{\boldsymbol{\beta}}$ in Algorithm 1 is consistent to $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ with convergence rate $r^{-1/2}$. Furthermore, the approximation error has an asymptotically normal distribution, that is*

$$\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) \xrightarrow{d} N(0, \mathbf{I}), \tag{21}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}$ with $\boldsymbol{\Psi}$ defined in (12),

$$\boldsymbol{\Gamma} = \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_{\delta i}^{\text{Lopt}}}\left[\int_0^\tau \left\{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\right\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\right]^{\otimes 2}, \tag{22}$$

and

$$\pi_{\delta i}^{\text{Lopt}} = (1-\delta)\pi_i^{\text{Lopt}} + \frac{\delta}{n}, \quad i = 1, \cdots, n. \tag{23}$$

Remark 5 *Note that the full data estimator $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ converges to the true parameter at a rate of $n^{-1/2}$, so the full data estimator $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ in (21) can be replaced by the true parameter since $r = o(n)$. Thus the asymptotic result in Theorem 3 can be used for inference on the true parameter. Since the subsampling rate is often very small when dealing with large-scale datasets, it is reasonable to apply the asymptotic normality in practice.*

The following proposition gives the unconditional convergence rate and asymptotic normality of the two-step subsample estimator towards the true parameter, which are very useful when we perform inference with respect to the true parameter.

Proposition 1 *Under assumptions 1-3, if $r = o(n)$, then as $r_0 \to \infty$, $r \to \infty$ and $n \to \infty$, the two-step estimator $\breve{\boldsymbol{\beta}}$ in Algorithm 1 is consistent to the true parameter $\boldsymbol{\beta}_0$ with convergence rate $r^{-1/2}$. i.e., we have $\|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(r^{-1/2})$. Moreover, $\breve{\boldsymbol{\beta}}$ is asymptotically normal, that is*

$$\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{I}),$$

where $\boldsymbol{\Sigma}$ is given in Theorem 3.

In view of Proposition 1 and Remark 5, we need to provide an estimate for the variance-covariance matrix of $\breve{\boldsymbol{\beta}}$ when conducting statistical inference for the true parameter. A simple method is to replace $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ with $\breve{\boldsymbol{\beta}}$ in the asymptotic variance-covariance matrix $\boldsymbol{\Sigma}$. However, this requires the calculation on the full data $\mathcal{D}_n$. To reduce computational cost, we propose to estimate the variance-covariance matrix of $\breve{\boldsymbol{\beta}}$ using the subsample $\mathcal{D}_r^*$ only with

$$\breve{\boldsymbol{\Sigma}} = \breve{\boldsymbol{\Psi}}^{-1}\breve{\boldsymbol{\Gamma}}\breve{\boldsymbol{\Psi}}^{-1}, \tag{24}$$

where

$$\breve{\boldsymbol{\Psi}} = \frac{1}{rn}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi_{\delta i}^{\text{app}*}}\left[\frac{S^{*(2)}(Y_i^*,\breve{\boldsymbol{\beta}})}{S^{*(0)}(Y_i^*,\breve{\boldsymbol{\beta}})} - \left\{\frac{S^{*(1)}(Y_i^*,\breve{\boldsymbol{\beta}})}{S^{*(0)}(Y_i^*,\breve{\boldsymbol{\beta}})}\right\}^{\otimes 2}\right],$$

$$\breve{\boldsymbol{\Gamma}} = \frac{1}{r^2n^2}\sum_{i=1}^{r}\frac{1}{\{\pi_{\delta i}^{\text{app}*}\}^2}\left[\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}^{0*}(t,\breve{\boldsymbol{\beta}})\}d\hat{M}_i^*(t,\breve{\boldsymbol{\beta}})\right]^{\otimes 2},$$

$S^{*(k)}(Y_i^*,\breve{\boldsymbol{\beta}}) = (rn)^{-1}\sum_{j=1}^{r}\pi_{\delta j}^{\text{app}*-1}I(Y_j^* \geq Y_i^*)\mathbf{X}_j^{*\otimes k}\exp(\breve{\boldsymbol{\beta}}'\mathbf{X}_j^*)$ for $k = 0, 1, 2$, $\bar{\mathbf{X}}^{0*}(t,\breve{\boldsymbol{\beta}})$ is defined in (16), and $d\hat{M}_i^*(t,\breve{\boldsymbol{\beta}}) = dN_i^*(t) - I(Y_i^* \geq t)\exp(\breve{\boldsymbol{\beta}}'\mathbf{X}_i^*)d\hat{\Lambda}_0^{\text{UNIF}}(t,\breve{\boldsymbol{\beta}})$. We will assess the performance of formula (24) by numerical simulations.

Lastly, the cumulative hazard function $\Lambda_0(t)$ plays an important role for predicting the survival probability of an individual in many biomedical applications. It has an expression of $S(t|\mathbf{X}) = P(T > t|\mathbf{X}) = \exp\{-\exp(\boldsymbol{\beta}'\mathbf{X})\Lambda_0(t)\}$ with Cox's model. The Breslow estimator $\hat{\Lambda}_0(t,\boldsymbol{\beta})$ is the maximum likelihood estimator of $\Lambda_0(t)$, where

$$\begin{aligned}\hat{\Lambda}_0(t,\boldsymbol{\beta}) &= \sum_{i=1}^{n}\int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n} I(Y_j \geq s)\exp(\boldsymbol{\beta}'\mathbf{X}_j)} \\ &= \sum_{i=1}^{n}\frac{\Delta_i I(Y_i \leq t)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)\exp(\boldsymbol{\beta}'\mathbf{X}_j)}.\end{aligned} \tag{25}$$

Based on the subsample estimator $\breve{\boldsymbol{\beta}}$, it is easy to obtain a Breslow type estimator $\hat{\Lambda}_0(t,\breve{\boldsymbol{\beta}})$ by replacing $\boldsymbol{\beta}$ with $\breve{\boldsymbol{\beta}}$ in (25), i.e., an estimated cumulative hazard function is based on the entire dataset but with a subsampling-based estimator of $\boldsymbol{\beta}$. As pointed out by an reviewer, the computation burden of this Breslow estimator is not heavy if the observed failure times are sorted in an increasing order, because it has an explicit expression and no optimization process is required. The Breslow type estimator has a computation complexity $O(n\log(n)) + O(n)$, where $O(n\log(n))$ is due to the sorting of the full data and $O(n)$ is from the summation.

5 Numerical Studies

5.1 Simulation

In this section, we conduct simulations to evaluate the performance of our proposed subsampling method. We generate failure times T_i's from Cox's model with a baseline hazard function $\lambda_0(t) = 0.5t$ and the true parameter $\boldsymbol{\beta}_0 = (-1, -0.5, 0, 0.5, 1)'$ with $p = 5$. We consider four settings for the covariate $\mathbf{X}_i = (X_{i1}, \cdots, X_{i5})'$.

Case I : components of $\mathbf{X}_i$ are independent uniform random variables over $(-1, 1)$.

Case II: $\mathbf{X}_i$ follows $0.5N(-\mathbf{1}, \boldsymbol{\Upsilon}) + 0.5N(\mathbf{1}, \boldsymbol{\Upsilon})$, where $\Upsilon_{jk} = 0.5^{|j-k|}$, i.e., $\mathbf{X}_i$ follows a mixture of two multivariate normal distributions.

Case III: components of $\mathbf{X}_i$ are independent exponential random variables with probability density function $f(x) = 2e^{-2x}I(x > 0)$.

Case IV: $\mathbf{X}_i$ follows a multivariate t distribution with degree of freedom 10, mean zero and covariance matrix $\boldsymbol{\Upsilon}$ where $\Upsilon_{jk} = 0.5^{|j-k|}$.

The censoring times C_i's are independently generated from a uniform distribution over $(0, c_0)$ with c_0 being chosen so that the censoring rate (CR) is about 20% and 60%, respectively. Results were calculated based on 1000 replications of the simulation. We set the full data sample size to $n = 10^6$, and consider the subsample sizes of $r = 400, 600, 800$, and 1000, respectively.

We evaluate the proposed method using the empirical mean squared error (MSE), defined as

$$\text{MSE}(\breve{\boldsymbol{\beta}}) = \frac{1}{1000}\sum_{b=1}^{1000} \|\breve{\boldsymbol{\beta}}^{(b)} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\|^2, \tag{26}$$

where $\breve{\boldsymbol{\beta}}^{(b)}$ is the estimate from the bth subsample with $\delta = 0.1$.

We studied the effect of the pilot subsample size r_0 first. Table 1 presents the MSEs of subsampling-based estimator by varying the pilot subsample size $r_0 = 300$, 400 and 500. We see that the influence of r_0 on $\breve{\boldsymbol{\beta}}$ is not significant if we choose a reasonably large pilot subsample. Hence, we suggest to use r_0= 300 for settings similar to the simulation setup. Users may adopt a larger pilot subsample if the dimension of the problem is higher or if the censoring rate is higher.

Table 1: The MSE of subsampling-based estimator with different pilot subsample size r_0.

		CR=20%				CR=60%			
	r_0	$r = 400$	$r = 600$	$r = 800$	$r = 1000$	$r = 400$	$r = 600$	$r = 800$	$r = 1000$
Case I	300	0.0320	0.0215	0.0159	0.0130	0.0590	0.0392	0.0279	0.0229
	400	0.0321	0.0211	0.0163	0.0123	0.0586	0.0385	0.0279	0.0220
	500	0.0322	0.0203	0.0161	0.0124	0.0573	0.0365	0.0279	0.0219
Case II	300	0.0340	0.0214	0.0165	0.0127	0.0592	0.0374	0.0284	0.0221
	400	0.0332	0.0219	0.0164	0.0128	0.0599	0.0381	0.0299	0.0227
	500	0.0338	0.0222	0.0167	0.0125	0.0593	0.0381	0.0281	0.0224
Case III	300	0.0418	0.0272	0.0206	0.0164	0.0804	0.0510	0.0379	0.0294
	400	0.0401	0.0262	0.0191	0.0155	0.0800	0.0515	0.0367	0.0310
	500	0.0392	0.0257	0.0187	0.0149	0.0769	0.0500	0.0359	0.0290
Case IV	300	0.0167	0.0108	0.0083	0.0065	0.0226	0.0157	0.0108	0.0086
	400	0.0152	0.0102	0.0080	0.0060	0.0241	0.0150	0.0112	0.0088
	500	0.0151	0.0100	0.0075	0.0060	0.0224	0.0144	0.0107	0.0084

Next we investigated how the MSEs behave as a function of δ. From the expression of $\boldsymbol{\pi}_\delta^{\text{app}} = \{\pi_{\delta i}^{\text{app}}\}_{i=1}^n$ given in (19), we know the sampling distribution $\boldsymbol{\pi}_\delta^{\text{app}}$ is close to the optimal subsampling distribution when δ is small, while it is close to the uniform subsampling distribution if δ is close to 1. In Table 2, we present the MSEs of the subsampling estimator for different values of δ: 0, 0.1, 0.3, and 0.5. It is seen that $\delta = 0.1$ produce the best result most frequently among. Hence, we recommend $\delta = 0.1$ when implementing our method in practical applications with similar settings.

Table 2: The MSE of subsampling-based estimator with different mixing rate δ.

		CR=20%				CR=60%			
	r	$\delta=0$	$\delta=0.1$	$\delta=0.3$	$\delta=0.5$	$\delta=0$	$\delta=0.1$	$\delta=0.3$	$\delta=0.5$
Case I	400	0.0326	0.0320	0.0322	0.0346	0.0596	0.0590	0.0613	0.0670
	600	0.0216	0.0215	0.0223	0.0230	0.0393	0.0392	0.0408	0.0443
	800	0.0163	0.0159	0.0166	0.0176	0.0290	0.0279	0.0290	0.0317
	1000	0.0131	0.0130	0.0132	0.0141	0.0222	0.0229	0.0243	0.0263
Case II	400	0.0347	0.0340	0.0357	0.0376	0.0594	0.0592	0.0623	0.0691
	600	0.0219	0.0214	0.0225	0.0242	0.0372	0.0374	0.0394	0.0432
	800	0.0164	0.0165	0.0170	0.0180	0.0284	0.0284	0.0302	0.0335
	1000	0.0127	0.0127	0.0134	0.0141	0.0222	0.0221	0.0231	0.0248
Case III	400	0.0418	0.0418	0.0437	0.0463	0.0924	0.0804	0.0837	0.0891
	600	0.0272	0.0272	0.0281	0.0300	0.0535	0.0510	0.0537	0.0582
	800	0.0209	0.0206	0.0211	0.0223	0.0359	0.0379	0.0404	0.0431
	1000	0.0164	0.0164	0.0169	0.0177	0.0295	0.0294	0.0306	0.0336
Case IV	400	0.0174	0.0167	0.0166	0.0178	0.0241	0.0226	0.0232	0.0256
	600	0.0106	0.0108	0.0106	0.0112	0.0154	0.0157	0.0157	0.0169
	800	0.0082	0.0083	0.0083	0.0084	0.0111	0.0108	0.0113	0.0128
	1000	0.0065	0.0065	0.0065	0.0067	0.0091	0.0086	0.0089	0.0097

We considered the proposed subsampling method with approximated optimal subsampling probabilities in Algorithm 1 with $\delta = 0.1$ ("Lopt estimator"), and the uniform subsampling method ("UNIF estimator"). We calculated the empirical biases (Bias), the mean estimated standard errors (SE) calculated using (24), the empirical standard errors (ESE), and the empirical 95% coverage probability (CP) towards the true parameter $\boldsymbol{\beta}$. The pilot sample size is $r_0 = 300$.

We present the estimation results about β_1 in Tables 3 and 4, indicating that both Lopt and UNIF estimators are asymptotically unbiased. The SE and ESE are similar and the coverage probabilities are close to the nominal level, which support the asymptotic normality of the proposed estimator and demonstrate that the subsample-based variance-

Table 3: Simulation results of the subsample estimator $\breve{\beta}_1$ with CR = 20%‡.

		Lopt				UNIF			
	r	Bias	ESE	SE	CP	Bias	ESE	SE	CP
Case I	400	-0.0021	0.0860	0.0901	0.960	-0.0069	0.1121	0.1148	0.947
	600	0.0008	0.0717	0.0728	0.947	-0.0033	0.0896	0.0933	0.961
	800	-0.0015	0.0603	0.0629	0.958	-0.0041	0.0787	0.0810	0.963
	1000	0.0021	0.0551	0.0559	0.948	-0.0039	0.0672	0.0722	0.961
Case II	400	-0.0008	0.0836	0.0844	0.946	-0.0156	0.1032	0.1039	0.952
	600	-0.0016	0.0662	0.0681	0.949	-0.0159	0.0857	0.0847	0.950
	800	-0.0037	0.0594	0.0588	0.949	-0.0149	0.0769	0.0731	0.935
	1000	-0.0020	0.0538	0.0523	0.944	-0.0090	0.0672	0.0649	0.939
Case III	400	-0.0025	0.1047	0.0997	0.937	-0.0109	0.1341	0.1297	0.943
	600	-0.0002	0.0827	0.0808	0.936	-0.0029	0.1132	0.1048	0.944
	800	-0.0029	0.0704	0.0696	0.938	-0.0037	0.0969	0.0906	0.939
	1000	-0.0015	0.0615	0.0621	0.952	0.0008	0.0872	0.0808	0.938
Case IV	400	0.0011	0.0602	0.0625	0.953	-0.0093	0.0763	0.0758	0.942
	600	-0.0010	0.0496	0.0505	0.958	-0.0072	0.0600	0.0616	0.955
	800	0.0015	0.0421	0.0433	0.953	-0.0057	0.0505	0.0531	0.950
	1000	-0.0003	0.0369	0.0386	0.966	-0.0058	0.0461	0.0473	0.953

‡ The (Bias, SE) of full data MPL estimator $\hat{\beta}_1$ is Case I: (Bias, SE) = (0.0001, 0.0021), Case II: (Bias, SE) = (-0.0036, 0.0020), Case III: (Bias, SE) = (0.0004, 0.0027), Case IV: (Bias, SE) = (-0.0001, 0.0014).

Table 4: Simulation results of the subsample estimator $\breve{\beta}_1$ with CR = 60%‡.

