Electronic Journal of Statistics
Vol. 9 (2015) 492–534
ISSN: 1935-7524
DOI: 10.1214/15-EJS1004

# Structured estimation for the nonparametric Cox model[*]

**Jelena Bradic**

*Department of Mathematics*
*University of California, San diego*
*e-mail:* jbradic@ucsd.edu

**and**

**Rui Song**

*Department of Statistics*
*North Carolina State University*
*e-mail:* rsong@ncsu.edu

**Abstract:** In this paper, we study theoretical properties of the non-parametric Cox proportional hazards model in a high dimensional non-asymptotic setting. We establish the finite sample oracle $l_2$ bounds for a general class of group penalties that allow possible hierarchical and overlapping structures. We approximate the log partial likelihood with a quadratic functional and use truncation arguments to reduce the error. Unlike the existing literature, we exemplify differences between bounded and possibly unbounded non-parametric covariate effects. In particular, we show that bounded effects can lead to prediction bounds similar to the simple linear models, whereas unbounded effects can lead to larger prediction bounds. In both situations we do not assume that the true parameter is necessarily sparse. Lastly, we present new theoretical results for hierarchical and smoothed estimation in the non-parametric Cox model. We provide two examples of the proposed general framework: a Cox model with interactions and an ANOVA type Cox model.

Received June 2014.

## Contents

[*] Jelena Bradic was supported in part by the NSF grant DMS-1205296.

## 1. Introduction

Prediction of an instantaneous rate of occurrence of events when covariates are high dimensional plays a critical role in contemporary genetics studies underlying the causes of many incurable diseases. The challenge of high-dimensionality and arrival of high throughput bioinformatics data give rise to the surge of interest in the statistical literature.

Ever since Cox's seminal work on proportional hazards models [9, 10] many significant steps have been taken toward analyzing and quantifying regression estimators with censored data – notably the work of [33, 1] and [35], among others. [2]'s seminal work established asymptotic properties of a class of estimators maximizing partial likelihood. It also introduced a martingale decomposition of the score vector of Cox's partial likelihood. Such martingale techniques were then further developed for a class of truncated regression models [18], additive risk models [22] and competing risks models [25]. Despite the substantial body of existing work on proportional hazards estimators, research on high dimensional proportional hazards estimators has mostly been limited to completely specified models and exactly sparse estimators [5, 15]. Several recent papers have shed new light on high dimensional, but not-necessarily, sparse estimators [14, 21, 17], by presenting a finite sample framework for the statistical analysis when $p \gg n$. The partial likelihoods studied in those papers are assumed to be finite sample versions of global convex and quadratic functions. When $p \gg n$, many instances of such likelihoods will only possess quadratic curvature over small, local and dimensionality independent regions. In this paper, we present new theory that does not rely on such an assumption and we exemplify differences between the two regimes, strongly quadratic curvature and local quadratic curvature. Broadly, we are interested in allowing model-misspecification and in allowing covariates whose dimension grows together with $p$ and $n$.

We consider the following nonparametric Cox model. Conditional on the p-dimensional covariate $\mathbf{x}$, the hazard function is modeled as

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp\{g(\mathbf{x})\}, \tag{1}$$

for a baseline hazard function $\lambda_0(t)$ and the relative risk function $g(\mathbf{x})$. In order to estimate $g$ when the dimensionality of the covariates $p$ is much greater than the number of samples $n$, it is commonly assumed that function $g(\mathbf{x})$ exhibits some form of sparsity. In the Cox model $g(\mathbf{x}) = \boldsymbol{\beta}^{*T}\mathbf{x}$ for an exactly sparse vector. However, the linear modeling assumption on the log hazard ratio does not admit a good interpretation when $g(\mathbf{x})$ is nonlinear in nature.

To handle this challenge, we divide the function $g(\mathbf{x})$ into an additive and non-additive component, where the additive component can be well approximated through a sparse structured additive function. We aim to solve the following problem

$$\min_{\boldsymbol{\beta}\in\mathcal{B}} \|f_{\boldsymbol{\beta}} - g\|^2, \tag{2}$$

where $\mathcal{B}$ is a convex set, $f_{\boldsymbol{\beta}}$ is a parametric approximation of $g$. To the best of our knowledge, previous literature on the proportional hazards regression all focused on bounded sets $\mathcal{B}$. On the contrary, we consider a convex set $\mathcal{B}$ that has diameter dependent on the dimensionality $p$, which grows exponentially with the sample size $n$. In practice, the relative risk function $g$ is unknown and it is impossible to perfectly solve (2). Our goal is therefore to recover an approximate solution to this problem. That is, we wish to construct an estimator $\hat{\boldsymbol{\beta}}$ such that

$$\|f_{\hat{\boldsymbol{\beta}}} - g\|^2 - K \min_{\boldsymbol{\beta}\in\mathcal{B}} \|f_{\boldsymbol{\beta}} - g\|^2 \tag{3}$$

with a constant $K \geq 1$ that does not depend on $p$ and is as small as possible. An inequality that provides an upper bound on the (random) quantity in the above display, in a certain probabilistic sense, is referred as an oracle inequality later on. Observe that this is not an additive model since we do not assume that the risk function $g$ is of the form $f_{\boldsymbol{\beta}}$ for some $\boldsymbol{\beta} \in \mathcal{B}$. Consequently, the bias term $\min_{\boldsymbol{\beta}\in\mathcal{B}} \|f_{\boldsymbol{\beta}} - g\|^2$ may not vanish and the goal is to approximate the structured combination with the smallest possible bias.

There is a vast literature on establishing oracle inequalities for (3). Optimal rates of estimation for Gaussian linear regression can be found in [36]. Approximate sparse models, can be handled for both the linear and generalized linear case [38]. Lack of unique sparse modeling of the relative risk in our setting, poses unique theoretical challenges. Both the gradient and the Hessian of the partial likelihood of model (1) are indexed by $\boldsymbol{\beta}$. As the true relative risk does not have parametric form, both the score and the Hessian lose its martingale structure. Such problems are similar in nature to model misspecification. When the baseline hazard is known, the covariates are fixed and bounded; [21] developed bounds using the additive Lasso penalty. Existing literature on the theoretical analysis of the proportional hazards model (1), does not address model misspecification that can be structured in nature, or that can allow random designs of size that depend on $p$. We approach this problem by introducing a very general class of penalty functions.

Existing literature that establishes oracle inequalities for (3), typically assumes that the loss function satisfies a quadratic margin behavior (see Assumption B of [38] or Condition 2 of [32]). Such an assumption is related to the strong convexity of the likelihood function. Using the same assumption, [17] showed oracle inequalities using Kullback-Leibler loss for the additive Cox model. In context of additive hazards models, [14] designed a loss function that satisfies quadratic margin and use fixed, bounded design to establish oracle inequalities. However, in proportional hazards models (1), when the diameter of the set $\mathcal{B}$

grows with $p$, such quadratic margin assumption can be easily violated. Moreover, new theoretical challenges arise when departing from the quadratic margin; the flat geometry of the loss function prevents the access to the informative oracle bounds. In the context of linear models, [31] developed restricted strong convexity arguments. For the nonparametric Cox model (1), we develop concentration arguments, to show that the geometry of the loss does play a significant role only in small, ellipsoidal neighborhoods of the sparse, non-degenerate models. Within such neighborhoods, we sandwich the loss function with two quadratic losses and analyze the two losses independently.

When $p \gg n$, it has been established [20, 7, 24] that vectors that maximize the likelihood over either a finite, sparse set or its convex hull can achieve oracle inequalities (3). Censored high dimensional data are often collected from clinical studies where genomic formations are highly complex with a large number of possible interactions. Despite its importance, the structured sparsity is rarely studied in the context of censored observations. [40] gave an interesting empirical study in the cases of $p \leq n$. This article focuses on censored data with structured (group or hierarchical) sparsity. In particular, our results easily extend to situations where group LASSO, hierarchical LASSO, group Ridge, Elastic Net and block $l_1/l_\infty$ penalty are employed.

The contributions of our paper are three fold. First, we establish two new oracle inequalities (OI) for the high-dimensional nonparametric Cox model (1) that explicitly bound the squared estimation error under a random design. Second, we develop techniques that allow deviations from the exact sparsity and that introduce model misspecification. Third, we show new bounds for hierarchical and smooth selection in the context of additive Cox models. In particular we discuss the complete CAP family as introduced in [43] and penalties based both on sparsity and smoothness constraints, the elastic net penalty [45], for example.

The rest of the paper is organized as follows. In Section 2, we define a new class of group penalty functions. Section 3 contains our theoretical results. New bounds on the distance between the least squares and the partial likelihood loss function are presented in Section 4. Section 5 is left for two examples of the Hierarchical Lasso and the Elastic Net penalties used in the Cox model with interactions and an ANOVA Cox model, respectively.

We use the following notation. A $pd$ dimensional vector $\mathbf{x}$ is represented as $\mathbf{x} = (\mathbf{x}_1^T, \dots, \mathbf{x}_p^T)^T$ with $\mathbf{x}_j = (x_{j1}, \dots, x_{jp})^T$. For a $d$ dimensional vector $\mathbf{x}$, norm $\|\mathbf{x}\|_{\gamma_j} = (\sum_{k=1}^d |x_k|^{\gamma_j})^{1/\gamma_j}$, with $\gamma_j \geq 1$. The Höelder conjugate of $\gamma_j$ is denoted by $\gamma_j^*$ and satisfies $1/\gamma_j + 1/\gamma_j^* = 1$. The Euclidean functional norm, $\|\cdot\|^2$, is defined as $\|f(\mathbf{X})\|^2 = \frac{1}{n}\sum_{i=1}^n f^2(\mathbf{X}_i)$. Throughout the paper, we use $\otimes$ to denote the outer product between vectors, that is $\mathbf{x}^{\otimes 2}$ denotes $\mathbf{x}\mathbf{x}^T$, for any vector $\mathbf{x}$.

## 2. Convex group selection

We begin by setting up the notation behind point process models. Let $T$ denote the event time, let $D$ denote the censoring time, and $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_p^T)^T$ denote the $pd$-dimensional covariate vector with $\mathbf{X}_j = (X_{j1}, \ldots, X_{jd})$. Define $Z = \min(T, D)$ and $\delta = \mathbb{1}\{T \le D\}$ as the observed event time and censoring indicator, respectively. We consider an i.i.d sample $\{(\mathbf{X}_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ from the population $(\mathbf{X}, Z, \delta)$, where $\mathbf{X}_i = (\mathbf{X}_{i1}^T, \ldots, \mathbf{X}_{ip}^T)^T$ and $\mathbf{X}_{ij} = (X_{ij,1}, \ldots, X_{ij,d})^T$. Let the event time, $T$, and the censoring time, $D$, be independent conditional on the covariates. We denote with $t_1 < \cdots < t_N$ the ordered failure times and with $\mathcal{R}_q = \{i \in \{1, \ldots, n\} : Z_i \ge t_q\}$ at risk set at each failure time $t_q$. We define counting processes $N_i(t) = 1\{Z_i \le t, \delta_i = 1\}$, $\bar{N}(t) = n^{-1}\sum_{i=1}^n N_i(t)$ and predictable processes $Y_i = 1\{Z_i \ge t\} \in [0, 1]$. It holds that $dN_i(t) = dM_i(t) + d\Lambda_i(t)$ with a martingale sequence $M_i$ and a compensator

$$d\Lambda_i(t) = \lambda_0(t)\exp\{g(\mathbf{X}_i)\}dY_i(t), \tag{4}$$

where $g(x)$ is the unknown function of interest. Moreover, we use $\Lambda_0(\tau) = \int_0^\tau \lambda_0(t)dt$ to denote the integrated baseline function.

To approximate $g(\mathbf{X})$, we define two collections of functions, the first includes univariate functions $\{f_1(x), \ldots, f_p(x)\}$ whereas the second includes a collection of candidate dictionary functions $\{\Psi_1(x), \ldots, \Psi_d(x)\}$. Examples of dictionary functions include wavelets, splines, step functions, frames etc. We aim to approximate $g(\mathbf{X})$ with a linear combination of univariate functions $f_j$, each of which we approximate with a linear combination of dictionary functions $\Psi_k(x)$. Specifically, we approximate $g(\mathbf{X})$ with

$$f_{\mathbf{b}}(\mathbf{X}_i) = \sum_{j=1}^p f_j(X_{ij}) = \sum_{j=1}^p \sum_{k=1}^d b_{jk}\Psi_k(X_{ij}) = \mathbf{b}^T\mathbf{\Psi}(\mathbf{X}_i),$$

where similarly as before we used $\mathbf{\Psi}(\mathbf{X}_i) = \left(\mathbf{\Psi}(X_{i1})^T, \ldots, \mathbf{\Psi}(X_{ip})^T\right)^T$ with $\mathbf{\Psi}(X_{ij}) = (\Psi_1(X_{ij}), \ldots, \Psi_d(X_{ij}))^T$. The candidate functions are known a priori with $|\Psi_k(x)| \le C < \infty$, but need not be orthogonal. Note that we do not make assumptions on the number of candidate functions $d$ or $p$ and we allow both to grow and be much larger than $n$.

Let $\tau$ denote the end of the study time, we define the empirical risk function $\mathcal{R}_n(\mathbf{b}) = -\mathcal{L}_n(\mathbf{b}, \tau)$ and $\mathcal{L}_n(\mathbf{b}, \tau)$ denotes the log partial likelihood associated to the additive component using the counting process notation:

$$\mathcal{L}_n(\mathbf{b}, \tau) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau f_{\mathbf{b}}(\mathbf{X}_i)dN_i(t) - \int_0^\tau \log \mathcal{S}_n^{(0)}(\mathbf{b}, t)d\bar{N}(t), \tag{5}$$

with

$$\mathcal{S}_n^{(l)}(\mathbf{b}, t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)\mathbf{\Psi}^{\otimes l}(\mathbf{X}_i)\exp\{f_{\mathbf{b}}(\mathbf{X}_i)\}, \qquad l = 0, 1, 2.$$

We denote population equivalents of $\mathcal{S}_n^{(l)}(\mathbf{b}, t)$ with $s^{(l)}(\mathbf{b}, t) = E_{Y,X}\mathcal{S}_n^{(l)}(\mathbf{b}, t)$. We also define $\mathcal{S}_n^{(0)}(g, t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)\exp\{g(\mathbf{X}_i)\}$ to denote the censored empirical average of the unknown hazard function. Later on we denote $\mathcal{L}_n(\mathbf{b}, \tau)$ as $\mathcal{L}_n(\mathbf{b})$ for simplicity.

We fix some vector $\boldsymbol{\beta}^* \in \mathcal{B}$ such that $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \dots, \boldsymbol{\beta}_p^*)^T$, $\boldsymbol{\beta}_j^* \in \mathbb{R}^d$ and are such that $\boldsymbol{\beta}_j^* \neq 0$, for $j \in \mathcal{M}_*$, $\|\boldsymbol{\beta}_j^*\|_{\gamma_j} = 0, j \in \mathcal{M}_*^c$. The set $\mathcal{M}_*$ is any subset of $\{1, \dots, p\}$ that has at most $s$ elements, i.e., $|\mathcal{M}_*| \leq s$. Such a vector $\boldsymbol{\beta}^*$ possesses structured or grouped sparsity. We choose vector $\boldsymbol{\beta}^*$ among all structured-sparse vectors, such that $f_{\boldsymbol{\beta}^*}$ is the closest, in the Euclidean distance, from the unknown function $g$, i.e. such that $\|f_{\boldsymbol{\beta}^*} - g\|^2 = \min_{\mathbf{b}} \|f_{\mathbf{b}} - g\|^2$. Notice that $\|f_{\boldsymbol{\beta}^*} - g\|^2 = 0$ if and only if $f_{\boldsymbol{\beta}^*} = g$, i.e. if the hazard risk has additive structure. In general the bias term $\min_{\boldsymbol{\beta} \in \mathcal{B}} \|f_{\boldsymbol{\beta}} - g\|^2$ does not vanish, and our goal is to imitate the structured vector $\boldsymbol{\beta}^*$.

The Cox regression model is a very flexible tool when analyzing the effect of several "risk" factors on time to event problems. Very frequently a "risk" factor may have several levels and can be expressed via a number of dummy variables [26]. The dummy variables corresponding to the same factor form a natural group that we would like to preserve at estimation. If a parameter of one of the factor levels enters a model, we would like to encourage other associated factors to enter the model. Additionally, it is of interest to determine the role of the interactions between various "risk" factors on the outcome of the patients with familiar event of interest [19]. Such studies require hierarchical models with multiple interaction terms. If a parameter of one of the interactions enters the model, then it does not force the main effects to be present in the model. Hierachical penalties have a more suitable format that forces main effects to be present in the model if the interactions are.

With this in mind, we consider a class of estimators $\widehat{\boldsymbol{\beta}}$ that solve the following penalized problem

$$\widehat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{pd}}{\arg\min} \left\{ -\mathcal{L}_n(\mathbf{b}) + \lambda_n P(\mathbf{b}) \right\}, \tag{6}$$

where we define the group penalty function (GPF) as

$$P(\mathbf{b}) = \sum_{j=1}^{p} d^{1/\gamma_j^*} \cdot \rho\left(\|\mathbf{b}_j\|_{\gamma_j}\right), \tag{7}$$

with a convex function $\rho$. We consider convex functions $\rho$ that are sparsity encouraging and that satisfy $\rho'(0+) > 0$. The scaling, $d^{1/\gamma_j^*}$, ensures that the penalty term and the number of parameters within each group are of the same order. The GPF includes a wide variety of grouping and hierarchical structures: $\rho$ determines how groups relate to one another, while $\{l_{\gamma_j}\}_{j=1}^p$ norms dictate the relationship of the coefficients among each group, $j$. For $\rho = l_1$ and any $\gamma_j$, the penalty function reduces to the CAP family of [43]; for $\rho = l_1$ and $\gamma_j = 2$, it becomes the group Lasso penalty of [42]; for $\rho = l_1$ and $\gamma_j = \infty$ it reduces to the block $l_1/l_\infty$ penalty of [30]. The problem can be reparametrized to include a variety of scaling factors in the penalty function. For example,

$$\rho\left(\|\mathbf{b}_j^T \mathbf{R}_j\|_{\gamma_j}\right),$$

with proper weights $\mathbf{R}_j$, or

$$\rho\left(\|\mathbf{R}_j\mathbf{b}_j\|_{\gamma_j} + \sqrt{\mathbf{b}_j^T\mathbf{M}_j\mathbf{b}_j}\right),$$

with smoothing matrix $\{\mathbf{M}_j\}_{kl} = \int \Psi_k^{''}(x_j)\Psi_l^{''}(x_j)dx_j$ [27]. In section 5, we discuss these cases in details.

The following property of the introduced GPF is important in establishing finite sample bounds. We leave the proof to the Appendix.

**Lemma 1.** *Let $\mathbf{v}$ denote a vector in $\mathbb{R}^{pd}$ decomposed as $(\mathbf{v}_1^T,\dots,\mathbf{v}_p^T)^T \in$, with $\mathbf{v}_{n,j} \in \mathbb{R}^d$. Let $\mathcal{E}_{n,j} = \{\|\mathbf{v}_j\|_{\gamma_j^*} \le \lambda_n d^{1/\gamma_j^*}\rho'(0+)\}$.*

*Then, if all the events $\mathcal{E}_{n,j}$ hold with $j = 1,\dots,p$, we have that the GPF family (7) with convex functions $\rho$ satisfies*

$$\boldsymbol{\beta}^{*T}\mathbf{v} = \min_{\mathbf{x}\in\mathbb{R}^{pd}} \left\{\lambda_n P(\mathbf{x}) - (\mathbf{x}-\boldsymbol{\beta}^*)^T\mathbf{v}\right\} \tag{8}$$

$$|\boldsymbol{\beta}^{*T}\mathbf{v}| = \min_{\mathbf{x}\in\mathbb{R}^{pd}} \left\{\lambda_n P(\mathbf{x}) - |(\mathbf{x}-\boldsymbol{\beta}^*)^T\mathbf{v}|\right\}.$$

□

## 3. Main results

In this section, we present the main results and establish the non-asymptotic oracle inequalities of $\widehat{\boldsymbol{\beta}}$ in terms of the $l_2$ prediction error. Our results differ from the previous literature in terms of the penalty function and the measure of prediction error. We present non-asymptotic prediction properties that allow the number of covariates to diverge with $n$ while allowing complicated group structures in the model. Most of existing theoretical derivations in literature are based on the assumption of bounded covariates. More precisely, we define a constant $M_p$ such that

$$M_p := \sup_{\boldsymbol{b}\in\mathcal{B}} \exp\{f_{\mathbf{b}}(\mathbf{X}_i)\}. \tag{9}$$

We note that $M_p$ is often bounded random quantity, especially in studies where the dimension of the covariates is considered as fixed, or the function $f_b$ is bounded. In high dimensional settings where $p \ge n$, $M_p$ could diverge with $p$ and $n$ and should be carefully considered. For example in cases where $\mathcal{B}$ is a $p$-dimensional ball of diameter $r$ and $\mathbf{X}_i$s are i.i.d. standard Gaussian, then $\log M_p = r^2\sqrt{\log p/n}$ and is unbounded for all $r$ such that $n^{1/4} = o(r)$. Moreover, most of finite sample studies rest on a fixed design setup, a condition rarely satisfied in large genomics studies in the presence of censoring. Most of this paper is dedicated to develop theory that allows $M_p$ in equation (9) to diverge with $p$ in a random design setting. We present two finite sample results, where the first is rested on assuming bounding $M_p$ (Theorem 1) whereas, the second isn't requiring such a condition (Theorem 2).

