Published in final edited form as:
Stat Sin. 2014 January 1; 24(1): 25–42. doi:10.5705/ss.2012.240.

Non-Asymptotic Oracle Inequalities for the High-Dimensional Cox Regression via Lasso

Shengchun Kong and **Bin Nan***
Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109-2029

Shengchun Kong: kongsc@umich.edu; Bin Nan: bnan@umich.edu

Abstract

We consider finite sample properties of the regularized high-dimensional Cox regression via lasso. Existing literature focuses on linear models or generalized linear models with Lipschitz loss functions, where the empirical risk functions are the summations of independent and identically distributed (iid) losses. The summands in the negative log partial likelihood function for censored survival data, however, are neither iid nor Lipschitz.We first approximate the negative log partial likelihood function by a sum of iid non-Lipschitz terms, then derive the non-asymptotic oracle inequalities for the lasso penalized Cox regression using pointwise arguments to tackle the difficulties caused by lacking iid Lipschitz losses.

Keywords and phrases

Cox regression; finite sample; lasso; oracle inequality; variable selection

1. Introduction

Since it was introduced by Tibshirani (1996), the lasso regularized method for high-dimensional regression models with sparse coefficients has received a great deal of attention in the literature. Properties of interest for such regression models include the finite sample oracle inequalities. Among the extensive literature of the lasso method, Bunea, Tsybakov, and Wegkamp (2007) and Bickel, Ritov, and Tsybakov (2009) derived the oracle inequalities for prediction risk and estimation error in a general nonparametric regression model, including the high-dimensional linear regression as a special example, and van de Geer (2008) provided oracle inequalities for the generalized linear models with Lipschitz loss functions, e.g., logistic regression and classification with hinge loss. Bunea (2008) and Bach (2010) also considered the lasso regularized logistic regression. For censored survival data, the lasso penalty has been applied to the regularized Cox regression in the literature, see e.g. Tibshirani (1997) and Gui and Li (2005), among others. Recently, Bradic, Fan, and Jiang (2011) studied the asymptotic properties of the lasso regularized Cox model. However, its finite sample non-asymptotic statistical properties have not yet been established in the literature to the best of our knowledge, largely due to lacking iid Lipschitz losses from the partial likelihood. Nonetheless, the lasso approach has been studied extensively in the literature for other models, see e.g. Martinussen and Scheike (2009) and Gaiffas and Guilloux (2012), among others, for the additive hazards model.

*Supported in part by NSF Grant DMS-1007590 and NIH grant R01-AG036802.

We consider the non-asymptotic statistical properties of the lasso regularized high-dimensional Cox regression. Let T be the survival time and C the censoring time. Suppose we observe a sequence of iid observations $(\mathbf{X}_i, Y_i, \Delta_i)$, $i = 1, \ldots, n$, where $\mathbf{X}_i = (X_{i1}, \cdots X_{im})$ are the m-dimensional covariates in $\mathscr{X}$ $Y_i = T_i \wedge C_i$, and $\Delta_i = I_{\{T_i \leq C_i\}}$. Due to a large amount of parallel material, we follow closely the notation in van de Geer (2008). Let

$$\mathscr{F} = \left\{ f_\theta(\mathbf{x}) = \sum_{k=1}^{m} \theta_k x_k, \theta \in \Theta \subset \mathbf{R}^{\mathbf{m}} \right\}.$$

Consider the Cox model (Cox (1972)):

$$\lambda(t|\mathbf{X}) = \lambda_0(t) e^{f_\theta(\mathbf{x})},$$

where θ is the parameter of interest and λ_0 is the unknown baseline hazard function. The negative log partial likelihood function for θ is

$$l_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left\{ f_\theta(\mathbf{X}_i) - \log \left[\frac{1}{n} \sum_{j=1}^{n} 1(Y_j \geq Y_i) e^{f_\theta(\mathbf{X}_j)} \right] \right\} \Delta_i. \quad (1.1)$$

The corresponding estimator with lasso penalty is denoted by

$$\widehat{\theta}_n := \arg \min_{\theta \in \Theta} \{ l_n(\theta) + \lambda_n I(\theta) \},$$

where $I(\theta) := \sum_{k=1}^{m} \sigma_k |\theta_k|$ is the weighted l_1 norm of the vector $\theta \in \mathbf{R}^{\mathbf{m}}$. van de Geer (2008) considered σ_k to be the square-root of the second moment of the k-th covariate X_k, either at the population level (fixed) or at the sample level (random). For normalized X_k, $\sigma_k = 1$. We consider fixed weights σ_k, $k = 1, \cdots, m$. The results for random weights can be easily obtained from the case with fixed weights following van de Geer (2008), and we leave the detailed calculation to interested readers.

Clearly the negative log partial likelihood (1.1) is a sum of non-iid random variables. For ease of calculation, consider an intermediate function as a "replacement" of the negative log partial likelihood function

$$\tilde{l}_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \{ f_\theta(\mathbf{X}_i) - \log \mu(Y_i; f_\theta) \} \Delta_i \quad (1.2)$$

that has the iid structure, but with an unknown population expectation

$$\mu(t; f_\theta) = E_{\mathbf{X},Y} \left\{ 1(Y \geq t) e^{f_\theta(\mathbf{X})} \right\}.$$

The negative log partial likelihood function (1.1) can then be viewed as a "working" model for the empirical loss function (1.2). The corresponding loss function is

$$\gamma_{f_\theta} = \gamma(f_\theta(\mathbf{X}), Y, \Delta) := -\{ f_\theta(\mathbf{X}) - \log \mu(Y; f_\theta) \} \Delta, \quad (1.3)$$

with expected loss

$$l(\theta) = -E_{Y,\Delta,\mathbf{X}}\left[\{f_\theta(\mathbf{X}) - \log\mu(Y;f_\theta)\}\,\Delta\right] = P\gamma_{f_\theta}, \quad (1.4)$$

where P denotes the distribution of $(Y, \Delta, \mathbf{X})$. Define the target function $\bar{f}$ as

$$\bar{f} := \arg\min_{f\in\mathscr{F}} P\gamma_f := f_{\bar{\theta}},$$

where $\bar{\theta} = \arg\min_{\theta\in\Theta} P\gamma_{f_\theta}$. It is well-known that $P\gamma_{f_\theta}$ is convex with respect to θ for the regular Cox model, see for example, Andersen and Gill (1982), thus the above minimum is unique if the Fisher information matrix of θ at $\bar{\theta}$ is non-singular. Define the excess risk of f by

$$\mathscr{E}(f) := P\gamma_f - P\gamma_{\bar{f}}.$$

It is desirable to show similar non-asymptotic oracle inequalities for the Cox regression model as in, for example, van de Geer (2008) for generalized linear models. That is, with large probability,

$$\mathscr{E}(f_{\hat{\theta}_n}) \le const. \times \min_{\theta\in\Theta}\{\mathscr{E}(f_\theta) + \mathscr{V}_\theta\}.$$

Here $\mathscr{V}_\theta$ is called the "estimation error", which is typically proportional to λ_n^2 times the number of nonzero elements in θ.

Note that the summands in the negative log partial likelihood function (1.1) are not iid, and the intermediate loss function $\gamma(\cdot, Y, \Delta)$ given in (1.3) is not Lipschitz. Hence the general result of van de Geer (2008) that requires iid Lipschitz loss functions does not apply to the Cox regression. We tackle the problem using pointwise arguments to obtain the oracle bounds of two types of errors: one is between empirical loss (1.2) and expected loss (1.4) without involving the Lipschitz requirement of van de Geer (2008), and one is between the negative log partial likelihood (1.1) and empirical loss (1.2) which establishes the iid approximation of non-iid losses. These steps distinguish our work from that of van de Geer (2008); we rely on the Mean Value Theorem with van de Geer's Lipschitz condition replaced by the similar, but much less restrictive, boundedness assumption for regression parameters in Bühlmann (2006).

