Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

Oracle inequalities for weighted group lasso in high-dimensional misspecified Cox models

Yijun Xiao[1], Ting Yan[2*], Huiming Zhang[3*] and Yuanyuan Zhang[4]

*Correspondence:
tingyanty@mail.ccnu.edu.cn;
huimingzhang@um.edu.mo
[2]Department of Statistics, Central China Normal University, Wuhan, China
[3]Department of Mathematics, Faculty of Science and Technology, University of Macau, Macau, China
Full list of author information is available at the end of the article

Abstract

We study the nonasymptotic properties of a general norm penalized estimator, which include Lasso, weighted Lasso, and group Lasso as special cases, for sparse high-dimensional misspecified Cox models with time-dependent covariates. Under suitable conditions on the true regression coefficients and random covariates, we provide oracle inequalities for prediction and estimation error based on the group sparsity of the true coefficient vector. The nonasymptotic oracle inequalities show that the penalized estimator has good sparse approximation of the true model and enables to select a few meaningful structure variables among the set of features.

Keywords: Proportional hazard model; Partial likelihood; Time-dependent data; Weighted group Lasso; Oracle inequalities; Suprema of empirical processes

1 Introduction

In recent years, high-throughput and nonparametric complex data have been frequently collected in gene-biology, signal processing, neuroscience, and other scientific fields. With massive data in regression problem, we encounter the situation that both the number of covariates p and the sample size n are increasing, and p is a function of n, i.e., $p =: p(n)$. The curse of dimensionality with computational complexity forces us to make the variable selection since the true regression coefficients β^* often are sparse with few nonzero components. Thus only a subset of the variable is preferable as important feature. The sparse set of nonzero coordinates in β^* also aims to choose the best model. A popular approach is to penalize the log-likelihood by adding a penalty function, which will intuitively lead to choosing a sparse model. One popular proposed method is Lasso (least absolute shrinkage and selection operator), which was introduced in Tibshirani [23] as a modification of the least square method in linear models. With the development of data science, high-dimensional statistics, including various regularization methods (such as group Lasso, weighted Lasso) have been sprung up by statisticians' efforts for over two decades.

Ever since the methodology of Lasso linear models, studying various penalty functions (from data independent to data-driven penalty) and loss functions (from smooth to nonsmooth, from Lipschitz to non-Lipschitz) remains hot in high-dimensional statistics, even though Lasso regularization has been thoroughly analyzed. However, arising in much

Springer

practical application, predictors may have group structures. Yuan and Lin [29] study the problem of selecting grouped variables for accurate prediction in linear regressions, and their proposed group Lasso is an extension of Lasso for the purpose of the accuracy of estimation. When considering the variable selection in Cox models, massive data sets bring researchers unprecedented computational challenges, see Tibshirani [24]. Fan and Li [9] study the SCAD penalized partial likelihood approach for the Cox models, and the proposed estimator enjoys the oracle property if a proper regularization parameter is chosen. Zhang and Lu [34] consider different penalties for different coefficients (the adaptive Lasso), and their idea is that "unimportant variables receive larger penalties than important ones so that important variables tend to be retained in the selection process, whereas unimportant variables are more likely to be dropped". Theoretical properties, including consistency and rate of convergence of this estimator called adaptive Lasso, are also shown by Zhang and Lu [34] when the number of covariates is fixed.

A potential characterization, which appeared in large-scale gene data associated with survival time, is that we only have a few (maybe several) significant predictors among p (maybe thousands) covariates and $p \gg n$ apparently. For example, the survival of patients with diffuse large-B-cell lymphoma(DLBCL) after chemotherapy is affected by molecular features of the tumors, which is measured by high-dimensional microarray gene expression. Rosenwald et al. [20] adopt Cox models to identify individual genes whose expression correlated with the outcome, and the data contain $n = 240$ patients and $p = 7399$ gene expression levels associated with a good or an adverse outcome. The main challenge is that directly utilizing low-dimensional (classical and traditional) statistical inference and computing methods for these data is prohibitive. Fortunately, the regularized partial likelihood method can perform parameter estimation and variable selection to enhance the prediction accuracy and interpretability of the Cox models.

There is the fact that the Lasso estimator is not asymptotically normal, and accurate and limit distribution of Lasso estimate is hard to derive and does not have explicit form, see Knight and Fu [15]. To avoid this trouble, a popular method is to derive the nonasymptotic *oracle inequality* based on some regularity conditions. Early in 2004, oracle inequalities for prediction error were derived without sparsity or restricted eigenvalue conditions for Lasso-type estimators [see Greenshtein and Ritov [10], Bartlett et al. [3]].

In the classical consistency analysis, the model size p is fixed, and the sample size n goes to infinity. While we need nonasymptotic error bounds in high-dimensional statistical consistency analysis when both model size p and sample size n go to infinity.

Let β^* be the true regression coefficient obtained from regression data $\{X_i, Y_i\}_{i=1}^n$, where X_i is p-dimensional covariates and $Y_i \in \mathbb{R}$ is the response. A modern problem, which will be the focus of this paper, is the behavior of $\hat{\beta}$ when its dimension grows with the number of samples. There are two types of *statistical guarantees* of a penalized estimate that are of interest in this setting (as mentioned by Bartlett et al. [3]):

1. *Prediction error* (Persistence): $\hat{\beta}$ performs well on future samples

$$\left(\text{i.e., } \mathrm{E}\left[X\left(\hat{\beta} - \beta^*\right)\right]^2 \text{ (or its empirical version) is small, called persistence}\right).$$

2. ℓ_1*-estimated error*: $\hat{\beta}$ approximates some "true" parameter β^*

$$\left(\text{i.e., } \left\|\hat{\beta} - \beta^*\right\|_1 \text{ is small with high probability}\right).$$

The two types of statistical guarantees can be obtained by following error bounds (say oracle inequalities)

$$\left\|\hat{\beta}-\beta^*\right\|_1 \le O_p(s\lambda_n), \qquad \mathrm{E}\left[X\left(\hat{\beta}-\beta^*\right)\right]^2 \le O_p\left(s\lambda_n^2\right),$$

where $\lambda_n \to 0$ is a tuning parameter and $s := \|\beta^*\|_0$.

Deriving oracle inequalities is a powerful mathematical skill that provides deep insight into the nonasymptotic fluctuation of an estimator compared to the ideal unknown parameter (it is called an oracle). Under linear models with group sparsity covariables, Lounici et al. [18] show oracle inequalities for estimation error (in terms of mixed $(2,p)$-norm) and prediction error (for fixed design). Blazere et al. [5] study the properties of group Lasso estimator in sparse high-dimensional generalized linear models (GLMs) with group sparsity of the covariates, and the oracle inequalities for the prediction and estimation error. Structured sparsity has recently attracted attention to the high-dimensional data. [36] focus on the oracle inequalities for GLMs with overlapping group structures. Zhou et al. There have been considerable developments in oracle inequalities, not limited to the linear models and GLMs. Lemler [17] introduces a data-driven weighted Lasso to estimate Cox models by approximating the intensity (without using partial likelihood), and oracle inequalities in terms of an appropriate empirical K-L divergence are obtained. By focusing on misspecified Cox models with their partial likelihood, Kong and Nan [16] derive the nonasymptotic oracle inequalities for the weighted Lasso penalized negative log partial likelihood function. Similar results have been proposed for Cox models with time-dependent covariances, see Huang et al. [13] for using martingale analysis of KKT conditions. Honda and Hardle [11] consider group SCAD-type and the adaptive group Lasso estimator to do variable selection for Cox models with varying coefficients, and the L_2 convergence rate is obtained for increasing-dimension setting $p/n \to 0$.

Contributions:

- The existing work on weighted group Lasso penalized Cox models has little attention on theoretical results. Yan and Huang [28] propose a weighted group Lasso method that selects important time-dependent variables with a group structure. We propose the oracle inequalities for the prediction and estimation error under the random design, which is different to Huang et al. [13] and Kong and Nan [16] (they do not consider the random design and prediction error).
- Huang et al. [13] do not give a clear definition of the true coefficient, our true coefficient in the oracle inequalities is defined by the minimizer of the expected loss function. It is applicable for misspecified Cox models.
- We provide unified nonasymptotic results in terms of oracle inequalities for prediction and estimation error, and this provides a theoretical justification for the consistency of weighted group Lasso estimator in Cox models (time-dependent covariates and random design).

The sections are presented as follows. Section 2 gives a brief review of Cox models. Section 3 presents the weighted group Lasso penalty for misspecified Cox models. Section 4 shows the oracle inequalities for prediction and estimation for weighted group Lasso penalized partial likelihood for misspecified Cox models, while detailed proofs are included in Sect. 5.

2 A brief review of Cox models

The celebrated Cox models have provided a tremendously successful tool for exploring the association of covariates with failure time and survival distributions. In order to match the drop-out situation in clinical trails, we consider that the continuous survival time T_i^* is governed by random right censoring. For subject i, let $T_i := T_i^* \wedge C_i$ be the observed survival time which is right-censored by C_i. And the censored indicator is denoted by $\Delta_i = 1(T_i^* \le C_i)$. Let $\{z_i(t)\}_{i=1}^n$ be the p-dimensional time-dependent covariates, where $z_i(t) := (z_{i1}(t), \dots, z_{ip}(t))^\tau$. Here we assume that the censoring is noninformative. The time-dependent covariates may degenerate to time-independent covariates, i.e., $z_{ik}(t) \equiv z_{ik}$ for some index k. For example, the CD4 count (relate to longitudinal process) is time-dependent. The time-independent covariates are baseline covariates (i.e., internal variables), which includes treatment indicator ages, sex, treatment indicator, and so on.

Suppose that we observe n independent and identically distributed (i.i.d.) data

$$\left\{T_i, \Delta_i, \left\{z_i(t)\right\}_{0\le t\le\tau}\right\}_{i=1}^n, \tag{2.1}$$

which is sampling from the random population $(T, \Delta, \{z(t)\}_{0\le t\le\tau})$.

Let $S(t|\mathcal{Z}) = P(T > t|\mathcal{Z})$ be the conditional survival function, where $\mathcal{Z}$ is the sigma algebra generated by some covariate variables. The relation of conditional distribution function and $S(t|\mathcal{Z})$ is $F(t|\mathcal{Z}) = P(T \le t|\mathcal{Z}) = 1 - S(t|\mathcal{Z})$. Denote $f(t|\mathcal{Z}) = \frac{d}{dt}F(t|\mathcal{Z})$ as the conditional probability density function. Different from the linear model for modeling conditional mean or the quantile regression for modeling conditional quantiles, the Cox models (also called proportional hazards regression or Cox regressions) aim to model the conditional hazard rate defined by

$$h(t|\mathcal{Z}) := \lim_{h\to 0} \frac{P(t \le T < t + h|T \ge t, \mathcal{Z})}{h} = \frac{f(t|\mathcal{Z})}{S(t|\mathcal{Z})} = -\frac{\partial \log S(t|\mathcal{Z})}{\partial t}. \tag{2.2}$$

The $h(t|\mathcal{Z})$ is the conditional hazard rate at time t conditional on survival until time t or later (i.e., $T \ge t$). From (2.2), the $S(t|\mathcal{Z})$ can be represented as the exponential integral of the cumulative hazard function defined by $H(t) = \int_0^t h(s)\,ds$, i.e., $S(t|Z) = \exp\{-\int_0^t h(s)\,\mathrm{d}s\} \equiv e^{-H(t)}$.

Having obtained the covariates $\{z_i(t)\}_{i=1}^n$, our aim is to model the conditional hazard function of survival time $\{T_i\}_{i=1}^n$ in a finite time interval $[0, \tau]$ by the following semiparametric regressions:

$$h_i(t) := h(t|z_i) = h_0(t)\exp\left\{z_i^\tau(t)\beta^*\right\} \quad \text{for } 0 \le t \le \tau < \infty, \tag{2.3}$$

where $h_0(t)$ is an unknown baseline hazard function, and $\beta^* \in \mathbb{R}^p$ is an unknown parameter which needs to be estimated.

By profiling our the term $h_0(t)$, Cox [6] suggests that the inference on β^* is based on the random likelihood function

$$L_n(\beta; T, z, \Delta) = \prod_{i=1}^n \left\{ \frac{e^{z_i^\tau(T_i)\beta}}{\sum_{j\in R_i} e^{z_j^\tau(T_i)\beta}} \right\}^{\Delta_i}, \tag{2.4}$$

where $R_i = \{j : T_j \ge T_i\}$ is the risk set (set of individuals whose survival times are greater than T_i). In a later paper, Cox [7] strictly derives the so-called partial likelihood function.

Suppose that the observed time is a continuous variable, and there is no tie in the observation time. The joint likelihood for the i.i.d. data (2.1) can be written as follows:

$$\begin{aligned} L_n(\beta, z, \Delta) &= \prod_{i:\Delta_i=1} f(T_i|z_i) \prod_{i:\Delta_i=0} \big(1 - F(T_i|z_i)\big) \\ &= \prod_{i=1}^{n} \big[f(T_i|z_i)\big]^{\Delta_i} \big[S(T_i|z_i)\big]^{1-\Delta_i} = \prod_{i=1}^{n} \big[h(T_i|z_i)\big]^{\Delta_i} S(T_i|z_i) \\ &= \prod_{i=1}^{n} \big\{e^{z_i^\tau(T_i)\beta} h_0(T_i)\big\}^{\Delta_i} \exp\left\{-\int_0^{T_i} h_0(s) e^{z_i^\tau(s)\beta}\, \mathrm{d}s\right\} \\ &= \exp\left\{\sum_{i=1}^{n} \left[\Delta_i \big\{z_i^\tau(T_i)\beta + \log h_0(T_i)\big\} - \int_0^{T_i} h_0(s) e^{z_i^\tau(s)\beta}\, \mathrm{d}s\right]\right\}, \end{aligned} \tag{2.5}$$

which contains the unknown $h_0(\cdot)$.

The key to deriving (2.4) is by specifying a reasonable estimator $\hat{h}_0(\cdot)$ for $h_0(\cdot)$ in (2.3). Assume that $h_0(\cdot)$ is discrete with mass $h_0(T_{(1)}), \ldots, h_0(T_{(k)})$ at the ordered observed survival time $T_{(1)} < \cdots < T_{(k)}$. Denote $\{z_{(o)}(T_{(o)}) : o = 1, \ldots, k\}$ as the k covariates corresponding to the ordered observed survival times $T_{(o)}$. The baseline cumulative hazard function $H_0(t)$ is modeled non-parametrically as the step function $H_0(t) = \sum_{o=1}^{k} h_0(T_{(o)}) I(T_{(o)} \le t)$, and hence $\sum_{i=1}^{n} \int_0^{T_i} h_0(s) e^{z_i^\tau(s)\beta}\, \mathrm{d}s = \sum_{i=1}^{n} \sum_{o=1}^{k} h_0(T_{(o)}) I(T_{(o)} \le T_i) e^{z_i^\tau(T_{(o)})\beta}$.

From (2.5), the joint log-likelihood function is expressed as follows:

$$\begin{aligned} &\log L_n(\beta; T, z, \Delta) \\ &= \sum_{o=1}^{k} \big\{z_{(o)}^\tau(T_{(o)})\beta + \log h_0(T_{(o)})\big\} - \sum_{o=1}^{k} \sum_{i=1}^{n} I(T_{(o)} \le T_i) h_0(T_{(o)}) e^{z_i^\tau(T_{(o)})\beta} \\ &= \sum_{o=1}^{k} \big\{z_{(o)}^\tau(T_{(o)})\beta + \log h_0(T_{(o)})\big\} - \sum_{o=1}^{k} \sum_{\{j:T_j \ge T_{(o)}\}} h_0(T_{(o)}) e^{z_j^\tau(T_{(o)})\beta}, \end{aligned} \tag{2.6}$$

where $\{j : T_j \ge T_{(o)}\}$ denotes the set of individual js who are "at risk" for failure at time $T_{(o)}$.

Taking derivative on $\log L_n(\beta; T, z, \Delta)$ with respect to $h_0(T_{(o)}), o = 1, \ldots, k$, we get

$$\hat{h}_0(T_{(o)}) = \left[\sum_{\{j:T_j \ge T_{(o)}\}} e^{z_j^\tau(T_{(o)})\beta}\right]^{-1},$$

which is also called Breslow's estimator for the baseline hazard function.

Plugging $\hat{h}_0(T_{(o)})$ into (2.6), we have

$$\begin{aligned} \log L_n(\beta; T, z, \Delta) &\propto \sum_{o=1}^{k} \left[z_{(o)}^\tau(T_{(o)})\beta - \log \sum_{\{j:T_j \ge T_{(o)}\}} e^{z_j^\tau(T_{(o)})\beta}\right] \\ &\propto \sum_{i=1}^{n} \left\{z_i^\tau(T_i)\beta - \log\left[\sum_{j=1}^{n} 1(T_j \ge T_i) \exp\big\{z_j^\tau(T_i)\beta\big\}\right]\right\} \Delta_i, \end{aligned}$$

which gives (2.4).