To present the results we define

$$\mathbf{E}_n(\mathbf{b},t) = S_n^{(1)}(\mathbf{b},t)/S_n^{(0)}(\mathbf{b},t), \mathbf{V}_n(\mathbf{b},t) = S_n^{(2)}(\mathbf{b},t)/S_n^{(0)}(\mathbf{b},t) - \Big(\mathbf{E}_n(\mathbf{b},t)\Big)^{\otimes 2}.$$

The gradient and the Hessian of the log partial likelihood are of the form:

$$\bigtriangledown \mathcal{L}_n(\mathbf{b}) = -n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \left(\mathbf{E}_n(\mathbf{b}, t) - \mathbf{\Psi}(\mathbf{X}_i)\right) dN_i(t),$$

$$-\bigtriangledown^2 \mathcal{L}_n(\mathbf{b}) = n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \mathbf{V}_n(\mathbf{b}, t) dN_i(t).$$

**Remark 1.** We note that $\bigtriangledown \mathcal{L}_n(\boldsymbol{\beta}^*)$ can be decomposed as

$$\bigtriangledown \mathcal{L}_n(\boldsymbol{\beta}^*) = \mathbf{h}_n(\boldsymbol{\beta}^*) - n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \left(\mathbf{E}_n(\mathbf{b}, t) - \mathbf{\Psi}(\mathbf{X}_i)\right) \lambda_0(t) \exp\{g(\mathbf{X}_i)\} dY_i(t),$$

where $\mathbf{h}_n(\boldsymbol{\beta}^*) = -n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \left(\mathbf{E}_n(\mathbf{b}, t) - \mathbf{\Psi}(\mathbf{X}_i)\right) dM_i(t)$. In the Cox models where there is a unique true parameter $\boldsymbol{\beta}^*$, the last term in the above display is zero. In our case, however, that term does not disappear as the compensator does not vanish.

The following condition replaces the conditions used in the asymptotic analysis of the estimation properties of the Cox model in [11], where $p < n$, and those presented in Condition 2 of [5], where $p > n$.

**Condition 1.** The random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent, identically distributed random variables such that

(i) the nonparametric function of interest satisfies $E \exp\{g(\mathbf{X}_i)\} < \infty$,

(ii) the process $Y(t)$ is left continuous with right hand limits and such that $D := P(Y(\tau) = 1) > 0$ and $\Lambda_0(\tau) < \infty$.

Before we state the main oracle inequality, we provide concentration of measure for the gradient of the log partial likelihood at the sparse vector $\boldsymbol{\beta}^*$. To that end, we need a preliminary result providing concentration of measure for the vector $\mathbf{E}_n(\boldsymbol{\beta}^*, t)$.This is summarized in the following Lemma.

**Lemma 2.** *If Condition 1 is satisfied, then there exists a constant $W > 0$ independent of $p, n$, and $d$, such that for every sequence of positive numbers $r_n$,*

$$P\left( \sup_{0 \le t \le \tau} \left\| \mathbf{E}_n(\boldsymbol{\beta}^*, t) - \frac{s^{(1)}(\boldsymbol{\beta}^*, t)}{s^{0}(\boldsymbol{\beta}^*, t)} \right\|_{\infty} \ge c r_n + \sqrt{\frac{\log 2d}{n u^2}} \right)$$
$$\le \frac{3}{8ed} W^2 e^{-n r_n^2 D^2 / u^2 e^{2m^* C}} + e^{-nD^2/2},$$

*for $\log u = \|\boldsymbol{\beta}^*\|_1$ and $c = 1 + 2\exp\{m^* C - \log D + C \log u\}$, with $m^*$ being the minimal signal strength defined as $m^* = \min\{\|\boldsymbol{\beta}_j^*\|_{\gamma_j} : j \in \mathcal{M}_*\}$. Recall that, functions $\Psi_k$ are such that $|\Psi_k(x)| \le C$, for a positive constant $C < \infty$.* □

**Remark 2.** For fixed dimension $p$, it is typically assumed in the literature (see [11]) that there exist population functions $s^{(l)}(\boldsymbol{\beta}, t), l = 0, 1, 2$ such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left| S_n^{(l)}(\boldsymbol{\beta}, t) - s^{(l)}(\boldsymbol{\beta}, t) \right| \xrightarrow{P} 0, \qquad l = 0, 1, 2, \tag{10}$$

when $n \to \infty$. For growing dimension $p$, [5], take on a relation of the previous condition and only assume the convergence to hold for the true parameter $\boldsymbol{\beta}^*$. Since we deal with diverging $p$, $s$, and aim to provide finite sample bounds, conditions of [5] are unsatisfactory in our case. Lemma 2 proves an exact rate of convergence without imposing any of the asymptotic conditions like (10).

The next result gives tail probabilities that will be used to control the approximation error. They both depend on the GPF and require nontrivial proofs. Our theoretical derivations are further complicated due to the lack of martingale structure in the score vector $\bigtriangledown \mathcal{L}_n(\boldsymbol{\beta}^*)$. We have the following result:

**Lemma 3.** *If Condition 1 is satisfied, then for a constant $M = 1/(\tau\lambda_0(\tau)\Lambda_0(\tau)C)$ and a sequence of positive numbers $\lambda_n$ and all $j = 1, \ldots, p$,*

$$P\left(\lambda_0(\tau)\left|\int_0^\tau S_n^{(0)}(g,t)dt\right| \|\boldsymbol{\Psi}(X_{ij})\|_{\gamma_j^*} \geq \lambda_n d^{1/\gamma_j^*}\rho'(0+)\right) \tag{11}$$
$$\leq e^{-\frac{n^2M^2\lambda_n^2\rho'(0+)^2}{2\theta^2+2M\sqrt{n}\lambda_n\rho'(0+)y/3}} + P(\max_{1\leq i\leq n}\exp\{g(\mathbf{X}_i)\} > y),$$

*for a truncation value $y$ such that $\theta^2 \geq \sum_{i=1}^n E\exp\{2g(\mathbf{X}_i)\}1\{\exp\{2g(\mathbf{X}_i)\} \leq y\}$. Moreover, there exists a constant $W > 0$ independent of $p, n,$ and $d$, such that for $C_{\lambda_n,n,p,d}$ defined as*

$$\min\left\{\frac{C\lambda_n\rho'(0+)}{2\lambda_0(\tau)}, \frac{C\lambda_0(\tau)D^2d^{2/\gamma_j^*}\log d}{u^4e^{2m^*C}}, \frac{D^2}{2n}, \frac{M^2\lambda_n^2\rho'(0+)^2}{2\theta^2+2M\sqrt{n}\lambda_n\rho'(0+)y/3}\right\}, \tag{12}$$

*we have*

$$P\left(\|\mathbf{h}_{n,j}(\boldsymbol{\beta}^*)\|_\infty \geq \lambda_n d^{1/\gamma_j^*}\rho'(0+)\right) \tag{13}$$
$$\leq 2pd\left(\max\left\{\left(3+\frac{3W^2}{8ed}\right)e^{-n^2C_{\lambda_n,n,p,d}}, e^{-n\frac{\lambda_n^2d^{2/\gamma_j^*}\rho'^2(0+)}{16c_1^2C^2u^2}}\right\}\right.$$
$$\left.+ P(\max_{1\leq i\leq n}\exp\{g(\mathbf{X}_i)\} > y)\right).$$

□

For clarity of exposition, the proof is relegated to the Appendix B.

**Remark 3.** The result of (11) shows that the penalty term absorbs the misspecification term that emerges due to lack of a unique model. Additionally, the result (12) controls the martingale part of the score vector $\bigtriangledown\mathcal{L}_n(\boldsymbol{\beta}^*)$. The proof is challenged by the lack of typical assumption (10) and $M_p < C_0$, with $M_p$ defined in equation (9), that is used to control the jump size and variation of the martingale. Instead, we develop finite sample tail bounds for both the jump sizes and predictable variation of the martingale $\mathbf{h}_n(\boldsymbol{\beta}^*)$.

Let $\mathbf{A}$ denote a $pd \times pd$ matrix. To establish the sparse oracle inequality, we define the **Restricted Eigenvalue constant** $\mathbf{RE}(\mu, s, \rho, \boldsymbol{\gamma}, \mathbf{A})$ as follows:

$$\mathbf{RE}(\mu, s, \rho, \boldsymbol{\gamma}, \mathbf{A}) \ := \min_{\mathbf{x} \in \mathbb{C}_{\mu,\rho}, \boldsymbol{x} \neq 0} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\sum_{j \in \mathcal{M}_*} \rho(\|\boldsymbol{x}_j\|_{\gamma_j})^2}, \tag{14}$$

where

$$\mathbb{C}_{\mu,\rho} = \left\{ \mathbf{b} \in \mathbb{R}^{pd} : P(\mathbf{b}_{\mathcal{M}_*^c}) \leq \mu P(\mathbf{b}_{\mathcal{M}_*}) \right\}, \tag{15}$$

for $\mathcal{M}_* = \{j \in \{1, \ldots, p\} : \|\boldsymbol{\beta}_j^*\|_{\gamma_j} \neq 0\}$, $|\{\mathcal{M}_*\}| \leq s$.

The set $\mathbb{C}_{\mu,\rho}$ consists of all vectors that have support similar to the sparse vector, $\boldsymbol{\beta}^*$. In particular, vectors with more than $s$ non-zero elements also belong to the cone $\mathbb{C}_{\mu,\rho}$. We only require that their components positioned outside of $\mathcal{M}_*$ are smaller in size than their components positioned inside $\mathcal{M}_*$. For example, if $\rho = l_1$ the set $\mathbb{C}_{\mu,\rho}$ is a cone formed by all vectors $\mathbf{b}$ satisfying $\|\mathbf{b}_{\mathcal{M}_*^c}\|_1 \leq \mu \|\mathbf{b}_{\mathcal{M}_*}\|_1$ as defined in [4]. For $\rho = l_1$ and $\gamma_j = 2$, $\mathbb{C}_{\mu,\rho}$ is the cone formed by all vectors $\mathbf{b}$ satisfying $\|\mathbf{b}_{\mathcal{M}_*^c}\|_2 \leq \mu \|\mathbf{b}_{\mathcal{M}_*}\|_2$ as defined in [24]. Its geometry changes with the penalty function, $\rho$, and the chosen $\gamma_j' s$.

Thus, we use the notation $\mathbf{RE}(\mu, s, \rho, \boldsymbol{\gamma}, \mathbf{A})$ to describe its dependence on the sparsity size, $s$, the choice $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$, the vector of norms used to describe the "smoothness" of each $f_j$ and the choice of the matrix $\mathbf{A}$. For the linear models, $\mathbf{A}$ takes the form of $\mathbf{X}^T\mathbf{X}$ for fixed designs [4] and $\boldsymbol{\Sigma}$ for random Gaussian designs with covariance matrix $\boldsymbol{\Sigma}$ [8]. For the parametric Cox model where $g(\mathbf{X}_i) = \boldsymbol{\beta}^{*T}\mathbf{X}_i$, [15] consider the case of $\mathbf{A} = - \bigtriangledown^2 \mathcal{L}_n(\boldsymbol{\beta}^*) = n^{-1} \sum_{i=1}^n \int_0^\tau \mathbf{V}_n(\boldsymbol{\beta}^*, t) dM_i(t)$ for both fixed and random designs. We postpone further discussion of this constant to the Appendix A.

Let us introduce two constants $0 \leq \upsilon_1 \leq 1$ and $0 \leq \upsilon_2 \leq 1$ satisfying

$$\begin{aligned} &\upsilon_1 \exp\{-2C\upsilon_1\} \leq 16\lambda_n^2 \rho'(0+) \frac{\bar{d}}{\zeta^2}, \\ &\upsilon_2 \exp\{-2C\upsilon_2\} - 4\lambda_n \frac{\bar{d}}{\zeta^2 \rho'^2(0+)} \sqrt{\upsilon_2} \leq 16\lambda_n^2 \frac{\bar{d}^{3/2}}{\zeta^3 \rho'^{3/2}(0+)}, \end{aligned} \tag{16}$$

where $\bar{d} = \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}$. These constants are used to bound the size of the GPF functions.

**Lemma 4.** *Let $\hat{\boldsymbol{\beta}}$ be defined as in (6) with penalty function GPF defined in (7). Let Condition 1 hold and let $\zeta = \zeta(s)$ be a positive constant. Then, with probability $1 - \delta$, for $\delta$ in (19) and all $\mathbf{b} \in \mathcal{B}$,*

$$2\lambda_n \sum_{j \in \mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\widehat{\boldsymbol{\beta}}_j - \mathbf{b}_j\|_{\gamma_j}) \leq 64\lambda_n^2 \frac{\bar{d}}{\zeta^2} \exp\{2C\upsilon_1\} + 32\lambda_n^2 \frac{\bar{d}}{\zeta^2} \exp\{2C\upsilon_2\},$$

*for $0 \leq \upsilon_1, \upsilon_2 \leq 1$ satisfying (16).* □

We also define a sequence of weight vectors $\boldsymbol{\omega}(\mathbf{b}) = (\omega_1(\mathbf{b}), \ldots, \omega_n(\mathbf{b}))^T$ as follows,

$$\omega_i(\mathbf{b}) = \sum_{q=1}^N \frac{\exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_i)\} 1\{i \in \mathcal{R}_q\}}{\sum_{l \in \mathcal{R}_q} \exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_l)\}}. \tag{17}$$

With these preparations we are ready to state the main result for the case of bounded covariate effects. Let $M_p$ be defined in (9) and let $\mathcal{E}_0$ be the event such that $M_p < C_0$, for a constant $C_0$ independent of $p$ or $n$.

**Theorem 1.** *Let $\hat{\boldsymbol{\beta}}$ be defined as in (6) and penalty function $P(\mathbf{b})$ defined in (7). Let $\zeta = \zeta(s)$ be a positive constant. Then, for any non-negative constant $A > 0$ and $\log u = \|\boldsymbol{\beta}^*\|_1$, with*

$$\lambda_n \geq \frac{8Aun^{1/4}\lambda_0(\tau)}{d\rho'(0+)}\sqrt{\frac{\log pd}{n}},$$

$$\|f_{\widehat{\boldsymbol{\beta}}} - g\|^2 \leq \min_{b\in\mathcal{B}}\left\{(1+\underline{\omega}^{-1})\|f_{\mathbf{b}} - g\|^2 + 32\lambda_n^2\frac{\bar{d}}{\zeta^2\underline{\omega}}\Big(2\exp\{2C\upsilon_1\}\exp\{2C\upsilon_2\}\Big)\right\}, \tag{18}$$

*with probability no less than $1-\delta-P(\mathcal{E}_0^c)$, $\delta > 0$, where*

$$\begin{aligned}\delta = 2pd\max&\left\{\left(3+\frac{3W^2}{8ed}\right)\exp\left\{-n^2C_{\lambda_n,n,p,d}\right\}, \exp\left\{-n\frac{\lambda_n^2 d^{2/\gamma_j^*}\rho'^2(0+)}{16c_1^2C^2u^2}\right\},\right. \qquad (19)\\ &\left.\exp\left\{-n^2\frac{M^2\lambda_n^2\rho'(0+)^2}{2\theta^2+2M\sqrt{n}\lambda_n\rho'(0+)y/3}\right\}\right\} + 4pdP\Big(\max_{1\leq i\leq n}\exp\{g(\mathbf{X}_i)\} > y\Big)\\ &+ P\Big(\mathbf{RE}(7,s,\rho,\boldsymbol{\gamma},-\bigtriangledown^2\mathcal{L}_n(\boldsymbol{\beta}^*)) \leq \zeta^2\Big),\end{aligned}$$

*for $\theta, y, M, u, m^*$ as in Lemma 3, and $\bar{d} = \sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}$, $0 \leq \upsilon_1 \leq 1$ and $0 \leq \upsilon_2 \leq 1$ satisfying (16) and*

$$\underline{\omega} = \min\left\{\omega_i(\boldsymbol{\beta}^* + c(\mathbf{b}-\boldsymbol{\beta}^*)) : \mathbf{b}{\in}\mathbb{C}_{7,\rho}, c\in(0,1), i\in 1,\ldots,n\right\}$$

*for $\boldsymbol{\omega}(\mathbf{b})$ in (17), $\mathbb{C}_{7,\rho}$ in (15).*

*Proof of Theorem 1.* This proof requires careful analysis of the possible model misspecification and uses results from Propositions 2 and 3 stated in Section 4. To that end, we define an empirical functional norm $\|\cdot\|_{n,\mathbf{b}^*}$ for functions $f_{\mathbf{b}} : R^{pd} \to R$, $\mathbf{b} \in R^{pd}$ with a fixed $c \in (0,1)$ and $\mathbf{b}^* = c\mathbf{b} + (1-c)\boldsymbol{\beta}^* \in R^{pd}$,

$$\begin{aligned}\|f_{\mathbf{b}}\|_{n,\mathbf{b}^*}^2 = &\frac{1}{n}\sum_{i=1}^n\int_0^\tau Y_i(t)\omega_i(\mathbf{b}^*,t)f_{\mathbf{b}}^2(\mathbf{X}_i)d\bar{N}(t) \qquad (20)\\ &-\Big[\frac{1}{n}\sum_{i=1}^n\int_0^\tau Y_i(t)\omega_i(\mathbf{b}^*,t)f_{\mathbf{b}}(\mathbf{X}_i)d\bar{N}(t)\Big]^2,\end{aligned}$$

for nonnegative weight process

$$\omega_i(\mathbf{b}^*,t) = \exp\{f_{\mathbf{b}^*}(\mathbf{X}_i)\}/S_n^{(0)}(\mathbf{b}^*,t), \tag{21}$$

and $\bar{M}(t) = n^{-1}\sum_{i=1}^{n} M_i(t)$. This norm is connected to the curvature of the partial likelihood and is further discussed in Section 4. From the Taylor expansion and some algebra, we have that the following representation holds for all $\mathbf{b}$:

$$\begin{aligned} -\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \mathcal{L}_n(\mathbf{b}) \quad &= \quad \frac{1}{2}\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}_{\widehat{\boldsymbol{\beta}}}} - \frac{1}{2}\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} \\ &+ \quad (\mathbf{b} - \widehat{\boldsymbol{\beta}})^T \mathbf{h}_n(\boldsymbol{\beta}^*) + \boldsymbol{\nu}_n(\hat{\boldsymbol{\beta}}, \mathbf{b}, g), \end{aligned}$$

for $\mathbf{b}_{\widehat{\boldsymbol{\beta}}} = c\widehat{\boldsymbol{\beta}} + (1-c)\boldsymbol{\beta}^*$ and $\mathbf{b}^* = \tilde{c}\mathbf{b} + (1-\tilde{c})\boldsymbol{\beta}^*$ with a particular choice of $c \in (0,1)$ and $\tilde{c} = \tilde{c}(\mathbf{b}) \in (0,1)$, $\mathbf{h}_n(\boldsymbol{\beta}^*)$ as in Lemma 3 and

$$\begin{aligned} &\boldsymbol{\nu}_n(\hat{\boldsymbol{\beta}}, \mathbf{b}, g) \qquad (22) \\ = \quad & -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} \lambda_0(t) Y_i(t) \exp\{g(\mathbf{X}_i)\} \left(\log \mathcal{S}_n^{(0)}(\hat{\boldsymbol{\beta}}, t) - \log \mathcal{S}_n^{(0)}(\mathbf{b}, t)\right) dt \\ + \quad & \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} \lambda_0(t) Y_i(t) \exp\{g(\mathbf{X}_i)\} \left(\hat{\boldsymbol{\beta}}^T \boldsymbol{\Psi}(\mathbf{X}_i) - \mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_i)\right) dt. \end{aligned}$$

where in the last expression we used the Doob Mayer decomposition $dN_i = dM_i + d\Lambda_i$ with $d\Lambda_i = \lambda_0(t)Y_i(t)\exp\{g(\mathbf{X}_i)\}dt$.

