Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2016, Vol. 52, No. 2, 981–1008
DOI: 10.1214/14-AIHP662

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

Oracle inequalities for the Lasso in the high-dimensional Aalen multiplicative intensity model

Sarah Lemler

LAMME, University of Évry Val d'Essonne, France. E-mail: sarah.lemler@genopole.cnrs.fr

Received 12 October 2013; revised 11 June 2014; accepted 30 October 2014

Abstract. In a general counting process setting, we consider the problem of obtaining a prognostic on the survival time adjusted on covariates in high-dimension. Towards this end, we construct an estimator of the whole conditional intensity. We estimate it by the best Cox proportional hazards model given two dictionaries of functions. The first dictionary is used to construct an approximation of the logarithm of the baseline hazard function and the second to approximate the relative risk. We introduce a new data-driven weighted Lasso procedure to estimate the unknown parameters of the best Cox model approximating the intensity. We provide non-asymptotic oracle inequalities for our procedure in terms of an appropriate empirical Kullback divergence. Our results rely on an empirical Bernstein's inequality for martingales with jumps and properties of modified self-concordant functions.

Résumé. Dans le cadre général d'un processus de comptage, nous intéressons à la façon d'obtenir un pronostic sur la durée de survie en fonction des covariables en grande dimension. Pour ce faire, nous construisons un estimateur de l'intensité conditionnelle. Nous l'estimons par le meilleur modèle de Cox étant donné deux dictionnaires de fonctions. Le premier dictionnaire est utilisé pour construire le logarithme du risque de base et le second, pour approximer le risque relatif. Nous introduisons une nouvelle procédure Lasso pondéré avec une pondération basée sur les données pour estimer les paramètres inconnus du meilleur modèle de Cox approximant l'intensité. Nous établissons une inégalité oracle non-asymptotique en divergence de Kullback empirique, qui est la fonction de perte la plus appropriée à notre procédure. Nos résultats reposent sur une inégalité de Bernstein pour les martingales à sauts et sur des propriétés des fonctions self-concordantes.

MSC: 62N02; 62G05; 62G08; 60E15

Keywords: Survival analysis; Right-censored data; Intensity; Cox proportional hazards model; Semiparametric model; Non-parametric model; High-dimensional covariates; Lasso; Non-asymptotic oracle inequalities; Empirical Bernstein's inequality

1. Introduction

We consider one of the statistical challenges brought by the recent advances in biomedical technology to clinical applications. For example, in Dave et al. [16], the considered data relate 191 patients with follicular lymphoma. The observed variables are the survival time, that can be right-censored, clinical variables, as the age or the disease stage, and 44,929 levels of gene expression. In this high-dimensional right-censored setting, there are two clinical questions. One is to determine prognostic biomarkers, the second is to predict the survival from follicular lymphoma adjusted on covariates. We focus our interest on the second (see Gourlay [20] and Steyerberg [33]). As a consequence, we consider the statistical question of estimating the whole conditional intensity. To adjust on covariates, the most popular semi-parametric regression model is the Cox proportional hazards model (see Cox [15]): the conditional hazard rate function of the survival time T given the vector of covariates $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ is defined by

$$\lambda_0(t, \mathbf{Z}) = \alpha_0(t) \exp\bigl(\boldsymbol{\beta}_0^T \mathbf{Z}\bigr), \tag{1}$$

where $\boldsymbol{\beta}_0 = (\beta_{0_1}, \ldots, \beta_{0_p})^T$ is the vector of regression coefficients and α_0 is the baseline hazard function. The unknown parameters of the model are $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and the function α_0. To construct an estimator of λ_0, one usually considers the partial likelihood introduced by Cox [15] to derive an estimator of $\boldsymbol{\beta}_0$ and then plug this estimator to obtain the well-known Breslow estimator of α_0. We propose in this paper an alternative one-step strategy.

1.1. *Framework*

Before describing our strategy, let us clarify our framework. We consider the general setting of counting processes. For $i = 1, \ldots, n$, let N_i be a marked counting process and Y_i a predictable random process with values in $[0, 1]$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ be the filtration defined by

$$\mathcal{F}_t = \sigma\big\{N_i(s), Y_i(s), 0 \leq s \leq t, \mathbf{Z}_i, i = 1, \ldots, n\big\},$$

where $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,p})^T \in \mathbb{R}^p$ is the $\mathcal{F}_0$-measurable random vector of covariates of individual i. Let $\Lambda_i(t)$ be the compensator of the process $N_i(t)$ with respect to $(\mathcal{F}_t)_{t\geq 0}$, so that $M_i(t) = N_i(t) - \Lambda_i(t)$ is a $(\mathcal{F}_t)_{t\geq 0}$-martingale.

