https://doi.org/10.1016/j.jmva.2018.04.005


# On the sign consistency of the Lasso for the high-dimensional Cox model

Shaogao Lv, Huazhen Lin

*Center of Statistical Research, School of Statistics,*

*Southwestern University of Finance and Economics, Chengdu, China, 611130*

and Jian Huang

*Department of Applied Mathematics*

*The Hong Kong Polytechnic University, Hung Hom, Hong Kong*

*and Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa, USA*

**Abstract**

In this paper we investigate the behavior of the $\ell_1$-penalized partial likelihood estimation for the sparse high dimensional Cox's proportional hazards model. Particular, we investigate how the $\ell_1$-penalized partial likelihood estimation recovers the sparsity pattern and the conditions under which the sign support consistency is guaranteed. We establish sign recovery consistency and $\ell_\infty$-error bounds for the Lasso partial likelihood estimator under mild and interpretable conditions, including mutual incoherence conditions. More importantly, we show that the conditions of the incoherence and bounds on the minimal non-zero coefficients are necessary, which providing significant and instructional implications for understanding the proposed methods.

**Key Words and Phrases:** Sparse recovery; Cox proportional hazard model; Oracle property; Empirical Process; Mutual coherence; the Lasso.

## 1 Introduction

High-dimensional data, including high-throughput genomic data and credit risk data, are becoming increasingly available as data collection technology evolves. Finding significant genetic risk factors for clinical outcomes, such as the age of disease onset or time to death, is fundamental to modern biostatistics, since survival outcomes are most often clinical endpoints. In view of the central role of the Cox proportional model in survival analysis, its widespread applications and the proliferation of ultra-high-dimensional covariates, it is quite interesting to study high-dimensional theory in the Cox model.

When the number of features $p$ far exceeds the sample size, without any additional structure, it is known that many standard approaches, for example, least squared error, classification by

linear or quadratic discriminant analysis and principal component analysis, are not consistent unless the ratio $p/n$ goes to zero. To handle high dimensional problems, various sparse models which aim to select only a small set of relevant variables from a huge number of features, have become more and more popular owing to the model interpretability, theory support and efficient computation.

Regularization method has been a powerful tool of sparse modelling and variable selection. Depending on the type of penalty functions that are used, the regularization methods can be grouped into two classes: convex and non-convex. A typical example of a convex penalty is the $\ell_1-$penalty or Lasso-type penalty, which gives rise to the $\ell_1-$regularization methods (Tibshirani, 1996). The convexity of Lasso-type penalty methods makes the implementation efficient and facilitates the theoretical analysis. Lasso-type method has been viewed as a standard approach to solve sparse problems. During the last decade, much work on the Lasso-type methods have appeared greatly beyond linear regression, and particularly Tibshirani (1997) initially proposed the Lasso programme for the Cox model. In recent years, we have witnessed several work on the use of $\ell_1$-constraints for the Cox related models in the presence of sparsity pattern. For example, Kong et al. (2014) took an approach of V., D. Geer (2008) to derive prediction and $\ell_1$ estimation error bounds for the Lasso in the Cox models. Lemler (2012) considered the joint estimation of the baseline hazard function and regression coefficients in the Cox model, and established theoretical guarantees for the prediction performance and error bounds of the Lasso estimator under very weak conditions and some incoherent conditions, respectively. Under natural extensions of the compatibility and cone invertibility of the Hessian matrix, Huang et al. (2013) established $\ell_q$ estimation error bounds of the Lasso estimator for the sparse Cox proportional hazard model with time-dependent covariates when $q \geq 1$.

Despite the aforementioned developments, some important high-dimensional theory that can provide strong performance guarantees for the Lasso estimator in the sparse Cox model is still lacking. Specifically, issues related to sparse pattern recovery has not been addressed to date. The problem of sparse pattern recovery can be stated as follows: given sparse, how to recover the positions of its non-zero entries, and further how to guarantee the sign support consistency of estimator. This problem, also known as support recovery or model selection consistency, arises in a variety of contexts, including compressed sensing (Donoho, 2006), sparse approximation (DeVore and Lorentz, 1993) and structure estimation in graphical models (Merinshausen and Buhlmann, 2006). Efficient discovery of sparsity patterns is a central concern in high dimensional data, and considerable effort for least square setting has been devoted, such as Zhao and Yu (2006), Merinshausen and Buhlmann (2006), Zhang and huang (2008), and Wainwright (2009).

It is known that $\ell_q$ estimation consistency does not guarantee exact recovery of the underlying sparsity pattern, and studying the problem of sparse pattern recovery for survival models is more involved than that for least square setting, since high dimensional inference for the Cox model involves not only the dynamic processes, semi-parametric nature and censoring mechanism of survival data, but also the non-quadratic Hessian matrix. A mathematical and

systematic study of statistical inference on high dimensional survival models has only started in recent years. In particular, Lin and Lv (2013) considered the additive hazard model in high dimensional setting, and investigated the sparse pattern recovery and estimation problems based on a class of general penalties, including the Lasso. A critical difference from the Cox model, the additive hazard model leads to a least square-type loss function and hence the key quantity (the Hessian matrix) for statistical inference is independent of the estimating coefficient. By contrast, the Hessian matrix involved in the Cox model is a function of estimating coefficient, which is more involved and needs new technical tools. For the high-dimensional Cox model, Bradic et al. (2011) considered estimation as well as variable selection and oracle properties using general concave penalties, including the Lasso as well. Although theirs results are quite broad and provide deep insights on the performance of various regularization methods in the Cox model, they required very strict conditions for deriving theoretical results, for example, they imposed a *random* condition on a large *empirical* covariance matrix in a *neighborhood* of the true regression coefficient. Moreover, all of the above mentioned work only provide sufficient conditions to guarantee oracle properties, and they do not consider the necessity of those conditions they imposed. Since most of conditions for high dimensional inference are often hard to verify, providing sufficient and necessary conditions for statistical inference has significant and instructional implications for understanding the proposed methods.

In view of the growing importance of finding significant genetic risk factors for survival outcomes, this paper aims at establishing a complete characterization for support recovery consistency of the Lasso Cox estimator. Under mild, interpretable conditions, we first establish the oracle property of the Lasso estimator, specially we provide sufficient conditions for sign recovery to succeed with high probability. Two of critical conditions to guarantee sign recovery consistency refer to *mutual incoherence condition* (Zhao and Yu, 2006) and *the minimum value of the true coefficient*. Furthermore, we can show that, the sign recovery will fail with probability at least 1/2 if either of two above conditions does not hold. To the best of our knowledge, there is no work on this topic in survival data.

The rest of this paper is organized as follows. In Section 2, we review the Cox proportional hazard model and then introduce the penalized partial likelihood with the $\ell_1$ penalty. The theoretical properties of the Lasso Cox estimator are studied in Section 3. In Section 4, we give an outline of proof of theoretical guarantee of performance of the method. Most of proof and technical details are relegated to the Appendices.

**Notation:** We collect here some standard notation used throughout the paper. For a vector $\boldsymbol{\alpha}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, we use the usual inner product of $\mathbb{R}^p$, given by $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle = \boldsymbol{\alpha}'\boldsymbol{\beta}$, and $\|\boldsymbol{\alpha}\|_2 = \sqrt{\langle \boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle}$ is denoted to be its $\ell_2$ norm. Similarly the $\ell_1$-norm is given by $\|\boldsymbol{\alpha}\|_1 = \sum_{j=1}^p |\alpha_j|$ with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)'$. For some subset $A \subseteq \{1, ..., p\}$, we denote by $\|\boldsymbol{\alpha}_A\|_\infty = \max_{j \in A} |\alpha_j|$. For a vector $z \in \mathbb{R}^p$ and a subset $S \subseteq \{1, ..., p\}$, we write $z_S \in \mathbb{R}^S$ to denote the vector $z$ restricted to $S$. Given sequence $f(n)$ and $g(n)$, the notion $f(n) = O_p(g(n))$ means that there exists a constant

$c$ such that $f(n) \leq cg(n)$; Similarly, $f(n) = o_p(g(n))$ means that $f(n)/g(n)$ goes to zero as $n \to \infty$, and sometimes we write it by $f(n) \ll g(n)$. For a matrix $M = (M_{ij})$, we define the spectral norm, given by $\|M\|_2 = \max_{\|x\|_2=1} \|Mx\|_2$. Also we write $\|M\|_\infty = \max_j \sum_i |M_{ij}|$.

# 2 Problem formulation

In this section, we formulate the setting and the estimator, and state some key quantities used for our analysis.

## 2.1 The sparse Cox's proportional hazards model

We now briefly reviewing the Cox's proportional hazards model. Following Andersen and Gill (1982), consider an $n$-dimensional counting process $\mathbf{N}^{(n)}(t) = (N_1(t), N_2(t), ..., N_n(t))'$, $t \geq 0$, where $N_i(t)$ is the number of observed events in the time interval $[0, t]$ for the $i$-th individual. The sample paths of $N_1(\cdot), ..., N_n(\cdot)$ are step functions, zero at $t = 0$, with jumps of size 1 only. Furthermore, no two components jump at the same time. For $t \geq 0$, let $\mathcal{F}_t$ be the $\sigma$-filtration representing all the information available up to time $t$. Assume that for $\{\mathcal{F}_t, t \geq 0\}$, $\mathbf{N}^{(n)}(\cdot)$ has predictable compensator $\mathbf{\Lambda}^{(n)} = (\Lambda_1, \Lambda_2, ..., \Lambda_n)'$ with

$$d\Lambda_i(t) = Y_i(t) \exp\left(\mathbf{Z}_i(t)'\boldsymbol{\beta}^o\right) d\Lambda_0(t), \tag{2.1}$$

where $\mathbf{Z}_i(t) = (Z_{i1}(t), Z_{i2}(t), ..., Z_{ip}(t))'$ is a $p$-dimensional vector-valued predictable covariate process, and $\boldsymbol{\beta}^o$ is the true regression coefficient associated with $p$-dimensional features. $\Lambda_0(\cdot)$ is an unknown baseline cumulative hazard function and, for each $i$, $Y_i(t) \in \{0, 1\}$ is a predictable at risk indicator process, which can be constructed from data. In this setting, we can always use the natural filtration of the processes, that is, $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), \mathbf{Z}_i(s); s \leq t, i = 1, ..., n\}$. In the case of Cox model, we assumes that the vector $\boldsymbol{\beta}^o$ is sparse, in the sense that the cardinality $s_o = |S(\boldsymbol{\beta}^o)|$ satisfies $s_o \ll p$, where $S(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$.

