ORACLE INEQUALITIES FOR THE LASSO IN THE COX MODEL

By Jian Huang[1], Tingni Sun, Zhiliang Ying[2], Yi Yu
and Cun-Hui Zhang[3]

University of Iowa, University of Pennsylvania, Columbia University, Fudan University and Rutgers University

We study the absolute penalized maximum partial likelihood estimator in sparse, high-dimensional Cox proportional hazards regression models where the number of time-dependent covariates can be larger than the sample size. We establish oracle inequalities based on natural extensions of the compatibility and cone invertibility factors of the Hessian matrix at the true regression coefficients. Similar results based on an extension of the restricted eigenvalue can be also proved by our method. However, the presented oracle inequalities are sharper since the compatibility and cone invertibility factors are always greater than the corresponding restricted eigenvalue. In the Cox regression model, the Hessian matrix is based on time-dependent covariates in censored risk sets, so that the compatibility and cone invertibility factors, and the restricted eigenvalue as well, are random variables even when they are evaluated for the Hessian at the true regression coefficients. Under mild conditions, we prove that these quantities are bounded from below by positive constants for time-dependent covariates, including cases where the number of covariates is of greater order than the sample size. Consequently, the compatibility and cone invertibility factors can be treated as positive constants in our oracle inequalities.

1. Introduction. The Cox (1972) proportional hazards model is widely used in the regression analysis of censored survival data, notably in identifying risk factors in epidemiological studies and treatment effects in clinical trials when the outcome variable is time to event. In a traditional biomedical study, the number of covariates p is usually relatively small as compared with the sample size n. Theoretical properties of the maximum partial likelihood estimator in the fixed p and large n setting have been well established. For example, Tsiatis (1981) proved the asymptotic normality of the maximum partial likelihood estimator. Andersen and

Gill (1982) formulated the Cox model in the context of the more general counting process framework and studied the asymptotic properties of the maximum partial likelihood estimator using martingale techniques. These results provide a solid foundation for the applications of the Cox model in a diverse range of problems where time to event is the outcome of interest.

In recent years, technological advancement has resulted in the proliferation of massive high-throughput and high-dimensional genomic data in studies that attempt to find genetic risk factors for disease and clinical outcomes, such as the age of disease onset or time to death. Finding genetic risk factors for survival is fundamental to modern biostatistics, since survival is an important clinical endpoint. However, in such problems, the standard approach to the Cox model is not applicable, since the number of potential genetic risk factors is typically much larger than the sample size. In addition, traditional variable selection methods such as subset selection are not computationally feasible when $p \gg n$.

The ℓ_1-penalized least squares estimator, or the lasso, was introduced by Tibshirani (1996). In the wavelet setting, the ℓ_1-penalized method was introduced by Chen, Donoho and Saunders (1998) as basis pursuit. Since then, the lasso has emerged as a widely used approach to variable selection and estimation in sparse, high-dimensional statistical problems. It has also been extended to the Cox model [Tibshirani (1997)]. Gui and Li (2005) implemented the LARS algorithm [Efron et al. (2004)] to approximate the lasso in the Cox regression model and applied their method to survival data with microarray gene expression covariates. Their work demonstrated the effectiveness of the lasso for variable selection in the Cox model in a $p \gg n$ setting.

There exists a substantial literature on the lasso and other penalized methods for survival models with a fixed number of covariates p. Zhang and Lu (2007) considered an adaptive lasso for the Cox model and showed that, under certain regularity conditions and with a suitable choice of the penalty parameter, their method possesses the asymptotic oracle property when the maximum partial likelihood estimator is used as the initial estimator. Fan and Li (2002) proposed the use of the smoothly clipped absolute deviation (SCAD) penalty [Fan (1997), Fan and Li (2001)] for variable selection and estimation in the Cox model which may include a frailty term. With a suitable choice of the penalty parameter, they showed that a local maximizer of the penalized log-partial likelihood has an asymptotic oracle property under certain regularity conditions on the Hessian of the log-partial likelihood and the censoring mechanism.

Extensive efforts have been focused upon the analysis of regularization methods for variable selection in the $p \gg n$ setting. In particular, considerable progress has been made in theoretical understanding of the lasso. However, most results are developed in the linear regression model. Greenshtein and Ritov (2004) studied the prediction performance of the lasso in high-dimensional linear regression. Meinshausen and Bühlmann (2006) showed that, for neighborhood selection in

the Gaussian graphical model, under a neighborhood stability condition and certain additional regularity conditions, the lasso is consistent even when $p/n \to \infty$. Zhao and Yu (2006) formalized the neighborhood stability condition in the context of linear regression as a strong irrepresentable condition on the design matrix. Oracle inequalities for the prediction and estimation error of the lasso was developed in Bunea, Tsybakov and Wegkamp (2007), Zhang and Huang (2008), Meinshausen and Yu (2009), Bickel, Ritov and Tsybakov (2009), Zhang (2009) and Ye and Zhang (2010), among many others.

A number of papers analyzed penalized methods beyond linear regression. Fan and Peng (2004) established oracle properties for a local solution of concave penalized estimator in a general setting with $n \gg p \to \infty$. van de Geer (2008) studied the lasso in high-dimensional generalized linear models (GLM) and obtained prediction and ℓ_1 estimation error bounds. Negahban et al. (2009) studied penalized M-estimators with a general class of regularizers, including an ℓ_2 error bound for the lasso in GLM under a restricted convexity and other regularity conditions. Bradic, Fan and Jiang (2011) made significant progress by extending the results of Fan and Li (2001) to a more general class of penalties in the Cox regression model with large p under different sets of regularity conditions. Huang and Zhang (2012) studied weighted absolute penalty and its adaptive, multistage application in GLM.

In view of the central role of the Cox model in survival analysis, its widespread applications and the proliferation of $p \gg n$ data, it is of great interest to understand the properties of the related lasso approach. The main goal of the present paper is to establish theoretical properties for the lasso in the Cox model when $p \gg n$. Specifically, we extend certain basic inequalities from linear regression to the case of the Cox regression. We generalize the compatibility and cone invertibility factors from the linear regression model and establish oracle inequalities for the lasso in the Cox regression model in terms of these factors at the true parameter value. Moreover, we prove that the compatibility and cone invertibility factors can be treated as constants under mild regularity conditions.

A main feature of our results is that they are derived under the more general counting process formulation of the Cox model with potentially a larger number of time-dependent covariates than the sample size. This formulation is useful because it "permits a regression analysis of the intensity of a recurrent event allowing for complicated censoring patterns and time-dependent covariates" [Andersen and Gill (1982)].

A second main feature of our results is that the regularity conditions on the counting processes and time-dependent covariates are directly imposed on the compatibility and cone invertibility factors of the Hessian of the negative log-partial likelihood evaluated at the true regression coefficients. Under such regularity conditions, the lasso estimator is proven to live in a neighborhood where the ratio between the estimated and true hazards is uniformly bounded away from zero and infinity. This allows unbounded and near zero ratios between the true and

baseline hazards. Our analysis can be also used to prove oracle inequalities based on the restricted eigenvalue. However, since the compatibility and cone invertibility factors are greater than the corresponding restricted eigenvalue [van de Geer and Bühlmann (2009), Ye and Zhang (2010)], the presented results are sharper.

A third main feature of our results is that the compatibility and cone invertibility factors used, and the smaller corresponding restricted eigenvalue, are proven to be greater than a fixed positive constant under mild conditions on the counting processes and time-dependent covariates, including cases where $p \gg n$. In the Cox regression model, the Hessian matrix is based on weighted averages of the cross-products of time-dependent covariates in censored risk sets, so that the compatibility and cone invertibility factors and the restricted eigenvalue are random variables even when they are evaluated for the Hessian at the true regression coefficients. Under mild conditions, we prove that these quantities are bounded from below by positive constants as certain truncated population versions of them. Thus the compatibility and cone invertibility factors can be treated as constants in our oracle inequalities.

The main results of this paper and the analytical methods used for deriving them are identical to those in its predecessor submitted in November 2011, with Section 4 as an exception. The difference in Section 4 is that the lower bound for the compatibility and cone invertibility factors and the restricted eigenvalue is improved to allow time-dependent covariates.

During the revision process of our paper, we became aware of a number of papers on hazard regression with censored data. Kong and Nan (2012) took an approach of van de Geer (2008) to derive prediction and ℓ_1 error bounds for the lasso in the Cox proportional hazards regression under a quite different set of conditions from us. For example, they required an ℓ_1 bound on the regression coefficients to guarantee a uniformly bounded ratio between hazard functions under consideration. Lemler (2012) considered the joint estimation of the baseline hazard function and regression coefficients in the Cox model. As a result, Lemler's (2012) error bounds for regression coefficients are of greater order than ours when the intrinsic dimension of the unknown baseline hazard function is of greater order than the number of nonzero regression coefficients. Gaïffas and Guilloux (2012) considered a quadratic loss function in place of a negative log-likelihood function in an additive hazards model. A nice feature of the additive hazards model is that the quadratic loss actually produces unbiased linear estimation equations so that the analysis of the lasso is similar to that of linear regression. The oracle inequalities in these three papers and ours can be all viewed as nonasymptotic. Unlike our paper, none of these three papers consider time-dependent covariates or constant lower bounds of the restricted eigenvalue or related key factors for the analysis of the lasso.

The rest of this paper is organized as follows. In Section 2 we provide basic notation and model specifications. In Section 3 we develop oracle inequalities for the lasso in the Cox model. In Section 4 we study the compatibility and cone

invertibility factors and the corresponding restricted eigenvalue of the Hessian of the log-partial likelihood in the Cox model. In Section 5 we make some additional remarks. All proofs are provided either right after the statement of the result or deferred to the Appendix.