The process N_i satisfies the Aalen multiplicative intensity model: for all $t \geq 0$,

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \mathbf{Z}_i) Y_i(s)\, \mathrm{d}s, \tag{A_1}$$

where λ_0 is an unknown non-negative function called intensity.

This general setting, introduced by Aalen [1], embeds several particular examples as censored data, marked Poisson processes and Markov processes (see Andersen et al. [2] for further details).

Remark 1.1 (Censoring case). *In the specific case of right censoring, let $(T_i)_{i=1,\ldots,n}$ be independent and identically distributed (i.i.d.) survival times of n individuals and $(C_i)_{i=1,\ldots,n}$ their i.i.d. censoring times. We observe $\{(X_i, \mathbf{Z}_i, \delta_i)\}_{i=1,\ldots,n}$ where $X_i = \min(T_i, C_i)$ is the event time, $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,p})^T$ is the vector of covariates and $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$ is the censoring indicator. The survival times T_i are supposed to be conditionally independent of the censoring times C_i given some vector of covariates $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,p})^T \in \mathbb{R}^p$ for $i = 1, \ldots, n$. With these notations, the $(\mathcal{F}_t)$-adapted processes Y_i and N_i are respectively defined as the at-risk process $Y_i(t) = \mathbb{1}_{\{X_i \geq t\}}$ and the counting process $N_i(t) = \mathbb{1}_{\{X_i \leq t, \delta_i = 1\}}$ which jumps when the ith individual dies.*

We observe the i.i.d. data $(\mathbf{Z}_i, N_i(t), Y_i(t), i = 1, \ldots, n, 0 \leq t \leq \tau)$, where $[0, \tau]$ is the time interval between the beginning and the end of the study.

$$\text{We assume that } A_0 = \sup_{1\leq i\leq n} \left\{ \int_0^\tau \lambda_0(s, \mathbf{Z}_i)\, \mathrm{d}s \right\} < \infty. \tag{A_2}$$

This is the standard assumption in statistical estimation of intensities of counting processes, see Andersen et al. [2] for instance. We also precise that, in the following, we work conditionally to the covariates and from now on, all probabilities $\mathbb{P}$ and expectations $\mathbb{E}$ are conditional to the covariates. Our goal is to estimate λ_0 non-parametrically in a high-dimensional setting, i.e. when the number of covariates p is larger than the sample size n ($p \gg n$).

1.2. *Previous results*

In high-dimensional regression, the benchmarks for results are the ones obtained in the additive regression model. In this setting, Tibshirani [35] has introduced the Lasso procedure, which consists in minimizing an ℓ_1-penalized criterion. The Lasso estimator has been widely studied for this model, with consistency results (see Meinshausen and Bühlmann [31]) and variable selection results (see Zhao and Yu [42], Zhang and Huang [39]). Recently, attention has been directed on establishing non-asymptotic oracle inequalities for the Lasso (see Bunea et al. [11,12], Bickel et al. [7], Massart and Meynet [30], Bartlett [5] and Koltchinskii [23] among others).

In the setting of survival analysis, the Lasso procedure has been first considered by Tibshirani [36] and applied to the partial log-likelihood. More generally, other procedures have been introduced for the parametric part of the Cox model: the adaptive Lasso, the smooth clipped absolute deviation penalizations and the Dantzig selector are respectively considered in Zou [44], Zhang and Lu [40], Fan and Li [17] and Antoniadis et al. [3]. Non-parametric approaches are considered in Letué [27], Hansen et al. [21] and Comte et al. [14]. Lasso procedures for the alternative Aalen additive model have been introduced in Martinussen and Scheike [28] and Gaïffas and Guilloux [18].