		Lopt				UNIF			
	r	Bias	ESE	SE	CP	Bias	ESE	SE	CP
Case I	400	0.0074	0.1119	0.1219	0.971	-0.0094	0.1509	0.1717	0.965
	600	0.0066	0.0926	0.0984	0.966	-0.0142	0.1221	0.1404	0.979
	800	0.0036	0.0769	0.0849	0.967	-0.0057	0.1089	0.1216	0.968
	1000	0.0056	0.0708	0.0758	0.962	0.0001	0.0927	0.1083	0.979
Case II	400	-0.0086	0.1032	0.1012	0.947	-0.0242	0.1398	0.1409	0.952
	600	-0.0033	0.0783	0.0813	0.961	-0.0141	0.1164	0.1139	0.945
	800	-0.0046	0.0728	0.0699	0.937	-0.0053	0.0979	0.0984	0.948
	1000	-0.0007	0.0623	0.0622	0.950	-0.0088	0.0869	0.0878	0.945
Case III	400	-0.0033	0.1562	0.1647	0.959	-0.0282	0.2186	0.2190	0.949
	600	-0.0004	0.1244	0.1322	0.963	-0.0199	0.1845	0.1777	0.942
	800	-0.0015	0.1113	0.1141	0.962	-0.0075	0.1497	0.1533	0.948
	1000	0.0027	0.0972	0.1015	0.958	-0.0085	0.1358	0.1364	0.954
Case IV	400	0.0037	0.0685	0.0672	0.939	-0.0122	0.0997	0.0975	0.952
	600	-0.0002	0.0585	0.0540	0.922	-0.0026	0.0784	0.0774	0.944
	800	0.0022	0.0469	0.0465	0.948	-0.0061	0.0701	0.0663	0.936
	1000	0.0018	0.0422	0.0414	0.944	-0.0047	0.0599	0.0588	0.945

‡ The (Bias, SE) of full data MPL estimator $\hat{\beta}_1$ is Case I: (Bias, SE) = (0.0026, 0.0029), Case II: (Bias, SE) = (-0.0013, 0.0028), Case III: (Bias, SE) = (0.0012, 0.0042), Case IV: (Bias, SE) = (0.0021, 0.0019).

covariance matrix estimator in (24) is accurate. Both subsample estimators get better as the sampling size r increases. In addition, the Bias and ESE of the Lopt estimator are much smaller than those of UNIF estimator with the same subsample size r. This agrees with the conclusion in Theorem 3. Results for other regression coefficients are similar and thus are omitted. Furthermore, we report the Bias and SE of the full data MPL estimator towards the true parameter in the footnotes of Tables 3 and 4.

We calculated the five-number summary statistics of π_δ^{app}'s for the censored and uncensored observations separately to demonstrate the impact of censoring on optimal subsampling probabilities numerically. Table 5 reports the results, including the Minimum, Lower-hinge (the first quartile), Median, Upper-hinge (the third quartile), and the Maximum. Uncensored observations have larger subsampling probabilities than censored observations in general, i.e., uncensored observations are more likely to be selected into a subsample by our subsampling method.

Furthermore, we assessed the computational efficiency of our optimal subsampling method. For comparison, we also considered the UNIF, full data estimator and SGD estimator (Tarkhan and Simon, 2020), where the full data estimator was calculated with the R function `coxph` and the SGD estimator was obtained by the R function `bigSurvSGD` (using default settings). The computations were carried out using R (R Core Team, 2021) on a desktop computer with 64GB memory. We restricted the calculations to access one CPU core and recorded the average CPU time from 100 repetitions. Table 6 reports the results for Case I, where the subsample size is $r = 1000$. The computational speed of the Lopt estimator is much faster than that of the full data estimator with `coxph`. The computational burden of the full data method gets heavier as the increase of full data sample size. In other words, subsampling is desirable in Cox's regression because it reduce the computational cost significantly. The UNIF estimator is faster to compute than the Lopt estimator, because it does not need the step of calculating the sampling probabilities, but it has a lower estimation efficiency as we have seen in previous results. Note that the SGD estimator is slower than the full data estimator in terms of computation speed. We point out that the main aim of the SGD estimator was to deal with large datasets where `coxph` cannot be used (due to out-of-memory issues) rather than speeding up the calculations. In Table 7, we present more

Table 5: The five-number summary statistics for $\pi_{\delta}^{\text{app}}$ $(\times 10^6)^{\ddagger}$.

			Minimum	Lower-hinge	Median	Upper-hinge	Maximum
CR=20%	Case I	$\pi_{\delta,c}^{\text{app}}$	0.1000	0.1894	0.4307	1.0141	28.3688
		$\pi_{\delta,u}^{\text{app}}$	0.1194	0.5507	0.8835	1.2732	21.5598
	Case II	$\pi_{\delta,c}^{\text{app}}$	0.1000	0.1621	0.3547	0.8058	28.1762
		$\pi_{\delta,u}^{\text{app}}$	0.1182	0.5451	0.8588	1.3208	40.5288
	Case III	$\pi_{\delta,c}^{\text{app}}$	0.1000	0.1698	0.3853	0.8781	58.9218
		$\pi_{\delta,u}^{\text{app}}$	0.1240	0.5132	0.8090	1.2228	93.0236
	Case IV	$\pi_{\delta,c}^{\text{app}}$	0.1000	0.1854	0.4018	0.8461	33.6698
		$\pi_{\delta,u}^{\text{app}}$	0.1430	0.5188	0.7948	1.2636	57.4029
CR=60%	Case I	$\pi_{\delta,c}^{\text{app}}$	0.1000	0.1759	0.4512	0.9822	21.7717
		$\pi_{\delta,u}^{\text{app}}$	0.1475	0.7905	1.2350	1.7154	24.5950
	Case II	$\pi_{\delta,c}^{\text{app}}$	0.1000	0.1801	0.3659	0.7793	27.9329
		$\pi_{\delta,u}^{\text{app}}$	0.1163	0.8631	1.3230	1.9279	43.5112
	Case III	$\pi_{\delta,c}^{\text{app}}$	0.1000	0.1607	0.3527	0.7858	54.5352
		$\pi_{\delta,u}^{\text{app}}$	0.1181	0.8581	1.2927	1.8870	61.6812
	Case IV	$\pi_{\delta,c}^{\text{app}}$	0.1000	0.1259	0.2200	0.4839	54.4077
		$\pi_{\delta,u}^{\text{app}}$	0.1724	0.8910	1.4241	2.1763	167.3048

‡ $\pi_{\delta,c}^{\text{app}}$ and $\pi_{\delta,u}^{\text{app}}$ denote the mixed approximated optimal subsampling probabilities for censored and uncensored samples, respectively; $\delta = 0.1$.

comparisons between Lopt and UNIF methods when the CPU computation times are similar. It is seen that the Lopt and UNIF may have similar estimation efficiency using similar CPU times. However, the UNIF uses larger sample sizes and thus larger memory. The optimal subsampling method achieves the same estimation efficiency with less computing resources in these scenarios.

Table 6: The CPU time for Case I with $r = 1000$ (in seconds)†.

		n		
	Methods	10^6	5×10^6	10^7
CR = 20%	UNIF	0.17	0.25	0.34
	Lopt	0.39	1.22	2.28
	full data	6.75	45.71	100.65
	SGD	94.19	603.81	1294.70
CR = 60%	UNIF	0.10	0.17	0.27
	Lopt	0.32	1.17	2.33
	full data	6.16	45.87	99.54
	SGD	99.50	530.84	1112.85

† "full data": calculated with R function `coxph`; "SGD": calculated with R function `bigSurvSGD`.

Finally, we compared the two subsampling probabilities derived from the L-optimality criterion (Lopt) and A-optimality criterion (Aopt), respectively. By Remark 4, the optimal subsampling probabilities under the A-optimality criterion are obtained by minimizing $\mathrm{tr}(\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is given in (11). Using a similar deduction as that of (18), we can obtain the approximated optimal subsampling probabilities under the A-optimality criterion:

$$\pi_i^{\mathrm{Aopt}} = \frac{\|\boldsymbol{\Psi}^{0*-1} \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\} d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0)\|}{\sum_{j=1}^n \|\boldsymbol{\Psi}^{0*-1} \int_0^\tau \{\mathbf{X}_j - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\} d\hat{M}_j(t, \tilde{\boldsymbol{\beta}}_0)\|}, \quad i = 1, \cdots, n, \tag{27}$$

where

$$\boldsymbol{\Psi}^{0*} = \frac{1}{r_0} \sum_{i=1}^{r_0} \Delta_i^{0*} \left[\frac{S^{0*(2)}(Y_i^{0*}, \tilde{\boldsymbol{\beta}}_0)}{S^{0*(0)}(Y_i^{0*}, \tilde{\boldsymbol{\beta}}_0)} - \left\{ \frac{S^{0*(1)}(Y_i^{0*}, \tilde{\boldsymbol{\beta}}_0)}{S^{0*(0)}(Y_i^{0*}, \tilde{\boldsymbol{\beta}}_0)} \right\}^{\otimes 2} \right],$$

Table 7: Comparisons of CPU times between Lopt and UNIF (in seconds)†.

		CR=20%			CR=60%		
		CPU	r	MSE	CPU	r	MSE
Case I	Lopt	0.4304	1000	0.01215	0.3834	1100	0.02047
	UNIF	0.4823	1700	0.01113	0.3631	2100	0.01930
Case II	Lopt	0.3670	800	0.01515	0.4591	1400	0.01602
	UNIF	0.3415	1400	0.01519	0.4568	2300	0.01757
Case III	Lopt	0.3842	850	0.01667	0.4972	1400	0.02158
	UNIF	0.4465	1500	0.01675	0.5259	2300	0.02424
Case IV	Lopt	0.4749	1100	0.00575	0.4097	1100	0.00819
	UNIF	0.5260	1650	0.00565	0.3726	2000	0.00882

† "CPU" denotes average CPU time from 100 repetitions; the full data size is $n = 10^6$.

and $S^{0*(k)}(Y_i^{0*}, \tilde{\boldsymbol{\beta}}_0) = (r_0)^{-1} \sum_{j=1}^{r_0} I(Y_j^{0*} \geq Y_i^{0*}) \mathbf{X}_j^{0*\otimes k} \exp(\tilde{\boldsymbol{\beta}}_0' \mathbf{X}_j^{0*})$ for $k = 0, 1, 2$. The corresponding mixed subsampling probabilities with A-optimality criterion are

$$\pi_{\delta i}^{\text{Aopt}} = (1 - \delta)\pi_i^{\text{Aopt}} + \frac{\delta}{n}, \quad i = 1, \cdots, n.$$

In Figures 1 and 2, we report the empirical MSEs of subsample estimators with Lopt, Aopt and UNIF methods, where $\delta = 0.1$. The results indicates that Lopt and Aopt have similar performance. In addition, the UNIF has the largest MSE compared with Lopt and Aopt methods. It is clear that the speed of Aopt is slower than Lopt, because there is an additional term $\boldsymbol{\Psi}^{0*}$ involved in (27). As a summary, we recommend using the Lopt for our subsampling method in practical applications.

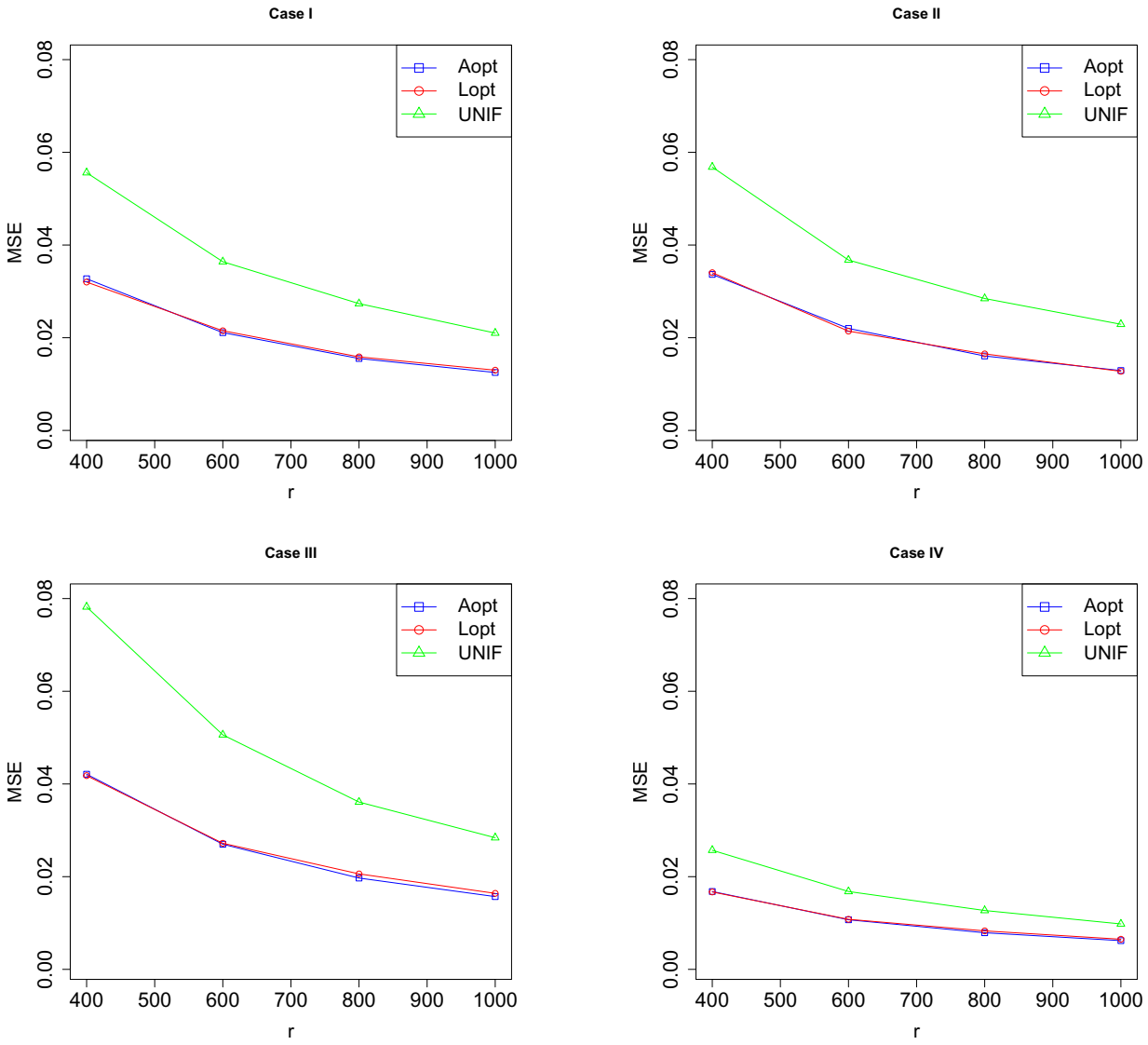

Figure 1: The MSEs for different subsampling methods with CR= 20%.