Following the counting process framework in Andersen and Gill [2], let $N_i(t) = 1(T_i \le t, \Delta_i = 1)$ be the counting process, and denote $Y_i(t) =: 1(T_i \ge t)$ to be the at-risk process for subject i. The σ-filtration is defined by $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), z_i(s), s \le t, i = 1, \dots, n\}$, which represents the information that occurs up to time t. Let $\mathrm{d}N_i(s) := 1\{T_i \in [s, s + \mathrm{d}s], \Delta_i = 1\}$. The negative log-partial-likelihood (2.4) for data (2.1) is rewritten as follows:

$$\begin{aligned}
&\ell_n(\beta; T, z, \Delta) \\
&\quad := -\frac{1}{n}\sum_{i=1}^{n}\left\{z_i^\tau(T_i)\beta - \log\left[\sum_{j=1}^{n} 1(T_j \ge T_i)\exp\left\{z_j^\tau(T_i)\beta\right\}\right]\right\}\Delta_i \\
&\quad \propto -\frac{1}{n}\left(\sum_{i=1}^{n}\int_0^t z_i^\tau(u)\beta\,\mathrm{d}N_i(u) - \int_0^t \log\left[\frac{1}{n}\sum_{j=1}^{n} 1(T_j \ge u)\exp\left\{z_j^\tau(u)\beta\right\}\right]\mathrm{d}\overline{N}(u)\right) \\
&\quad = -\frac{1}{n}\sum_{i=1}^{n}\int_0^t \left[z_i^\tau(u)\beta - \log R_n(u,\beta)\right]\mathrm{d}N_i(u), \qquad (2.7)
\end{aligned}$$

where $R_n(u,\beta) = \frac{1}{n}\sum_{j=1}^{n} 1(T_j \ge u)\exp\{z_j^\tau(u)\beta\}$ is the empirical relative risk function.

The negative log-partial likelihood function (2.7), as the summands are neither independent nor Lipschitz, can be approximated by the following intermediate empirical loss function:

$$\begin{aligned}
\tilde{\ell}_n(\beta; T, z, \Delta) &= -\frac{1}{n}\sum_{i=1}^{n}\left\{z_i^\tau(T_i)\beta - \log R(T_i,\beta)\right\}\Delta_i \\
&= \ell_n(\beta; T, z, \Delta) + \frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{R_n(T_i,\beta)}{R(T_i,\beta)}\right\}\Delta_i \qquad (2.8)
\end{aligned}$$

with expected relative risk function defined by $R(t,\beta) = \mathrm{E}[1(T \ge t)\exp\{z^\tau(t)\beta\}]$.

We define the loss function by $l(\beta; T, z, \Delta) := -[z^\tau(t)\beta - \log R(t,\beta)]\Delta$.

Let $\overline{N}(t) := \sum_{i=1}^{n} N_i(t)$. The gradient of $\ell_n(\beta; T, z, \Delta)$ can be written as

$$\nabla\ell_n(\beta; T, z, \Delta) := \frac{\partial\ell_n(\beta; T, z, \Delta)}{\partial\beta} = -\frac{1}{n}\sum_{i=1}^{n}\int_0^t \left[z_i(u) - \overline{z}_n(u,\beta)\right]\mathrm{d}N_i(u), \qquad (2.9)$$

where $\bar{z}_n(u,\beta) = \frac{1}{n}\sum_{j=1}^{n}\frac{Y_j(u)\mathrm{e}^{z_j^\tau(u)\beta}}{R_n(u,\beta)}z_j(u)$ is the random weighted sum of covariates.

The $\nabla\ell_n(\beta; T, z, \Delta)$ is called score process, which is a martingale adapted to the filtration $\mathcal{F}_t$. Furthermore, the Hessian matrix of $\ell_n(\beta; T, z, \Delta)$ is

$$\nabla^2\ell_n(\beta; T, z, \Delta) = \frac{1}{n}\int_0^t V_n(u,\beta)\,\mathrm{d}\overline{N}(u),$$

where $V_n(u,\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(u)\mathrm{e}^{z_i^\tau(u)\beta}}{R_n(u,\beta)}[z_i(s) - \overline{z}_n(u,\beta)][z_i(s) - \overline{z}_n(u,\beta)]^\tau$ is the random weighted sample covariance matrix. Readers can refer to technical details required to make the counting process rigorous in Andersen et al. [1].

3 Weighted group lasso for misspecified Cox models

In this section, we present the concepts and mathematics notations for the penalized misspecified Cox models with the group structure.

Many high-dimensional variables in microarrays data and other scientific applications have a natural group structure. It is better to divide p variables into small sets of variables based on biological knowledge, see Kanehisa and Goto [14], Wang et al. [27]. Suppose that the p-dimensional covariate X is divided into G_n groups each of size d_g for $g \in \{1, \dots, G_n\}$,

$$X_i = \left(X_i^1, \dots, X_i^g, \dots, X_i^{G_n}\right)^\tau, \quad i = 1, \dots, n,$$

where $X_i^g = (X_{i,1}^g, \dots, X_{i,d_g}^g)^T$ and $\sum_{g=1}^{G_n} d_g = p$.

It is allowed that the number of groups increases with the sample size n and $G_n \gg n$. We define the two quantities

$$d_{\max} := \max_{g \in \{1,\dots,G_n\}} d_g \quad \text{and} \quad d_{\min} := \min_{g \in \{1,\dots,G_n\}} d_g,$$

which are crucial constants in the theoretical analysis.

For $\beta \in \mathbb{R}^p$, let β^g be the sub-vector of β whose indexes correspond to the index set of the gth group of X. Given a proper tuning parameter λ, we are interested in weighted group Lasso estimator which achieves group sparsity. It is obtained as the solution of the convex optimization problem:

$$\hat{\beta}_n = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \ell_n(\beta; T, z, \Delta) + \lambda \sum_{g=1}^{G_n} w_g \left\| \beta^g \right\|_2 \right\}, \tag{3.1}$$

where $\| \cdot \|_2$ refers to the Euclidian norm and w_g is a given weighted function.

If all d_g are of size one and $w_g = 1$, then $\sum_{g=1}^{G_n} w_g \|\beta^g\|_2$ reduces to $\|\beta\|_1$ which is essentially a Lasso problem; If all d_g are of size one and $\{w_j\}_{j=1}^p$ are data-dependent weights (the weights only depend on observed data). Let $W = \operatorname{diag}\{w_1, \dots, w_p\}$, thus the weighted group Lasso penalty $\sum_{g=1}^{G_n} d_g \|\beta^g\|_2$ becomes weighted Lasso penalty $\|W\beta\|_1$. Increasing λ leads to the shrinkage of β^g tending to zero, which indicates that some blocks of β diminish to zero simultaneously, and groups of predictors are eliminated from the model. Typically in the reference, they usually choose $w_g := \sqrt{d_g}$ to penalize more heavily groups of large size. For adaptive group Lasso in Cox models, Yan and Huang [28] use $w_g = \sqrt{d_g}/\|\tilde{\beta}^g\|$, where d_g is the size of group g and $\tilde{\beta}^g$ is some consistent estimator of β^g.

Taking the subdifferential of the objective function (3.1), we get the first order condition:

$$\begin{cases} \frac{\partial \ell_n(\beta; T, z, \Delta)}{\partial \beta_g}|_{\beta_g = \hat{\beta}_g} = \lambda w_k \frac{\hat{\beta}_g}{\|\hat{\beta}_g\|_2} & \text{if } \hat{\beta}_g \neq 0, \\ \| \frac{\partial \ell_n(\beta; T, z, \Delta)}{\partial \beta_g}|_{\beta_g = \hat{\beta}_g} \|_2 \leq \lambda w_k & \text{if } \hat{\beta}_g = 0. \end{cases} \tag{3.2}$$

(It is also called Karush–Kuhn–Tucker (KKT) condition, see Sect. 2.2 of Huang et al. [13] for un-group version.) From the adaptive estimation point of view, the weights in equation (3.1) can be determined from the observed data, where KKT conditions (3.2) hold with high probability, for example, $1 - p^r, r < 0$. Applying the concentration inequalities to martingale, the data-driven weights $\{w_j\}_{j=1}^p$ are obtained from the KKT conditions with high probability, see Huang et al. [12] and the references therein. The motif of this work is to derive nonasymptotic oracle inequalities in a mathematical view. The choice of optimal adaptive weight and statistical inferences (confidence interval, testing the coefficient, FDR control) is left for future studies.

In the high-dimensional settings, we study the estimation and prediction of the oracle inequalities for the weighted group Lasso even when the number of groups is extremely greater than the sample size, i.e., $G_n \gg n$. Define $H^* = \{g : \beta_g^* \neq 0\}$ as the group index set corresponding to the nonzero sub-vectors of β^*.

Let $X_1, \ldots, X_n$ be a random sample from a measure $\mathbb{P}$ on a measurable space $(\mathcal{X}, \mathcal{A})$. We denote the empirical distribution as a discrete uniform measure $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ_x is the probability distribution that is degenerate at x.

The expected loss function is defined by

$$\ell(\beta; T, z, \Delta) = -\mathrm{E}\big[\big\{z^\tau(T)\beta - \log R(T, \beta)\big\}\Delta\big] =: \mathrm{E}l(\beta; T, z, \Delta).$$

Corresponding to the form of estimator, the true parameter of the misspecified Cox models is the minimizer of the expected loss function

$$\beta^* = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \mathbb{P}l(\beta; T, z, \Delta) = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} -\mathrm{E}\big\{\big[z^\tau(T)\beta - \log R(T, \beta)\big]\Delta\big\}, \tag{3.3}$$

where $R(t, \beta) = \mathrm{E}[1(T \geq t)\exp\{z^\tau(t)\beta\}]$.

Definition (3.3) was pioneeringly studied in Struthers and Kalbfleisch [21] by clarifying the true parameter as a solution of estimating equation neatly mentioned in the proof of Lemma 3.1 in Andersen and Gill [2].

Here, the expectation of the random variables in the model is unknown, thus as well as β^*. By solving the optimization problem in (3.3), β^* satisfies

$$\beta^* = \left\{\beta \in \mathbb{R}^p : \mathbb{P}\dot{l}(\beta; T, z, \Delta) = -\mathrm{E}\left[\left\{z(T) - \frac{\mathrm{E}[Y(t)z(T)\mathrm{e}^{z^\tau(T)\beta}]}{\mathrm{E}[Y(t)\mathrm{e}^{z^\tau(T)\beta}]}\right\}\Delta\right] = 0\right\}. \tag{3.4}$$

In order to get the unique solution in (3.4), we require that the Hessian matrix for expected loss function

$$\begin{aligned}\mathrm{E}\ddot{l}(\beta; T, z, \Delta) = \mathrm{E}\Bigg[\Bigg\{&\frac{\mathrm{E}[Y(t)z(T)z^\tau(T)\mathrm{e}^{z^\tau(T)\beta}]}{\mathrm{E}[Y(t)\mathrm{e}^{z^\tau(T)\beta}]} \\ &- \frac{\mathrm{E}[Y(t)z(T)\mathrm{e}^{z^\tau(T)\beta}]\mathrm{E}[Y(t)z^\tau(T)\mathrm{e}^{z^\tau(T)\beta}]}{(\mathrm{E}[Y(t)\mathrm{e}^{z^\tau(T)\beta}])^2}\Bigg\}\Delta\Bigg]\end{aligned} \tag{3.5}$$

is nonpositive definite.

We aim to estimate sparse β^* and to predict the hazard function $h(t|z_i(t))$ conditionally on a given process $z_i(t)$. To facilitate the technical proof, additional assumptions are required.

- (H.1): The covariates $\{z_{ij}(t)\}$ are almost surely bounded by a positive constant L, i.e.,

$$\sup_{0 \leq t \leq \tau} \max_{1 \leq i \leq n, 1 \leq j \leq p} \big|z_{ij}(t)\big| \leq L, \quad \text{a.s.}$$

- (H.2): Assume that the parameter space is compact, $\|\beta^*\|_1 \leq B$, where B is a positive constant.

- (H.3): There exists a large constant M such that $\hat{\beta}$ is in the weighted ℓ_2-ball

$$\mathcal{S}_M(\beta^*) := \left\{ \beta \in \mathbb{R}^p : \sum_{g=1}^{G_n} w_g \left\| \beta^g - \beta^{*g} \right\|_2 \le M \right\}.$$

- (H.4): Under $\Delta = 1$, there exists a constant $c_l > 0$ and $c_u < \infty$ such that $\ddot{l}(\beta; t, z, \Delta)$ is uniformly positive definite for all $\beta \in \mathcal{S}_M(\beta^*)$

$$c_u z(t) z^\tau(t) \succ \mathrm{E}\left[\ddot{l}(\beta; T, z, \Delta) | z(t)\right] \succ c_l z(t) z^\tau(t) \quad \text{a.s.}$$

(H.1) and (H.2) are standard assumptions in deriving consistency property for regularized GLMs, see Blazere et al. [5], Zhang and Wu [33]. (H.2) is also used in Zhao et al. [35] for the increasing dimensional Cox models with interval-censored data. (H.3) has been addressed by Kong and Nan [16]. (H.4) makes sure the object function for a minimizer of population expected loss is strongly convex, a similar assumption is used in Andersen and Gill [2], Fan and Li [9].

As mentioned by one reviewer, we often assume that the data are generated from the model with some baseline hazard function and some true parameter β^*. In (3.3), the true parameter is defined as the minimizer of true loss function. We present it in detail from Theorem 1 in Struthers and Kalbfleisch [21].

Lemma 3.1 (Consistency) *Let the expectation* E *be taken with respect to randomness of* $\{(T_i, \Delta_i, z_i(t))\}_{i=1}^n$ *from the true model. Consider the following notations for* $r = 0, 1, 2$:

$$S^{(r)}(t) = n^{-1} \sum_{i=1}^n Y_i(t) h_0(t) e^{z_i^\tau(t)\beta^*} z_i(t)^{\otimes r}, \qquad s^{(r)}(t) = \mathrm{E}\left[S^{(r)}(t)\right],$$

$$S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^n Y_i(t) e^{z_i^\tau(t)\beta} z_i(t)^{\otimes r}, \qquad s^{(r)}(\beta, t) = \mathrm{E}\left[S^{(r)}(\beta, t)\right],$$

where, for a column vector a, $a^{\otimes 2}$ *refers to the matrix* aa^T, $a^{\otimes 1}$ *refers to the vector* a, *and* $a^{\otimes 0}$ *refers to the scalar* 1. *Consider the following conditions.*

Condition 3.1 There exists a neighborhood $\mathcal{S}_M(\beta^*)$ of β^* such that, for each $t < \infty$,

$$\sup_{x \in [0,t], \beta \in \mathcal{S}_M(\beta^*)} \left| S^{(0)}(\beta, x) - s^{(0)}(\beta; x) \right| \to 0, \quad \text{in probability as } n \to \infty.$$

Condition 3.2 (a). The $s^{(0)}(\beta, x)$ is bounded away from zero on $\mathcal{S}_M(\beta^*) \times [0, t]$, and $s^{(0)}(\beta, x)$ and $s^{(1)}(\beta, x)$ are bounded on $\mathcal{S}_M(\beta^*) \times [0, t]$; (b). For each $t < \infty$, we have $\int_0^t s^{(2)}(x)\, dx < \infty$.

- When the data are generated from the correctly specified Cox models (2.3), under Conditions 3.1 and 3.2, we have that the maximum partial likelihood estimator $\hat{\beta}$ is a consistent estimator for β^*, where β^* is the solution to the equation $h(\beta) = 0$ with

$$h(\beta) := \int_0^\infty s^{(1)}(t)\, dt - \int_0^\infty \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} s^{(0)}(t)\, dt.$$

- When the model is misspecified, i.e., suppose that the true hazard function is $h_i(t) \neq h_0(t)e^{z_i^\tau(t)\beta^*}$. If $S^{(r)}(t)$ and $s^{(r)}(t)$ are replaced by $S_m^{(r)}(t) := n^{-1}\sum_{i=1}^n Y_i(t)h_i(t)z_i(t)^{\otimes r}$ and $s_m^{(r)}(t) := \mathrm{E}[S_m^{(r)}(t)]$ in Conditions 3.1 and 3.2, then the solution of the equation $h_m(\beta) = 0$ with

$$h_m(\beta) := \int_0^\infty s_m^{(1)}(t)\,dt - \int_0^\infty \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)} s_m^{(0)}(t)\,dt \tag{3.6}$$

  is the *pseudo-true parameter* β^*.

Since $dM_i(t) := dN_i(t) - 1(T_i \geq t)h_0(t)e^{z_i^\tau(t)\beta^*}\,dt$ is mean-zero $\mathcal{F}_t$-martingale from the theory in Andersen and Gill [2], by comparing the empirical version (2.9) and the population version (3.4) with the limits $s^{(0)}(t)$, $s^{(1)}(t)$ and $s^{(0)}(\beta,t)$, $s^{(1)}(\beta,t)$, we can see that (3.4) coincides with (3.6). Moreover, our assumptions (H.1)–(H.5) verify Conditions 3.1 and 3.2 without confliction if the uniform law of large numbers is applied by using the compactness of β^* and the boundedness of covariates.