From the definition of the penalized estimator as the minimizer of penalized empirical risk in (6), we obtain $-\mathcal{L}_n(\widehat{\boldsymbol{\beta}}) + \lambda_n P(\widehat{\boldsymbol{\beta}}) \leq -\mathcal{L}_n(\mathbf{b}) + \lambda_n P(\mathbf{b})$, i.e.

$$\begin{aligned} \|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}_{\widehat{\boldsymbol{\beta}}}} \quad &\leq \quad \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} + 2(\mathbf{b} - \widehat{\boldsymbol{\beta}})^T \mathbf{h}_n(\boldsymbol{\beta}^*) + 2\boldsymbol{\nu}_n(\hat{\boldsymbol{\beta}}, \mathbf{b}, g) \\ &+ \quad 2\lambda_n(P(\mathbf{b}) - P(\widehat{\boldsymbol{\beta}})). \qquad (23) \end{aligned}$$

According to (22) we decompose $\boldsymbol{\nu}_n(\mathbf{b}, \hat{\boldsymbol{\beta}}, g)$ in two parts, one that can be tied up with the estimation error and another that can be tied up with the penalty term. To that end, we observe that

$$\begin{aligned} &\boldsymbol{\nu}_n(\hat{\boldsymbol{\beta}}, \mathbf{b}, g) \leq \lambda_0(\tau)\left|\int_0^{\tau} S_n^{(0)}(g,t)dt\right| \times \\ &\left(\sup_{t\in[0,\tau]}\left|\log \mathcal{S}_n^{(0)}(\hat{\boldsymbol{\beta}}, t) - \log \mathcal{S}_n^{(0)}(\mathbf{b}, t)\right| + (\hat{\boldsymbol{\beta}} - \mathbf{b})^T \max_{1\leq i\leq n} \boldsymbol{\Psi}(\mathbf{X}_i)\right). \end{aligned}$$

We denote $\mathcal{S}_n^{(0)}(g,t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp\{g(\mathbf{X}_i)\}$, equivalent of $\mathcal{S}_n^{(0)}(\mathbf{b}, t)$ at the true, unknown function $g(\mathbf{x})$ and with $\lambda_0(\tau)$ denoting the value of the baseline hazard function at the end of the study time $\tau$. Observe that $\log \mathcal{S}_n^{(0)}(\mathbf{b}, t)$ is a positively weighted log-sum-exp function for any value of $\mathbf{b}$, therefore it is Lipschitz continuous (with constant 1 with respect to the $l_\infty$ norm),

$$\sup_{t\in[0,\tau]}\left|\log \mathcal{S}_n^{(0)}(\hat{\boldsymbol{\beta}}, t) - \log \mathcal{S}_n^{(0)}(\mathbf{b}, t)\right| \leq \max_{1\leq i\leq n} |\hat{\boldsymbol{\beta}}^T \boldsymbol{\Psi}(\mathbf{X}_i) - \mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_i)|.$$

Furthermore, utilizing Condition 1 we obtain

$$\boldsymbol{\nu}_n(\hat{\boldsymbol{\beta}}, \mathbf{b}, g) \leq 2\lambda_0(\tau) \left| \int_0^\tau S_n^{(0)}(g,t)dt \right| \max_{1\leq i\leq n} \left| (\hat{\boldsymbol{\beta}} - \mathbf{b})^T \boldsymbol{\Psi}(\mathbf{X}_i) \right|.$$

Combining with the previous result, we get

$$\begin{aligned} \|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}_{\widehat{\boldsymbol{\beta}}}} &\leq \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} + 2(\mathbf{b} - \widehat{\boldsymbol{\beta}})^T \mathbf{h}_n(\boldsymbol{\beta}^*) \\ &+ 2\lambda_0(\tau) \left| \int_0^\tau S_n^{(0)}(g,t)dt \right| \max_{1\leq i\leq n} |(\hat{\boldsymbol{\beta}} - \mathbf{b})^T \boldsymbol{\Psi}(\mathbf{X}_i)| + 2\lambda_n (P(\mathbf{b}) - P(\widehat{\boldsymbol{\beta}})), \end{aligned} \tag{24}$$

for any $\mathbf{b}$ and $\mathbf{b}^*$, $\mathbf{b}_{\widehat{\boldsymbol{\beta}}}$ fixed and defined as before. To establish oracle inequality we need to tightly control the last three terms on the right hand side of the previous inequality. The first of those is a martingale score vector at the additive part, the second measures model misspecification, whereas the third quantifies the size of the penalty function. Model misspecification is controlled by the penalty term. To that end we use the result of Lemma 1.

Utilizing Lemma 1 with $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}} - \mathbf{b}$ and $\mathbf{v}_n = 2\mathbf{h}_n(\boldsymbol{\beta}^*)$, the following holds from the first equality in (8)

$$\boldsymbol{\beta}^{*T} \mathbf{h}_n(\boldsymbol{\beta}^*) \leq -(\boldsymbol{\Delta} - \boldsymbol{\beta}^*)^T \mathbf{h}_n(\boldsymbol{\beta}^*) + \lambda_n P(\boldsymbol{\Delta}),$$

that is

$$4\boldsymbol{\Delta}^T \mathbf{h}_n(\boldsymbol{\beta}^*) \leq \lambda_n P(\boldsymbol{\Delta}), \tag{25}$$

on the event $\mathcal{E}_n$ defined as

$$\mathcal{E}_n = \bigcap_{j=1}^p \left\{ 2\|\mathbf{h}_{n,j}(\boldsymbol{\beta}^*)\|_{\gamma_j^*} \leq \lambda_n d^{1/\gamma_j^*} \rho'(0+) \right\}. \tag{26}$$

Moreover, utilizing Lemma 1 again, but now with $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}} - \mathbf{b}$ and $\mathbf{v}_n = 4\gamma_n \boldsymbol{\Psi}(\mathbf{X}_i)$, the following holds from the second equality in (8)

$$|\boldsymbol{\beta}^{*T} 4\gamma_n \boldsymbol{\Psi}(\mathbf{X}_i)| \leq -|(\boldsymbol{\Delta} - \boldsymbol{\beta}^*)^T 4\gamma_n \boldsymbol{\Psi}(\mathbf{X}_i)| + \lambda_n P(\boldsymbol{\Delta}),$$

on the event $\mathcal{D}_{n,i}$ defined as

$$\mathcal{D}_{n,i} = \bigcap_{j=1}^p \left\{ 4\lambda_0(\tau) \left| \int_0^\tau S_n^{(0)}(g,t)dt \right| \|\boldsymbol{\Psi}(X_{ij})\|_{\gamma_j^*} \leq \lambda_n d^{1/\gamma_j^*} \rho'(0+) \right\}. \tag{27}$$

After rearranging the terms and noticing that $|\boldsymbol{\Delta}^T \boldsymbol{\Psi}(\mathbf{X}_i)| \leq |\boldsymbol{\beta}^{*T} \boldsymbol{\Psi}(\mathbf{X}_i)| + |(\boldsymbol{\Delta} - \boldsymbol{\beta}^*)^T \boldsymbol{\Psi}(\mathbf{X}_i)|$, we get

$$4\gamma_n |\boldsymbol{\Delta}^T \boldsymbol{\Psi}(\mathbf{X}_i)| \leq \lambda_n P(\boldsymbol{\Delta}), \tag{28}$$

and with it that $4\gamma_n \max_{1\leq i\leq n} |\boldsymbol{\Delta}^T \boldsymbol{\Psi}(\mathbf{X}_i)| \leq \lambda_n P(\boldsymbol{\Delta})$, on the event $\mathcal{D}_{n,i}$.

Recall that $\mathcal{E}_0$ is the event that $M_P < C_0$ for a random element $M_p$ defined in (9) and a constant $C_0$ independent of $p$. Therefore, we can combine (24) with (25) and (28), to obtain that for all $\mathbf{b}$, conditionally on the event

$$\mathcal{E}_0 \cap \mathcal{E}_n \cap \bigcap_{i=1}^{n} \mathcal{D}_{n,i},$$

the following inequality holds

$$\begin{aligned}\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}_{\widehat{\boldsymbol{\beta}}}} &\leq \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} \\ &\quad + 2\lambda_n \sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*} \left(\rho(\|\widehat{\boldsymbol{\beta}}_j - \mathbf{b}_j\|_{\gamma_j}) + \rho(\|\mathbf{b}\|_{\gamma_j}) - \rho(\|\widehat{\boldsymbol{\beta}}_j\|_{\gamma_j})\right),\end{aligned}$$

for all $\mathbf{b}_{\widehat{\boldsymbol{\beta}}} = c\widehat{\boldsymbol{\beta}} + (1-c)\boldsymbol{\beta}^*$ and $\mathbf{b}^* = \tilde{c}\mathbf{b} + (1-\tilde{c})\boldsymbol{\beta}^*$. Let us fix $\mathbf{b}_{\widehat{\boldsymbol{\beta}}}$ and $\mathbf{b}^*$ henceforth. From the triangular inequality for the GPF, we have $\rho(\|\mathbf{b}_j\|_{\gamma_j}) \leq \rho(\|\widehat{\boldsymbol{\beta}}_j - \mathbf{b}_j\|_{\gamma_j}) + \rho(\|\widehat{\boldsymbol{\beta}}_j\|_{\gamma_j})$ leading to

$$\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}_{\widehat{\boldsymbol{\beta}}}} \leq \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} + 2\lambda_n \sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*} \rho(\|\boldsymbol{\Delta}_j\|_{\gamma_j}). \tag{29}$$

Secondly, we control the penalty term in (29) in Lemma 4 whose proof is presented in Appendix D.

Utilizing further the bound between the norms $\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} \leq \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_n$ (proved in Proposition 3 in Section 4) in combination to (29), we obtain

$$\underline{\omega}\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2 \leq \|f_{\boldsymbol{\beta}^*} - f_{\mathbf{b}}\|^2 + 64\lambda_n^2 \frac{\bar{d}}{\zeta^2} \exp\{2C\upsilon_1\} + 32\lambda_n^2 \frac{\bar{d}}{\zeta^2} \exp\{2C\upsilon_2\},$$

with $\upsilon_1, \upsilon_2$ defined above in (16). Moreover, from the definition of the vector $\boldsymbol{\beta}^*$ and the triangular inequality we have

$$\|f_{\widehat{\boldsymbol{\beta}}} - g\|^2 \leq \|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2 + \min_{\mathbf{b}\in\mathcal{B}} \|f_{\mathbf{b}} - g\|^2,$$

which in combination to the previous inequality provides

$$\|f_{\widehat{\boldsymbol{\beta}}} - g\|^2 \leq (1 + \frac{1}{\underline{\omega}}) \min_{\mathbf{b}\in\mathcal{B}} \|f_{\mathbf{b}} - g\|^2 + 64\lambda_n^2 \frac{\bar{d}}{\zeta^2\underline{\omega}} \exp\{2C\upsilon_1\} + 32\lambda_n^2 \frac{\bar{d}}{\zeta^2\underline{\omega}} \exp\{2C\upsilon_2\}.$$

The theorem follows easily if we bound the probability of the event $\mathcal{E}_0 \cap \mathcal{E}_n \cap \bigcap_{i=1}^{n} \mathcal{D}_{n,i}$, which is given in Lemma 3. Hence, the proof is completed. □

**Remark 4.** A typical assumption in the literature of oracle inequalities of the type (18) requires profile likelihood to be bounded below with a quadratic function (see quadratic margin condition of Assumption B of [38]). Instead of assuming such a margin, we directly derive lower and upper quadratic processes that sandwich the partial likelihood process. In case of Gaussian linear models, these two quadratic processes equal to the typical $l_2$ loss function.

Theorem 1 establishes new finite sample oracle inequality with possible deviations from exact sparsity. The first term on the right hand side of (18) measures how far is the true function of interest $g(x)$ from the sparse additive approximation $f_{\boldsymbol{\beta}^*}$ and is only equal to zero if $g = f_{\boldsymbol{\beta}^*}$ almost surely. Typically, similar results appeared in problems with fixed design [17] or if one considers estimation errors related to the Kullback-Leibler divergence that are quadratic in nature [14, 21]. In contrast, our results hold for a log partial likelihood of non-quadratic type and a class of general random designs and general group penalties. The last two quantities of the RHS of (18) represent the convergence rate for the appropriate choices of $\lambda_n$.

We now comment on the size of the constant $\underline{\omega}^{-1}$ appearing in the bound (18). Each weight, $\omega_i(\mathbf{b})$, is a sum of the conditional probabilities that observation $i$ had an event at time $t_q$, given that at least one event occurred at time $t_q$.

**Proposition 1.** *Let $\eta > 0$ and $c_2 \in \mathbb{R}$ be constants such that for all $q = 1, \ldots, N$,*

$$\lambda_{\min}\begin{pmatrix} \sum_{l\in\mathcal{R}_q} \boldsymbol{\Psi}^{\otimes 2}(\mathbf{X}_l) + \eta\mathbb{I}_{pd} & \sum_{l\in\mathcal{R}_q} \boldsymbol{\Psi}(\mathbf{X}_l) \\ \sum_{l\in\mathcal{R}_q} \boldsymbol{\Psi}^T(\mathbf{X}_l) & c_2 - \eta b_n \end{pmatrix} = \delta^\star, \tag{30}$$

*where $\mathbb{I}_{pd}$ is a unit matrix. Then, for all $i \in \{1, \ldots, n\}$ and $b_n > 0$, the solution to the optimization problem*

$$\min_{\mathbf{b}\in\mathbb{R}^{pd}} \left\{\omega_i(\mathbf{b}) : \|\mathbf{b}\|_2^2 \le b_n\right\} \tag{31}$$

*is attained and the minimum $\omega_{\min}$ satisfies*

$$\omega_{\min}\delta^\star = \sum_{q=1}^N \min\left\{0, \lambda_{\min}\left(\boldsymbol{\Psi}(\mathbf{X}_i)^{\otimes 2}\right)1(i \in \mathcal{R}_q)\right\}.$$

The conditions of Proposition 1 are not restrictive and are easily verifiable for well posed problems. For $\kappa_i = \min\{\mathbf{v}\boldsymbol{\Psi}^{\otimes}(\mathbf{X}_i)\mathbf{v}^T, \|\mathbf{v}\|_2 \le 1, \mathbf{v} \in \mathbb{C}_{7,\rho}\}$ and by Cauchy's interlacing theorem of Hermitian matrices, for Propositon 1 to hold it suffices that the random covariates $\mathbf{X}_i$ satisfy $\min_{i\in\mathbb{R}_q} \kappa_i > 0$. In that case, we conclude that $\underline{\omega}$ satisfies

$$\delta^\star\underline{\omega} \ge \sum_{q=1}^N \min\{0, \min_{i\in\mathbb{R}_q} \kappa_i\}.$$

For $\kappa^q = \min\{\sum_{i\in\mathbb{R}_q} \mathbf{v}\boldsymbol{\Psi}^{\otimes}(\mathbf{X}_i)\mathbf{v}^T, \|\mathbf{v}\|_2 \le 1, \mathbf{v} \in \mathbb{C}_{7,\rho}\}$ we obtain an upper bound on the leading constant of Theorem 1,

$$1 + \varepsilon \le \left(\sum_{q=1}^N \frac{\kappa_i}{\kappa^q} 1(i \in \mathcal{R}_q)\right)^{-1}, \tag{32}$$

with the right-hand side bounded away from infinity almost surely.

**Remark 5.** Note that if $P(M_p < C_0) \to 1$, the proposed estimator achieves Gaussian-like oracle rates similar to those of penalized least squares methods used in Gaussian linear models (see further discussion in Section 5). In other words, the first term in (18) is negligible and the rate is driven by the last two terms.

In the next result, we take a novel approach and establish high-dimensional sparse oracle results for possible unbounded covariate effect, i.e. without requiring Condition 9.

**Remark 6.** The proof of the rest of the section consists of two parts. First, we localize our penalized estimator to a small elliptical neighborhood around $\boldsymbol{\beta}^*$. With an appropriate choice of the tuning parameter, $\lambda_n$, we show that the diameter of the convex neighborhood becomes independent of the dimensionality and it shrinks to zero asymptotically (see Lemma 5). Second, using such local neighborhood structure, we sandwich the partial likelihood process with a lower and upper quadratic processes as in Theorem 1. The proof is completed by the analysis of those lower and upper bounds.

The localization step is presented in the following Lemma 5.

**Lemma 5.** *For $\log p \le n$ and $s \le \log n$, let $\hat{\boldsymbol{\beta}}$ be defined as in (6) with penalty function $P(\mathbf{b})$ defined in (7) and $\boldsymbol{\beta}^*$ the true sparse parameter. Let Condition 1 hold and let $\zeta = \zeta(s)$ be a positive constant. If $\lambda_n$ satisfies*

$$\lambda_n \sum_{j \in \mathcal{M}_*} d^{1+2/\gamma_j^*} \ge c\zeta^4,$$

*then with probability $1 - \delta$, for $\delta$ in (19),*

$$\sum_{j=1}^{p} d^{1/\gamma_j^*} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_{\gamma_j} \le 16\sqrt{2} C e^{C+\upsilon_1} r_n, \tag{33}$$

*for $r_n = \frac{\lambda_n}{\zeta^2} \sum_{j \in \mathcal{M}_*} d^{2/\gamma_j^*}$ and $0 \le \upsilon_1 \le 1$ satisfying (16).* □

Next, we present the main result of this section.

**Theorem 2.** *For $\log(pd) \le n$ let $\hat{\boldsymbol{\beta}}$ be defined as in (6) and penalty function $P(\mathbf{b})$ defined in (7). Let Condition 1 hold and let $\zeta = \zeta(s)$ be a positive constant. Then, for non-negative constant $A > 0$ and $u$ defined in Theorem 1,*

$$\lambda_n \ge \frac{8Aun^{1/4}}{d\rho'(0+)} \sqrt{\frac{\log pd}{n}} \qquad \textit{and} \qquad \lambda_n \sum_{j \in \mathcal{M}_*} d^{1+2/\gamma_j^*} \ge c\zeta^4, \tag{34}$$

*with probability no less than $1 - \delta$, $\delta > 0$ and satisfying (19), there exists $\varepsilon > 1$ such that,*

$$\|f_{\widehat{\boldsymbol{\beta}}} - g\|^2 \le \min_{\mathbf{b} \in \mathcal{B}} \left\{ (1+\varepsilon)\|f_{\mathbf{b}} - g\|^2 + 32\lambda_n^2 \varepsilon \frac{\bar{d}}{\zeta^2} \Big( 2\exp\{2C\upsilon_1\} \exp\{2C\upsilon_2\} \Big) \right\}, \tag{35}$$

*with* $\bar{d} = \sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}$, $0 \le \upsilon_1 \le 1$ *and* $0 \le \upsilon_2 \le 1$ *satisfying* (16) *and* $\underline{\underline{\omega}} = \min_i \omega_i(\boldsymbol{\beta}^*)$ *with*

$$\varepsilon = \underline{\underline{\omega}}^{-1} \exp\left\{ Ce^C 26 \frac{\lambda_n}{\zeta^2} \sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*} \right\}.$$

*Proof.* We proceed by first restricting the parameter space to an elliptical neighborhood that is not expanding with dimensionality $p$. Then, we apply Lemma 4 (stated in the Proof of Theorem 1) and Proposition 3 (stated and proved in Section 4) to finalize the proof.

From Proposition 2, we have that

$$\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\widehat{\boldsymbol{\beta}}^*} \ge e^{-2a_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}} \|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\boldsymbol{\beta}^*}, \tag{36}$$

with $\widehat{\boldsymbol{\beta}}^* = c\widehat{\boldsymbol{\beta}} + (1-c)\boldsymbol{\beta}^*$, some $c \in (0,1)$ and $a_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*} = 2\max_{1\le i\le n} |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{\Psi}(\mathbf{X}_i)|$. The exponential term in the previous equation needs to be tightly controlled for $p \gg n$. The proof of the theorem is then finalized by finding nontrivial bounds for the empirical norms, $\|\cdot\|_{n,\cdot}$ as defined in (20), while allowing $p \gg n$. Let $p \ge n$ and $\log p \le n$. We establish that by bounding the appropriate norm of the error vector $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and obtaining the bound, which is log linear in dimensionality $p$. The result is summarized in Lemma 5, whose proof is given in Appendix D.