## 2.2 Maximum partial likelihood estimator with the Lasso

This paper focuses on the maximum partial likelihood estimator with the $\ell_1$ penalty. For this purposes, define logarithm of the Cox partial likelihood for survival experience at time $t$,

$$C(\boldsymbol{\beta}; t) = \sum_{i=1}^{n} \int_0^t \mathbf{Z}_i(s)'\boldsymbol{\beta} dN_i(s) - \int_0^t \log\Big[\sum_{i=1}^{n} Y_i(s) e^{\mathbf{Z}_i(s)'\boldsymbol{\beta}}\Big] d\bar{N}(s),$$

where $\bar{N} = \sum_{i=1}^n N_i$. The Lasso programme for the Cox model is to minimize a $\ell_1$-penalized negative log-partial likelihood criterion, given by

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) = \ell(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.2}$$

where $\ell(\boldsymbol{\beta}) = -C(\boldsymbol{\beta};\tau)/n$ and $\lambda$ is a penalty parameter. $\tau$ is a finite experiment time, the data up to that are frequently used. Since the minimization problem for (2.2) is a convex programme, a global estimator of the Lasso programme (2.2) always exists, denoted by $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \{\mathcal{L}(\boldsymbol{\beta}, \lambda)\}. \tag{2.3}$$

### 2.3 Additional useful notation

To facilitate our theoretical analysis, we introduce the gradient vector and Hessian matrix of the Cox model. The following notation on a vector/matrix is useful before giving several critical quantities for our analysis. For a vector $\mathbf{v}$, we write $\mathbf{v}^{\otimes 0} = 1 \in \mathbb{R}$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}'$. For any vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we define

$$S^{(k)}(t,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i(t)^{\otimes\kappa} Y_i(t) e^{\mathbf{Z}_i(t)'\boldsymbol{\beta}}, \quad \kappa = 0,1,2. \quad \bar{\mathbf{Z}}_n(t,\boldsymbol{\beta}) = \frac{S^{(1)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})},$$
$$V_n(t,\boldsymbol{\beta}) = \sum_{i=1}^{n} w_{ni}(t,\boldsymbol{\beta})(\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_n(t,\boldsymbol{\beta}))^{\otimes 2} = \frac{S^{(2)}(t,\boldsymbol{\beta})}{S^{(0)}(t,\boldsymbol{\beta})} - \bar{\mathbf{Z}}_n(t,\boldsymbol{\beta})^{\otimes 2},$$

where $w_{ni}(t,\boldsymbol{\beta}) = Y_i(t)\exp(\mathbf{Z}_i(t)'\boldsymbol{\beta})/[nS^{(0)}(t,\boldsymbol{\beta})]$. Note that $S^{(0)}$ is a scalar, $S^{(1)}$ and $\bar{\mathbf{Z}}$ are $p-$dimensional vectors, and $S^{(2)}$ and $V_n$ are $p \times p$ matrices. It has been shown as in Andersen et al. (1982) that the gradient of $\ell(\boldsymbol{\beta})$ is represented as

$$\dot{\ell}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \big[\mathbf{Z}_i(s) - \bar{\mathbf{Z}}_n(s,\boldsymbol{\beta})\big] dN_i(s), \tag{2.4}$$

and the Hessian matrix of $\ell(\boldsymbol{\beta})$ is

$$\ddot{\ell}(\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau V_n(s,\boldsymbol{\beta}) dN_i(s). \tag{2.5}$$

Since $S^{(k)}$ and $V_n$ depend on random sample, we need to define their population counterparts for theoretical analysis. Define

$$\mathbf{s}^{(k)}(t,\boldsymbol{\beta}) = \mathbb{E}\big[\mathbf{Z}(t)^{\otimes\kappa} Y(t) e^{\mathbf{Z}(t)'\boldsymbol{\beta}}\big], \quad \kappa = 0,1,2,$$
$$\mathbf{e}(t,\boldsymbol{\beta}) = \mathbf{s}^{(1)}(t,\boldsymbol{\beta})/\mathbf{s}^{(0)}(t,\boldsymbol{\beta}).$$

Then we can define the population counterparts of $\ddot{\ell}(\boldsymbol{\beta})$ by

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \mathbb{E}\Big[\int_0^\tau \Big(\frac{\mathbf{s}^{(2)}(s,\boldsymbol{\beta})}{\mathbf{s}^{(0)}(s,\boldsymbol{\beta})} - \mathbf{e}(s,\boldsymbol{\beta})^{\otimes 2}\Big) dN(s)\Big].$$

The matrix $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}(\boldsymbol{\beta}^o)$ characterizes the covariance structure of the model (2.1) and will play a critical role in our high dimensional analysis.

In this paper, we are concerned with the problem of signed support recovery, defined explicitly as follows. For any vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we define its sign vector

$$\operatorname{sign}(\beta_j) := \begin{cases} +1 & \beta_j > 0 \\ -1 & \beta_j < 0 \\ 0 & \beta_j = 0. \end{cases}$$

Despite the empirical success of the Lasso estimators for various sparse models, relatively little theory is available to explain why they work and which conditions are essential to guarantee their empirical performance. One of contributions of this paper is to provide sufficient and necessary conditions for the Lasso scheme (2.3), so that the signed support $\operatorname{sign}(\hat{\boldsymbol{\beta}}) = \operatorname{sign}(\boldsymbol{\beta}^o)$ is guaranteed with high probability.

# 3 Primal dual witness technique and statistical theory

In this section, we first construct a biased oracle estimator based on the primal-dual witness proof technique (PDW). Then we give two useful lemmas to describe the solutions of the Lasso programme (2.3). In Subsection 3.2, we introduce some technical conditions for building analysis framework, and then establish estimation error of the biased oracle estimator. Based on these results, we state the weak oracle property of the Lasso estimator in Subsection 3.3, and in Subsection 3.4 we show that the mutual coherent condition and the minimum value of the true nonzero coefficients are essential to guarantee the corrected sign recovery.

## 3.1 Primal dual witness construction

We now outline the main steps of the PDW, developed by Wainwright (2009), which we will use to establish support recovery for the program (2.3). Define the active set $S = S(\boldsymbol{\beta}_0) = \{j : \beta_j^o \neq 0\}$, and its supplementary $S^c = \{j : \beta_j^o = 0\}$. Without loss of generality, we assume that the last $p - s_o$ components of $\boldsymbol{\beta}^o$ are 0, i.e. $\boldsymbol{\beta}^o = ((\boldsymbol{\beta}_1^o)', \mathbf{0}')'$. A vector $z \in \mathbb{R}^p$ is a subgradient for the $\ell_1$-norm evaluated at $\boldsymbol{\beta}$, written as $z \in \partial\|\boldsymbol{\beta}\|_1$, if the elements satisfy the following relation:

$$z_j = \operatorname{sign}(\beta_j) \text{ if } \beta_j \neq 0, \text{ and } z_j \in [-1, 1], \text{ otherwise.}$$

Recall that the key steps of the PDW argument consists of the following procedures:

*Step 1:* Acting as if the true sparsity structure is known in advance, the biased oracle estimator is defined as $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_S, \mathbf{0})$, where

$$\check{\boldsymbol{\beta}}_S = \min_{\boldsymbol{\beta}_S \in \mathbb{R}^S} \big\{ \ell\big((\boldsymbol{\beta}_S, \mathbf{0})\big) + \lambda \|\boldsymbol{\beta}_S\|_1 \big\}. \tag{3.1}$$

Note that we here impose the additional constraint such that $S(\check{\boldsymbol{\beta}}) \subseteq S(\boldsymbol{\beta}^o) = S$. The solution to this restricted convex program (3.1) is guaranteed to be unique under the invertibility condition on $\ddot{\ell}_{SS}(\boldsymbol{\beta}_S, \mathbf{0})$.

*Step 2:* We choose $\check{z}_S \in \mathbb{R}^S$ as an element of the subdifferential of the $\|\cdot\|_1$ norm evaluated at $\check{\boldsymbol{\beta}}_S$. By definition of subgradient, we see that $\|\check{z}_S\|_\infty \leq 1$.

*Step 3:* We solve for a vector $\check{z}_{S^c} \in \mathbb{R}^{p-s_o}$ to satisfy the zero subgradient condition

$$\dot{\ell}(\check{\boldsymbol{\beta}}) + \lambda \check{z} = 0, \tag{3.2}$$

where $\check{z} = (\check{z}_S, \check{z}_{S^c})$. Then we check whether or not the dual feasibility condition $|\check{z}_j| \leq 1$ for all $j \in S^c$ is satisfied. (For ensuring uniquenss, we need to check strict dual feasibility of $\check{z}_S$, that is, $\max_{j \in S^c} |\check{z}_j| < 1$.)

*Step 4:* We check whether the sign consistency condition $\check{z}_S = \text{sign}(\boldsymbol{\beta}_1^o)$ is satisfied.

Note that, in high dimensions, the Lasso programme (2.3) is not guaranteed to be strictly convex, so there may be multiple solutions generated by (2.3). We next consider the properties of the lasso estimators based on (2.3).

**Lemma 1.** *(a) A vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is optimal if and only if there exists a subgradient vector $\hat{z} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big[\mathbf{Z}_i(s) - \bar{\mathbf{Z}}_n(s, \hat{\boldsymbol{\beta}})\big] dN_i(s) - \lambda \hat{z} = 0. \tag{3.3}$$

*(b) Suppose that the subgradient vector satisfies the strict dual feasibility condition $|\hat{z}_j| < 1$ for all $j \notin S(\hat{\boldsymbol{\beta}})$, then any optimal solution $\tilde{\boldsymbol{\beta}}$ to the Lasso programme (2.3) satisfies $\tilde{\beta}_j = 0$ for all $j \notin S(\hat{\boldsymbol{\beta}})$.*

*(c) Under the conditions of part (b). If the $|S(\hat{\boldsymbol{\beta}})| \times |S(\hat{\boldsymbol{\beta}})|$ matrix $\ddot{\ell}_{S(\hat{\boldsymbol{\beta}})S(\hat{\boldsymbol{\beta}})}(\boldsymbol{\beta})$ is invertible, then $\hat{\boldsymbol{\beta}}$ is the unique optimal solution of the Lasso program.*

The above lemma shows that the solution of (2.3) is unique to the support recovery , provided that the strict dual feasibility condition in (3.3) holds and the restricted Hessian matrix over $\mathbb{R}^{S(\hat{\boldsymbol{\beta}})}$ is invertible.

### 3.2 Estimation error of the oracle estimator $\check{\boldsymbol{\beta}}$

In this subsection, we are ready to establish estimation error for the Lasso Cox estimator $\check{\boldsymbol{\beta}}$. To this end, we need the following assumptions to facilitate our theoretical analysis.

**Assumption 1.** (1) Suppose that $\{Y_i(t), \mathbf{Z}_i(t), N_i(t),\ t \geq 0\}$, $i = 1, \cdots, n$ are i.i.d. time-dependent sample from the underlying process $\{Y(t), \mathbf{Z}(t), N(t),\ t \geq 0\}$. (2) $\mathbb{P}[\max_i\{N_i(\tau)\} \leq 1] = 1$. (3) No two components of $N_i$'s jump at the same time.

**Assumption 2.** (1) There exists some constant $K$ such that

$$\sup_{0\le t\le\tau}\max_{j\le p}|Z_{ij}(t)-Z_{i'j}(t)|\le K,\quad \text{for all}\, i<i'\le n.$$

(2) $\Lambda_0(\tau)<\infty$. (3) $P\{Y(\tau)=1\}>0$. (4) The sample path of $Z_j(\cdot)$, $j=1,...,p$, are of uniformly bounded variation.

**Assumption 3.** Suppose that $\big\|\boldsymbol{\Sigma}_{SS}\big\|_2=O_p(1)$ and $\big\|(\boldsymbol{\Sigma}_{SS})^{-1}\big\|_2=O_p(1)$.

Assumption 1 and parts (2) and (3) of Assumption 2 are standard for survival models (Andersen and Gill, 1982); part (1) of Assumption 2 controls the behavior of the covariates and such condition for linear regression has been imposed on the deterministic Gram design. Actually, our analysis still holds under a relaxed condition such as sub-Gaussian ensembles, but this will complicate our proof; part (4) of Assumption 2 is a mild technique condition that will control entropy integrals involved empirical process, which has been imposed in Lin et al. (2013). Assumption 3 is also a standard condition for the Cox model, which holds obviously when $s_o$ is fixed. This condition has also been required by Bradic et al. (2011) under similar settings.

To estimate $\|\check{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_\infty$, we only need to consider the subvector in the first $s_o$ component, that is $\|\check{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_1^o\|_\infty$, because $\check{\boldsymbol{\beta}}_{S^c}=\boldsymbol{\beta}^o_{S^c}=\mathbf{0}$. Hence, the consistency of the estimator $\check{\boldsymbol{\beta}}$ can be obtained from the consistent result of Theorem 3.1 (Huang et al., 2013) over the restricted space $\mathbb{R}^S$.

**Lemma 2** (Estimation Error)**.** *Consider the local estimator $\check{\boldsymbol{\beta}}_S$ generated by the restricted Lasso programme* (3.1)*. Suppose that Assumptions 1-3 hold over $\mathbb{R}^S$, then, in the event $\big\|\dot{\ell}\big(\boldsymbol{\beta}^o\big)_S\big\|_\infty=O_p(\lambda)$, we have*

$$\|\check{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_1^o\|_2=O_p\big(\sqrt{s_o}\lambda\big).$$

*Moreover, let $\lambda=O_p\big(\sqrt{\log(p/\delta)/n}\big)$ with a small $0<\delta<1$, then with probability at least $1-\delta$,*

$$\|\check{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_1^o\|_2=O_p\big(\sqrt{s_o\log(p/\delta)/n}\big).$$

The proof of Lemma 2 can be found in Appendix A. By Lemma 1(a), we can see that, in order to prove $\check{\boldsymbol{\beta}}$ is an optimum of the Lasso programme (2.3), we still need to show that $\check{z}$ is one of subgradients of the $\ell_1$-norm at $\check{\boldsymbol{\beta}}$.