4 Oracle inequalities for estimation and prediction

As a powerful mathematical skill, oracle inequalities provide deep insight into the nonasymptotic fluctuation of an estimator compared to the unknown true parameter. A comprehensive theory of oracle inequalities in high-dimensional regressions has been developed for Lasso and its generalization, see Chap. 7 of Wainwright [26].

4.1 Key of nonasymptotic analysis

In this section, nonasymptotic oracle inequalities for weighted group Lasso estimates of Cox models are sought, as well as assumptions of the required restricted eigenvalue (such as group stabil condition). The proof leans on several steps:

- *Step1*: To avoid ill behavior of Hessian, propose the *restricted eigenvalue condition* or other analogous conditions about the design matrix.
- *Step2*: Find the tuning parameter based on high-probability event (*KKT conditions* or other KKT-like conditions).
- *Step3*: According to some restricted eigenvalue assumptions and tuning parameter selection, derive the oracle inequalities via *the definition of weighted group Lasso optimality* and *the minimizer under unknown expected risk function* and *some basic inequalities*. There are three sub-steps:
  - (i) Under the KKT-like conditions, show that the error vector $\hat\beta - \beta^*$ is in some restricted set with structure sparsity, and moreover check that $\hat\beta - \beta^*$ is in a big compact set;
  - (ii) Show that likelihood-based divergence of $\hat\beta$ and β^* can be lower bounded by some quadratic distance between $\hat\beta$ and β^*;
  - (iii) By some elementary inequalities and (ii), show that $\sum_{g=1}^{G_n} w_g\|\hat\beta_n^g - \beta^{*g}\|_2$ is in a smaller compact set with radius of optimal rate (proportional to λ).

As mentioned by one reviewer, our general framework of the proof is quite standard, but consecutive steps of defining some high-probability events rely on nontrivial new results. For simplicity, we introduce and use the notation in empirical processes, see van der Vaart and Wellner [25].

Let $X_1,\ldots,X_n$ be a random sample from a measure $\mathbb{P}$ on a measurable space $(\mathcal{X},\mathcal{A})$. We denote the empirical distribution as a discrete uniform measure $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$, where δ_x is the probability distribution that degenerates at x.

Given a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we write $\mathbb{P}_n f$ for the expectation of f under the empirical measure $\mathbb{P}_n$, and Pf for the expectation under P. Thus

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad Pf = \int f\,dP.$$

The $\mathbb{P}_n f$ is called *empirical processes* index by n. In fact, we treat $\mathbb{P}_n$ and P as operators rather than the measure.

It follows from (2.8) and $\mathbb{P}_n l(\beta; T, z, \Delta) := \tilde{\ell}_n(\beta; T, z, \Delta)$ that

$$\ell_n(\beta; T, z, \Delta) = \mathbb{P}_n l(\beta; T, z, \Delta) - \frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{R_n(T_i, \beta)}{R(T_i, \beta)}\right\}\Delta_i. \tag{4.1}$$

4.2 Define some events with high probability

Using the definition of $\hat{\beta}_n$ in (3.3), we have

$$\ell_n(\hat{\beta}_n; T, z, \Delta) + \lambda\sum_{g=1}^{G_n} w_g \left\|\hat{\beta}_n^g\right\|_2 \le \ell_n\big(\beta^*; T, z, \Delta\big) + \lambda\sum_{g=1}^{G_n} w_g \left\|\beta^{*g}\right\|_2. \tag{4.2}$$

Hence we get

$$\begin{aligned}
&\mathbb{P}\big(l(\hat{\beta}_n; T, z, \Delta) - l\big(\beta^*; T, z, \Delta\big)\big) + \lambda\sum_{g=1}^{G_n} w_g \left\|\hat{\beta}_n^g\right\|_2 \\
&\quad\le \big[\ell_n\big(\beta^*; T, z, \Delta\big) - \mathbb{P}l\big(\beta^*; T, z, \Delta\big)\big] - \big[\ell_n(\hat{\beta}_n; T, z, \Delta) - \mathbb{P}l(\hat{\beta}_n; T, z, \Delta)\big] \\
&\qquad + \lambda\sum_{g=1}^{G_n} w_g \left\|\beta^{*g}\right\|_2.
\end{aligned} \tag{4.3}$$

Then, by (4.1), the first and second terms in the right-hand side of (4.3) are

$$\begin{aligned}
&\big[\ell_n\big(\beta^*; T, z, \Delta\big) - \mathbb{P}l\big(\beta^*; T, z, \Delta\big)\big] \\
&\quad= \big[\mathbb{P}_n l\big(\beta^*; T, z, \Delta\big) - \mathbb{P}l\big(\beta^*; T, z, \Delta\big)\big] - \frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{R_n(T_i, \beta^*)}{R(T_i, \beta^*)}\right\}\Delta_i, \\
&\big[\ell_n(\hat{\beta}_n; T, z, \Delta) - \mathbb{P}l(\hat{\beta}_n; T, z, \Delta)\big] \\
&\quad= \big[\mathbb{P}_n l(\hat{\beta}_n; T, z, \Delta) - \mathbb{P}l(\hat{\beta}_n; T, z, \Delta)\big] - \frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{R_n(T_i, \hat{\beta}_n)}{R(T_i, \hat{\beta}_n)}\right\}\Delta_i.
\end{aligned}$$

It implies

$$\begin{aligned}
&\big[\ell_n\big(\beta^*; T, z, \Delta\big) - \mathbb{P}l\big(\beta^*; T, z, \Delta\big)\big] - \big[\ell_n(\hat{\beta}_n; T, z, \Delta) - \mathbb{P}l(\hat{\beta}_n; T, z, \Delta)\big] \\
&\quad= (\mathbb{P}_n - \mathbb{P})\big(l\big(\beta^*; T, z, \Delta\big) - l(\hat{\beta}_n; T, z, \Delta)\big) - D_n\big(\hat{\beta}, \beta^*\big),
\end{aligned} \tag{4.4}$$

where

$$D_n\big(\hat{\beta}, \beta^*\big) := \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{R_n(T_i, \beta^*)}{R(T_i, \beta^*)}\right\}\Delta_i - \frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{R_n(T_i, \hat{\beta}_n)}{R(T_i, \hat{\beta}_n)}\right\}\Delta_i\right].$$

To obtain oracle inequalities for the weighed group Lasso applied to misspecified Cox models, it is necessary to study the rate of convergence of the empirical process $(\mathbb{P}_n - \mathbb{P})(l(\beta^*; T, z, \Delta) - l(\hat{\beta}_n; T, z, \Delta))$ and $D_n(\hat{\beta}, \beta^*)$. The centralized empirical loss $(\mathbb{P}_n - \mathbb{P})(l(\beta^*; T, z, \Delta) - l(\hat{\beta}_n; T, z, \Delta))$ and the normalized error $D_n(\hat{\beta}, \beta^*)$ represent the fluctuation between the expected loss and sample loss. It will be shown that

$$(\mathbb{P}_n - \mathbb{P})\big(l\big(\beta^*; T, z, \Delta\big) - l(\hat{\beta}_n; T, z, \Delta)\big) \quad \text{and} \quad D_n\big(\hat{\beta}, \beta^*\big)$$

have stochastic Lipschitz properties with respect to $\sum_{g=1}^{G_n} w_g \|\hat{\beta}_n^g - \beta^{*g}\|_2$.

The concentration inequalities are essential tools to obtain an upper bound of (4.4), which is proportional to a regularization parameter that ensures good statistical properties of the regularized estimator with high probability.

Define $F(s,z)$ as the joint distribution of $(T_i, z_i^\tau(t))$. Let $\tilde{\beta} := (\tilde{\beta}_1, \dots, \tilde{\beta}_p)^T$ with the components $\{\tilde{\beta}_j\}_{j=1}^p$ between $\{\hat{\beta}_j\}_{j=1}^p$ and $\{\beta_j^*\}_{j=1}^p$, respectively, via first-order Taylor's expansions of the function

$$f_t(\beta) = \log R(t, \beta) = \log \mathrm{E}\big[1(T_i \geq t) e^{z_i^\tau(t)\beta}\big] = \log \int 1(s \geq t) e^{z_i^\tau(t)\beta}\, dF(s,z)$$

with derivative

$$\frac{df_t(\beta)}{d\beta_j} = \frac{\int z_{ij}^\tau(s) 1(s \geq t) e^{z_i^\tau(t)\beta}\, dF(s,z)}{\int 1(s \geq t) e^{z_i^\tau(t)\beta}\, dF(s,z)}, \quad j = 1, 2, \dots, p.$$

Plugging $t = T_i$, we have componentwise Taylor's expansion

$$\log R\big(T_i, \beta^*\big) - \log R(T_i, \hat{\beta}_n) = \frac{\int z_{ij}(s) 1(s \geq T_i) e^{z_i^\tau(T_i)\tilde{\beta}}\, dF(s,z)}{\int 1(s \geq T_i) e^{z_i^\tau(T_i)\tilde{\beta}}\, dF(s,z)} \big(\beta_j^* - \hat{\beta}_j\big), \quad j = 1, 2, \dots, p.$$

Considering the first term in (4.4), we have

$$\begin{aligned}
&(\mathbb{P}_n - \mathbb{P})\big(l\big(\beta^*; T, z, \Delta\big) - l(\hat{\beta}_n; T, z, \Delta)\big) \\
&\quad = -\frac{1}{n}\sum_{i=1}^n \big[z_i^\tau(T_i)\beta^* - \log R\big(T_i, \beta^*\big)\big]\Delta_i + \frac{1}{n}\sum_{i=1}^n \big[z_i^\tau(T_i)\hat{\beta}_n - \log R(T_i, \hat{\beta}_n)\big]\Delta_i \\
&\qquad - \mathrm{E}\big\{\big[z^\tau(T)\beta^* - \log R\big(T, \beta^*\big)\big]\Delta\big\} + \mathrm{E}\big\{\big[z^\tau(T)\hat{\beta}_n - \log R(T, \hat{\beta}_n)\big]\Delta\big\} \\
&\quad = \frac{-1}{n}\sum_{i=1}^n \sum_{j=1}^p \big(\beta^* - \hat{\beta}\big)\big[z_{ij}(T_i)\Delta_i - \mathrm{E}\big(z_{ij}(T_i)\Delta_i\big)\big] \\
&\qquad + \frac{-1}{n}\sum_{i=1}^n \sum_{j=1}^p \big(\beta_j^* - \hat{\beta}_j\big)\bigg(\frac{\int z_{ij}(s) 1(s \geq T_i) e^{z_i^\tau(T_i)\tilde{\beta}}\, dF(s,z)}{\int 1(s \geq T_i) e^{z_i^\tau(T_i)\tilde{\beta}}\, dF(s,z)} \\
&\qquad - \mathrm{E}\frac{\int z_{ij}(s) 1(s \geq T_i) e^{z_i^\tau(T_i)\tilde{\beta}}\, dF(s,z)}{\int 1(s \geq T_i) e^{z_i^\tau(T_i)\tilde{\beta}}\, dF(s,z)}\bigg) \\
&\quad = \sum_{g=1}^{G_n} \big(\beta_g^* - \hat{\beta}_g\big)\frac{-1}{n}\sum_{i=1}^n \bigg[\frac{z_{ig}^\tau(T_i)}{w_g}\Delta_i - \mathrm{E}\big(z_{ig}^\tau(T_i)\Delta_i\big)\bigg]
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{g=1}^{G_n}\bigl(\beta_g^*-\hat{\beta}_g\bigr)\frac{-1}{n}\sum_{i=1}^{n}\biggl(\frac{\int z_{ig}(s)1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}{\int 1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}\\
&-\mathrm{E}\frac{\int z_{ig}(s)1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}{\int 1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}\biggr)\\
&\leq\sum_{g=1}^{G_n}w_g\bigl\|\beta_g^*-\beta_g\bigr\|_2\cdot\biggl\|\frac{1}{n}\sum_{i=1}^{n}\biggl[\Delta_i\frac{z_{ig}^T(T_i)}{w_g}-\mathrm{E}\biggl(\frac{z_{ig}^\tau(T_i)}{w_g}\Delta_i\biggr)\biggr]\biggr\|_2\\
&+\sum_{g=1}^{G_n}w_g\bigl\|\beta_g^*-\beta_g\bigr\|_2\cdot\biggl\|\frac{1}{n}\sum_{i=1}^{n}\biggl(\frac{\int z_{ig}(s)1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}{w_g\int 1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}\\
&-\mathrm{E}\frac{\int z_{ig}(s)1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}{w_g\int 1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}\biggr)\biggr\|_2. \qquad (4.5)
\end{aligned}
$$

To get the stochastic Lipschitz properties, we define the following two events:

$$
\begin{aligned}
\mathcal{A}_1&=\bigcap_{g=1}^{G_n}\biggl\{\biggl\|\frac{1}{n}\sum_{i=1}^{n}\biggl[\frac{z_{ig}^\tau(T_i)}{w_g}\Delta_i-\mathrm{E}\biggl(\frac{z_{ig}^\tau(T_i)}{w_g}\Delta_i\biggr)\biggr]\biggr\|_2\leq\lambda_{a1}\biggr\},\\
\mathcal{A}_2&=\bigcap_{g=1}^{G_n}\biggl\{\biggl\|\frac{1}{n}\sum_{i=1}^{n}\biggl(\frac{\int z_{ig}(s)1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}{w_g\int 1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}\\
&\quad-\mathrm{E}\frac{\int z_{ig}(s)1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}{w_g\int 1(s\geq T_i)e^{z_i^\tau(T_i)\tilde{\beta}}\,dF(s,z)}\biggr)\biggr\|_2\leq\lambda_{a2}\biggr\}.
\end{aligned}
$$

The random sum in event $\mathcal{A}_2$ is not independent, which renders this problem more challenging. We need to check a uniform version of the event $\mathcal{A}_2$ in terms of β. Concentration inequalities for suprema empirical processes are powerful to check that event $\mathcal{A}_2$ holds with high probability. It will be derived from Talagrand's sharper bounds for suprema empirical processes, which is a generalization of Dvoretzky–Kiefer–Wolfowitz inequality, see Talagrand [22]. Like an index function class for the empirical distribution function, boundedness assumption (H.1) on the components of $z(t)$ guarantees the conditions for concentrations of suprema empirical processes.

Next, an upper bound is obtained for the centralized empirical process $(\mathbb{P}_n-\mathbb{P})[l(\beta^*;T,z,\Delta)-l(\hat{\beta}_n;T,z,\Delta)]$.

Proposition 4.1 *Assume that* (*H*.1)–(*H*.3) *are true. On the event* $\mathcal{A}=\mathcal{A}_1\cap\mathcal{A}_2$, *we have* $P(\mathcal{A})\geq 1-2d_{\max}(2G_n)^{1-A^2}$. *Moreover, the upper bound* (4.6) *holds with the probability as least* $1-2d_{\max}(2G_n)^{1-A^2}$,

$$
(\mathbb{P}_n-\mathbb{P})\bigl(l\bigl(\beta^*;T,z,\Delta\bigr)-l(\hat{\beta}_n;T,z,\Delta)\bigr)\leq\lambda_a\sum_{g=1}^{G_n}w_g\bigl\|\hat{\beta}_n^g-\beta^{*g}\bigr\|_2, \qquad (4.6)
$$

where $\lambda_a := \lambda_{a1} + \lambda_{a2}$ *with*

$$\begin{aligned}
&\lambda_{a1} = \frac{L\sqrt{2d_{\max}}}{w_{\min}}\sqrt{\frac{\log(2G_n)}{n}} \quad \textit{and} \\
&\lambda_{a2} = \frac{2L\sqrt{2d_{\max}}}{w_{\min}}\left(\sqrt{\frac{\log 2p}{n}} + Ae^{2LB}\sqrt{\frac{\log(2G_n)}{n}}\right).
\end{aligned} \tag{4.7}$$

This proposition states that the difference between the centralized empirical processes is bounded from above by the tuning parameter multiplied by the weighted group Lasso norm of the difference between the estimated parameter and the true parameter β^*.

For the normalized error $D_n(\beta, \beta^*)$, set

$$\mathcal{B} = \left\{ \sup_{\beta \in \mathcal{S}_M(\beta^*)} \frac{|D_n(\beta, \beta^*)|}{\sum_{g=1}^{G_n} w_g \|\beta^g - \beta^{*g}\|_2} \le \lambda_a \right\},$$

where $D_n(\beta, \beta^*) := \frac{1}{n}[\sum_{i=1}^n \{\log \frac{R_n(T_i,\beta^*)}{R(T_i,\beta^*)}\} - \sum_{i=1}^n \{\log \frac{R_n(T_i,\beta)}{R(T_i,\beta)}\}]\Delta_i$ and λ_{a2} is a suitable tuning parameter.