Consequently we have

$$\begin{aligned} a_{\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*} &\le 2\sum_{j=1}^{p} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_{\gamma_j} \max_{1\le i\le n} \left( \sum_{k=1}^{d} (\Psi_k(X_{ij}))^{\gamma_j^*} \right)^{1/\gamma_j^*} \\ &\le 32\sqrt{2}Ce^C \frac{\lambda_n}{\zeta^2} \sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}. \end{aligned}$$

Remember that $C \ge \max_{k,i,j} |\Psi_k(X_{ij})|$. Hence, we have successfully localized the error vector $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ in a convex neighborhood whose diameter is not increasing with the dimensionality $p$.

Utilizing further Lemma 4 and Proposition 3 with equation (29), we obtain

$$\begin{aligned} &\underline{\underline{\omega}}\|f_{\hat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|_n^2 \\ \le\ & e^{32\sqrt{2}Ce^C r_n} \left\{ \min_{\mathbf{b}\in\mathcal{B}} \|g - f_{\mathbf{b}}\|^2 + 64\lambda_n^2 \frac{\bar{d}}{\zeta^2} \exp\{2C\upsilon_1\} + 32\lambda_n^2 \frac{\bar{d}}{\zeta^2} \exp\{2C\upsilon_2\} \right\}, \end{aligned}$$

with $\upsilon_1, \upsilon_2$ defined above in Lemma 4. The proof is completed by applying the triangle inequality. □

Let us comment on the size of $\varepsilon$ appearing on the RHS in Theorem 2. If $s \le \log n$, with the help of Proposition 1, we can conclude that $\varepsilon \ge 0$ and

$$\varepsilon \le \exp\{32\sqrt{2}Ce^C r_n\},$$

where $r_n \to 0$ and is such that $r_n\zeta^2 = \lambda_n \sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}$ with the constant $C$ defined as the upper bound on the dictionary functions $\Psi$.

**Remark 7.** The difference in the rates of convergence between Theorems 1 and 2 reflects the dimensionality and geometry of the problem. In comparison to Theorem 1, Theorem 2 differs in the presence of the exponential term of the order of $e^{r_n}$. This additional term is coming from the complex likelihood structure for possibly unbounded covariate effects for which we establish that local Lipchitz constant is proportional to $e^{r_n}$.

Results of (34) and (35) imply dimensionality restrictions on the hazard regression problem (6). Results of Theorem 2, rely on the choice of the number of basis functions, $d$, and it requires $d^{-1}\log(pd) < \sqrt{n}e^{-\|\boldsymbol{\beta}^*\|_1}/s$. This on the other hand, implies certain bound on $s, p, n$ and $d$. In the case of $d = O(n^{1/2})$ we have

$$\frac{s\log p}{n}e^{\|\boldsymbol{\beta}^*\|_1} + \frac{s\log n}{2n}e^{\|\boldsymbol{\beta}^*\|_1} < 1,$$

which is more restrictive than the bound appearing in linear regression models with $s\log p/n < 1$[4].

We also note that previous results do not require exact sparsity to hold, that is, they do not assume $\boldsymbol{\beta}^*$ is the true underlying parameter. In summary, the results of Theorems 1 and 2 are quite general. They cover a wide range of penalty functions with a choice of $\gamma_j$'s and are applicable to Lasso, group Lasso, group ridge, CAP penalty, elastic net and many more. Two specific examples will be discussed in Section 5.

## 4. Sandwich bounds for the log partial likelihood

This section gathers some results that were crucial in obtaining Theorems 1 and 2. The novel ideas of the major results are to quantify the distance between the log partial likelihood $-\mathcal{L}_n(\mathbf{b})$ and the approximate quadratic expansion of the log partial likelihood.

Without loss of generality, $-\mathcal{L}_n(\mathbf{b})$ can be written as $\mathcal{R}_n(\mathbf{b}) = -\mathcal{L}_n(\mathbf{b}) + \mathcal{L}_n(\boldsymbol{\beta}^*) - \mathcal{L}_n(\boldsymbol{\beta}^*)$. By Taylor expansion around $\boldsymbol{\beta}^*$, we have that there exists a $c \in (0,1)$ and $\mathbf{b}^* = c\mathbf{b} + (1-c)\boldsymbol{\beta}^*$ such that

$$\mathcal{R}_n(\mathbf{b}) = -(\mathbf{b}-\boldsymbol{\beta}^*)^T\left\{\bigtriangledown\mathcal{L}_n(\boldsymbol{\beta}^*)\right\} - \frac{1}{2}(\mathbf{b}-\boldsymbol{\beta}^*)^T\left\{\bigtriangledown^2\mathcal{L}_n(\mathbf{b}^*)\right\}(\mathbf{b}-\boldsymbol{\beta}^*) - \mathcal{L}_n(\boldsymbol{\beta}^*).$$

Together with the previous Taylor expansion, the empirical risk function can be decomposed as follows. For every $\mathbf{b}$, there exists a $c \in (0,1)$ and $\mathbf{b}^* = c\mathbf{b} + (1-c)\boldsymbol{\beta}^*$ such that $\mathcal{R}_n(\mathbf{b})$ admits the following quadratic representation:

$$\mathcal{R}_n(\mathbf{b}) = -(\mathbf{b}-\boldsymbol{\beta}^*)^T\left\{\bigtriangledown\mathcal{L}_n(\boldsymbol{\beta}^*)\right\} + \frac{1}{2}\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} - \mathcal{L}_n(\boldsymbol{\beta}^*).$$

Because no two counting processes, $N_i(t)$ and $N_j(t)$, jump at the same time, the following holds:

$$\|f_{\mathbf{b}}\|^2_{n,\mathbf{b}^*} = \frac{1}{n}\sum_{i=1}^n \int_0^\tau Y_i(t)\omega_i(\mathbf{b}^*,t)(f_{\mathbf{b}}(\mathbf{X}_i) - \bar{f}^*_{\mathbf{b}}(t))^2 d\bar{N}(t), \tag{37}$$

where $\bar{f}^*_{\mathbf{b}}(t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)\omega_i(\mathbf{b}^*, t) f_{\mathbf{b}}(X_i)$ can be understood as a process of empirical weighted averages of $f_{\mathbf{b}}$. If Condition 1 is satisfied, then, there exists a $c \in (0,1)$ such that the introduced empirical norm is a proper norm. To be specific, the norm is nonnegative, $\|f_{\mathbf{b}}\|^2_{n,\mathbf{b}^*} = 0$ for every such $\mathbf{b}^*$ if and only if $\mathbf{b} = 0$. In addition, the norm satisfies the triangular inequality that $\|f_{\mathbf{b}_1} - f_{\mathbf{b}_2}\|_{n,\mathbf{b}^*} \le \|f_{\mathbf{b}_1}\|_{n,\mathbf{b}^*} + \|f_{\mathbf{b}_2}\|_{n,\mathbf{b}^*}$ for every $\mathbf{b}_1, \mathbf{b}_2$ and fixed $\mathbf{b}^*$.

Since the squared Euclidean norm $\|\cdot\|$ represents a natural benchmark, we seek to understand the lower and upper bounds of the $\|\cdot\|_{n,\cdot}$ norm using the $l_2$ empirical norm $\|\cdot\|$ in the next result.

**Proposition 2.** *Let $\underline{\underline{\omega}}$ be defined as in Theorem 2. For any vector $\mathbf{v}$ define*

$$a_{\mathbf{v}} = \max_{1\le i,q\le n} |\mathbf{v}^T[\Psi(\mathbf{X}_i) - \Psi(\mathbf{X}_q)]|.$$

*Then, the following sandwich bound holds almost surely for every vector $\mathbf{b}$ and corresponding vector $\mathbf{b}^* = c\mathbf{b} + (1-c)\boldsymbol{\beta}^*$,*

$$\underline{\underline{\omega}} e^{-2a_{\mathbf{b}-\boldsymbol{\beta}^*}} \ \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2 \le \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} \le e^{2a_{\mathbf{b}-\boldsymbol{\beta}^*}} \ \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2, \tag{38}$$

*uniformly for every $c \in (0,1)$.*

A similar result appeared independently in the recent work of [15] (see Lemma 4.3). Such a result shares similarities to self-concordant arguments of [3], but the last arguments do not cover cases of $p \gg n$.

**Proposition 3.** *Let $N$ represent the number of distinct events. Then, uniformly for every $\mathbf{b} \in [-b_n, b_n]$, with $b_n > 0$ satisfying the condition of Proposition 1, and $\mathbf{b}^* = c\mathbf{b} + (1-c)\boldsymbol{\beta}^*$, with $c \in (0,1)$, the following holds almost surely:*

$$n^{-1}\sum_{q=1}^N \frac{\min\left\{0, \min_{i\in\mathbb{R}_q} \lambda_{\min}\left(\mathbf{\Psi}(\mathbf{X}_i)\mathbf{\Psi}^T(\mathbf{X}_i)\right)\right\}}{\lambda_{\min}\left(\sum_{l\in\mathbb{R}_q} \mathbf{\Psi}(\mathbf{X}_l)\mathbf{\Psi}^T(\mathbf{X}_l)\right)} \|f_{\mathbf{b}}\|^2 \le \|f_{\mathbf{b}}\|^2_{n,\mathbf{b}^*} \le \|f_{\mathbf{b}}\|^2, \tag{39}$$

*Moreover, if $b_n$ is bounded and $\min_{1\le i\le n} \lambda_{\min}\left(\mathbf{\Psi}(\mathbf{X}_i)\mathbf{\Psi}^T(\mathbf{X}_i)\right) > 0$, then the left-hand bound in (6) is strictly positive almost surely.*

Propositions 1–3 are critical in establishing the main result in terms of nontrivial lower and upper bounds. We utilize Propositions 1 & 3 for low- and 2 for high-dimensional problems, respectively. With the help of all four results, we are able to obtain the main results in Section 3.

## 5. Examples

In this section, we present two examples of GPFs (7) (that allows hierarchical structures within and among groups) and establish their theoretical properties in the Cox model setup. To the best of our knowledge, similar results do not exist in the current literature. For simplicity in the presentation, the results of this section focus on the exact sparse models with $\boldsymbol{\beta}^*$ representing the unknown true parameter.

### 5.1. Hierarchical selection and CAP

Our results apply to a general class of additive models, where the groups in the additive Cox model may share some, but not necessarily all features across groups. For example, the effect of one gene can be shared by many different pathways, and thus studying hierarchical gene selection is of significant importance.

One model that has very specific hierarchical structure is a Cox model with pairwise interaction terms. Given the covariates, the hazard rate function $\lambda(t|\mathbf{X})$ of each patient is related to the covariates $\mathbf{X}_1, \dots, \mathbf{X}_p$ and their cross products $\mathbf{X}_j\mathbf{X}_k$ with the following model

$$\lambda_i(t|\mathbf{X}) = \lambda_0(t) \exp\left\{\sum_{j=1}^{p} \beta_j X_{ij} + \sum_{j\neq k} \Theta_{jk} X_{ij} X_{ik}\right\},$$

where $\boldsymbol{\Theta} = \boldsymbol{\Theta}^T \in \mathbb{R}^{p\times p}$. In the display above, we refer to the additive part as the main effect terms and the quadratic part as the "interaction" terms. In this model, the total number of parameters is $1/2(p^2 + 3p)$ and each $f_j$ takes on a bilinear form $\beta_j X_{ij} + \sum_{k=1}^{p} \Theta_{jk} X_{ij} X_{ik}$ with $\mathbf{b}_j = (\beta_j, \boldsymbol{\Theta}_j^T)$ and $\boldsymbol{\Psi}(X_{ij}) = (X_{ij}, X_{ij}\mathbf{X}_i)$.

More generally, going back to the notation of previous chapters, let all covariates be decomposed into $G$ possibly overlapping groups $\{\Gamma_j\}_{j=1}^{G}$ in such a way that $\bigcup_j \Gamma_j = \{1, \dots, p\}$ and $\Gamma_j \cap \Gamma_k \neq \emptyset$, for $j \neq k$. Then, each $f_j$ can be approximated by $\mathbf{b}_{\Gamma_j}^T \boldsymbol{\Psi}_{\Gamma_j}$, where $\Gamma_j$ is a set of covariates that belongs to group $j$. In previous example the set $\Gamma_j$ was $\{j, 1, \dots, p\}$.

The regularized estimator, $\widehat{\boldsymbol{\beta}}$, is then defined as the minimizer of penalized partial likelihood $\mathcal{PL}_n(\mathbf{b}) + P(\mathbf{b})$, where

$$\mathcal{PL}_n(\mathbf{b}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left[\sum_{j=1}^{G} \mathbf{b}_{\Gamma_j}^T \boldsymbol{\Psi}_{\Gamma_j} + \log\left(\frac{1}{n}\sum_{i=1}^{n} Y_i(t) \exp\left\{\sum_{j=1}^{G} \mathbf{b}_{\Gamma_j}^T \boldsymbol{\Psi}_{\Gamma_j}\right\}\right)\right] dN_i(t)$$

with the penalty function $P(\mathbf{b})$ defined as

$$P(\mathbf{b}) = \sum_{j=1}^{G} \lambda_{n,j} |\Gamma_j|^{1/\gamma_j^*} \|\mathbf{b}_{\Gamma_j}\|_{\gamma_j},$$

where $|\Gamma_j|$ denotes for the cardinality of that set. Note that this penalty includes the classical group Lasso penalty, where one would select all $\gamma_j = 2$.

**Corollary 1.** *Let conditions of Theorem 2 be satisfied. Then, for some constant $A > 4$ and the choice of the tuning parameters*

$$\sqrt{d}\lambda_{n,j} \geq A \min\left\{\zeta^2, \sqrt{\frac{\log G}{n}} |\Gamma_j|^{-2/\gamma_j^*}\right\},$$

*we have*

$$
\begin{aligned}
&P\left(\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2 \le \underline{\underline{\omega}}^{-1} e^{C(r_n e^C + 2\upsilon_1)} 26 \frac{\sum_{j\in\mathcal{M}_*} \lambda_{n,j}^2 |\Gamma_j|^{2/\gamma_j^*}}{\zeta^2}\right) \\
&\ge 1 - 6\{Gd\}^{1-A} - P\left(\mathbf{RE}(7, s, \rho, \boldsymbol{\gamma}, -\bigtriangledown^2 \mathcal{P}\mathcal{L}_n(\boldsymbol{\beta}^*)) \le \zeta^2\right) \qquad (40) \\
&- 4pdP\left(\max_{1\le i\le n} \exp\{g(\mathbf{X}_i)\} > y\right),
\end{aligned}
$$

*for* $r_n = \zeta^{-2} \sum_{j\in\mathcal{M}_*} \lambda_{n,j} |\Gamma_j|^{2/\gamma_j^*}$*, and* $0 \le \upsilon_1 \le 1$ *satisfying*

$$\upsilon_1 e^{-2C\upsilon_1} \le 16\rho'(0+) \frac{\sum_{j\in\mathcal{M}_*} \lambda_{n,j}^2 |\Gamma_j|^{2/\gamma_j^*}}{\zeta^2}. \qquad (41)$$

The proof of this result is omitted because it is a simple modification of that of Theorem 2 with $\lambda_n$ being adaptive to each group $\Gamma_j$. The oracle inequality of Corollary 1 discusses finite-sample properties of the whole CAP family proposed in the seminal work of [43]. In particular, the block $l_1/l_\infty$ penalty introduced in [30] is a member of the CAP family. [30] present $l_\infty$ bounds on the estimation error of block $l_1/l_\infty$ penalty in the linear models. Corollary 1 provides its finite sample $l_2$ error bounds for the sparse additive Cox model with possibly overlapping groups.

In more details, we obtain with high probability

$$\|f_{\widehat{\boldsymbol{\beta}}_{l_1/l_\infty}} - f_{\boldsymbol{\beta}^*}\|^2 \le 26 e^{C(r_n e^C + 2\upsilon_1)} \frac{\sum_{j\in\mathcal{M}_*} \lambda_{n,j}^2 |\Gamma_j|^2}{\zeta^2},$$

for the block $l_1/l_\infty$ penalty, $0 \le \upsilon_1 \le 1$ satisfying

$$\upsilon_1 e^{-2C\upsilon_1} \le 16\rho'(0+) \sum_{j\in\mathcal{M}_*} \lambda_{n,j}^2 |\Gamma_j|^2/\zeta^2,$$

and $r_n = \zeta^{-2} \sum_{j\in\mathcal{M}_*} \lambda_{n,j} |\Gamma_j|^2$. In case of a parametric model with $g(\mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\beta}^*$, or the interaction model discussed in the example above, we observe that Gaussian random designs with $\boldsymbol{\Sigma}$ such that $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$ will make the last two terms in (40) negligible.

Moreover, non-overlapping groups gained significant attention with importance of multi-task learning [24]. Similar setup has not been investigated in models related to (1). Corollary 1, provides a finite sample bound for the multi-task learning i.e. $l_1/l_2$ penalty as well. In more details, we obtain with high probability

$$\|f_{\widehat{\boldsymbol{\beta}}_{l_1/l_2}} - f_{\boldsymbol{\beta}^*}\|^2 \le 26 e^{C(r_n e^C + 2\upsilon_1)} \frac{d \sum_{j\in\mathcal{M}_*} \lambda_{n,j}^2}{\zeta^2}$$

with $0 \le \upsilon_1 \le 1$ satisfying

$$\upsilon_1 e^{-2C\upsilon_1} \le 16d\rho'(0+) \sum_{j\in\mathcal{M}_*} \lambda_{n,j}^2/\zeta^2$$

and $r_n = d\zeta^{-2}\sum_{j\in\mathcal{M}_*}\lambda_{n,j}$. If in addition $P(\mathcal{E}_0)\to 1$, the upper bound above, up to a constant, matches the minimax rate of [24] developed for high dimensional linear models.

### 5.2. Smooth selection

Throughout the previous sections, we simplified the technical details and left out the smoothing component of the penalty. Although selection of groups of features is important, smoothing splines become of interest when considering non-parametric estimation. Because of knot selection, there are potential questions of stability of estimation. Adding pre-described smoothing requirements for the choice of $\Psi$ has become a standard technique for avoiding instability.

Next we present one example of the ANOVA Cox model where the hazard rate function of interest is a nonparametric function of covariates. Consider the multivariate nonparametric regression problem

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\left\{f(\mathbf{X})\right\},$$

where $f$ is the function of estimation interest. A popular model for high dimensional problems above is a smoothing spline analysis of variance model

$$\begin{aligned}\lambda_i(t|\mathbf{X}) = \lambda_0(t)\exp\Bigg\{&\sum_{j=1}^{p} g_j(X_{ij}) + \sum_{j=1}^{p}\sum_{k=1,j<k}^{p} g_{jk}(X_{ij},X_{ik})\\ &+\sum_{j=1}^{p}\sum_{k=1,j<k}^{p}\sum_{l=1,j<k<l}^{p} g_{jkl}(X_{ij},X_{ik},X_{il}) + \cdots\Bigg\},\end{aligned}$$

and in particular its truncated version

$$\lambda_i(t|\mathbf{X}) = \lambda_0(t)\exp\left\{\sum_{j=1}^{p} g_j(X_{ij}) + \sum_{j<k} g_{jk}(X_{ij},X_{ik}) + r\right\},$$

where $r$ is the truncation term. The identifiability of the terms is assured by the side conditions through averaging operators [23]. The form of FGP penalty is similar to the common smoothing spline and allows multiple smoothing parameters for each function $f_{jk}$ independently. With the notation of previous chapters, in this model $g_j$ can be approximated using a set of basis functions $\{B_q\}_{q=1}^{d}$, as $g_j = \sum_{q=1}^{d}\beta_{jq}B_q(X_{ij})$ and similarly $g_{jk} = \sum_{q=1}^{d}\Theta_{jkq}B_q(X_{ij})B_q(X_{ik})$, with $\beta_{jk}$ and $\Theta_{jkq}$ as the unknown parameters. In this model, the total number of unknown parameters is $1/2(p^2+3p)d$ and each $f_j$ takes on a bilinear form $\sum_{q=1}^{d}\beta_{jq}B_q(X_{ij}) + \sum_{q=1}^{d}\sum_{k=1}^{p}\Theta_{jkq}B_q(X_{ij})B_q(X_{ik})$. Hence, in notation of the previous chapters, $b_{jk}$ is now a vector $\mathbf{b}_j = \text{vec}(\beta_{jk},(\Theta_{j1k},\ldots,\Theta_{jpk})^T)$ and $\Psi_k(X_{ij}) = \text{vec}(B_k(X_{ij}), B_k(X_{ij})(B_k(X_{i1}),\ldots,B_k(X_{ip}))^T)$. In the display above, $\text{vec}(\mathbf{A}) = (A_{11},\ldots,A_{1p},\ldots,A_{k1},\ldots,A_{kp})$, for a matrix $\mathbf{A}$.