2. Cox model with the ℓ_1 penalty. Following Andersen and Gill (1982), consider an n-dimensional counting process $\mathbf{N}^{(n)}(t) = (N_1(t), \ldots, N_n(t))$, $t \geq 0$, where $N_i(t)$ counts the number of observed events for the ith individual in the time interval $[0, t]$. The sample paths of $N_1, \ldots, N_n$ are step functions, zero at $t = 0$, with jumps of size $+1$ only. Furthermore, no two components jump at the same time. For $t \geq 0$, let $\mathcal{F}_t$ be the σ-filtration representing all the information available up to time t. Assume that for $\{\mathcal{F}_t, t \geq 0\}$, $\mathbf{N}^{(n)}$ has predictable compensator $\mathbf{\Lambda}^{(n)} = (\Lambda_1, \ldots, \Lambda_n)$ with

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

where $\boldsymbol{\beta}^o$ is a p-vector of true regression coefficients, Λ_0 is an unknown baseline cumulative hazard function and, for each i, $Y_i(t) \in \{0, 1\}$ is a predictable at risk indicator process that can be constructed from data, and $\mathbf{Z}_i(t) = (Z_{i,1}(t), \ldots, Z_{i,p}(t))'$ is a p-dimensional vector-valued predictable covariate process. In this setting the σ-filtration can be the natural $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), \mathbf{Z}_i(s); s \leq t, i = 1, \ldots, n\}$ or a richer one. We are interested in the problem of variable selection in sparse, high-dimensional settings where p, the number of possible covariates, is large, but the number of important covariates is relatively small.

2.1. *Maximum partial likelihood estimator with ℓ_1 penalty.* Define logarithm of the Cox partial likelihood for survival experience at time t,

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

where $\bar{N} = \sum_{i=1}^n N_i$. The log-partial likelihood function is

$$C(\boldsymbol{\beta}, \infty) = \lim_{t \to \infty} C(\boldsymbol{\beta}, t).$$

Let $\ell(\boldsymbol{\beta}) = -C(\boldsymbol{\beta}; \infty)/n$. The maximum partial likelihood estimator is the value that minimizes $\ell(\boldsymbol{\beta})$.

An approach to variable selection in sparse, high-dimensional settings for the Cox model is to minimize an ℓ_1-penalized negative log-partial likelihood criterion,

$$(2.2) \qquad L(\boldsymbol{\beta}; \lambda) = \ell(\boldsymbol{\beta}) + \lambda|\boldsymbol{\beta}|_1$$

[Tibshirani (1997)], where $\lambda \geq 0$ is a penalty parameter. Henceforth, we use notation $|\boldsymbol{\beta}|_q = \{\sum_{i=1}^p |\beta_i|^q\}^{1/q}$ for $1 \leq q < \infty$, $|\boldsymbol{\beta}|_\infty = \max_{1 \leq i \leq p} |\beta_i|$ and $|\boldsymbol{\beta}|_0 = \#\{j : \beta_j \neq 0\}$. For a given λ, the ℓ_1-penalized maximum partial likelihood estimator, or the lasso estimator hereafter, is defined as

$$(2.3) \qquad \hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \lambda).$$

2.2. *The Karush–Kuhn–Tucker conditions.* The lasso estimator can be characterized by the Karush–Kuhn–Tucker (KKT) conditions. Since the log-partial likelihood belongs to an exponential family, $\ell(\boldsymbol{\beta})$ must be convex in $\boldsymbol{\beta}$ and so is $L(\boldsymbol{\beta}; \lambda)$. It follows that a vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ is a solution to (2.3) if and only if the following KKT conditions hold:

$$
(2.4) \qquad \begin{cases} \dot{\ell}_j(\hat{\boldsymbol{\beta}}) = -\lambda \operatorname{sgn}(\hat{\beta}_j), & \text{if } \hat{\beta}_j \neq 0, \\ |\dot{\ell}_j(\hat{\boldsymbol{\beta}})| \leq \lambda, & \text{if } \hat{\beta}_j = 0, \end{cases}
$$

where $\dot{\ell}(\boldsymbol{\beta}) = (\dot{\ell}_1(\boldsymbol{\beta}), \ldots, \dot{\ell}_p(\boldsymbol{\beta}))' = \partial \ell(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ is the gradient of $\ell(\boldsymbol{\beta})$. The necessity and sufficiency of (2.4) can be proved by subdifferentiation of the convex penalized loss (2.2). This does not require strict convexity.

The KKT conditions indicate that the lasso in the Cox regression model may be analyzed in a similar way to the lasso in linear regression. As can be seen in the subsequent developments, such analysis can be carried out by proving that $|\dot{\ell}(\boldsymbol{\beta}^o)|_\infty$ is sufficiently small and the Hessian of $\ell(\boldsymbol{\beta})$ does not vanish for a sparse $\boldsymbol{\beta}$ at the true $\boldsymbol{\beta} = \boldsymbol{\beta}^o$. The (local) martingales for the counting process will play a crucial role to ensure that these requirements are satisfied.

2.3. *Additional notation.* Since the Λ_i are compensators,

$$
M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\mathbf{Z}_i'(s)\boldsymbol{\beta}^o) \, d\Lambda_0(s), \qquad 1 \leq i \leq n, t \geq 0,
$$

are (local) martingales with predictable variation/covariation processes

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

For a vector v, let $v^{\otimes 0} = 1 \in \mathbb{R}$, $v^{\otimes 1} = v$ and $v^{\otimes 2} = vv'$. Define

$$
S^{(k)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i^{\otimes k}(t) Y_i(t) e^{\mathbf{Z}_i'(t)\boldsymbol{\beta}}, \qquad k = 0, 1, 2,
$$

$$
R_n(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n Y_i(t) e^{\mathbf{Z}_i'(t)\boldsymbol{\beta}}, \qquad \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}) = \frac{S^{(1)}(t, \boldsymbol{\beta})}{S^{(0)}(t, \boldsymbol{\beta})},
$$

$$
V_n(t, \boldsymbol{\beta}) = \sum_{i=1}^n w_{ni}(t, \boldsymbol{\beta}) \big(\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta})\big)^{\otimes 2} = \frac{S^{(2)}(t, \boldsymbol{\beta})}{S^{(0)}(t, \boldsymbol{\beta})} - \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta})^{\otimes 2},
$$

where $w_{ni}(t, \boldsymbol{\beta}) = Y_i(t) \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta}) / [nS^{(0)}(t, \boldsymbol{\beta})]$. By differentiation and rearrangement of terms, it can be shown as in Andersen and Gill (1982) that the gradient of $\ell(\boldsymbol{\beta})$ is

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

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

$$(2.6) \qquad \ddot{\ell}(\boldsymbol{\beta}) \equiv \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}'} = \frac{1}{n} \int_0^\infty V_n(s, \boldsymbol{\beta}) \, d\bar{N}(s).$$

3. Oracle inequalities. In this section, we derive oracle inequalities for the estimation error of lasso in the Cox regression model. Let $\boldsymbol{\beta}^o$ be the vector of true regression coefficients, and define $\mathcal{O} = \{j : \beta_j^o \neq 0\}$, $\mathcal{O}^c = \{j : \beta_j^o = 0\}$ and $d_o = |\mathcal{O}|$, where $|\mathcal{U}|$ for a set $\mathcal{U}$ denotes its cardinality.

Making use of the KKT conditions (2.4), we first develop a basic inequality involving the symmetric Bregman divergence and ℓ_1 estimation error in the support $\mathcal{O}$ of $\boldsymbol{\beta}^o$ and its complement. The symmetric Bregman divergence, defined as

$$D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\dot{\ell}(\hat{\boldsymbol{\beta}}) - \dot{\ell}(\boldsymbol{\beta}))$$

can be viewed as symmetric, partial Kullback–Leibler distance between the partial likelihood at $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. Thus, $D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ can be viewed as a measure of prediction performance. The basic inequality, given in Lemma 3.1 below, serves as a vehicle for establishing the desired oracle inequalities.

LEMMA 3.1. *Let $\hat{\boldsymbol{\beta}}$ be defined as in (2.3), $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o$ and $z^* = |\dot{\ell}(\boldsymbol{\beta}^o)|_\infty$. Then the following inequalities hold*:

$$(3.1) \qquad (\lambda - z^*)|\tilde{\boldsymbol{\theta}}_{\mathcal{O}^c}|_1 \leq D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) + (\lambda - z^*)|\tilde{\boldsymbol{\theta}}_{\mathcal{O}^c}|_1 \leq (\lambda + z^*)|\tilde{\boldsymbol{\theta}}_{\mathcal{O}}|_1,$$

where $\tilde{\boldsymbol{\theta}}_{\mathcal{O}}$ and $\tilde{\boldsymbol{\theta}}_{\mathcal{O}^c}$ denote the subvectors of $\tilde{\boldsymbol{\theta}}$ of components in $\mathcal{O}$ and $\mathcal{O}^c$, respectively. In particular, for any $\xi > 1$, $|\tilde{\boldsymbol{\theta}}_{\mathcal{O}^c}|_1 \leq \xi|\tilde{\boldsymbol{\theta}}_{\mathcal{O}}|_1$ in the event $z^* \leq (\xi - 1)/(\xi + 1)\lambda$.

It follows from Lemma 3.1 that in the event $z^* \leq (\xi - 1)/(\xi + 1)\lambda$, the estimation error $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o$ belongs to the cone

$$(3.2) \qquad \mathscr{C}(\xi, \mathcal{O}) = \{\mathbf{b} \in \mathbb{R}^p : |\mathbf{b}_{\mathcal{O}^c}|_1 \leq \xi|\mathbf{b}_{\mathcal{O}}|_1\}.$$

In linear regression, the invertibility of the Gram matrix in the same cone, expressed in terms of restricted eigenvalues and related quantities, has been used to control the estimation error of the lasso. In what follows, we prove that a direct extension of the compatibility and cone invertibility factors can be used to control the estimation error of the lasso in the Cox regression.