Observe that

$$\begin{aligned}
D_n(\beta, \beta^*) &:= \left| \frac{1}{n}\left[\sum_{i=1}^n \left\{ \log \frac{R_n(T_i,\beta^*)}{R(T_i,\beta^*)} \right\} - \sum_{i=1}^n \left\{ \log \frac{R_n(T_i,\beta)}{R(T_i,\beta)} \right\} \right] \Delta_i \right| \\
&= \left| \frac{1}{n}\sum_{i=1}^n \left[\log \frac{1}{n}\sum_{j=1}^n \frac{1(T_j \ge T_i)e^{z_i^\tau(T_i)\beta}}{R(T_i,\beta)} - \log \frac{1}{n}\sum_{j=1}^n \frac{1(T_j \ge T_i)e^{z_i^\tau(T_i)\beta^*}}{R(T_i,\beta^*)} \right] \Delta_i \right| \\
&\le \sup_{0\le t\le\tau} \left| \log \frac{1}{n}\sum_{j=1}^n \frac{1(T_j \ge t)e^{z_i^\tau(t)\beta}}{R(t,\beta)} - \log \frac{1}{n}\sum_{j=1}^n \frac{1(T_j \ge t)e^{z_i^\tau(t)\beta^*}}{R(t,\beta^*)} \right| \\
&=: \left| \log \frac{1}{n}\sum_{i=1}^n \frac{1(T_i \ge t_s)e^{z_i^T(t_s)\beta}}{R(t_s,\beta)} - \log \frac{1}{n}\sum_{i=1}^n \frac{1(T_i \ge t_s)e^{z_i^T(t_s)\beta^*}}{R(t_s,\beta^*)} \right|
\end{aligned} \tag{4.8}$$

for certain random variable t_s on a compact set $[0, \tau]$.

By the first order Taylor's expansion of the function $g_{t_s}(\beta) := \log(\frac{1}{n}\sum_{i=1}^n \frac{1(T_i \ge t_s)e^{z_i^\tau(t_s)\beta}}{R(t_s,\beta)})$, let the corresponding mean value $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ be between β_j^* and β_j for each $j = 1, 2, \ldots, p$. We have

$$\begin{aligned}
&D_n(\beta, \beta^*) \\
&= \left| \sum_{j=1}^p (\beta_j^* - \beta_j) \left[\sum_{i=1}^n \frac{1(T_i \ge t_s)e^{z_i^\tau(t_s)\tilde{\beta}}}{R(t_s,\tilde{\beta})} \right]^{-1} \right. \\
&\quad \times \sum_{i=1}^n \left\{ \frac{1(T_i \ge t_s)z_{ij}(t_s)e^{z_i^\tau(t_s)\tilde{\beta}}R(t_s,\tilde{\beta})}{R^2(t_s,\tilde{\beta})} \right. \\
&\quad \left. \left. - \frac{1(T_i \ge t_s)e^{z_i^\tau(t_s)\tilde{\beta}}\mathrm{E}[1(T \ge t_s)z_{ij}(t_s)e^{z^\tau(t_s)\tilde{\beta}}]}{R^2(t_s,\tilde{\beta})} \right\} \right| \\
&= \left| \sum_{j=1}^p (\beta_j^* - \beta_j) \left\{ \frac{\frac{1}{n}\sum_{i=1}^n 1(T_i \ge t_s)z_{ij}(t_s)e^{z_i^\tau(t_s)\tilde{\beta}}}{\frac{1}{n}\sum_{i=1}^n 1(T_i \ge t_s)e^{z_i^\tau(t_s)\tilde{\beta}}} - \frac{\mathrm{E}[1(T \ge t_s)z_{ij}(T)e^{z^\tau(t_s)\tilde{\beta}}]}{\mathrm{E}[1(T \ge t_s)e^{z^\tau(t_s)\tilde{\beta}}]} \right\} \right|
\end{aligned}$$

$$
\begin{aligned}
&= \left| \sum_{g=1}^{G} w_g \left(\beta_g^* - \beta_g\right)^T \left\{ \frac{\frac{1}{n}\sum_{i=1}^{n} 1(T_i \geq t_s) \frac{z_{ig}(t_s)}{w_g} e^{z_i^\tau(t_s)\tilde{\beta}}}{\frac{1}{n}\sum_{i=1}^{n} 1(T_i \geq t_s) e^{z_i^\tau(t_s)\tilde{\beta}}} - \frac{\mathrm{E}[1(T \geq t_s)\frac{z_{ig}(t_s)}{w_g} e^{z^\tau(t_s)\tilde{\beta}}]}{\mathrm{E}[1(T \geq t_s) e^{z^\tau(t_s)\tilde{\beta}}]} \right\} \right| \\
&\leq \sum_{g=1}^{G} w_g \left\|\beta_g^* - \beta_g\right\|_2 \left\| \frac{\sum_{i=1}^{n} 1(T_i \geq t_s) \frac{z_{ig}(t_s)}{w_g} e^{z_i^\tau(t_s)\tilde{\beta}}}{\sum_{i=1}^{n} 1(T_i \geq t_s) e^{z_i^\tau(t_s)\tilde{\beta}}} \right. \\
&\quad \left. - \frac{\mathrm{E}[1(T \geq t_s)\frac{z_{ig}(t_s)}{w_g} e^{z^\tau(t_s)\tilde{\beta}}]}{\mathrm{E}[1(T \geq t_s) e^{z^\tau(t_s)\tilde{\beta}}]} \right\|_2 . \qquad (4.9)
\end{aligned}
$$

From the following decomposition and inequality

$$
\left\| \frac{a_n}{b_n} - \frac{a}{b} \right\|_2 = \left\| \frac{1}{b_n} \left[(a_n - a) + \frac{a}{b}(b_n - b) \right] \right\|_2 \leq \frac{1}{|b_n|} \left(\|a_n - a\|_2 + \frac{\|a\|_2}{|b|} |b_n - b| \right),
$$

which implies that

$$
\begin{aligned}
&D_n(\beta, \beta^*) \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} 1(T_i \geq t_s) e^{z_i^\tau(t_s)\tilde{\beta}} \right|^{-1} \left\{ \sum_{g=1}^{G} w_g \left\|\beta_g^* - \beta_g\right\|_2 \right. \\
&\quad \times \left[\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{1(T_i \geq t_s) z_{ig}(t_s)}{w_g} e^{z_i^\tau(t_s)\tilde{\beta}} - \mathrm{E}\left[1(T \geq t_s) \frac{z_{ig}(t_s)}{w_g} e^{z^\tau(t_s)\tilde{\beta}} \right] \right\|_2 \right. \\
&\quad \left. + \frac{\|\mathrm{E}[1(T \geq t_s)\frac{z_{ig}(t_s)}{w_g} e^{z^\tau(t_s)\tilde{\beta}}]\|_2}{|\mathrm{E}[1(T \geq t_s) e^{z^\tau(t_s)\tilde{\beta}}]|} \right. \\
&\quad \left. \times \left| \frac{1}{n} \sum_{i=1}^{n} 1(T_i \geq t_s) e^{z_i^\tau(t_s)\tilde{\beta}} - \mathrm{E}\left[1(T \geq t_s) e^{z^\tau(t_s)\tilde{\beta}} \right] \right| \right] \Bigg\} \\
&\leq \left| \frac{1}{n} \sum_{i=1}^{n} 1(T_i \geq t_s) e^{z_i^\tau(t_s)\tilde{\beta}} \right|^{-1} \sum_{g=1}^{G} w_g \left\|\beta_g^* - \beta_g\right\|_2 \\
&\quad \times \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{1(T_i \geq t_s) z_{ig}(t_s)}{w_g} e^{z_i^\tau(t_s)\tilde{\beta}} - \mathrm{E}\left[1(T \geq t_s) \frac{z_{ig}(t_s)}{w_g} e^{z^\tau(t_s)\tilde{\beta}} \right] \right\|_2 \right. \\
&\quad \left. + \frac{L\sqrt{d_g}}{w_{\min}} \left| \frac{1}{n} \sum_{i=1}^{n} 1(T_i \geq t_s) e^{z_i^\tau(t_s)\tilde{\beta}} - \mathrm{E}\left[1(T \geq t_s) e^{z^\tau(t_s)\tilde{\beta}} \right] \right| \right\}, \qquad (4.10)
\end{aligned}
$$

where the last inequality is from

$$
\frac{\|\mathrm{E}[1(T \geq t_s)\frac{z_{ig}(t_s)}{w_g} e^{z^\tau(t_s)\tilde{\beta}}]\|}{|\mathrm{E}[1(T \geq t_s) e^{z^\tau(t_s)\tilde{\beta}}]|} = \frac{1}{w_g} \sqrt{\sum_{j=1}^{d_g} \left(\mathrm{E}\left[\frac{1(T \geq t_s) z_{ij}(t_s) e^{z^\tau(t_s)\tilde{\beta}}}{\mathrm{E}[1(T \geq t_s) e^{z^\tau(t_s)\tilde{\beta}}]} \right] \right)^2} \leq \frac{L\sqrt{d_g}}{w_{\min}}
$$

by using assumptions (H.1)–(H.2).

If we have $\hat{\beta} \in \mathcal{S}_M(\beta^*)$ for some finite M, thus $\tilde{\beta} \in \mathcal{S}_M(\beta^*)$ by

$$
\sum_{g=1}^{G_n} w_g \left\|\tilde{\beta}^g - \beta^{g*}\right\|_2 \leq \sum_{g=1}^{G_n} w_g \sqrt{\sum_{j=1}^{d_g} t_j^2 \left|\hat{\beta}_j - \beta_j^*\right|^2} \leq \sum_{g=1}^{G_n} w_g \left\|\hat{\beta}_n^g - \beta^{*g}\right\|_2 \leq M.
$$

Note that summation (4.10) contains a common random variable t_s which renders (4.10) to be a dependent summation. In order to bound the quotient and the two centralized summations, we denote three events by $\mathcal{B}_0$, $\mathcal{B}_1$, $\mathcal{B}_2$, respectively:

$$\mathcal{B}_0 = \left\{ \sup_{\substack{t_s\in[0,\tau],\\ \beta\in\mathcal{S}_M(\beta^*)}} \frac{1}{n}\sum_{j=1}^{n} 1(T_j \geq t_s)\mathrm{e}^{z_i^\tau(t_s)\beta} \geq U \right\},$$

$$\mathcal{B}_1 = \left\{ \bigcap_{g=1}^{G_n} \left\| \frac{1}{n}\sum_{i=1}^{n} 1(T_i \geq t_s)\frac{z_{ig}(t_s)\mathrm{e}^{z_i^\tau(t_s)\beta}}{w_g} - \mathrm{E}\left[1(T \geq t_s)\frac{z_{ig}(t_s)\mathrm{e}^{z^\tau(t_s)\beta}}{w_g}\right] \right\|_2 \leq \lambda_{b1}U \right\},$$

and

$$\mathcal{B}_2 = \left\{ \sup_{\substack{t_s\in[0,\tau],\\ \beta\in\mathcal{S}_M(\beta^*)}} \left| \frac{1}{n}\sum_{i=1}^{n} 1(T_i \geq t_s)\mathrm{e}^{z_i^\tau(t_s)\tilde{\beta}} - \mathrm{E}\left[1(T \geq t_s)\mathrm{e}^{z^\tau(t_s)\tilde{\beta}}\right] \right| \leq \lambda_{b2}U \right\}. \tag{4.11}$$

To solve the problem, we need the concentration inequalities for the suprema of the empirical processes in $\{\mathcal{B}_l\}_{l=0}^2$ uniformly in $t \in [0,\tau]$ and $\beta \in \mathcal{S}_M(\beta^*)$, see Sect. 2.14 of van der Vaart and Wellner [25].

Let $\mathcal{B} = \mathcal{B}_0 \cap \mathcal{B}_1 \cap \mathcal{B}_2$. We aim to show that each event in $\{\mathcal{B}_l\}_{l=0}^2$ holds with high probability. Thus $\mathcal{B}$ is also a high probability event via utilizing the basic inequality $P(\mathcal{B}) \geq P(\mathcal{B}_0) + P(\mathcal{B}_1) + P(\mathcal{B}_2) - 2$.

Based on (4.10), we obtain the following local stochastic Lipschitz condition under the event $\mathcal{B}$:

$$\frac{|D_n(\hat{\beta},\beta^*)|}{\sum_{g=1}^{G_n} w_g\|\hat{\beta}_n^g - \beta^{*g}\|_2} \leq \sup_{\beta\in\mathcal{S}_M(\beta^*)} \frac{|D_n(\beta,\beta^*)|}{\sum_{g=1}^{G_n} w_g\|\beta^g - \beta^{*g}\|_2} \leq \lambda_b,$$

where λ_b can be viewed as the local stochastic Lipschitz constant.

The following proposition is a similar but significant improvement of Corollary 2 in Kong and Nan [16] from the Lasso to the group Lasso case and from the fixed design to the random design.

Proposition 4.2 *Let $p_\tau := P(T_1 \geq \tau) > 0$, and $D^2(\sqrt{2})$ be a universal constant. Under (H.1)–(H.3) and some constant $A^2 > 2$, we have $P(\mathcal{B}) \geq 1 - 2e^{-np_\tau^2/2} - \frac{d_{\max}D^2(\sqrt{2})A^2\log(G_n)}{4n} \times G_n^{2-A^2} - \frac{D^2(\sqrt{2})A^2\log p}{4n}p^{-A^2}$ with*

$$\lambda_{b1} = \frac{2\sqrt{2}LA\mathrm{e}^{2LB}\sqrt{d_{\max}}}{p_\tau w_{\min}}\sqrt{\frac{\log(G_n)}{n}} \quad \text{and} \quad \lambda_{b2} = \frac{\sqrt{2}A\mathrm{e}^{2LB}}{p_\tau}\sqrt{\frac{\log p}{n}}. \tag{4.12}$$

Moreover, let $\lambda_b := \lambda_{b1} + \lambda_{b2}$, we have

$$D_n(\hat{\beta},\beta^*) \leq \lambda_b\sum_{g=1}^{G_n} w_g\|\hat{\beta}_n^g - \beta^{*g}\|_2$$

with probability at least $1 - 2e^{-np_\tau^2/2} - \frac{d_{\max}D^2(\sqrt{2})A^2\log}{4n}G_n^{2-A^2} - \frac{D^2(\sqrt{2})A^2\log p}{4n}p^{-A^2}$.

If the true model is sparse and $\log p = o(n)$, then the two propositions above illustrate that $P(\mathcal{A}), P(\mathcal{B}) \to 1$ as $p, n \to \infty$.

4.3 Sharp oracle inequalities from restricted eigenvalue conditions

In this section, we give sharp bounds for estimation and prediction errors for Cox models using a weaker condition similar to the *restricted eigenvalue condition* of Bickel et al. [4].

Consider linear models $\{\mathrm{E}[Y_i|X_i] = X_i^{\tau}(t)\beta^*\}_{i=1}^n$ with random covariate vectors $\{X_i\}_{i=1}^n$. The key condition to derive oracle inequalities rests on the correlation between the covariates, i.e., on the behavior of the sample covariance matrix $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$, which is necessarily singular when $p > n$. Let S be any subset of $\{1, 2, \ldots, p\}$. The restricted eigenvalue condition (RE in short) of $p \times p$ matrix Σ_n is defined by

$$RE(\eta, S, \Sigma_n) = \inf_{0 \neq b \in C(\eta, S)} \frac{(b^T \Sigma_n b)^{1/2}}{\|b\|_2} > 0, \tag{4.13}$$

where $C(\eta, S) = \{b \in \mathbb{R}^p : \|b_{S^c}\|_1 \leq \eta \|b_S\|_1\}$, $\eta > 0$.

It should be noted that if we omit the sparse restricted set $C(\eta, S)$, (4.13) leads to $\frac{b^T \Sigma_n b}{\|b\|_2^2} > RE^2(\eta, S, \Sigma_n)$. Thus it means that the smallest eigenvalue of the sample covariance matrix $\mathbf{\Sigma}_n$ is positive, which is impossible when $p > n$ ($\mathbf{\Sigma}_n$ is not full rank). To avoid the low rank of $\mathbf{\Sigma}_n$, Bickel et al. [4] consider the restricted eigenvalue condition under the sparse restricted set $C(\eta, S)$ as considerable relation in the sparse high-dimensional estimation. The restricted eigenvalue is from the restricted strong convexity, which enforces a type of strong convexity condition for the negative log-likelihood function of linear models under certain sparse restrict set.