The article is organized as follows. In Section 2, we provide assumptions that are used throughout the paper. In Section 3, we define several useful quantities followed by the main result. We then provide a detailed proof in Section 4 by introducing a series of lemmas and corollaries useful for deriving the oracle inequalities for the Cox model. To avoid duplicate material as much as possible, we refer to the preliminaries and some results in van de Geer (2008) from place to place in the proofs without providing much detail.

2. Assumptions

We impose five basic assumptions. Let $\|\cdot\|$ be the $L_2(P)$ norm and $\|\cdot\|_\infty$ the sup norm.

Assumption A. $K_m := \max_{1\le k\le m}\{\|X_k\|_\infty/\sigma_k\} < \infty$.

Assumption B. There exists an $\eta > 0$ and strictly convex increasing G, such that for all $\theta \in \Theta$ with $\|f_\theta - f\|_\infty \le \eta$, one has $\mathscr{E}(f_\theta) \ge G(\|f_\theta - f\|)$.

In particular, G can be chosen as a quadratic function with some constant C_0, i.e., $G(u) = u^2/C_0$, then the convex conjugate of function G, denoted by H, such that $uv \le G(u) + H(v)$ is also quadratic.

Assumption C. There exists a function $D(\cdot)$ on the subsets of the index set $\{1, \cdots m\}$, such that for all $\mathscr{K} \subset \{1, \cdots, m\}$, and for all $\theta \in \Theta$ and $\tilde{\theta} \in \Theta$, we have $\sum_{k\in\mathscr{K}} \sigma_k \left|\theta_k - \tilde{\theta}_k\right| \le \sqrt{D(\mathscr{K})}\|f_\theta - f_{\tilde{\theta}}\|$. Here, $D(\mathscr{K}$ is chosen to be the cardinal number of $\mathscr{K}$

Assumption D. $L_m := \sup_{\theta\in\Theta}\sum_{k=1}^m |\theta_k| < \infty$.

Assumption E. The observation time stops at a finite time $\tau > 0$, with $\xi := P(Y \ge \tau) > 0$.

Assumptions A, B, and C are identical to those in van de Geer (2008) with her ψ_k the identity function. Assumptions B and C can be easily verified for the random design setting where **X** is random (van de Geer (2008)) together with the usual assumption of non-singular Fisher information matrix at θ (and its neighborhood) for the Cox model. Assumption D has a similar flavor to the assumption (A2) in Bühlmann (2006) for the persistency property of boosting method in high-dimensional linear regression models, but is much less restrictive in the sense that L_m is allowed to depend on m in contrast with the fixed constant in Bühlmann (2006). Here it replaces the Lipschitz assumption in van de Geer (2008). Assumption E is commonly used for survival models with censored data, see for example, Andersen and Gill (1982). A straightforward extension of Assumption E is to allow τ (thus ξ) to depend on n.

From Assumptions A and D, we have, for any $\theta \in \Theta$,

$$e^{|f_\theta(\mathbf{X}_i)|} \le e^{K_m L_m \sigma_{(m)}} := U_m < \infty \quad (2.1)$$

for all i, where $\sigma_{(m)} = \max_{1\le k\le m} \sigma_k$. Note that U_m is allowed to depend on m.

3. Main result

Let $I(\theta) := \sum_{k=1}^m \sigma_k |\theta_k|$ be the l_1 norm of θ. For any θ and $\tilde{\theta}$ in Θ, denote

$$I_1(\theta \left| \tilde{\theta}\right.) := \sum_{k:\tilde{\theta}_k\neq 0} \sigma_k|\theta_k|, I_2(\theta|\tilde{\theta}) := I(\theta) - I_1(\theta \left| \tilde{\theta}\right.).$$

Consider the estimator

$$\widehat{\theta}_n := \arg\min_{\theta\in\Theta}\{l_n(\theta) + \lambda_n I(\theta)\}.$$

3.1. Useful quantities

We first define a set of useful quantities that are involved in the oracle inequalities.

- $\bar{a}_n = 4a_n$, $a_n = \sqrt{\frac{2K_m^2 \log(2m)}{n}} + \frac{K_m \log(2m)}{n}$.
- $r_1 > 0$, $b > 0$, $d > 1$, and $1 > \delta > 0$ are arbitrary constants.
- $d_b := d\left(\frac{b+d}{(d-1)b} \vee 1\right)$.
- $\bar{\lambda}_{n,0} = \bar{\lambda}^A_{n,0} + \bar{\lambda}^B_{n,0}$, where

$$\bar{\lambda}^A_{n,0} := \bar{\lambda}^A_{n,0}(r_1) := \bar{a}_n \left(1 + 2r_1 \sqrt{2\left(K_m^2 + \bar{a}_n K_m\right)} + \frac{4r_1^2 \bar{a}_n K_m}{3}\right), \bar{\lambda}^B_{n,0} := \bar{\lambda}^B_{n,0}(r_1) := \frac{2K_m U_m^2}{\xi}\left(2\bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}}\right).$$

- $\lambda_n := (1 + b)\bar{\lambda}_{n,0}$.
- $\delta_1 = (1 + b)^{-N_1}$ and $\delta_2 = (1 + b)^{-N_2}$ are arbitrary constants for some N_1 and N_2, where $N_1 \in \mathbf{N} := \{1, 2, \ldots\}$ and $N_2 \in \mathbf{N} \cup \{0\}$.
- $d(\delta_1, \delta_2) = 1 + \frac{1 + (d^2 - 1)\delta_1}{(d-1)(1-\delta_1)}\delta_2$.
- W is a fixed constant given in Lemma 4.3 for a class of empirical processes.
- $D_\theta := D(\{k : \theta_k \neq 0, k = 1, \ldots, m\})$ is the number of nonzero θ_k's, where $D(\cdot)$ is given in Assumption C.
- $\mathscr{V}_\theta := 2\delta H\left(\frac{2\lambda_n \sqrt{D_\theta}}{\delta}\right)$, where H is the convex conjugate of function G defined in Assumption B.
- $\theta^*_n := \arg\min_{\theta \in \Theta}\{\mathscr{E}(f_\theta) + \mathscr{V}_\theta\}$.
- $\varepsilon^*_n := (1+\delta)\mathscr{E}(f_{\theta^*_n}) + \mathscr{V}_{\theta^*_n}$.
- $\zeta^*_n := \frac{\varepsilon^*_n}{\lambda_{n,0}}$.
- $\theta(\varepsilon^*_n) := \arg\min_{\theta \in \Theta, I(\theta - \theta^*_n) \leq d_b \zeta^*_n / b}\{\delta\mathscr{E}(f_\theta) - 2\lambda_n I_1(\theta - \theta^*_n | \theta^*_n)\}$

In the above, the dependence of θ^*_n on the sample size n is through $\mathscr{V}_\theta$ that involves the tuning parameter λ_n. We also impose conditions as in van de Geer (2008):

Condition I(b, δ). $\|f_{\theta^*_n} - \overline{f}\|_\infty \leq \eta$.

Condition II(b, δ, d). $\|f_{\theta(\varepsilon^*_n)} - \overline{f}\|_\infty \leq \eta$.

In both conditions, η is given in Assumption B.

3.2. Oracle inequalities

We now provide our theorem on oracle inequalities for the Cox model lasso estimator, with detailed proof given in the next section. The key idea of the proof is to find bounds of differences between empirical errors of the working model (1.2) and between approximation errors of the partial likelihood, denoted as Z_θ and R_θ in the next section.