For the cone in (3.2) and a given $p \times p$ nonnegative-definite matrix $\bar{\boldsymbol{\Sigma}}$, define

$$(3.3) \qquad \kappa(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}) = \inf_{0 \neq \mathbf{b} \in \mathscr{C}(\xi, \mathcal{O})} \frac{d_o^{1/2}(\mathbf{b}'\bar{\boldsymbol{\Sigma}}\mathbf{b})^{1/2}}{|\mathbf{b}_{\mathcal{O}}|_1}$$

as the compatibility factor [van de Geer (2007), van de Geer and Bühlmann (2009)], and

$$(3.4) \qquad F_q(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}) = \inf_{0 \neq \mathbf{b} \in \mathscr{C}(\xi, \mathcal{O})} \frac{d_o^{1/q}\mathbf{b}'\bar{\boldsymbol{\Sigma}}\mathbf{b}}{|\mathbf{b}_{\mathcal{O}}|_1|\mathbf{b}|_q}$$

as the weak cone invertibility factor [Ye and Zhang (2010)]. These quantities are closely related to the restricted eigenvalue [Bickel, Ritov and Tsybakov (2009), Koltchinskii (2009)],

$$
(3.5) \qquad \text{RE}(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}) = \inf_{0 \neq \mathbf{b} \in \mathscr{C}(\xi, \mathcal{O})} \frac{(\mathbf{b}' \bar{\boldsymbol{\Sigma}} \mathbf{b})^{1/2}}{|\mathbf{b}|_2}.
$$

In linear regression, the Hessian of the squared loss $|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|_2^2/(2n)$ is taken as $\bar{\boldsymbol{\Sigma}}$, and the oracle inequalities established in the papers cited in the above paragraph can be summarized as follows: in the event $z^* = |\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^o)/n|_\infty \leq \lambda(\xi - 1)/(\xi + 1)$,

$$
(3.6) \quad |\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)|_2^2/n \leq \frac{4(1 + 1/\xi)^{-2}\lambda^2 d_o}{\kappa^2(\xi, \mathcal{O}; \mathbf{X}'\mathbf{X}/n)}, \qquad |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_1 \leq \frac{2\xi d_o \lambda}{\kappa^2(\xi, \mathcal{O}; \mathbf{X}'\mathbf{X}/n)}
$$

and

$$
(3.7) \qquad
\begin{aligned}
|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_2 &\leq \frac{2(1 + 1/\xi)^{-1} d_o^{1/2} \lambda}{\text{RE}^2(\xi, \mathcal{O}; \mathbf{X}'\mathbf{X}/n)}, \\
|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_q &\leq \frac{2(1 + 1/\xi)^{-1} d_o^{1/q} \lambda}{F_q(\xi, \mathcal{O}; \mathbf{X}'\mathbf{X}/n)}, \qquad q \geq 1.
\end{aligned}
$$

In the Cox regression model, we still take the Hessian of the log-partial likelihood as $\bar{\boldsymbol{\Sigma}}$, in fact the Hessian at the true $\boldsymbol{\beta}^o$, so that (3.3) and (3.4) become

$$
(3.8) \qquad \kappa(\xi, \mathcal{O}) = \kappa(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o)), \qquad F_q(\xi, \mathcal{O}) = F_q(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o)).
$$

The reason for using these factors is that they yield somewhat sharper oracle inequalities than the restricted eigenvalue. It follows from $|\mathbf{b}_{\mathcal{O}}|_1 \leq d_o^{1/2}|\mathbf{b}|_2$ that $F_2(\xi, \mathcal{O}) \geq \kappa(\xi, \mathcal{O}) \text{RE}(\xi, \mathcal{O})$ and $\kappa(\xi, \mathcal{O}) \geq \text{RE}(\xi, \mathcal{O})$. Therefore, the first inequality of (3.7) is subsumed by the second with $q = 2$. Moreover, it is possible to have $\kappa(\xi, \mathcal{O}) \gg \text{RE}(\xi, \mathcal{O})$ [van de Geer and Bühlmann (2009)], and consequently, the ℓ_2 error bound based on the cone invertibility factor may be of sharper order that the one based on the restricted eigenvalue.

The following theorem extends (3.6) and (3.7) from the linear regression model to the proportional hazards regression model with

$$
(3.9) \qquad \max_{i < i' \leq n} \sup_{0 \leq t < \infty} \max_{j \leq p} |Z_{i,j}(t) - Z_{i',j}(t)| \leq K.
$$

Let $\xi > 1$, $\mathcal{O} = \{j : \beta_j^o \neq 0\}$, $\kappa(\xi, \mathcal{O})$ and $F_q(\xi, \mathcal{O})$ be as in (3.8).

THEOREM 3.1. *Let $\tau = K(\xi + 1)d_o \lambda/\{2\kappa^2(\xi, \mathcal{O})\}$ with a certain $K > 0$. Suppose condition (3.9) holds and $\tau \leq 1/e$. Then, in the event $|\dot{\ell}(\boldsymbol{\beta}^o)|_\infty \leq (\xi - 1)/(\xi + 1)\lambda$,*

$$
(3.10) \quad D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq \frac{4e^\eta (1 + 1/\xi)^{-2}\lambda^2 d_o}{\kappa^2(\xi, \mathcal{O})}, \qquad |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_1 \leq \frac{e^\eta (\xi + 1)d_o \lambda}{2\kappa^2(\xi, \mathcal{O})}
$$

and

$$(3.11) \qquad |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_q \leq \frac{e^{\eta} 2(1 + 1/\xi)^{-1} d_o^{1/q} \lambda}{F_q(\xi, \mathcal{O})}, \qquad q \geq 1,$$

where $\eta \leq 1$ is the smaller solution of $\eta e^{-\eta} = \tau$.

Compared with (3.6) and (3.7), the new inequalities (3.10) and (3.11) contain an extra factor $e^{\eta} \leq e$. This is due to the nonlinearity in the Cox regression score equation. Aside from this factor, the error bounds for the Cox regression have the same form as those for linear regression, except for an improvement of a factor of $4\xi/(1 + \xi) \geq 2$ for the ℓ_1 oracle inequality.

The theorem assumes condition (3.9), which asserts $|\mathbf{Z}_i(t) - \mathbf{Z}_{i'}(t)|_{\infty} \leq K$ uniformly in $\{t, i, i'\}$. This condition is a consequence of the uniform boundedness of the individual covariates, and is reasonable in most practical situations (e.g., single-nucleotide polymorphism data). In the case where the covariates are normal variables with uniformly bounded variance, the condition holds with $K = K_{n,p}$ of $\sqrt{\log(np)}$ order.

From an analytical perspective, an important feature of (3.10) and (3.11) is that the constant factors (3.3) and (3.4) are both defined with the true $\boldsymbol{\beta}^o$ in (3.8). No condition is imposed on the gradient and Hessian of the log-partial likelihood for $\boldsymbol{\beta} \neq \boldsymbol{\beta}^o$. In other words, the key condition $\tau < 1/e$, expressed in terms of $\{K, d^o, \lambda\}$ and the compatibility factor $\kappa^2(\xi, \mathcal{O})$ at the true $\boldsymbol{\beta}^o$, is sufficient to guarantee the error bounds in Theorem 3.1. Thus, our results are much simpler to state and conditions easier to verify than existing ones requiring regularity conditions in a neighborhood of $\boldsymbol{\beta}^o$ in the Cox regression model. This feature of Theorem 3.1 plays a crucial role in our derivation of lower bounds for $\kappa^2(\xi, \mathcal{O})$ and $F_q(\xi, \mathcal{O})$ for time-dependent covariates in Section 4. We note that the local martingale structure is valid only at the true $\boldsymbol{\beta}^o$.

To prove Theorem 3.1, we develop a sharpened version of an inequality of Hjort and Pollard (1993). This inequality, given in Lemma 3.2 below, explicitly controls the symmetric Bregman-divergence and Hessian of the log-partial likelihood in a neighborhood of $\boldsymbol{\beta}$. Based on this relationship, Theorem 3.1 is proved using the definition of the quantities in (3.8) and the membership of the error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o$ in the cone $\mathscr{C}(\xi, \mathcal{O})$ (3.2). For two symmetric matrices A and B, $A \leq B$ means $B - A$ is nonnegative-definite.

LEMMA 3.2.   *Let $\ell(\boldsymbol{\beta})$ and its Hessian $\ddot{\ell}(\boldsymbol{\beta})$ be as in (2.2) and (2.6). Then*

$$(3.12) \quad e^{-\eta_{\mathbf{b}}} \mathbf{b}' \ddot{\ell}(\boldsymbol{\beta}) \mathbf{b} \leq D^s(\boldsymbol{\beta} + \mathbf{b}, \boldsymbol{\beta}) = \mathbf{b}' [\dot{\ell}(\boldsymbol{\beta} + \mathbf{b}) - \dot{\ell}(\boldsymbol{\beta})] \leq e^{\eta_{\mathbf{b}}} \mathbf{b}' \ddot{\ell}(\boldsymbol{\beta}) \mathbf{b},$$

where $\eta_{\mathbf{b}} = \max_{s \geq 0} \max_{i,j} |\mathbf{b}' \mathbf{Z}_i(s) - \mathbf{b}' \mathbf{Z}_j(s)|$. Moreover,

$$(3.13) \qquad\qquad e^{-2\eta_{\mathbf{b}}} \ddot{\ell}(\boldsymbol{\beta}) \leq \ddot{\ell}(\boldsymbol{\beta} + \mathbf{b}) \leq e^{2\eta_{\mathbf{b}}} \ddot{\ell}(\boldsymbol{\beta}).$$

Under the conditions of Theorem 3.1, the factors $e^{\pm\eta\mathbf{b}}$ and $e^{\pm 2\eta\mathbf{b}}$ in the inequalities in Lemma 3.2 are bounded by positive constants. These factors lead to the factor e^{η} for $\eta \leq 1$ (and thus $e^{\eta} \leq e$) in the upper bounds in (3.10) and (3.11).

Since the oracle inequalities in Theorem 3.1 are guaranteed to hold only within the event $|\dot{\ell}(\boldsymbol{\beta}^o)|_{\infty} \leq (\xi - 1)/(\xi + 1)\lambda$, a probabilistic upper bound is needed for $|\dot{\ell}(\boldsymbol{\beta}^o)|_{\infty}$. Lemma 3.3 below provides such a probability bound. Similar inequalities can be found in de la Peña (1999).

LEMMA 3.3. (i) *Let* $f_n(t) = n^{-1} \sum_{i=1}^{n} \int_0^t a_i(s)\{dN_i(t) - Y_i(s)\exp(\mathbf{Z}_i'(s) \times \boldsymbol{\beta}^o) \, d\Lambda_0(s)\}$ *with* $[-1, 1]$*-valued predictable processes* $a_i(s)$. *Then, for all* $C_0 > 0$,

$$(3.14) \quad \mathrm{P}\left\{ \max_{t>0} |f_n(t)| > C_0 x, \sum_{i=1}^{n} \int_0^{\infty} Y_i(t) \, dN_i(t) \leq C_0^2 n \right\} \leq 2e^{-nx^2/2}.$$

(ii) *Suppose that* $\max_{i \leq n} \sup_{t \geq 0} \max_{j \leq p} |Z_{i,j}(t) - \bar{Z}_{n,j}(t, \boldsymbol{\beta}^o)|_{\infty} \leq K$, *where* $\bar{Z}_{n,j}(t, \boldsymbol{\beta}^o)$ *are the components of* $\bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^o)$. *Let* $\dot{\ell}(\boldsymbol{\beta})$ *be the gradient in* (2.5). *Then, for all* $C_0 > 0$,

$$(3.15) \quad \mathrm{P}\left\{ |\dot{\ell}(\boldsymbol{\beta}^o)|_{\infty} > C_0 K x, \sum_{i=1}^{n} \int_0^{\infty} Y_i(t) \, dN_i(t) \leq C_0^2 n \right\} \leq 2p e^{-nx^2/2}.$$

In particular, if $\max_{i \leq n} N_i(1) \leq 1$, *then* $\mathrm{P}\{|\dot{\ell}(\boldsymbol{\beta}^o)|_{\infty} > Kx\} \leq 2p e^{-nx^2/2}$.