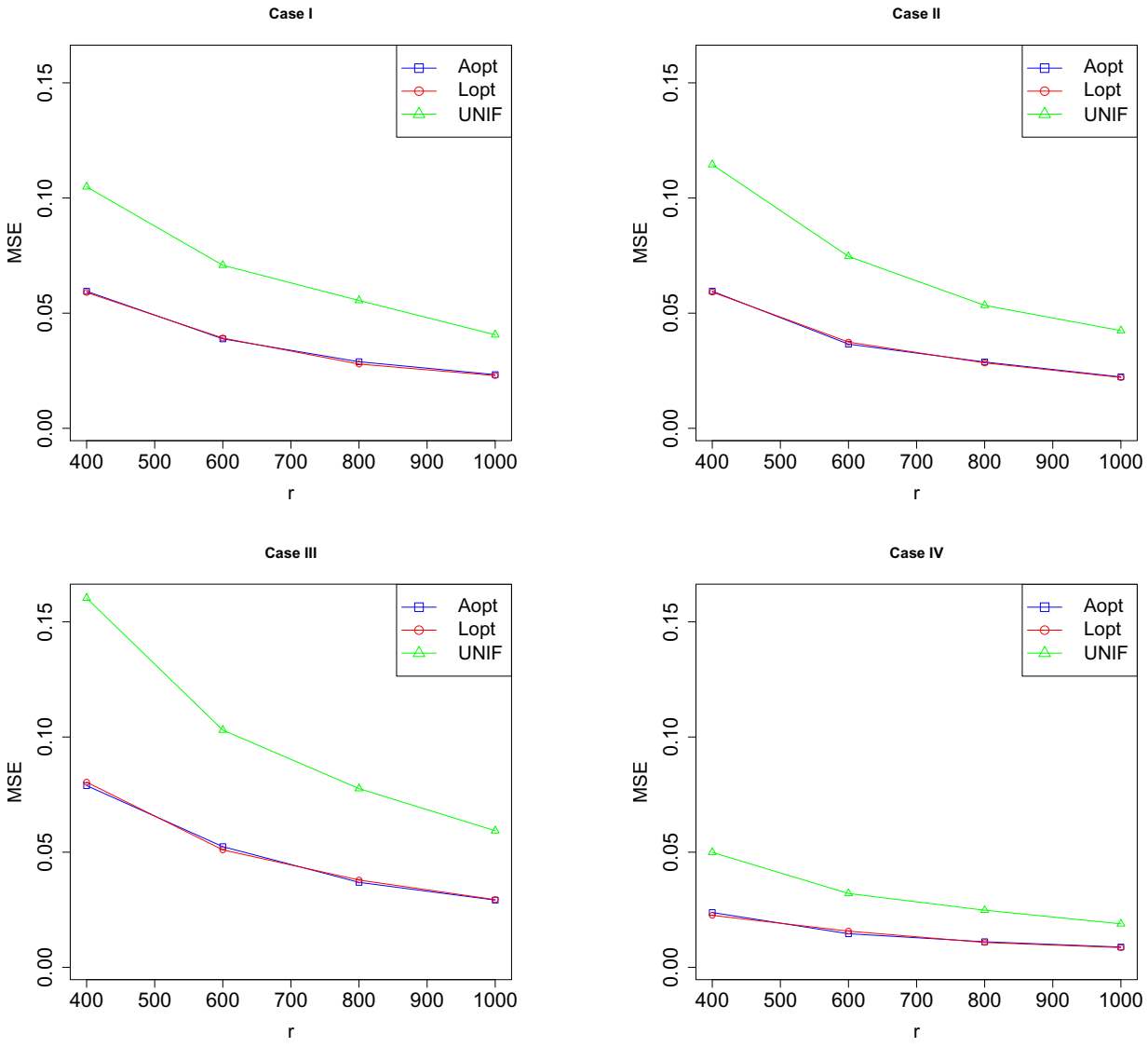

Figure 2: The MSEs for different subsampling methods with CR= 60%.

5.2 Application

In this section, we apply the proposed subsampling method to a real-world data example about the USA airline, where the dataset is publicly available from DVN (2008). We are interested in analysing those arrival delayed airlines, where there were totally 57,729,435

arrival delayed commercial flights within the USA from October 1987 to April 2008. For the ith airline, the failure time T_i is defined as the delayed time from scheduled arrival time to actual arrival time (in minutes). Among those 57,729,435 arrival delayed flights, 33,142,872 subjects experienced an actual arrival within 15 minutes (the delayed arrival time is less than 15 minutes). The censoring rate is about 42.6%. For analysis, the risk factors $\mathbf{X}_i = (X_{i1}, X_{i2})'$ in Cox's model are departure status (departure on time or ahead of schedule = 0 and departure delayed = 1) and distance (continuous, in thousands of miles), respectively.

Table 8: Estimation results for the airline arrival delay data with one subsample.

		Lopt			UNIF		
	β	Est	SE	CI	Est	SE	CI
$r = 400$	β_1	-1.0009	0.1303	(-1.2562, -0.7456)	-1.2038	0.1353	(-1.4691, -0.9385)
	β_2	-0.3188	0.1194	(-0.5527, -0.0848)	-0.2471	0.1296	(-0.5009, 0.0069)
$r = 600$	β_1	-1.0633	0.1036	(-1.2663, -0.8602)	-1.0865	0.1298	(-1.3409, -0.8322)
	β_2	-0.2341	0.0869	(-0.4045, -0.0637)	-0.3207	0.1195	(-0.5549, -0.0865)
$r = 800$	β_1	-1.1121	0.0826	(-1.2741, -0.9502)	-1.0575	0.1001	(-1.2536, -0.8613)
	β_2	-0.3418	0.0689	(-0.4769, -0.2067)	-0.2674	0.1031	(-0.4694, -0.0654)
$r = 1000$	β_1	-1.1099	0.0745	(-1.2559, -0.9638)	-1.1147	0.0857	(-1.2826, -0.9468)
	β_2	-0.2498	0.0598	(-0.3669, -0.1326)	-0.2505	0.0882	(-0.4233, -0.0777)

For comparison, we calculated the full data estimator $\hat{\boldsymbol{\beta}}_{\text{MPL}} = (-1.1301, -0.2396)'$ with `coxph`, and the corresponding SEs are 0.00035 and 0.00034, respectively. Hence, departure status and distance have negative effects on the hazard rate of airline's arrival. That is to say, it is expected that a departure delayed airline with long distance would owns a longer arrival delay. In addition, we calculated the Lopt estimator with $\delta = 0.1$ and the UNIF estimator. We present the results on the subsampling-based estimator (Est), the SE and the 95% confidence interval (CI) based on one subsample in Table 8. Both Lopt and UNIF estimators are close to $\hat{\boldsymbol{\beta}}_{\text{MPL}}$, especially when the subsample size is large (e.g. $r = 1000$). The

SE of the Lopt estimator is much smaller than that of the UNIF estimator, which supports the theoretical conclusion in Theorem 2. To further validate the usefulness of our method, we report the Bias, SE and ESE of the subsample-based estimates based on 1000 subsamples in Table 9, where the Bias denotes the average bias of the subsampling estimator with respect to the full-data MPL estimator. Both subsample-based estimates are unbiased, and the SE is close to ESE indicating that the estimated variance-covariance matrix in (24) works well. The results in Table 9 again demonstrate that the Lopt estimator is more efficient than the UNIF estimator. Finally, the full data MPL estimator with `coxph` needs 265.58 seconds, where the computer is the same as that used in the simulation. For $r = 1000$, the Lopt only requires 9.03 seconds to output the subsample estimators and their SEs (UNIF needs 1.08 seconds). i.e., the Lopt method has a much faster computation efficiency than the `coxph` when we face with large-scale survival dataset in practice.

Table 9: The Bias and (ESE, SE) for subsample estimates in real data.

	β	Lopt	UNIF
$r = 400$	β_1	0.0030 (0.1228, 0.1289)	-0.0059 (0.1351, 0.1466)
	β_2	0.0022 (0.1033, 0.1094)	-0.0035 (0.1346, 0.1419)
$r = 600$	β_1	0.0003 (0.1017, 0.1016)	-0.0055 (0.1129, 0.1087)
	β_2	-0.0018 (0.0833, 0.0804)	-0.0019 (0.1103, 0.1061)
$r = 800$	β_1	0.0019 (0.0827, 0.0854)	-0.0026 (0.0982, 0.0932)
	β_2	-0.0002 (0.0680, 0.0692)	-0.0008 (0.0938, 0.0912)
$r = 1000$	β_1	0.0002 (0.0799, 0.0781)	-0.0009 (0.0859, 0.0886)
	β_2	-0.0027 (0.0616, 0.0657)	-0.0030 (0.0849, 0.0859)

6 Concluding Remarks

In this paper, we have studied the statistical properties of a general subsampling algorithm for Cox's model with massive survival data. We provided the optimal subsampling probabilities, and established asymptotic properties of the two-step subsample-based parameter

estimator conditional on the full data. Extensive simulations and a real data example have been used to validate the practical usefulness of our method. Note that the proposed approach is appropriate when the outcome of interest is common and the dataset includes enough observed events, i.e., our subsample method is suitable for the regular time-to-event data. Faced with massive survival datasets with rare events, Keret and Gorfine (2020) proposed a novel and interesting subsampling procedure to deal with computational challenges in massive data Cox regression. Their procedure is based on counting process type score function, while we derive the asymptotic distribution of subsample-based estimator from the martingale-type subsample score function. Keret and Gorfine (2020) avoided the need of estimating the cumulative baseline hazard function for the optimal subsampling probabilities, which was in contrast to our approach. In addition, our proposed subsample estimator approximates the maximizer of the full data partial likelihood, and the approximation error is not significantly affected by the correctness of the Cox model. In other words, if the proportional hazards assumption is violated, the subsample estimator is still close to the full data estimator, but the full data estimator may not be the best estimator any more.

There are four important topics for further research. First, it is desirable to investigate how the proportional hazards assumption can be adequately checked based on subsamples. Second, the numerator of π_i^{app} only involves the ith subject and the pilot subsample, which sheds light on the feasibility of distributed or parallel algorithms when calculating the optimal sampling probabilities. For example, by splitting the full data into multiple blocks, it is possible to calculate the terms $\| \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\} d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) \|$ with distributed computing environments. In this case, the computational speed of optimal subsampling method would be significantly improved. Third, in many practical applications, observed data are often corrupted by outliers (Meng *et al.*, 2021). Therefore, it is interesting to study the optimal subsampling method for Cox's model with the presence of outliers. Forth, the tuning parameter δ perform well with a value of 0.1 in our numerical results, but there is not theoretically justification to show that this value will work in all scenarios. How to select δ attentively requires further investigations.

Supplementary Materials

Supplement The supplementary PDF file contains proofs of all the theoretical results in this paper.

R codes The zip file contains the R codes used to perform the subsampling methods described in the article, where the readme file describes details about the codes.

Acknowledgement

The authors would like to thank the Editor, the Associate Editor and two reviewers for their constructive and insightful comments that greatly improved the manuscript. We also thank Aliasghar Tarkhan for providing some helpful comments on the usage of R package "bigSurvSGD". The work of Wang was supported by National Science Foundation (NSF), USA grant CCF-2105571. The work of Sun was supported in part by the National Natural Science Foundation of China (Grant No. 12171463).

Conflict of Interest

The authors declare that there are no conflicts of interest.

References

(2008). Data Expo 2009: Airline on time data.

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021). Optimal subsampling algorithms for big data regressions. *Statistica Sinica* **31**, 749–772.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 4, 1100–1120.

Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press.

Bai, Y., Li, C., Lin, Z., Wu, Y., Miao, Y., Liu, Y., and Xu, Y. (2021). Efficient data loader for fast sampling-based gnn training on large graphs. *IEEE Transactions on Parallel and Distributed Systems* **10**, 2541–2556.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics* **46**, 3, 1352–1382.

Chen, X., Cheng, J. Q., and Xie, M. (2021b). Divide-and-conquer methods for big data analysis. *arXiv:2102.10771v1* .

Chen, X., Liu, W., and Zhang, Y. (2022). First-order newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association* **117**, 1858–1874.

Cox, D. R. (1972). Regression models and life-tables (with discussions). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 2, 269–276.

Fan, J., Guo, Y., and Wang, K. (2021). Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1969238* .

Fang, E. X., Ning, Y., and Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 5, 1415–1437.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. New York: John Wiley and Sons.

Han, L., Tan, K. M., Yang, T., and Zhang, T. (2020). Local uncertainty sampling for large-scale multiclass logistic regression. *The Annals of Statistics* **48**, 1770–1788.

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

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. (2013). Oracle inequalities for the lasso in the cox model. *The Annals of Statistics* **41**, 3, 1142–1165.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* **114**, 526, 668–681.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. New Jersey: Wiley-Interscience.

Keret, N. and Gorfine, M. (2020). Optimal cox regression subsampling procedure with rare events. *arXiv:2012.02122v1* .

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

Kleinbaum, D. G. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text*. New York: Springer Science+Business Media, Inc.

Lee, S. and Ng, S. (2020). An econometric perspective on algorithmic subsampling. *Annual Review of Economics* **12**, 45–80.

Li, R., Chang, C., Justesen, J. M., Tanigawa, Y., Qiang, J., Hastie, T., Rivas, M. A., and Tibshirani, R. (2020). Fast lasso method for large-scale and ultrahigh-dimensional cox model with applications to uk biobank. *Biostatistics* DOI: 10.1093/biostatistics/kxaa038.

Li, T. and Meng, C. (2021). Modern subsampling methods for large-scale least squares regression. *International Journal of Cyber-Physical Systems* **2**, 1–28.

Lin, L., Li, W., and Lu, J. (2020). Unified rules of renewable weighted sums for various online updating estimations. *arXiv:2008.08824v1* .

Liu, H., You, J., and Cao, J. (2021). Functional L-optimality subsampling for massive data. *arXiv:2104.03446v1* .

Luo, L. and Song, P. X. (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of The Royal Statistical Society Series B* **82**, 69–97.

Luo, L., Zhou, L., and Song, P. X.-K. (2022). Real-time regression analysis of streaming clustered data with possible abnormal data batches. *Journal of the American Statistical Association* DOI:10.1080/01621459.2022.2026778.

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

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (2021). Lowcon: A design based subsampling approach in a misspecified linear model. *Journal of Computational and Graphical Statistics* **30**, 694–708.

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Shi, C., Lu, W., and Song, R. (2018). A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association* **113**, 524, 1698–1709.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* **39**, 1–13.

Tarkhan, A. and Simon, N. (2020). Bigsurvsgd: Big survival data analysis via stochastic gradient descent. *arXiv:2003.00116v1* .

Therneau, T. M. (2021). *A Package for Survival Analysis in R*. R package version 3.2-13.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–160.

Volgushev, S., Chao, S.-K., and Cheng, G. (2019). Distributed inference for quantile regression processes. *The Annals of Statistics* **47**, 3, 1634–1662.

Wang, C., Chen, M.-H., Schifano, E., Wu, J., and Yan, J. (2016). Statistical methods and computing for big data. *Statistics and Its Interface* **9**, 399–414.

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

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

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

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

Wang, J., Zou, J., and Wang, H. (2022a). Sampling with replacement vs poisson sampling: a comparative study in optimal subsampling. *IEEE Transactions on Information Theory* **68**, 6605–6630.

Wang, K., Wang, H., and Li, S. (2022b). Renewable quantile regression for streaming datasets. *Knowledge-Based Systems* DOI: 10.1016/j.knosys.2021.107675.

Wang, T. and Zhang, H. (2022). Optimal subsampling for multiplicative regression with massive data. *Statistica Neerlandica* **76**, 418–449.

Wang, Y., Hong, C., Palmer, N., Di, Q., Schwartz, J., Kohane, I., and Cai, T. (2021). A fast divide-and-conquer sparse cox regression. *Biostatistics* **22**, 381–401.

Wu, J., Chen, M. H., Schifano, E. D., and Yan, J. (2021). Online updating of survival analysis. *Journal of Computational and Graphical Statistics* **30**, 1209–1223.

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

Xu, J., Ying, Z., and Zhao, N. (2020). Scalable estimation and inference with large-scale or online survival data. *arXiv preprint arXiv:2001.01434* .

Xue, Y., Wang, H., Yan, J., and Schifano, E. D. (2019). An online updating approach for testing the proportional hazards assumption with streams of survival data. *Biometrics* **76**, 1, 171–182.

Yang, Z., Wang, H., and Yan, J. (2022). Optimal subsampling for parametric accelerated failure time models with massive survival data. *Statistics in Medicine* **41**, 5421–5431.

Yao, Y. and Wang, H. (2019). Optimal subsampling for softmax regression. *Statistical Papers* **60**, 585–599.

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

Yao, Y., Zou, J., and Wang, H. (2021). Optimal poisson subsampling for softmax regression. *Journal of Systems Science and Complexity* , accepted.

Yu, J., Ai, M., and Ye, Z. (2023). A review on design inspired subsampling for big data. *Statistical Papers* 1–44.

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association* **117**, 265–76.

Zhang, A., Zhang, H., and Yin, G. (2020). Adaptive iterative hessian sketch via a-optimal subsampling. *Statistics and Computing* **30**, 1075–1090.

Zhang, H. and Wang, H. (2021). Distributed subdata selection for big data via sampling-based approach. *Computational Statistics & Data Analysis* **153**, 107072.