Let the smoothing matrix, $\mathbf{M}_j \in \mathbb{R}^{d\times d}$, contain the inner products of the second derivatives of the B-spline basis functions, i.e.,

$$\{\mathbf{M}_j\}_{kl} = \int \Psi_k^{''}(x_j)\Psi_l^{''}(x_j)dx_j, \qquad \mathbf{M}_j = \mathbf{R}_j^T\mathbf{R}_j,$$

$k, l = 1, \dots, d$, and $\mathbf{R}_j \in \mathbb{R}^{d\times d}$ is a matrix obtained from Cholesky decomposition of $\mathbf{M}_j$. More generally, we show that the work of the previous sections extends to this situation with only a few adaptations. Let us define the penalized smoothed estimator as

$$\widehat{\boldsymbol{\beta}}_{\mathbb{S}} = \arg\min_{\mathbf{b}}\left\{-\mathcal{L}_n(\mathbf{b})+\lambda_n\sum_{j=1}^p \sqrt{d}\rho\left(\|\mathbf{b}_j^T\mathbf{R}_j\|_{\gamma_j} + \sqrt{\mathbf{b}_j^T\mathbf{M}_j\mathbf{b}_j}\right)\right\}, \quad \text{for } \gamma_j \geq 2,$$

for a convex and subadditive choice of $\rho$. Then, we can rewrite the problem as

$$\widehat{\boldsymbol{\beta}}_{\text{s}} = \arg\min_{\tilde{\mathbf{b}}}\left\{-\mathcal{L}_n(\tilde{\mathbf{b}}) + \lambda_n\sum_{j=1}^p \sqrt{d}\rho\left(\|\tilde{\mathbf{b}}_j\|_{\gamma_j} + \|\tilde{\mathbf{b}}_j\|_2\right)\right\},$$

with $\tilde{\mathbf{b}}_j = \mathbf{R}_j\mathbf{b}_j$ and

$$\tilde{\mathcal{L}_n}(\tilde{\mathbf{b}}) = \frac{1}{n}\sum_{i=1}^n\int_0^\tau\Bigg[\sum_{j=1}^p \tilde{\mathbf{b}_j}^T\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij}) \\ - \log\Bigg(\frac{1}{n}\sum_{i=1}^n Y_i(t)\exp\Bigg\{\sum_{j=1}^p \tilde{\mathbf{b}_j}^T\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij})\Bigg\}\Bigg)\Bigg]dN_i(t).$$

A crucial part of extending the previous results to this novel setting requires extending the results of Lemma 1 and Propositions 2 and 3 to the new penalty structure. Details of the proof are presented in the Appendix D. To state the results we define a set

$$\mathcal{T}_n = \left\{\|\tilde{\mathbf{h}}_{n,j}(\boldsymbol{\beta}^*)\|_{\gamma_j^*} \leq 2\lambda_n \max_j\{d^{1/\gamma_j^*}\sqrt{d}\}\min_j \lambda_{\min}(\mathbf{R}_j)\rho'(0+), \forall j \in \{1,\dots,p\}\right\}$$

for a new score vector $\tilde{\mathbf{h}}_{n,j}(\boldsymbol{\beta}^*) = -\frac{1}{n}\sum_{i=1}^n\int_0^\tau(\widetilde{\mathbf{E}}_{n,j}(\boldsymbol{\beta}^*,t) - \mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij}))dM_i(t)$ and

$$\widetilde{\mathbf{E}}_{n,j}(\boldsymbol{\beta}^*,t) = \frac{1}{n}\sum_{i=1}^n\frac{Y_i(t)\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij})}{\frac{1}{n}\sum_{l=1}^n Y_l(t)\exp\left\{\sum_{j=1}^p \boldsymbol{\beta}_j^{*T}\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{lj})\right\}} \\ \times\exp\left\{\sum_{j=1}^p \boldsymbol{\beta}_j^{*T}\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij})\right\}. \tag{42}$$

Let $N$ represent the number of distinct events and let

$$\underline{\underline{\omega}}_{\mathbb{S}} := \min_{i\in\{1,\dots,n\}, i\in\cup_{q=1}^n\mathbb{R}_q}\left\{\frac{\sum_{q=1}^N\exp\{\sum_{j=1}^p\boldsymbol{\beta}_j^{*T}\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij})\}1\{i\in\mathbb{R}_q\}\}}{\sum_{l\in\mathbb{R}_q}\exp\{\sum_{j=1}^p\boldsymbol{\beta}_j^{*T}\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{lj})\}}\right\}.$$

**Lemma 6.** *Let* $\mathbf{v} = (\mathbf{v}_1^T, \ldots, \mathbf{v}_p^T)^T \in \mathbb{R}^{pd}$*, with* $\mathbf{v}_{n,j} \in \mathbb{R}^d$*. Then, on the event* $\mathcal{T}_n$*, the penalty function* $P(\tilde{\mathbf{b}}) = \sum_{j=1}^p \sqrt{d}\rho(\|\tilde{\mathbf{b}}_j\|_{\gamma_j} + \|\tilde{\mathbf{b}}_j\|_2)$ *satisfies* (8)*. Moreover, the following two statements hold almost surely:*

*(i) Let* $\|f_{\mathbf{b}}\|^2_{n,\underline{\underline{\omega}}_{\mathbb{S}},\mathbf{b}^*}$ *be defined in the same way as* $\|f_{\mathbf{b}}\|^2_{n,\mathbf{b}^*}$ *is* (37) *and replace* $\omega$ *with* $\underline{\underline{\omega}}_{\mathbb{S}}$*. Then, uniformly for every* $\mathbf{b} \in [-b_n, b_n]$*, with* $b_n > 0$ *and* $\mathbf{b}^* = c\mathbf{b} + (1-c)\boldsymbol{\beta}^*$*, with* $c \in (0,1)$*,*

$$\underline{\underline{\omega}}_{\mathbb{S}}\|f_{\mathbf{b}}\|^2 \le \|f_{\mathbf{b}}\|^2_{n,\underline{\underline{\omega}}_{\mathbb{S}},\mathbf{b}^*} \le \|f_{\mathbf{b}}\|^2.$$

*(ii) Let* $a_{\mathbf{v}}$ *be a constant defined in Proposition* 2*. Then, the following sandwich bound holds almost surely for every vector* $\mathbf{b}$ *and corresponding vector* $\mathbf{b}^* = c\mathbf{b} + (1-c)\boldsymbol{\beta}^*$*,*

$$\begin{aligned} &\underline{\underline{\omega}} e^{-2a_{\mathbf{b}-\boldsymbol{\beta}^*} \max_{1\le j\le p} \lambda_{\min}^{-1}(\mathbf{R}_j)} \;\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2 \\ &\le \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\underline{\underline{\omega}}_{\mathbb{S}},\mathbf{b}^*} \\ &\le e^{2a_{\mathbf{b}-\boldsymbol{\beta}^*} \max_{1\le j\le p} \lambda_{\min}^{-1}(\mathbf{R}_j)} \;\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2, \end{aligned}$$

*uniformly for every* $c \in (0,1)$*.* □

With the help of the results presented in earlier sections and this Lemma, we have the following Corollary.

**Corollary 2.** *Let conditions of Theorem* 2 *be satisfied. Let* $\mathbf{M}_j$ *be well defined with* $\underline{\lambda} = \min_{1\le j\le p} \lambda_{\min}(\mathbf{R}_j) > 0$*. Then, for some constant* $A > 4$ *and the choice of the tuning parameters*

$$\lambda_n d^2 \ge A \min\left\{\zeta^2, \sqrt{\frac{\log(pd)}{n}}\right\},$$

*we have*

$$\begin{aligned} &P\left(\|f_{\widehat{\boldsymbol{\beta}}_{\mathbb{S}}} - f_{\boldsymbol{\beta}^*}\|^2 \le \underline{\underline{\omega}}^{-1} e^{C(r_n e^C + 2\upsilon_1)} 32\sqrt{2}\frac{s\lambda_n^2 d}{\zeta^2} \sum_{j\in\mathcal{M}_*} \mathbf{R}_j\mathbf{R}_j^T\right) \qquad (43) \\ &\ge 1 - 6\{pd\}^{1-A} - P\Big(\mathbf{RE(}7, s, \rho, \boldsymbol{\gamma}, -\bigtriangledown^2 \tilde{\mathcal{L}}_n(\boldsymbol{\beta}^*)\mathbf{)} \le \zeta^2\Big) \\ &- 4pdP\Big(\max_{1\le i\le n} \exp\{g(\mathbf{X}_i)\} > y\Big), \end{aligned}$$

*for* $r_n = \zeta^{-2}s\lambda_n d$*, and* $0 \le \upsilon_1 \le 1$ *satisfying*

$$\upsilon_1 e^{-2C\upsilon_1} \le 16\lambda_n^2\underline{\lambda}\frac{sd}{\zeta^2}\rho'(0+).$$

The result above is a finite sample one on prediction properties of a nonparametric smoothing estimator for the high-dimensional Cox model. A particular example of a smooth selection is the Elastic net penalty [45]. Although our previous results easily apply to this penalty (by specifying $\gamma_j = 1$ and $\rho = l_1$), its efficient implementation in the Cox model was only recently proposed in [41], but its theoretical properties have not been previously studied. Although tackled as the last problem, the importance of the obtained finite sample bounds for smooth selection lies in the inadmissibility of such results with techniques that already exist in the literature. In particular, in the case of Elastic-Net penalty we obtain with high probability

$$\|f_{\widehat{\boldsymbol{\beta}}_{\text{elastic net}}} - f_{\boldsymbol{\beta}^*}\|^2 \leq \underline{\underline{\omega}}^{-1} e^{C(r_n e^C + 2\upsilon_1)} 32\sqrt{2}\frac{s^2\lambda_n^2}{\zeta^2}$$

for $r_n = \zeta^{-2}s\lambda_n$, and $0 \leq \upsilon_1 \leq 1$ satisfying $\upsilon_1 e^{-2C\upsilon_1} \leq 16\lambda_n^2\frac{s}{\zeta^2}\rho'(0+)$. Such a result holds under the assumption that the random design $\mathbf{X}$ is such that last two terms of the equation (41) are negligible.

## Discussion

In this paper, we propose a new method for analyzing the theoretical oracle risk properties of likelihood functions that are not necessarily of a quadratic nature. By sandwiching the likelihood with two other processes, we establish that it is sufficient to analyze the risk properties of the bounding processes alone. To the best of our knowledge, minimax rates, have not been established for any survival model so far despite their importance. Equivalents of traditional information theoretic tools, such as Fano's lemma, are not easy to understand in the Cox model setup. Our proposed method of sandwiching the likelihood with two quadratic likelihoods may be useful in establishing minimax rates.

## Appendix A: The restricted eigenvalue condition

The restricted eigenvalue condition, $\mathbf{RE}(\mu, s, \rho, \boldsymbol{\gamma}, \mathbf{A})$, defined in (12) represents a generalization of the cone constraint condition that appears in work on Lasso problems [4]. Equivalent definitions were proposed for various hazard rate models [21, 14, 17, 15]. We refer to [6] for comparisons of different kinds of compatibility and restricted eigenvalue conditions and their relationships for sparse linear models. The usual scaling factor of $\sqrt{n}$ disappears in the definition of the restricted eigenvalue condition because it is included in the definition of the empirical norm, $\|f(\cdot)\|_{n,\cdot}^2$. Compared to the RE condition in [4], the denominators differ in that the $l_2$ norm is replaced with an $l_{1,\boldsymbol{\gamma}}$ norm. In the least squares procedures, $\bigtriangledown^2\mathcal{L}_n(\boldsymbol{\beta}^*) = -\mathbf{X}^T\mathbf{X}$ and the restricted eigenvalue conditions are defined on the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Condition (14) can be seen as a rescaling of the minimum eigenvalue problem in the classical $\mathbf{RE}$ condition needed for the complex likelihood structures.

Determining the class of matrices that satisfy the $\mathbf{RE}(\mu, s, \rho, \boldsymbol{\gamma}, -\bigtriangledown^2\mathcal{L}_n(\boldsymbol{\beta}^*))$ condition is an important open question. Heuristically we can argue in the following manner. First, we observe that with respect to time, $\int_0^\tau \mathbf{V}_n(0,t)d\bar{N}(t)$ has a martingale structure. With respect to $\boldsymbol{\beta}^*$, it is a function of the matrix $\sum_{i=1}^n\sum_{q=1}^n \boldsymbol{\Psi}^T(\mathbf{X}_i)\boldsymbol{\Psi}(\mathbf{X}_q)$. Using Condition 1 and the boundedness of the $\Psi$ functions, matrix $\int_0^\tau \mathbf{V}_n(0,t)d\bar{N}(t)$ will belong to a random matrix ensemble with sub-Gaussian tails, studied in [44]. Dependence through time was established not to be essential in [15], where a lower bound for **RE** was shown to be independent of time. Moreover, we can combine both results to conclude that for large enough sample size, there exists a positive constant $\zeta_1$ such that with overwhelming probability

$$\min_{\boldsymbol{\Delta}\in\mathbb{C}_{\mu,\rho},\boldsymbol{\Delta}\neq 0} \frac{\|\boldsymbol{\Delta}^T\{-\bigtriangledown^2 \mathcal{L}_n(\boldsymbol{\beta}^*)\}\boldsymbol{\Delta}\|_2}{\|\boldsymbol{\Delta}_{\mathcal{M}_*}\|_{1,\boldsymbol{\gamma}}^2} \geq \zeta_2,$$

for all random designs **X** with bounded moments.

## Appendix B: Preliminary lemmas

The following lemma provided exponential inequality for a martingale sequence and can be found in [37] as Lemma 2.1

**Lemma 7.** *Let $(\Omega, \mathcal{F}, P)$ be a probability triple and let $M_t$ be a sequence of locally square integrable martingales w.r.t. the filtration $\mathcal{F}_t$. Suppose that $|M_t - M_{t-}| \leq K$ for all $t > 0$ and some $0 < K < \infty$. Then, for each $a > 0, b > 0$.*

$$P\left(M_t \geq a \text{ and } \langle M, M\rangle_t \leq b^2 \text{ for some } t\right) \leq \exp\left\{-\frac{a^2}{2(aK + b^2)}\right\},$$

*where $\langle M, M\rangle_t$ denotes predictable variation of the martingale sequence $M_t$.* □

The following lemma provides an exponential inequality for a unbounded supermartingale sequence and can be found in [13] as Corollary 2.3.

**Lemma 8.** *Let $(\Omega, \mathcal{F}, P)$ be a probability triple. Assume that $(\xi_i, \mathcal{F}_i)_{i=1,\dots,n}$ are supermartingale differences i.e. $E(\xi_i|\mathcal{F}_{i-1}) \leq 0$. Let $b > 0$ and*

$$V_k^2(b) = \sum_{i=1}^k E\left(\xi_i^2 1\{\xi_i \leq b\}|\mathcal{F}_{i-1}\right), \qquad k = 1, \dots, n.$$

*Then, for any $a \geq 0$, $b > 0$ and $c > 0$*

$$P\left(\sum_{i=1}^k \xi_i \geq a \text{ and } V_k^2(b) \leq c^2 \text{ for some } k\right) \\ \leq \exp\left\{-\frac{a^2}{2(c^2 + \frac{1}{3}ab)}\right\} + P(\max_{1\leq i\leq n} \xi_i > b).$$

□

## Appendix C: Proofs of propositions

*Proof of Proposition 1.* Without loss of generality, let us represent the optimization problem (31) as a quadratically constrained minimum of the ratio of two quadratic functions of the following form

$$\begin{array}{lc} \min_{\mathbf{b}} & \sum_{q=1}^{N} \frac{(\mathbf{b}^T \mathbf{A}_1^i \mathbf{b} + 2\mathbf{a}_1^{i\,T} \mathbf{b} + c_1) 1(i \in \mathcal{R}_q)}{\mathbf{b}^T \mathbf{A}_2 \mathbf{b} + 2\mathbf{a}_2^T \mathbf{b} + c_2}, \\ \text{s.t} & \|\mathbf{b}\|_2^2 \le r_n, \\ & \mathbf{b} \in \mathbb{R}^{pd}, \end{array} \tag{44}$$

where $\mathbf{A}_1^i = \boldsymbol{\Psi}(\mathbf{X}_i)\boldsymbol{\Psi}(\mathbf{X}_i)^T$, $\mathbf{A}_2 = \sum_{l \in \mathcal{R}_q} \boldsymbol{\Psi}(X_l)\boldsymbol{\Psi}^T(\mathbf{X}_l)$ and $a_1^i = \boldsymbol{\Psi}(\mathbf{X}_i)$, $a_2 = \sum_{l \in \mathcal{R}_q} \boldsymbol{\Psi}(X_l)$. Constants $c_1$ and $c_2$ are residuals of the Maclaurin series expansions of the functions $\exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_i)\}$ and $\sum_{l \in \mathcal{R}_q} \exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_l)\}$. This makes $\mathbf{A}_1^i$ and $\mathbf{a}_1^i$ second order and first order approximations of $\exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_i)\}$, around $\mathbf{0}$.

Condition (30) implies that for any feasible point $\mathbf{b}$, the above optimization problem is well defined. Multiplying (30) by $(\mathbf{b}^T, 1)$ from the left and $(\mathbf{b}^T, 1)^T$ from the right results in

$$\sum_{l \in \mathcal{R}_q} \exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_l)\} + \eta(\|\mathbf{b}\|_2^2 - r_n) \ge \delta(\|\mathbf{b}\|_2^2) + 1,$$

which implies that $\sum_{l \in \mathcal{R}_q} \exp\{\mathbf{b}^T \boldsymbol{\Psi}(\mathbf{X}_l)\} \ge \delta(\|\mathbf{b}\|_2^2) + 1 \ge \delta > 0$.