The following theorem states an upper bound of the estimation error, which follows directly from Theorem 3.1 and Lemma 3.3.

THEOREM 3.2. *Suppose* (3.9) *holds and* $N_i(\infty) \leq 1$ *for all* $i \leq n$ *and* $t \geq 0$. *Let* $\xi > 1$ *and* $\lambda = \{(\xi + 1)/(\xi - 1)\}K\sqrt{(2/n)\log(2p/\varepsilon)}$ *with a small* $\varepsilon > 0$ *(e.g.,* $\varepsilon = 1\%$*). Let* $C_\kappa > 0$ *satisfying* $\tau = K(\xi + 1) d_o \lambda/(2C_\kappa^2) \leq 1/e$. *Let* $\eta \leq 1$ *be the smaller solution of* $\eta e^{-\eta} = \tau$. *Then, for any* $C_{F,q} > 0$,

$$D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq \frac{4e^{\eta}\xi^2\lambda^2 d_o}{(1 + \xi)^2 C_\kappa^2}, \qquad |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_1 \leq \frac{e^{\eta}(\xi + 1)d_o\lambda}{2C_\kappa^2},$$

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_q \leq \frac{2e^{\eta}\xi d_o^{1/q}\lambda}{(\xi + 1)C_{F,q}}$$

all hold with at least probability $\mathrm{P}\{\kappa(\xi, \mathcal{O}) \geq C_\kappa, F_q(\xi, \mathcal{O}) \geq C_{F,q}\} - \varepsilon$.

It is noteworthy that this theorem gives an upper bound of the estimation error for all the ℓ_q norms with $q \geq 1$. From this theorem, for the ℓ_q error $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_q$ with $q \geq 1$ to be small with high probability, we need to ensure that $d_o\lambda \to 0$ as $n \to \infty$. This requires $p = \exp(o(n/d_o^2))$. If d_o is bounded, then p can be as large as $e^{o(n)}$.

Bradic, Fan and Jiang (2011) considered estimation as well as variable selection and oracle properties for general concave penalties, including the lasso. Their broader scope seems to have led to more elaborate statements and some key conditions that are more difficult to verify than those of Theorems 3.1 and 3.2, for example, their Condition 2(i) on a uniformly small spectrum bound between $S^{(2)}(t, \boldsymbol{\beta}_1)$ and its population version for a sparse $\boldsymbol{\beta}_1$ in a neighborhood of $\boldsymbol{\beta}^o$.

PROOF OF THEOREM 3.2. Let $C_0 = 1$ and $x = \lambda(\xi - 1)/\{K(\xi + 1)\} = \sqrt{(2/n)\log(2p/\varepsilon)}$ in Lemma 3.3. The probability of the event $|\dot{\ell}(\boldsymbol{\beta}^o)|_\infty > (\xi - 1)/(\xi + 1)\lambda$ is at most ε. The desired result follows directly from Theorem 3.1. □

4. Compatibility and invertibility factors and restricted eigenvalues. In Section 3, the oracle inequalities in Theorems 3.1 and 3.2 are expressed in terms of the compatibility and weak cone invertibility factors. However, as mentioned in the Introduction, these quantities are still random variables. This section provides sufficient conditions under which they can be treated as constants. Since these factors appear in the denominator of error bounds, it suffices to bound them from below. We also derive a lower bound for the restricted eigenvalue to facilitate further analysis of the Cox model in high-dimension. We will prove that these quantities are bounded from below by the population version of their certain truncated versions.

Compared with linear regression, our problem poses two additional difficulties in the Cox model: (a) time dependence of covariates, and (b) stochastic integration of the Hessian over random risk sets. Fortunately, the compatibility and weak cone invertibility factors in Theorems 3.1 and 3.2 involve only the Hessian of the log-partial likelihood at the true $\boldsymbol{\beta}^o$, so that a martingale argument can be used.

To simplify the statement of our results, we use $\phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$ to denote any of the following quantities:

$$(4.1) \qquad \phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}) = \kappa^2(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}), \qquad F_q(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}) \quad \text{or} \quad \text{RE}^2(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}),$$

where $\kappa(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$, $F_q(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$, and $\text{RE}(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$ are as in (3.3), (3.4) and (3.5), respectively. If we make a claim about $\phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$, we mean that the claim holds for any quantity it represents. Let $\phi_{\min}$ denote the smallest eigenvalue. The following lemma provides some key properties of $\phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$ used in the derivation of its lower bounds.

LEMMA 4.1. *Let* $\kappa(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$, $F_q(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$, $\text{RE}(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$ *and* $\phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$ *be as in* (4.1). *Let* $\bar{\Sigma}_{jk}$ *be the elements of* $\bar{\boldsymbol{\Sigma}}$ *and* $\boldsymbol{\Sigma}$ *be another nonnegative-definite matrix with elements* Σ_{jk}.

(i) *For* $1 \le q \le 2$,

$$\min\{\kappa^2(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}), (1 + \xi)^{2/q - 1} F_q(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})\} \ge \text{RE}^2(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}) \ge \phi_{\min}(\bar{\boldsymbol{\Sigma}}).$$

(ii) $\phi(\xi, \mathcal{O}; \bar{\Sigma}) \geq \phi(\xi, \mathcal{O}; \Sigma) - d^o(\xi + 1)^2 \max_{1 \leq j \leq k \leq p} |\bar{\Sigma}_{jk} - \Sigma_{jk}|$.

(iii) If $\bar{\Sigma} \geq \Sigma$, then $\phi(\xi, \mathcal{O}; \bar{\Sigma}) \geq \phi(\xi, \mathcal{O}; \Sigma)$.

PROOF. By the Hölder inequality, $|\mathbf{b}|_q \leq |\mathbf{b}|_1^{2/q-1} |\mathbf{b}|_2^{2-2/q}$ for all $1 \leq q \leq 2$. Since $|\mathbf{b}|_1 \leq (1 + \xi)|\mathbf{b}_{\mathcal{O}}|_1$ in the cone and $|\mathbf{b}_{\mathcal{O}}|_1 \leq d_o^{1/2}|\mathbf{b}|_2$, we have

$$|\mathbf{b}_{\mathcal{O}}|_1 |\mathbf{b}|_q / d_o^{1/q} \leq (1 + \xi)^{2/q-1} |\mathbf{b}_{\mathcal{O}}|_1^{2/q} |\mathbf{b}|_2^{2-2/q} / d_o^{1/q} \leq (1 + \xi)^{2/q-1} |\mathbf{b}|_2^2.$$

This and $|\mathbf{b}_{\mathcal{O}}|_1 \leq d_o^{1/2}|\mathbf{b}|_2$ yields part (i) by the definition of the quantities involved. Part (ii) follows from $|\mathbf{b}'\bar{\Sigma}\mathbf{b} - \mathbf{b}'\Sigma\mathbf{b}| \leq |\mathbf{b}|_1^2 \max_{j,k} |\bar{\Sigma}_{jk} - \Sigma_{jk}|$ and $|\mathbf{b}|_1 \leq (\xi + 1)|\mathbf{b}_{\mathcal{O}}|_1 \leq (\xi + 1)d_o^{1/q}|\mathbf{b}|_q$. Part (iii) follows immediately from the definition of the quantities in (4.1). $\square$

It follows from Lemma 4.1(ii) and (iii) that quantities of type $\phi(\xi, \mathcal{O}; \bar{\Sigma})$ in (4.1) can be bounded from below in two ways. The first is to bound the matrix $\bar{\Sigma}$ from below and the second is to approximate $\bar{\Sigma}$ under the supreme norm for its elements. In the $p \gg n$ setting, our problem is essentially the rank deficiency of $\bar{\Sigma}$ to begin with, so that its lower bound is still rank deficient. However, a lower bound of the random matrix $\bar{\Sigma} = \ddot{\ell}(\boldsymbol{\beta}^o)$, for example, a certain truncated version of it, may have a smaller variability to allow an approximation by its population version. This is our basic idea. In fact, our analysis takes advantage of this argument twice to remove different sources of randomness.