Zhang, T., Ning, Y., and Ruppert, D. (2021). Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics* **30**, 106–114.

Zhao, T., Cheng, G., and Liu, H. (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics* **44**, 4, 1400–1437.

Zuo, L., Zhang, H., Wang, H., and Liu, L. (2021a). Sampling-based estimation for massive survival data with additive hazards model. *Statistics in Medicine* **40**, 441–450.

Zuo, L., Zhang, H., Wang, H., and Sun, L. (2021b). Optimal subsample selection for massive logistic regression with distributed data. *Computational Statistics* **36**, 2535–2562.

Supplementary Materials for

Approximating Partial Likelihood Estimators via Optimal Subsampling

Haixiang Zhang, Lulu Zuo, HaiYing Wang and Liuquan Sun

A Proofs

In this section, we give the proof details of Theorems 1-3 and Proposition 1. For these goals, we first need the following lemmas.

Lemma S.1 *(Xu* et al.*, 2009) Suppose that as* $n \to \infty$*,*

$$\sup_{t\in[0,\tau]} |h_n(t) - h(t)| \to 0, \qquad \sup_{t\in[0,\tau]} |g_n(t) - g(t)| \to 0,$$

where h *is continuous on* $[0,\tau]$*,* $g_n(\cdot)$ *and* $g(\cdot)$ *are left-continuous on* $[0,\tau]$*, with their total variations bounded by a constant that is independent of* n*. Then,* $n \to \infty$*,*

$$\sup_{t\in[0,\tau]} \left| \int_0^t h_n(u) dg_n(u) - \int_0^t h(u) dg(u) \right| \to 0,$$

and

$$\sup_{t\in[0,\tau]} \left| \int_0^t h_n(u) dg_n(u) - \int_0^t h_n(u) dg(u) \right| \to 0.$$

Lemma S.2 *Suppose the assumptions 1-4 hold, then as* $n \to \infty$ *and* $r \to \infty$*, conditional on* $\mathcal{D}_n$*, for any* $\boldsymbol{\beta} \in \Theta$ *we have*

$$\mathbf{U}^*(\boldsymbol{\beta}) = \dot{\ell}(\boldsymbol{\beta}) + O_{P|\mathcal{D}_n}(r^{-1/2}), \tag{S.1}$$

and

$$\dot{\ell}^*(\boldsymbol{\beta}) = \mathbf{U}^*(\boldsymbol{\beta}) + o_{P|\mathcal{D}_n}(r^{-1/2}), \tag{S.2}$$

where Θ is a compact set containing the true value of $\boldsymbol{\beta}$, $\dot{\ell}(\boldsymbol{\beta})$ and $\dot{\ell}^*(\boldsymbol{\beta})$ are given in (3) and (8), respectively. Moreover,

$$\mathbf{U}^*(\boldsymbol{\beta}) = -\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i^*(t,\boldsymbol{\beta})$$

with $dM_i^*(t,\boldsymbol{\beta}) = dN_i^*(t) - I(Y_i^* \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i^*)\lambda_0(t)dt$, $i = 1, \cdots, r$.

Proof. For $i = 1, \cdots, r$, denote

$$\zeta_i^*(\boldsymbol{\beta}) = -\frac{1}{n\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i^*(t,\boldsymbol{\beta}).$$

Conditional on $\mathcal{D}_n$, $\zeta_1^*(\boldsymbol{\beta})$, $\cdots$, $\zeta_r^*(\boldsymbol{\beta})$ are independent and identically distributed random vectors, it is straightforward to derive that

$$\begin{aligned} E\{\zeta_1^*(\boldsymbol{\beta})|\mathcal{D}_n\} &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta}) \\ &= \dot{\ell}(\boldsymbol{\beta}). \end{aligned}$$

Note that $\mathbf{U}^*(\boldsymbol{\beta}) = r^{-1}\sum_{i=1}^{r}\zeta_i^*(\boldsymbol{\beta})$, then $E\{\mathbf{U}^*(\boldsymbol{\beta})|\mathcal{D}_n\} = E\{\zeta_1^*(\boldsymbol{\beta})|\mathcal{D}_n\} = \dot{\ell}(\boldsymbol{\beta})$.

Let $\mathbf{U}_j^*(\boldsymbol{\beta})$ be the jth component of $\mathbf{U}^*(\boldsymbol{\beta})$ for $j = 1, \cdots, p$, then we have

$$\begin{aligned} Var\{\mathbf{U}_j^*(\boldsymbol{\beta})|\mathcal{D}_n\} &= \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_i}\left[\int_0^\tau \{\mathbf{X}_{ij} - \bar{\mathbf{X}}_j(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta})\right]^2 \\ &\quad -\frac{1}{n^2 r}\left[\sum_{i=1}^{n}\int_0^\tau \{\mathbf{X}_{ij} - \bar{\mathbf{X}}_j(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta})\right]^2 \\ &\leq \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_i}\left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta})\right\|^2 \\ &\leq \max_{1\leq i\leq n}\left\{\frac{1}{n\pi_i}\right\}\frac{1}{nr}\sum_{i=1}^{n}\left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta})\right\|^2 \\ &= O_P(r^{-1}). \end{aligned}$$

Here the last equality is from the assumption 4, together with

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta})\right\|^2 = O_P(1),$$

which can be deduced by the boundedness of $\mathbf{X}_i$'s in $\mathcal{D}_n$, together with the assumptions 1 and 3. The Markov's inequality implies that $\mathbf{U}_j^*(\boldsymbol{\beta}) - \dot{\ell}_j(\boldsymbol{\beta}) = O_{P|\mathcal{D}_n}(r^{-1/2})$. Therefore, $\mathbf{U}^*(\boldsymbol{\beta}) = \dot{\ell}(\boldsymbol{\beta}) + O_{P|\mathcal{D}_n}(r^{-1/2})$, i.e., the conclusion given in (S.1) is established.

For the sake of proving (S.2), we rewrite the expression of $\dot{\ell}^*(\boldsymbol{\beta})$ as

$$\begin{aligned}\dot{\ell}^*(\boldsymbol{\beta}) &= -\frac{1}{rn}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t,\boldsymbol{\beta}) + \bar{\mathbf{X}}(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}^*(t,\boldsymbol{\beta})\} dM_i^*(t,\boldsymbol{\beta}) \\ &= \mathbf{U}^*(\boldsymbol{\beta}) + \underbrace{\frac{1}{rn}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau \{\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\} dM_i^*(t,\boldsymbol{\beta})}_{\mathbf{R}^*(\boldsymbol{\beta})}. \end{aligned} \tag{S.3}$$

Recall that $\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) = S^{*(1)}(t,\boldsymbol{\beta})/S^{*(0)}(t,\boldsymbol{\beta})$, where $S^{*(1)}(t,\boldsymbol{\beta}) = (nr)^{-1}\sum_{i=1}^r \pi_i^{*-1} I(Y_i^* \geq t)\mathbf{X}_i^* \exp(\boldsymbol{\beta}'\mathbf{X}_i^*)$ and $S^{*(0)}(t,\boldsymbol{\beta}) = (nr)^{-1}\sum_{i=1}^r \pi_i^{*-1} I(Y_i^* \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i^*)$. Conditional on $\mathcal{D}_n$, $(\mathbf{X}_i^*, Y_i^*, \pi_i^*)$'s are independent and identically distributed variables. For a subsample $\mathcal{D}_r^* = \{Z_i^*\}_{i=1}^r$ with $Z_i^* = (\mathbf{X}_i^*, \Delta_i^*, Y_i^*, \pi_i^*)$, we define a subsample empirical measure given the full data $\mathcal{D}_n$,

$$\mathbb{P}_{r|\mathcal{D}_n} = \frac{1}{r}\sum_{i=1}^{r}\delta_{Z_i^*},$$

where δ_Z is a measure that assigns mass 1 at Z and 0 elsewhere. For a measurable function $f : \mathcal{D}_n \mapsto \mathbb{R}$, we denote

$$\mathbb{P}_{r|\mathcal{D}_n} f = \frac{1}{r}\sum_{i=1}^{r} f(Z_i^*).$$

Using the conditional empirical measure $\mathbb{P}_{r|\mathcal{D}_n}$, we can rewrite the $S^{*(k)}(t,\boldsymbol{\beta})$ as

$$S^{*(k)}(t,\boldsymbol{\beta}) = \mathbb{P}_{r|\mathcal{D}_n}\{(n\pi^*)^{-1} I(Y^* \geq t)\mathbf{X}^{*\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}, \quad k = 0, 1, 2.$$

In order to use the technique of empirical process (van der Vaart and Wellner, 1996), we denote $\mathbf{P}_{\mathcal{D}_n}$ as taking expectation conditional on the full data $\mathcal{D}_n$. e.g.

$$\mathbf{P}_{\mathcal{D}_n} f(Z^*) = E\{f(Z^*) \mid \mathcal{D}_n\} = \sum_{i=1}^{n}\pi_i f(Z_i). \tag{S.4}$$

From (S.4), we have the following expressions:

$$\begin{aligned}\mathbf{P}_{\mathcal{D}_n}\{(n\pi^*)^{-1} I(Y^* \geq t)\mathbf{X}^{*\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}^*)\} &= E\left[\frac{1}{n\pi^*} I(Y^* \geq t)\mathbf{X}^{*\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}^*)\,\middle|\,\mathcal{D}_n\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq t)\mathbf{X}_i^{\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}_i) \\ &= S^{(k)}(t,\boldsymbol{\beta}).\end{aligned}$$

By Kosorok (2008) and the assumptions 3 and 4, we know $\{(n\pi)^{-1}I(Y \geq t)\mathbf{X}^{\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}) : t \in [0,\tau], \boldsymbol{\beta} \in \Theta\}$ and $\{(n\pi)^{-1}N(t) : t \in [0,\tau]\}$ are Donsker, where $k = 0$, 1 and 2. Therefore, conditional on $\mathcal{D}_n$ we have

$$\|S^{*(k)}(t,\boldsymbol{\beta}) - S^{(k)}(t,\boldsymbol{\beta})\| \xrightarrow{P} 0 \text{ uniformly towards } t. \tag{S.5}$$

Because $S^{(0)}(t,\boldsymbol{\beta})$ is bounded away from zero (Andersen and Gill, 1982), then conditional on $\mathcal{D}_n$,

$$\sup_{t\in[0,\tau]}\left\|\frac{\mathbb{P}_{r|\mathcal{D}_n}\{(n\pi^*)^{-1}I(Y^* \geq t)\mathbf{X}^*\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}}{\mathbb{P}_{r|\mathcal{D}_n}\{(n\pi^*)^{-1}I(Y^* \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}} - \frac{\mathbf{P}_{\mathcal{D}_n}\{(n\pi^*)^{-1}I(Y^* \geq t)\mathbf{X}^*\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}}{\mathbf{P}_{\mathcal{D}_n}\{(n\pi^*)^{-1}I(Y^* \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}}\right\| \xrightarrow{P} 0.$$

i.e., as $r \to \infty$,

$$\|\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\| \xrightarrow{P} 0 \text{ uniformly towards } t. \tag{S.6}$$

Combining (S.3) and (S.6), as $r \to \infty$ some calculations lead to

$$\begin{aligned}
\mathbf{R}^*(\boldsymbol{\beta}) &= \frac{1}{rn}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau \{\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i^*(t) \\
&= \underbrace{\int_0^\tau \{\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}d\bar{N}_r^*(t)}_{\mathbf{R}_1^*(\boldsymbol{\beta})} - \underbrace{\int_0^\tau \{\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}d\bar{\Lambda}_r^*(t)}_{\mathbf{R}_2^*(\boldsymbol{\beta})},
\end{aligned}$$

where $\bar{N}_r^*(t) = \frac{1}{rn}\sum_{i=1}^r \frac{1}{\pi_i^*}N_i^*(t)$ and $\bar{\Lambda}_r^*(t) = \frac{1}{rn}\sum_{i=1}^r \frac{1}{\pi_i^*}\int_0^t I(Y_i^* \geq u)\exp(\boldsymbol{\beta}'\mathbf{X}_i^*)\lambda_0(u)du$. Note that $\bar{N}_r^*(t)$ and $\bar{\Lambda}_r^*(t)$ are two nondecreasing processes, due to (S.6) we have

$$\begin{aligned}
\|\mathbf{R}_1^*(\boldsymbol{\beta})\| &= \left\|\int_0^\tau \{\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}d\bar{N}_r^*(t)\right\| \\
&\leq \int_0^\tau \|\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\|d\bar{N}_r^*(t) \\
&= \bar{N}_r^*(\tau)o_P(1),
\end{aligned}$$

and

$$\begin{aligned}
\|\mathbf{R}_2^*(\boldsymbol{\beta})\| &= \left\|\int_0^\tau \{\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}d\bar{\Lambda}_r^*(t)\right\| \\
&\leq \int_0^\tau \|\bar{\mathbf{X}}^*(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\|d\bar{\Lambda}_r^*(t) \\
&= \bar{\Lambda}_r^*(\tau)o_P(1).
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{R}^*(\boldsymbol{\beta}) &= \{\bar{N}_r^*(\tau) - \bar{\Lambda}_r^*(\tau)\}o_P(1) \\ &= \left\{\frac{1}{rn}\sum_{i=1}^{r}\frac{1}{\pi_i^*}M_i^*(\tau)\right\}o_P(1).\end{aligned}$$

In view of the martingale property $E\{M(\tau)\} = 0$, conditional on $\mathcal{D}_n$ we observe the following two facts:

$$E\left\{\frac{1}{rn}\sum_{i=1}^{r}\frac{1}{\pi_i^*}M_i^*(\tau)\,\middle|\,\mathcal{D}_n\right\} = \frac{1}{n}\sum_{i=1}^{n}M_i(\tau) = o_P(1),$$

and

$$\begin{aligned}Var\left\{\frac{1}{rn}\sum_{i=1}^{r}\frac{1}{\pi_i^*}M_i^*(\tau)\,\middle|\,\mathcal{D}_n\right\} &= \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_i}M_i^2(\tau) - \frac{1}{r}\left\{\frac{1}{n}\sum_{i=1}^{n}M_i(\tau)\right\}^2 \\ &\leq \max_{1\leq i\leq n}\left\{\frac{1}{n\pi_i}\right\}\frac{1}{rn}\sum_{i=1}^{n}M_i^2(\tau) + o_P(r^{-1}) \\ &= O_P(r^{-1}),\end{aligned}$$

where the last equality is due to the assumptions 1, 3 and 4. By the Markov's inequality, we have

$$\frac{1}{rn}\sum_{i=1}^{r}\frac{1}{\pi_i^*}M_i^*(\tau) = O_{P|\mathcal{D}_n}(r^{-1/2}). \tag{S.7}$$

Therefore, we know that

$$\mathbf{R}^*(\boldsymbol{\beta}) = O_{P|\mathcal{D}_n}(r^{-1/2})o_P(1) = o_{P|\mathcal{D}_n}(r^{-1/2}). \tag{S.8}$$

Combining (S.3) and (S.8), we get $\dot{\ell}^*(\boldsymbol{\beta}) = \mathbf{U}^*(\boldsymbol{\beta}) + o_{P|\mathcal{D}_n}(r^{-1/2})$. This ends the proof.

Lemma S.3 *If the assumptions 1-4 hold, as $n \to \infty$ and $r \to \infty$, conditional on $\mathcal{D}_n$, we have*

$$\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = O_{P|\mathcal{D}_n}(r^{-1/2}), \tag{S.9}$$

and

$$\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \boldsymbol{\Psi} + o_P(1), \tag{S.10}$$

where $\boldsymbol{\Psi}$ *is given in (12), and*

$$\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi_i^*}\left[\frac{S^{*(2)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{*(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{*(1)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{*(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}\right]. \tag{S.11}$$

Proof. In view of (S.1) and (S.2), we know $\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \dot{\ell}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n}(r^{-1/2})$. From Cox (1975), the full data maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ satisfying $\dot{\ell}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = 0$, hence the conclusion given in (S.9) holds.