Let us fix an $i \in \mathbb{R}_q$ for some $q$. Now, let us define

$$d_1 = \inf\big\{ f(\mathbf{b}) : \|\mathbf{b}\| \le r_n, \mathbf{b}^T \mathbf{A}_1^i \mathbf{b} + 2\mathbf{a}_1^{i\,T} \mathbf{b} + c_1 \ge 0 \big\}, \tag{45}$$

$$d_2 = \inf\big\{ f(\mathbf{b}) : \|\mathbf{b}\| \le r_n, \mathbf{b}^T \mathbf{A}_1^i \mathbf{b} + 2\mathbf{a}_1^{i\,T} \mathbf{b} + c_1 \le 0 \big\}, \tag{46}$$

with $f(\mathbf{b}) = (\mathbf{b}^T \mathbf{A}_1^i \mathbf{b} + 2\mathbf{a}_1^{i\,T} \mathbf{b} + c_1)/(\mathbf{b}^T \mathbf{A}_2 \mathbf{b} + 2\mathbf{a}_2^T \mathbf{b} + c_2)$. Then using the relation that

$$\inf\{f(\mathbf{b}) : \mathbf{b} \in \mathcal{C}_1 \cup \mathcal{C}_2\} = \min\Big\{ \inf_{\mathbf{b} \in \mathcal{C}_1} f(\mathbf{b}), \inf_{\mathbf{b} \in \mathcal{C}_2} f(\mathbf{b}) \Big\},$$

we have that the optimal solution to (44) is equal to $\min\{d_1, d_2\}$. By definition, $d_1$ is nonnegative. It remains to establish that $d_2$ is finite. Indeed, for every $\mathbf{b}$ satisfying $\|\mathbf{b}\|_2^2 \le r_n$ and $\mathbf{b}^T \mathbf{A}_1^i \mathbf{b} + 2\mathbf{a}_1^{i\,T} \mathbf{b} + c_1 \le 0$, we have

$$d_2 \ge f(\mathbf{b}) \ge \frac{\mathbf{b}^T \mathbf{A}_1^i \mathbf{b} + 2\mathbf{a}_1^{i\,T} \mathbf{b} + c_1}{\delta(\|\mathbf{b}\|_2^2) + 1} \ge \frac{1}{\delta} \lambda_{\min} \begin{pmatrix} \mathbf{A}_1^i & \mathbf{a}_1 \\ \mathbf{a}_1^T & c_1 \end{pmatrix}.$$

□

*Proof of Proposition 2.* To see that the equation (38) is correct, we adopt the following reasoning. First, note that $\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|_{n,\mathbf{b}^*}^2$ is equal to

$$n^{-1} \int_0^\tau \frac{\sum_{i,q=1}^n w_i w_q (a_i - a_q)^{\otimes 2} e^{(1-c)a_i - \bar{c}} e^{(1-c)a_q - \bar{c}}}{\sum_{i,q=1}^n 2 w_i w_q e^{(1-c)a_i - \bar{c}} e^{(1-c)a_q - \bar{c}}} d\bar{N}(t),$$

with $a_i = (\mathbf{b} - \boldsymbol{\beta}^*)^T (\Psi(\mathbf{X}_i) - \mathbf{E}_n(\boldsymbol{\beta}^*, t))$ and $w_i = Y_i(t) \exp\{\boldsymbol{\beta}^{*T} \Psi(\mathbf{X}_i)\}$ and $\bar{c} = (1-c)(\max_i a_i + \min_i a_i)/2$. If we let $\eta = a_{\mathbf{b} - \boldsymbol{\beta}^*}$, we can see that $\max_i |(1-$

$c)a_i - \bar{c}| \leq \eta/2$. Using this notation, $e^{(1-c)a_i-\bar{c}} \geq e^{-\eta/2}$ and $e^{(1-c)a_i-\bar{c}} \leq e^{\eta/2}$ leading to

$$\begin{aligned}\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} &\geq \exp\{-2\eta\}n^{-1}\int_0^\tau \frac{\sum_{i,q=1}^n w_i w_q (a_i - a_q)^{\otimes 2}}{\sum_{i,q=1}^n 2w_i w_q} d\bar{N}(t) \\ &= \exp\{-2\eta\}\ \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\boldsymbol{\beta}^*}.\end{aligned}$$

The upper bound follows the same reasoning, and thus it is omitted. The lower bound of the RHS of previous inequality follows by repeating the same steps as in Proposition 3 and the definition of the weight vectors, $\omega_i(\boldsymbol{\beta}^*)$, in (17),

$$\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} \geq \underline{\underline{\omega}}\exp\{-2\eta\}\ \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_n.$$

The upper bound follows directly from Proposition 3 by taking $\mathbf{b}^* = \boldsymbol{\beta}^*$ to obtain

$$\|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} \leq \exp\{-2\eta\}\ \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|^2_n.$$

□

*Proof of Proposition 3.* Let $N$ denote the cardinality of the set $\{i = 1, \ldots, n : N_i(\tau) = 1\}$. The weight process, $\omega_i(\mathbf{b}, t)$ as defined in (21), satisfies the following normalization uniformly over $\mathbf{b}$ and $t$,

$$\frac{1}{n}\sum_{i=1}^n Y_i(t)\omega_i(\mathbf{b}, t) = 1.$$

For each $\mathbf{b}$, there exists at least one $i \in \{1, \ldots, n\}$ such that $\omega_i(\mathbf{b}, t) > 0$ and that for all $i$, for which $\exists t \in [0, \tau]$, $Y_i(t) = 1$, we have that $\omega_i(\mathbf{b}, t) \leq n$, for all $t$.

Let us denote

$$\omega_i(\mathbf{b}) = \int_0^\tau Y_i(t)\omega_i(\mathbf{b}, t)d\bar{N}(t),$$

with $\omega_i(\mathbf{b}, t)$ defined as in (21). If $t_1 < \cdots < t_N$ are ordered failure times and $\mathcal{R}_j = \{i \in \{1, \ldots, n\} : Z_i \geq t_j\}$ is at risk set, then $\omega_i(\mathbf{b})$ has the following representation:

$$\omega_i(\mathbf{b}) = \sum_{j=1}^N \frac{\exp\{\mathbf{b}^T\boldsymbol{\Psi}(\mathbf{X}_i)\}1\{i \in \mathbb{R}_j\}}{\sum_{l\in\mathbb{R}_j}\exp\{\mathbf{b}^T\boldsymbol{\Psi}(\mathbf{X}_l)\}},$$

which matches the definition provided in Theorem 1 equation (17). Note that $\omega_i \geq 0$ and $\omega_i > 0$ for $i \in \{1, \ldots, n\}$. Using the previous notation, we have

$$\|f_{\mathbf{b}}\|^2_{n,\mathbf{b}^*} = \frac{1}{n}\sum_{i=1}^n f^2_{\mathbf{b}}(X_i)\omega_i(\mathbf{b}^*) - \left(\frac{1}{n}\sum_{i=1}^n f_{\mathbf{b}}(X_i)\omega_i(\mathbf{b}^*)\right)^2,$$

With this notation at hand, we have that

$$\frac{1}{n}\sum_{i=1}^n \omega_i(\mathbf{b}^*) = \frac{1}{n}\sum_{j=1}^N\sum_{i\in\mathbb{R}_j}\frac{\exp\{\mathbf{b}^T\boldsymbol{\Psi}(\mathbf{X}_i)\}}{\sum_{l\in\mathbb{R}_j}\exp\{\mathbf{b}^T\boldsymbol{\Psi}(\mathbf{X}_l)\}} = \frac{N}{n}.$$

Since $\omega_i(\mathbf{b}^*) \geq 0$, and are defined as conditional probabilities we have $\omega = \max\{\omega_i(\mathbf{b}) : i \in \{1,\dots,n\}, \mathbf{b} \in R^{pd}\} \leq 1$. We are then able to conclude that $1 \geq \bar{\omega} = \max\{\omega_i(\mathbf{b}) : i \in \{1,\dots,n\}, \mathbf{b} \in R^{pd}\} \geq 1/n \geq \underline{\omega} = \min\{\omega_i(\mathbf{b}) : i \in \{1,\dots,n\}, \mathbf{b} \in R^{pd}\}$. Hence,

$$\|f_{\mathbf{b}}\|^2_{n,\mathbf{b}^*} \leq \bar{\omega}\|f_{\mathbf{b}}\|^2_n - \underline{\omega}\left(\frac{1}{n}\sum_{i\in\mathbf{I}} f_{\mathbf{b}}(X_i)\right)^2 \leq \|f_{\mathbf{b}}\|^2_n.$$

To obtain the left-hand side of (6), remember from previous exposition we have

$$\|f_{\mathbf{b}}\|^2_{n,\mathbf{b}^*} = \frac{1}{n}\sum_{i=1}^n \omega_i(\mathbf{b}_{\mathbf{b}})(f_{\mathbf{b}}(\mathbf{X}_i) - \bar{f}^*_{\mathbf{b}})^2$$

with $\bar{f}^*_{\mathbf{b}} = \frac{1}{n}\sum_{i=1}^n \omega_i(\mathbf{b}_{\mathbf{b}}) f_{\mathbf{b}}(X_i)$ and $\omega_i(\mathbf{b}_{\mathbf{b}})$ following the definition in (17). Hence, by centering the data so that the sample mean is equal to zero, i.e., $\frac{1}{n}\sum_{i=1}^n f_{\mathbf{b}}(\mathbf{X}_i) = 0$, we have

$$\begin{aligned}\|f_{\mathbf{b}}\|^2_{n,\mathbf{b}^*} &\geq \underline{\omega}\frac{1}{n}\sum_{i\in I}\left(f^2_{\mathbf{b}}(\mathbf{X}_i) + \{\bar{f}^*_{\mathbf{b}}\}^2\right) + 2\underline{\omega}\ \bar{f}^*_{\mathbf{b}}\ \left(\frac{1}{n}\sum_{i=1}^n f_{\mathbf{b}}(\mathbf{X}_i)\right)\\ &\geq \underline{\omega}\frac{1}{n}\sum_{i=1}^n f^2_{\mathbf{b}}(\mathbf{X}_i) = \underline{\omega}\|f_{\mathbf{b}}\|^2_n.\end{aligned}$$

The result of the Proposition 3 follows easily after applying Proposition 1 on the interval $[-b_n, b_n]$ and following discussion after Proposition 1. □

## Appendix D: Proofs of lemmas

*Proof of Lemma 1.* Let us first concentrate on the first statement of (8). This can be seen from the following reasoning. Let us define a function

$$f(\mathbf{b}) := -(\mathbf{b} - \boldsymbol{\beta}^*)^T\mathbf{v}_n + \lambda_n P(\mathbf{b}) - \mathcal{L}_n(\boldsymbol{\beta}^*),$$

where $\mathbf{v}_n \in \mathbb{R}^{pd}$. First, we establish that zero is a local minimum of function $f(\mathbf{b})$ for all $\mathbf{b}$ such that $\|\mathbf{b}_j\|_1 \leq 1$. Note that

$$f(\mathbf{b}) - f(\mathbf{0}) = \sum_{j=1}^p\left(-\mathbf{b}_j^T\mathbf{v}_{n,j} + \lambda_n d^{1/\gamma_j^*}\rho(\|\mathbf{b}_j\|_{\gamma_j})\right),$$

and conditional on the event $\mathcal{E}_{n,j} = \{\|\mathbf{v}_{n,j}\|_{\gamma_j^*} \leq \lambda_n d^{1/\gamma_j^*}\rho'(0+)\}$,

$$-\mathbf{b}_j^T\mathbf{v}_{n,j} + \lambda_n d^{1/\gamma_j^*}\rho(\|\mathbf{b}_j\|_{\gamma_j}) \geq \|\mathbf{b}_j\|_{\gamma_j}\left(-\|\mathbf{v}_{n,j}\|_{\gamma_j^*} + \lambda_n d^{1/\gamma_j^*}\rho'(0+)\right) \geq 0,$$

where we have utilized the Höelder inequality. Therefore, we can conclude that $f(\mathbf{b}) - f(\mathbf{0}) \geq 0$ if the event $\mathcal{E}_n = \cap_{j=1}^p \mathcal{E}_{n,j}$. Because $f$ is a convex function, we can conclude that 0 is a global minimum as well. Note that we don't require unicity of minimum.

We are left to prove the second statement of (8). We proceed in the similar way by first defining an appropriate function to minimize over. To that end, let us define

$$f(\mathbf{b}) := -|(\boldsymbol{\beta}^* - \mathbf{b})^T \mathbf{v}_n| + \lambda_n P(\mathbf{b}) - \mathcal{L}_n(\boldsymbol{\beta}^*),$$

where $\mathbf{v}_n \in \mathbb{R}^{pd}$. By the same reasoning as above, it suffices to notice that

$$\begin{aligned} & -\Big(|(\boldsymbol{\beta}^* - \mathbf{b})^T \mathbf{v}_{n,j}| - |{\boldsymbol{\beta}^*}^T \mathbf{v}_{n,j}|\Big) + \lambda_n d^{1/\gamma_j^*} \rho(\|\mathbf{b}_j\|_{\gamma_j}) \\ \geq \quad & -|\mathbf{b}^T \mathbf{v}_{n,j}| + \lambda_n d^{1/\gamma_j^*} \rho(\|\mathbf{b}_j\|_{\gamma_j}) \\ \geq \quad & \|\mathbf{b}_j\|_{\gamma_j} \Big(-\|\mathbf{v}_{n,j}\|_{\gamma_j^*} + \lambda_n d^{1/\gamma_j^*} \rho'(0+)\Big) \geq 0, \end{aligned}$$

by first using $|x-y| \geq ||x| - |y||$ and then Höelder inequality. □

*Proof of Lemma 2.* We make use of the following decomposition

$$\begin{aligned} & \|\mathbf{E}_n(\boldsymbol{\beta}^*, t) - \mathbf{e}(\boldsymbol{\beta}^*, t)\|_\infty = \\ \leq \quad & \max_{1\leq j\leq p, 1\leq k\leq d} \tfrac{\left|\{S_n^{(1)}\}_{jk}(\boldsymbol{\beta}^*,t) - \{s^{(1)}\}_{jk}(\boldsymbol{\beta}^*,t)\right|}{|s^{(0)}(\boldsymbol{\beta}^*,t)|} \\ + \quad & \max_{1\leq j\leq p, 1\leq k\leq d} \left|\{s^{(1)}\}_{jk}(\boldsymbol{\beta}^*, t)\right| \left|\tfrac{1}{S_n^{(0)}(\boldsymbol{\beta}^*,t)} - \tfrac{1}{s^{(0)}(\boldsymbol{\beta}^*,t)}\right| \end{aligned} := I_1 + I_2 \qquad (47)$$

We will prove maximal inequalities for each of the two terms in the above inequality.

First, consider classes of functions indexed by $t$:

$$\mathcal{F} = \{1\{z > t\} \exp\{f_{\boldsymbol{\beta}^*}(x)\}/u : t \in [0, \tau]\},$$

and

$$\mathcal{G}^k = \{1\{z > t\} \Psi_k(x) \exp\{f_{\boldsymbol{\beta}^*}(x)\}/u : t \in [0, \tau]\}.$$

Since $\boldsymbol{\beta}^*$ is a $s$-sparse vector we have that $u = \exp\{\sum_{j \in \mathcal{M}_*} \|\boldsymbol{\beta}_j^*\|_1\}$. We proceed by calculating theirs bracketing number. Noticing that previous classes are products of a class of indicator functions and a class of bounded functions we have that

$$\mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_2) \leq 2/\epsilon^2, \qquad \mathcal{N}_{[]}(\epsilon, \mathcal{G}^k, L_2) \leq 2/\epsilon^2,$$

By direct consequence of Theorem 2.14.9 of [39] we obtain that there exists a constant $W$ such that

$$\begin{aligned} P\Bigg(\sqrt{n} \sup_{t \in [0,\tau]} \Bigg| \frac{1}{n} \sum_{i=1}^n Y_i(t) \exp\{f_{\boldsymbol{\beta}^*}(\mathbf{X}_i)\}/u \\ - E_{Y,X} Y_i(t) \exp\{f_{\boldsymbol{\beta}^*}(\mathbf{X}_i)\}/u\} \Bigg| \geq r \Bigg) \leq \frac{1}{2e} W^2 e^{-r^2} \end{aligned}$$

and

$$P\Big(\sqrt{n}\sup_{t\in[0,\tau]}\Big|\frac{1}{n}\sum_{i=1}^{n}Y_i(t)\Psi_k(\mathbf{X}_i)\exp\{f_{\boldsymbol{\beta}^*}(\mathbf{X}_i)\}/u \\ - E_{Y,X}Y_i(t)\Psi_k(\mathbf{X}_i)\exp\{f_{\boldsymbol{\beta}^*}(\mathbf{X}_i)\}/u\Big|\geq r\Big)\leq \frac{1}{2e}W^2e^{-r^2},$$

for every fixed $k\in\{1,\dots,d\}$. By replacing $r$ with $\sqrt{n}r_n$ in the first and utilizing union bound and replacing $r$ with $\sqrt{nr_n^2+\log 2d}$ in the second we obtain

$$P\left(\sup_{t\in[0,\tau]}\left|S_n^{(0)}(\boldsymbol{\beta}^*,t)-s^{(0)}(\boldsymbol{\beta}^*,t)\right|\geq ur_n\right)\leq\frac{1}{2e}W^2e^{-nr_n^2}, \tag{48}$$

$$P\left(\sup_{t\in[0,\tau]}\|S_n^{(1)}(\boldsymbol{\beta}^*,t)-s^{(1)}(\boldsymbol{\beta}^*,t)\|_\infty\geq u\left(\sqrt{r_n^2+\frac{\log 2d}{n}}\right)\right)\leq\frac{1}{4de}W^2e^{-nr_n^2}, \tag{49}$$

Second, from the definition of $s^{(0)}(\boldsymbol{\beta}^*,t)$ and Condition 1 (iii) we observe that there exists a constant $0<D<1$ with $D=P(Y(\tau)=1)$ and

$$\inf_{t\in[0,\tau]}\frac{1}{n}\sum_{i=1}^{n}E_{Y,X}Y_i(t)\exp\{f_{\boldsymbol{\beta}^*}(\mathbf{X}_i)\}\geq\exp\{-m^*C\}P(Y(t)=1) \\ > D\exp\{-m^*C\}$$

with $C$ being an upper bound on $|\Psi_k(x)|$ and $m^*$ defined as minimum signal strength in the additive component of the hazards model (1).

According to (47) and (48) we have

$$I_2\leq\frac{\sup_{t\in[0,\tau]}\left\|s^{(1)}(\boldsymbol{\beta}^*,t)\right\|_\infty}{D\exp\{-m^*C\}}\frac{ur_n}{\inf_{t\in[0,\tau]}S_n^{(0)}(\boldsymbol{\beta}^*,t)}$$

with probability $\frac{1}{2ed}W^2e^{-r_n^2}$, and according to (47) and (49) we have

$$I_1\leq\frac{u\left(\sqrt{r_n^2+\frac{\log 2d}{n}}\right)\exp\{m^*C\}}{D}\leq\frac{u\left(r_n+\sqrt{\frac{\log 2d}{n}}\right)\exp\{m^*C\}}{D},$$

with probability no smaller than $1-\frac{1}{4e}W^2e^{-nr_n^2}$. To further bound $I_2$ we show that $|S_n^{(0)}(\boldsymbol{\beta}^*,t)|$ is bounded away from zero with high probability. To that end, we employ Massart's Dvoretzky-Kiefer-Wolfowitz inequality bounding how close an empirically determined distribution function is to the distribution function from which the empirical samples are drawn. Hence,

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i\geq\tau)\geq\frac{1}{2}P(Z_1\geq\tau)\right)$$

$$\geq P\left(\sup_{t\in[0,\tau]}\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i\geq t)-P(Z_1\geq\tau)\right|\leq\sqrt{n}/2\ P(Z_1\geq\tau)\right)$$
$$\geq 1-2e^{-nD^2/2}. \tag{50}$$

Remeber that $S_n^{(0)}(\boldsymbol{\beta}^*,t)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Z_i\geq t\}\exp\{f_{\boldsymbol{\beta}^*}(\mathbf{X}_i)\}$ and observe that for all $t\leq\tau$ we have $\{Z_i\geq t\}\supset\{Z_i\geq\tau\}$. Hence,

$$S_n^{(0)}(\boldsymbol{\beta}^*,t)\geq\exp\{-m^*C\}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Z_i\geq\tau\},\qquad\text{for all } t\leq\tau.$$

Together with (50) we have

$$P\left(\inf_{t\in[0,\tau]}S_n^{(0)}(\boldsymbol{\beta}^*,t)\geq\exp\{-m^*C\}D/2\right)\geq\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i\geq\tau)\geq D/2\right)$$
$$\geq 1-2e^{-nD^2/2}.$$

Next, we bound $\sup_{t\in[0,\tau]}\|s^{(1)}(\boldsymbol{\beta}^*,t)\|_\infty$. Observe that

$$\sup_{t\in[0,\tau]}\|s^{(1)}(\boldsymbol{\beta}^*,t)\|_\infty\leq\sup_{t\in[0,\tau]}E_X(P\{Z_1\geq t|\mathbf{X}_1\}\exp\{f(\boldsymbol{\beta}^*(\mathbf{X}_1))\})$$
$$\leq E_X(\exp\{\boldsymbol{\beta}^{*T}\boldsymbol{\Psi}(\mathbf{X}_1)\})\leq\exp\{C\log u\}$$

With all of the above notice that

$$I_2\leq\frac{2ur_n\exp\{2m^*C\}\exp\{C\log u\}}{D^2}$$

with probability no smaller than $1-\frac{1}{2ed}W^2e^{-nr_n^2}-2e^{-nD^2/2}$. Hence, we conclude that

$$\|\mathbf{E}_n(\boldsymbol{\beta}^*,t)-\mathbf{e}(\boldsymbol{\beta}^*,t)\|_\infty\leq\frac{u\left(r_n+\sqrt{\frac{\log 2d}{n}}\right)\exp\{m^*C\}}{D}$$
$$+\frac{2ur_n\exp\{2m^*C\}\exp\{C\log u\}}{D^2},$$

with probability no smaller than $1-\frac{3}{8ed}W^2e^{-nr_n^2}-e^{-nD^2/2}$. □

*Proof of Lemma 3.*

**Bounding $\mathcal{D}_{n,i}^c$**

Recall that

$$\mathcal{D}_{n,i}=\bigcap_{j=1}^{p}\left\{4\lambda_0(\tau)\left|\int_0^\tau S_n^{(0)}(g,t)dt\right|\|\boldsymbol{\Psi}(X_{ij})\|_{\gamma_j^*}\leq\lambda_n d^{1/\gamma_j^*}\rho'(0+)\right\}.$$