### 3.3 Weak oracle property

An estimator is said to have the weak oracle property if it is both estimation consistency and model selection consistency, proposed originally by Lv et al. (2009). By the part (a) and (b) of Lemma 1, to derive the weak oracle property of the Lasso estimator defined in (2.3), we shall provide a sufficient condition to ensure that $\hat{z}\in\partial\|\hat{\boldsymbol{\beta}}\|_1$ and $\max_{j\in S(\hat{\boldsymbol{\beta}})^c}|\hat{z}_j|<1$.

**Assumption 4.** There exists $\gamma\in(0,1]$ such that $\big\|\boldsymbol{\Sigma}_{S^cS}(\boldsymbol{\Sigma}_{SS})^{-1}\big\|_\infty\le 1-\gamma$.

Assumption 4 is an analog of mutual incoherent conditions related to the covariance matrix $\mathbf{\Sigma}$, which have been considered in linear regression (Zhao and Yu, 2006) and in additive hazards regression (Lin and Lv, 2013). It should be pointed out, the restriction on the correlation structure in Assumption 4 is more complex, compared to linear or general additive models. In fact, the matrix $\mathbf{\Sigma}$ depends on the failure process and censoring mechanism, as well as the distribution of covariates. Although Assumption 4 is stringent in some sense, we will show in Theorem 3 that such incoherent condition is necessary to guarantee support recovery of the Cox model via the Lasso. As a consequence, the ability of conducting variable selection via the Lasso programme is affected by the interplay of several factors, including the the distribution of covariates, failure process and censoring mechanism.

We now state general theoretical results regarding the proposed estimator, which plays a critical role in deriving oracle properties.

**Theorem 1.** *Under Assumptions 1-3 as above.*

*(a) If Steps 1 through 3 of the PDW method succeed with strict dual feasibility in Step 3, then the Lasso* (2.3) *has a unique solution given by* $\check{\boldsymbol{\beta}}$ *with* $S(\check{\boldsymbol{\beta}}) \subseteq S$.

*(b) If Steps 1 through 4 of the PDW method succeed with strict dual feasibility in Step 3. When* $|\check{\beta}_j| > 0$ *is satisfied for* $j \in S$, $\check{\boldsymbol{\beta}}$ *is the unique solution of the Lasso* (2.3)*, such that the corrected signed support holds, i.e.* $sign(\check{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}^o)$.

*(c) Conversely, if either Step 3 or 4 of the PDW method fails, then the Lasso fails to recover the corrected signed support.*

It is seen from Theorem 1, the main task in the primal-dual witness construction lies in verifying the dual feasibility condition in Step 3, and the sign consistency condition in Step 4.

Note also that part(b) of Theorem 1 hold only when $|\check{\beta}_j| > 0$ $(j \in S)$ is satisfied, which can be verified as long as the lower bound of $\{|\beta_j^o|, j \in S\}$ is not small sufficiently. This will be shown clearly in Theorem 2 below.

Combining Theorem 1 and technical lemmas presented in Appendix, we can state the weak oracle property of the proposed estimator.

**Theorem 2.** *Under Assumptions 1-4 and* $\mathbb{E}\big[\big\|\mathbf{Z}(t)_{SS}^{\otimes 2}\big\|_2\big]$ *is bounded uniformly on* $t \in [0, \tau]$. *If* $n \gg s_o^3 \log(p)$, *then with probability at least* $1 - \delta$, *the Lasso programme* (2.3) *has a unique optimum given by* $\check{\boldsymbol{\beta}}$, *such that*

$(a)$ *(Sparity)* $\check{\boldsymbol{\beta}}_{S^c} = \mathbf{0}$;
$(b)$ *($\ell_\infty$-loss)* $\|\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^o\|_\infty = O_p\big(\sqrt{s_o \log(p/\delta)/n}\big)$.

*If additionally, we have a lower bound of the form* $\min_{j \in S} |\beta_j^o| \gg \sqrt{s_o \log(p/\delta)/n}$, *then it is guaranteed that* $\check{\boldsymbol{\beta}}$ *is sign-consistent for* $\boldsymbol{\beta}^o$.

The proof of Theorem 2 is relegated to Section 4.1. Due to semi-parametric structure and censoring mechanism, we require constraints $n \gg s_o^3 \log(p)$ to guarantee sparse recovery. This sufficient condition on $p$, $s_o$ and $n$ implies that the Lasso Cox estimator can handle a nonpolynomially growing dimension of covariates as high as $p = o_p(\exp(n/s_o^3))$, and the dimension of the true sparse model growing as $s_o = o_p(n^{1/3})$. Theorem 2 guarantees that $\check{\boldsymbol{\beta}}$ is the unique sign-consistent estimator for $\boldsymbol{\beta}^o$, as long as $\min_{j\in S} |\beta_j^o| \gg \sqrt{s_o \log(p)/n}$ are satisfied.

### 3.4 Necessary conditions for sign recovery consistency

Now we turn to the results related to the failure of the sign recovery consistency, providing that either mutual incoherence condition or the constraint on minimum value of the true coefficient is violated.

**Theorem 3.** *Suppose that Assumptions 1-3 hold.*

*(a) If mutual incoherence condition, that is Assumption 4, is violated, namely,*

$$\max_{j\in S^c} \left| e_j' \boldsymbol{\Sigma}_{S^c S} (\boldsymbol{\Sigma}_{SS})^{-1} sign(\boldsymbol{\beta}_1^o) \right| = 1 + \nu, \quad \nu > 0, \tag{3.4}$$

*then for any $\lambda > 0$ and sufficiently large $n$, the probability in which sign support recovery fails is no less than $1/2$, that is,*

$$\mathbb{P}\big[ sign(\hat{\boldsymbol{\beta}}) \neq sign(\boldsymbol{\beta}^o) \big] \geq 1/2.$$

*(b) For each $j \in S$, we define the quantity*

$$h(\lambda) := \lambda e_j' (\boldsymbol{\Sigma}_{SS})^{-1} sign(\boldsymbol{\beta}_1^o).$$

*If there exists $j \in S$, we have the inclusion $\beta_j^o \in (0, h(\lambda))$ or the inclusion $\beta_j^o \in (h(\lambda), 0)$. Then there also holds*

$$\mathbb{P}\big[ sign(\hat{\boldsymbol{\beta}}) \neq sign(\boldsymbol{\beta}^o) \big] \geq 1/2.$$

Theorem 3(a) implies that, the mutual incoherence condition in Assumption 4 is essential to ensure the corrected sign recovery for the Lasso partial likelihood estimator. The same conclusion has been found in linear regression model with high dimensions (Wainwright, 2009) and linear regression model with fixed dimension (Zhao and Yu, 2006). For sign consistency, Theorem 3(b) indicates that the value $\min_{j\in S} |\beta_j^o|$ cannot decay to zero faster than the penalty parameter $\lambda$.

## 4 Proof for main theorems

A key step for the proof of Theorem 2 is to verify the strict dual feasibility condition In PDW procedure. To this end, We first derive two sufficient conditions to guarantee the strict dual

feasibility condition. Then, we shall prove that these two conditions are satisfied under our Assumptions 1-4. To prove Theorem 3, we need an application of martingale central limit theorem.

### 4.1 Proof of Theorem 2

In this section, we first derive conditions that allows us to establish the strict dual feasibility conditions required so as to apply Theorem 1.

By the zero-sbugradient condition (3.2), we rewrite it as

$$\dot{\ell}(\check{\boldsymbol{\beta}}) - \dot{\ell}(\boldsymbol{\beta}^o) + \dot{\ell}(\boldsymbol{\beta}^o) + \lambda \check{z} = 0.$$

Let $\widehat{Q} := \int_0^1 \ddot{\ell}\big(\boldsymbol{\beta}^o + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\big)d\theta$, then one gets $\widehat{Q}[\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o] + \dot{\ell}(\boldsymbol{\beta}^o) + \lambda \check{z} = 0$. Since $S(\check{\boldsymbol{\beta}}) \subseteq S$, the above equality can be rewritten with a block form

$$\begin{bmatrix} \widehat{Q}_{SS} & \widehat{Q}_{SS^c} \\ \widehat{Q}_{S^cS} & \widehat{Q}_{S^cS^c} \end{bmatrix} \begin{bmatrix} \check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^o \\ 0 \end{bmatrix} + \begin{bmatrix} \dot{\ell}(\boldsymbol{\beta}^o)_S \\ \dot{\ell}(\boldsymbol{\beta}^o)_{S^c} \end{bmatrix} + \lambda \begin{bmatrix} \check{z}_S \\ \check{z}_{S^c} \end{bmatrix} = 0. \tag{4.1}$$

By computing the top block of the equation (4.1), we have

$$\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^o = -(\widehat{Q}_{SS})^{-1}\big[\dot{\ell}(\boldsymbol{\beta}^o)_S + \lambda \check{z}_S\big]. \tag{4.2}$$

Furthermore, a simple algebraic manipulations for the bottom of (4.1) yields that

$$\check{z}_{S^c} = \frac{1}{\lambda}\left\{\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big[\dot{\ell}(\boldsymbol{\beta}^o)_S + \lambda \check{z}_S\big] - \dot{\ell}(\boldsymbol{\beta}^o)_{S^c}\right\}. \tag{4.3}$$

Based on the equality (4.3), we have the following result:

**Proposition 1.** *Under the PDW construction of* (3.2)*, the strict dual feasibility holds, provided that $\lambda$ is chosen to satisfy the following inequalities*

$$\big\|\dot{\ell}(\boldsymbol{\beta}^o)\big\|_\infty \leq \frac{\gamma}{8+2\gamma}\lambda, \tag{4.4}$$

*and*

$$\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty \leq 1 - \gamma/2. \tag{4.5}$$

*Proof.* Since $\|\check{z}_S\|_\infty \leq 1$ by definition, the equation (4.3) is applied to yield that

$$\begin{aligned} \|\check{z}_{S^c}\|_\infty &\leq \frac{1}{\lambda}\big\|\dot{\ell}(\boldsymbol{\beta}^o)\big\|_\infty + \frac{1}{\lambda}\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^o)_S\big\|_\infty + \big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty \\ &\leq \frac{1}{\lambda}\big\|\dot{\ell}(\boldsymbol{\beta}^o)\big\|_\infty + \big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty\Big(1 + \frac{1}{\lambda}\big\|\dot{\ell}(\boldsymbol{\beta}^o)_S\big\|_\infty\Big) \\ &\leq 1 - \frac{\gamma}{4}, \end{aligned}$$

provided that (4.4) and (4.5) hold simultaneously. □

By Proposition 1, in order to verify the strict dual feasibility condition, we need to give an appropriate bound for $\|\dot{\ell}(\boldsymbol{\beta}^o)\|_\infty$ and provide sufficient conditions to guarantee that (4.5) holds.

To bound $\|\dot{\ell}(\boldsymbol{\beta}^o)\|_\infty$, we introduce additional notation on martingales. Since $\boldsymbol{\Lambda}^{(n)}$ is the predictable compensator of $\mathbf{N}^{(n)}$ and $N_i(t)$ are independent processes,

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp\big(\mathbf{Z}_i(s)'\boldsymbol{\beta}^o\big) d\Lambda_0(s), \quad 1 \le i \le n,\, t \ge 0, \tag{4.6}$$

are local martingales on $[0, \tau)$ with predictable variation/covariation processes

$$\langle M_i, M_i\rangle(t) = \int_0^t Y_i(s) \exp\big(\mathbf{Z}_i(s)'\boldsymbol{\beta}^o\big) d\Lambda_0(s), \quad \text{and} \quad \langle M_i, M_j\rangle(t) = 0,\ i \ne j.$$

The following lemma is provided by Huang et al. (2013), and another similar lemma can be found in (De. La. Pena, 1999).