Based on the subsample $\mathcal{D}_r^* = \{(\mathbf{X}_i^*, \Delta_i^*, Y_i^*, \pi_i^*)\}_{i=1}^r$, we introduce an auxiliary term

$$\mathbf{V}^*(\boldsymbol{\beta}) = -\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dN_i^*(t,\boldsymbol{\beta}), \tag{S.12}$$

where $N_i^*(t) = I(\Delta_i^* = 1, Y_i^* \le t)$, and $\bar{\mathbf{X}}(t,\boldsymbol{\beta})$ is given in (4). Some calculations lead to the following expression:

$$\dot{\mathbf{V}}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi_i^*}\left[\frac{S^{(2)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{(1)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}\right]. \tag{S.13}$$

Conditional on $\mathcal{D}_n$, it is straightforward to deduce that

$$\begin{aligned} E\{\dot{\mathbf{V}}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})|\mathcal{D}_n\} &= \frac{1}{n}\sum_{i=1}^{n}\Delta_i\left[\frac{S^{(2)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{(1)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}\right] \\ &= \boldsymbol{\Psi}. \end{aligned}$$

For any $1 \le j_1, j_2 \le p$, denote $\dot{\mathbf{V}}_{j_1j_2}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})$ and $\boldsymbol{\Psi}_{j_1j_2}$ as any components of $\dot{\mathbf{V}}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})$ and $\boldsymbol{\Psi}$, respectively. Then we have

$$\begin{aligned} Var\{\dot{\mathbf{V}}_{j_1j_2}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})|\mathcal{D}_n\} &= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\left[\frac{S_{j_1j_2}^{(2)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{(1)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}_{j_1j_2}^{\otimes 2}\right]^2 - \frac{1}{r}\boldsymbol{\Psi}_{j_1j_2}^2 \\ &\le \frac{1}{rn^2}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_i}\left\|\frac{S^{(2)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{(1)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}\right\|^2 \end{aligned}$$

$$
\begin{aligned}
&\leq \max_{1\leq i\leq n}\left\{\frac{1}{n\pi_i}\right\}\frac{1}{rn}\sum_{i=1}^{n}\Delta_i\left\|\frac{S^{(2)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}-\left\{\frac{S^{(1)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\}^{\otimes 2}\right\|^2\\
&= O_{P|\mathcal{D}_n}(r^{-1}).
\end{aligned}
$$

Here the last equality is from the assumption 4, along with

$$
\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left\|\frac{S^{(2)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}-\left\{\frac{S^{(1)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\}^{\otimes 2}\right\|^2 = O_{P|\mathcal{D}_n}(1),
$$

which is derived from the boundedness of $\mathbf{X}_i$'s in $\mathcal{D}_n$, the assumption 3 and $S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})$ is bounded away from zero. By the Markov's inequality, we get

$$
\dot{\mathbf{V}}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}}) = \boldsymbol{\Psi} + O_{P|\mathcal{D}_n}(r^{-1/2}). \tag{S.14}
$$

Conditional on $\mathcal{D}_n$, some calculations lead to

$$
\begin{aligned}
\|\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})-\dot{\mathbf{V}}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\| &\leq \frac{1}{rn}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi_i^*}\left\|\frac{S^{*(2)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{*(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}-\frac{S^{(2)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\|\\
&\quad+\frac{1}{rn}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi_i^*}\left\|\left\{\frac{S^{*(1)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{*(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\}^{\otimes 2}-\left\{\frac{S^{(1)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\}^{\otimes 2}\right\|\\
&= \left\{\frac{1}{rn}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi_i^*}\right\}o_P(1)\\
&\leq \max_{1\leq i\leq n}\left\{\frac{1}{n\pi_i}\right\}o_P(1)\\
&= o_P(1),
\end{aligned} \tag{S.15}
$$

which is due to (S.5) and the assumption 4. Thus, we have

$$
\|\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})-\dot{\mathbf{V}}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\| = o_P(1). \tag{S.16}
$$

By the triangle inequality, it is easy to derive that

$$
\begin{aligned}
\|\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})-\boldsymbol{\Psi}\| &\leq \|\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})-\dot{\mathbf{V}}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\|+\|\dot{\mathbf{V}}^*(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})-\boldsymbol{\Psi}\|\\
&= o_P(1),
\end{aligned}
$$

where the last equality is owing to (S.14) and (S.16). This ends the proof.

Proof of Theorem 1. First we establish the asymptotic normality of subsample-based estimator $\tilde{\boldsymbol{\beta}}$ towards $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ given $\mathcal{D}_n$. As $n \to \infty$ and $r \to \infty$, it follows from (S.1) and (S.2) that $\dot{\ell}^*(\boldsymbol{\beta}) - \dot{\ell}(\boldsymbol{\beta}) \to 0$ in probability conditional on $\mathcal{D}_n$. Because the parameter space Θ is compact, the full data estimator $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ is an unique solution to $\dot{\ell}(\boldsymbol{\beta}) = 0$ (Andersen and Gill, 1982). From Theorem 5.9 and its remark of van der Vaart (1998), conditional on $\mathcal{D}_n$ in probability, as $n \to \infty$ and $r \to \infty$, we can obtain the following conclusion:

$$\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| = o_{P|\mathcal{D}_n}(1). \tag{S.17}$$

i.e., for any $\epsilon > 0$, we have $\lim_{r\to\infty} P(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| > \epsilon|\mathcal{D}_n) = 0$. According to Xiong and Li (2008), a random sequence converges to zero in conditional probability also indicates that it converges to zero in unconditional probability. For notational simplicity, throughout the proofs we will use $o_P(1)$ instead of $o_{P|\mathcal{D}_n}(1)$. In other word, we have $\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| = o_P(1)$.

By the Taylor expansion, as $r \to \infty$ we can derive that

$$0 = \dot{\ell}_j^*(\tilde{\boldsymbol{\beta}}) = \dot{\ell}_j^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + \frac{\partial \dot{\ell}_j^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})}{\partial \boldsymbol{\beta}'}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) + R_j, \tag{S.18}$$

where $\dot{\ell}_j^*(\boldsymbol{\beta})$ is the partial derivative of $\ell^*(\boldsymbol{\beta})$ with respect to β_j, and

$$R_j = (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}})' \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}_j^*\{\hat{\boldsymbol{\beta}}_{\text{MPL}} + uv(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} v du dv (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}).$$

Due to the assumptions 3-4, some direct calculations lead to the following conclusion:

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\Theta} \left\| \frac{\partial^2 \dot{\ell}_j^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right\| &\leq \frac{K}{nr} \sum_{i=1}^{r} \frac{\Delta_i^*}{\pi_i^*} \\
&\leq \max_{1\leq i\leq r} \left\{ \frac{1}{n\pi_i^*} \right\} \frac{K}{r} \sum_{i=1}^{r} \Delta_i^* \\
&= O_{P|\mathcal{D}_n}(1),
\end{aligned}$$

where K is a positive constant. Therefore, $R_j = O_{P|\mathcal{D}_n}(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\|^2)$. Based on (S.10), we know $\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \boldsymbol{\Psi} + O_{P|\mathcal{D}_n}(r^{-1/2}) = O_{P|\mathcal{D}_n}(1)$. In view of (S.9) and (S.18), we get

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}} &= -\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1}\{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n}(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\|^2)\} \\
&= O_{P|\mathcal{D}_n}(r^{-1/2}) + o_{P|\mathcal{D}_n}(\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\|) \\
&= O_{P|\mathcal{D}_n}(r^{-1/2}).
\end{aligned} \tag{S.19}$$

Subsequently, we need to prove the asymptotic normality of $\tilde{\boldsymbol{\beta}}$ towards $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ given $\mathcal{D}_n$. Recall that

$$\mathbf{U}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = -\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_i^*(t,\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \sum_{i=1}^{r}\xi_i^*,$$

where

$$\xi_i^* = -\frac{1}{nr\pi_i^*}\int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_i^*(t,\hat{\boldsymbol{\beta}}_{\text{MPL}}), \quad i = 1,\cdots,r.$$

Given $\mathcal{D}_n$, $\xi_1^*,\cdots,\xi_r^*$ are independent and identically distributed random variables with

$$\begin{aligned} E(\xi_i^*|\mathcal{D}_n) &= -\frac{1}{nr}\sum_{i=1}^{n}\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}}) \\ &= \dot{\ell}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = 0, \end{aligned}$$

and

$$\begin{aligned} Var(\xi_i^*|\mathcal{D}_n) &= E\left(\frac{1}{n^2r^2\pi_i^{*2}}\left[\int_0^\tau \left\{\mathbf{X}_i^* - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right\}dM_i^*(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right]^{\otimes 2}\,\middle|\,\mathcal{D}_n\right) \\ &= \frac{1}{n^2r^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\left[\int_0^\tau \left\{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right\}dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right]^{\otimes 2}. \end{aligned}$$

For every $\epsilon > 0$, we have

$$\begin{aligned} &E\left(\sum_{i=1}^{r}\|\xi_i^*\|^2 I(\|\xi_i^*\| > \epsilon)\,\middle|\,\mathcal{D}_n\right) \\ &\le \frac{1}{\epsilon}\sum_{i=1}^{r}E(\|\xi_i^*\|^3|\mathcal{D}_n) \\ &= \frac{1}{r^2\epsilon}\left\{\frac{1}{n^3}\sum_{i=1}^{n}\frac{1}{\pi_i^2}\left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right\|^3\right\} \\ &\le \frac{1}{r^2\epsilon}\max_{1\le i\le n}\left\{\frac{1}{n^2\pi_i^2}\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right\|^3\right\} \\ &= o_P(1), \quad \text{as } r\to\infty. \end{aligned}$$

Here the last equality is from the assumption 4, and

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\}dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right\|^3 = O_{P|\mathcal{D}_n}(1),$$

which is due to the boundedness of $\mathbf{X}_i$'s in $\mathcal{D}_n$ and the assumptions 1 and 3. Therefore, the Lindeberg-Feller conditions are satisfied in probability. By the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart (1998)), as $r \to \infty$ and conditional on $\mathcal{D}_n$, we get

$$\boldsymbol{\Gamma}^{-1/2}\mathbf{U}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{S.20}$$

where

$$\boldsymbol{\Gamma} = \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_i}\left[\int_0^\tau \left\{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\right\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\right]^{\otimes 2} = O_{P|\mathcal{D}_n}(r^{-1}). \tag{S.21}$$

Conditional on $\mathcal{D}_n$, it follows from Theorem 2.7 of van der Vaart (1998), together with (S.2) and (S.20) that as $r \to \infty$,

$$\begin{aligned} \boldsymbol{\Gamma}^{-1/2}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) &= \boldsymbol{\Gamma}^{-1/2}\mathbf{U}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + o_P(1) \\ &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{S.22}$$

Based on (S.10) and (S.19), we can deduce the following conclusion:

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}} = -\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n}(r^{-1}). \tag{S.23}$$

It follows from the assumption 2 and (S.21) that

$$\boldsymbol{\Sigma} = \boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1} = O_{P|\mathcal{D}_n}(r^{-1}). \tag{S.24}$$

By (S.24), it can be deduced that

$$\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1} - \boldsymbol{\Psi}^{-1} = -\boldsymbol{\Psi}^{-1}\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) - \boldsymbol{\Psi}\}\boldsymbol{\Psi}^{-1} = o_P(1). \tag{S.25}$$

From (S.23), (S.24), (S.25) and Lemma S.3, we get

$$\begin{aligned} \boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) &= -\boldsymbol{\Sigma}^{-1/2}\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n}(r^{-1/2}) \\ &= -\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) - \boldsymbol{\Sigma}^{-1/2}[\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1} - \boldsymbol{\Psi}^{-1}]\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n}(r^{-1/2}) \\ &= -\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Gamma}^{-1/2}\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + o_P(1). \end{aligned} \tag{S.26}$$

Furthermore, we observe that

$$\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^{1/2}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^{1/2})' = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Gamma}^{1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}^{-1/2} = \mathbf{I}. \tag{S.27}$$

By (S.22), (S.26), (S.27) and the Slutsky's theorem, conditional on $\mathcal{D}_n$, as $n \to \infty$ and $r \to \infty$,

$$\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}).$$

That is to say, for any $\mathbf{x} \in \mathbb{R}^p$ we have $P\{\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) \le \mathbf{x} | \mathcal{D}_n\} \to \Phi(\mathbf{x})$ in probability, where $\Phi(\mathbf{x})$ is the cumulative distribution function of the standard multivariate normal distribution.

Proof of Theorem 2. Note that

$$\begin{aligned}
tr(\boldsymbol{\Gamma}) &= tr\left(\frac{1}{n^2 r} \sum_{i=1}^{n} \frac{1}{\pi_i} \left[\int_0^\tau \left\{ \mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right]^{\otimes 2} \right) \\
&= \frac{1}{rn^2} \sum_{i=1}^{n} \frac{1}{\pi_i} \left\| \int_0^\tau \left\{ \mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\|^2 \\
&= \frac{1}{rn^2} \sum_{i=1}^{n} \pi_i \sum_{i=1}^{n} \frac{1}{\pi_i} \left\| \int_0^\tau \left\{ \mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\|^2 \\
&\ge \frac{1}{rn^2} \left\{ \sum_{i=1}^{n} \left\| \int_0^\tau \left\{ \mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\| \right\}^2,
\end{aligned}$$

where the last inequality is from the Cauchy-Schwarz inequality, and its equality holds if and only if $\pi_i = \varsigma \| \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \|$ for some $\varsigma > 0$. Due to $\sum_{i=1}^n \pi_i = 1$, we know $\varsigma = \{\sum_{j=1}^n \| \int_0^\tau \{\mathbf{X}_j - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_j(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \|\}^{-1}$. Therefore, the optimal subsampling probabilities are

$$\pi_i^{Lopt} = \frac{\| \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \|}{\sum_{j=1}^n \| \int_0^\tau \{\mathbf{X}_j - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_j(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \|}, \quad i = 1, \cdots, n.$$

This completes the proof.

Lemma S.4 *Under the assumptions 1-3, as $n \to \infty$ and $r_0 \to \infty$, conditional on $\mathcal{D}_n$ we have $\hat{\Lambda}_0^{\text{UNIF}}(t, \tilde{\boldsymbol{\beta}}_0) = \Lambda_0(t) + O_{P|D_n}(r_0^{-1/2})$, i.e., for any $\epsilon > 0$, with probability approaching one, there exists a finite Δ_ϵ and r_ϵ, such that*

$$P\left(\left| \hat{\Lambda}_0^{\text{UNIF}}(t, \tilde{\boldsymbol{\beta}}_0) - \Lambda_0(t) \right| \ge r_0^{-1/2} \Delta_\epsilon \,\middle|\, \mathcal{D}_n \right) < \epsilon, \tag{S.28}$$

for all $r_0 \geq r_\epsilon$, where $\hat{\Lambda}_0^{\text{UNIF}}(t, \boldsymbol{\beta})$ is a uniform subsample Breslow-type estimator defined in (17).

Proof. For any $t \in [0, \tau]$ and $\boldsymbol{\beta} \in \Theta$, conditional on $\mathcal{D}_n$ we need to prove the following two expressions:

$$\frac{1}{r_0}\sum_{i=1}^{r_0} I(Y_i^{0*} \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i^{0*}) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i) + O_{P|\mathcal{D}_n}(r_0^{-1/2}), \tag{S.29}$$

and

$$\frac{1}{r_0}\sum_{i=1}^{r_0} \frac{\Delta_i^{0*} I(Y_i^{0*} \leq t)}{n^{-1}\sum_{j=1}^{n} I(Y_j \geq Y_i^{0*})\exp(\boldsymbol{\beta}'\mathbf{X}_j)} = \hat{\Lambda}_0(t, \boldsymbol{\beta}) + O_{P|\mathcal{D}_n}(r_0^{-1/2}), \tag{S.30}$$

where $\hat{\Lambda}_0(t, \boldsymbol{\beta})$ is the full data Breslow estimator given in (25), and $\mathcal{D}_{r_0}^* = \{(\mathbf{X}_i^{0*}, \Delta_i^{0*}, Y_i^{0*})\}_{i=1}^{r_0}$ is a uniform subsample from the full data $\mathcal{D}_n$.