By simple union bound we see that

$$P(\mathcal{D}_{n,i}^c) \leq \sum_{j=1}^{p} P\left(\lambda_0(\tau)\left|\int_0^\tau S_n^{(0)}(g,t)dt\right| \|\mathbf{\Psi}(X_{ij})\|_{\gamma_j^*} \geq \lambda_n d^{1/\gamma_j^*}\rho'(0+)\right). \quad (51)$$

First, observe that the definition of $S_n^{(0)}(g,t)$ allows the following bound

$$\begin{aligned}\left|\int_0^\tau S_n^{(0)}(g,t)d\Lambda_0(t)\right| &\leq \Lambda_0(\tau)\frac{1}{n}\sum_{i=1}^{n}\exp\{g(\mathbf{X}_i)\}\int_0^\tau Y_i(t)dt\\ &\leq \tau\Lambda_0(\tau)\frac{1}{n}\sum_{i=1}^{n}\exp\{g(\mathbf{X}_i)\},\end{aligned}$$

whereas the boundedness of $\Psi_k$ allows $\|\mathbf{\Psi}(X_{ij})\|_{\gamma_j^*} = (\sum_{k=1}^d \Psi_k^{\gamma_j^*}(X_{ij}))^{1/\gamma_j^*} \leq d^{1/\gamma_j^*}C$ to hold. With this in mind, we observe that

$$P(\mathcal{D}_{n,i}^c) \leq \sum_{j=1}^{p} P\left(\tau\Lambda_0(\tau)\lambda_0(\tau)C^{1/\gamma_j^*}\frac{1}{n}\sum_{i=1}^{n}\exp\{g(\mathbf{X}_i)\} \geq \lambda_n\rho'(0+)\right). \quad (52)$$

Previous inequality is a tail probability of a sum of i.i.d.positive random variables where $g$ is the unknown function of interest.By large-deviation inequality of non-negative random variables (Lemma 8 in the Appendix B), we obtain

$$P\left(\sum_{i=1}^{n}\exp\{g(\mathbf{X}_i)\} \geq \sqrt{n}\gamma_n\right) \leq e^{-\frac{n\gamma_n^2}{2\theta^2+2\gamma_n y/3}} + P(\max_{1\leq i\leq n}\exp\{g(\mathbf{X}_i)\} > y), \quad (53)$$

for a sequence of non-negative numbers $\gamma_n$ and a truncation value $y$ such that

$$\theta^2 \geq \sum_{i=1}^{n} E\exp\{2g(\mathbf{X}_i)\}1\{\exp\{2g(\mathbf{X}_i)\} \leq y\}. \quad (54)$$

By choosing $\gamma_n = M\sqrt{n}\lambda_n\rho'(0+)$ with $M = 1/(\tau\lambda_0(\tau)\Lambda_0(\tau)C)$, we obtain that

$$P(\mathcal{D}_{n,i}^c) \leq e^{-\frac{n^2M^2\lambda_n^2\rho'(0+)^2}{2\theta^2+2M\sqrt{n}\lambda_n\rho'(0+)y/3}} + P(\max_{1\leq i\leq n}\exp\{g(\mathbf{X}_i)\} > y).$$

**Bounding $\mathcal{E}_n^c$**

Notice that the set of interest, $\mathcal{E}_n^c$, is a subset of

$$\bigcup_{j=1}^{p}\left\{\|\mathbf{h}_{n,j}(\boldsymbol{\beta}^*)\|_\infty \geq \lambda_n d^{1/\gamma_j^*}\rho'(0+)\right\},$$

where $\|\mathbf{h}_{n,j}(\boldsymbol{\beta}^*)\|_\infty = \max_{1\leq k\leq d}|\{\mathbf{h}_n\}_{jk}(\boldsymbol{\beta}^*)|$. According to the definition of $\mathbf{h}_n(\boldsymbol{\beta}^*)$

$$\mathbf{h}_n(\boldsymbol{\beta}^*) = -n^{-1}\sum_{i=1}^{n}\int_0^\tau (\mathbf{E}_n(\boldsymbol{\beta}^*,t) - \mathbf{\Psi}(\mathbf{X}_i))\, dM_i(t), \quad (55)$$

we decompose $\mathbf{h}_n(\boldsymbol{\beta}^*)$ as follows

$$\mathbf{h}_n(\boldsymbol{\beta}^*) := \upsilon + \nu \tag{56}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\mathbf{E}_n(\boldsymbol{\beta}^*) - \mathbf{e}(\boldsymbol{\beta}^*, t))\, dM_i(t) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\mathbf{e}(\boldsymbol{\beta}^*, t) - \boldsymbol{\Psi}(\mathbf{X}_i))\, dM_i(t).$$

We will consider each term separately. First, to control $\upsilon_{jk}$'s we develop a finite sample result in Lemma 2 whose proof can be found in the Appendix D.

Next, we bound $|\boldsymbol{\Delta}\upsilon_{jk}|$ and the predictable variation of the martingale $\upsilon_{jk}$. By Lemma 2, with high probability, the jumps are bounded by

$$\begin{aligned} |\boldsymbol{\Delta}\upsilon_{jk}| &= \frac{1}{n}\,|\{\mathbf{E}_n(\boldsymbol{\beta}^*)\}_{jk} - \{\mathbf{e}(\boldsymbol{\beta}^*)\}_{jk}| \\ &\leq \frac{1}{n}\sup_{0\leq t\leq\tau}\|\{\mathbf{E}_n(\boldsymbol{\beta}^*, t)\} - \{\mathbf{e}(\boldsymbol{\beta}^*, t)\}\|_\infty \\ &\leq \frac{w_n}{n}, \end{aligned} \tag{57}$$

with $w_n = cr_n + \sqrt{\frac{\log d}{nu^2}}$. The predictable variation process can be bounded as follows

$$\begin{aligned} \langle\boldsymbol{\Delta}\upsilon_{jk}\rangle_2 &= \frac{1}{n^2}\int_0^\tau [\{\mathbf{E}_n(\boldsymbol{\beta}^*, t)\}_{jk} - \{\mathbf{e}(\boldsymbol{\beta}^*, t)\}_{jk}]^2\, d\langle\bar{M}(t)\rangle \\ &\leq \frac{1}{n}\sup_{0\leq t\leq\tau}\|\{\mathbf{E}_n(\boldsymbol{\beta}^*, t)\} - \{\mathbf{e}(\boldsymbol{\beta}^*, t)\}\|_\infty^2 \int_0^\tau S_n^{(0)}(g, t)d\Lambda_0(t). \end{aligned}$$

The first term on the RHS of the above equation can be bounded above with high probability using Lemma 2 with $w_n$. For the last term we use the result in (53) to conclude that

$$\langle\boldsymbol{\Delta}\upsilon_{jk}\rangle_2 \leq \frac{\tau\Lambda_0(\tau)}{n\sqrt{n}} w_n^2\gamma_n, \tag{58}$$

for a sequence of non-negative numbers $\gamma_n$, with probability larger than or equal to

$$1 - e^{-\frac{n\gamma_n^2}{2\theta^2 + 2\gamma_n y/3}} - P(\max_{1\leq i\leq n}\exp\{g(\mathbf{X}_i)\} > y)$$

for any truncation value $y$ satisfying (54).

Then, observe that for any three events $A_1, A_2, A_3$,

$$P(A_1) = P(A_1 \cap A_2) + P(A_1|A_2^c)P(A_2^c) \leq P(A_1 \cap A_2) + P(A_2^c)$$

and similarly $P(A_1 \cap A_2) \leq P(A_1 \cap A_2 \cap A_3) + P(A_3^c)$, leading to

$$P(A_1) \leq P(A_1 \cap A_2 \cap A_3) + P(A_2^c) + P(A_3^c).$$

Let $A_1 = \{|\upsilon_{jk}| \geq q_n\}$,$A_2 = \{|\boldsymbol{\Delta}\upsilon_{jk}| \leq \frac{w_n}{n}\}$ and $A_3 = \{\langle \boldsymbol{\Delta}\upsilon_{jk}\rangle_2 \leq \frac{\tau\Lambda_0(\tau)}{n\sqrt{n}} w_n^2 \gamma_n\}$. By large deviation inequality for martingales of bounded jumps and variation in Lemma 7, there exists a sequence of positive numbers $q_n$ such that

$$P\left(|\upsilon_{jk}| \geq q_n\right) \leq 2e^{-\frac{nq_n^2}{Kq_n + K_1^2}} + P\Big(|\boldsymbol{\Delta}\upsilon_{jk}| \geq \frac{w_n}{n}\Big) + P\Big(\langle \boldsymbol{\Delta}\upsilon_{jk}\rangle_2 \geq \frac{\tau\Lambda_0(\tau)}{n\sqrt{n}} w_n^2\gamma_n\Big).$$

By Lemma 2 and equations (57) and (58) we have

$$\begin{aligned} P\left(|\upsilon_{jk}| \geq q_n\right) &\leq 2e^{-\frac{nq_n^2}{Kq_n + K_1^2}} + \frac{3W^2}{8ed} e^{-\frac{nr_n^2 D^2}{u^2 e^{2m^*C}}} + e^{-\frac{nD^2}{2}} \\ &\quad + e^{-\frac{n\gamma_n^2}{2\theta^2 + 2\gamma_n y/3}} + P(\max_{1\leq i\leq n} \exp\{g(\mathbf{X}_i)\} > y)). \end{aligned}$$

for $K = w_n/n$ and $K_1^2 = \gamma_n w_n^2 \tau\Lambda_0(\tau)/n\sqrt{n}$. The choice of $\gamma_n$ is driven by (53) where we considered $\gamma_n = M\sqrt{n}\lambda_n\rho'(0+)$ with $M = 1/(\tau\lambda_0(\tau)\Lambda_0(\tau)C)$. For a $q_n = \frac{1}{2}\lambda_n d^{1/\gamma_j^*}\rho'(0+)$, $Kq_n \leq K_1^2$ as long as

$$2\omega_n \geq C\lambda_0(\tau)d^{1/\gamma_j^*}.$$

With $w_n = cr_n + \sqrt{\frac{\log d}{nu^2}}$ the choice of $r_n = C\lambda_0(\tau)\sqrt{n}d^{1/\gamma_j^*}\sqrt{\frac{\log d}{u^2}}$, suffices to guarantee the above inequality. For such choices of $\omega_n, \gamma_n$ and $q_n$ we have

$$\begin{aligned} P\left(|\upsilon_{jk}| \geq \frac{1}{2}\lambda_n d^{1/\gamma_j^*}\rho'(0+)\right) &\leq 2e^{-\frac{nq_n^2}{2K_1^2}} + \frac{3W^2}{8ed} e^{-\frac{nr_n^2 D^2}{u^2 e^{2m^*C}}} \\ &\quad + e^{-\frac{nD^2}{2}} + e^{-\frac{n\gamma_n^2}{2\theta^2 + 2\gamma_n y/3}} + P(\max_{1\leq i\leq n} \exp\{g(\mathbf{X}_i)\} > y)). \end{aligned} \tag{59}$$

The right-hand-side of (59) can be simplified to

$$\begin{aligned} & e^{-\frac{n^2 C\lambda_n\rho'(0+)}{2\lambda_0(\tau)}} + \frac{3W^2}{8ed} e^{-\frac{n^2 C\lambda_0(\tau)D^2 d^{2/\gamma_j^*} \log d}{u^4 e^{2m^*C}}} + e^{-\frac{nD^2}{2}} \\ &\quad + e^{-\frac{n^2 M^2 \lambda_n^2 \rho'(0+)^2}{2\theta^2 + 2M\sqrt{n}\lambda_n\rho'(0+)y/3}} + P(\max_{1\leq i\leq n} \exp\{g(\mathbf{X}_i)\} > y)), \end{aligned}$$

which can be further bounded by

$$\leq \left(3 + \frac{3W^2}{8ed}\right) e^{-n^2 C_{\lambda_n,n,p,d}} + P(\max_{1\leq i\leq n} \exp\{g(\mathbf{X}_i)\} > y)),$$

for a constant $C_{\lambda_n,n,p,d}$ defined as

$$\min\left\{\frac{C\lambda_n\rho'(0+)}{2\lambda_0(\tau)}, \frac{C\lambda_0(\tau)D^2 d^{2/\gamma_j^*}\log d}{u^4 e^{2m^*C}}, \frac{D^2}{2n}, \frac{M^2\lambda_n^2\rho'(0+)^2}{2\theta^2 + 2M\sqrt{n}\lambda_n\rho'(0+)y/3}\right\}$$

Second, to control the $\nu$ term in (56), we observe that according to Lemma 2, there exists a constant $0 < D = P(Y(\tau) = 1) \leq 1$ such that for the $u$ as defined in Condition 1 (iii) we have

$$\sup_{t\in[0,\tau]} \|\mathbf{e}(\boldsymbol{\beta}^*,t)\|_\infty \le \frac{C\sup_{t\in[0,\tau]} s^{(0)}(\boldsymbol{\beta}^*,t)}{D\exp\{-m^*C\}} \le Cu.$$

Thus, each $\nu_{jk}/u$ is a sum of a sequence of i.i.d bounded random variables. However, across $k$'s, i.e., group elements, $\nu_{jk}/u$ are not independent random variables. By Hoeffding's inequality,

$$P\left(\max_{1\le k\le d} |\nu_{jk}| \ge 2\|M\|_n Cut_n\right) \le 2e^{-nt_n^2},$$

where $\|M\|_n$ is proportional to $E\sqrt{\frac{1}{n}\sum_{i=1}^n M_i^2(\tau)}$. Because $\bar{M}$ is a bounded martingale, we can conclude that there exists a constant $c_1 > 0$ such that $\|M\|_n \le c_1$.

Hence, for $t_n = \lambda_n d^{1/\gamma_j^*}\rho'(0+)/4c_1Cu$ we obtain

$$P\left(\max_{1\le k\le d} |\nu_{jk}| \ge \frac{1}{2}\lambda_n d^{1/\gamma_j^*}\rho'(0+)\right) \le 2e^{-n\frac{\lambda_n^2 d^{2/\gamma_j^*}\rho'^2(0+)}{16c_1^2C^2u^2}}. \tag{60}$$

Utilizing (59) and (60) we obtain a bound on the size of the set $\mathcal{E}_n^c$ as follows,

$$\begin{aligned} P\left(\mathcal{E}_n^c\right) \quad \le \quad & 2pd\left(\max\left\{\left(3+\frac{3W^2}{8ed}\right)e^{-n^2C_{\lambda_n,n,p,d}}, e^{-n\frac{\lambda_n^2 d^{2/\gamma_j^*}\rho'^2(0+)}{16c_1^2C^2u^2}}\right\}\right. \\ & \left. + P(\max_{1\le i\le n}\exp\{g(\mathbf{X}_i)\} > y)\right). \end{aligned}$$

□

*Proof of Lemma 4.* Let $\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}} - \mathbf{b}$.

We consider two cases: (i) $4\lambda_n\sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*}\rho(\|\boldsymbol{\Delta}_j\|_{\gamma_j}) \ge \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|_{n,\mathbf{b}^*}^2$, and (ii) $4\lambda_n\sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*}\ \rho(\|\boldsymbol{\Delta}_j\|_{\gamma_j}) \le \|f_{\mathbf{b}} - f_{\boldsymbol{\beta}^*}\|_{n,\mathbf{b}^*}^2$.

**Case (i)** From (29), we have

$$\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|_{n,\mathbf{b}_{\widehat{\boldsymbol{\beta}}}}^2 + \lambda_n\sum_{j=1}^p d^{1/\gamma_j^*}\rho(\|\boldsymbol{\Delta}_j\|_{\gamma_j}) \le 8\lambda_n\sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*}\rho(\|\boldsymbol{\Delta}_j\|_{\gamma_j}).$$

This implies that $\sum_{j\in\mathcal{M}_*^c} d^{1/\gamma_j^*}\rho(\|\boldsymbol{\Delta}_j\|_{\gamma_j}) < 7\sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*}\rho(\|\boldsymbol{\Delta}_j\|_{\gamma_j})$ or that $\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}} - \mathbf{b} \in \mathbb{C}_{7,\rho}$ as defined in RE condition. For such $\boldsymbol{\Delta}$, from the RE condition in (14) we have with

$$\bar{d} = \sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*},$$

$$\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|_{n,\mathbf{b}_{\widehat{\boldsymbol{\beta}}}}^2 \le 8\lambda_n\sqrt{\bar{d}}\sqrt{\sum_{j\in\mathcal{M}_*}\rho^2(\|\boldsymbol{\Delta}_j\|_{\gamma_j})} \le 8\lambda_n\frac{\sqrt{\bar{d}}}{\zeta}\sqrt{\boldsymbol{\Delta}^T\bigtriangledown^2\mathcal{L}_n(\boldsymbol{\beta}^*)\boldsymbol{\Delta}}.$$

The left hand side can be further bounded using Proposition 2 and triangle inequality with

$$8\lambda_n \frac{\sqrt{\bar{d}}}{\zeta}\left(\exp\{a_{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\}\|f_{\boldsymbol{\beta}^*}-f_{\hat{\boldsymbol{\beta}}}\|_{n,\mathbf{b}_{\hat{\boldsymbol{\beta}}}}+\exp\{a_{\mathbf{b}-\boldsymbol{\beta}^*}\}\|f_{\mathbf{b}}-f_{\boldsymbol{\beta}^*}\|_{n,\mathbf{b}^*}\right).$$

Furthermore, with the simple inequality $ab \le b^2/2 + a^2/2$, we can further upper bound the left hand side above, to obtain

$$\begin{aligned}\|f_{\hat{\boldsymbol{\beta}}}-f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}_{\hat{\boldsymbol{\beta}}}} &\le 16\lambda_n^2\frac{\bar{d}}{\zeta^2}\exp\{2a_{\mathbf{b}-\boldsymbol{\beta}^*}\}+\|f_{\mathbf{b}}-f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*}\\ &+ 32\lambda_n^2\frac{\bar{d}}{\zeta^2}\exp\{2a_{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\}+\frac{1}{2}\|f_{\boldsymbol{\beta}^*}-f_{\hat{\boldsymbol{\beta}}}\|^2_{n,\mathbf{b}_{\hat{\boldsymbol{\beta}}}}.\end{aligned}$$

Combining all of the above we obtain

$$\begin{aligned}\|f_{\hat{\boldsymbol{\beta}}}-f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}_{\hat{\boldsymbol{\beta}}}} &\le 2\|f_{\mathbf{b}}-f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*}\\ &+ 64\lambda_n^2\frac{\bar{d}}{\zeta^2}\exp\{2a_{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}\}+32\lambda_n^2\frac{\bar{d}}{\zeta^2}\exp\{2a_{\mathbf{b}-\boldsymbol{\beta}^*}\}.\end{aligned}$$

To upper bound the LHS of the previous inequality we bound the two exponential terms independently.

First: let $b = \boldsymbol{\beta}^*$ in (29). Then, by using the RE condition and all equations above we obtain that $y_1 = \sum_{j\in\mathcal{M}^*} d^{1/\gamma_j^*}\rho(\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*\|_{\gamma_j}) \ge 0$ and $\upsilon_1 = \sum_{j\in\mathcal{M}^*}\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*\|_{\gamma_j} \ge 0$ are such that $a_{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*} \le C\upsilon_1$ and

$$y_1\exp\{-2C\upsilon_1\} \le 16\lambda_n^2\frac{\bar{d}}{\zeta^2}.$$

From convexity of $\rho$ we know that $\rho(\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*\|_{\gamma_j}) \ge \rho'(0+)\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*\|_{\gamma_j}$, hence $\upsilon_1, \upsilon_2$ satisfy $\upsilon_1 \ge \rho'(0+)\upsilon_2$. Combining all the above, $\upsilon_1$ solves

$$\upsilon_1\exp\{-2C\upsilon_1\} \le 16\lambda_n^2\rho'(0+)\frac{\bar{d}}{\zeta^2}. \tag{61}$$

Second, we consider the case of general $\mathbf{b}$ possibly different from $\boldsymbol{\beta}^*$. In such cases,

$$\|f_{\mathbf{b}}-f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*} \le 4\lambda_n\sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*}\rho(\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^*\|_{\gamma_j})+4\lambda_n\sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*}\rho(\|\mathbf{b}_j-\boldsymbol{\beta}_j^*\|_{\gamma_j})$$

Then, by utilizing Proposition 2 on the left and Caushy-Shwarz inequality to the right, we notice that

$$\begin{aligned}&\exp\{-2a_{\mathbf{b}-\boldsymbol{\beta}^*}\}\zeta^2\sum_{j\in\mathcal{M}_*}\rho^2(\|\mathbf{b}_j-\boldsymbol{\beta}_j^*\|_{\gamma_j})\\ &\le \|f_{\mathbf{b}}-f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*}\\ &\le 4\lambda_n\bar{d}\sqrt{\upsilon_2}+4\lambda_n\bar{d}\sqrt{\sum_{j\in\mathcal{M}_*}\rho^2(\|\mathbf{b}_j-\boldsymbol{\beta}_j^*\|_{\gamma_j})}.\end{aligned}$$

To that end, let us denote with $y_2 = \sum_{j\in\mathcal{M}_*}\rho^2(\|\mathbf{b}_j - \boldsymbol{\beta}^*_j\|_{\gamma_j}) \ge 0$ and $\upsilon_2 = \sum_{j\in\mathcal{M}_*}\|\mathbf{b}_j-\boldsymbol{\beta}^*_j\|^2_{\gamma_j}\ge 0$ and observe that $a_{\mathbf{b}-\boldsymbol{\beta}^*}\le C\upsilon_2$

$$\zeta^2 y_2 \exp\{-2C\upsilon_2\} - 4\lambda_n \bar{d}\sqrt{y_2} \le 4\lambda_n\bar{d}\sqrt{\upsilon_1}.$$

Utilizing the equation $\upsilon_1$ satisfies and the convexity of $\rho$ we have

$$\upsilon_2\exp\{-2C\upsilon_2\} - 4\lambda_n\frac{\bar{d}}{\zeta^2{\rho'}^2(0+)}\sqrt{\upsilon_2}\le 16\lambda_n^2\frac{\bar{d}^{3/2}}{{\rho'}^{3/2}(0+)\zeta^3}. \tag{62}$$

Although $\upsilon_2$ depends on $\mathbf{b}$, we observe that the previous inequality holds uniformly over $\mathbf{b}$ hence we have suppressed the dependence on $\mathbf{b}$ in the notation of $\upsilon_2$.