Theorem 3.1. *Suppose Assumptions A-E and Conditions I(b, δ) and II(b, δ, d) hold. With*

$$\Delta(b,\delta,\delta_1,\delta_2) := d(\delta_1,\delta_2)\frac{1-\delta^2}{\delta b} \vee 1,$$

we have, with probability at least

$$1-\left\{\log_{1+b}\frac{(1+b)^2\Delta(b,\delta,\delta_1,\delta_2)}{\delta_1\delta_2}\right\}\left\{\left(1+\frac{3}{10}W^2\right)\exp(-n\bar{a}_n^2 r_1^2)+2\exp(-n\xi^2/2)\right\},$$

that

$$\mathscr{E}(f_{\widehat{\theta}_n}) \le \frac{1}{1-\delta}\varepsilon_n^* and I(\widehat{\theta}_n-\theta_n^*) \le d(\delta_1,\delta_2)\frac{\zeta_n^*}{b}.$$

4. Proofs

4.1. Preparations

Denote the empirical probability measure based on the sample $\{(\mathbf{X}_i, Y_i, \Delta_i) : i = 1, \ldots n\}$ by P_n. Let $\varepsilon_1, \cdots, \varepsilon_n$ be a Rademacher sequence, independent of the training data $(\mathbf{X}_1, Y_1, \Delta_1), \cdots, (\mathbf{X}_n, Y_n, \Delta_n)$. For some fixed $\theta^* \in \Theta$ and some $M > 0$, denote $\mathscr{F}_M := \{f_\theta : \theta \in \Theta, I(\theta-\theta^*) \le M\}$. Later we take $\theta^*=\theta_n^*$, which is the case of interest. For any θ where $I(\theta-\theta^*) \le M$, denote

$$Z_\theta(M) := \left|(P_n-P)[\gamma_{f_\theta}-\gamma_{f_{\theta^*}}]\right| = \left|\left[\tilde{l}_n(\theta)-l(\theta)\right]-\left[\tilde{l}_n(\theta^*)-l(\theta^*)\right]\right|.$$

Note that van de Geer (2008) sought to bound $\sup_{f\in\mathscr{F}_M} Z_\theta(M)$, thus the contraction theorem of Ledoux and Talagrand (1991) (Theorem A.3 in van de Geer (2008)) was needed, which holds for Lipschitz functions. We find that the calculation in van de Geer (2008) does not apply to the Cox model due to the lack of Lipschitz property. However, the pointwise argument is adequate for our purpose because only the lasso estimator or the difference between the lasso estimator $\widehat{\theta_n}$ and the oracle θ_n^* is of interest. Note the notational difference between an arbitrary θ^* in the above $Z_\theta(M)$ and the oracle θ_n^*.

Lemma 4.1. *Under Assumptions A, D, and E, for all θ satisfying $I(\theta-\theta^*) \le M$, we have $EZ_\theta(M) \le a_n\bar{M}$.*

Proof. By the symmetrization theorem, see e.g. van der Vaart and Wellner (1996) or Theorem A.2 in van de Geer (2008), for a class of only one function we have

$$EZ_\theta(M) \le 2E\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\{[f_\theta(\mathbf{X}_i) - \log\mu(Y_i;f_\theta)]\Delta_i - [f_{\theta^*}(\mathbf{X}_i) - \log\mu(Y_i;f_{\theta^*})]\Delta_i\}\right|\right) \le 2E\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\{f_\theta(\mathbf{X}_i) - f_{\theta^*}(\mathbf{X}_i)\}\Delta_i\right|\right)$$
$$+2E\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\{\log\mu(Y_i;f_\theta) - \log\mu(Y_i;f_{\theta^*})\}\Delta_i\right|\right) = A + B.$$

For *A* we have

$$A \le 2\left(\sum_{k=1}^{m}\sigma_k\left|\theta_k - \theta_k^*\right|\right) E\left(\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta_i X_{ik}/\sigma_k\right|\right).$$

Applying Lemma A.1 in van de Geer (2008), we obtain

$$E\left(\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta_i\frac{X_{ik}}{\sigma_k}\right|\right) \le a_n.$$

Thus we have

$$A \le 2a_n M. \quad (4.1)$$

For *B*, instead of using the contraction theorem that requires Lipschitz, we use the Mean Value Theorem:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\{\log\mu(Y_i;f_\theta) - \log\mu(Y_i;f_{\theta^*})\}\Delta_i\right|$$
$$= \left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta_i\sum_{k=1}^{m}\frac{1}{\mu(Y_i;f_{\theta^{**}})}\int_{Y_i}^{\infty}\int_{\mathcal{X}}(\theta_k - \theta_k^*)x_k e^{f_{\theta^{**}}(x)}dP_{\mathbf{X},Y}(\mathbf{x},y)\right|$$
$$= \left|\sum_{k=1}^{m}\sigma_k(\theta_k - \theta_k^*)\frac{1}{n}\sum_{i=1}^{n}\frac{\varepsilon_i\Delta_i}{\mu(Y_i;f_{\theta^{**}})\sigma_k}\int_{Y_i}^{\infty}\int_{\mathcal{X}}x_k e^{f_{\theta^{**}}(x)}dP_{\mathbf{X},Y}(\mathbf{x},y)\right| \le \left|\sum_{k=1}^{m}\sigma_k(\theta_k - \theta_k^*)\right|\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta_i F_{\theta^{**}}(k,Y_i)\right| \le M\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta_i F_{\theta^{**}}(k,Y_i)\right|,$$

where θ** is between θ and θ*, and

$$F_{\theta^{**}}(k,t) = \frac{E[1(Y \ge t)X_k e^{f_{\theta^{**}}(\mathbf{X})}}{\mu(t;f_{\theta^{**}})\sigma_k} \quad (4.2)$$

satisfying

$$|F_{\theta^{**}}(k,t)| \leq \frac{(\|X_k\|_\infty/\sigma_k)E[1(Y \geq t)e^{f_{\theta^{**}}(\mathbf{X})}]}{\mu(t;f_{\theta^{**}})} \leq K_m.$$

Since for all *i*,

$$E[\varepsilon_i\Delta_i F_{\theta^{**}}(k,Y_i)]=0, \|\varepsilon_i\Delta_i F_{\theta^{**}}(k,Y_i)\|_\infty \leq K_m, \text{and} \frac{1}{n}\sum_{i=1}^n E[\varepsilon_i\Delta_i F_{\theta^{**}}(k,Y_i)]^2 \leq \frac{1}{n}\sum_{i=1}^n E[F_{\theta^{**}}(k,Y_i)]^2 \leq EK_m^2=K_m^2,$$

following Lemma A.1 in van de Geer (2008), we obtain

$$B \leq 2a_n M. \quad (4.3)$$

Combining (4.1) and (4.3), the upper bound for $EZ_\theta(M)$ is achieved.

We now can bound the tail probability of $Z_\theta(M)$ using the Bousquet's concentration theorem noted as Theorem A.1 in van de Geer (2008).