All of the existing results in the Cox model are based on the partial log-likelihood, which does not answer the clinical question associated with a prognosis. Antoniadis et al. [3] have established asymptotic estimation inequalities in the Cox proportional hazard model for the Dantzig estimator (see Bickel et al. [7] for a comparison between these two estimators in an additive regression model). In Bradic et al. [8], asymptotic estimation inequalities for the Lasso estimator have also been obtained in the Cox model. More recently, Kong and Nan [24] and Bradic and Song [9] have established non-asymptotic oracle inequalities for the Lasso in the generalized Cox model

$$\lambda_0(t, \mathbf{Z}) = \alpha_0(t) \exp\big(f_0(\mathbf{Z})\big), \tag{2}$$

where α_0 is the baseline hazard function and f_0 a function of the covariates. However, the focus in both papers is on the Cox partial log-likelihood, the obtained results are either on $f_{\hat{\beta}_L} - f_0$ or on $\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0$ for $f_0(\mathbf{Z}) = \boldsymbol{\beta}_0^T \mathbf{Z}$ and the problem of estimating the whole intensity λ_0 is not considered, as needed for the prediction of the survival time.

1.3. *Our contribution*

The first motivation of the present paper is to address the problem of estimating λ_0 defined in (A_1) regardless of an underlying model. We use an agnostic learning approach, see Kearns et al. [22], to construct an estimator that mimics the performance of the best Cox model, whether this model is true or not. More precisely, we will consider candidates for the estimation of λ_0 of the form

$$\lambda_{\beta,\gamma}(t, \mathbf{Z}) = \alpha_\gamma(t) \mathrm{e}^{f_\beta(\mathbf{Z})} \quad \text{for } (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N,$$

where f_β and α_γ are respectively linear combinations of functions of two dictionaries $\mathbb{F}_M$ and $\mathbb{G}_N$. The estimator of λ_0 is defined as the candidate which minimizes a weighted ℓ_1-penalized total log-likelihood as opposed to the Cox partial log-likelihood. The second motivation of the paper is to obtain non-asymptotic oracle inequalities for Lasso estimators of the complete intensity λ_0. Indeed, in practice, one cannot consider that the asymptotic regime has been reached, cf. in Dave et al. [16] for example. In addition, Comte et al. [14] established non-asymptotic oracle inequalities for the whole intensity but not in a high-dimensional setting and to the best of our knowledge, no non-asymptotic results for the estimation of the whole intensity in high dimension exist in the literature.

Towards this end, we will proceed in two steps. In a first step, we assume that λ_0 verifies Model (2), where α_0 is assumed to be known. In this particular case, the only non-parametric function to estimate is f_0 and we estimate it by a linear combination of functions of the dictionary $\mathbb{F}_M$. In this setting, we obtain non-asymptotic oracle inequalities for the Cox model when α_0 is supposed to be known. In a second step, we consider the general problem of estimating the whole intensity λ_0. We state non-asymptotic oracle inequalities both in terms of empirical Kullback divergence and weighted empirical quadratic norm for our Lasso estimators, thanks to properties of modified self-concordant functions (see Bach [4]).

These results are obtained via three ingredients: a new Bernstein's inequality, a modified Restricted Eigenvalue condition on the expectation of the weighted Gram matrix and modified self-concordant functions. Let us be more precise. We establish empirical versions of Bernstein's inequality involving the optional variation for martingales with jumps (see Gaïffas and Guilloux [18] and Hansen et al. [21] for related results). This allows us to define a fully data-driven weighted ℓ_1-penalization. For the resulting estimator, we work under a modified Restricted Eigenvalue condition according to which the expectation of a weighted Gram matrix fullfilled the Restricted Eigenvalue condition (see Bickel et al. [7]). This new version of the Restricted Eigenvalue condition is both new and weaker than the comparable condition in the Cox model. Finally, we extend the notion of self-concordance (see Bach [4]) to the problem at hands in order to connect our weighted empirical quadratic norm and our empirical Kullback divergence. In this context, we state the first fast non-asymptotic oracle inequality for the whole intensity.

The paper is organized as follows. In Section 2, we describe the framework and the Lasso procedure for estimating the intensity. The estimation risk that we consider and its associated loss function are presented. In Section 3, prediction and estimation oracle inequalities in the particular Cox model with known baseline hazard function are stated. In Section 4, non-asymptotic oracle inequalities with different convergence rates are given for a general intensity. Section 5 is devoted to statement of empirical Bernstein's inequalities associated with our processes. Proofs are gathered in Section 6.