**Case (ii)** From (29), we have

$$\begin{aligned}\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}_{\widehat{\boldsymbol{\beta}}}} &\le 2\|f_{\mathbf{b}}-f_{\boldsymbol{\beta}^*}\|^2_{n,\mathbf{b}^*}\\ &\le 64\lambda_n^2\frac{\bar{d}}{\zeta^2}\exp\{2C\upsilon_1\}+32\lambda_n^2\frac{\bar{d}}{\zeta^2}\exp\{2C\upsilon_2\}+2\min_{\mathbf{b}\in\mathcal{B}}\|g-f_{\mathbf{b}}\|^2.\end{aligned}$$

□

*Proof of Lemma 5.* Following the same steps as in the proof of Lemma 4, we obtain easily that $\boldsymbol{\Delta}\in\mathbb{C}_3$ for $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*$ (exact steps are omitted). Combined with assumption RE$(7, s, \boldsymbol{\gamma})$ (14), it leads to

$$\|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\widehat{\boldsymbol{\beta}}^*} \le 32\frac{\lambda_n^2}{\zeta^2}e^{2\upsilon_1}\sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}, \tag{63}$$

for $0\le\upsilon_1\le 1$ satisfying (16). This result gives a preliminary step towards the final statement. The right- hand side is a complicated random norm (introduced in (20)). The rest of the proof establishes tight non-trivial lower bounds on its size. Together with Proposition 2, we have

$$\begin{aligned} & 32\frac{\lambda_n^2}{\zeta^2}e^{2\upsilon_1}\sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*} \ge \|f_{\widehat{\boldsymbol{\beta}}} - f_{\boldsymbol{\beta}^*}\|^2_{n,\widehat{\boldsymbol{\beta}}^*}\\ = \quad & \frac{\int_0^\tau(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^T\mathbf{V}_n(\mathbf{b}_{\widehat{\boldsymbol{\beta}}},t)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)d\bar{N}(t)}{\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}_*}-\boldsymbol{\beta}^*_{\mathcal{M}_*}\|^2_{1,\boldsymbol{\gamma}}}\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}_*}-\boldsymbol{\beta}^*_{\mathcal{M}_*}\|^2_{1,\boldsymbol{\gamma}}\\ \ge \quad & e^{-2a_{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*}}\zeta^2\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}_*}-\boldsymbol{\beta}^*_{\mathcal{M}_*}\|^2_{1,\boldsymbol{\gamma}}, \end{aligned} \tag{64}$$

where we used the notation $\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}_*}-\boldsymbol{\beta}^*_{\mathcal{M}_*}\|^2_{1,\boldsymbol{\gamma}} = \sum_{j\in\mathcal{M}_*}\|\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}^*_j\|^2_{\gamma_j}$, and

$$a_{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*} = \max_{1\le q,i\le n}|(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^T\Psi(\mathbf{X}_i)-\Psi(\mathbf{X}_q)|\le 2C\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1.$$

The rest of the proof is based on the analysis of the upper bound for the norm $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1$. The goal is to first find the worst case upper bound that satisfies (64). Therefore, the desired upper bound is the optimal solution of the following optimization problem

$$\begin{aligned} \max \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & e^{-\|\mathbf{x}\|_1}\|\mathbf{x}\|_{1,\gamma} \le z, \end{aligned}$$

for $z = 16\frac{\lambda_n^2}{\zeta^4}e^{2v_1}\sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}$. Because $\|\mathbf{x}\|_{1,\gamma} \ge d^{-1}\|\mathbf{x}\|_1$, the optimal value of the previous problem is upper bounded by the optimal value of the following problem

$$\begin{aligned} \max \quad & u \\ \text{s.t.} \quad & e^{-u}u \le zd, \\ & u \ge 0. \end{aligned}$$

Function $e^{-u}u$ is neither convex nor concave. It is concave up to $u = 2$ and then convex with exponentially rate of convergence towards zero. When $zd > 1/e$, the optima is reached at $u = 1$. When $zd < 2e^{-2}$, $u \to \infty$ exponentially fast. Thus, for $\lambda_n$ satisfying

$$e^{-1}\zeta^4 \le 32\lambda_n^2 e^{2v_1} d \sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*},$$

we have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le 1$. Under such conditions for some constant $c_0 > 1$, $a_{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*} \le 2C$, and we utilize (64) to conclude

$$\|\widehat{\boldsymbol{\beta}}_{\mathcal{M}_*} - \boldsymbol{\beta}^*_{\mathcal{M}_*}\|^2_{1,\boldsymbol{\gamma}} \le 32e^{2C+2v_1}\frac{\lambda_n^2}{\zeta^4}\sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}.$$

From Cauchy Schwartz inequality, we have that $\sum_{j\in\mathcal{M}_*} d^{1/\gamma_j^*}\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_{\gamma_j}$ is less than or equal to $\sqrt{\sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}}\sqrt{\sum_{j\in\mathcal{M}_*}\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|^2_{\gamma_j}}$ which is according to inequality above upper bounded with $4e^C\frac{\lambda_n}{\zeta^2}\sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}$. Knowing that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \in \mathbb{C}_3$ and using the convexity of $\rho$, we have $\|\rho(\widehat{\boldsymbol{\beta}}_{\mathcal{M}_*^c} - \boldsymbol{\beta}^*_{\mathcal{M}_*^c})\|_1 \le 3\|\rho(\widehat{\boldsymbol{\beta}}_{\mathcal{M}_*} - \boldsymbol{\beta}^*_{\mathcal{M}_*})\|_1$ and thus

$$\sum_{j=1}^p d^{1/\gamma_j^*}\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_{\gamma_j} \le 16\sqrt{2}e^{C+v_1}\frac{\lambda_n}{\zeta^2}\sum_{j\in\mathcal{M}_*} d^{2/\gamma_j^*}.$$

□

*Proof of Lemma 6*. Let the event $\mathcal{T}_n$ be defined as

$$\{\|\tilde{\mathbf{h}}_{n,j}(\boldsymbol{\beta}^*)\|_{\gamma_j^*} \le 2\lambda_n \max\{d^{1/\gamma_j^*}\sqrt{d}\}\min_{1\le j\le p}\lambda_{\min}(\mathbf{R}_j)\rho'(0+), \forall j \in \{1,\dots,p\}\},$$

with $\tilde{\mathbf{h}}_{n,j}(\boldsymbol{\beta}^*) = -\frac{1}{n}\sum_{i=1}^n\int_0^\tau(\widetilde{\mathbf{E}}_{n,j}(\boldsymbol{\beta}^*,t) - \mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij}))dM_i(t)$,

$$\begin{aligned} &\widetilde{\mathbf{E}}_{n,j}(\boldsymbol{\beta}^*,t) \\ &\quad = \frac{1}{n}\sum_{i=1}^n \frac{Y_i(t)\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij})}{\frac{1}{n}\sum_{l=1}^n Y_l(t)\exp\{\sum_{j=1}^p \boldsymbol{\beta}_j^{*T}\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{lj})\}}\exp\{\sum_{j=1}^p \boldsymbol{\beta}_j^{*T}\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij})\} \end{aligned}$$

We first adapt the results of Lemma 1 with the following few steps

$$\begin{aligned}&-\mathbf{b}_j^T\tilde{\mathbf{h}}_{n,j}(\boldsymbol{\beta}^*)+\lambda_n\sqrt{d}\rho\left(\|\mathbf{b}_j\|_{\gamma_j}+\|\mathbf{b}_j\|_2\right)\\ \geq\ &\|\mathbf{b}_j\|_{\gamma_j}\left(-\|\tilde{\mathbf{h}}_{n,j}(\boldsymbol{\beta}^*)\|_{\gamma_j^*}+\lambda_n\sqrt{d}\rho'(0+)\left(1+\frac{\|\mathbf{b}_j\|_2}{\|\mathbf{b}_j\|_{\gamma_j}}\right)\right).\end{aligned}$$

For $\gamma_j \geq 2$, we know that $\|\mathbf{b}_j\|_{\gamma_j} \leq \|\mathbf{b}_j\|_2$. This relation leads to the conclusion that previous quantity is lower bounded with

$$\geq \|\mathbf{b}_j\|_{\gamma_j}\left(-\|\mathbf{h}_{n,j}(\boldsymbol{\beta}^*)\|_{\gamma_j^*}+2\lambda_n\sqrt{d}\rho'(0+)\right),$$

which leads us to conclude that the results of Lemma 1 hold for this particular penalty. Size of the set $\mathcal{T}_n$ is easily deducible by adapting the very last proof of Theorem 1 (exact details are omitted).

To prove equivalent results to those of Section 4, we need to define new constants corresponding to $a_{\mathbf{v}}$ and $\underline{\underline{\omega}}$. First, the equivalent of $\mathbf{V}_n(\mathbf{b})$ has extra $\mathbf{R}_j^{-1}$ terms, which will factor into f $a_i$ terms (of Proposition 2) as $(\mathbf{b}-\boldsymbol{\beta}^*)(\mathbf{R}^{-1}\boldsymbol{\Psi}(\mathbf{X}_i)-\mathbf{E}_n(\boldsymbol{\beta}^*,t))$. $\mathbf{R}$ is a diagonal block matrix

$$\mathbf{R}=\begin{pmatrix}\mathbf{R}_1 & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_2 & \cdots & \mathbf{0}\\ \vdots & & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{R}_p\end{pmatrix}.$$

Second, as $\bar{\mathbf{v}}_j=\mathbf{v}_j\mathbf{R}_j^{-1}$

$$\bar{a}_{\mathbf{v}}=\max_{1\leq i,q\leq n}\left|\bar{\mathbf{v}}^T(\boldsymbol{\Psi}(\mathbf{X}_i)-\boldsymbol{\Psi}(\mathbf{X}_q))\right|\leq\max_{1\leq j\leq p}r(\mathbf{R}_j^{-1})a_{\mathbf{v}}$$

with spectral radius

$$r(\mathbf{R}_j^{-1})=\max_{k=1,\ldots,d}|\lambda_k(\mathbf{R}_j^{-1})|=\max_{k=1,\ldots,d}|\lambda_k(\mathbf{R}_j)|^{-1}=\lambda_{\min}^{-1}(\mathbf{R}_j).$$

Then, $\bar{a}_{\mathbf{v}} \leq \max_{1\leq j\leq p}\lambda_{\min}^{-1}(\mathbf{R}_j)a_{\mathbf{v}}$. Thus, the result of Proposition 2 follows with $\eta$ equal to $\max_{1\leq j\leq p}\lambda_{\min}^{-1}(\mathbf{R}_j)a_{\mathbf{v}}$.

The definition of the weights, $\omega_i(\mathbf{b})$, in the proof of Proposition 3 will be changed to address the new weighting matrix, $\mathbf{R}_j$,. Once they are redefined with

$$\underline{\underline{\omega}}_{\mathbb{S}}:=\min_{i\in\{1,\ldots,n\},i\in\cup_{q=1}^n\mathbb{R}_q}\left\{\frac{\sum_{q=1}^N\exp\{\sum_{j=1}^p\boldsymbol{\beta}_j^{*T}\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{ij})\}1\{i\in\mathbb{R}_q\}\}}{\sum_{l\in\mathbb{R}_q}\exp\{\sum_{j=1}^p\boldsymbol{\beta}_j^{*T}\mathbf{R}_j^{-1}\boldsymbol{\Psi}(X_{lj})\}}\right\},$$

the exact steps of the proof of Proposition 3 will follow easily, and thus we omit the details here. □

## References

[1] Aalen, O. O. (1980), A model for nonparametric regression analysis of counting processes, In *Lecture Notes in Statistics*, **2**, N. Klonecki, A. Kosek and J. Rosinski, eds., 1–25. MR0577267

[2] Andersen, P. K. and Gill, R. D. (1982), Cox's regression model for counting processes: A large sample study, *Annals of Statistics*, **10**(4), 1100–1120. MR0673646

[3] Bach, F. (2010), Self-concordant analysis for logistic regression, *Electronic Journal of Statistics*, **4**, 384–414. MR2645490

[4] Bickel, P., Ritov, Y. and Tsybakov, A. (2009), Simultaneous analysis of Lasso and Dantzig selector, *The Annals of Statistics*, **37**, 1705–1732. MR2533469

[5] Bradic, J., Fan, J. and Jiang, J. (2011), Regularization for Cox proportional hazards model with NP dimensionality, *The Annals of Statistics*, **36**(9), 3092–3120. MR3012402

[6] Bühlmann, P. and van de Geer, S. (2011), Statistics for High Dimensional Data, Springer, p. 556. MR2807761

[7] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007), Aggregation for Gaussian regression, *The Annals of Statistics*, **35**(4), 1674–1697. MR2351101

[8] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007), Sparsity oracle inequalities for the Lasso, *Electron. J. Statist.*, **1**, 169–194. MR2312149

[9] Cox, D. R. (1972), Regression model and life tables (with discussion), *Journal of the Royal Statistical Society, Series B*, **34**, 187–220. MR0341758

[10] Cox, D. R. (1975), Partial likelihood, *Biometrika*, **62**, 269–276. MR0400509

[11] Fleming, T. R. and Harrington, D. P. (2005), Counting Processes and Survival Analysis, JohnWiley & Sons. MR1100924

[12] Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of American Statistical Association*, **96**(456), 1348–1360. MR1946581

[13] Fan, X., Grama, I. and Liu, Q. (2012), Hoefding's inequality for supermartingales, arXiv:1109.4359v5.

[14] Gaïffas, S. and Guilloux, A. (2012), High dimensional additive Hazards models and the Lasso, *Electronic Journal of Statistics*, **6**, 522–546. MR2988418

[15] Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013), Oracle inequalities for the Lasso in the Cox model, *Annals of Statistics*, **41**(3), 1055–1692. MR3113806

[16] Jacobsen, M. (1989), Right censoring and martingale methods for failure time data, *The Annals of Statistics*, **17**(3), 1133–1156. MR1015142

[17] Kong, S. and Nan, B. (2012), Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso, *Forthcoming in Statistica Sinica*. MR3184591

[18] Lai, T. L. and Ying, Z. (1992), Asymptotically efficient estimation in censored and truncated regression models, *Statistica Sinica*, **2**, 17–46. MR1152296

[19] Landrum, L., Java, J., Mathews, C. A., Lanneau Jr., G. S., Copeland, L. J., Armstrong, D. K. and Walker, J. L. (2013), Prognostic factors for stage III epithelial ovarian cancer treated with intraperitoneal chemotherapy: A gynecologic oncology group study, *Gynecologic Oncology*, **130**(1), 12–18.

[20] Letue, F. (2000), Modele de Cox: Estimation par selection de modele et modele de chocs bivarie. PhD thesis.

[21] Lemler, S. (2012), Oracle inequalities for the Lasso for the conditional hazard rate in a high-dimensional setting, arXiv:1206.5628.

[22] Lin, D. Y. and Ying, Z. (1994), Semiparametric analysis of the additive risk model, *Biometrika*, **81**(1), 61–71. MR1279656

[23] Lin, Y. and Zhang, H. H. (2006), Component selection and smoothing in multivariate nonparametric regression, *Ann. Statist.*, **34**(5), 2272–2297. MR2291500

[24] Lounici, K., Pontil, M., Tsybakov, A. B. and van de Geer, S. (2011), Oracle inequalities and optimal inference under group sparsity, *The Annals of Statistics*, **39**(4), 2164–2204. MR2893865

[25] Lunn, M. and McNeil, D. (1995), Applying Cox regression to competing risks, *Biometrics*, **51**(2), 524–532.

[26] Martins-Filho, A., Jammal, M. P., Nomelini, R. S. and Murta, E. F. (2014), The immune response in malignant ovarian neoplasms, *European Journal of Gynaecological Oncology*, **35**(5), 487–491.

[27] Meier, L., van de Geer, S. and Bühlmann, L. (2009), High dimensional aditive modeling, *The Annals of Statistics*, **37**(6b), 3779–3821. MR2572443

[28] Meinshausen, N. and Yu, B. (2009), Lasso-type recovery of sparse representations for high-dimensional data, *The Annals of Statistics*, **37**, 246–270. MR2488351

[29] Müller, H. and Yao, F. (2008), Functional additive models, *J. Am. Statist. Ass.*, **103**, 1534–1544. MR2504202

[30] Negahban, D. and Wainwright, M. J. (2011), Simultaneous support recovery in high dimensions: Benefits and perils of block $l_1/l_\infty$-regularization, *IEEE Transactions on Information Theory*, **57**(6), 3841–3863. MR2817058

[31] Negahban, S., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, *Statistical Science*, **27**(4), 538–557. MR3025133

[32] Rigollet, P. (2012), Kullback-Leibler aggregation and misspecified generalized linear models, *The Annals of Statistics* (to appear), arXiv:0911.2919. MR2933661

[33] Taulbee, J. D. (1979), A general model for the hazard rate with covariates, *Biometrics*, **35**, 439–450.

[34] Tibshirani, R. (1997), The Lasso methods for variable selection in the Cox model, *Statistics in Medicine*, **16**, 385–395.

[35] Tibshirani, R. and Ciampi, A. (1983), A family of proportional- and additive-Hazards models for survival data, *Biometrics*, **39**(1), 141–147. MR0712745

[36] Tsybakov, A. B. (2003), Optimal rates of aggregation, In *COLT*, B. Scholkopf and M. K. Warmuth, eds., *Lecture Notes in Computer Science*, **2777**, 303–313.

[37] van de Geer, S. (1995), Exponential inequalities for martingales with application to maximum likelihood estimation for counting processes, *Annals of Statistics*, **23**(5), 1779–1801. MR1370307

[38] van de Geer, S. (2008), High-dimensional generalized linear models and the Lasso, *The Annals of Statistics*, **36**, 614–645. MR2396809

[39] Van der Vaart, A. W. and Wellner, J. A. (1996), Weak convergence and empirical processes, *Springer Series in Statistics*, Springer-Verlag, New York. MR1385671 (97g:60035)

[40] Wang, S., Nan, B., Zhou, N. and Zhu, J. (2009), Hierarchically penalized Cox regression with grouped variables, *Biometrika*, **96**(2), 307–322. MR2507145

[41] Wu, Y. (2012), Elastic net for Cox's proportional hazards model with a solution path algorithm, *Statistica Sinica*, **22**, 271–294. MR2933176

[42] Yuan, M. and Lin, L. (2006), Model selection and estimation in regression with grouped variables, *Journal of Royal Statistical Society B*, **68**(1), 49–67. MR2212574

[43] Zhao, P., Rocha, G. and Yu, B. (2009), The composite absolute penalties family for grouped and hierarchical variable selection, *The Annals of Statistics*, **37**(6a), 3468–3497. MR2549566

[44] Zhou, S. (2009), Restricted eigenvalue conditions on subgaussian random matrices. *Technical report, ETH Zurich.*

[45] Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320. MR2137327