Conditional on $\mathcal{D}_n$, it is straightforward to clarify that

$$E\left\{\frac{1}{r_0}\sum_{i=1}^{r_0} I(Y_i^{0*} \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i^{0*}) \,\middle|\, \mathcal{D}_n\right\} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i),$$

and

$$\begin{aligned}
&Var\left\{\frac{1}{r_0}\sum_{i=1}^{r_0} I(Y_i^{0*} \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i^{0*}) \,\middle|\, \mathcal{D}_n\right\} \\
&= \frac{1}{nr_0}\sum_{i=1}^{n} I(Y_i \geq t)\exp(2\boldsymbol{\beta}'\mathbf{X}_i) - \frac{1}{n^2 r}\left\{\sum_{i=1}^{n} I(Y_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i)\right\}^2 \\
&\leq \frac{1}{r_0 n}\sum_{i=1}^{n}\exp(2\boldsymbol{\beta}'\mathbf{X}_i) \\
&= O_{P|\mathcal{D}_n}(r^{-1}),
\end{aligned}$$

where the last equality is from the assumption 3. The Markov's inequality leads to (S.29).

For $i = 1, \cdots, r_0$, we denote

$$\eta_i^{0*} = \frac{\Delta_i^{0*} I(Y_i^{0*} \leq t)}{n^{-1}\sum_{j=1}^{n} I(Y_j \geq Y_i^{0*})\exp(\boldsymbol{\beta}'\mathbf{X}_j)}.$$

Conditional on $\mathcal{D}_n$, we have

$$E\left(\frac{1}{r_0}\sum_{i=1}^{r_0}\eta_i^{0*} \,\middle|\, \mathcal{D}_n\right) = \sum_{i=1}^{n}\frac{\Delta_i I(Y_i \leq t)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)\exp(\boldsymbol{\beta}'\mathbf{X}_j)}$$

$$= \hat{\Lambda}_0(t, \boldsymbol{\beta}),$$

and

$$\begin{aligned} Var\left(\frac{1}{r_0}\sum_{i=1}^{r_0}\eta_i^{0*}\,\middle|\,\mathcal{D}_n\right) &= E\left\{\frac{1}{r_0}\sum_{i=1}^{r_0}\eta_i^{0*} - \hat{\Lambda}_0(t,\boldsymbol{\beta})\,\middle|\,\mathcal{D}_n\right\}^2 \\ &= \frac{1}{r_0}E\left\{(\eta_i^{0*})^2 - 2\eta_i^{0*}\hat{\Lambda}_0(t,\boldsymbol{\beta}) + \hat{\Lambda}_0^2(t,\boldsymbol{\beta})\,\middle|\,\mathcal{D}_n\right\} \\ &= \frac{1}{r_0}E\left\{(\eta_i^{0*})^2\,|\,\mathcal{D}_n\right\} - \frac{1}{r_0}\hat{\Lambda}_0^2(t,\boldsymbol{\beta}) \\ &\le \frac{1}{r_0 n}\left[\sum_{i=1}^{n}\frac{\Delta_i I(Y_i \le t)}{\{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\mathbf{X}_j'\boldsymbol{\beta})\}^2}\right] \\ &= O_{P|\mathcal{D}_n}(r_0^{-1}). \end{aligned}$$

Here the last equality is due to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I(Y_i \le t)}{\{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\mathbf{X}_j'\boldsymbol{\beta})\}^2} = O_P(1),$$

which is from the assumption 3. As a result, the Markov's inequality ensures that (S.30) holds, i.e., $r_0^{-1}\sum_{i=1}^{r_0}\eta_i^{0*} = \hat{\Lambda}_0(t,\boldsymbol{\beta}) + O_{P|\mathcal{D}_n}(r_0^{-1/2})$. In addition, some direct calculations yield that

$$\begin{aligned} \hat{\Lambda}_0^{\text{UNIF}}(t,\boldsymbol{\beta}) &= \frac{1}{r_0}\sum_{i=1}^{r_0}\left\{\frac{\Delta_i^{0*} I(Y_i^{0*} \le t)}{r_0^{-1}\sum_{j=1}^{r_0} I(Y_j^{0*} \ge Y_i^{0*})\exp(\boldsymbol{\beta}'\mathbf{X}_j^{0*})} - \eta_i^{0*} + \eta_i^{0*}\right\} \\ &= \frac{1}{r_0}\sum_{i=1}^{r_0}\Delta_i^{0*} I(Y_i^{0*} \le t)\left\{\frac{1}{r_0^{-1}\sum_{j=1}^{r_0} I(Y_j^{0*} \ge Y_i^{0*})\exp(\boldsymbol{\beta}'\mathbf{X}_j^{0*})}\right. \\ &\quad \left. - \frac{1}{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i^{0*})\exp(\boldsymbol{\beta}'\mathbf{X}_j)}\right\} + \frac{1}{r_0}\sum_{i=1}^{r_0}\eta_i^{0*} \\ &= \left\{\frac{1}{r_0}\sum_{i=1}^{r_0}\Delta_i^{0*} I(Y_i^{0*} \le t)\right\} O_{P|\mathcal{D}_n}(r_0^{-1/2}) + \hat{\Lambda}_0(t,\boldsymbol{\beta}) + O_{P|\mathcal{D}_n}(r_0^{-1/2}), \quad \text{(S.31)} \end{aligned}$$

where the last equality is owing to (S.29) and (S.30).

Given $\mathcal{D}_n$, it can be deduced that

$$E\left\{\frac{1}{r_0}\sum_{i=1}^{r_0}\Delta_i^{0*} I(Y_i^{0*} \le t)\,\middle|\,\mathcal{D}_n\right\} = \frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i \le t),$$

and

$$Var\left\{\frac{1}{r_0}\sum_{i=1}^{r_0}\Delta_i^{0*} I(Y_i^{0*} \le t)\,\middle|\,\mathcal{D}_n\right\} = E\left\{\frac{1}{r_0}\sum_{i=1}^{r_0}\Delta_i^{0*} I(Y_i^{0*} \le t) - \frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i \le t)\,\middle|\,\mathcal{D}_n\right\}^2$$

$$= \frac{1}{r_0}\left[\frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i \le t) - \left\{\frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i \le t)\right\}^2\right]$$
$$= O_{P|\mathcal{D}_n}(r_0^{-1}).$$

Hence, we get

$$\frac{1}{r_0}\sum_{i=1}^{r_0}\Delta_i^{0*} I(Y_i^{0*} \le t) = \frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i \le t) + O_{P|\mathcal{D}_n}(r_0^{-1/2})$$
$$= O_{P|\mathcal{D}_n}(1). \tag{S.32}$$

It follows from (S.31) and (S.32) that

$$\hat{\Lambda}_0^{\text{UNIF}}(t, \boldsymbol{\beta}) = \hat{\Lambda}_0(t, \boldsymbol{\beta}) + O_{P|\mathcal{D}_n}(r_0^{-1/2}). \tag{S.33}$$

In addition, we investigate the distance between $\hat{\Lambda}_0(t, \tilde{\boldsymbol{\beta}}_0)$ and $\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})$:

$$\begin{aligned}
&|\hat{\Lambda}_0(t, \tilde{\boldsymbol{\beta}}_0) - \hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})| \\
&= \left|\sum_{i=1}^{n}\frac{\Delta_i I(Y_i \le t)}{\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\tilde{\boldsymbol{\beta}}_0'\mathbf{X}_j)} - \sum_{i=1}^{n}\frac{\Delta_i I(Y_i \le t)}{\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\hat{\boldsymbol{\beta}}_{\text{MPL}}'\mathbf{X}_j)}\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i \le t)\frac{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i)\{\exp(\tilde{\boldsymbol{\beta}}_0'\mathbf{X}_j) - \exp(\hat{\boldsymbol{\beta}}_{\text{MPL}}'\mathbf{X}_j)\}}{\{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\tilde{\boldsymbol{\beta}}_0'\mathbf{X}_j)\}\{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\hat{\boldsymbol{\beta}}_{\text{MPL}}'\mathbf{X}_j)\}}\right| \\
&\le \left|\frac{1}{n}\sum_{i=1}^{n}\frac{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\xi'\mathbf{X}_j)\|\mathbf{X}_j\|}{\{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\tilde{\boldsymbol{\beta}}_0'\mathbf{X}_j)\}\{n^{-1}\sum_{j=1}^{n} I(Y_j \ge Y_i)\exp(\hat{\boldsymbol{\beta}}_{\text{MPL}}'\mathbf{X}_j)\}}\right| \|\tilde{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| \\
&= O_P(r_0^{-1/2}),
\end{aligned} \tag{S.34}$$

where ξ is on the segment between $\tilde{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_{\text{MPL}}$, and the last equality is due to assumption 3 together with $\|\tilde{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| = O_{P|\mathcal{D}_n}(r_0^{-1/2})$.

Based on Andersen and Gill (1982), the convergence rate of full data Breslow estimator $\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})$ to $\Lambda_0(t)$ is $O_P(n^{-1/2})$, i.e. $\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) - \Lambda_0(t) = O_P(n^{-1/2})$. This together with (S.33) and (S.34) ensures that

$$\begin{aligned}
|\hat{\Lambda}_0^{\text{UNIF}}(t, \tilde{\boldsymbol{\beta}}_0) - \Lambda_0(t)| &\le |\hat{\Lambda}_0^{\text{UNIF}}(t, \tilde{\boldsymbol{\beta}}_0) - \hat{\Lambda}_0(t, \tilde{\boldsymbol{\beta}}_0)| + |\hat{\Lambda}_0(t, \tilde{\boldsymbol{\beta}}_0) - \hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})| \\
&\quad + |\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) - \Lambda_0(t)| \\
&= O_{P|\mathcal{D}_n}(r_0^{-1/2}) + O_{P|\mathcal{D}_n}(r_0^{-1/2}) + O_P(n^{-1/2})
\end{aligned}$$

$$= \quad O_{P|\mathcal{D}_n}(r_0^{-1/2}).$$

Therefore, the convergence rate given in (S.28) is established. This ends the proof.

Lemma S.5 *Suppose the assumptions 1-3 hold, as* $r_0 \to \infty$, $r \to \infty$, *and* $n \to \infty$, *conditional on* $\mathcal{D}_n$ *and* $\tilde{\boldsymbol{\beta}}_0$, *we have*

$$\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) = \dot{\ell}(\boldsymbol{\beta}) + O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2}), \tag{S.35}$$

and

$$\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) = \mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) + o_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2}), \tag{S.36}$$

where $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})$ *is given in (20), and*

$$\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) = -\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi^{app*}_{\delta i}}\int_0^\tau \{\mathbf{X}^*_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM^*_i(t,\boldsymbol{\beta})$$

with $\mathbf{X}^*_i$, $\pi^{app*}_{\delta i}$ *and* $M^*_i(t,\boldsymbol{\beta})$ *being given in (20),* $i = 1, \cdots, r$.

Proof. Given $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, it is direct to deduce the unbiasedness of $\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})$ towards the score $\dot{\ell}(\boldsymbol{\beta})$, i.e.,

$$\begin{aligned} E\{\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0\} &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta}) \\ &= \dot{\ell}(\boldsymbol{\beta}). \end{aligned}$$

Denote $\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\boldsymbol{\beta})$ as the jth component of $\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta})$ with $1 \le j \le p$, then we get

$$\begin{aligned} Var\{\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\boldsymbol{\beta})|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0\} &\le \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi^{\text{app}}_{\delta i}}\left[\int_0^\tau \{\mathbf{X}_{ij} - \bar{\mathbf{X}}_j(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta})\right]^2 \\ &\le \frac{1}{nr\delta}\sum_{i=1}^{n}\left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i(t,\boldsymbol{\beta})\right\|^2 \\ &= O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1}), \end{aligned}$$

where the last equality is from the assumptions 1 and 3. This together with the Markov's inequality can ensure that (S.35) holds.

In addition, some direct calculations lead to the following expressions:

$$\begin{aligned}
\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) &= -\frac{1}{r}\sum_{i=1}^{r}\int_0^\tau \frac{1}{n\pi^{\text{app}*}_{\delta i}}\left\{\mathbf{X}^*_i - \bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})\right\} dN^*_i(t) \\
&= -\frac{1}{r}\sum_{i=1}^{r}\int_0^\tau \frac{1}{n\pi^{\text{app}*}_{\delta i}}\left\{\mathbf{X}^*_i - \bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})\right\} dM^*_i(t,\boldsymbol{\beta}) \\
&= -\frac{1}{r}\sum_{i=1}^{r}\int_0^\tau \frac{1}{n\pi^{\text{app}*}_{\delta i}}\left\{\mathbf{X}^*_i - \bar{\mathbf{X}}(t,\boldsymbol{\beta}) + \bar{\mathbf{X}}(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})\right\} dM^*_i(t,\boldsymbol{\beta}) \\
&= \mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) + \underbrace{\frac{1}{r}\sum_{i=1}^{r}\int_0^\tau \frac{1}{n\pi^{\text{app}*}_{\delta i}}\left\{\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\right\} dM^*_i(t,\boldsymbol{\beta})}_{\mathbf{R}^*_\delta(\boldsymbol{\beta})}. \qquad \text{(S.37)}
\end{aligned}$$

For k =0, 1, and 2, we denote

$$S^{*(k)}_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta}) = \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n\pi^{\text{app}*}_{\delta i}} I(Y^*_i \geq t)\mathbf{X}^{*\otimes k}_i \exp(\boldsymbol{\beta}'\mathbf{X}^*_i). \qquad \text{(S.38)}$$

For a subsample $\mathcal{D}^*_r = \{Z^*_i\}_{i=1}^r$ with $Z^*_i = (\mathbf{X}^*_i, \Delta^*_i, Y^*_i, \pi^{\text{app}*}_{\delta i})$, we define a subsample empirical measure conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$,

$$\mathbb{P}_{r|\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n} = \frac{1}{r}\sum_{i=1}^{r}\delta_{Z^*_i},$$

and

$$\mathbb{P}_{r|\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n} f = \frac{1}{r}\sum_{i=1}^{r} f(Z^*_i).$$

Based on the conditional empirical measure $\mathbb{P}_{r|\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n}$, we can rewrite $S^{*(k)}_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})$ as

$$S^{*(k)}_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta}) = \mathbb{P}_{r|\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n}[\{n\pi^{\text{app}*}_\delta\}^{-1} I(Y^* \geq t)\mathbf{X}^{*\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}^*)],$$

where $k = 0$, 1 and 2. For convenience, we denote $\mathbf{P}_{\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n}$ as taking expectation conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$. e.g.

$$\mathbf{P}_{\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n} f(Z^*) = E\left\{f(Z^*)\,\middle|\,\tilde{\boldsymbol{\beta}}_0, \mathcal{D}_n\right\} = \sum_{i=1}^{n}\pi^{\text{app}}_{\delta i} f(Z_i). \qquad \text{(S.39)}$$

By (S.39), we can deduce the following expressions:

$$\begin{aligned}\mathbf{P}_{\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n}[\{n\pi_\delta^{\text{app}*}\}^{-1}I(Y^*\geq t)\mathbf{X}^{*\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}^*)] &= E\left[\frac{1}{n\pi_\delta^{\text{app}*}}I(Y^*\geq t)\mathbf{X}^{*\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}^*)\,\middle|\,\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}I(Y_i\geq t)\mathbf{X}_i^{\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}_i)\\ &= S^{(k)}(t,\boldsymbol{\beta}).\end{aligned}$$

Due to Kosorok (2008) and the assumption 3, we get $\{(n\pi_\delta^{\text{app}*})^{-1}I(Y^*\geq t)\mathbf{X}^{*\otimes k}\exp(\boldsymbol{\beta}'\mathbf{X}^*) : t\in[0,\tau],\boldsymbol{\beta}\in\Theta\}$ and $\{(n\pi_\delta^{\text{app}*})^{-1}N(t) : t\in[0,\tau]\}$ are Donsker, where k =0, 1 and 2. Therefore, conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$ we have

$$\|S^{*(k)}_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})-S^{(k)}(t,\boldsymbol{\beta})\|\xrightarrow{P}0 \text{ uniformly towards } t. \tag{S.40}$$