A shortcoming for (4.13) is that we cannot assume that $RE(\eta, S, \Sigma_n) > 0$ happens with high probability 1. Instead, we replace Σ_n by a non-random version: $\Sigma = \mathrm{E}\Sigma_n$. Observe that $\frac{b^T \Sigma_n b}{\|b_S\|_2^2} \geq \frac{b^T \Sigma_n b}{\|b\|_2^2} > 0$ if (4.13) holds. So $b^T \Sigma_n b \geq k\|b_S\|_2^2 > k\|b_S\|_2^2 - \varepsilon$ for a constant $k > 0$ and a relax constant $\varepsilon > 0$. Technically, for group penalty, here we use a condition which is a modified version of the restricted eigenvalue conditions presented in Blazere et al. [5] for generalized linear models. Define by $H^* = \{g : \beta^{*g} \neq 0\}$ the index set of the groups and $\gamma^* := |H^*|$

Definition (Group stabil condition) Let $c_0, \varepsilon, k > 0$ be given constants. Let Σ be the $p \times p$ non-random matrix, which satisfies the group stabil condition $GS(c_0, \varepsilon, k, H^*)$ if there exists $k > 0$ such that

$$\delta^T \Sigma \delta \geq k \sum_{g \in H^*} \left\|\delta^g\right\|_2^2 - \varepsilon, \quad \forall \delta \in S\big(c_0, H^*\big), \tag{4.14}$$

where the restricted set is defined as $S(c_0, H^*) := \{\delta : \sum_{g \in H^{*c}} w_g \|\delta^g\|_2 \leq c_0 \sum_{g \in H^*} w_g \|\delta^g\|_2\}$.

$S(c_0, H^*)$ is a restricted cone set with group sparsity, which is similar to the condition used by Lounici et al. [18] to prove oracle inequalities for group Lasso in linear models. The ε is an error or relax term that can be set to zero, and we can view k as the smallest generalized eigenvalue of Σ.

If we assume that the group stabil condition is satisfied for the covariance matrix $\Sigma := \mathrm{E}[z(t)z^{\tau}(t)]$ under the restricted cone set $S(c_0, H^*)$ with $\delta = \hat{\beta}_n - \beta^*$, then we check that $\hat{\beta}_n - \beta^* \in S(1, H^*)$ holds with high probability. With the preparation above, we are now able to present the main result of this paper, which provides sharper and minimax optimal bounds for the estimation and prediction error when the true model is sparse and $\log p$ is small as compared to n.

Theorem 4.1 *Let* $\gamma^* := \sum_{g\in H^*} d_g$, $p_\tau := P(T_1 \geq \tau) > 0$ *and* $D^2(\sqrt{2})$ *be a universal constant. Assume that* $(H.1)$–$(H.4)$ *and group stabil condition* $GS(1, \varepsilon_n, k, H^*)$ *are satisfied for* $\Sigma := \mathrm{E}[z(t)z^\tau(t)]$. *If* λ *is chosen such that*

$$\lambda \geq \lambda_{a1} + \lambda_{a2} + \lambda_{b1} + \lambda_{b2} \quad \text{given by (4.7) and (4.12).}$$

Then, with probability at least $(A^2 > 2)$

$$1 - 2d_{\max}(2G_n)^{-A^2/2} - 2e^{-np_\tau^2/2} - \frac{d_{\max}D^2(\sqrt{2})A^2\log(G_n)}{4n}G_n^{2-A^2} - \frac{D^2(\sqrt{2})A^2\log p}{4n}p^{-A^2},$$

we have $\hat{\beta}_n - \beta^* \in S(1, H^*)$ *and*

$$\sum_{g=1}^{G_n} w_g \left\|\hat{\beta}_n^g - \beta^{*g}\right\|_2 \leq \frac{8\gamma^*\lambda}{kc_l} + \frac{c_l\varepsilon_n}{2\lambda},$$

where $c_1 > 0$ *is a constant given in* $(H.4)$.

Moreover, if a new covariate $z^*(t)$ (*the test data*) *is an independent copy of* $z(t)$ (*as the training data*) *and* E^* *represents expectation only about* $z^*(t)$, *then the square prediction error under* $\Delta = 1$ *is*

$$\mathrm{E}^*\left[z^{*\tau}(t)\left(\hat{\beta}_n - \beta^*\right)\right]^2 \leq \frac{32\gamma^*\lambda^2}{kc_1^2} + 2\varepsilon_n$$

under the event $\mathcal{A} \cap \mathcal{B}$.

Consider $\varepsilon_n = 0$. The obtained results are for the fixed design which is analogous to the bounds in Lounici et al. [18] who show the optimal convergence rate of the group Lasso estimator for linear models under the fixed design. Note that if $\gamma^* = O(1)$ then the bound on the estimation error is of the order $O(\sqrt{\frac{\log p}{n}}) + O(\sqrt{\frac{\log(G_n)}{n}})$ and the weighted group Lasso estimator still remains consistent for the $\ell_{2,1}$-estimation error and for the square prediction error under the group stabil condition if the number of groups increases almost as fast as $e^{o(n)}$. The terms $\sqrt{\log p}$ and $\sqrt{\log G_n}$ are the price to pay for the unknown group sparsity of β^*. If the relax error ε_n is a big order of λ, it leads to the convergence rate ε_n for the estimation error $\sum_{g=1}^{G_n} w_g\|\hat{\beta}_n^g - \beta^{*g}\|_2$.

From Theorem 4.1, if all $d_g = 1$, it enables us to derive analogous results for un-weighted Lasso penalty in what follows.

Corollary 4.1 *Let* $\gamma^* := \|\beta^*\|_0$, $p_\tau := P(T_1 \geq \tau) > 0$, *and* $D^2(\sqrt{2})$ *be a universal constant in the proof. Assume that* $(H.1)$–$(H.4)$ *and condition* $GS(1, \varepsilon_n, k)$ *are fulfilled for* $\Sigma := \mathrm{E}[z(t)z^\tau(t)]$. *If* λ *is chosen such that*

$$\lambda \geq \sqrt{2}L\left(3 + 2Ae^{2LB}\right)\sqrt{\frac{\log(2p)}{n}} + \frac{\sqrt{2}(2L+1)A\mathrm{e}^{2LB}}{p_\tau}\sqrt{\frac{\log p}{n}}.$$

Then, with probability at least

$$1 - 2(2p)^{-A^2/2} - 2e^{-np_\tau^2/2} - \frac{D^2(\sqrt{2})A^2\log p}{4n}p^{2-A^2} - \frac{D^2(\sqrt{2})A^2\log p}{4n}p^{-A^2} \quad \left(A^2 > 2\right),$$

we have $\hat{\beta}_n - \beta^* \in S(1, H^*)$ *and*

$$\left\|\hat{\beta}_n - \beta^*\right\|_1 \leq \frac{8\gamma^*\lambda}{kc_l} + \frac{c_l\varepsilon_n}{2\lambda}, \qquad \mathrm{E}^*\left[z^{*\tau}(t)\left(\hat{\beta}_n - \beta^*\right)\right]^2 \leq \frac{32\gamma^*\lambda^2}{kc_1^2} + 2\varepsilon_n.$$

Corollary 4.1 presents an upper bound of the ℓ_1-estimation error, which is similar to the existing result in Theorem 3.2 in Huang et al. [13] for classical Lasso penalized Cox models. The advantages of Corollary 4.1 are that the restricted eigenvalue condition is not stochastic and Theorem 3.2 in Huang et al. [13] requires further analysis of the restricted eigenvalue condition to guarantee a high-probability event. Another significant difference is that oracle inequalities in Huang et al. [13] require that the sample size is larger than a given constant. Our oracle inequalities are valid for any finite n under the given high-probability event.

5 Proofs

5.1 Proofs of Theorem 4.1

The proof is based on the following three steps.

Step1: Check $\hat{\beta}_n - \beta^* \in S(1, H^*)$.

Using Proposition 4.1 and Proposition 4.2 to bound the empirical process on the event $\mathcal{A} \cap \mathcal{B}$ by (4.4), we have

$$\begin{aligned}
&(\mathbb{P}_n - \mathbb{P})\left(\ell\left(\beta^*; T, z, \Delta\right) - \ell(\hat{\beta}_n; T, z, \Delta)\right) \\
&\quad = (\mathbb{P}_n - \mathbb{P})\left(l\left(\beta^*; T, z, \Delta\right) - l(\hat{\beta}_n; T, z, \Delta)\right) - D_n\left(\hat{\beta}, \beta^*\right) \\
&\quad \leq (\lambda_a + \lambda_b) \sum_{g=1}^{G_n} w_g \left\|\hat{\beta}_n^g - \beta^{*g}\right\|_2 = \lambda \sum_{g=1}^{G_n} w_g \left\|\hat{\beta}_n^g - \beta^{*g}\right\|_2.
\end{aligned} \tag{5.1}$$

From (4.3), (5.1) implies

$$\begin{aligned}
&\mathbb{P}\left(l(\hat{\beta}_n; T, z, \Delta) - l\left(\beta^*; T, z, \Delta\right)\right) + \lambda \sum_{g=1}^{G_n} w_g \|\hat{\beta}_n\|_2 \\
&\quad \leq \lambda \sum_{g=1}^{G_n} w_g \left\|\hat{\beta}_n^g - \beta^{*g}\right\|_2 + \lambda \sum_{g=1}^{G_n} w_g \left\|\beta^{*g}\right\|_2.
\end{aligned} \tag{5.2}$$

By adding $\lambda \sum_{g=1}^{G_n} w_g \|\hat{\beta}_n^g - \beta^{*g}\|_2$ to both sides of inequality (5.2), on $\mathcal{A} \cap \mathcal{B}$, we can obtain that

$$\begin{aligned}
&\lambda \sum_{g=1}^{G_n} w_g \left\|\hat{\beta}_n^g - \beta^{*g}\right\|_2 + \mathbb{P}\left(l(\hat{\beta}_n; T, z, \Delta) - l\left(\beta^*; T, z, \Delta\right)\right) \\
&\quad \leq \lambda \sum_{g=1}^{G_n} w_g \left(\left\|\hat{\beta}_n^g - \beta^{*g}\right\|_2 + \left\|\beta^{*g}\right\|_2 - \left\|\hat{\beta}_n^g\right\|_2\right).
\end{aligned} \tag{5.3}$$

If $g \notin H^*$, then $\|\hat{\beta}_n^g - \beta^{*g}\|_2 + \|\beta^{*g}\|_2 - \|\hat{\beta}_n^g\| = 0$, and otherwise $\|\beta^{*g}\|_2 - \|\hat{\beta}_n^g\|_2 \le \|\hat{\beta}_n^g - \beta^{*g}\|_2$. So the last term in inequality (5.3) can be rewritten as

$$\lambda \sum_{g=1}^{G_n} w_g \|\hat{\beta}_n^g - \beta^{*g}\|_2 + \mathbb{P}\big(l(\hat{\beta}_n; T, z, \Delta) - l\big(\beta^*; T, z, \Delta\big)\big) \le 2\lambda \sum_{g \in H^*} w_g \|\hat{\beta}_n^g - \beta^{*g}\|_2. \quad (5.4)$$

By the definition of β^*, we have $\mathbb{P}(l(\hat{\beta}_n; T, z, \Delta) - l(\beta^*; T, z, \Delta)) > 0$ and therefore

$$\sum_{g \notin H^*} w_g \|\hat{\beta}_n^g - \beta^{*g}\|_2 \le \sum_{g \in H^*} w_g \|\hat{\beta}_n^g - \beta^{*g}\|_2,$$

i.e., $\hat{\beta}_n - \beta^* \in S(1, H^*)$.

Step2: Find a lower bound for $\mathbb{P}(l(\hat{\beta}_n; T, z, \Delta) - l(\beta^*; T, z, \Delta))$.

The next proposition provides the desired lower bound.

Proposition 5.1 *Under* (*H*.4), *conditioning on* $\Delta = 1$, *we have*

$$\mathbb{P}\big(l(\hat{\beta}_n; T, z, \Delta) - l\big(\beta^*; T, z, \Delta\big)\big) \ge \frac{c_l}{2} \mathrm{E}^*\big[z_i^{*\tau}(t)\big(\hat{\beta}_n - \beta^*\big)\big]^2 \quad (5.5)$$

with $c_l > 0$ *is a constant given in* (*H*.4).

Proof By the second order Taylor's expansion of the function $\beta \mapsto l(\beta; T, z, \Delta)$, let the corresponding mean value $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ be between β_j^* and β_j for each $j = 1, 2, \ldots, p$.

Let $z^\tau(t)\tilde{\beta}$ be the intermediate point between $z^\tau(t)\beta^*$ and $z^\tau(t)\hat{\beta}_n$ given by a second order Taylor's expansion of $l(\beta^*; T, z, \Delta)$. Then, conditioning on $\Delta = 1$, we have

$$\begin{aligned}
&\mathbb{P}\big(l(\hat{\beta}_n; T, z, \Delta) - l\big(\beta^*; T, z, \Delta\big)\big) \\
&\quad = \mathrm{E}^*\big[\mathrm{E}\big\{l(\beta; T, z, \Delta) - l\big(\beta^*; T, z, \Delta\big) | z_i^\tau(t)\big\}\big]\big|_{\beta=\hat{\beta}_n} \\
&\quad = \mathrm{E}^*\mathrm{E}\big[\big\{l\big(\beta; T, z^*, \Delta\big) - l\big(\beta^*; T, z^*, \Delta\big) | z_i^{*\tau}(t)\big\}\big]\big|_{\beta=\hat{\beta}_n} \\
&\quad = \mathrm{E}^*\mathrm{E}\left\{[(\beta - \beta^*)^\tau \dot{l}(\beta^*, z^*, \Delta) + \frac{1}{2}(\beta - \beta^*)^\tau \ddot{l}(\tilde{\beta}, z^*, \Delta)(\beta - \beta^*)\right\}\Big|_{\beta=\hat{\beta}_n}, \\
&\qquad \tilde{\beta} \in \mathcal{S}_M(\beta^*) \\
&\quad = \big\{(\beta - \beta^*)^\tau \mathrm{E}^*\mathrm{E}\big[\dot{l}(\beta^*, z^*, \Delta)\big]\big\}\big|_{\beta=\hat{\beta}_n} + \frac{1}{2}(\beta - \beta^*)^\tau \mathrm{E}^*\mathrm{E}\big\{\ddot{l}(\tilde{\beta}, z^*, \Delta)\big\}(\beta - \beta^*)\big|_{\beta=\hat{\beta}_n} \\
&\quad \big[\text{By (H.4)}\big] \\
&\quad \ge \frac{c_l}{2} \mathrm{E}^*\mathrm{E}\big\{\big[z^{*\tau}(t)\big(\hat{\beta}_n - \beta^*\big)\big]^2\big\} = \frac{c_l}{2} \mathrm{E}^*\big[z_i^{*\tau}(t)\big(\hat{\beta}_n - \beta^*\big)\big]^2, \quad (5.6)
\end{aligned}$$

where the second last equality is obtained by estimating the equation in (3.4). □

From Proposition 5.1 and (5.4), it deduced that

$$\lambda \sum_{g=1}^{G_n} w_g \|\hat{\beta}_n^g - \beta^{*g}\|_2 + \frac{c_l}{2} \mathrm{E}^*\big[z^{*\tau}(t)\big(\hat{\beta}_n - \beta^*\big)\big]^2 \le 2\lambda \sum_{g \in H^*} w_g \|\hat{\beta}_n^g - \beta^{*g}\|_2. \quad (5.7)$$

Step 3: Squeeze error bounds from group stabil condition

Let Σ be the $p \times p$ covariance matrix whose entries are $\mathrm{E}[z_j(t)z_k(t)] = \mathrm{E}^*[z_j(t)z_k(t)]$. We have

$$\mathrm{E}^*\big[z^{*\tau}(t)\big(\hat{\beta}_n - \beta^*\big)\big]^2 = \big(\hat{\beta}_n - \beta^*\big)^\tau \mathrm{E}^*\big[z(t)z^\tau(t)\big]\big(\hat{\beta}_n - \beta^*\big) = \big(\hat{\beta}_n - \beta^*\big)^\tau \Sigma\big(\hat{\beta}_n - \beta^*\big)$$

since we assume that $\Sigma := \mathrm{E}[z(t)z^\tau(t)]$ satisfies the group stabil condition $GS(1, \varepsilon_n, k, H^*)$ after $\hat{\beta}_n - \beta^* \in S(1, H^*)$ is verified. Multiplying $c_l/2$ in (4.14), we have

$$\frac{c_l}{2}\big(\hat{\beta}_n - \beta^*\big)^T \Sigma\big(\hat{\beta}_n - \beta^*\big) \geq \frac{kc_l}{2} \sum_{g \in H^*} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2^2 - \frac{c_l \varepsilon_n}{2}.$$

Then substitute the above inequality to (5.7), by using the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
&\lambda \sum_{g=1}^{G_n} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2 + \frac{kc_l}{2} \sum_{g \in H^*} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2^2 \\
&\quad \leq 2\lambda \sqrt{\sum_{g \in H^*} d_g} \sqrt{\sum_{g \in H^*} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2^2} + \frac{c_l \varepsilon_n}{2}.
\end{aligned}$$

Now the fact that $2xy \leq tx^2 + y^2/t$ for all $t > 0$ leads to the following inequality:

$$\begin{aligned}
&\lambda \sum_{g=1}^{G_n} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2 + \frac{kc_l}{2} \sum_{g \in H^*} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2^2 \\
&\quad \leq 4t\lambda^2 \gamma^* + \frac{1}{t} \sum_{g \in H^*} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2^2 + \frac{c_l \varepsilon_n}{2}.
\end{aligned} \tag{5.8}$$

Putting $t := \frac{2}{kc_l}$ in (5.8), we have the oracle inequality

$$\sum_{g=1}^{G_n} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2 \leq \frac{8\gamma^* \lambda}{kc_l} + \frac{c_l \varepsilon_n}{2\lambda}.$$

Finally, for the prediction oracle inequality, it is deduced from (5.7) that

$$\begin{aligned}
&\lambda \sum_{g=1}^{G_n} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2 + \frac{c_l}{2} \mathrm{E}^*\big[z^{*\tau}(t)\big(\hat{\beta}_n - \beta^*\big)\big]^2 \\
&\quad \leq 2\lambda \left(\sum_{g=1}^{G_n} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2 - \sum_{g \notin H^*} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2 \right).
\end{aligned} \tag{5.9}$$

Therefore,

$$\frac{c_l}{2} \mathrm{E}^*\big[z^{*\tau}(t)\big(\hat{\beta}_n - \beta^*\big)\big]^2 \leq 2\lambda \sum_{g=1}^{G_n} w_g \big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2.$$

Note that the term $\sum_{g\notin H^*} w_g\|\hat{\beta}_n^g - \beta^{*g}\|_2 = \sum_{g\notin H^*} w_g\|\hat{\beta}_n^g\|_2$ that we have discarded for the first inequality sign in the above expression is very small on the set $\{g : \beta^{*g} = 0\}$.