2. Estimation procedure

2.1. *The estimation criterion and the loss function*

To estimate the intensity λ_0, we consider the total empirical log-likelihood. By Jacod's formula (see Andersen et al. [2]), the log-likelihood based on the data $(\mathbf{Z}_i, N_i(t), Y_i(t), i = 1, \dots, n, 0 \le t \le \tau)$ is given by

$$C_n(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^\tau \log\lambda(t,\mathbf{Z}_i)\,\mathrm{d}N_i(t) - \int_0^\tau \lambda(t,\mathbf{Z}_i)Y_i(t)\,\mathrm{d}t\right\}.$$

Our estimation procedure is based on the minimization of this empirical risk. To this empirical risk, we associate the empirical Kullback divergence defined by

$$\begin{aligned}\widetilde{K}_n(\lambda_0,\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(\log\lambda_0(t,\mathbf{Z}_i) - \log\lambda(t,\mathbf{Z}_i)\big)\lambda_0(t,\mathbf{Z}_i)Y_i(t)\,\mathrm{d}t \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(\lambda_0(t,\mathbf{Z}_i) - \lambda(t,\mathbf{Z}_i)\big)Y_i(t)\,\mathrm{d}t. \end{aligned} \tag{3}$$

We refer to van de Geer [37] and Senoussi [32] for close definitions. We notice in addition, that this loss function is closed to the Kullback–Leibler information considered in the density framework (see Stone [34] and Le Pennec and Cohen [25]). The following proposition justifies the choice of this criterion.

Proposition 2.1. *The empirical Kullback divergence $\widetilde{K}_n(\lambda_0,\lambda)$ is non-negative and equals zero if and only if $\lambda = \lambda_0$ almost surely on the interval $[0, \tau \wedge \sup\{t\colon \exists i \in \{1,\dots,n\}, Y_i(t) \neq 0\}]$.*

Remark 2.2 (Censoring case). *In the specific case of right censoring, the proposition holds true on $[0, \tau \wedge \max_{1\le i\le n} X_i]$. In this case, we can specify that $\mathbb{P}([0,\tau] \subset [0, \max_{1\le i\le n} X_i]) = 1 - (1 - S_T(\tau))^n(1 - S_C(\tau))^n$, where S_T and S_C are the survival functions of the survival time T and the censoring time C respectively. From (A_2), $S_T(\tau) > 0$ and if τ is such that $S_C(\tau) > 0$, then $\mathbb{P}([0,\tau] \subset [0, \max_{1\le i\le n} X_i])$ is large. See Gill [19] for a discussion on the role of τ.*

In the following, we consider that we estimate $\lambda_0(t)$ for t in $[0, \tau \wedge \sup\{t\colon \exists i \in \{1,\dots,n\}, Y_i(t) \neq 0\}]$. Let introduce the weighted empirical quadratic norm defined for all function h on $[0,\tau] \times \mathbb{R}^p$ by

$$\|h\|_{n,\Lambda} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(h(t,\mathbf{Z}_i)\big)^2\,\mathrm{d}\Lambda_i(t)}, \tag{4}$$

where Λ_i is defined in (A_1). Notice that, in this definition, the higher the intensity of the process N_i is, the higher the contribution of individual i to the empirical norm is. This norm is connected to the empirical Kullback divergence, as it will be shown in Proposition 6.4. Finally, for a vector $\mathbf{b}$ in $\mathbb{R}^M$, we define, $\|\mathbf{b}\|_1 = \sum_{j=1}^M |b_j|$ and $\|\mathbf{b}\|_2^2 = \sum_{j=1}^M b_j^2$.

2.2. *Weighted Lasso estimation procedure*

The estimation procedure is based on the choice of two finite sets of functions, called dictionaries. Let $\mathbb{F}_M = \{f_1, \dots, f_M\}$ where $f_j : \mathbb{R}^p \to \mathbb{R}$ for $j = 1, \dots, M$, and $\mathbb{G}_N = \{\theta_1, \dots, \theta_N\}$ where $\theta_k : \mathbb{R}_+ \to \mathbb{R}$ for $k = 1, \dots, N$, be two dictionaries. Typically the size of the dictionary $\mathbb{F}_M$ used to estimate the function of the covariates in a high-dimensional setting is large, i.e. $M \gg n$, whereas to estimate a function on $\mathbb{R}_+$, we consider a dictionary $\mathbb{G}_N$ with size N of the order of n. The sets $\mathbb{F}_M$ and $\mathbb{G}_N$ can be collections of functions such as wavelets, splines, step functions, coordinate functions etc. They can also be collections of several estimators computed using different tuning parameters. To make sure that no identification problems appear by using two dictionaries, it is assumed that only the dictionary $\mathbb{G}_N = \{\theta_1, \dots, \theta_N\}$ can contain the constant function, not $\mathbb{F}_M = \{f_1, \dots, f_M\}$. The candidates for the estimator of λ_0 are of the form