Because $S^{(0)}(t,\boldsymbol{\beta})$ is bounded away from zero (Andersen and Gill, 1982), then conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$,

$$\sup_{t\in[0,\tau]}\left\|\frac{\mathbb{P}_{r|\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n}\{(n\pi_\delta^{\text{app}*})^{-1}I(Y^*\geq t)\mathbf{X}^*\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}}{\mathbb{P}_{r|\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n}\{(n\pi_\delta^{\text{app}*})^{-1}I(Y^*\geq t)\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}}-\frac{\mathbf{P}_{\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n}\{(n\pi_\delta^{\text{app}*})^{-1}I(Y^*\geq t)\mathbf{X}^*\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}}{\mathbf{P}_{\tilde{\boldsymbol{\beta}}_0,\mathcal{D}_n}\{(n\pi_\delta^{\text{app}*})^{-1}I(Y^*\geq t)\exp(\boldsymbol{\beta}'\mathbf{X}^*)\}}\right\|\xrightarrow{P}0.$$

i.e., as $r\to\infty$,

$$\|\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})-\bar{\mathbf{X}}(t,\boldsymbol{\beta})\|\xrightarrow{P}0 \text{ uniformly towards } t. \tag{S.41}$$

Recall that $dM_i^*(t,\boldsymbol{\beta})=dN_i^*(t)-I(Y_i^*\geq t)\exp(\boldsymbol{\beta}'\mathbf{X}_i^*)\lambda_0(t)dt$, some derivations result in the following expressions:

$$\begin{aligned}\mathbf{R}^*_\delta(\boldsymbol{\beta}) &= \frac{1}{rn}\sum_{i=1}^{r}\frac{1}{\pi_{\delta i}^{\text{app}*}}\int_0^\tau\{\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})-\bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dM_i^*(t,\boldsymbol{\beta}) \qquad \text{(S.42)}\\ &= \underbrace{\int_0^\tau\{\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})-\bar{\mathbf{X}}(t,\boldsymbol{\beta})\}d\bar{N}^*_{r\delta}(t)}_{\mathbf{R}^*_{1\delta}(\boldsymbol{\beta})}-\underbrace{\int_0^\tau\{\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})-\bar{\mathbf{X}}(t,\boldsymbol{\beta})\}d\bar{\Lambda}^*_{r\delta}(t)}_{\mathbf{R}^*_{2\delta}(\boldsymbol{\beta})},\end{aligned}$$

where $\bar{N}^*_{r\delta}(t)=\frac{1}{rn}\sum_{i=1}^r\frac{1}{\pi_{\delta i}^{\text{app}*}}N_i^*(t)$ and $\bar{\Lambda}^*_{r\delta}(t)=\frac{1}{rn}\sum_{i=1}^r\frac{1}{\pi_{\delta i}^{\text{app}*}}\int_0^t I(Y_i^*\geq u)\exp(\boldsymbol{\beta}'\mathbf{X}_i^*)\lambda_0(u)du$. Notice that $\bar{N}^*_{r\delta}(t)$ and $\bar{\Lambda}^*_{r\delta}(t)$ are two nondecreasing processes, due to (S.41) we have

$$\begin{aligned}\|\mathbf{R}^*_{1\delta}(\boldsymbol{\beta})\| &= \left\|\int_0^\tau\{\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})-\bar{\mathbf{X}}(t,\boldsymbol{\beta})\}d\bar{N}^*_{r\delta}(t)\right\|\\ &\leq \int_0^\tau\|\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta})-\bar{\mathbf{X}}(t,\boldsymbol{\beta})\|d\bar{N}^*_{r\delta}(t)\end{aligned}$$

$$= \bar{N}^*_{r\delta}(\tau) o_P(1),$$

and

$$\begin{aligned}
\|\mathbf{R}^*_{2\delta}(\boldsymbol{\beta})\| &= \left\| \int_0^\tau \{\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\} d\bar{\Lambda}^*_{r\delta}(t) \right\| \\
&\leq \int_0^\tau \|\bar{\mathbf{X}}^*_{\tilde{\boldsymbol{\beta}}_0}(t,\boldsymbol{\beta}) - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\| d\bar{\Lambda}^*_{r\delta}(t) \\
&= \bar{\Lambda}^*_{r\delta}(\tau) o_P(1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{R}^*_\delta(\boldsymbol{\beta}) &= \{\bar{N}^*_{r\delta}(\tau) - \bar{\Lambda}^*_{r\delta}(\tau)\} o_P(1) \\
&= \left\{ \frac{1}{rn} \sum_{i=1}^r \frac{1}{\pi^{\text{app}*}_{\delta i}} M^*_i(\tau) \right\} o_P(1).
\end{aligned}$$

Conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, we get

$$E\left\{ \frac{1}{r} \sum_{i=1}^r \frac{1}{n\pi^{\text{app}*}_{\delta i}} M^*_i(\tau) \,\middle|\, \mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0 \right\} = \frac{1}{n} \sum_{i=1}^n M_i(\tau) = o_P(1),$$

and

$$\begin{aligned}
Var\left\{ \frac{1}{r} \sum_{i=1}^r \frac{1}{n\pi^{\text{app}*}_{\delta i}} M^*_i(\tau) \,\middle|\, \mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0 \right\} &= \frac{1}{n^2 r} \sum_{i=1}^n \frac{1}{\pi^{\text{app}}_{\delta i}} M^2_i(\tau) - \frac{1}{r} \left\{ \frac{1}{n} \sum_{i=1}^n M_i(\tau) \right\}^2 \\
&\leq \frac{1}{rn\delta} \sum_{i=1}^n M^2_i(\tau) + o_P(r^{-1}) \\
&= O_P(r^{-1}),
\end{aligned}$$

where the last equality is from the assumptions 1 and 3. By the Markov's inequality, we know

$$\frac{1}{r} \sum_{i=1}^r \frac{1}{n\pi^{\text{app}*}_{\delta i}} M^*_i(\tau) = O_{P|\mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0}(r^{-1/2}). \tag{S.43}$$

Conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, due to (S.42) and (S.43) we get

$$\mathbf{R}^*_\delta(\boldsymbol{\beta}) = o_P(1) O_{P|\mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0}(r^{-1/2}) = o_{P|\mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0}(r^{-1/2}).$$

This together with (S.37) leads to the conclusion given in (S.36), which completes the proof.

Lemma S.6 *Under the assumptions 1-3, as $r_0 \to \infty$, $r \to \infty$ and $n \to \infty$, conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, we have*

$$\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2}), \tag{S.44}$$

and

$$\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \boldsymbol{\Psi} + o_P(1), \tag{S.45}$$

where $\boldsymbol{\Psi}$ is given in (12), and

$$\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi_{\delta i}^{app*}}\left[\frac{S^{*(2)}_{\tilde{\boldsymbol{\beta}}_0}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{*(0)}_{\tilde{\boldsymbol{\beta}}_0}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{*(1)}_{\tilde{\boldsymbol{\beta}}_0}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{*(0)}_{\tilde{\boldsymbol{\beta}}_0}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}\right].$$

Proof. Conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, it follows from (S.35) and (S.36) that $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \dot{\ell}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2})$. Due to $\dot{\ell}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = 0$, then we get $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2})$.

To prove (S.45), we introduce a term as

$$\mathbf{V}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = -\frac{1}{nr}\sum_{i=1}^{r}\frac{1}{\pi_{\delta i}^{\text{app}*}}\int_0^{\tau}\{\mathbf{X}_i^* - \bar{\mathbf{X}}(t,\boldsymbol{\beta})\}dN_i^*(t,\boldsymbol{\beta}).$$

Furthermore, some direct calculations lead to the following expression:

$$\dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \frac{1}{nr}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi_{\delta i}^{\text{app}*}}\left[\frac{S^{(2)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{(1)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}\right].$$

Then, we get

$$\begin{aligned} E\{\dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}})|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0\} &= \frac{1}{n}\sum_{i=1}^{n}\Delta_i\left[\frac{S^{(2)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{(1)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}\right] \\ &= \boldsymbol{\Psi}. \end{aligned}$$

Let $\dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0,j_1j_2}(\hat{\boldsymbol{\beta}}_{\text{MPL}})$ be any component of $\dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}})$ with $1 \le j_1, j_2 \le p$, we can deduce that

$$Var\{\dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0,j_1j_2}(\hat{\boldsymbol{\beta}}_{\text{MPL}})|\mathcal{F}_n,\tilde{\boldsymbol{\beta}}_0\} = \frac{1}{rn^2}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_{\delta i}^{\text{app}}}\left[\frac{S^{(2)}_{j_1j_2}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})} - \left\{\frac{S^{(1)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\text{MPL}})}\right\}^{\otimes 2}_{j_1j_2}\right]^2 - \frac{1}{r}\boldsymbol{\Psi}_{j_1j_2}$$

$$\leq \frac{1}{r\delta}\left[\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left\|\frac{S^{(2)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}-\left\{\frac{S^{(1)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\}^{\otimes 2}\right\|^2\right]$$
$$= O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1}),$$

where the last equality is due to the assumptions 1-3. Conditional on $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, the Markov's inequality implies

$$\dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}}) = \boldsymbol{\Psi} + O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{1/2}). \tag{S.46}$$

In view of (S.40), we can derive that

$$\begin{aligned}
\|\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}}) - \dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\| &\leq \frac{1}{rn}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi^{\mathrm{app}*}_{\delta i}}\left\|\frac{S^{*(2)}_{\tilde{\boldsymbol{\beta}}_0}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{*(0)}_{\tilde{\boldsymbol{\beta}}_0}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}-\frac{S^{(2)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\| \\
&\quad + \frac{1}{rn}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi^{\mathrm{app}*}_{\delta i}}\left\|\left\{\frac{S^{*(1)}_{\tilde{\boldsymbol{\beta}}_0}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{*(0)}_{\tilde{\boldsymbol{\beta}}_0}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\}^{\otimes 2}-\left\{\frac{S^{(1)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}{S^{(0)}(Y_i^*,\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})}\right\}^{\otimes 2}\right\| \\
&= \left\{\frac{1}{nr}\sum_{i=1}^{r}\frac{\Delta_i^*}{\pi^{\mathrm{app}*}_{\delta i}}\right\}o_P(1) \\
&\leq \left\{\frac{1}{r\delta}\sum_{i=1}^{r}\Delta_i^*\right\}o_P(1) \\
&= o_P(1).
\end{aligned} \tag{S.47}$$

Accordingly, it follows from the triangle inequality that

$$\begin{aligned}
\|\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}}) - \boldsymbol{\Psi}\| &\leq \|\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}}) - \dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}})\| + \|\dot{\mathbf{V}}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}}) - \boldsymbol{\Psi}\| \\
&= o_P(1),
\end{aligned}$$

where the last equality is due to (S.46) and (S.47). Hence, we obtain $\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\mathrm{MPL}}) = \boldsymbol{\Psi} + o_P(1)$. This finishes the proof.

Proof of Theorem 3. By (S.35) and (S.36), we get $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\boldsymbol{\beta}) = \dot{\ell}^*(\boldsymbol{\beta}) + o_P(1)$ as $r_0 \to \infty$ and $r \to \infty$. We observe that the full data estimator $\hat{\boldsymbol{\beta}}_{\mathrm{MPL}}$ is a unique solution to $\dot{\ell}^*(\boldsymbol{\beta}) = 0$, and the two-step subsample estimator $\breve{\boldsymbol{\beta}}$ satisfies $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\breve{\boldsymbol{\beta}}) = 0$. Conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, we

know from Theorem 5.9 and its remark of van der Vaart (1998) that

$$\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| = o_P(1). \tag{S.48}$$

By the Taylor's theorem, we obtain

$$0 = \dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\breve{\boldsymbol{\beta}}) = \dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + \frac{\partial \dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\hat{\boldsymbol{\beta}}_{\text{MPL}})}{\partial \boldsymbol{\beta}'}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) + R_{\tilde{\boldsymbol{\beta}}_0,j}, \tag{S.49}$$

where $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\cdot)$ is the jth component of $\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\cdot)$, and

$$R_{\tilde{\boldsymbol{\beta}}_0,j} = (\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}})' \int_0^1 \int_0^1 \frac{\partial^2 \dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}\{\hat{\boldsymbol{\beta}}_{\text{MPL}} + uv(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}})\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} v du dv (\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}).$$

From the assumptions 1 and 3, we get

$$\begin{aligned} \sup_{\boldsymbol{\beta} \in \Theta} \left\| \frac{\partial^2 \dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0,j}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right\| &\leq \frac{K}{nr} \sum_{i=1}^{r} \frac{\Delta_i^*}{\pi_{\delta i}^{\text{app}*}} \\ &\leq \frac{K}{r\delta} \sum_{i=1}^{r} \Delta_i^* \\ &= O_{P|\mathcal{D}_n}(1), \end{aligned}$$

where K is a positive constant. Hence, $R_{\tilde{\boldsymbol{\beta}}_0,j} = O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\|^2)$.

By (S.45), the assumption 2 and the continuous mapping theorem (Theorem 2.3 of van der Vaart (1998)), conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, as $r \to \infty$ we get

$$\begin{aligned} \{\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1} &= \{\boldsymbol{\Psi} + o_P(1)\}^{-1} \\ &= O_P(1). \end{aligned} \tag{S.50}$$

Conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, it follows from (S.44), (S.49) and (S.50) that

$$\begin{aligned} \breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}} &= -\{\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1}\{\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\|^2)\} \\ &= O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2}) + o_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(\|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\|) \\ &= O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2}). \end{aligned} \tag{S.51}$$

Therefore, $\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}} = o_P(1)$, i.e., $\breve{\boldsymbol{\beta}}$ is consistent to $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ as $r \to \infty$.

We start to prove the asymptotic normality of the error term $\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}$ conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$. Recall that

$$\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) = \sum_{i=1}^{r} \xi_i^{*\tilde{\boldsymbol{\beta}}_0},$$

where

$$\xi_i^{*\tilde{\boldsymbol{\beta}}_0} = -\frac{1}{nr\pi_{\delta i}^{\text{app}*}} \int_0^\tau \{\mathbf{X}_i^* - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_i^*(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}), \quad i = 1, \cdots, r.$$

Conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, $\xi_1^{*\tilde{\boldsymbol{\beta}}_0}, \cdots, \xi_r^{*\tilde{\boldsymbol{\beta}}_0}$ are independent and identically distributed random variables with

$$\begin{aligned} E(\xi_i^{*\tilde{\boldsymbol{\beta}}_0} | \mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0) &= -\frac{1}{nr} \sum_{i=1}^n \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} Var(\xi_i^{*\tilde{\boldsymbol{\beta}}_0} | \mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0) &= E\left(\frac{1}{n^2 r^2 \{\pi_{\delta i}^{\text{app}*}\}^2} \left[\int_0^\tau \left\{ \mathbf{X}_i^* - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\} dM_i^*(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right]^{\otimes 2} \middle| \mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0 \right) \\ &= \frac{1}{n^2 r^2} \sum_{i=1}^n \frac{1}{\pi_{\delta i}^{\text{app}}} \left[\int_0^\tau \left\{ \mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right]^{\otimes 2}. \end{aligned}$$

For every $\epsilon > 0$, we can deduce that

$$\begin{aligned} & E\left(\sum_{i=1}^r \|\xi_i^{*\tilde{\boldsymbol{\beta}}_0}\|^2 I(\|\xi_i^{*\tilde{\boldsymbol{\beta}}_0}\| > \epsilon) \middle| \mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0 \right) \\ &\leq \frac{1}{\epsilon} \sum_{i=1}^r E(\|\xi_i^{*\tilde{\boldsymbol{\beta}}_0}\|^3 | \mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0) \\ &= \frac{1}{r^2 \epsilon} \left\{ \frac{1}{n^3} \sum_{i=1}^n \frac{1}{(\pi_{\delta i}^{\text{app}})^2} \left\| \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\|^3 \right\} \\ &\leq \frac{1}{\delta^2 r^2 \epsilon} \left\{ \frac{1}{n} \sum_{i=1}^n \left\| \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) \right\|^3 \right\} \\ &= o_P(1), \quad \text{as } r \to \infty, \end{aligned}$$

where δ is a factor controlling the mixture proportion in (19), and the last equality is from the assumptions 1 and 3. Therefore, the Lindeberg-Feller conditions are satisfied in probability. By the Lindeberg-Feller central limit theorem (Proposition 2.27 of van der Vaart (1998)), as $r_0 \to \infty$, $r \to \infty$, $n \to \infty$, conditional on $\mathcal{F}_n$ and $\tilde{\boldsymbol{\beta}}_0$, we have

$$\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\beta}}_0}^{-1/2} \mathbf{U}_{\tilde{\boldsymbol{\beta}}_0}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) \xrightarrow{d} N(0, \mathbf{I}), \tag{S.52}$$

where

$$\mathbf{\Gamma}_{\tilde{\boldsymbol{\beta}}_0} = \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{1}{\pi_{\delta i}^{\text{app}}}\left[\int_0^\tau \left\{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right\} dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right]^{\otimes 2} = O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1}).$$