**Lemma 3.** *(i) Let* $f_n(t) = \frac{1}{n}\int_0^t a_i(s) dM_i(s)$ *with* $[-1, 1]$*-valued predictable processes* $a_i(s)$*. Then, for all* $c > 0$*,*

$$P\Big\{ \max_{t\in[0,\tau]} |f_n(t)| > cx,\ \sum_{i=1}^n \int_0^\tau Y_i(t) dN_i(t) \le c^2 n \Big\} \le 2\exp\Big(-\frac{nx^2}{2}\Big).$$

*(ii) Suppose that* $\sup_{t\ge 0} \max_{i\le n, j\le p} |Z_{ij}(t) - \bar{Z}_{nj}(t, \boldsymbol{\beta}^o)| \le K$*, where* $\bar{Z}_{nj}(t, \boldsymbol{\beta}^o)$ *are the components of* $\bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o)$*. Then, for all* $c > 0$*,*

$$P\Big\{ \|\dot{\ell}(\boldsymbol{\beta}^o)\|_\infty > cKx,\ \sum_{i=1}^n \int_0^\tau Y_i(t) dN_i(t) \le c^2 n \Big\} \le 2p\exp\Big(-\frac{nx^2}{2}\Big).$$

*If additionally* $P\{\max_{i\le n} N_i(\tau) \le 1\} = 1$*, we can take* $c = 1$ *and yields that*

$$P\Big\{ \|\dot{\ell}(\boldsymbol{\beta}^o)\|_\infty > Kx \Big\} \le 2p\exp\Big(-\frac{nx^2}{2}\Big).$$

It is shown in Lemma 3 above, with high probability we can take $\lambda = c_2\big(\sqrt{\log(p)/n}\big)$ with a suitable constant $c_2$, so that the equation (4.4) is satisfied.

On the other hand, to verify (4.5) in Proposition 1 under Assumption 4, we need to measure how close between $\widehat{Q}$ and $\boldsymbol{\Sigma}$. To be precise, since $\widehat{Q}$ is an empirical counterpart of the matrix $\boldsymbol{\Sigma}$ (note also depends on $\check{\boldsymbol{\beta}}$), a key step to verify (4.5) is to show that $\big(\widehat{Q}_{SS}\big)^{-1}$ and $\widehat{Q}_{S^cS}$ are close to $\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}$ and $\boldsymbol{\Sigma}_{S^cS}$, respectively. These intermediate results are provided by the following proposition.

**Proposition 2.** *(Concentration of empirical matrices) Suppose that Assumptions 1-3 hold and*

$\mathbb{E}\big[\big\|\mathbf{Z}(t)_{SS}^{\otimes 2}\big\|_2\big]$ *is bounded uniformly. Then, if* $s_o^3 \ll n$*, with probability at least* $1-\delta$*, there holds*

$$\big\|\big(\widehat{Q}_{SS}\big)^{-1} - \big(\mathbf{\Sigma}_{SS}\big)^{-1}\big\|_2 = s_o O_p\Big(\sqrt{\frac{s_o}{n}} + \sqrt{\frac{\log(4s_o^2/\delta)}{n}}\Big). \tag{4.7}$$

*If additionally Assumption 4 also holds, and* $n \gg s_o^3 \log(p)$*. Then, with the same probability as above, there holds*

$$\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty \leq 1 - \gamma/2.$$

Thus, under Assumptions 1-4, the equation (4.5) is verified by Proposition 2. Until now, two conditions involved in Proposition 1 are both verified, so the strict dual feasibility of Step 3 in PDW procedure holds. Then by part(a) of Theorem 1, we conclude that $\check{\boldsymbol{\beta}}$ is the unique solution of the Lasso programme (2.3) with $S(\check{\boldsymbol{\beta}}) \subseteq S$. Next we shall obtain the inequality in Theorem 2 based on the formula in (4.2). In fact, by (4.7) we can treat $\big(\widehat{Q}_{SS}\big)^{-1}$ as some constant. Meanwhile, recall that $\lambda = c_2\big(\sqrt{\log(p/\delta)/n}\big)$ is taken as before, the desired inequality in Theorem 2 follows from (4.2) immediately.

In addition, if $\min_{j\in S}|\beta_j^o| \gg \sqrt{s_o \log(p/\delta)/n}$, without loss of generality, we assume that $\beta_j^o > 0$. In this case, by part(b) of Theorem 2, we have that $\check{\beta}_j \geq \beta_j^o - O_p\big(\sqrt{s_o\log(p/\delta)/n}\big) > 0$ for any $j \in S$. The similar argument also holds for $\beta_j^o < 0$. As a result, we conclude that $\mathrm{sign}(\check{\boldsymbol{\beta}}_S) = \mathrm{sign}(\boldsymbol{\beta}_1^o)$. Since we have proved that $|\check{\beta}_j| > 0$ for all $j \in S$, this follows that $\check{z}_S = \mathrm{sign}(\check{\boldsymbol{\beta}}_S) = \mathrm{sign}(\boldsymbol{\beta}_1^o)$ by the definition of subgradient. So Step 4 in PDW is verified. Totally, all the conditions of part(b) in Theorem 1 are satisfied, and based on this we obtain the second part of Theorem 2. □

## 4.2 Proof of Theorem 3

By part (c) of Theorem 1, it suffices to show that the dual feasibility check in Step 3, or the sign consistency check in Step 4 of the PDW must fail with probability at least 1/2. It may be assumed that $\check{z}_S = \mathrm{sign}(\boldsymbol{\beta}_1^o)$, otherwise, the sign consistency condition fails. Then, it remains to show that under this condition, the dual feasibility condition in Step 3 fails with probability at least 1/2. From (4.3), we recall that

$$\check{z}_{S^c} = \frac{1}{\lambda}\Big\{\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big[\dot{\ell}(\boldsymbol{\beta}^o)_S + \lambda\mathrm{sign}(\boldsymbol{\beta}_1^o)\big] - \dot{\ell}(\boldsymbol{\beta}^o)_{S^c}\Big\}. \tag{4.8}$$

Let $j \in S^c$ be the index corresponding to the maximum which is achieved in the violating condition (3.4). On one hand, among the proof of Proposition 2, we have show that $\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}$ converges to $\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}$ in $\|\cdot\|_\infty$-norm. Then, the violation condition (3.4) implies that

$$\big|e_j'\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\mathrm{sign}(\boldsymbol{\beta}_1^o)\big| \geq 1 + \nu/2$$

with high probability tending to one. Without loss of generality, we may assume that $e_j'\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\text{sign}(\boldsymbol{\beta}_1^o) \geq 1+\nu/2$. By (4.8), we have

$$\check{z}_j = e_j'\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\text{sign}(\boldsymbol{\beta}_1^o) - \frac{1}{\lambda}e_j'\big\{\dot{\ell}(\boldsymbol{\beta}^o)_{S^c} - \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^o)_S\big\}.$$

Hence, to prove $\check{z}_j > 1$, it suffices to show that $\mathbb{P}[W_j \leq \frac{\nu}{4}] \geq 1/2$, where

$$W_j := \frac{1}{\lambda}\big\{\dot{\ell}(\boldsymbol{\beta}^o)_j - e_j'\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^o)_S\big\}.$$

For this purpose, from the proof of Theorem 3.2 (Andersen and Gill, 1982), an application of the martingale central limit theorem yields that $\dot{\ell}(\boldsymbol{\beta}^o)_j$ and $\dot{\ell}(\boldsymbol{\beta}^o)_S$ are both asymptotically standard normal with zero mean. Note that

$$e_j'\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^o)_S = e_j'\big[\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1} - \boldsymbol{\Sigma}_{S^cS}(\boldsymbol{\Sigma}_{SS})^{-1}\big]\dot{\ell}(\boldsymbol{\beta}^o)_S + e_j'\big(\boldsymbol{\Sigma}_{S^cS}(\boldsymbol{\Sigma}_{SS})^{-1}\big)\dot{\ell}(\boldsymbol{\beta}^o)_S.$$

It follows from the proof of Proposition 2 that $\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1} - \boldsymbol{\Sigma}_{S^cS}(\boldsymbol{\Sigma}_{SS})^{-1}\|_\infty = o_p(1)$. Then the first term in the above equality converges to zero in probability, and this in turn means that $e_j'\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^o)_S$ is also an asymptotically standard normal with zero mean. Totally, $W_j$ is also an asymptotically standard normal with zero mean, so $\mathbb{P}\big[W_j \leq \frac{\nu}{4}\big] > \mathbb{P}\big[W_j \leq 0\big] = 1/2$ as n goes to infinity. Thus we complete the proof of part (a) in Theorem 3.

To prove part (b) of Theorem 3, we first recall from (4.2) that

$$\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^o = -(\widehat{Q}_{SS})^{-1}\big[\dot{\ell}(\boldsymbol{\beta}^o)_S + \lambda\check{z}_S\big].$$

To recover the correct signed support, we must have $\check{z}_S = \text{sign}(\boldsymbol{\beta}_1^o)$. Besides, by Proposition 2, we have shown that $\big\|\big(\widehat{Q}_{SS}\big)^{-1} - \big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\big\|_\infty = o_p(1)$. Thus, for $j \in S$ specified in part (b) of Theorem 3, $\check{\beta}_j$ can be expressed by

$$\begin{aligned}\check{\beta}_j &= e_j'\big[\big(\boldsymbol{\Sigma}_{SS}\big)^{-1} - \big(\widehat{Q}_{SS}\big)^{-1}\big]\big[\dot{\ell}(\boldsymbol{\beta}^o)_S + \lambda\text{sign}(\boldsymbol{\beta}_1^o)\big] + \beta_j^o - e_j'\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\big[\dot{\ell}(\boldsymbol{\beta}^o)_S + \lambda\text{sign}(\boldsymbol{\beta}_1^o)\big] \\ &= o_p(1) + \big\{\beta_j^o - h(\lambda)\big\} - e_j'\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\dot{\ell}(\boldsymbol{\beta}^o)_S.\end{aligned}$$

Without loss of generality, we only consider the situation $\beta_j^o \in \big(0, h(\lambda)\big)$. As mentioned above, $e_j'\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\dot{\ell}(\boldsymbol{\beta}^o)_S$ is an asymptotically standard normal with zero mean, which yields that

$$\mathbb{P}[\check{\beta}_j < 0] > \mathbb{P}[e_j'\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\dot{\ell}(\boldsymbol{\beta}^o)_S \geq 0] = 1/2, \quad n \to \infty,$$

where we used the fact that $\beta_j^o - h(\lambda) < 0$. As a result, Step 4 of the PDW fails, that is, $\check{z}_S \neq \text{sign}(\boldsymbol{\beta}_1^o)$. This together with part (c) of Theorem 1 concludes the proof. □

## Appendix A: Useful Lemmas in Empirical Processes

In this paper, a main ingredient from the theoretical point of view is that the randomness of the problem should be taken care of. For example, the covariate $\mathbf{Z}_i(t)$ is a random function of $t$, and classical random matrix theory are invalid to our time-dependent data. In this situation, we need to consider the behavior of the empirical process. This paper adopts the notation of Rademacher complexity to characterize the functional capacity. Let us recall Rademacher random variables, which are independent $\{-1, 1\}$-valued random variables with probability $1/2$ of taking either value. Let $X_1$, ..., $X_n$ be i.i.d. random sample from the distribution $\rho$ and $\sigma_1$, ..., $\sigma_n$ be i.i.d. Rademacher random variables, we define the empirical Rademacher average on the function space $\mathcal{G}$ as

$$\widehat{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(X_i), \quad g \in \mathcal{G},$$

and the population Rademacher average $R_n(\mathcal{G})$ is given by

$$R_n(\mathcal{G}) = \sup_{g\in\mathcal{G}} \mathbb{E}_{\sigma,\rho}[\widehat{R}_n(g)],$$

where $\mathbb{E}_{\sigma,\rho}$ means taking expectation with respect to all the random variables (i.e. the data and the Rademacher variables). Now we give some basic properties of Rademacher average.

**Property.** (1) Given any function class $\mathcal{G}$ and constants $a, b \in \mathbb{R}$, denote the function class $\{h|h(x) = ag(x) + b\}$ by $a\mathcal{G} + b$. Then

$$R_n(a\mathcal{G} + b) = |a|R_n(\mathcal{G}). \tag{4.9}$$

(2) Another useful property of Rademacher average is that it can be bounded by the so-called Dudley's entropy integral, namely, there exists some constant $c_0$ such that

$$R_n(\mathcal{G}) \leq \frac{c_0}{\sqrt{n}} \int_0^{\|\mathcal{G}\|_\infty} \sqrt{\log N(\mathcal{G}, \varepsilon, d_n)}d\varepsilon, \tag{4.10}$$

where $N(\mathcal{G}, \varepsilon, d_n)$ is the empirical covering number of $\mathcal{G}$ with the radius $\varepsilon$. Here the metric is defined as $(d_n(f, g))^2 = \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - g(x_i)|^2$ associated with available sample points $\{x_i\}_{i=1}^n$. In the literature of empirical process that, using this Rademacher average can remove the unnecessary $\log n$ factor of the VC bound, as well as refined constants.