$$\lambda_{\beta,\gamma}(t, \mathbf{Z}_i) = \alpha_\gamma(t) \mathrm{e}^{f_\beta(\mathbf{Z}_i)} \quad \text{with } \log \alpha_\gamma = \sum_{k=1}^{N} \gamma_k \theta_k \quad \text{and} \quad f_\beta = \sum_{j=1}^{M} \beta_j f_j,$$

where $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N$.

The dictionaries $\mathbb{F}_M$ and $\mathbb{G}_N$ are chosen such that the two following assumptions are fulfilled:

$$\text{For all } j \text{ in } \{1, \dots, M\}, \quad \|f_j\|_{n,\infty} = \max_{1 \le i \le n} \big|f_j(Z_i)\big| < \infty. \tag{A$_3$}$$

$$\text{For all } k \text{ in } \{1, \dots, N\}, \quad \|\theta_k\|_\infty = \max_{t \in [0,\tau]} \big|\theta_k(t)\big| < \infty. \tag{A$_4$}$$

We consider a weighted Lasso procedure for estimating λ_0.

Estimation procedure 2.3. *The Lasso estimator of λ_0 is defined by $\lambda_{\hat{\beta}_L, \hat{\gamma}_L}$, where*

$$(\hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\gamma}}_L) = \operatorname*{arg\,min}_{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N} \big\{ C_n(\lambda_{\beta,\gamma}) + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\gamma}) \big\},$$

with

$$\mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j |\beta_j| \quad \textit{and} \quad \mathrm{pen}(\boldsymbol{\gamma}) = \sum_{k=1}^{N} \delta_k |\gamma_k|.$$

The positive data-driven weights $\omega_j = \omega(f_j, n, M, \nu, x)$, $j = 1, \dots, M$ and $\delta_k = \delta(\theta_k, n, N, \tilde{\nu}, y)$, $k = 1, \dots, N$ are defined as follows. Let $x > 0$, $y > 0$, $\varepsilon > 0$, $\tilde{\varepsilon} > 0$, $c = 2\sqrt{2(1+\varepsilon)}$, $\tilde{c} = 2\sqrt{2(1+\tilde{\varepsilon})}$ and $(\nu, \tilde{\nu}) \in (0,3)^2$ such that $\nu > \Phi(\nu)$ and $\tilde{\nu} > \Phi(\tilde{\nu})$, where $\Phi(u) = \exp(u) - u - 1$. With these notations, the weigths are defined by

$$\omega_j = c\sqrt{\frac{\hat{W}_n^\nu(f_j)(x + \log M)}{n}} + 2\frac{x + \log M}{3n}\|f_j\|_{n,\infty}, \tag{5}$$

$$\delta_k = \tilde{c}\sqrt{\frac{\hat{T}_n^{\tilde{\nu}}(\theta_k)(y + \log N)}{n}} + 2\frac{y + \log N}{3n}\|\theta_k\|_\infty, \tag{6}$$

for

$$\hat{W}_n^\nu(f_j) = \frac{\nu/n}{\nu/n - \Phi(\nu/n)}\hat{V}_n(f_j) + \frac{x/n}{\nu/n - \Phi(\nu/n)}\|f_j\|_{n,\infty}^2, \tag{7}$$

$$\hat{T}_n^{\tilde{\nu}}(\theta_k) = \frac{\tilde{\nu}/n}{\tilde{\nu}/n - \Phi(\tilde{\nu}/n)}\hat{R}_n(\theta_k) + \frac{y/n}{\tilde{\nu}/n - \Phi(\tilde{\nu}/n)}\|\theta_k\|_\infty^2, \tag{8}$$

where $\hat{V}_n(f_j)$ and $\hat{R}_n(\theta_k)$ are the "observable" empirical variance of f_j and θ_k respectively, given by

$$\hat{V}_n(f_j) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(f_j(\mathbf{Z}_i)\big)^2 \,\mathrm{d}N_i(s) \quad \text{and} \quad \hat{R}_n(\theta_k) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(\theta_k(s)\big)^2 \,\mathrm{d}N_i(s).$$