Corollary 4.1. *Under Assumptions A, D, and E, for all $M > 0$, $r_1 > 0$ and all θ satisfying $I(\theta - \theta^*) \leq M$, it holds that*

$$P(Z_\theta(M) \geq \overline{\lambda}_{n,0}^A M) \leq \exp(-n\overline{a}_n^2 r_1^2).$$

Proof. Using the triangular inequality and the Mean Value Theorem, we obtain

$$\begin{aligned}
|\gamma_{f_\theta} - \gamma_{f_{\theta^*}}| \leq\ & |f_\theta(\mathbf{X}) \\
& - f_{\theta^*}(\mathbf{X})|\Delta \\
& + |\log\mu(Y;f_\theta) \\
& - \log\mu(Y;f_{\theta^*})|\Delta \leq \sum_{k=1}^m \sigma_k|\theta_k \\
& - \theta_k^*|\frac{|X_k|}{\sigma_k} \\
& + |\log\mu(Y;f_\theta) \\
& - \log\mu(Y;f_{\theta^*})| \leq MK_m \\
& + \sum_{k=1}^m \sigma_k|\theta_k \\
& - \theta_k^*| \cdot \max_{1\leq k\leq m} |F_{\theta^{**}}(k,Y) \leq 2MK_m,
\end{aligned}$$

where θ^{**} is between θ and θ^*, and $F_{\theta^{**}}(k, Y)$ is defined in (4.2). So we have

$$\|\gamma_{f_\theta} - \gamma_{f_{\theta^*}}\|_\infty \leq 2MK_m, \text{and} P(\gamma_{f_\theta} - \gamma_{f_{\theta^*}})^2 \leq 4M^2K_m^2.$$

Therefore, in view of Bousquet's concentration theorem and Lemma 4.1, for all $M > 0$ and $r_1 > 0$,

$$P\left(Z_\theta(M) \geq \overline{a}_n M\left(1+2r_1\sqrt{2(K_m^2+\overline{a}_n K_m)}+\frac{4r_1^2\overline{a}_n K_m}{3}\right)\right) \leq \exp\left(-n\overline{a}_n^2 r_1^2\right).$$

Now for any θ satisfying $I(\theta - \theta^*) \leq M$, we bound

$$\begin{aligned} R_\theta(M) :=& \left|\left[l_n(\theta) - \tilde{l}_n(\theta)\right]\right. \\ &\left. - \left[l_n(\theta^*) - \tilde{l}_n(\theta^*)\right]\right| \\ =& \frac{1}{n}\sum_{i=1}^n \left| \left[\log\frac{1}{n}\sum_{j=1}^n \frac{1(Y_j \geq Y_i)e^{f_\theta(\mathbf{X}_j)}}{\mu(Y_i;f_\theta)}\right.\right. \\ &\left.\left. - \log\frac{1}{n}\sum_{j=1}^n \frac{1(Y_j \geq Y_i)e^{f_{\theta^*}(\mathbf{X}_j)}}{\mu(Y_i;f_{\theta^*})}\right]\Delta_i\right| \leq \sup_{0\leq t\leq\tau}\left|\log\frac{1}{n}\sum_{j=1}^n \frac{1(Y_j \geq t)e^{f_\theta(\mathbf{X}_j)}}{\mu(t;f_\theta)}\right. \\ &\left. - \log\frac{1}{n}\sum_{j=1}^n \frac{1(Y_j \geq t)e^{f_{\theta*}(\mathbf{X}_j)}}{\mu(t;f_{\theta*})}\right|. \end{aligned}$$

Here recall that τ is given in Assumption E. By the Mean Value Theorem, we have

$$\begin{aligned} R_\theta(M) \leq& \sup_{0\leq t\leq\tau}\left|\sum_{k=1}^m (\theta_k\right. \\ &\left. - \theta_k^*)\left\{\frac{\sum_{j=1}^n 1(Y_j \geq t)e^{f_{\theta^{**}}(\mathbf{X}_j)}}{\mu(t;f_{\theta^{**}})}\right\}^{-1}\left\{\frac{\sum_{j=1}^n 1(Y_j \geq t)X_{jk}e^{f_{\theta^{**}}(\mathbf{X}_j)}}{\mu(t;f_{\theta^{**}})}\right. \\ &\left.\left. - \frac{\sum_{j=1}^n 1(Y_j \geq t)e^{f_{\theta^{**}}(\mathbf{X}_j)}E[1(Y \geq t)X_k e^{f_{\theta^{**}}(\mathbf{X})}]}{\mu(t;f_{\theta^{**}})^2}\right\}\right| \\ =& \sup_{0\leq t\leq\tau}\left|\sum_{k=1}^m \sigma_k(\theta_k\right. \\ &\left. - \theta_k^*)\left\{\frac{\sum_{j=1}^n 1(Y_j \geq t)(X_{jk}/\sigma_k)e^{f_{\theta^{**}}(\mathbf{X}_j)}}{\sum_{j=1}^n 1(Y_j \geq t)e^{f_{\theta**}(\mathbf{X}_j)}} - \frac{E[1(Y \geq t)(X_k/\sigma_k)e^{f_{\theta^{**}}(\mathbf{X})}]}{E[1(Y \geq t)e^{f_{\theta**}(\mathbf{X})}]}\right\}\right| \leq M \sup_{0\leq t\leq\tau}\left[\frac{1}{n}\sum_{i=1}^n 1(Y_i \geq t)e^{f_{\theta^{**}}(\mathbf{X}_i)}\right]^{-1} \text{su} \\ &\left. - E[1(Y \geq t)(X_k/\sigma_k)e^{f_{\theta^{**}}(\mathbf{X})}]\right| + K_m\left|\frac{1}{n}\sum_{i=1}^n 1(Y_i \geq t)e^{f_{\theta^{**}}(\mathbf{X}_i)} - E[1(Y \geq t)e^{f_{\theta^{**}}(\mathbf{X})}]\right|, \end{aligned} \quad (4.4$$

where θ** is between θ and θ* and, by (2.1), we have

$$\sup_{0\leq t\leq\tau}\left[\frac{1}{n}\sum_{i=1}^n 1(Y_i \geq t)e^{f_{\theta^{**}}(\mathbf{X}_i)}\right]^{-1} \leq U_m\left[\frac{1}{n}\sum_{i=1}^n 1(Y_i \geq \tau)\right]^{-1}. \quad (4.5)$$

Lemma 4.2. *Under Assumption E, we have*

$$P\left(\frac{1}{n}\sum_{i=1}^n 1(Y_i \geq \tau) \leq \frac{\xi}{2}\right) \leq 2e^{-n\xi^2/2}.$$

Proof. This is obtained directly from Massart (1990) for the Kolmogorov statistic by taking $r=\xi\sqrt{n}/2$ in the following:

$$P\left(\frac{1}{n}\sum_{i=1}^{n}1(Y_i\geq\tau)\leq\frac{\xi}{2}\right)\leq P\left(\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}1(Y_i\geq\tau)-\xi\right|\geq r\right)\leq P\left(\sup_{0\leq t\leq\tau}\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}1(Y_i\geq t)-P(Y\geq t)\right|\geq r\leq 2e^{-2r^2}.\right.$$

Lemma 4.3. *Under Assumptions A, D, and E, for all* θ *we have*

$$P\left(\sup_{0\leq t\leq\tau}\left|\frac{1}{n}\sum_{i=1}^{n}1(Y_i\geq t)e^{f_\theta(\mathbf{X}_i)}-\mu(t;f_\theta)\right|\geq U_m\overline{a}_n r_1\right)\leq\frac{1}{5}W^2e^{-n\overline{a}_n^2r_1^2},\quad(4.6)$$

where W is a fixed constant.