Let $P_n(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$ and $P(g) = \mathbb{E}_\rho[g(X)]$. We now state the fundamental result involving Rademacher averages from Ledoux et al. (2011).

**Lemma 4.** *If $\mathcal{G} \subseteq \{g : X \to [c, c + 1]\}$ for any given constant $c$. For any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\sup_{g\in\mathcal{G}} |P(g) - P_n(g)| \leq 2R_n(\mathcal{G}) + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Note that by (4.9), we can extend the result of Lemma 4 to general case with any bounded

function space.

A major challenge in our proof is to characterize the concentration of the large matrix/ vector $S^{(\kappa)}(t,\boldsymbol{\beta}^o) = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i(t)^{\otimes\kappa} Y_i(t) e^{\mathbf{Z}_i(t)'\boldsymbol{\beta}^o}$, $\kappa = 0, 1, 2$. Note that each entry of the stochastic matrices is not a sum of independent terms, since it may vary with the time $t$. Hence the functional complexity in the law of large number has to be considered. To this end, we rely on Lemma 4 concerning the Rademacher complexity for empirical processes as our primary mathematical tools.

**Lemma 5.** *Under Assumption 2, for all $l, k = 1, ..., p$, there exists some universal constant $c_0 > 0$, such that for any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\sup_{t\in[0,\tau]} \left|S^{(0)}(t,\boldsymbol{\beta}^o) - s^{(0)}(t,\boldsymbol{\beta}^o)\right| \le c_0 s_o e^{K\|\boldsymbol{\beta}_1^o\|_1}\sqrt{\frac{\log(2/\delta)}{n}}, \tag{4.11}$$

$$\sup_{t\in[0,\tau]} \left|S_l^{(1)}(t,\boldsymbol{\beta}^o) - s_l^{(1)}(t,\boldsymbol{\beta}^o)\right| \le c_0 s_o e^{K\|\boldsymbol{\beta}_1^o\|_1}\sqrt{\frac{\log(2/\delta)}{n}}, \tag{4.12}$$

$$\sup_{t\in[0,\tau]} \left|S_{lk}^{(2)}(t,\boldsymbol{\beta}^o) - s_{lk}^{(2)}(t,\boldsymbol{\beta}^o)\right| \le c_0 s_o e^{K\|\boldsymbol{\beta}_1^o\|_1}\sqrt{\frac{\log(2/\delta)}{n}}, \tag{4.13}$$

*where $S_l^{(1)}(\cdot)$ is the $j$-th element of $S^{(1)}(\cdot)$ and $S_{lk}^{(2)}(\cdot)$ is the $(l,k)$-th entry of $S^{(2)}(\cdot)$, and the similar manner is valid to $s_l^{(2)}(\cdot)$ and $s_{lk}^{(2)}(\cdot)$.*

*Proof.* We only prove (4.13), and the other two inequalities follow similarly. For any given $l, k$, we write $g(X_i) = w_i(t) Z_{il}(t) Z_{ik}(t)$, where $w_i(t) = Y_i(t) e^{\mathbf{Z}_i(t)'\boldsymbol{\beta}^o}$, $t \in [0,\tau]$. That is, $S_{lk}^{(2)}(t,\boldsymbol{\beta}^o) = P_n(g)$ and $\mathbb{E}[S_{lk}^{(2)}(t,\boldsymbol{\beta}^o)] = s_{lk}^{(2)}(t,\boldsymbol{\beta}^o) = P(g)$. Since $\max_{i,l}|Z_{il}(t)| \le K$ for all $t \in [0,\tau]$, it follows that $w_i(t) \le e^{K\|\boldsymbol{\beta}_1^o\|_1}$. Let $\tilde{g}(X_i) = g(X_i)/(2K^2 e^{K\|\boldsymbol{\beta}_1^o\|_1})$, then $\tilde{g}(X_i) \in [-1/2, 1/2]$. Based on Lemma 4 with $c = -1/2$, the following inequality holds with probability at least $1 - \delta$

$$|P(\tilde{g}) - P_n(\tilde{g})| \le 2R_n(\mathcal{G}) + \sqrt{\frac{\log(2/\delta)}{2n}}, \quad \forall\, \tilde{g} \in \mathcal{G}, \tag{4.14}$$

where $\mathcal{G} = \{w(t) Z_l(t) Z_k(t)/(2K^2 e^{K\|\boldsymbol{\beta}_1^o\|_1}),\ t \in [0,\tau]\}$ as the hypotheses set. By (4.10), it suffices to show that the class of functions $\mathcal{G}$ has bounded uniform entropy integral. Since a function of bounded variation can be expressed as the difference of two increasing functions, it follows from Lemma 9.10 of Kosorok (2008) that $\mathcal{Z}_l = \{Z_l(t)/K : t \in [0,\tau]\}$ is a VC-hull class associated with a VC class of index 2. Then, by Corollary 2.6.12 of Van der Vaart et al. (1996), the entropy of $\mathcal{Z}_l$ satisfies $\log N(\mathcal{Z}_l, \varepsilon, d_n) = O_p(1/\varepsilon)$. Also, by Example 19.16 of Van der Vaart (1998), $\mathcal{Y} = Y(t) : t \in [0,\tau]$ is a VC class and hence has bounded uniform entropy integral. Thus, by Theorem 9.15 of Kosorok (2008), $\mathcal{Y}\mathcal{Z}_l\mathcal{Z}_k$ has bounded uniform entropy integral. It remains to consider the set $\mathcal{H} = \left\{\exp\left(\mathbf{Z}(t)'\boldsymbol{\beta}^o - K\|\boldsymbol{\beta}_1^o\|_1\right),\ t \in [0,\tau]\right\}$. For any two functions $f(t_m) = \exp\left(\mathbf{Z}(t_m)'\boldsymbol{\beta}^o - K\|\boldsymbol{\beta}_1^o\|_1\right)$, $t_m \in [0,\tau] (m = 1, 2)$. It is easy to check that $|f(t_1) - f(t_2)| \le \|\boldsymbol{\beta}_1^o\|_\infty \sum_{j\in S}\left|Z_j(t_1) - Z_j(t_2)\right|$, and it follows that $\log N(\mathcal{H}, \varepsilon, d_n) \le s_o \max_{l\in S}\left\{\log N(\mathcal{Z}_l, \varepsilon/(\|\boldsymbol{\beta}_1^o\|_\infty s_o), d_n)\right\} = s_o^2 O_p(1/\varepsilon)$. Then, applying Theorem 9.15 of Kosorok

(2008) again, we conclude that the uniform entropy integral of $\mathcal{Y}\mathcal{Z}_l\mathcal{Z}_k\mathcal{H}$ is bounded by the order of $s_o$. Consequently, our desired result follows immediately from (4.14) and (4.10). □

## Appendix B: Proof For Lemma 1 And Theorem 1

**Proof of Lemma 1.**

Equivalently, the convex program (2.3) can be reformulated as the $\ell_1$-constrained minimization, i.e. $\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \ell(\boldsymbol{\beta})$, subject to $\|\boldsymbol{\beta}\|_1 \leq C$, where the penalty parameter $\lambda$ and constraint level $C$ are in one-to-one correspondence via Lagrangian duality. So by Weierstrass's theorem, the minimum is always achieved. Furthermore, by a standard condition for optimality in a convex program on the open set $\mathbb{R}^p$, a point $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ minimizing (2.3) if and only if there exists a subgradient $\hat{z} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$ such that $\dot{\ell}(\hat{\boldsymbol{\beta}}) + \lambda\hat{z} = 0$. Thus part (a) is derived from (2.4).

To prove part (b), by standard duality theory (Bertsekas, 1995), given the subgradient $\hat{z} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$, any optimal solution $\tilde{\boldsymbol{\beta}}$ to the group Lasso must satisfy the complementary slackness condition $\hat{z}^\tau\tilde{\boldsymbol{\beta}} = \|\tilde{\boldsymbol{\beta}}\|_1$. Since $|\hat{z}_k| \leq 1$ for all $k$, $|\hat{z}_j| < 1$ (for some $j$) implies $\tilde{\beta}_j = 0$. This establishes Lemma 1 (b).

Finally, since the invertibility of $\ddot{\ell}_{S(\hat{\boldsymbol{\beta}})S(\hat{\boldsymbol{\beta}})}(\boldsymbol{\beta})$ implies that $\ddot{\ell}_{S(\hat{\boldsymbol{\beta}})S(\hat{\boldsymbol{\beta}})}(\boldsymbol{\beta})$ is strictly positive definite, then when restricted to vectors of the form $(\boldsymbol{\beta}_{S(\hat{\boldsymbol{\beta}})}, \mathbf{0})$, the group-Lasso program (2.3) is strictly convex, and so its optimum is uniquely determined, which ultimately yields part (c). □

**Lemma 6.** *(a) If Assumption 1(3) holds and* $\lambda \geq \frac{\log(n)}{n}$ *is satisfied, all the optimum of the restricted programme* (3.1) *stay within a bounded domain of* $\mathbb{R}^S$.

*(b) Under Assumptions* 1−3. *For any bounded vector* $\mathbf{b}_S \in \mathbb{R}^S$, $\ddot{\ell}_{SS}(\boldsymbol{\beta}^o + (\mathbf{b}_S, \mathbf{0}))$ *is invertible.*

The part(a) of Lemma 6 tells us that, it is enough to consider all the possible solutions of (3.1) within a bounded domain, not the whole space $\mathbb{R}^S$. Furthermore, under additional conditions, part (b) of Lemma 6 implies that the solution of (3.1) is unique. This serves the proof of part (a) of Theorem 1 below.

**Proof of Lemma 6.** To prove part (a) of Lemma 6, by definition of $\check{\boldsymbol{\beta}}_S$, we have that

$$\ell\big((\check{\boldsymbol{\beta}}_S, \mathbf{0})\big) + \lambda\|\check{\boldsymbol{\beta}}_S\|_1 \leq \ell\big(\mathbf{0}\big) = \frac{1}{n}\int_0^\tau \log\Big(\sum_{i=1}^n Y_i(s)\Big)d\bar{N}(s), \tag{4.15}$$

where $\bar{N}(s) = \sum_{i=1}^n N_i(s)$. Since no two counting processes $N_i$ jump at the same time, we have $|d\bar{N}(s)| = \big|\sum_{i=1}^n dN_i(s)\big| \leq 1$, where $dN_i(s) = N_i(t) - N_i(t^-)$ denotes the jump of $N_i(\cdot)$ at time $t$. By (4.15), it follows that $\|\check{\boldsymbol{\beta}}_S\|_1 \leq \frac{\tau\log(n)}{n\lambda} < \infty$, provided that $\lambda \geq \frac{\log(n)}{n}$. On the other hand, under Assumption 2 and $\mathbb{P}[\max_i\{N_i(\tau)\} \leq 1] = 1$, by Lemma 7 below, we have that

$$\big\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big\|_2 = o_p(1),$$

provided that $s_o^4 \ll n$. Then, repeating the process as that from (4.34) to (4.35), this together with Assumption 3 implies that $\ddot{\ell}(\boldsymbol{\beta}^o)_{SS}$ is invertible with $\big\|\big(\ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big)^{-1}\big\|_2 = O_p(1)$. Since $\ddot{\ell}(\boldsymbol{\beta}^o)_{SS}$ is symmetric, we see that $\ddot{\ell}(\boldsymbol{\beta}^o)_{SS}$ is strictly positive definite. Besides, over the restricted set $\mathbb{R}^S$, from Lemma 3.2 of Huang et al. (2013), we see that

$$\ddot{\ell}(\boldsymbol{\beta}^o + (\mathbf{b}_S, \mathbf{0}))_{SS} - e^{-2\eta_b}\ddot{\ell}(\boldsymbol{\beta}^o)_{SS} \text{ is nonnegative-definite,}$$

where $\eta_b = \max_t \max_{ij} |\mathbf{b}_S'\mathbf{Z}_{iS}(t) - \mathbf{b}_S'\mathbf{Z}_{jS}(t)|$. Since $\mathbf{b}_S$ and $\mathbf{Z}_S$ are both bounded by assumption, $e^{-2\eta_b} > 0$ holds. This further implies that $\ddot{\ell}(\boldsymbol{\beta}^o + (\mathbf{b}_S, \mathbf{0}))_{SS}$ is also strictly positive definite. Note that $\ddot{\ell}(\boldsymbol{\beta}^o + (\mathbf{b}_S, \mathbf{0}))_{SS} = \ddot{\ell}\big|_S(\boldsymbol{\beta}_1^o + \mathbf{b}_S)$, and then part (b) of Lemma 6 is proved. □

**Proof of Theorem 1**

*Proof.* Since $\check{\boldsymbol{\beta}}_S$ is an interior point over $\mathbb{R}^S$, it must be a zero-gradient point for the restricted program (3.1), hence $(\dot{\ell})\big|_S((\check{\boldsymbol{\beta}}_S, \mathbf{0})) + \lambda\check{z}_S = 0$. By the chain rule, this implies that $\big(\dot{\ell}(\check{\boldsymbol{\beta}})\big)_S + \lambda\check{z}_S = 0$, where $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_S, 0_{S^c})$. Besides, since $\max_{j\in S^c}|\check{z}_j| \le 1$ by assumption, we treat $\check{z}$ as one extended subgradient of $\check{\boldsymbol{\beta}}$. Then, $\check{\boldsymbol{\beta}}$ is an optimal point of the group-Lasso scheme (2.3) based on part (a) of Lemma 1. Furthermore, since $S(\check{\boldsymbol{\beta}}) \subseteq S$ be definition, the part(b) of Lemma 6 implies that $\ddot{\ell}_{S(\check{\boldsymbol{\beta}})S(\check{\boldsymbol{\beta}})}(\boldsymbol{\beta})$ is also invertible. Thus, by the parts (b) and (c) of Lemma 1, we conclude that $\check{\boldsymbol{\beta}}$ is the unique solution of (2.3) satisfying $S(\check{\boldsymbol{\beta}}) \subseteq S$.