Then using oracle inequality for $\sum_{g=1}^{G_n} w_g\|\hat{\beta}_n^g - \beta^{*g}\|_2$ leads to

$$\frac{c_l}{2}\mathrm{E}^*\big[z^{*\tau}(t)\big(\hat{\beta}_n - \beta^*\big)\big]^2 \le 2\lambda\sum_{g=1}^{G_n} w_g\big\|\hat{\beta}_n^g - \beta^{*g}\big\|_2 = 2\lambda\left(\frac{8\gamma^*\lambda}{kc_l} + \frac{c_l\varepsilon_n}{2\lambda}\right) = \frac{16\gamma^*\lambda^2}{kc_l} + c_l\varepsilon_n.$$

Finally we conclude the proof by using Propositions 4.1 and 4.2. We show that the desired oracle inequalities hold with high probability under the event $\mathcal{A}\cap\mathcal{B}$.

5.2 Proofs of the propositions

5.2.1 Proof of Proposition 4.1

First we show that the summation is satisfied by applying Hoeffding's inequality, see Wainwright [26].

Lemma 5.1 (Hoeffding's inequality) *Let $X_1,\dots,X_n$ be independent random variables on $\mathbb{R}$ satisfying bound condition $a_i \le X_i \le b_i$ for $i = 1,2,\dots,n$. Then we have*

$$P\left(\left|\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\right| \ge t\right) \le 2\exp\left\{\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}.$$

For $\mathcal{A}_1 = \bigcap_{g=1}^{G_n}\{\|\frac{1}{n}\sum_{i=1}^{n}[\frac{z_{ig}^{\tau}(T_i)}{w_g}\Delta_i - \mathrm{E}(\frac{z_{ig}^{\tau}(T_i)}{w_g}\Delta_i)]\|_2 \le \lambda_{a1}\}$, let $W_i^g := \frac{z_{ig}^{\tau}(T_i)}{w_g}\Delta_i - \mathrm{E}(\frac{z_{ig}^{\tau}(T_i)}{w_g}\Delta_i)$ and

$$W_{ij}^g := \frac{z_{ij}^{\tau}(T_i)}{w_g}\Delta_i - \mathrm{E}\left(\frac{z_{ij}^{\tau}(T_i)}{w_g}\Delta_i\right), \quad j = 1,\dots,d_g; i = 1,\dots,n.$$

We have

$$P\big(\mathcal{A}_1^c\big) \le \sum_{g=1}^{G_n} P\left\{\left\|\frac{1}{n}\sum_{i=1}^{n} W_i^g\right\|_2^2 > \lambda_{a1}^2\right\} \le \sum_{g=1}^{G_n}\sum_{j=1}^{d_g} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} W_{ij}^g\right| > \frac{\lambda_{a1}}{\sqrt{d_g}}\right\} \tag{5.10}$$

due to $\{\|\frac{1}{n}\sum_{i=1}^{n} W_i^g\|_2^2 > \lambda_{a1}^2\} \subset \bigcup_{j\in\mathrm{Group}_g, |\mathrm{Group}_g|=d_g}\{|\frac{1}{n}\sum_{i=1}^{n} W_{ij}^g|^2 > \frac{\lambda_{a1}^2}{d_g}\}$.

Applying Hoeffding's inequality with $a_i = \frac{-L}{nw_{\min}} \le \frac{1}{n}\frac{z_{ij}^{\tau}(T_i)}{w_g}\Delta_i \le \frac{L}{nw_{\min}} = b_i$, we obtain

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} W_{ij}^g\right| > \frac{\lambda_{a1}}{\sqrt{d_g}}\right\} \le 2\exp\left(-\frac{nw_{\min}^2\lambda_{a1}^2}{2L^2 d_g}\right) \le 2\exp\left(-\frac{nw_{\min}^2\lambda_{a1}^2}{2L^2 d_{\max}}\right). \tag{5.11}$$

Finally, from (5.10) and (5.11), it is deduced that

$$P\big(\mathcal{A}_1^c\big) \le 2d_{\max}G_n\exp\left(-\frac{nw_{\min}^2\lambda_{a1}^2}{2L^2 d_{\max}}\right) =: d_{\max}(2G_n)^{1-A^2}, \tag{5.12}$$

which gives $\lambda_{a1} = \frac{L\sqrt{2d_{\max}}}{w_{\min}}\sqrt{\frac{\log(2G_n)}{n}}$.

For $\mathcal{A}_2'$, we resort to McDiarmid's concentration inequalities with bounded difference condition for random vectors, see Wainwright [26].

Lemma 5.2 *Suppose that $X_1,\dots,X_n$ are independent random vectors all taking values in the set A, and assume that $f : A^n \to \mathbb{R}$ is a function satisfying the* bounded difference condition

$$\sup_{x_1,\dots,x_n,x_k'\in A} \left|f(x_1,\dots,x_n) - f\left(x_1,\dots,x_{k-1},x_k',x_{k+1},\dots,x_n\right)\right| \le c_k.$$

Then, for all $t > 0$,

$$P\left[\left|f(X_1,\dots,X_n) - \mathrm{E}\left\{f(X_1,\dots,X_n)\right\}\right| \ge t\right] \le 2\exp\left(2t^2 \Big/ \sum_{i=1}^n c_i^2\right).$$

If there are no absolute signs in the above event, then the upper bound is changed by $\exp(2t^2/\sum_{i=1}^n c_i^2)$.

Similar to the treatment of $\mathcal{A}_1$, let

$$Z_i^g(\beta) := \frac{\int z_{ig}^\tau(s)1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{w_g\int 1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)} - \mathrm{E}\left(\frac{\int z_{ig}^\tau(s)1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{w_g\int 1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}\right)$$

and

$$Z_{ij}^g(\beta) := \frac{\int z_{ij}^\tau(s)1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{w_g\int 1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)} - \mathrm{E}\left(\frac{\int z_{ij}^\tau(s)1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{w_g\int 1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}\right),$$
$$j = 1,\dots,d_g; i = 1,\dots,n.$$

Then $\mathcal{A}_2 := \bigcap_{g=1}^{G_n}\{\|\frac{1}{n}\sum_{i=1}^n Z_i^g\|_2 \le \lambda_{a2}\}$. We have

$$\begin{aligned} P\left(\mathcal{A}_2^c\right) &\le \sum_{g=1}^{G_n} P\left\{\left\|\frac{1}{n}\sum_{i=1}^n Z_i^g(\beta)\right\|_2^2 > \lambda_{a1}^2\right\} \le \sum_{g=1}^{G_n}\sum_{j=1}^{d_g} P\left\{\left|\frac{1}{n}\sum_{i=1}^n Z_{ij}^g(\beta)\right| > \frac{\lambda_{a2}}{\sqrt{d_g}}\right\} \\ &\le \sum_{g=1}^{G_n}\sum_{j=1}^{d_g} P\left\{\sup_{\beta\in\mathcal{S}_M(\beta^*)}\left|\frac{1}{n}\sum_{i=1}^n Z_{ij}^g(\beta)\right| > \frac{\lambda_{a2}}{\sqrt{d_{\max}}}\right\}. \end{aligned} \tag{5.13}$$

Let

$$\begin{aligned} f(z_1,\dots,z_n) = \sup_{\beta\in S_M(\beta^*)}\Bigg|\frac{1}{n}\sum_{i=1}^n\Bigg\{&\frac{\int z_{ij}^\tau(T_i)1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{w_g\int 1(s \ge T_i)e^{z_i^\tau(s)\beta}\,dF(s,z)} \\ &- \mathrm{E}\left(\frac{\int z_{ij}^\tau(T_i)1(s \ge T_i)e^{z_i^\tau(s)\beta}\,dF(s,z)}{w_g\int 1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}\right)\Bigg\}\Bigg| \end{aligned}$$

and

$$\begin{aligned} &f(z_1,\dots,z_{k-1},\tilde{z}_k,z_{k+1},\dots,z_n) \\ &\quad= \sup_{\beta\in S_M(\beta^*)}\Bigg|\frac{1}{n}\sum_{i=1,i\ne k}^n\Bigg\{\frac{\int z_{ij}^\tau(s)1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{w_g\int 1(s \ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)} \end{aligned}$$

$$
\begin{aligned}
&- \mathrm{E}\left(\frac{\int z_{ij}^{\tau}(s) 1(s \geq T_i) e^{z_i^{\tau}(T_i)\beta}\, dF(s,z)}{w_g \int 1(s \geq T_i) e^{z_i^{\tau}(T_i)\beta}\, dF(s,z)} \right) \Bigg\} \\
&+ \frac{1}{n} \Bigg\{ \frac{\int z_{kj}^{\tau}(s) 1(s \geq T_k) e^{z_k^{\tau}(T_k)\beta}\, dF(s,z)}{w_g \int 1(s \geq T_k) e^{z_k^{\tau}(T_k)\beta}\, dF(s,z)} \\
&- \mathrm{E}\left(\frac{\int z_{kj}^{\tau}(s) 1(s \geq T_k) e^{z_k^{\tau}(T_k)\beta}\, dF(s,z)}{w_g \int 1(s \geq T_k) e^{z_k^{\tau}(T_k)\beta}\, dF(s,z)} \right) \Bigg\} \Bigg|. \qquad (5.14)
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{k-1}, \tilde{z}_k, z_{k+1}, \ldots, z_n) \qquad (5.15) \\
&\leq \sup_{\beta \in S_M(\beta^*)} \Bigg| \frac{1}{n} \Bigg\{ \frac{\int z_{kj}^{\tau}(s) 1(s \geq T_k) e^{z_k^{\tau}(T_k)\beta}\, dF(s,z)}{w_g \int 1(s \geq T_k) e^{z_k^{\tau}(T_k)\beta}\, dF(s,z)} \\
&- \mathrm{E}\left(\frac{\int z_{kj}^{\tau}(s) 1(s \geq T_k) e^{z_k^{\tau}(T_k)\beta}\, dF(s,z)}{w_g \int 1(s \geq T_k) e^{z_k^{\tau}(T_k)\beta}\, dF(s,z)} \right) \Bigg\} \\
&- \frac{1}{n} \Bigg\{ \frac{\int \tilde{z}_{kj}^{\tau}(s) 1(s \geq T_k) e^{\tilde{z}_k^{\tau}(T_k)\beta}\, dF(s,z)}{w_g \int 1(s \geq T_k) e^{\tilde{z}_k^{\tau}(T_k)\beta}\, dF(s,z)} \\
&- \mathrm{E}\left(\frac{\int \tilde{z}_{kj}^{\tau}(s) 1(s \geq T_k) e^{\tilde{z}_k^{\tau}(T_k)\beta}\, dF(s,z)}{w_g \int 1(s \geq T_k) e^{\tilde{z}_k^{\tau}(T_k)\beta}\, dF(s,z)} \right) \Bigg\} \Bigg|. \qquad (5.16)
\end{aligned}
$$

Note that, for $j = 1, \ldots, d_g$ and $i = 1, \ldots, n$, we have

$$
\begin{aligned}
-\frac{Le^{2LB}}{w_{\min}} &= -\frac{\int L 1(s \geq T_i) e^{LB}\, dF(s,z)}{w_{\min} \int 1(s \geq T_i) e^{-LB} F(s,z)} \leq \frac{\int z_{ij}^{\tau}(s) 1(s \geq T_i) e^{z_i^{\tau}(s)\beta}\, dF(s,z)}{w_g \int 1(s \geq T_i) e^{z_i^{\tau}(s)\beta}\, dF(s,z)} \\
&\leq \frac{\int L 1(s \geq T_i) e^{LB}\, dF(s,z)}{w_{\min} \int 1(s \geq T_i) e^{-LB}\, dF(s,z)} = \frac{Le^{2LB}}{w_{\min}}. \qquad (5.17)
\end{aligned}
$$

For fixed j, (5.15) gives

$$
\left| f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{k-1}, \tilde{z}_k, z_{k+1}, \ldots, x_n) \right| \leq \frac{4Le^{2LB}}{nw_{\min}}
$$

for all $z_1, \ldots, z_n, \tilde{z}_k$.

Lemma 5.2 implies

$$
P\left\{ \sup_{\beta \in S_M(\beta^*)} \left| \frac{1}{n} \sum_{i=1}^{n} Z_{ij}^{g}(\beta) \right| \geq \mathrm{E}\left(\sup_{\beta \in S_M(\beta^*)} \left| \frac{1}{n} \sum_{i=1}^{n} Z_{ij}^{g}(\beta) \right| \right) + t \right\} \leq \exp\left(-\frac{nt^2 w_{\min}^2}{8L^2 e^{4LB}} \right).
$$

It is sufficient to estimate the sharper upper bounds of $\mathrm{E}(\sup_{\beta \in S_M(\beta^*)} |\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^{g}(\beta)|)$ by the symmetrization theorem and the contraction theorem below, which can be found in van der Vaart and Wellner [25], Wainwright [26].

Lemma 5.3 (Symmetrization theorem) *Let $\varepsilon_1, \dots, \varepsilon_n$ be a Rademacher sequence with uniform distribution on $\{-1, 1\}$, independent of $X_1, \dots, X_n$ and $f \in \mathcal{F}$. Then we have*

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\left[f(X_i) - \mathrm{E}\left\{f(X_i)\right\}\right]\right|\right] \le 2\mathrm{E}\left[\mathrm{E}_\epsilon\left\{\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right\}\right],$$

where $\mathrm{E}[\cdot]$ *refers to the expectation w.r.t.* $X_1, \dots, X_n$ *and* $\mathrm{E}_\epsilon\{\cdot\}$ *w.r.t.* $\epsilon_1, \dots, \epsilon_n$.

Using the symmetrization theorem, we have

$$\begin{aligned}
&\mathrm{E}\left[\sup_{\beta\in S_M(\beta^*)}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\int z_{ij}(s)1(s\ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{w_g\int 1(s\ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}\right.\right.\right.\\
&\qquad\left.\left.\left. - \mathrm{E}\left(\frac{\int z_{ij}^\tau(s)1(s\ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{w_g\int 1(s\ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}\right)\right\}\right|\right]\\
&\le 2\mathrm{E}\left[\mathrm{E}_\epsilon\left\{\sup_{\beta\in S_M(\beta^*)}\left|\sum_{i=1}^{n}\frac{\epsilon_i\int z_{ij}(s)1(s\ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{nw_g\int 1(s\ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}\right|\right\}\right]\\
&\le \frac{2}{nw_{\min}}\mathrm{E}\left(\sup_{\beta\in S_M(\beta^*)}\left|\sum_{i=1}^{n}w_i(\beta)\epsilon_i\right|\right), \qquad (5.18)
\end{aligned}$$

where $w_i(\beta) := \frac{\int z_{ij}(s)1(s\ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}{\int 1(s\ge T_i)e^{z_i^\tau(T_i)\beta}\,dF(s,z)}$ for $i = 1, 2, \dots, n$.