Proof. For a class of functions indexed by t, $\mathscr{F}=\{1(y\geq t)e^{f\theta(x)}/U_m : t\in[0,\tau], y\in\mathbf{R}, e^{f\theta(x)}\leq U_m\}$, we calculate its bracketing number. For any nontrivial $\in$ satisfying $1>\in>0$, let t_i be the i-th $\lceil 1/\varepsilon\rceil$ quantile of Y, so

$$P(Y\leq t_i)=i\varepsilon, i=1,\ldots,\lceil 1/\varepsilon\rceil-1,$$

where $\lceil x\rceil$ is the smallest integer that is greater than or equal to x. Furthermore, take $t_0=0$ and $t_{\lceil 1/\varepsilon\rceil}=+\infty$. For $i=1,\cdots,\lceil 1/\varepsilon\rceil$, define brackets $[L_i, U_i]$ with

$$L_i(x,y)=1(y\geq t_i)e^{f_\theta(x)}/U_m, U_i(x,y)=1(y>t_{i-1})e^{f_\theta(x)}/U_m$$

such that $L_i(x,y)\leq 1(y\geq t)e^{f\theta(x)}/U_m\leq U_i(x,y)$ when $t_{i-1}<t\leq t_i$. Since

$$\{E[U_i-L_i]^2\}^{1/2}\leq\left\{E\left[\frac{e^{f_\theta(\mathbf{X})}}{U_m}\{1(Y\geq t_i)-1(Y>t_{i-1})\}\right]^2\right\}^{1/2}\leq\{P(t_{i-1}<Y\leq t_i)\}^{1/2}=\sqrt{\varepsilon},$$

we have $N_{[]}(\sqrt{\varepsilon},\mathscr{F},L_2)\leq\lceil 1/\varepsilon\rceil\leq 2/\varepsilon$, which yields

$$N_{[]}(\varepsilon,\mathscr{F},L_2)\leq\frac{2}{\varepsilon^2}=\left(\frac{K}{\varepsilon}\right)^2,$$

where $K=\sqrt{2}$. Thus, from Theorem 2.14.9 in van der Vaart and Wellner (1996), we have for any $r>0$,

$$P\left(\sqrt{n}\sup_{0\leq t\leq\tau}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1(Y_i\geq t)e^{f_\theta(\mathbf{X}_i)}}{U_m}-\frac{\mu(t;f_\theta)}{U_m}\right|\geq r\right)\leq\frac{1}{2}W^2r^2e^{-2r^2}\leq\frac{1}{5}W^2e^{-r^2},$$

where W is a constant that only depends on K. Note that $r^2e^{-r^2}$ is bounded by e^{-1}. With $r=\sqrt{n}\overline{a}_n r_1$, we obtain (4.6).

Lemma 4.4. *Under Assumptions A, D, and E, for all θ we have*

$$P\left(\sup_{0\le t\le\tau}\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}1(Y_i\ge t)\frac{X_{ik}}{\sigma_k}e^{f_\theta(\mathbf{X}_i)}-E\left[1(Y\ge t)\frac{X_k}{\sigma_k}e^{f_\theta(\mathbf{X})}\right]\right|\ge K_mU_m\left[\overline{a}_nr_1+\sqrt{\frac{\log(2m)}{n}}\right]\right)\le\frac{1}{10}W^2e^{-n\overline{a}_n^2r_1^2}. \tag{4.7}$$

Proof. Consider the classes of functions indexed by *t*,

$$\mathscr{G}^k=\{1(y\ge t)e^{f_\theta(x)}x_k/(\sigma_kK_mU_m):t\in[0,\tau],y\in\mathbf{R},\left|e^{f_\theta(x)}x_k/\sigma_k\right|\le K_mU_m\},k=1,\ldots,m.$$

Using the argument in the proof of Lemma 4.3, we have

$$N_{[]}(\varepsilon,\mathscr{G}^k,L_2)\le\left(\frac{K}{\varepsilon}\right)^2,$$

where $K=\sqrt{2}$, and then for any *r* > 0,

$$P\left(\sqrt{n}\sup_{0\le t\le\tau}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1(Y_i\ge t)e^{f_\theta(\mathbf{X}_i)}X_{ik}}{\sigma_kK_mU_m}-E\left[\frac{1(Y\ge t)e^{f_\theta(\mathbf{X})}X_k}{\sigma_kK_mU_m}\right]\right|\ge r\right)\le\frac{1}{5}W^2e^{-r^2}.$$

Thus we have

$$\begin{aligned}P\Bigg(&\sqrt{n}\sup_{0\le t\le\tau}\max_{1\le k\le m}\left|\frac{1}{n}\sum_{i=1}^{n}1(Y_i\ge t)e^{f_\theta(\mathbf{X}_i)}X_{ik}/(\sigma_kU_mK_m)\right.\\&\left.-E[1(Y\ge t)e^{f_\theta(\mathbf{X})}X_k/(\sigma_kU_mK_m)]\right|\ge r\le P\Bigg(\bigcup_{k=1}^{m}\sqrt{n}\sup_{0\le t\le\tau}\Big|\frac{1}{n}\sum_{i=1}^{n}1(Y_i\ge t)e^{f_\theta(\mathbf{X}_i)}X_{ik}/(\sigma_kU_mK_m)-E[1-(Y\ge t)e^{f_\theta(}\\&\left.-E[1(Y\ge t)e^{f_\theta(\mathbf{X})}X_k/(\sigma_kU_mK_m)]\right|\ge r\le\frac{m}{5}W^2e^{-r^2}\\=&\frac{1}{10}W^2e^{\log(2m)-r^2}.\end{aligned}$$

Let $\log(2m)-r^2=-n\overline{a}_n^2r_1^2$, so $r=\sqrt{n\overline{a}_n^2r_1^2+\log(2m)}$. Since

$$\sqrt{\overline{a}_n^2r_1^2+\frac{\log(2m)}{n}}\le\overline{a}_nr_1+\sqrt{\frac{\log(2m)}{n}},$$

, we obtain (4.7).

Corollary 4.2. *Under Assumptions A, D, and E, for all M > 0, r_1 > 0, and all θ that satisfy $I(\theta-\theta^*)\le M$, we have*

$$P(R_\theta(M) \geq \overline{\lambda}_{n,0}^B M) \leq 2\exp(-n\xi^2/2) + \frac{3}{10} W^2 \exp(-n\overline{a}_n^2 r_1^2). \quad (4.8)$$

Proof. From (4.4) and (4.5) we have

$$P(R_\theta(M) \leq \overline{\lambda}_{n,0}^B \cdot M) \geq P(E_1^c \cap E_2^c \cap E_3^c),$$

where the events E_1, E_2 and E_3 are defined as

$$E_1 = \left\{ \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq \tau) \leq \xi/2 \right\},$$

$$E_2 = \left\{ \sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) e^{f_{\theta^{**}}(\mathbf{X}_i)} - \mu(t; f_{\theta^{**}}) \right| \geq U_m \overline{a}_n r_1 \right\},$$

$$E_3 = \left\{ \max_{1 \leq k \leq m} \sup_{0 \leq t \leq \tau} \left| \frac{1}{n} \sum_{i=1}^n 1(Y_i \geq t) \frac{X_{ik}}{\sigma_k} e^{f_{\theta^{**}}(\mathbf{X}_i)} - E\left[1(Y \geq t) \frac{X_k}{\sigma_k} e^{f_{\theta^{**}}(\mathbf{X})} \right] \right| \geq K_m U_m (\overline{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}}) \right\}.$$

Thus

$$P(R_\theta(M) \geq \overline{\lambda}_{n,0}^B \cdot M) \leq P(E_1) + P(E_2) + P(E_3),$$

and the result follows from Lemmas 4.2, 4.3 and 4.4.

Now with $\theta^* = \theta_n^*$, we have the following results.

Lemma 4.5. *Suppose Conditions I(b, δ) and II(b, δ, d) are met. Under Assumptions B and C, for all $\theta \in \Theta$ with $I(\theta - \theta_n^*) \leq d_b \zeta_n^*/b$, it holds that*

$$2\lambda_n I_1(\theta - \theta_n^*) \leq \delta \mathscr{E}(f_\theta) + \varepsilon_n^* - \mathscr{E}(f_{\theta_n^*}).$$

Proof. The proof is exactly the same as that of Lemma A.4 in van de Geer (2008), with the λ_n defined in Subsection 3.1.