If additionally, the sign consistency condition in Step 4 is satisfied. Then since $\check{z}_S$ was chosen as an element of the subdifferential $\partial\|\check{\boldsymbol{\beta}}_S\|_1$ in Step 2, we must have $\mathrm{sign}(\check{\boldsymbol{\beta}}_S) = \mathrm{sign}(\boldsymbol{\beta}_1^o)$, provided that $\check{\beta}_j \neq 0$ for all $j \in S$. Then this implies Lemma 1(b) by noting that $S(\check{\boldsymbol{\beta}}) = S(\boldsymbol{\beta}^o)$.

To prove part (c), it suffices to prove the following equivalent assertion: if there exists a group-Lasso solution $\hat{\boldsymbol{\beta}}$ with $S(\hat{\boldsymbol{\beta}}) = S$ and $\mathrm{sign}(\hat{\boldsymbol{\beta}}_S) = \mathrm{sign}(\boldsymbol{\beta}_1^o)$, then the PDW method succeeds in producing a dual feasible vector $\check{z}$ with $\check{z}_S = \mathrm{sign}(\boldsymbol{\beta}_1^o)$. First, by (2.4), it is easy to verify that $\dot{\ell}(\hat{\boldsymbol{\beta}})_S = (\dot{\ell})\big|_S(\hat{\boldsymbol{\beta}}_S)$. So $\hat{\boldsymbol{\beta}}_S$ is an optimal point to the restricted program (3.1). Also note that $\ddot{\ell}_{SS}((\boldsymbol{\beta}_S, \mathbf{0}))$ is invertible as proved in part (b) of Lemma 6, the vector $\hat{\boldsymbol{\beta}}_S$ must be the unique solution to (3.1). Note that $\mathrm{sign}(\hat{\boldsymbol{\beta}}_S) = \mathrm{sign}(\boldsymbol{\beta}_1^o)$ by condition, and we conclude that $\check{z}_S = \mathrm{sign}(\hat{\boldsymbol{\beta}}_S)$ is only subgradient that can be chosen in Step 2. Since $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S, \mathbf{0})$ is an optimal Lasso solution by assumption, then there must exist a dual feasible vector $\check{z}_{S^c}$ such that $(\mathrm{sign}(\hat{\boldsymbol{\beta}}_S), \check{z}_{S^c})$ satisfies the zero subgradient condition (3.2). □

## Appendix C: Proof for Concentration of Hessian Matrix

**Lemma 7.** *Under Assumption* 2 *and Assumption 1(2), the following inequality holds with probability at least* $1 - \delta$*,*

$$\big\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big\|_2 = O_p\Big(s_o^{3/2}\sqrt{\frac{\log(s_o^2/\delta)}{n}}\Big). \tag{4.16}$$

*Proof.* Let $L_{lk}$ and $\mathbf{\Sigma}_{lk}$ be the $(l,k)$-th entries of the matrices $\ddot{\ell}(\boldsymbol{\beta}^o)$ and $\mathbf{\Sigma}$, respectively. First of all, we write

$$\begin{aligned} L_{lk} - \mathbf{\Sigma}_{lk} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\frac{S_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)}\Big]dN_i(t) - \mathbb{E}\int_0^\tau \Big[\frac{s_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{s^{(0)}(t,\boldsymbol{\beta}^o)}\Big]dN(t) \\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\frac{S_{l}^{(1)}(t,\boldsymbol{\beta}^o)S_{k}^{(1)}(t,\boldsymbol{\beta}^o)}{(S^{(0)}(t,\boldsymbol{\beta}^o))^2}\Big]dN_i(t) + \mathbb{E}\int_0^\tau \Big[\frac{s_{l}^{(1)}(t,\boldsymbol{\beta}^o)s_{k}^{(1)}(t,\boldsymbol{\beta}^o)}{(s^{(0)}(t,\boldsymbol{\beta}^o))^2}\Big]dN(t) \\ &= T_1 - T_2. \end{aligned}$$

To bound term $T_1$, note that

$$\frac{S_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{s_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{s^{(0)}(t,\boldsymbol{\beta}^o)} = \frac{S_{lk}^{(2)}(t,\boldsymbol{\beta}^o) - s_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{s_{lk}^{(2)}(t,\boldsymbol{\beta}^o)\big[S^{(0)}(t,\boldsymbol{\beta}^o) - s^{(0)}(t,\boldsymbol{\beta}^o)\big]}{S^{(0)}(t,\boldsymbol{\beta}^o)s^{(0)}(t,\boldsymbol{\beta}^o)}. \quad (4.17)$$

Since $s^{(0)}(t,\boldsymbol{\beta}^o)/e^{K\|\boldsymbol{\beta}_1^o\|_1}$ is bounded away from zero by Assumption 2(3), $S^{(0)}(t,\boldsymbol{\beta}^o)/e^{K\|\boldsymbol{\beta}_1^o\|_1}$ is also bounded away from zero by (4.11). Then, from (4.11) and (4.13) in Lemma 5, we have

$$\sup_t \Big|\frac{S_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{s_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{s^{(0)}(t,\boldsymbol{\beta}^o)}\Big| \le c_0 s_o\sqrt{\frac{\log(4/\delta)}{n}} \quad (4.18)$$

with probability at least $1-\delta/2$. Write

$$\begin{aligned} T_1 &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\frac{S_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{s_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{s^{(0)}(t,\boldsymbol{\beta}^o)}\Big]dN_i(t) + (P_n - P)\int_0^\tau \Big[\frac{s_{lk}^{(2)}(t,\boldsymbol{\beta}^o)}{s^{(0)}(t,\boldsymbol{\beta}^o)}\Big]dN(t) \\ &= T_{11} + T_{12}, \end{aligned} \quad (4.19)$$

Since $P(\max_i\{N_i(\tau)\} \le 1) = 1$, it follows from (4.18) that $|T_{11}| = O_p\Big(s_o\sqrt{\frac{\log(2/\delta)}{n}}\Big)$. Besides, note that the term $T_{22}$ is an i.i.d. and bounded sum, an application of Hoeffding inequality yields that, with probability at least $1-\delta/2$

$$|T_{12}| = O_p\Big(\sqrt{\frac{\log(4/\delta)}{n}}\Big).$$

Putting the bounds for $T_{11}$ and $T_{12}$ together, with probability at least $1-\delta$, there holds

$$|T_1| = O_p\Big(s_o\sqrt{\frac{\log(2/\delta)}{n}}\Big).$$

Similarly, we can rewrite $T_2$ as

$$\begin{aligned} T_2 &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\frac{S_{l}^{(1)}(t,\boldsymbol{\beta}^o)S_{k}^{(1)}(t,\boldsymbol{\beta}^o)}{(S^{(0)}(t,\boldsymbol{\beta}^o))^2} - \frac{s_{i}^{(1)}(t,\boldsymbol{\beta}^o)s_{j}^{(1)}(t,\boldsymbol{\beta}^o)}{(s^{(0)}(t,\boldsymbol{\beta}^o))^2}\Big]dN_i(t) \\ &\quad +(P_n - P)\int_0^\tau \Big[\frac{s_{l}^{(1)}(t,\boldsymbol{\beta}^o)s_{k}^{(1)}(t,\boldsymbol{\beta}^o)}{(s^{(0)}(t,\boldsymbol{\beta}^o))^2}\Big]dN(t) \\ &= T_{21} + T_{22}. \end{aligned}$$

To bound term $T_{21}$, note that

$$\frac{S_l^{(1)}(t,\boldsymbol{\beta}^o)S_k^{(1)}(t,\boldsymbol{\beta}^o)}{(S^{(0)}(t,\boldsymbol{\beta}^o))^2} - \frac{s_l^{(1)}(t,\boldsymbol{\beta}^o)s_k^{(1)}(t,\boldsymbol{\beta}^o)}{(s^{(0)}(t,\boldsymbol{\beta}^o))^2} = \frac{S_k^{(1)}(t,\boldsymbol{\beta}^o)}{(S^{(0)}(t,\boldsymbol{\beta}^o))^2}\big\{S_l^{(1)}(t,\boldsymbol{\beta}^o) - s_l^{(1)}(t,\boldsymbol{\beta}^o)\big\}$$
$$+\frac{s_l^{(1)}(t,\boldsymbol{\beta}^o)}{(S^{(0)}(t,\boldsymbol{\beta}^o))^2}\big\{S_k^{(1)}(t,\boldsymbol{\beta}^o) - s_k^{(1)}(t,\boldsymbol{\beta}^o)\big\} - \frac{s_l^{(1)}(t,\boldsymbol{\beta}^o)s_k^{(1)}(t,\boldsymbol{\beta}^o)}{(S^{(0)}(t,\boldsymbol{\beta}^o)s^{(0)}(t,\boldsymbol{\beta}^o))^2}\big\{[S^{(0)}(t,\boldsymbol{\beta}^o)]^2 - [s^{(0)}(t,\boldsymbol{\beta}^o)]^2\big\}.$$

By the same arguments as in the proof of (4.17), it follows that

$$\Big|\frac{S_l^{(1)}(t,\boldsymbol{\beta}^o)S_k^{(1)}(t,\boldsymbol{\beta}^o)}{(S^{(0)}(t,\boldsymbol{\beta}^o))^2} - \frac{s_l^{(1)}(t,\boldsymbol{\beta}^o)s_k^{(1)}(t,\boldsymbol{\beta}^o)}{(s^{(0)}(t,\boldsymbol{\beta}^o))^2}\Big| \le c_0 s_o\sqrt{\frac{\log(4/\delta)}{n}}$$

with probability at least $1-\delta/2$. This further implies that $|T_{21}| = O_p\Big(s_o\sqrt{\frac{\log(4/\delta)}{n}}\Big)$. Also, note that the term $T_{22}$ is an i.i.d. and bounded sum, an application of Hoeffding inequality yields that, with probability at least $1-\delta/2$, $|T_{22}| = O_p\Big(\sqrt{\frac{\log(4/\delta)}{n}}\Big)$. Then, combining the bounds for $T_{21}$ and $T_{22}$ yields that

$$T_2 = O_p\Big(s_o\sqrt{\frac{\log(4/\delta)}{n}}\Big).$$

Finally, putting the bounds for $T_1$ and $T_2$ together tells us that, with probability at least $1-\delta$

$$\big\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big\|_{\max} = O_p\Big(s_o\sqrt{\frac{\log(s_o^2/\delta)}{n}}\Big), \tag{4.20}$$

where $\|\cdot\|_{\max}$ is denoted to be the elementwise norm for a matrix. Note that $\|M\|_2 \le \sqrt{s_o}\max_{i,j}|M_{ij}|$ for any $M \in \mathbb{R}^{s_o\times s_o}$, the desired inequality follows from (4.20) immediately.