In addition, we need to consider the distance between $\mathbf{\Gamma}_{\tilde{\boldsymbol{\beta}}_0}$ and $\mathbf{\Gamma}$. More specifically,

$$\mathbf{\Gamma}_{\tilde{\boldsymbol{\beta}}_0} - \mathbf{\Gamma} = \frac{1}{nr}\sum_{i=1}^{n}\left[\int_0^\tau \left\{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right\} dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right]^{\otimes 2}\left\{\frac{1}{n\pi_{\delta i}^{\text{app}}} - \frac{1}{n\pi_{\delta i}^{\text{Lopt}}}\right\}, \quad \text{(S.53)}$$

where $\pi_{\delta i}^{\text{app}}$ and $\pi_{\delta i}^{\text{Lopt}}$ are given in (19) and (23), respectively. For notational convenience, we denote

$$\phi_i = \left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}^{0*}(t,\tilde{\boldsymbol{\beta}}_0)\} d\hat{M}_i(t,\tilde{\boldsymbol{\beta}}_0)\right\|,$$

and

$$\psi_i = \left\|\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_i(t,\hat{\boldsymbol{\beta}}_{\text{MPL}})\right\|.$$

Then, we can rewrite the expressions of $\pi_{\delta i}^{\text{app}}$ and $\pi_{\delta i}^{\text{Lopt}}$ with

$$\pi_{\delta i}^{\text{app}} = (1-\delta)\frac{\phi_i}{\sum_{j=1}^n \phi_j} + \delta\frac{1}{n},$$

and

$$\pi_{\delta i}^{\text{Lopt}} = (1-\delta)\frac{\psi_i}{\sum_{j=1}^n \psi_j} + \delta\frac{1}{n},$$

respectively. Note that

$$\begin{aligned}\left|\frac{1}{n\pi_{\delta i}^{\text{app}}} - \frac{1}{n\pi_{\delta i}^{\text{Lopt}}}\right| &= \frac{1}{n\pi_{\delta i}^{\text{app}}\pi_{\delta i}^{\text{Lopt}}}|\pi_{\delta i}^{\text{app}} - \pi_{\delta i}^{\text{Lopt}}| \\ &\le \frac{n(1-\delta)}{\delta^2}\left|\frac{\phi_i}{\sum_{j=1}^n \phi_j} - \frac{\psi_i}{\sum_{j=1}^n \psi_j}\right| \\ &= \frac{(1-\delta)}{\delta^2}\left|\frac{\phi_i n^{-1}\sum_{j=1}^n \psi_j - \psi_i n^{-1}\sum_{j=1}^n \phi_j}{(n^{-1}\sum_{j=1}^n \phi_j)(n^{-1}\sum_{j=1}^n \psi_j)}\right|.\end{aligned}$$

For any $t \in [0,\tau]$, we observe that

$$\begin{aligned}\bar{\mathbf{X}}^{0*}(t,\tilde{\boldsymbol{\beta}}_0) - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}}) &= \bar{\mathbf{X}}^{0*}(t,\tilde{\boldsymbol{\beta}}_0) - \bar{\mathbf{X}}(t,\tilde{\boldsymbol{\beta}}_0) + \bar{\mathbf{X}}(t,\tilde{\boldsymbol{\beta}}_0) - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}}) \\ &= \bar{\mathbf{X}}(t,\tilde{\boldsymbol{\beta}}_0) - \bar{\mathbf{X}}(t,\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2}), \quad \text{(S.54)}\end{aligned}$$

where the last equality is from (S.41). In view of the fact that $\tilde{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_{\text{MPL}} = O_{P|\mathcal{D}_n}(r_0^{-1/2})$, for any $t \in [0, \tau]$ we obtain

$$\begin{aligned}
\bar{\mathbf{X}}(t, \tilde{\boldsymbol{\beta}}_0) - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) &= \frac{S^{(1)}(t, \tilde{\boldsymbol{\beta}}_0)}{S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0)} - \frac{S^{(1)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})}{S^{(0)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})} \\
&= \frac{1}{S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0) S^{(0)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})} \{S^{(0)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) S^{(1)}(t, \tilde{\boldsymbol{\beta}}_0) - S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0) S^{(1)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} \\
&= \frac{1}{S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0) S^{(0)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})} \{S^{(0)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) S^{(1)}(t, \tilde{\boldsymbol{\beta}}_0) - S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0) S^{(1)}(t, \tilde{\boldsymbol{\beta}}_0) \\
&\quad + S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0) S^{(1)}(t, \tilde{\boldsymbol{\beta}}_0) - S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0) S^{(1)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} \\
&= \frac{1}{S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0) S^{(0)}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})} \{S^{(1)}(t, \xi_1)'(\hat{\boldsymbol{\beta}}_{\text{MPL}} - \tilde{\boldsymbol{\beta}}_0) S^{(1)}(t, \tilde{\boldsymbol{\beta}}_0) \\
&\quad - S^{(0)}(t, \tilde{\boldsymbol{\beta}}_0) S^{(2)}(t, \xi_2)(\hat{\boldsymbol{\beta}}_{\text{MPL}} - \tilde{\boldsymbol{\beta}}_0)\} \\
&= O_{P|\mathcal{D}_n}(r_0^{-1/2}), \qquad \text{(S.55)}
\end{aligned}$$

where ξ_1 and ξ_2 are on the segment between $\hat{\boldsymbol{\beta}}_{\text{MPL}}$ and $\tilde{\boldsymbol{\beta}}_0$, and the last equality is from the assumption 3. It follows from (S.54), (S.55) and typically $r_0 < r$ that for any $t \in [0, \tau]$

$$\|\bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0) - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\| = O_{P|\mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0}(r_0^{-1/2}). \qquad \text{(S.56)}$$

Recall that $\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) = N_i(t) - \int_0^t I(\tilde{T}_i \geq u) \exp(\tilde{\boldsymbol{\beta}}_0' \mathbf{X}_i) d\hat{\Lambda}_0^{\text{UNIF}}(u, \tilde{\boldsymbol{\beta}}_0)$ and $M_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) = N_i(t) - \int_0^t I(\tilde{T}_i \geq u) \exp(\hat{\boldsymbol{\beta}}_{\text{MPL}}' \mathbf{X}_i) d\Lambda_0(u)$, it is straightforward to derive that

$$\begin{aligned}
\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) - M_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) &= \exp(\hat{\boldsymbol{\beta}}_{\text{MPL}}' \mathbf{X}_i) \Lambda_0(min\{t, Y_i\}) - \exp(\tilde{\boldsymbol{\beta}}_0' \mathbf{X}_i) \hat{\Lambda}_0^{\text{UNIF}}(min\{t, Y_i\}, \tilde{\boldsymbol{\beta}}_0) \\
&= \exp(\hat{\boldsymbol{\beta}}_{\text{MPL}}' \mathbf{X}_i) \Lambda_0(min\{t, Y_i\}) - \exp(\tilde{\boldsymbol{\beta}}_0' \mathbf{X}_i) \Lambda_0(min\{t, Y_i\}) \\
&\quad + \exp(\tilde{\boldsymbol{\beta}}_0' \mathbf{X}_i) \Lambda_0(min\{t, Y_i\}) - \exp(\tilde{\boldsymbol{\beta}}_0' \mathbf{X}_i) \hat{\Lambda}_0^{\text{UNIF}}(min\{t, Y_i\}, \tilde{\boldsymbol{\beta}}_0) \\
&= \exp(\xi' \mathbf{X}_i) \mathbf{X}_i' (\tilde{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_{\text{MPL}}) \Lambda_0(min\{t, Y_i\}) \\
&\quad - \exp(\tilde{\boldsymbol{\beta}}_0' \mathbf{X}_i) \{\hat{\Lambda}_0^{\text{UNIF}}(min\{t, Y_i\}, \tilde{\boldsymbol{\beta}}_0) - \Lambda_0(min\{t, Y_i\})\} \\
&= O_{P|\mathcal{D}_n}(r_0^{-1/2}), \qquad \text{(S.57)}
\end{aligned}$$

where ξ is between $\tilde{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_{\text{MPL}}$, the last equality is due to $\tilde{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_{\text{MPL}} = O_{P|\mathcal{D}_n}(r_0^{-1/2})$, the assumption 3 and Lemma S.4.

Furthermore, some direct calculations lead to the following expressions:

$$\int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\} d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0)$$

$$
\begin{aligned}
&= \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) + \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\}\{d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) - dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) + dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} \\
&= \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) + \mathbf{R}_1 + \mathbf{R}_2 + \mathbf{R}_3,
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{R}_1 &= \int_0^\tau \{\mathbf{X}_i - \bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\}\{d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) - dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\}, \\
\mathbf{R}_2 &= \int_0^\tau \{\bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\}\{d\hat{M}_i(t, \tilde{\boldsymbol{\beta}}_0) - dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}})\}, \\
\mathbf{R}_3 &= \int_0^\tau \{\bar{\mathbf{X}}(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}) - \bar{\mathbf{X}}^{0*}(t, \tilde{\boldsymbol{\beta}}_0)\} dM_i(t, \hat{\boldsymbol{\beta}}_{\text{MPL}}).
\end{aligned}
$$

Combining the boundedness of $\mathbf{X}_i$'s in $\mathcal{D}_n$, the assumptions 1-3, (S.56), (S.57) and Lemma S.1, we can deduce that $\mathbf{R}_1 = o_P(1)$ as $r_0 \to \infty$. In a similar way, we have $\mathbf{R}_2 = o_P(1)$ and $\mathbf{R}_3 = o_P(1)$. For $i = 1, \cdots, n$, we know

$$
\phi_i = \psi_i + o_P(1), \text{ as } r_0 \to \infty, \tag{S.58}
$$

indicating that conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$,

$$
\frac{1}{n}\sum_{i=1}^n \phi_i = \frac{1}{n}\sum_{i=1}^n \psi_i + o_P(1). \tag{S.59}
$$

Moreover, both (S.58) and (S.59) lead to

$$
\begin{aligned}
\left| \frac{\phi_i}{n}\sum_{j=1}^n \psi_j - \frac{\psi_i}{n}\sum_{j=1}^n \phi_j \right| &\leq |\phi_i - \psi_i| \frac{1}{n}\sum_{j=1}^n \psi_j + \psi_i \left| \frac{1}{n}\sum_{j=1}^n \psi_j - \frac{1}{n}\sum_{j=1}^n \phi_j \right| \\
&= o_P(1).
\end{aligned}
$$

Therefore, as $r_0 \to \infty$ and $r \to \infty$, we get the following conclusion:

$$
\left| \frac{1}{n\pi_{\delta i}^{\text{app}}} - \frac{1}{n\pi_{\delta i}^{\text{Lopt}}} \right| = o_P(1). \tag{S.60}
$$

Combining the assumptions 1-3, (S.53) and (S.60), conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$ we have

$$
\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\beta}}_0} - \boldsymbol{\Gamma} = o_{P|\mathcal{D}_n, \tilde{\boldsymbol{\beta}}_0}(r^{-1}). \tag{S.61}
$$

Based on (S.36), (S.50), (S.51), the Slutsky's theorem, and Theorem 2.7 of van der Vaart (1998), we can derive that

$$\begin{aligned}\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\beta}}_0}^{-1/2}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) &= \boldsymbol{\Gamma}_{\tilde{\boldsymbol{\beta}}_0}^{-1/2}\mathbf{U}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + o_P(1)\\ &\xrightarrow{d} N(0,\mathbf{I}).\end{aligned}$$

From (S.50) and (S.51),

$$\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}} = -\{\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1}). \tag{S.62}$$

Due to the assumption 2 and (S.45), we can derive that

$$\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1} - \boldsymbol{\Psi}^{-1} = -\boldsymbol{\Psi}^{-1}\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}}) - \boldsymbol{\Psi}\}\boldsymbol{\Psi}^{-1} = o_P(1). \tag{S.63}$$

In view of (S.62) and (S.63), we have

$$\begin{aligned}&\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) = -\boldsymbol{\Sigma}^{-1/2}\{\ddot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2})\\ &= -\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) - \boldsymbol{\Sigma}^{-1/2}[\{\ddot{\ell}^*(\hat{\boldsymbol{\beta}}_{\text{MPL}})\}^{-1} - \boldsymbol{\Psi}^{-1}]\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + O_{P|\mathcal{D}_n}(r^{-1/2})\\ &= -\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^{1/2}_{\tilde{\boldsymbol{\beta}}_0}\boldsymbol{\Gamma}^{-1/2}_{\tilde{\boldsymbol{\beta}}_0}\dot{\ell}^*_{\tilde{\boldsymbol{\beta}}_0}(\hat{\boldsymbol{\beta}}_{\text{MPL}}) + o_P(1).\end{aligned} \tag{S.64}$$

From (S.61), as $r_0 \to \infty$ and $r \to \infty$,

$$\begin{aligned}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^{1/2}_{\tilde{\boldsymbol{\beta}}_0})(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^{1/2}_{\tilde{\boldsymbol{\beta}}_0})' &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\beta}}_0}\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}^{-1/2}\\ &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}^{-1/2} + o_P(1)\\ &= \mathbf{I} + o_P(1).\end{aligned}$$

Conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, the Slutsky's theorem, together with (S.52) and (S.64) ensures that as $r_0 \to \infty$ and $r \to \infty$,

$$\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) \xrightarrow{d} N(0,\mathbf{I}).$$

This ends the proof.

Proof of Proposition 1. Conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, it is direct to derive that

$$\|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \le \|\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}\| + \|\hat{\boldsymbol{\beta}}_{\text{MPL}} - \boldsymbol{\beta}_0\|$$

$$= O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2}) + O_P(n^{-1/2}),$$

where the equality is due to Eq. (S.51). i.e., $\|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_{P|\mathcal{D}_n,\tilde{\boldsymbol{\beta}}_0}(r^{-1/2})$. It follows from Proposition 2 of Wang *et al.* (2022) that

$$\|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(r^{-1/2}).$$

Next, we prove the asymptotic normality of $\breve{\boldsymbol{\beta}}$ with respect to the true parameter. Note that

$$\begin{aligned} r^{1/2}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= r^{1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) + r^{1/2}(\hat{\boldsymbol{\beta}}_{\text{MPL}} - \boldsymbol{\beta}_0) \\ &= r^{1/2}(\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}) + o_P(1), \end{aligned}$$

where the last equality is due to $\hat{\boldsymbol{\beta}}_{\text{MPL}} - \boldsymbol{\beta}_0 = o_P(n^{-1/2})$ and the assumption $r = o(n)$. Hence, conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$, the asymptotic distribution of $\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ is the same as that of $\breve{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{MPL}}$. That is to say, conditional on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$ we have

$$\boldsymbol{\Sigma}^{-1/2}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{I}), \tag{S.65}$$

where $\boldsymbol{\Sigma}$ is given in Theorem 3. Based on Proposition 2 of Wang *et al.* (2022), the asymptotic normality in (S.65) also holds without conditioning on $\mathcal{D}_n$ and $\tilde{\boldsymbol{\beta}}_0$. This completes the proof.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 4, 1100–1120.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 2, 269–276.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

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

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Wang, J., Zou, J., and Wang, H. (2022). Sampling with replacement vs poisson sampling: a comparative study in optimal subsampling. *IEEE Transactions on Information Theory* **68**, 6605–6630.

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

Xu, Q., Paik, M. C., Luo, X., and Tsai, W.-Y. (2009). Reweighting estimators for cox regression with missing covariates. *Journal of the American Statistical Association* **104**, 1155–1167.