According to our plan described in the previous paragraph, we first choose a suitable truncation of $\bar{\Sigma} = \ddot{\ell}(\boldsymbol{\beta}^o)$ as a lower bound of the matrix. This is done by truncating the maximum event time under consideration. It follows from (2.6) that for $t^* > 0$,

$$(4.2) \quad \ddot{\ell}(\boldsymbol{\beta}^o) \geq \ddot{\ell}(\boldsymbol{\beta}^o; t^*) \qquad \text{where } \ddot{\ell}(\boldsymbol{\beta}^o; t^*) = n^{-1} \int_0^{t^*} V_n(s, \boldsymbol{\beta}^o) \, d\bar{N}(s).$$

This allows us to remove the randomness from the counting process by replacing the average counting measure $n^{-1} d\bar{N}(t)$ by its compensator $R_n(s, \boldsymbol{\beta}^o) \, d\Lambda_0(s)$, where Λ_0 is the baseline cumulative hazard function. This approximation of $\ddot{\ell}(\boldsymbol{\beta}^o; t^*)$ can be written as

$$(4.3) \qquad \qquad \bar{\Sigma}(t^*) = \int_0^{t^*} V_n(s, \boldsymbol{\beta}^o) R_n(s, \boldsymbol{\beta}^o) \, d\Lambda_0(s).$$

To completely remove the randomness with $\bar{\Sigma}(t^*)$, we apply the method again by truncating the weights $e^{\mathbf{Z}_i'(t)\boldsymbol{\beta}^o}$ with $R_n(s, \boldsymbol{\beta}^o)$. For $M > 0$, define

$$(4.4) \qquad \qquad \bar{\Sigma}(t^*; M) = \int_0^{t^*} \hat{\mathbf{G}}_n(s; M) \, d\Lambda_0(s),$$

where $\hat{\mathbf{G}}_n(t; M) = n^{-1} \sum_{i=1}^{n} \{\mathbf{Z}_i - \bar{\mathbf{Z}}_n(t; M)\}^{\otimes 2} Y_i(t) \min\{M, \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta}^o)\}$ with

$$\bar{\mathbf{Z}}_n(t; M) = \frac{\sum_{i=1}^{n} \mathbf{Z}_i(t) Y_i(t) \min\{M, \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta}^o)\}}{\sum_{i=1}^{n} Y_i(t) \min\{M, \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta}^o)\}}.$$

We will prove that the matrix (4.4) is a lower bound of (4.3). Suppose $\{Y_i(t), \mathbf{Z}_i(t), t \geq 0\}$ are i.i.d. stochastic processes from $\{Y(t), \mathbf{Z}(t), t \geq 0\}$. The population version of (4.4) is then

$$(4.5) \qquad\qquad \boldsymbol{\Sigma}(t^*; M) = E \int_0^{t^*} \mathbf{G}_n(s; M) \, d\Lambda_0(s),$$

where $\mathbf{G}_n(t; M) = n^{-1} \sum_{i=1}^{n} \{\mathbf{Z}_i - \boldsymbol{\mu}(t; M)\}^{\otimes 2} Y_i(t) \min\{M, \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta}^o)\}$ with

$$\boldsymbol{\mu}(t; M) = \frac{E[\mathbf{Z}(t) Y(t) \min\{M, \exp(\mathbf{Z}'\boldsymbol{\beta}^o)\}]}{E[Y(t) \min\{M, \exp(\mathbf{Z}'\boldsymbol{\beta}^o)\}]}.$$

The analysis outlined above leads to the following main result of this section. For $\xi \geq 1$ and $\mathcal{O} \subset \{1, \ldots, p\}$ with $|\mathcal{O}| = d_o$, let $\phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$ represent all quantities of interest given in (4.1), $\kappa(\xi, \mathcal{O})$ and $F_q(\xi, \mathcal{O})$ be the compatibility and weak cone invertibility factors in (3.8) with the Hessian $\ddot{\ell}(\boldsymbol{\beta}^o)$ in (2.6) at the true $\boldsymbol{\beta}$, and RE$(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$ be as in (3.5). Let $L_n(t) = \sqrt{(2/n) \log t}$.

THEOREM 4.1. *Suppose* $\{Y_i(t), \mathbf{Z}_i(t), t \geq 0\}$ *are i.i.d. processes from* $\{Y(t), \mathbf{Z}(t), t \geq 0\}$ *with* $\sup_t P\{|\mathbf{Z}_i(t) - \mathbf{Z}(t)|_\infty \leq K\} = P\{\max_i N_i(\infty) \leq 1\} = 1$. *Let* $\{t^*, M\}$ *be positive constants and* $r_* = EY(t^*) \min\{M, \exp(\mathbf{Z}'(t^*)\boldsymbol{\beta}^o)\}$. *Then,*

$$\phi(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o))$$

$$(4.6) \qquad \geq \phi(\xi, \mathcal{O}; \boldsymbol{\Sigma}(t^*; M))$$

$$\qquad\qquad - d_o(\xi + 1)^2 K^2 \{C_1 L_n(p(p+1)/\varepsilon) + C_2 t_{n,p,\varepsilon}^2\}$$

with at least probability $1 - 3\varepsilon$, *where* $C_1 = 1 + \Lambda_0(t^*)$, $C_2 = (2/r_*)\Lambda_0(t^*)$ *and* $t_{n,p,\varepsilon}$ *is the solution of* $p(p+1) \exp\{-nt_{n,p,\varepsilon}^2/(2 + 2t_{n,p,\varepsilon}/3)\} = \varepsilon/2.221$. *Consequently, for* $1 \leq q \leq 2$,

$$\min\{\kappa^2(\xi, \mathcal{O}), (1 + \xi)^{2/q-1} F_q(\xi, \mathcal{O})\}$$

$$(4.7) \qquad \geq \text{RE}^2(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o))$$

$$\qquad\qquad \geq \rho_* - d_o(\xi + 1)^2 K^2 \{C_1 L_n(p(p+1)/\varepsilon) + C_2 t_{n,p,\varepsilon}^2\}$$

with at least probability $1 - 3\varepsilon$, *where* $\rho_* = \phi_{\min}(\boldsymbol{\Sigma}(t^*; M))$ *with the matrix in* (4.5).

Theorem 4.1 implies that the compatibility and cone invertibility factors and the restricted eigenvalue can be all treated as constants in high-dimensional Cox model with time-dependent covariates. We note that $C_2 t_{n,p,\varepsilon}^2$ is of smaller order

than $L_n(p(p+1)/\varepsilon)$ so that the lower bounds in (4.6) and (4.7) depend on the choice of t^* and M marginally through C_1 and ρ_*. If $d^o\sqrt{(\log p)/n}$ is sufficiently small as assumed in Theorem 3.2, the right-hand side of (4.7) can be treated as $\rho_*/2$. It is reasonable to treat ρ_* as a constant since it is the smallest eigenvalue of a population integrated covariance matrix in (4.5).

In the proof of Theorem 4.1, the martingale exponential inequality in Lemma 3.3 is used to bound the difference between (4.2) and (4.3). The following Bernstein inequality for V-statistics is used to bound the difference between (4.4) and (4.5). This inequality can be viewed as an extension of the Hoeffding (1963) inequality for sums of bounded independent variables and nondegenerate U-statistics.

LEMMA 4.2. *Let X_i be a sequence of independent stochastic processes and $f_{i,j}$ be functions of X_i and X_j with $|f_{i,j}| \leq 1$. Suppose $f_{i,j}$ are degenerate in the sense of $E[f_{i,j}|X_i] = E[f_{i,j}|X_j] = 0$ for all $i \neq j$. Let $V_n = \sum_{i=1}^n \sum_{j=1}^n f_{i,j}$. Then,*

$$P\{\pm V_n > (nt)^2\} \leq \frac{2\varepsilon_n(t)(1+\varepsilon_n(t))}{(1+\varepsilon_n^2(t))^2} \leq 2.221 \exp\left(-\frac{nt^2/2}{1+t/3}\right),$$

where $\varepsilon_n(t) = e^{-(nt^2/2)/(1+t/3)}$.

Our discussion focuses on the quantities in (4.1) for the Hessian matrix $\bar{\Sigma} = \ddot{\ell}(\boldsymbol{\beta}^o)$ evaluated at the true vector of coefficients. Still, through Lemma 3.2, Theorem 4.1 also provides lower bounds for these quantities at any $\boldsymbol{\beta}$ not far from the true $\boldsymbol{\beta}^o$ in terms of the ℓ_1 distance. We formally state this result in the following corollary.

COROLLARY 4.1. *Let $\phi(\xi, \mathcal{O}; \bar{\Sigma})$ represent any quantities in (4.1). Then,*

$$e^{-2\eta_{\mathbf{b}}}\phi(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o)) \leq \phi(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o + \mathbf{b})) \leq e^{2\eta_{\mathbf{b}}}\phi(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o)),$$

where

$$\eta_{\mathbf{b}} = \sup_s \max_{i,j} |\mathbf{b}'\mathbf{Z}_i(s) - \mathbf{b}'\mathbf{Z}_j(s)|.$$

Consequently, when $|\mathbf{Z}_i(s) - \mathbf{Z}_j(s)|_\infty \leq K$,

$$\begin{aligned}
\inf&\{\phi(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta})): |\boldsymbol{\beta} - \boldsymbol{\beta}^o|_1 \leq \eta/(2K)\} \\
&\geq e^{-\eta}\phi(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o)) \\
&\geq e^{-\eta}[\rho_* - d_o(\xi+1)^2 K^2\{C_1 L_n(p(p+1)/\varepsilon) + C_2 t_{n,p,\varepsilon}^2\}]
\end{aligned}$$

under the conditions of Theorem 4.1.

It is worthwhile to point out that unlike typical "small ball" analysis based on Taylor expansion, Corollary 4.1 provides nonasymptotic control of the quantities in an ℓ_1 ball of constant size. Since $\mathbf{b}'\bar{\boldsymbol{\Sigma}}\mathbf{b}$ appears in the numerator of the quantities represented by $\phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}})$, Corollary 4.1 follows immediately from Theorem 4.1 and (3.13). It implies that the Hessian has sufficient invertibility properties in the analysis of the lasso when the estimator is not far from the true $\boldsymbol{\beta}^o$ in ℓ_1 distance. On the other hand, if the Hessian has sufficient invertibility properties in a ball of fixed size, nonasymptotic error bounds for the lasso estimator can be established. This "chicken and egg" problem is directly solved in the proof of Theorem 3.1.

5. Concluding remarks.

This paper deals with the Cox proportional hazards regression model when the number of time-dependent covariates p is potentially much larger than the sample size n. The ℓ_1 penalty is used to regularize the log-partial likelihood function. Error bounds parallel to those of the lasso in linear regression are established. In establishing these bounds, we extend the notion of the restricted eigenvalue and compatibility and cone invertibility factors to the Cox model. We show that these quantities indeed provide useful error bounds.

An important issue is the choice of the penalty level λ. Theorem 3.2 requires a λ slightly larger than $K\sqrt{(2/n)\log p}$, where K is a uniform upper bound for the range of individual real covariates. This indicates that the lasso is tuning insensitive since the theoretical choice does not depend on the unknowns. In practice, cross validation can be used to fine tune the penalty level λ. Theoretical investigation of the performance of the lasso with cross-validated λ, an interesting and challenging problem in and of itself even in the simpler linear regression model, is beyond the scope of this paper.

General concave penalized estimators in the Cox regression model have been considered in Bradic, Fan and Jiang (2011) where oracle inequalities and properties of certain local solutions are considered. Zhang and Zhang (2012) has provided a unified treatment of global and local solutions for concave penalized least squares estimators in linear regression. Since this unified treatment relies on an oracle inequality for the global solution based on the cone invertibility factor, the results in this paper point to a possible extension of such a unified treatment of global and local solutions of general concave regularized methods in the Cox regression model.