Lemma 4.6. *Suppose Conditions I(b, δ) and II(b, δ, d) are met. Consider any random $\tilde{\theta} \in \Theta$ with $l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta}) \leq l_n(\theta_n^*) + \lambda_n I(\theta_n^*)$. Let $1 < d_0 \leq d_b$. It holds that*

$$P\left(I(\tilde{\theta} - \theta_n^*) \leq d_0 \frac{\zeta_n^*}{b} \right) \leq P\left(I(\tilde{\theta} - \theta_n^*) \leq \left(\frac{d_0 + b}{1 + b} \right) \frac{\zeta_n^*}{b} \right) + \left(1 + \frac{3}{10} W^2 \right) \exp\left(-n\overline{a}_n^2 r_1^2 \right) + 2\exp(-n\xi^2/2).$$

Proof. The idea is similar to the proof of Lemma A.5 in van de Geer (2008). Let $\tilde{\mathscr{E}} = \mathscr{E}(f_{\tilde{\theta}})$ and $\mathscr{E}^* = \mathscr{E}(f_{\theta_n^*})$. We will use short notation: $I_1(\theta) = I_1(\theta|\theta_n^*)$ and $I_2(\theta) = I_2(\theta|\theta_n^*)$. Since $l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta}) \le l_n(\theta_n^*) + \lambda_n I(\theta_n^*)$, on the set where $I(\tilde{\theta} - \theta_n^*) \le d_0\zeta_n^*/b$ and $Z_{\tilde{\theta}}(d_0\zeta_n^*/b) \le d_0\zeta_n^*/b \cdot \overline{\lambda}_{n,0}^A$, we have

$$\begin{aligned} R_{\tilde{\theta}}(d_0\zeta_n^*/b) \ge & [l_n(\theta_n^*) \\ & + \lambda_n I(\theta_n^*)] \\ & - [l_n(\tilde{\theta}) \\ & + \lambda_n I(\tilde{\theta})] - \lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) - [\tilde{l}_n(\theta_n^*) \\ & - \tilde{l}_n(\tilde{\theta})] \ge \\ & - \lambda_n I(\theta_n^*) \\ & + \lambda_n I(\tilde{\theta}) \\ & - [\tilde{l}_n(\theta_n^*) \\ & - \tilde{l}_n(\tilde{\theta})] \ge \\ & - \lambda_n I(\theta_n^*) \\ & + \lambda_n I(\tilde{\theta}) \\ & - [l(\theta_n^*) \\ & - l(\tilde{\theta}) \\ & - d_0\zeta_n^*/b \\ & \cdot \overline{\lambda}_{n,0}^A \ge -\lambda_n I(\theta_n^*) \\ & + \lambda_n I(\tilde{\theta}) \\ & - \mathscr{E}^* + \tilde{\mathscr{E}} - d_0\lambda_{n,0}^A\zeta_n^*/b. \end{aligned} \tag{4.9}$$

By (4.8) we know that $R_{\tilde{\theta}}(d_0\zeta_n^*/b)$ is bounded by $d_0\overline{\lambda}_{n,0}^B\zeta_n^*/b$ with probability at least $1 - \frac{3}{10}W^2\exp(-n\overline{a}_n^2 r_1^2) - 2\exp(-n\xi^2/2)$, then we have

$$\tilde{\mathscr{E}} + \lambda_n I(\tilde{\theta}) \le \overline{\lambda}_{n,0}^B d_0\zeta_n^*/b + \mathscr{E}^* + \lambda_n I(\theta_n^*) + \overline{\lambda}_{n,0}^A d_0\zeta_n^*/b.$$

Since $I(\tilde{\theta}) = I_1(\tilde{\theta}) + I_2(\tilde{\theta})$ and $I(\theta_n^*) = I_1(\theta_n^*)$, using the triangular inequality, we obtain

$$\tilde{\mathscr{E}} + (1+b)\overline{\lambda}_{n,0}I_2(\tilde{\theta}) \le \overline{\lambda}_{n,0}d_0\zeta_n^*/b + \mathscr{E}^* + (1+b)\overline{\lambda}_{n,0}I_1(\theta_n^*) - (1+b)\overline{\lambda}_{n,0}I_1(\tilde{\theta}) \le \overline{\lambda}_{n,0}d_0\zeta_n^*/b + \mathscr{E}^* + (1+b)\overline{\lambda}_{n,0}I_1(\tilde{\theta} - \theta_n^*). \tag{4.10}$$

Adding $(1+b)\overline{\lambda}_{n,0}I_1(\tilde{\theta} - \theta_n^*)$ to both sides and from Lemma 4.5,

$$\tilde{\mathscr{E}} + (1+b)\overline{\lambda}_{n,0}I(\tilde{\theta} - \theta_n^*) \le \lambda_{n,0}d_0\frac{\zeta_n^*}{b} + \mathscr{E}^* + 2(1+b)\overline{\lambda}_{n,0}I_1(\tilde{\theta} - \theta_n^*) \le (\overline{\lambda}_{n,0}d_0 + b\overline{\lambda}_{n,0})\frac{\zeta_n^*}{b} + \delta\tilde{\mathscr{E}} = (d_0+b)\overline{\lambda}_{n,0}\frac{\zeta_n^*}{b} + \delta\tilde{\mathscr{E}}.$$

Because $0 < \delta < 1$, it follows that

$$I(\tilde{\theta}-\theta_n^*) \le \frac{d_0+b}{1+b}\frac{\zeta_n^*}{b}.$$

Hence,

$$P\left(\left\{I(\tilde{\theta}-\theta_n^*) \le d_0\frac{\zeta_n^*}{b}\right\} \cap \left\{Z_{\tilde{\theta}}(d_0\zeta_n^*/b) \le d_0\overline{\lambda}_{n,0}^A\frac{\zeta_n^*}{b}\right\} \cap \left\{R_{\tilde{\theta}}(d_0\zeta_n^*/b) \le d_0\overline{\lambda}_{n,0}^B\frac{\zeta_n^*}{b}\right\}\right) \le P\left(I(\tilde{\theta}-\theta_n^*) \le \frac{d_0+b}{1+b}\frac{\zeta_n^*}{b}\right),$$

which yields the desired result.

Corollary 4.3. *Suppose Conditions I(b, δ) and II(b, δ, d) are met. Consider any random* $\tilde{\theta} \in \Theta$ *with* $l_n(\tilde{\theta})+\lambda_n I(\tilde{\theta}) \le l_n(\theta_n^*)+\lambda_n I(\theta_n^*)$. *Let* $1 < d_0 \le d_b$. *It holds that*

$$P\left(I(\tilde{\theta}-\theta_n^*) \le d_0\frac{\zeta_n^*}{b}\right) \le P\left(I(\tilde{\theta}-\theta_n^*) \le [1+(d_0-1)(1+b)^{-N}]\frac{\zeta_n^*}{b}\right) + N\left\{\left(1+\frac{3}{10}W^2\right)\exp(-n\overline{a}_n^2 r_1^2)+2\exp(-n\xi^2/2)\right\}.$$

Proof. Repeat Lemma 4.6 *N* times.

Lemma 4.7. *Suppose Conditions I(b, δ) and II(b, δ, d) hold. If* $\tilde{\theta}_s = s\widehat{\theta}_n+(1-s)\theta_n^*$, *where*

$$s = \frac{d\zeta_n^*}{d\zeta_n^*+bI(\widehat{\theta}_n-\theta_n^*)},$$

then for any integer N, with probability at least

$$1-N\left\{\left(1+\frac{3}{10}W^2\right)\exp(-n\overline{a}_n^2 r_1^2)+2\exp(-n\xi^2/2)\right\},$$

we have

$$I(\tilde{\theta}_s-\theta_n^*) \le (1+(d-1)(1+b)^{-N}\frac{\zeta_n^*}{b}.$$

Proof. Since the negative log partial likelihood $l_n(\theta)$ and the lasso penalty are both convex with respect to θ, applying Corollary 4.3, we obtain the above inequality. This proof is similar to the proof of Lemma A.6 in van de Geer (2008).