For any $w_i(\beta)$, we can find a sequence of random vectors $\{a_i\}_{i=1}^n \in \mathbb{R}^p$ with $\|a_i\|_\infty = 1$ and vector $b \in \mathbb{R}^p$ with $\|b\|_1 \le L$ such that

$$-L \le w_i(\beta) = a_i^T b \le \|a_i\|_\infty \|b\|_1 = \|b\|_1 \le L.$$

Then we have

$$\begin{aligned}
&\frac{2}{nw_{\min}}\mathrm{E}\left(\sup_{\beta\in S_M(\beta^*)}\left|\sum_{i=1}^{n}w_i(\beta)\epsilon_i\right|\right)\\
&\le \frac{2}{nw_{\min}}\mathrm{E}\left(\sup_{\beta\in S_M(\beta^*)}\left|\sum_{i=1}^{n}\sum_{j=1}^{p}\epsilon_i a_{ij}b_j\right|\right)\\
&= \frac{2}{nw_{\min}}\mathrm{E}\left(\sup_{\beta\in S_M(\beta^*)}\left|\sum_{j=1}^{p}\left(\sum_{i=1}^{n}\epsilon_i a_{ij}\right)b_j\right|\right)\\
&\quad\left[\text{By Hölder's inequality}\right]\\
&\le \frac{2}{nw_{\min}}\mathrm{E}\left(\sup_{\|b\|_1\le L}\max_{1\le j\le p}\left|\sum_{i=1}^{n}\epsilon_i a_{ij}\right|\cdot\|b\|_1\right)\\
&\le \frac{2L}{nw_{\min}}\mathrm{E}\left(\max_{1\le j\le p}\left|\sum_{i=1}^{n}\epsilon_i a_{ij}\right|\right) = \frac{2L}{nw_{\min}}\mathrm{E}\left(\mathrm{E}_\epsilon\max_{1\le j\le p}\left|\sum_{i=1}^{n}\epsilon_i a_{ij}\right|\right).
\end{aligned}$$

Next, we are going to use the following maximal inequality for bounded variables; see [31] for more discussions.

Lemma 5.4 (Maximal inequality) *Let $X_1,\dots,X_n$ be independent random vectors that take values in a measurable space $\mathcal{X}$ and $f_1,\dots,f_n$ be real-valued functions in $\mathcal{X}$ which satisfy, for all $j=1,\dots,p$ and all $i=1,\dots,n$,*

$$\mathrm{E}f_j(X_i)=0,\qquad \bigl|f_j(X_i)\bigr|\le a_{ij}.$$

Then

$$\mathrm{E}\left(\max_{1\le j\le p}\left|\sum_{i=1}^{n}f_j(X_i)\right|\right)\le\sqrt{2\log(2p)}\max_{1\le j\le p}\sqrt{\sum_{i=1}^{n}a_{ij}^2}.$$

By Proposition 5.4, with $\mathrm{E}[z_i\epsilon_i a_{ij}]=0$ and $\epsilon_i a_{ij}\le\max_{1\le i\le n}\|a_i\|_\infty=1$, we get

$$\begin{aligned}\mathrm{E}\left(\sup_{\beta\in S_M(\beta^*)}\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^g(\beta)\right|\right)&\le\frac{2L}{nw_{\min}}\mathrm{E}\left(\max_{1\le j\le p}\left|\sum_{i=1}^{n}\epsilon_i a_{ij}\right|\right)\\&\le\frac{2L}{nw_{\min}}\sqrt{2\log 2p}\sqrt{n}=\frac{2L}{w_{\min}}\sqrt{\frac{2\log 2p}{n}}.\end{aligned}$$

Then

$$\begin{aligned}&P\left\{\sup_{\beta\in S_M(\beta^*)}\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^g(\beta)\right|\ge\frac{2L}{w_{\min}}\sqrt{\frac{2\log 2p}{n}}+t\right\}\\&\quad\le P\left\{\sup_{\beta\in S_M(\beta^*)}\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^g(\beta)\right|\ge\mathrm{E}\left(\sup_{\beta\in S_M(\beta^*)}\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^g(\beta)\right|\right)+t\right\}\le\exp\left(-\frac{nt^2w_{\min}^2}{8L^2e^{4LB}}\right).\end{aligned}$$

Therefore, (5.13) can be further bounded by letting $\frac{\lambda_{a2}}{\sqrt{d_{\max}}}=\frac{2L}{w_{\min}}\sqrt{\frac{2\log 2p}{n}}+t$

$$\begin{aligned}P\bigl(\mathcal{A}_2^c\bigr)&\le\sum_{g=1}^{G_n}\sum_{j=1}^{d_g}P\left\{\sup_{\beta\in\mathcal{S}_M(\beta^*)}\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^g(\beta)\right|>\frac{\lambda_{a2}}{\sqrt{d_g}}\right\}\\&\le 2d_{\max}G_n\exp\left(-\frac{nt^2w_{\min}^2}{8L^2e^{4LB}}\right).\end{aligned}\tag{5.19}$$

Let $2d_{\max}G_n\exp(-\frac{nt^2w_{\min}^2}{8L^2e^{4LB}})=d_{\max}(2G_n)^{1-A^2}$, which gives

$$t=\frac{2\sqrt{2}ALe^{2LB}}{w_{\min}}\sqrt{\frac{\log(2G_n)}{n}}.$$

Finally, we have

$$P\bigl(\mathcal{A}_2^c\bigr)\le d_{\max}(2G_n)^{1-A^2}\tag{5.20}$$

by letting $\lambda_{a2}=\frac{2L\sqrt{2d_{\max}}}{w_{\min}}(\sqrt{\frac{\log 2p}{n}}+Ae^{2LB}\sqrt{\frac{\log(2G_n)}{n}})$. Together with (5.12), it gives

$$P(\mathcal{A})=P(\mathcal{A}_1\cap\mathcal{A}_2)\ge P(\mathcal{A}_1)+P(\mathcal{A}_2)-1\ge 1-2d_{\max}(2G_n)^{1-A^2}.$$

Then (4.6) is obtained by using (4.5) conditioning on the event $\mathcal{A}_1\cap\mathcal{A}_2$.

5.2.2 Proof of Proposition 4.2

For the event $\mathcal{B}_0$, we need the exponential concentration inequality for the uniform convergence of empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le x\}, \quad x \in \mathbb{R}.$$

Lemma 5.5 (DKW inequality, Massart [19]) *For $x \in \mathbb{R}$, the DKW inequality bounds the probability that the random function $F_n(x)$ differs from $F(x)$ by more than a given constant $\varepsilon > 0$:*

$$P\Big(\sup_{x\in\mathbb{R}}\big|F_n(x) - F(x)\big| > \varepsilon\Big) \le 2e^{-2n\varepsilon^2}.$$

[8] proves the inequality with an unspecified multiplicative constant multiples of the exponent in the tail bounds. Massart [19] shows that the DKW inequality has the multiplying constant 2. Let $p_\tau := P(T_1 \ge \tau) = 2U\mathrm{e}^{LB}$, so $U = p_\tau \mathrm{e}^{-LB}/2$. We have

$$\begin{aligned}
P\big(\mathcal{B}_0^c\big) &= P\Big(\sup_{t_s\in[0,\tau],\beta\in\mathcal{S}_M(\beta^*)} B_{1n}(t_s,\beta) \le U\Big) \\
&\le P\left(\sup_{t_s\in[0,\tau],\beta\in\mathcal{S}_M(\beta^*)} \frac{1}{n}\sum_{j=1}^{n} 1(T_j \ge t_s)\mathrm{e}^{z_i^\tau(t_s)\beta} \le U\right) \\
&\le P\left(\frac{1}{n}\sum_{j=1}^{n} 1(T_j \ge \tau)\mathrm{e}^{-LB} \le U\right) \\
&= P\left(\frac{1}{n}\sum_{j=1}^{n} 1(T_j \ge \tau) - \mathrm{E}\big[1(T_1 \ge \tau)\big] \le U\mathrm{e}^{LB} - P(T_1 \ge \tau)\right) \\
&\le P\left(\left|\frac{1}{n}\sum_{j=1}^{n} 1(T_j \ge \tau) - P(T_1 \ge \tau)\right| \ge \frac{p_\tau}{2}\right) \\
&\le P\left(\sup_{\tau\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^{n} 1(T_j \ge \tau) - \mathrm{E}\big[1(T_1 \ge \tau)\big]\right| \ge \frac{p_\tau}{2}\right) \le 2e^{-np_\tau^2/2}.
\end{aligned} \tag{5.21}$$

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of a normed space of real functions $f : \mathcal{X} \to \mathbb{R}$ in some set $\mathcal{X}$. Define the $L_r(Q)$-norm by $\|f\|_{L_r(Q)} = (\int |f|^r\, dQ)^{1/r}$. For probability measures Q, we have $L_r(Q)$-spaces endowed by the $L_r(Q)$-norm. Given two functions $l(\cdot)$ and $u(\cdot)$, the bracket $[l,u]$ is the set of all functions $f \in \mathcal{F}$ with $l(x) \le f(x) \le u(x)$ for all $x \in \mathcal{X}$. An ε-bracket is a bracket $[l,u]$ with $\|l-u\|_{L_r(Q)} < \varepsilon$, see van der Vaart and Wellner [25]. The bracketing number $N_{[]}(\varepsilon,\mathcal{F},L_r(Q))$ is the minimum number of ε-brackets covered by $\mathcal{F}$, i.e.,

$$N_{[]}\big(\varepsilon,\mathcal{F},L_r(Q)\big) = \inf\left\{n : \exists l_1,u_1,\ldots,l_n,u_n \text{ s.t. } \bigcup_{i=1}^{n}[l_i,u_i] = \mathcal{F} \text{ and } \|l_n - u_n\|_{L_r(Q)} < \varepsilon\right\}.$$

For the event $\mathcal{B}_1$, let $B_i^g(\beta) := 1(T_i \ge t_s)\frac{z_{ig}(t_s)\mathrm{e}^{z_i^\tau(t_s)\beta}}{w_g} - \mathrm{E}[1(T \ge t_s)\frac{z_{ig}(T)\mathrm{e}^{z^\tau(T)\beta}}{w_g}]$ and

$$B_{ij}^g(\beta) := 1(T_i \ge t_s)\frac{z_{ig}(t_s)\mathrm{e}^{z_i^\tau(t_s)\beta}}{w_g} - \mathrm{E}\left[1(T \ge t_s)\frac{z_{ig}(T)\mathrm{e}^{z^\tau(T)\beta}}{w_g}\right], \quad j = 1,\ldots,d_g; i = 1,\ldots,n.$$

Similar to the analysis of $\mathcal{A}_1$ and $\mathcal{A}_2$, we have

$$\begin{aligned}
P\big(\mathcal{B}_1^c\big) &\le \sum_{g=1}^{G_n} P\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} B_i^g(\beta) \right\|_2^2 > \lambda_{b1}^2 U^2 \right\} \\
&\le \sum_{g=1}^{G_n}\sum_{j=1}^{d_g} P\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} 1(T_i \ge t_s) \frac{z_{ij}(t_s)\mathrm{e}^{z_i^\tau(t_s)\beta}}{w_g} \right.\right. \\
&\quad \left.\left. - \mathrm{E}\left[1(T \ge t_s) \frac{z_{ij}(t_s)\mathrm{e}^{z^\tau(t_s)\beta}}{w_g} \right] \right| > \frac{\lambda_{b1} U}{\sqrt{d_g}} \right\} \\
&\le \sum_{g=1}^{G_n}\sum_{j=1}^{d_g} P\left\{ \sup_{\substack{t_s\in[0,\tau],\\ \beta\in\mathcal{S}_M(\beta^*)}} \left| \frac{1}{n}\sum_{i=1}^{n} 1(T_i \ge t_s) \frac{z_{ij}(t_s)\mathrm{e}^{z_i^\tau(t_s)\beta}}{w_g} \right.\right. \\
&\quad \left.\left. - \mathrm{E}\left[1(T \ge t_s) \frac{z_{ij}(t_s)\mathrm{e}^{z^\tau(t_s)\beta}}{w_g} \right] \right| > \frac{\lambda_{b1} U}{\sqrt{d_{\max}}} \right\}. \qquad (5.22)
\end{aligned}$$

Then we will apply sub-Gaussian concentration for suprema of the empirical processes as the following event:

$$\begin{aligned}
\mathcal{B}_{1gj} = \left\{ \sup_{\substack{t_s\in[0,\tau],\\ \beta\in\mathcal{S}_M(\beta^*)}} \left| \frac{1}{n}\sum_{i=1}^{n} 1(T_i \ge t_s) \frac{z_{ij}(t_s)\mathrm{e}^{z_i^\tau(t_s)\beta}}{w_g} - \mathrm{E}\left[1(T \ge t_s) \frac{z_{ij}(T)\mathrm{e}^{z^\tau(T)\beta}}{w_g} \right] \right| \le \frac{\lambda_{b1} U}{\sqrt{d_{\max}}} \right\}, \\
j = 1, \dots, d_g; g = 1, \dots, G_n,
\end{aligned}$$

with bracketing numbers $\{N_{[]}(\varepsilon, \mathcal{B}_{1gj}, L_2(P))\}$ relative to $L_2(P)$-norm, see Theorem 2.14.9 of van der Vaart and Wellner [25].

Lemma 5.6 (Sharper bounds for suprema of empirical processes, Talagrand [22]) *Consider a probability space* (Ω, Σ, P) *and n i.i.d. random variables* $X_1, \dots, X_n$, *valued in* Ω, *of law P. Let* $\mathcal{F}$ *be a class of measurable functions* $f : \mathcal{X} \mapsto [0, 1]$ *that satisfy*

$$N_{[]}\big(\varepsilon, \mathcal{F}, L_2(P)\big) \le \left(\frac{K}{\varepsilon}\right)^V \quad \textit{for every } 0 < \varepsilon < K.$$

Then, for every $t > 0$,

$$P\left(\sqrt{n} \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathrm{E}f(X_i) \right| \ge t \right) \le \left(\frac{D(K)t}{\sqrt{V}} \right)^V e^{-2t^2}$$

for a constant $D(K)$ *that depends on K only.*

The explicit constant $D(K)$ can be found in Zhang [30], who studies the tail bounds for the supremums of the empirical process $\{n^{-1/2}\sum_{i=1}^{n}[f(X_i) - \mathrm{E}f(X_i)]\}$, where $\{X_i\}$ is a sequence of (non-i.i.d, unbounded) independent random vectors with values in a general measurable space $(\mathcal{X}, \mathcal{A})$, and f is a measurable real function on $(\mathcal{X}, \mathcal{A})$.

In what follows, we assume that $z(t)$ is non-random. For $\{\mathcal{B}_{1gj}\}$ in (5.22), we have the function classes

$$\mathcal{F}_{1gj} = \left\{ f_{t,\beta}(x,z) = 1(x \geq t) \frac{[z_{1j}(t)e^{z^\tau(t)\beta} + Le^{LB}]w_{\min}}{2Le^{LB}w_g} : t \in [0,\tau], \beta \in \mathbb{R}^p \right\},$$
$$j = 1,\ldots,d_g; g = 1,\ldots,G_n,$$

so $0 \leq f_{t,\beta}(x,z) \leq 1$.

In $\mathcal{B}_2$, we focus on the class of functions $0 \leq g_{t,\beta}(x,z) \leq 1$,

$$\mathcal{G}_2 = \left\{ g_{t,\beta}(x,z) = 1(x \geq t) e^{z^\tau(t)\beta - LB} : t \in [0,\tau], \beta \in \mathbb{R}^p \right\}.$$

Let $\lceil x \rceil$ be the smallest integer that is greater than or equal to x. For any $\epsilon \in (0,1)$, let t_s be the sth $\lceil 1/\varepsilon \rceil$ quantile of T_1, thus

$$P(T_1 \leq t_s) = i\varepsilon, \quad s = 1,\ldots,\lceil 1/\varepsilon \rceil - 1, t_0 = 0, t_{\lceil 1/\varepsilon \rceil} = \infty.$$

For $\mathcal{F}_{1gj}$ and $\mathcal{G}_2$, we consider two types of brackets of the forms

$$\begin{aligned}&\left[L^{\mathcal{F}}_{jg,k}(x,z), U^{\mathcal{F}}_{jg,k}(x,z)\right] \\ &:= \left[1(x \geq s_k)\frac{(z_j e^{z^\tau\beta} + Le^{LB})w_{\min}}{2Le^{LB}w_g}, 1(x \geq s_{k-1})\frac{(z_j e^{z^\tau\beta} + Le^{LB})w_{\min}}{2Le^{LB}w_g}\right], \\ &j = 1,\ldots,d_g; g = 1,\ldots,G_n; z = (z_1,\ldots,z_p)^\tau := (L,\ldots,L)^\tau \in \mathbb{R}^p\end{aligned}$$

and

$$\left[L^{\mathcal{G}}_k(x,z), U^{\mathcal{G}}_k(x,z)\right] := \left[1(x \geq s_k)\frac{e^{z^\tau\beta}}{e^{LB}}, 1(x \geq s_{k-1})\frac{e^{z^\tau\beta}}{e^{LB}}\right]$$

for a grid of points $-\infty = s_0 < s_1 < \cdots < s_{\lceil 1/\varepsilon \rceil} = \infty$ with the property $F(s_k) - F(s_{k-1}) < \varepsilon$ for all i.