□

We here introduce the following useful lemma on matrices (Loh and Wainwright, 2014).

**Lemma 8.** *Let $A$, $B \in \mathbb{R}^p$ be invertible. For any matrix norm $||\cdot||$, there holds*

$$||A^{-1} - B^{-1}|| \le \frac{||A^{-1}||^2\cdot||A-B||}{1-|||A^{-1}||\cdot|||A-B||}.$$

**Proof of Proposition 2.** First of all, by the triangle inequality, we have that

$$\big\|\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\big\|_2 \le \big\|\widehat{Q}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big\|_2 + \big\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big\|_2. \tag{4.21}$$

Since the second term of (4.21) has been shown in Lemma 7, it suffices to bound the first one of (4.21). Recalling $\ddot{\ell}(\boldsymbol{\beta}) = \frac{1}{n}\int_0^\tau V_n(s,\boldsymbol{\beta})d\bar{N}(s)$, we have that

$$\begin{aligned}\widehat{Q} - \ddot{\ell}(\boldsymbol{\beta}^o) &= \int_0^1 \Big\{\ddot{\ell}\big(\boldsymbol{\beta}^o + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\big) - \ddot{\ell}(\boldsymbol{\beta}^o)\Big\}d\theta \\ &= \frac{1}{n}\int_0^1\int_0^\tau \Big\{V_n(s,\boldsymbol{\beta}^o + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)) - V_n(s,\boldsymbol{\beta}^o)\Big\}d\bar{N}(s)d\theta. \end{aligned}\tag{4.22}$$

Then, for any unit vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$,

$$\mathbf{a}' V_n(t, \boldsymbol{\beta}) \mathbf{b} = \sum_{i=1}^{n} w_{in}(t, \boldsymbol{\beta}) \mathbf{a}' \big( \mathbf{Z}_i(t) - \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}) \big) \cdot \mathbf{b}' \big( \mathbf{Z}_i(t) - \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}) \big), \quad \boldsymbol{\beta} \in \mathbb{R}^p.$$

Following the above formulation, for any $\boldsymbol{\delta} \in \mathbb{R}^p$, we have that

$$\mathbf{a}' \big( V_n(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) - V_n(t, \boldsymbol{\beta}^o) \big) \mathbf{b} = I_1 - I_2 - I_3 + I_4, \tag{4.23}$$

where

$$\begin{aligned}
I_1 &:= \sum_{i=1}^{n} \big[ w_{in}(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) - w_{in}(t, \boldsymbol{\beta}^o) \big] \mathbf{a}' \mathbf{Z}_i(t) \mathbf{b}' \mathbf{Z}_i(t), \\
I_2 &:= \sum_{i=1}^{n} w_{in}(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) \mathbf{a}' \mathbf{Z}_i(t) \mathbf{b}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) - \sum_{i=1}^{n} w_{in}(t, \boldsymbol{\beta}^o) \mathbf{a}' \mathbf{Z}_i(t) \mathbf{b}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o), \\
I_3 &:= \sum_{i=1}^{n} w_{in}(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) \mathbf{b}' \mathbf{Z}_i(t) \mathbf{a}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) - \sum_{i=1}^{n} w_{in}(t, \boldsymbol{\beta}^o) \mathbf{b}' \mathbf{Z}_i(t) \mathbf{a}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o), \\
I_4 &:= \sum_{i=1}^{n} w_{in}(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) \mathbf{a}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) \mathbf{b}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o + \boldsymbol{\delta}) - \sum_{i=1}^{n} w_{in}(t, \boldsymbol{\beta}^o) \mathbf{a}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o) \mathbf{b}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o).
\end{aligned}$$

Now we consider $I_1$. To simplify expression, let $\delta_i = \delta_i(t) = \exp\{\mathbf{Z}_i(t)' \boldsymbol{\delta}\}$ and $w_i = w_i(t) = Y_i(t) \exp\{\mathbf{Z}_i(t)' \boldsymbol{\beta}^o - K \|\boldsymbol{\beta}_1^o\|_1\}$. Since $\mathbf{Z}_i(t)$ is uniformly bounded by $K$, $\max_{i,t}\{(\mathbf{Z}_i(t))' \boldsymbol{\beta}^o\} \leq K \|\boldsymbol{\beta}_1^o\|_1$, which guarantees that $0 < w_i \leq 1$ for all $i, t$. In this case, $I_1$ can be rewritten as

$$I_1 = \sum_{i=1}^{n} \Big[ \frac{w_i \sum_{k \neq i} w_k (\delta_i - \delta_k)}{\sum_{k,l=1} w_k \delta_k w_l} \Big] \mathbf{a}' \mathbf{Z}_i(t) \mathbf{b}' \mathbf{Z}_i(t).$$

Similarly, $I_2 - I_4$ can be rewritten as the following formulas,

$$\begin{aligned}
I_2 &= \sum_{i=1}^{n} \Big[ \frac{w_i \sum_{k \neq i} w_k (\delta_i - \delta_k)}{\sum_{k,l=1} w_k \delta_k w_l} \Big] \mathbf{a}' \mathbf{Z}_i(t) \mathbf{b}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o) + \\
&\quad \Big( \frac{\sum_{i=1}^n w_i \mathbf{a}' \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k} \Big) \cdot \Big( \frac{\sum_{i=1}^n w_i \delta_i \mathbf{b}' \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k \delta_k} - \frac{\sum_{i=1}^n w_i \mathbf{b}' \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k} \Big), \\
I_3 &= \sum_{i=1}^{n} \Big[ \frac{w_i \sum_{k \neq i} w_k (\delta_i - \delta_k)}{\sum_{k,l=1} w_k \delta_k w_l} \Big] \mathbf{b}' \mathbf{Z}_i(t) \mathbf{a}' \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o) + \\
&\quad \Big( \frac{\sum_{i=1}^n w_i \mathbf{b}' \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k} \Big) \cdot \Big( \frac{\sum_{i=1}^n w_i \delta_i \mathbf{a}' \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k \delta_k} - \frac{\sum_{i=1}^n w_i \mathbf{a}' \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k} \Big), \\
I_4 &= \frac{\Big( \sum_{i=1}^n w_i \delta_i \mathbf{a}' \mathbf{Z}_i(t) \Big) \Big( \sum_{i=1}^n w_i \delta_i \mathbf{b}' \mathbf{Z}_i(t) \Big)}{\big( \sum_{k=1}^n w_k \delta_k \big)^2} - \frac{\Big( \sum_{i=1}^n w_i \mathbf{a}' \mathbf{Z}_i(t) \Big) \Big( \sum_{i=1}^n w_i \mathbf{b}' \mathbf{Z}_i(t) \Big)}{\big( \sum_{k=1}^n w_k \big)^2}.
\end{aligned}$$

Denote $\gamma_\delta = \gamma_\delta(t) = \max_{i,j}[\mathbf{Z}_i(t) - \mathbf{Z}_j(t)]'\boldsymbol{\delta} > 0$, and a direct computation yields that

$$|I_1| \leq \exp(\gamma_\delta)\gamma_\delta \sum_{i=1}^n \big(w_i \big|\mathbf{a}'\mathbf{Z}_i(t)^{\otimes 2}\mathbf{b}\big|\big) \Big/ \sum_{i=1}^n w_i. \tag{4.24}$$

Following Assumption 2(1), we have that $\gamma_\delta \leq K\|\boldsymbol{\delta}\|_1$. Thus, it remains for bounding $I_1$ to estimate $\sum_{i=1}^n \big(w_i\big|\mathbf{a}'\mathbf{Z}_i(t)^{\otimes 2}\mathbf{b}\big|\big) \Big/ \sum_{i=1} w_i$, which can be solved by estimating its numerator and denominator respectively.

Note that, taking a supermum over unit vectors $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^S$, by definition we have

$$\sup_{\|\mathbf{a}\|_2=1, \|\mathbf{b}\|_2=1} \Big\{\frac{1}{n}\sum_{i=1}^n w_i\big|\mathbf{a}'\mathbf{Z}_i(t)^{\otimes 2}\mathbf{b}\big|\Big\} = \Big\|\frac{1}{n}\sum_{i=1}^n w_i\big(\mathbf{Z}_i(t)^{\otimes 2}\big)_{SS}\Big\|_2,$$

which (4.24) further implies that

$$\tilde{I}_1(t) := \sup_{\|\mathbf{a}\|_2=1, \|\mathbf{b}\|_2=1} \{I_1\} \leq K\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1 \cdot \Big\|\frac{1}{n}\sum_{i=1}^n w_i\big(\mathbf{Z}_i(t)^{\otimes 2}\big)_{SS}\Big\|_2 \Big/ \frac{1}{n}\sum_{i=1}^n w_i. \tag{4.25}$$

Note that by the triangle inequality, one gets

$$\Big\|\frac{1}{n}\sum_{i=1}^n w_i\big(\mathbf{Z}_i(t)^{\otimes 2}\big)_{SS}\Big\|_2 \leq \Big\|\frac{1}{n}\sum_{i=1}^n w_i\big(\mathbf{Z}_i(t)^{\otimes 2}\big)_{SS} - \mathbb{E}[w\big(\mathbf{Z}(t)^{\otimes 2}\big)_{SS}]\Big\|_2 + \Big\|\mathbb{E}[w\big(\mathbf{Z}(t)^{\otimes 2}\big)_{SS}]\Big\|_2.$$

and by definition we have

$$\Big\|\frac{1}{n}\sum_{i=1}^n w_i\big(\mathbf{Z}_i(t)^{\otimes 2}\big)_{SS} - \mathbb{E}[w\big(\mathbf{Z}(t)^{\otimes 2}\big)_{SS}]\Big\|_2 \leq \sqrt{s_o}\Big\|\frac{1}{n}\sum_{i=1}^n w_i\big(\mathbf{Z}_i(t)^{\otimes 2}\big)_{SS} - \mathbb{E}[w\big(\mathbf{Z}(t)^{\otimes 2}\big)_{SS}]\Big\|_{\max}, \tag{4.26}$$

where $\|\cdot\|_{\max}$ is defined as that in Lemma 7. Moreover, by the same arguments as that for the proof of Lemma 5, with probability at least $1-\delta$, the following two inequities both hold

$$\Big|\frac{1}{n}\sum_{i=1}^n w_i - \mathbb{E}(w)\Big| = O_p\Big(\sqrt{\frac{\log(4/\delta)}{n}}\Big), \tag{4.27}$$

$$\Big\|\frac{1}{n}\sum_{i=1}^n w_i\big(\mathbf{Z}_i(t)^{\otimes 2}\big)_{SS} - \mathbb{E}[w\big(\mathbf{Z}(t)^{\otimes 2}\big)_{SS}]\Big\|_2 = O_p\Big(s_o^{1/2}\sqrt{\frac{\log(s_o^2/\delta)}{n}}\Big) \tag{4.28}$$

for all $t \in [0, \tau]$. Thus, since $\mathbb{E}(w) > 0$ by Assumption 2(3), from (4.27) we see that $\frac{1}{n}\big|\sum_{i=1}^n w_i\big|$ is bounded away from zero with high probability. Thus, plugging (4.27) and (4.28) into (4.25), we obtain that

$$\sup_t\{\tilde{I}_1(t)\} = O_p\Big(\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1\Big), \tag{4.29}$$

since $\mathbb{E}\big[\big\|[\mathbf{Z}(t)^{\otimes 2}_{SS}\big\|_2\big]$ is bounded uniformly by assumption. Besides, a similar argument shows

that

$$\tilde{I}_2(t) := \sup_{\|\mathbf{a}\|_2=1,\|\mathbf{b}\|_2=1} I_2(t) = O_p\Big(\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1\Big), \quad \text{for all}\, t \in [0,\tau],$$
$$\tilde{I}_3(t) := \sup_{\|\mathbf{a}\|_2=1,\|\mathbf{b}\|_2=1} I_3(t) = O_p\Big(\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1\Big), \quad \text{for all}\, t \in [0,\tau],$$
$$\tilde{I}_4(t) := \sup_{\|\mathbf{a}\|_2=1,\|\mathbf{b}\|_2=1} I_4(t) = O_p\Big(\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1\Big), \quad \text{for all}\, t \in [0,\tau].$$

Then, setting $\boldsymbol{\delta} = \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)$, and combining with (4.22), (4.37) and all the derived bounds of $\tilde{I}_1 - \tilde{I}_4$, we have that

$$\Big\|\big[V_n(s, \boldsymbol{\beta}^o + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)) - V_n(s, \boldsymbol{\beta}^o)\big]_{SS}\Big\|_2 = O_p\Big(\exp(K\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1)\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1\Big), \tag{4.30}$$

for all $s \in [0,\tau]$ and $\theta \in [0,1]$. Also, taking a supermum over unit vectors $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^S$, we see from (4.22) that

$$\big\|\widehat{Q}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big\|_2 \leq \frac{1}{n}\int_0^1 \int_0^\tau \Big\|\big[V_n(s, \boldsymbol{\beta}^o + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)) - V_n(s, \boldsymbol{\beta}^o)\big]_{SS}\Big\|_2 d\bar{N}(s)d\theta. \tag{4.31}$$

Since $N_i(\tau) \leq 1$ for all $i \leq n$, it follows from (4.31), (4.30) and the result of Lemma 2 that

$$\big\|\widehat{Q}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big\|_2 = s_o O_p\Big(\sqrt{\frac{\log(p/\delta)}{n}}\Big), \tag{4.32}$$

which converges to zero as $n$ goes to infinity and $s_o^2 \ll n$.