Lemma 4.8. *Suppose Conditions I(b, δ) and II(b, δ, d) are met. Let* $N_1 \in \mathbf{N} := \{1, 2, \ldots\}$ *and* $N_2 \in \mathbf{N} \cup \{0\}$. *With* $\delta_1 = (1+b)^{-N_1}$ *and* $\delta_2 = (1+b)^{-N_2}$, *for any n, with probability at least*

$$1-(N_1+N_2)\left\{\left(1+\frac{3}{10}W^2\right)\exp(-n\overline{a}_n^2 r_1^2)+2\exp(-n\xi^2/2)\right\},$$

we have

$$I(\widehat{\theta}_n - \theta_n^*) \le d(\delta_1, \delta_2)\frac{\zeta_n^*}{b},$$

where

$$d(\delta_1, \delta_2)=1+\frac{1+(d^2-1)\delta_1}{(d-1)(1-\delta_1)}\delta_2.$$

Proof. The proof is the same as that of Lemma A.7 in van de Geer (2008), with a slightly different probability bound.

4.2. Proof of Theorem 3.1

Proof. The proof follows the same ideas in the proof of Theorem A.4 in van de Geer (2008), with exceptions of pointwise arguments and slightly different probability bounds. Since this is our main result, we provide a detailed proof here despite the amount of overlaps.

Define $\hat{\mathscr{E}} := \mathscr{E}(f_{\hat{\theta}_n})$ and $\mathscr{E}^* := \mathscr{E}(f_{\theta_n^*})$; use the notation $I_1(\theta) := I_1(\theta|\theta_n^*)$ and $I_2(\theta) := I_2(\theta|\theta_n^*)$; set $c := \delta b/(1-\delta^2)$. Consider the cases (a) $c < d(\delta_1, \delta_2)$ and (b) $c \ge d(\delta_1, \delta_2)$.

(a) $c < d(\delta_1, \delta_2)$. Let J be an integer satisfying $(1+b)^{J-1} c \le d(\delta_1, \delta_2)$ and $(1+b)^J c > d(\delta_1, \delta_2)$. We consider the cases (a1) $c\zeta_n^*/b < I(\widehat{\theta}_n - \theta_n^*) \le d(\delta_1, \delta_2)\zeta_n^*/b$ and (a2) $I(\widehat{\theta}_n - \theta_n^*) \le c\zeta_n^*/b$.

(a1) If $c\zeta_n^*/b < I(\widehat{\theta}_n - \theta_n^*) \le d(\delta_1, \delta_2)\zeta_n^*/b$, then

$$(1+b)^{j-1}c\frac{\zeta_n^*}{b} < I(\widehat{\theta}_n - \theta_n^*) \le (1+b)^j c\frac{\zeta_n^*}{b}$$

for some $j \in \{1, \cdots, J\}$. Let $d_0 = c(1+b)^{j-1} \le d(\delta_1, \delta_2) \le d_b$. From Corollary 4.1, with probability at least $1 - \exp(-n\overline{a}_n^2 r_1^2)$ we have $Z_{\widehat{\theta}_n}((1+b)d_0\zeta_n^*/b) \le (1+b)d_0\overline{\lambda}_{n,0}^A\zeta_n^*/b)$.

Since $l_n(\tilde{\theta}_n)+\lambda_n I(\tilde{\theta}_n) \le l_n(\theta_n^*)+\lambda_n I(\theta_n^*)$, from (4.9) we have

$$\hat{\mathscr{E}}+\lambda_n I(\widehat{\theta}_n) \le R_{\widehat{\theta}_n}\left((1+b)d_0\frac{\zeta_n^*}{b}\right)+\mathscr{E}^*+\lambda_n I(\theta_n^*)+(1+b)\overline{\lambda}_{n,0}^A d_0\frac{\zeta_n^*}{b}.$$

By (4.8), $R_{\widehat{\theta}_n}((1+b)d_0\zeta_n^*/b)$ is bounded by $(1+b)\overline{\lambda}_{n,0}^B d_0\zeta_n^*/b)$ with probability at least

$$1 - \frac{3}{10}W^2\exp(-n\overline{a}_n^2 r_1^2) - 2\exp(-n\xi^2/2).$$

Then we have

$$\widehat{\mathscr{E}}+(1+b)\overline{\lambda}_{n,0}I(\widehat{\theta}_n)\leq(1+b)\overline{\lambda}^{B}_{n,0}d_0\frac{\zeta^*_n}{b}+\mathscr{E}^*+(1+b)\overline{\lambda}_{n,0}I(\theta^*_n)+(1+b)\overline{\lambda}^{A}_{n,0}d_0\frac{\zeta^*_n}{b}\leq(1+b)\overline{\lambda}_{n,0}I(\widehat{\theta}_n-\theta^*_n)+\mathscr{E}^*+(1+b)\overline{\lambda}_{n,0}I(\theta^*_n).$$

Since $I(\widehat{\theta}_n)=I_1(\widehat{\theta}_n)+I_2(\widehat{\theta}_n)$, $I(\widehat{\theta}_n-\theta^*_n)=I_1(\widehat{\theta}_n-\theta^*_n)+I_2(\widehat{\theta}_n)$, and $I(\theta^*_n)=I_1(\theta^*_n)$, by triangular inequality we obtain $\widehat{\mathscr{E}}\leq 2(1+b)\overline{\lambda}_{n,0}I_1(\widehat{\theta}_n-\theta^*_n)+\mathscr{E}^*$. From Lemma 4.5, $\widehat{\mathscr{E}}\leq\delta\widehat{\mathscr{E}}+\varepsilon^*_n-\mathscr{E}^*+\mathscr{E}^*=\delta\widehat{\mathscr{E}}+\varepsilon^*_n$. Hence, $\widehat{\mathscr{E}}\leq\varepsilon^*_n/(1-\delta)$.

(a2) If $I(\widehat{\theta}_n-\theta^*_n)\leq c\zeta^*_n/b$, from (4.10) with $d_0 = c$, with probability at least

$$1-\left\{\left(1+\frac{3}{10}W^2\right)\exp(-n\overline{a}^2_n r^2_1)+2\exp(-n\xi^2/2)\right\},$$

we have

$$\widehat{\mathscr{E}}+(1+b)\overline{\lambda}_{n,0}I(\widehat{\theta}_n)\leq\frac{\delta}{1-\delta^2}\overline{\lambda}_{n,0}\zeta^*_n+\mathscr{E}^*+(1+b)\overline{\lambda}_{n,0}I(\theta^*_n).$$

By the triangular inequality, Lemma 4.5 and (A4),

$$\widehat{\mathscr{E}}\leq\frac{\delta}{1-\delta^2}\overline{\lambda}_{n,0}\zeta^*_n+\mathscr{E}^*+(1+b)\overline{\lambda}_{n,0}I_1(\widehat{\theta}_n-\theta^*_n)\leq\frac{\delta}{1-\delta^2}\overline{\lambda}_{n,0}\frac{\varepsilon^*_n}{\overline{\lambda}_{n,0}}+\mathscr{E}^*+\frac{\delta}{2}\widehat{\mathscr{E}}+\frac{1}{2}\varepsilon^*_n-\frac{1}{2}\mathscr{E}^*=\left(\frac{\delta}{1-\delta^2}+\frac{1}{2}\right)\varepsilon^*_n+\frac{1}{2}\mathscr{E}^*+\frac{\delta}{2}\widehat{\mathscr{E}}\leq\left(\frac{\delta}{1-\delta^2}+\frac{1}{2}\right)\varepsilon^*_n+\frac{1}{2(1+\delta)}\varepsilon^*_n+\frac{\delta}{2}\widehat{\mathscr{E}}.$$