Then, for given j and g, the bracket functions satisfy

$$U^{\mathcal{F}}_{jg,k}(x,z) \leq f_{s,\beta}(x,z) \leq L^{\mathcal{F}}_{jg,k}(x,z), \quad k = 0,1,2,\ldots$$
$$U^{\mathcal{G}}_k(x,z) \leq g_{s,\beta}(x,z) \leq L^{\mathcal{G}}_k(x,z), \quad k = 0,1,2,\ldots$$

provided $s_{k-1} < s \leq s_k$.

For $\{\mathcal{B}_{1gj}\}$, the $L_2(P)$-norm of $U^{\mathcal{F}}_{jg,k}(x,z) - L^{\mathcal{F}}_{jg,k}(x,z)$ is

$$\begin{aligned}&\left\| U^{\mathcal{F}}_{jg,k}(x,z) - L^{\mathcal{F}}_{jg,k}(x,z) \right\|_{L_2(P)} \\ &= \left\{ \mathrm{E}_T\left[U^{\mathcal{F}}_{jg,k}(T,z) - L^{\mathcal{F}}_{jg,k}(T,z) \right]^2 \right\}^{1/2} \\ &\leq \left\{ \mathrm{E}_T\left[\frac{(z_j e^{z^\tau\beta} + Le^{LB})w_{\min}}{2Le^{LB}w_g} \left\{ 1(T \geq s_k) - 1(T > s_{k-1}) \right\} \right]^2 \right\}^{1/2} \\ &\leq \left\{ P(s_{k-1} < T \leq s_{k-1}) \right\}^{1/2} = \left\{ F(s_k) - F(s_{k-1}) \right\}^{1/2} < \sqrt{\varepsilon}.\end{aligned}$$

For $\mathcal{B}_2$, the $L_2(P)$-norm for $U_k^{\mathcal{G}}(x,z) - L_k^{\mathcal{G}}(x,z)$ is

$$\begin{aligned}\left\|U_k^{\mathcal{G}}(x,z) - L_k^{\mathcal{G}}(x,z)\right\|_{L_2(P)} &= \left\{\mathrm{E}_T\left[U_k^{\mathcal{G}}(T,z) - L_k^{\mathcal{G}}(T,z)\right]^2\right\}^{1/2} \\ &\le \left\{\mathrm{E}_T\left[\frac{(z_j \mathrm{e}^{z^\tau\beta} + L\mathrm{e}^{LB})w_{\min}}{2L\mathrm{e}^{LB}w_g}\left\{1(T \ge s_k) - 1(T > s_{k-1})\right\}\right]^2\right\}^{1/2} \\ &\le \left\{P(s_{k-1} < T \le s_{k-1})\right\}^{1/2} = \left\{F(s_k) - F(s_{k-1})\right\}^{1/2} < \sqrt{\varepsilon}.\end{aligned}$$

In both cases, by the definition of bracketing number, we get

$$N_{[]}\left(\sqrt{\varepsilon}, \mathcal{F}, L_2(P)\right) \le \lceil 1/\varepsilon \rceil \le 2/\varepsilon.$$

Hence, $N_{[]}(\varepsilon, \mathcal{F}, L_2(P)) \le 2/\varepsilon^2$.

For the event $\mathcal{B}_1$ with relation (5.22), we get $K = \sqrt{2}$ and $V = 2$ in Lemma 5.6. Then, conditioning on the random design z, with Lemma 5.6 we define

$$\begin{aligned}P(\mathcal{B}_{1gj}) &= P\left\{\left|\frac{1}{n}\sum_{i=1}^n 1(T_i \ge t_s)\frac{z_{ij}(t_s)}{w_g}\mathrm{e}^{z_i^\tau(t_s)\beta} - \mathrm{E}_T\left[1(T \ge t_s)z_{ij}\frac{z_{ij}(t_s)}{w_g}\mathrm{e}^{z_i^\tau(t_s)\beta}\right]\right| \le \frac{2Le^{LB}t}{w_{\min}}\right\} \\ &= \mathrm{E}_z P\left\{\sup_{\substack{t_s\in[0,\tau],\\ \beta\in\mathcal{S}_M(\beta^*)}}\left|\frac{1}{n}\sum_{i=1}^n \frac{1(T_i \ge t_s)z_{ij}(t_s)[\mathrm{e}^{z_i^\tau(t_s)\beta} + Le^{LB}]w_{\min}}{2Le^{LB}w_g}\right.\right. \\ &\quad \left.\left. - \mathrm{E}_T\left[\frac{1(T \ge t_s)z_{ij}(T)[\mathrm{e}^{z_i^\tau(t_s)\beta} + Le^{LB}]w_{\min}}{2Le^{LB}w_g}\right]\right| \le t \,\middle|\, z\right\} \\ &\le \mathrm{E}_z\frac{D^2(\sqrt{2})t^2}{2}e^{-2nt^2} = \frac{D^2(\sqrt{2})t^2}{2}e^{-2nt^2}.\end{aligned}$$

Note that $U = p_\tau \mathrm{e}^{-LB}/2$, thus we put $\frac{2Le^{LB}t}{w_{\min}} = \frac{\lambda_{b1}U}{\sqrt{d_{\max}}} = \frac{\lambda_{b1}p_\tau \mathrm{e}^{-LB}}{2\sqrt{d_{\max}}}$ in (5.22), which implies

$$P\left(\mathcal{B}_1^c\right) \le \sum_{g=1}^{G_n}\sum_{j=1}^{d_g} P(\mathcal{B}_{1gj}) \le d_{\max}G_n\frac{D^2(\sqrt{2})t^2}{2}e^{-2nt^2}$$

with $t = \frac{\lambda_{b1}p_\tau \mathrm{e}^{-2LB}w_{\min}}{4L\sqrt{d_{\max}}}$.

Let $d_{\max}G_n\frac{D^2(\sqrt{2})t^2}{2}e^{-2nt^2} = d_{\max}G_n\frac{D^2(\sqrt{2})t^2}{2}(G_n)^{1-A^2}$, it gives $t = \frac{A}{\sqrt{2}}\sqrt{\frac{\log(G_n)}{n}}$. Then we have

$$P\left(\mathcal{B}_1^c\right) \le \frac{d_{\max}G_nD^2(\sqrt{2})A^2\log(G_n)}{4n}(G_n)^{1-A^2} \tag{5.23}$$

with the tuning parameter λ_{b1} determined by

$$\lambda_{b1} = \frac{4tL\sqrt{d_g}}{p_\tau \mathrm{e}^{-2LB}w_{\min}} = \frac{2\sqrt{2}LA\mathrm{e}^{2LB}\sqrt{d_g}}{p_\tau w_{\min}}\sqrt{\frac{\log(G_n)}{n}}.$$

For the event $\mathcal{B}_2$, we have $K = \sqrt{2}$ and $V = 2$ in Lemma 5.6. Define

$$P\left(\mathcal{B}_2^c\right) = \mathrm{E}_z P\left\{\sup_{\substack{t_s\in[0,\tau],\\ \beta\in\mathcal{S}_M(\beta^*)}}\left|\frac{1}{n}\sum_{i=1}^n 1(T_i \ge t_s)\mathrm{e}^{z_i^\tau(t_s)\beta} - \mathrm{E}_T\left[1(T \ge t_s)\mathrm{e}^{z^\tau(T)\beta}\right]\right| \le e^{LB}t \,\middle|\, z\right\}$$

$$= P\left\{ \sup_{\substack{t_s\in[0,\tau],\\ \beta\in\mathcal{S}_M(\beta^*)}} \left| \frac{1}{n}\sum_{i=1}^{n} 1(T_i \geq t_s)\mathrm{e}^{z_i^\tau(t_s)\beta - LB} - \mathrm{E}_T\left[1(T \geq t_s)\mathrm{e}^{z^\tau(T)\beta - LB}\right] \right| \leq t \,\middle|\, z \right\}$$

$$\leq \frac{D^2(\sqrt{2})t^2}{2} e^{-2nt^2}.$$

Note that $U = p_\tau \mathrm{e}^{-LB}/2$, thus we set $e^{LB}t = \lambda_{b1}U = \frac{\lambda_{b1}p_\tau \mathrm{e}^{-LB}}{2}$ in (4.11). It gives $P(\mathcal{B}_2^c) \leq \frac{D^2(\sqrt{2})t^2}{2} e^{-2nt^2}$ with $t = \frac{\lambda_{b2}p_\tau \mathrm{e}^{-2LB}}{2}$.

Assign $\frac{D^2(\sqrt{2})t^2}{2} e^{-2nt^2} = \frac{D^2(\sqrt{2})t^2}{2} p^{-A^2}$, it implies $t = \frac{A}{\sqrt{2}}\sqrt{\frac{\log p}{n}}$. Therefore, the tuning parameter λ_{b2} is determined by

$$\lambda_{b2} = \frac{\sqrt{2}A\mathrm{e}^{2LB}}{p_\tau}\sqrt{\frac{\log p}{n}}$$

such that

$$P\left(\mathcal{B}_2^c\right) \leq \frac{D^2(\sqrt{2})t^2}{2} e^{-2nt^2} = \frac{D^2(\sqrt{2})A^2 \log p}{4n} p^{-A^2}. \tag{5.24}$$

Finally, we obtain by combining (5.21), (5.23), and (5.24)

$$P(\mathcal{B}) \geq P(\mathcal{B}_0) + P(\mathcal{B}_1) + P(\mathcal{B}_2) - 2$$

$$\geq 1 - 2e^{-np_\tau^2/2} - \frac{d_{\max} G_n D^2(\sqrt{2})A^2 \log(G_n)}{4n} G_n^{1-A^2} - \frac{D^2(\sqrt{2})A^2 \log p}{4n} p^{-A^2}.$$

6 Conclusions and future study

In this paper, we focus on the survival analysis problem by proportional hazard regressions, which includes situations when both the number of covariates p and sample size n are increasing, and $p \gg n$. When $p > n$, the classical partial likelihood estimation is over-parameterized and requires Lasso or weighted group Lasso regularization estimation to obtain a stable and satisfactory fitting of proportional hazard regressions. Under the group stabil condition, the sharp oracle inequalities for weighted group Lasso regularized misspecified Cox models are derived. The upper bound of their $\ell_{2,1}$-estimation error is determined by the tuning parameter with the rate $O(\sqrt{\frac{\log p}{n}}) + O(\sqrt{\frac{\log(G_n)}{n}})$. The obtained nonasymptotic oracle inequalities imply that the penalized estimator is consistent when $\log p/n \to 0$ under mild conditions. The rate is rate-optimality in the minimax sense.

In the future study, the statistical inferences (confidence interval and testing for the coefficient, FDR control) are left for further studies.

Acknowledgements

Ting Yan (tingyanty@mail.ccnu.edu.cn) and Huiming Zhang (huimingzhang@um.edu.mo) are co-coresponding authors. The authors are listed in alphabetical order and they contributed equally to this work. We would like to thank the two reviewers for taking the time to read our paper and for providing excellent suggestions and comments. The first author would like to show sincere gratefulness to the advisor Prof. Jinzhu Jia for his guidance of high-dimensional statistics. The authors also thank Prof. Hui Zhao for helpful discussions.

Funding

Yan Ting is partially supported by the National Natural Science Foundation of China (No. 11771171), the Fundamental Research Funds for the Central Universities. Zhang Huiming is supported in part by the University of Macau under UM Macao Talent Programme (UMMTP-2020-01).

Availability of data and materials
This is a purely mathematical paper. Data analysis is not applicable.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The authors completed the paper and approved the final manuscript.

Author details
[1]School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing, China. [2]Department of Statistics, Central China Normal University, Wuhan, China. [3]Department of Mathematics, Faculty of Science and Technology, University of Macau, Macau, China. [4]Center for Statistical Science, Tsinghua University, Beijing, China.

Publisher's Note

Received: 15 May 2020 Accepted: 16 November 2020 Published online: 30 November 2020

References

1. Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N.: Statistical Models Based on Counting Processes. Springer, Berlin (1993)
2. Andersen, P.K., Gill, R.D.: Cox's regression model for counting processes: a large sample study. Ann. Stat. **10**(4), 1100–1120 (1982)
3. Bartlett, P.L., Mendelson, S., Neeman, J.: L1-regularized linear regression: persistence and oracle inequalities. Probab. Theory Relat. Fields **154**(1), 193–224 (2012)
4. Bickel, P.J., Ritov, Y.A., Tsybakov, A.B.: Simultaneous analysis of lasso and Dantzig selector. Ann. Stat. **37**, 1705–1732 (2009)
5. Blazere, M., Loubes, J.M., Gamboa, F.: Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. IEEE Trans. Inf. Theory **60**(4), 2303–2318 (2014)
6. Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc., Ser. B, Methodol. **34**, 187–220 (1972)
7. Cox, D.R.: Partial likelihood. Biometrika **62**, 269–276 (1975)
8. Dvoretzky, A., Kiefer, J., Wolfowitz, J.: Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Stat. **27**(3), 642–669 (1956)
9. Fan, J., Li, R.: Variable selection for Cox's proportional hazards model and frailty model. Ann. Stat. **30**, 74–99 (2002)
10. Greenshtein, E., Ritov, Y.A.: Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli **10**(6), 971–988 (2004)
11. Honda, T., Hardle, W.K.: Variable selection in Cox regression models with varying coefficients. J. Stat. Plan. Inference **148**, 67–81 (2014)
12. Huang, H., Gao, Y., Zhang, H., Li, B.: Weighted lasso estimates for sparse logistic regression: non-asymptotic properties with measurement error. Acta Math. Sci. (2021, in press). arXiv preprint, arXiv:2006.06136
13. Huang, J., Sun, T., Ying, Z., Yu, Y., Zhang, C.H.: Oracle inequalities for the lasso in the Cox model. Ann. Stat. **41**(3), 1142–1165 (2013)
14. Kanehisa, M., Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**(1), 27–30 (2000)
15. Knight, K., Fu, W.: Asymptotics for lasso-type estimators. Ann. Stat. **28**, 1356–1378 (2000)
16. Kong, S., Nan, B.: Non-asymptotic oracle inequalities for the high-dimensional Cox regression via lasso. Stat. Sin. **24**(1), 25–42 (2014)
17. Lemler, S.: Oracle inequalities for the lasso in the high-dimensional Aalen multiplicative intensity model. Ann. Inst. Henri Poincaré Probab. Stat. **52**(2), 981–1008 (2016)
18. Lounici, K., Pontil, M., Van De Geer, S., Tsybakov, A.B.: Oracle inequalities and optimal inference under group sparsity. Ann. Stat. **39**(4), 2164–2204 (2011)
19. Massart, P.: The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. Ann. Probab. **18**, 1269–1283 (1990)
20. Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Giltnane, J.M.: The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N. Engl. J. Med. **346**(25), 1937–1947 (2002)
21. Struthers, C.A., Kalbfleisch, J.D.: Misspecified proportional hazard models. Biometrika **73**(2), 363–369 (1986)
22. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. **22**, 28–76 (1994)
23. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Methodol. **58**, 267–288 (1996)
24. Tibshirani, R.: The lasso method for variable selection in the Cox model. Stat. Med. **16**(4), 385–395 (1997)
25. van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, Berlin (1996)
26. Wainwright, M.J.: High-Dimensional Statistics: A Non-asymptotic Viewpoint, vol. 48. Cambridge University Press, Cambridge (2019)
27. Wang, S., Nan, B., Zhu, N., Zhu, J.: Hierarchically penalized Cox regression with grouped variables. Biometrika **96**(2), 307–322 (2009)
28. Yan, J., Huang, J.: Model selection for Cox models with time-varying coefficients. Biometrics **68**(2), 419–428 (2012)
29. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc., Ser. B, Stat. Methodol. **68**(1), 49–67 (2006)
30. Zhang, D.X.: Tail bounds for the suprema of empirical processes over unbounded classes of functions. Acta Math. Sin. **22**, 339–345 (2006)
31. Zhang, H., Chen, S.X.: Concentration inequalities for statistical inference. arXiv preprint, arXiv:2011.02258
32. Zhang, H., Jia, J.: Elastic-net regularized high-dimensional negative binomial regression: consistency and weak signals detection. Stat. Sin. (2021). https://doi.org/10.5705/ss.202019.0315

33. Zhang, H., Wu, X.: Compound Poisson point processes, concentration and oracle inequalities. J. Inequal. Appl. **2019**(1), 312 (2019)
34. Zhang, H.H., Lu, W.: Adaptive lasso for Cox's proportional hazards model. Biometrika **94**(3), 691–703 (2007)
35. Zhao, H., Wu, Q., Li, G., Sun, J.: Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. J. Am. Stat. Assoc. **115**, 204–216 (2020)
36. Zhou, S., Zhou, J., Zhang, B.: High-dimensional generalized linear models incorporating graphical structure among predictors. Electron. J. Stat. **13**(2), 3161–3194 (2019)