On the other hand, by Lemma 7 we recall that

$$\big\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^o)_{SS}\big\|_2 = O_p\Big(s_o^{3/2}\sqrt{\frac{\log(s_o^2/\delta)}{n}}\Big), \tag{4.33}$$

with probability at least $1-\delta/2$. Therefore, combining (4.32), (4.33) with (4.21), we obtain that

$$\big\|\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\big\|_2 = O_p\Big(s_o^{3/2}\sqrt{\frac{\log(s_o^2/\delta)}{n}}\Big). \tag{4.34}$$

Since $\big\|(\boldsymbol{\Sigma}_{SS})^{-1}\big\|_2 = O_p(1)$ by Assumption 3, we still need to show that $\big\|(\widehat{Q}_{SS})^{-1}\big\|_2 = O_p(1)$ with high probability, so that applying Lemma 8 with the $\|\cdot\|_2$-norm yields our desired result. To this end, we decompose $(\widehat{Q}_{SS})^{-1}$ as

$$(\widehat{Q}_{SS})^{-1} = (\boldsymbol{\Sigma}_{SS})^{-1/2}\big\{I + (\boldsymbol{\Sigma}_{SS})^{-1/2}\big(\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\big)(\boldsymbol{\Sigma}_{SS})^{-1/2}\big\}^{-1}(\boldsymbol{\Sigma}_{SS})^{-1/2},$$

and let $\mathcal{A} = I + (\boldsymbol{\Sigma}_{SS})^{-1/2}\big(\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\big)(\boldsymbol{\Sigma}_{SS})^{-1/2}$. Then $(\widehat{Q}_{SS})^{-1} = (\boldsymbol{\Sigma}_{SS})^{-1/2}\mathcal{A}^{-1}(\boldsymbol{\Sigma}_{SS})^{-1/2}$.

By the Bauer-Fike inequality, we have

$$|\lambda(\mathcal{A})-1| \le \|(\mathbf{\Sigma}_{SS})^{-1/2}\big(\widehat{Q}_{SS}-\mathbf{\Sigma}_{SS}\big)(\mathbf{\Sigma}_{SS})^{-1/2}\|_2 \le \|\mathbf{\Sigma}_{SS})^{-1/2}\|_2^2\|\widehat{Q}_{SS}-\mathbf{\Sigma}_{SS}\|_2.$$

Then by (4.34) and Assumption 3, $|\lambda(\mathcal{A})-1| = o_p(1)$. Hence $\lambda(\mathcal{A}^{-1}) = 1+o_p(1)$. Since $\mathcal{A}$ is symmetrical, $\|\mathcal{A}^{-1}\|_2 = O_p(1)$. This together with Assumption 2 yields that

$$\|(\widehat{Q}_{SS})^{-1}\|_2 \le \|(\mathbf{\Sigma}_{SS})^{-1/2}\|_2\|\mathcal{A}^{-1}\|_2\|(\mathbf{\Sigma}_{SS})^{-1/2}\|_2 = O_p(1). \tag{4.35}$$

As a result, by (4.34) and (4.35), the first part of Proposition 2 follows easily from Lemma 8.

Next, we turn to the second term of Proposition 2, that is, verify the expression (4.5). Under Assumption 4, we first notice that

$$\begin{aligned}\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty &\le \big\|\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big\|_\infty + \big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}-\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big\|_\infty \\ &\le 1-\gamma+\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}-\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big\|_\infty. \end{aligned}\tag{4.36}$$

Then it suffices to show that $\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}-\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big\|_\infty \le \frac{\gamma}{2}$. For this purpose, let $T := \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}-\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}$, and we split $T$ into two parts:

$$T := T_1+T_2,$$

where $T_1 = \big(\widehat{Q}_{S^cS}-\mathbf{\Sigma}_{S^cS}\big)(\widehat{Q}_{SS})^{-1}$, and $T_2 = \mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big(\widehat{Q}_{SS}-\mathbf{\Sigma}_{SS}\big)(\widehat{Q}_{SS})^{-1}$.

Similarly as before, by the union bound, Lemma 5 and the result of Theorem 2, we have

$$\max_{j\in S^c}\big\|e_j'\big(\widehat{Q}_{S^cS}-\mathbf{\Sigma}_{S^cS}\big)\big\|_2 = O_p\Big(s_o^{3/2}\sqrt{\frac{\log(p/\delta)}{n}}\Big).$$

Since by (4.35), $\|(\widehat{Q}_{SS})^{-1}\|_2$ can be treated as positive constant. Then if $s_o^3 \ll n$, we have

$$\|T_1\|_\infty \le \max_{j\in S^c}\big\|e_j'\big(\widehat{Q}_{S^cS}-\mathbf{\Sigma}_{S^cS}\big)\big\|_2\cdot\|(\widehat{Q}_{SS})^{-1}\|_2 = O_p\Big(s_o^{3/2}\sqrt{\frac{\log(p/\delta)}{n}}\Big). \tag{4.37}$$

To bound $T_2$, by Assumption 4 and (4.35), we have

$$\|T_2\|_\infty \le \|\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\|_\infty\|(\widehat{Q}_{SS})^{-1}\|_\infty\|\widehat{Q}_{SS}-\mathbf{\Sigma}_{SS}\|_\infty = \sqrt{s_o}O_p\big(\|\widehat{Q}_{SS}-\mathbf{\Sigma}_{SS}\|_2\big), \tag{4.38}$$

since $\|A\|_\infty \le \sqrt{s_o}\|A\|_2$ for any matrix $A\in\mathbb{R}^{S\times S}$. Then by bounds (4.34) and (4.38), we have

$$\|T_2\|_\infty = O_p\Big(s_o^2\sqrt{\frac{\log(s_o^2/\delta)}{n}}\Big). \tag{4.39}$$

So combined with inequalities (4.36), (4.37) and (4.39), we have

$$\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty \le 1-\gamma+\|T_1\|_\infty+\|T_2\|_\infty \le 1-\frac{\gamma}{2}, \tag{4.40}$$

provided that $\left(s_o^2\sqrt{\frac{\log(p/\delta)}{n}}\right) = o_p(1)$. □

### Appendix D: Proof for Lemma 2

**Proof of Lemma 2.** Since estimation error established in Theorem 3.1 of Huang et al. (2013) holds under the particular setting $p = s_o$, We can apply this result over $\mathbb{R}^S$ directly. To be precise, by Theorem 3.1 of Huang et al. (2013), in the event $\left\|\dot{\ell}(\boldsymbol{\beta}_1^o)\right\|_\infty \leq \lambda$, we have

$$\|\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^o\|_2 = O_p\Big(\sqrt{s_o}\lambda/F_2(\xi, S)\Big),$$

where $F_2(\xi, S)$ has been defined in equation (3.4) over there. Moreover, with high probability we have

$$F_2(\xi, S) \geq \lambda_{\min}\big(\ddot{\ell}(\boldsymbol{\beta}_1^o)\big) = \big\|(\ddot{\ell}(\boldsymbol{\beta}_1^o))^{-1}\big\|_2 \geq \frac{1}{2}\big\|(\boldsymbol{\Sigma}_{SS})^{-1}\big\|_2 = O_p(1),$$

where the first inequality follows from the definition of $F_2(\xi, S)$, and the second inequality follows from the results of Lemma 7 and Lemma 8, and the last one holds from Assumption 3. In addition, if $P\{\max_{i\leq n} N_i(\tau) \leq 1\} = 1$, as mentioned earlier, we have taken $\lambda = \sqrt{\log(p/\delta)/n}$, which completes the proof of Lemma 2. □

## References


A., Antoniadis, P., Fryzlewicz and F., Letue. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.*, *37*, 531–552.

K., Azuma. (1967). Weighted sums of certain dependent random variables. *Tôhoku Math. J.*, *19*, 357–367.

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

D. P., Bertsekas. (1995). Nonlinear Programming. *Athena Scientific*, *Belmont, MA.*

J., Bradic, J., Fan, and J., Jiang. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.*, *39*, 3092–3120.

D., Donoho. (2006). Compressing sensing. *IEEE. Trans. Info. Theory*, *52*, 1289–1306.

D. L., Donoho and J. M., Tanner. (2008). Counting faces of randomly-projected polytopes when the projection radically lower dimension. *J. Amer. Math. Soc.*, July.

R. A., DeVore and G. G., Lorentz. (1993). Constructive Approximation. *Springer-Verlag*, *New York.*

S. V., De Geer. (2002). Empirical Processes in M-Estimation. *Cambridge Unversity Press.*

S. A., Van De Geer. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, *36*, 614–645.

V. H., De. La. Pena. (1999). A general class of exponential inequlities for martingales and ratios. *Ann. probab.*, *27*, 537–564.

E. X., Fang, Y., Ning and H, Liu. (2015) Testing and confidence intervals for high dimensional proportional hazards model. *arXiv:1412.5158v1.*

J. Fan, and R. Li. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Stat. Assoc. 96*, 1348–1360.

J., Fan, and R., Li. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, *30*, 74–99.

S., Gaiffas, and A., Guilloux. (2012). High-dimensional additive hazards models and the Lasso. *Electron. J. Stat.*, *6*, 522–546.

J., Huang, T. N., Sun, Z. L., Ying, Y., Yu and C. H., Zhang. (2013). Oracle inequqlities for the Lasso in the Cox model. *Ann. Statist.*, *41*, 1142–1165.

W., Hoeffding. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* , *58*, 13–30.

J., Lv and Y., Fan. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, *37*, 3498–3528.

S., Lemler. (2012). Oracle inequalities for the Lasso for the conditional hazard rate in a high dimensional setting. Available at arXiv:1206.5628.

P. L., Loh and M. J., Wainwright. (2014). Support recovery without incoherence: A case for nonconvex regularization. *arXiv: 1412.5632v1* .

W., Lin and J. Lv. (2013). High dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.*, *108*, 247–264.

N., Merinshausen and P., Buhlmann. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, *34*, 1436–1462.

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

S., Kong and B., Nan. (2014). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Stat. Sin.*, *24*, 25–42.

R. Tibshirani. (1996) Regression selection and shrinkage via the Lasso, *J. R. Stat. Soc. Ser. B*, *58*, 267–288.

R., Tibshirani. (1997). The Lasso method for variable selection in the Cox model. *Stat. Med.*, *16*, 385–396.

A. W., Van der Vaart. (1998). Asymptotic Statistics. *New York: Cambridge University Press.*

A. W., Van der Vaart and J. A., Wellner. (1996). Weak Convergence and Empirical Processes: With Applications to Statistics. *New York: Springer.*

M. J., Wainwright. (2009). Sharp thresholds for high-dimensional and noisy sparity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory.* , *55*, 2183–2202.

P., Zhao and B., Yu. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, *7*, 2541–2563.

T., Zhang. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, *11*, 1081–1107.

C. H., Zhang and J., Huang. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.*, *36*, 1567–1494.

M., Ledoux and M., Talagrand. (2011). Probability in Banach Spaces: Isoperimetry and Processes. *Springer-Verlag Berlin Heidelberg.*