APPENDIX

Here we prove Lemmas 3.1, 3.2, 3.3 and 4.2 and Theorems 3.1 and 4.1.

PROOF OF LEMMA 3.1. Since $\ell(\boldsymbol{\beta})$ is a convex function, $D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \tilde{\boldsymbol{\theta}}'\{\dot{\ell}(\boldsymbol{\beta}^o + \tilde{\boldsymbol{\theta}}) - \dot{\ell}(\boldsymbol{\beta}^o)\} \geq 0$, so that the first inequality holds. Since $\tilde{\theta}_j = \hat{\beta}_j$ for

$j \in \mathcal{O}^c$, (2.4) gives

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}} \{ \dot{\ell}(\boldsymbol{\beta}^o + \tilde{\boldsymbol{\theta}}) &- \dot{\ell}(\boldsymbol{\beta}^o) \} \\
&= \sum_{j \in \mathcal{O}^c} \tilde{\theta}_j (\dot{\ell}(\boldsymbol{\beta}^o + \tilde{\boldsymbol{\theta}}))_j + \sum_{j \in \mathcal{O}} \tilde{\theta}_j (\dot{\ell}(\boldsymbol{\beta}^o + \tilde{\boldsymbol{\theta}}))_j + \tilde{\boldsymbol{\theta}}'(-\dot{\ell}(\boldsymbol{\beta}^o)) \\
&\leq \sum_{j \in \mathcal{O}^c} \hat{\beta}_j (-\lambda \operatorname{sgn}(\hat{\beta}_j)) + \sum_{j \in \mathcal{O}} |\tilde{\theta}_j| \lambda + |\tilde{\boldsymbol{\theta}}|_1 z^* \\
&= \sum_{j \in \mathcal{O}^c} -\lambda |\tilde{\theta}_j| + |\tilde{\boldsymbol{\theta}}_{\mathcal{O}}|_1 \lambda + z^* |\tilde{\boldsymbol{\theta}}_{\mathcal{O}}|_1 + z^* |\tilde{\boldsymbol{\theta}}_{\mathcal{O}^c}|_1 \\
&= (z^* - \lambda)|\tilde{\boldsymbol{\theta}}_{\mathcal{O}^c}|_1 + (\lambda + z^*)|\tilde{\boldsymbol{\theta}}_{\mathcal{O}}|_1.
\end{aligned}
$$

Thus the second inequality in (3.1) holds. Note that the inequality in the third line above requires $(\dot{\ell}(\boldsymbol{\beta}^o + \tilde{\boldsymbol{\theta}}))_j = -\lambda \operatorname{sgn}(\hat{\beta}_j)$ only in the set $\mathcal{O}^c \cap \{j : \hat{\beta}_j \neq 0\}$, since $\tilde{\theta}_j = \hat{\beta}_j - \beta_j^o = 0$ when $j \in \mathcal{O}^c$ and $\hat{\beta}_j = 0$. $\quad\square$

PROOF OF LEMMA 3.2. We use similar notation as in Hjort and Pollard (1993). Let $a_i = a_i(s) = \mathbf{b}'\{\mathbf{Z}_i(s) - \bar{\mathbf{Z}}_n(s, \boldsymbol{\beta})\}$, $w_i = w_i(s) = Y_i(s) \exp[\boldsymbol{\beta}' \mathbf{Z}_i(s)]$ and $c = c(s) = (\max_i a_i(s) + \min_i a_i(s))/2$. Clearly, $\max_i |a_i - c| \leq (1/2)\eta_{\mathbf{b}}$. By the definition of $\bar{\mathbf{Z}}_n(t, \boldsymbol{\beta})$,

$$
\begin{aligned}
\mathbf{b}'\{\bar{\mathbf{Z}}_n(s, \boldsymbol{\beta} + \mathbf{b}) &- \bar{\mathbf{Z}}_n(s, \boldsymbol{\beta})\} \\
&= \sum_i \mathbf{b}' \mathbf{Z}_i(s) w_i e^{\mathbf{b}' \mathbf{Z}_i(s)} \Big/ \sum_i w_i e^{\mathbf{b}' \mathbf{Z}_i(s)} - \sum_i \mathbf{b}' \mathbf{Z}_i(s) w_i \Big/ \sum_i w_i \\
&= \sum_i a_i w_i e^{a_i} \Big/ \sum_i w_i e^{a_i} - \sum_i a_i w_i \Big/ \sum_i w_i \\
&= \sum_{i,j} w_i w_j a_i (e^{a_i} - e^{a_j}) \Big/ \sum_{i,j} w_i w_j e^{a_i} \\
&= \sum_{i,j} w_i w_j (a_i - a_j)(e^{a_i - c} - e^{a_j - c}) \Big/ \sum_{i,j} 2 w_i w_j e^{a_i - c} \\
&\geq \exp\Big(-2 \max_i |a_i - c|\Big) \sum_{i,j} w_i w_j (a_i - a_j)^2 \Big/ \sum_{i,j} 2 w_i w_j \\
&\geq \exp(-\eta_{\mathbf{b}}) \sum_i w_i a_i^2 \Big/ \sum_i w_i,
\end{aligned}
$$

where the first inequality comes from $(e^y - e^x)/(y - x) \geq e^{-(|y| \vee |x|)}$ and, since $\sum_i w_i a_i = 0$, the second one from $\sum_{i,j} w_i w_j (a_i - a_j)^2 = 2 \sum_i w_i \sum_i w_i a_i^2$. Thus,

since $a_i^2 = \mathbf{b}'\{\mathbf{Z}_i(s) - \bar{\mathbf{Z}}_n(s, \boldsymbol{\beta})\}^{\otimes 2}\mathbf{b}$, (2.6) and (2.5) give

$$e^{-\eta_{\mathbf{b}}}\mathbf{b}'\ddot{\ell}(\boldsymbol{\beta})\mathbf{b} = \frac{e^{-\eta_{\mathbf{b}}}}{n}\int_0^\infty \sum_{i=1}^n w_i a_i^2 \left(\sum_{i=1}^n w_i\right)^{-1} d\bar{N}(s) \le \mathbf{b}'\{\dot{\ell}(\boldsymbol{\beta}+\mathbf{b}) - \dot{\ell}(\boldsymbol{\beta})\}.$$

This implies the lower bound in (3.12). Similarly, the lower bound in (3.13) follows from

$$\ddot{\ell}(\boldsymbol{\beta}+\mathbf{b}) = \frac{1}{n}\int_0^\infty \frac{\sum_{i,j} w_i w_j\{\mathbf{Z}_i(s)\mathbf{Z}_i'(s) - \mathbf{Z}_i(s)\mathbf{Z}_j'(s)\}e^{a_i+a_j}}{\sum_{i,j} w_i w_j e^{a_i+a_j}} d\bar{N}(s)$$

$$= \frac{1}{n}\int_0^\infty \frac{\sum_{i,j} w_i w_j(\mathbf{Z}_i(s) - \mathbf{Z}_j(s))^{\otimes 2} e^{(a_i-c)+(a_j-c)}}{\sum_{i,j} 2w_i w_j e^{(a_i-c)+(a_j-c)}} d\bar{N}(s)$$

and

$$\ddot{\ell}(\boldsymbol{\beta}) = \frac{1}{n}\int_0^\infty \frac{\sum_{i,j} w_i w_j(\mathbf{Z}_i(s) - \mathbf{Z}_j(s))^{\otimes 2}}{\sum_{i,j} 2w_i w_j} d\bar{N}(s).$$

The proof of the upper bounds in (3.12) and (3.13), nearly identical to the proof of the lower bounds, is omitted. $\square$

PROOF OF LEMMA 3.3.   Applying the union bound and changing the scale of the covariates if necessary, we assume without loss of generality that $p = K = 1$. In this case

$$\dot{\ell}(\boldsymbol{\beta}^o) = \frac{1}{n}\sum_{i=1}^n \int_0^\infty a_i(s)\,dN_i(s) = \frac{1}{n}\sum_{i=1}^n \int_0^\infty a_i(s)\,dM_i(s),$$

where $a_i(t) = Z_{i1}(t) - \bar{Z}_{n,1}(t), i = 1, \ldots, n$, are predictable and satisfy $|a_i(t)| \le 1$. Thus, (3.15) follows from (3.14).

Let t_j be the time of the jth jump of the process $\sum_{i=1}^n \int_0^\infty Y_i(t)\,dN_i(t), j = 1, \ldots, m$ and $t_0 = 0$. Then, t_j are stopping times. For $j = 0, \ldots, m$, define

$$(A.1) \qquad X_j = \sum_{i=1}^n \int_0^{t_j} a_i(s)\,dN_i(s) = \sum_{i=1}^n \int_0^{t_j} a_i(s)\,dM_i(s).$$

Since $M_i(s)$ are martingales and $a_i(s)$ are predictable, $\{X_j, j = 0, 1, \ldots\}$ is a martingale with the difference $|X_j - X_{j-1}| \le \max_{s,i} |a_i(s)| \le 1$. Let m be the greatest integer lower bound of $C_0^2 n$. By the martingale version of the Hoeffding (1963) inequality [Azuma (1967)],

$$P(|X_m| > nC_0 x) \le 2\exp(-n^2 C_0^2 x^2/(2m)) \le 2e^{-nx^2/2}.$$

By (A.1), $X_m = n\dot{\ell}(\boldsymbol{\beta}^o)$ if and only if $\sum_{i=1}^n \int_0^\infty Y_i(t)\,dN_i(t) \le m$. Thus, the left-hand side of (3.15) is no greater than $P(|X_m| > nC_0 x) \le 2e^{-nx^2/2}$. $\square$

PROOF OF LEMMA 4.2. For integers $j, m, i_1, \ldots, i_m$, let $\#(j; i_1, \ldots, i_m)$ be the number of appearances of j in the sequence $\{i_1, \ldots, i_m\}$. Since $f_{i,j}$ are degenerate,

$$E(\pm V_n)^m = \sum_{1 \le i_1, \ldots, i_{2m} \le n} (\pm 1)^m f_{i_1, i_2} \cdots f_{i_{2m-1}, i_{2m}}$$

$$\le \sum_{1 \le i_1, \ldots, i_{2m} \le n} \prod_{j=1}^n I\{\#(j; i_1, \ldots, i_{2m}) \ne 1\}.$$