Remark 2.4. *The general Lasso estimator for $\boldsymbol{\beta}$ is classically defined by*

$$\hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}\in\mathbb{R}^M}{\arg\min}\left\{ C_n(\lambda_{\boldsymbol{\beta}}) + \Gamma\sum_{j=1}^{M}|\beta_j| \right\},$$

with $\Gamma > 0$ a smoothing parameter. Usually, Γ is of order $\sqrt{\log M/n}$ (see Massart and Meynet [30] for the usual additive regression model and Antoniadis et al. [3] for the Cox model among other). The Lasso penalization for $\boldsymbol{\beta}$ corresponds to the simple choice $\omega_j = \Gamma$ where $\Gamma > 0$ is a smoothing parameter. Our weights could be compared with those of Bickel et al. [7] in the case of an additive regression model with a gaussian noise. They have considered a weighted Lasso with a penalty term of the form $\Gamma\sum_{j=1}^{M}\|f_j\|_n|\beta_j|$, with Γ of order $\sqrt{\log M/n}$ and $\|\cdot\|_n$ the usual empirical norm. We can deduce from the weights ω_j defined by (5) higher suitable weights that can be written $\Gamma^1_{n,M}\tilde{\omega}_j$ with $\tilde{\omega}_j = \sqrt{\hat{W}^v_n(f_j)}$ defined by (7), which is of order $\sqrt{\hat{V}_n(f_j)}$ and

$$\Gamma^1_{n,M} = c\sqrt{\frac{x+\log M}{n}} + 2\frac{x+\log M}{3n}\max_{1\le j\le M}\frac{\|f_j\|_{n,\infty}}{\sqrt{\hat{W}^v_n(f_j)}}.$$

The regularization parameter $\Gamma^1_{n,M}$ is still of order $\sqrt{\log M/n}$. The weights $\tilde{\omega}_j$ correspond to the estimation of the weighted empirical norm $\|\cdot\|_{n,\Lambda}$ that is not observable and play the same role than the empirical norm $\|f_j\|_n$ in Bickel et al. [7]. These weights are also of the same form as those of van de Geer [38] for the logistic model.

The idea of adding some weights in the penalization comes from the adaptive Lasso, although it is not the same procedure. Indeed, in the adaptive Lasso (see Zou [43]) one chooses $\omega_j = |\tilde{\beta}_j|^{-a}$ where $\tilde{\beta}_j$ is a preliminary estimator and $a > 0$ a constant. The idea behind this is to correct the bias of the Lasso in terms of variables selection accuracy (see Zou [43] and Zhang [41] for regression analysis and Zhang and Lu [40] for the Cox model). The weights ω_j can also be used to scale each variable at the same level, which is suitable when some variables have a large variance compared to the others.

Remark 2.5 (Towards practical issues). *The actual computation of the estimator $\lambda_{\hat{\boldsymbol{\beta}}_L,\hat{\boldsymbol{\gamma}}_L}$, although of the greatest interest, is beyond the scope of the present paper. However, we give here the principal steps to get it. Two types of algorithms could be considered: the cyclical coordinate descent or the proximal gradient descent. As far as we know, maximal algorithms have not yet been implemented for the Cox model (neither for the partial likelihood nor for the total likelihood). On the other hand, cyclical coordinate descent is implemented for the Cox model, e.g., in the R function glmnet, but only for the partial likelihood. In addition, our sum of weighted ℓ_1-penalizations is not usual and require attention when applying the proximal operator in both cyclical coordinate descend and proximal algorithm. Finally, the cross validation procedure will have to consider regularization parameters on a squared grid. This done, we would be able to compare unweighted to weighted procedures in terms of selection, estimation or prediction accuracies. Following the example in an other context of Hansen et al. [21], we shall expect our weighted procedure to outscore the unweighted one.*

3. Oracle inequalities for the Cox model when the baseline hazard function is known

As a first step, we suppose that the intensity satisfies the generalization of the Cox model (2) with a known baseline function α_0. In this context, only f_0 has to be estimated and λ_0 is estimated by

$$\lambda_{\hat{\beta}_L}(t,\mathbf{Z}_i) = \alpha_0(t)\mathrm{e}^{f_{\hat{\beta}_L}(\mathbf{Z}_i)} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}\in\mathbb{R}^M}{\arg\min}\big\{C_n(\lambda_{\boldsymbol{\beta}}) + \mathrm{pen}(\boldsymbol{\beta})\big\}. \tag{9}$$

In this section, we state non-asymptotic oracle inequalities for the prediction loss of the Lasso in terms of the Kullback divergence. These inequalities allow us to compare the prediction error of the estimator and the best approximation of the regression function by a linear combination of the functions of the dictionary in a non-asymptotic way.