Hence,

$$\widehat{\mathscr{E}}\leq\frac{2}{2-\delta}\left[\frac{\delta}{1-\delta^2}+\frac{1}{2}+\frac{1}{2(1+\delta)}\right]\varepsilon^*_n=\frac{1}{1-\delta}\varepsilon^*_n.$$

Furthermore, by Lemma 4.8, we have with probability at least

$$1-(N_1+N_2)\left\{\left(1+\frac{3}{10}W^2\right)\exp(-n\bar{a}_n^2r_1^2)+2\exp(-n\xi^2/2)\right\}$$

that $I(\widehat{\theta}_n-\theta_n^*)\le d(\delta_1,\delta_2)\frac{\zeta_n^*}{b}$, where

$$N_1=\log_{1+b}\left(\frac{1}{\delta_1}\right), N_2=\log_{1+b}\left(\frac{1}{\delta_2}\right).$$

(b) $c \ge d(\delta_1, \delta_2)$. On the set where $I(\widehat{\theta}_n-\theta_n^*)\le d(\delta_1,\delta_2)\zeta_n^*/b$, from equation (4.10) we have with probability at least

$$1-\left\{\left(1+\frac{3}{10}W^2\right)\exp(-n\bar{a}_n^2r_1^2)+2\exp(-n\xi^2/2)\right\}$$

that

$$\widehat{\mathscr{E}}+(1+b)\overline{\lambda}_{n,0}I(\widehat{\theta}_n)\le\overline{\lambda}_{n,0}d(\delta_1,\delta_2)\frac{\zeta_n^*}{b}+\mathscr{E}^*+(1+b)\overline{\lambda}_{n,0}I(\theta_n^*)\le\frac{\delta}{1-\delta^2}\overline{\lambda}_{n,0}\zeta_n^*+\mathscr{E}^*+(1+b)\overline{\lambda}_{n,0}I(\theta_n^*),$$

which is the same as (a2) and leads to the same result.

To summarize, let

$$A=\left\{\widehat{\mathscr{E}}\le\frac{1}{1-\delta}\varepsilon_n^*\right\}, B=\left\{I(\widehat{\theta}_n-\theta_n^*)\le d(\delta_1,\delta_2)\frac{\zeta_n^*}{b}\right\}.$$

Note that

$$J+1\le\log_{1+b}\left(\frac{(1+b)^2d(\delta_1,\delta_2)}{c}\right).$$

Under case (a), we have

$$
\begin{aligned}
P(A\cap B)=&P(\mathrm{a1})\\
&-P(A^c\cap \mathrm{a1})\\
&+P(\mathrm{a2})\\
&-P(A^c\cap \mathrm{a2})\geq P(\mathrm{a1})\\
&-J\left\{(1\right.\\
&+\frac{3}{10}W^2\exp(\\
&\quad -n\bar{a}_n^2r_1^2)\\
&+2\exp(\\
&\quad \left.-n\xi^2/2)\right\}\\
&+P(\mathrm{a2})\\
&-\left\{(1\right.\\
&+\frac{3}{10}W^2\exp(\\
&\quad -n\bar{a}_n^2r_1^2)\\
&\quad \left.+2\exp(-n\xi^2/2)\right\}\\
=&P(B)\\
&-(J+1)\left\{(1\right.\\
&+\frac{3}{10}W^2\exp(\\
&\quad -n\bar{a}_n^2r_1^2)\\
&+2\exp(\\
&-n\xi^2/2)\geq 1-(N_1+N_2+J+1)\left\{(1\right.\\
&+\frac{3}{10}W^2\exp(\\
&\quad -n\bar{a}_n^2r_1^2)\\
&+2\exp(\\
&\quad \left.-n\xi^2/2)\right\}\geq 1\\
&-\log_{1+b}\left\{\frac{(1+b)^2}{\delta_1\delta_2}\right.\\
&\left.\cdot\frac{d(\delta_1,\delta_2)(1-\delta^2)}{\delta b}\right\}\left\{(1\right.\\
&+\frac{3}{10}W^2\exp(\\
&\quad -n\bar{a}_n^2r_1^2)\\
&\quad \left.+2\exp(-n\xi^2/2)\right\}.
\end{aligned}
$$

Under case (b),

$$
\begin{aligned}
P(A\cap B) &= P(B) \\
&\quad - P(A^c\cap B) \geq P(B) \\
&\quad - \{(1 \\
&\quad + \frac{3}{10}W^2\exp(\\
&\quad - n\bar{a}_n^2 r_1^2) \\
&\quad + 2\exp(\\
&\quad - n\xi^2/2) \geq 1 - (N_1+N_2+2)\{(1 \\
&\quad + \frac{3}{10}W^2\exp(\\
&\quad - n\bar{a}_n^2 r_1^2) \\
&\quad + 2\exp(\\
&\quad - n\xi^2/2) = 1 - \log_{1+b}\left\{\frac{(1+b)^2}{\delta_1\delta_2}\right\}\{(1 \\
&\quad + \frac{3}{10}W^2\exp(\\
&\quad - n\bar{a}_n^2 r_1^2) \\
&\quad + 2\exp(-n\xi^2/2)\Big\}.
\end{aligned}
$$

We thus obtain the desired result.

References

Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann. Statist. 1982; 10:1100–1120.

Bach F. Self-concordant analysis for logistic regression. Electronic Journal of Statistics. 2010; 4:384–414.

Bickel P, Ritov Y, Tsybakov A. Simultaneous analysis of lasso and dantzig selector. Ann. Statist. 2009; 37:1705–1732.

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

Bühlmann P. Boosting for high-dimensional linear models. Ann. Statist. 2006; 34:559–583.

Bunea F. Honest variable selection in linear and logistic regression models via ℓ_1 and ℓ_1+ℓ_2 penalization. Electronic Journal of Statistics. 2008; 2:1153–1194.

Bunea F, Tsybakov AB, Wegkamp MH. Sparsity oracle inequalities for the Lasso. Electronic Journal of Statistics. 2007; 1:169–194.

Cox DR. Regression models and life tables (with discussion). J. Roy. Statist. Soc. B. 1972; 34:187–220.

Gaiffas S, Guilloux A. High-dimensional additive hazards models and the lasso. Electronic Journal of Statistics. 2012; 6:522–546.

Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics. 2005; 21:3001–3008. [PubMed: 15814556]

Ledoux, M.; Talagrand, M. Probability in Banach Spaces: Isoperimetry and Processes. Berlin: Springer; 1991.

Martinussen T, Scheike TH. Covariate selection for the semiparametric additive risk model. Scandinavian Journal of Statistics. 2009; 36:602–619.

Massart P. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. Ann. Probab. 1990; 18:1269–1283.

Tarigan B, van de Geer S. Classifiers of support vector machine type with ℓ_1 complexity regularization. Bernoulli. 2006; 12:1045–1076.

Tibshirani R. Regression shrinkage and selection via the Lasso. JRStat. Soc. B. 1996; 58:267–288.

Tibshirani R. The Lasso method for variable selection in the Cox model. Statistics in Medicine. 1997; 16:385–395. [PubMed: 9044528]

van de Geer S. High-dimensional generalized linear models and the lasso. Ann. Statist. 2008; 36:614–645.

van der Vaart and Wellner; Wellner, J. Weak Convergence and Empirical Processes: With Applications to Statistics. New York: Wiley; 1996.