This is due to the fact that all terms with exactly one appearance of an index j have zero expectation and all other terms are bounded by 1. Let E_0 be the expectation under which $i_1, \ldots, i_{2m}$ are i.i.d. uniform variables in $\{1, \ldots, n\}$ and $k_j = \#(j; i_1, \ldots, i_{2m})$. Since $(k_1, \ldots, k_n)$ is multinomial$(2m, 1/n, \ldots, 1/n)$, the above inequality can be written as

$$E(\pm V_n)^m \le n^{2m} E_0 \prod_{j=1}^n I\{k_j \ne 1\} = (2m)! \sum_{k_1 + \cdots + k_n = 2m} \prod_{j=1}^n \frac{I\{k_j \ne 1\}}{k_j!}.$$

Let $f_0(x) = \sum_{m=0}^\infty x^m / (2m)! = \cosh(|x|^{1/2}) I\{x \ge 0\} + \cos(|x|^{1/2}) I\{x < 0\}$ and $\lambda = t/(1 + t/3)$. It follows from the above moment inequality that

$$Ef_0(\pm \lambda^2 V_n) = \sum_{m=0}^\infty \lambda^{2m} E(\pm V_n)^m / (2m)!$$

$$\le \sum_{m=0}^\infty \lambda^{2m} \sum_{k_1 + \cdots + k_n = 2m} \prod_{j=1}^n \frac{I\{k_j \ne 1\}}{k_j!}$$

$$\le \sum_{m=0}^\infty \lambda^m \sum_{k_1 + \cdots + k_n = m} \prod_{j=1}^n \frac{I\{k_j \ne 1\}}{k_j!}$$

$$= \left(\sum_{k=0}^\infty \lambda^k I\{k \ne 1\} / k! \right)^n.$$

Since $\sum_{k=0}^\infty \lambda^k I\{k \ne 1\} / k! \le 1 + (\lambda^2/2)/(1 - \lambda/3) = 1 + \lambda t/2$, we find $Ef_0(\pm \lambda^2 V_n) \le e^{n\lambda t/2}$. Consequently, the monotonicity of $f(x) = \cosh(x^{1/2})$ for $x > 0$ and the lower bound $f(x) \ge -1$ allow us to apply the Markov inequality as follows:

$$P\{\pm V_n > (nt)^2\} \le P\{1 + f_0(\pm \lambda^2 V_n) > 1 + f_0((\lambda nt)^2)\}$$

$$\le \{1 + f_0((n\lambda t)^2)\}^{-1} E\{1 + f_0(\pm \lambda^2 V_n)\}$$

$$\le \{1 + \cosh(n\lambda t)\}^{-1} (1 + e^{n\lambda t/2})$$

$$= 2e^{-n\lambda t/2} (1 + e^{-n\lambda t/2}) / (1 + e^{-n\lambda t})^2.$$

The conclusion follows from $2 \max_{0 \leq x \leq 1}(1 + x)/(1 + x^2)^2 \leq 2.221$. $\quad\square$

PROOF OF THEOREM 3.1.   Let $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o \neq 0$ and $\mathbf{b} = \tilde{\boldsymbol{\theta}}/|\tilde{\boldsymbol{\theta}}|_1$. It follows from the convexity of $\ell(\boldsymbol{\beta}^o + x\mathbf{b})$, as a function of x, and Lemma 3.1 that, in the event $|\dot{\ell}(\boldsymbol{\beta}^o)|_\infty \leq (\xi - 1)/(\xi + 1)\lambda$,

$$(A.2) \qquad \mathbf{b}'\{\dot{\ell}(\boldsymbol{\beta}^o + x\mathbf{b}) - \dot{\ell}(\boldsymbol{\beta}^o)\} + \frac{2\lambda}{\xi + 1}|\mathbf{b}_{\mathcal{O}^c}|_1 \leq \frac{2\xi\lambda}{\xi + 1}|\mathbf{b}_{\mathcal{O}}|_1$$

for $x \in [0, |\tilde{\boldsymbol{\theta}}|_1]$ and $\mathbf{b} \in \mathscr{C}(\xi, \mathcal{O})$. Consider all nonnegative x satisfying (A.2). We need to establish a lower bound for

$$\mathbf{b}'\{\dot{\ell}(\boldsymbol{\beta}^o + x\mathbf{b}) - \dot{\ell}(\boldsymbol{\beta}^o)\} = \frac{1}{n}\int_0^\infty \mathbf{b}'\{\bar{Z}_n(s, \boldsymbol{\beta}^o + x\mathbf{b}) - \bar{Z}_n(s, \boldsymbol{\beta}^o)\}\, d\bar{N}(s).$$

Since $\eta_{x\mathbf{b}} = \max_{0 \leq s \leq 1} \max_{i,j} |x\mathbf{b}'\mathbf{Z}_i(s) - x\mathbf{b}'\mathbf{Z}_j(s)| \leq Kx|\mathbf{b}|_1 = Kx$, Lemma 3.2 yields

$$(A.3) \qquad \begin{aligned} x\mathbf{b}'\{\dot{\ell}(\boldsymbol{\beta}^o + x\mathbf{b}) - \dot{\ell}(\boldsymbol{\beta}^o)\} &\geq x^2 \exp(-\eta_{x\mathbf{b}})\mathbf{b}'\ddot{\ell}(\boldsymbol{\beta}^o)\mathbf{b} \\ &\geq x^2 \exp(-Kx)\mathbf{b}'\ddot{\ell}(\boldsymbol{\beta}^o)\mathbf{b}. \end{aligned}$$

This, combined with (A.2) and the definition of $\kappa(\xi, \mathcal{O})$, gives

$$\begin{aligned} xe^{-Kx}\kappa^2(\xi, \mathcal{O})|\mathbf{b}_{\mathcal{O}}|_1^2/d_o &\leq xe^{-Kx}\mathbf{b}'\ddot{\ell}(\boldsymbol{\beta}^o)\mathbf{b} \\ &\leq \frac{2\xi\lambda}{\xi + 1}|\mathbf{b}_{\mathcal{O}}|_1 - \frac{2\lambda}{\xi + 1}|\mathbf{b}_{\mathcal{O}^c}|_1 \\ &= 2\lambda|\mathbf{b}_{\mathcal{O}}|_1 - \frac{2\lambda}{\xi + 1} \\ &\leq \lambda(\xi + 1)|\mathbf{b}_{\mathcal{O}}|_1^2/2. \end{aligned}$$

In other words, any x satisfying (A.2) must satisfy

$$(A.4) \qquad Kx\exp(-Kx) \leq \frac{K(\xi + 1)\lambda d_o}{2\kappa^2(\xi, \mathcal{O})} = \tau.$$

Since $\mathbf{b}'\{\dot{\ell}(\boldsymbol{\beta}^o + x\mathbf{b}) - \dot{\ell}(\boldsymbol{\beta}^o)\}$ is an increasing function of x due to the convexity of ℓ, the set of all nonnegative x satisfying (A.2) is a closed interval $[0, \tilde{x}]$ for some $\tilde{x} > 0$. Thus, (A.4) implies $K\tilde{x} \leq \eta$, the smaller solution of $\eta e^{-\eta} = \tau$. This yields

$$|\tilde{\boldsymbol{\theta}}|_1 \leq \tilde{x} \leq \frac{\eta}{K} = \frac{e^\eta \tau}{K} = \frac{e^\eta(\xi + 1)\lambda d_o}{2\kappa^2(\xi, \mathcal{O})}$$

in (3.10). The first part of (3.10) follows from (3.3), (3.8), (3.12) and (3.1), due to

$$e^{-\eta}\kappa^2(\xi, \mathcal{O})|\tilde{\boldsymbol{\theta}}_{\mathcal{O}}|_1^2/d^o \leq e^{-\eta}\tilde{\boldsymbol{\theta}}'\ddot{\ell}(\boldsymbol{\beta}^o)\tilde{\boldsymbol{\theta}} \leq D^s(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq \frac{2\xi\lambda|\tilde{\boldsymbol{\theta}}_{\mathcal{O}}|_1}{\xi + 1}.$$

Finally, it follows from the definition of $F_q(\xi, \mathcal{O})$, (A.3) and (A.2) that, for $x = |\tilde{\boldsymbol{\theta}}|_1$,

$$xe^{-\eta} \leq \frac{xe^{-Kx}\mathbf{b}'\ddot{\ell}(\boldsymbol{\beta}^o)\mathbf{b}}{F_q(\xi, \mathcal{O})(|\mathbf{b}_{\mathcal{O}}|_1/d_o^{1/q})|\mathbf{b}|_q} \leq \frac{\mathbf{b}'\{\dot{\ell}(\boldsymbol{\beta}^o + x\mathbf{b}) - \dot{\ell}(\boldsymbol{\beta}^o)\}}{F_q(\xi, \mathcal{O})(|\mathbf{b}_{\mathcal{O}}|_1/d_o^{1/q})|\mathbf{b}|_q}$$

$$\leq \frac{2\xi\lambda d_o^{1/q}}{(\xi + 1)F_q(\xi, \mathcal{O})|\mathbf{b}|_q}.$$

This gives the second inequality in (3.11) due to $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o|_q = |\tilde{\boldsymbol{\theta}}|_1|\mathbf{b}|_q$. $\square$

PROOF OF THEOREM 4.1. Let

$$a(s) = \left(V_n(s, \boldsymbol{\beta}^o)\right)_{jk}/K^2$$

$$= \sum_{i=1}^n w_{ni}(t, \boldsymbol{\beta}^o)\{Z_{i,j}(s) - \bar{Z}_{n,j}(s)\}\{Z_{i,k}(s) - \bar{Z}_{n,k}(s)\}/K^2.$$

It follows from Lemma 3.3(i) with $a_i(s) = a(s)$ and $C_0 = 1$ that

$$P\left\{\left|\left(\int_0^{t^*} V_n(s, \boldsymbol{\beta}^o)\, d\bar{N}(s) - \int_0^{t^*} V_n(s, \boldsymbol{\beta}^o)R_n(s, \boldsymbol{\beta}^o)\, d\Lambda_0(s)\right)_{jk}\right| > K^2 x\right\}$$

$$\leq 2e^{-nx^2/2}.$$

Thus, $P\{\max_{j,k}|(\ddot{\ell}(\boldsymbol{\beta}^o; t^*) - \bar{\boldsymbol{\Sigma}}(t^*)|_{j,k} \geq K^2 L_n(p(p + 1)/\varepsilon)\} \leq \varepsilon$ by the union bound and the respective definitions of $\ddot{\ell}(\boldsymbol{\beta}^o; t^*)$ and $\bar{\boldsymbol{\Sigma}}(t^*)$ in (4.2) and (4.3). Consequently, by (4.2) and Lemma 4.1(iii) and (ii)

$$P\{\phi(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o)) \geq \phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}(t^*)) - d_o(\xi + 1)^2 K^2 L_n(p(p + 1)/\varepsilon)\}$$