3.1. *A slow oracle inequality*

In the following theorem, we state an oracle inequality in the Cox model with slow rate of convergence, i.e. with a rate of convergence of order $\sqrt{\log M/n}$. This inequality is obtained under a very light assumption on the dictionary $\mathbb{F}_M$.

Proposition 3.1. *Consider model* (2) *with known* α_0. *Let* $x > 0$ *be fixed,* ω_j *be defined by* (5) *and for* $\boldsymbol{\beta} \in \mathbb{R}^M$,

$$\operatorname{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j |\beta_j|.$$

Let $A_{\varepsilon,\nu}(x)$ *be some numerical positive constant depending only on* ε, ν *and* x. *Under assumption* (A_3), *with a probability larger than* $1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-x}$, *then*

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) \leq \inf_{\beta \in \mathbb{R}^M} \big(\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2 \operatorname{pen}(\boldsymbol{\beta}) \big). \tag{10}$$

This theorem states a non-asymptotic oracle inequality in prediction on the conditional hazard rate function in the Cox model. The ω_j are the order of $\sqrt{\log M/n}$ and the penalty term is of order $\|\boldsymbol{\beta}\|_1 \sqrt{\log M/n}$. This variance order is usually referred as a slow rate of convergence in high dimension (see Bickel et al. [7] for the additive regression model, Bertin et al. [6] and Bunea et al. [13] for density estimation).

3.2. *A fast oracle inequality*

Now, we are interested in obtaining a non-asymptotic oracle inequality with a fast rate of convergence of order $\log M/n$ and we need further assumptions in order to prove such result. In this subsection, we shall work locally, for $\mu > 0$, on the set $\Gamma_M(\mu) = \{\boldsymbol{\beta} \in \mathbb{R}^M\colon \ \|\log \lambda_{\boldsymbol{\beta}} - \log \lambda_0\|_{n,\infty} \leq \mu\}$, simply denoted $\Gamma(\mu)$ to simplify the notations and we consider the following assumption:

$$\text{There exists } \mu > 0, \text{ such that } \Gamma(\mu) \text{ contains a non-empty open set of } \mathbb{R}^M. \tag{A_5}$$

This assumption has already been considered by van de Geer [38] or Kong and Nan [24]. Roughly speaking, it means that one can find a set where we can restrict our attention for finding good estimator of f_0. This assumption is needed in order to connect, via the notion of self-concordance (see Bach [4]), the weighted empirical quadratic norm and the empirical Kullback divergence (see Proposition 6.2).

The weighted Lasso estimator becomes

$$\hat{\boldsymbol{\beta}}_L^{\mu} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \Gamma(\mu)} \big\{ C_n(\lambda_{\boldsymbol{\beta}}) + \operatorname{pen}(\boldsymbol{\beta}) \big\}. \tag{11}$$

By definition, this weighted Lasso estimator is obtained on a ball centered around the true function λ_0. However in assumption (A_5), we can always consider a large radius μ, which weakens it. This could not change the rate of convergence in the oracle inequalities ($\sim \log M/n$) but only the range of a constant. In the particular case in which $\log \lambda_{\boldsymbol{\beta}}$ for all $\boldsymbol{\beta} \in \mathbb{R}^M$ and $\log \lambda_0$ are bounded, there exists $\mu > 0$ such that $\|\log \lambda_{\boldsymbol{\beta}} - \log \lambda_0\|_{n,\infty} \leq \|\log \lambda_\beta\|_{n,\infty} + \|\log \lambda_0\|_{n,\infty} \leq \mu$.