(A.5)

$$\geq 1 - \varepsilon.$$

Let us take the sample mean of i-indexed quantities with weights $Y_i(t) \min\{M, \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta}^o)\}$, so that $\bar{\mathbf{Z}}_n(t; M)$ is the sample mean of $\mathbf{Z}_i(t)$. Since $V_n(t, \boldsymbol{\beta}^o)R_n(t, \boldsymbol{\beta}^o) = \hat{\mathbf{G}}_n(t; \infty)$,

$$\mathbf{u}'\hat{\mathbf{G}}_n(t; \infty)\mathbf{u} \geq \frac{1}{n}\sum_{i=1}^n [\mathbf{u}'\{\mathbf{Z}_i - \bar{\mathbf{Z}}_n(t; \infty)\}]^2 Y_i(t)\min\{M, \exp(\mathbf{Z}_i'(t)\boldsymbol{\beta}^o)\}$$

$$\geq \mathbf{u}'\hat{\mathbf{G}}_n(t; M)\mathbf{u}.$$

Thus, by the definition of $\bar{\boldsymbol{\Sigma}}(t^*; M)$ in (4.4) and Lemma 4.1(iii),

(A.6) $$\phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}(t^*)) \geq \phi(\xi, \mathcal{O}; \bar{\boldsymbol{\Sigma}}(t^*; M)).$$

In addition, the relationship between the sample second moment and variance gives

$$\mathbf{G}_n(t; M) = \hat{\mathbf{G}}_n(t; M) + \{\bar{\mathbf{Z}}_n(t; M) - \boldsymbol{\mu}(t; M)\}^{\otimes 2}$$

by the definition of $\mathbf{G}_n(t; M)$ and $\hat{\mathbf{G}}_n(t; M)$, so that (4.4) can be written as

$$\bar{\mathbf{\Sigma}}(t^*; M) = \int_0^{t^*} \mathbf{G}_n(s; M) \, d\Lambda_0(s)$$

(A.7)

$$- \int_0^{t^*} \{\bar{\mathbf{Z}}_n(t; M) - \boldsymbol{\mu}(t; M)\}^{\otimes 2} \, d\Lambda_0(s).$$

We first bound the second term on the right-hand side of (A.7). Define

$$R_n(t; M) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \min\{M, \exp(\mathbf{Z}'_i(t)\boldsymbol{\beta}^o)\},$$

$$\boldsymbol{\Delta}(t; M) = R_n(t; M)\{\bar{\mathbf{Z}}_n(t; M) - \boldsymbol{\mu}(t; M)\}$$

$$= \frac{1}{n} \sum_{i=1}^n Y_i(t) \min\{M, \exp(\mathbf{Z}'_i(t)\boldsymbol{\beta}^o)\}\{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t; M)\}.$$

Since $Y_i(t)$ is nonincreasing in t,

$$(A.8) \quad 0 \le \int_0^{t^*} \{\bar{\mathbf{Z}}_n(t; M) - \boldsymbol{\mu}(t; M)\}^{\otimes 2} \, d\Lambda_0(s) \le \frac{\int_0^{t^*} \boldsymbol{\Delta}^{\otimes 2}(t; M) \, d\Lambda_0(s)}{R_n^2(t^*, M)}.$$

Since $R_n(t^*, M)$ is the average of i.i.d. variables uniformly bounded by M and $E R_n(t^*, M) = r_*$, the Hoeffding (1963) inequality gives

$$P\{R_n(t^*, M) < r_*/2\} \le e^{-nr_*^2/(8M^2)}.$$

Since $\boldsymbol{\Delta}(t; M)$ is an average of i.i.d. mean zero vectors,

$$\left(n^2 \int_0^{t^*} \boldsymbol{\Delta}^{\otimes 2}(t; M) \, d\Lambda_0(s)\right)_{jk}$$

is a degenerate V-statistic for each (j, k). Moreover, since the summands of these V-statistics are all bounded by $K^2 \Lambda_0(t^*)$, Lemma 4.2 yields

$$\max_{1 \le j, k \le p} P\left\{\pm\left(\int_0^{t^*} \boldsymbol{\Delta}^{\otimes 2}(t; M) \, d\Lambda_0(s)\right)_{jk} > K^2 \Lambda_0(t^*)t^2\right\}$$

$$\le 2.221 \exp\left(\frac{-nt^2/2}{1 + t/3}\right).$$

Thus, by (A.7), (A.8), the above two probability bounds and Lemma 4.1(ii),

$$\phi(\xi, \mathcal{O}; \bar{\mathbf{\Sigma}}(t^*; M))$$

(A.9)

$$\ge \phi\left(\xi, \mathcal{O}; \int_0^{t^*} \mathbf{G}_n(s; M) \, d\Lambda_0(s)\right)$$

$$- d^o(\xi + 1)^2 K^2 \Lambda_0(t^*) t_{n,p,\varepsilon}^2 / (r_*/2)$$

with at least probability $1 - e^{-nr_*^2/(8M^2)} - \varepsilon$.

Finally, by (4.5), $\int_0^{t^*} \mathbf{G}_n(s; M) \, d\Lambda_0(s)$ is an average of i.i.d. matrices with mean $\mathbf{\Sigma}(t^*; M)$ and the summands of $(\int_0^{t^*} \mathbf{G}_n(s; M) \, d\Lambda_0(s))_{jk}$ are uniformly bounded by $K^2 \Lambda_0(t^*)$, so that the Hoeffding (1963) inequality gives

$$P\left\{ \max_{j,k} \left| \left(\int_0^{t^*} \mathbf{G}_n(s; M) \, d\Lambda_0(s) - \mathbf{\Sigma}(t^*, M) \right)_{jk} \right| > K^2 \Lambda_0(t^*) t \right\}$$
$$\leq p(p+1) e^{-nt^2/2}.$$

By (A.5), (A.6), (A.9), the above inequality with $t = L_n(p(p+1)/\varepsilon)$ and Lemma 4.1(ii),

$$\phi(\xi, \mathcal{O}; \ddot{\ell}(\boldsymbol{\beta}^o))$$
$$\geq \phi\left(\xi, \mathcal{O}; \int_0^{t^*} \mathbf{G}_n(s; M) \, d\Lambda_0(s) \right)$$
$$\quad - d_o(\xi+1)^2 K^2 \{ L_n(p(p+1)/\varepsilon) + (2/r_*)\Lambda_0(t^*) t_{n,p,\varepsilon}^2 \}$$
$$\geq \phi(\xi, \mathcal{O}; \mathbf{\Sigma}(t^*, M))$$
$$\quad - d_o(\xi+1)^2 K^2 \{ (1 + \Lambda_0(t^*)) L_n(p(p+1)/\varepsilon) + (2/r_*)\Lambda_0(t^*) t_{n,p,\varepsilon}^2 \}$$

with at least probability $1 - e^{-nr_*^2/(8M^2)} - 3\varepsilon$. Since

$$\phi(\xi, \mathcal{O}; \mathbf{\Sigma}(t^*, M)) \geq \mathrm{RE}^2(\xi, \mathcal{O}; \mathbf{\Sigma}(t^*, M)) \geq \rho_*$$

by Lemma 4.1(i) and the definition in (3.5), the conclusion follows. $\square$

REFERENCES

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646

AZUMA, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Math. J.* (2) **19** 357–367. MR0221571

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

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

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

COX, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34** 187–220. MR0341758

DE LA PEÑA, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.* **27** 537–564. MR1681153

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–451.

FAN, J. (1997). Comments on "Wavelets in statistics: A review," by A. Antoniadis. *J. Amer. Statist. Assoc.* **6** 131–138.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

GAÏFFAS, S. and GUILLOUX, A. (2012). High-dimensional additive hazards models and the Lasso. *Electron. J. Stat.* **6** 522–546. MR2988418

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

GUI, J. and LI, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21** 3001–3008.

HJORT, N. L. and POLLARD, D. (1993). Asymptotics for minimisers of convex processes. Preprint, Yale Univ.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

HUANG, J. and ZHANG, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *J. Mach. Learn. Res.* **13** 1839–1864. MR2956344

KOLTCHINSKII, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. MR2555200

KONG, S. and NAN, B. (2012). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. Available at arXiv:1204.1992.

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

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

NEGAHBAN, S., RAVIKUMAR, P., WAINWRIGHT, M. and YU, B. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Proceedings of the NIPS Conference*. Vancouver, Canada.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.

TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108. MR0600535

VAN DE GEER, S. A. (2007). The deterministic Lasso. Technical Report 140, ETH Zürich, Switzerland. Available at http://stat.ethz.ch/research/publ_archive/2007/140.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

ZHANG, T. (2009). On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.* **10** 555–568. MR2491749

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. MR2410017

ZHANG, C. H. and ZHANG, T. (2012). A general theory of concave regularization for high dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593.

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

J. HUANG
DEPARTMENT OF STATISTICS
  AND ACTUARIAL SCIENCE
UNIVERSITY OF IOWA
241 SH
IOWA CITY, IOWA 52242
USA
E-MAIL: jian-huang@uiowa.edu

T. SUN
STATISTICS DEPARTMENT
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
400 JON M. HUNTSMAN HALL
3730 WALNUT STREET
PHILADELPHIA, PENNSYLVANIA 19104-6340
USA
E-MAIL: tingni@wharton.upenn.edu

Z. YING
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
1255 AMSTERDAM AVENUE
NEW YORK, NEW YORK 10027
USA
E-MAIL: zying@stat.columbia.edu

Y. YU
SCHOOL OF MATHEMATICAL SCIENCES
FUDAN UNIVERSITY
SHANGHAI
CHINA
E-MAIL: yuyi@fudan.edu.cn

C.-H. ZHANG
DEPARTMENT OF STATISTICS
  AND BIOSTATISTICS
HILL CENTER, BUSCH CAMPUS
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: czhang@stat.rutgers.edu