To achieve a fast rate of convergence, one needs an additional assumption on the Gram matrix. See Bühlmann and van de Geer [10] and Bickel et al. [7] for detailed discussions on the different assumptions required for fast oracle inequalities. One of the weakest assumption is the Restricted Eigenvalue condition introduced by Bickel et al. [7]. We

choose to work under this Restricted Eigenvalue condition. Let us first introduce further notations:

$$\boldsymbol{\Delta} = \mathbf{D}\big(\hat{\boldsymbol{\beta}}_L^{\mu} - \boldsymbol{\beta}\big) \quad \text{with } \boldsymbol{\beta} \in \Gamma(\mu) \quad \text{and} \quad \mathbf{D} = \big(\operatorname{diag}(\omega_j)\big)_{1\le j\le M},$$
$$\mathbf{X} = \big(f_j(\mathbf{Z}_i)\big)_{i,j}, \quad \text{with } i \in \{1,\dots,n\} \quad \text{and} \quad j \in \{1,\dots,M\},$$
$$\mathbf{G}_n = \frac{1}{n}\mathbf{X}^T\mathbf{C}\mathbf{X} \quad \text{with } \mathbf{C} = \big(\operatorname{diag}\big(\Lambda_i(\tau)\big)\big)_{1\le i\le n}. \tag{12}$$

In the matrix $\mathbf{G}_n$, the covariates of individual i are re-weighted by its cumulative risk $\Lambda_i(\tau)$, which is consistent with the definition of the empirical norm in (4). Let also $J(\boldsymbol{\beta})$ be the sparsity set of vector $\boldsymbol{\beta} \in \Gamma(\mu)$ defined by $J(\boldsymbol{\beta}) = \{j \in \{1,\dots,M\}\colon \beta_j \neq 0\}$, and the sparsity index is then given by $|J(\boldsymbol{\beta})| = \operatorname{Card}\{J(\boldsymbol{\beta})\}$. For $J \subset \{1,\dots,M\}$, we denote by $\boldsymbol{\beta}_J$ the vector $\boldsymbol{\beta}$ restricted to the set J: $(\beta_J)_j = \beta_j$ if $j \in J$ and $(\beta_J)_j = 0$ if $j \in J^c$ where $J^c = \{1,\dots,M\} \setminus J$.

Usually, in order to obtain a fast oracle inequality, we need to assume a Restricted Eigenvalue condition on the Gram matrix $\mathbf{G}_n$. However, since $\mathbf{G}_n$ is random in our case, we impose the Restricted Eigenvalue condition to $\mathbb{E}(\mathbf{G}_n)$, where the expectation is taken conditionally to the covariates.

For some integer $s \in \{1,\dots,M\}$ and a constant $a_0 > 0$, the following condition holds:

$$0 < \boldsymbol{\kappa}_0(s,a_0) = \min_{\substack{J\subset\{1,\dots,M\},\\ |J|\le s}} \min_{\substack{\mathbf{b}\in\mathbb{R}^M\setminus\{0\},\\ \|\mathbf{b}_{J^c}\|_1 \le a_0\|\mathbf{b}_J\|_1}} \frac{(b^T\mathbb{E}(\mathbf{G}_n)b)^{1/2}}{\|\mathbf{b}_J\|_2}. \tag{RE(s, a$_0$)}$$

The integer s here plays the role of an upper bound on the sparsity $|J(\boldsymbol{\beta})|$ of a vector of coefficients $\boldsymbol{\beta}$, so that the square submatrices of size less than $2s$ of the expectation of the weighted Gram matrix are positive definite.

This assumption is weaker than the classical one and the following lemma implies that if the Restricted Eigenvalue condition is verified for $\mathbb{E}(\mathbf{G}_n)$, then the empirical version of the Restricted Eigenvalue condition applied to $\mathbf{G}_n$ holds true with large probability. This modified Restricted Eigenvalue condition is new and this is the first time to our best knowledge that a fast-non-asymptotic oracle inequality has been established under such a condition.

Lemma 3.2. *Let $L > 0$ such that $\max_{1\le j\le M}\max_{1\le i\le n}|f_j(\mathbf{Z}_i)| \le L$. Under assumptions (A$_2$) and (RE(s, a$_0$)), we have*

$$0 < \boldsymbol{\kappa} = \min_{\substack{J\subset\{1,\dots,M\},\\ |J|\le s}} \min_{\substack{\mathbf{b}\in\mathbb{R}^M\setminus\{0\},\\ \|\mathbf{b}_{J^c}\|_1 \le a_0\|\mathbf{b}_J\|_1}} \frac{(\mathbf{b}^T\mathbf{G}_n\mathbf{b})^{1/2}}{\|\mathbf{b}_J\|_2} \quad \text{and} \quad \boldsymbol{\kappa} = (1/\sqrt{2A_0})\boldsymbol{\kappa}_0(s,a_0), \tag{13}$$

with probability larger than $1-\pi_n$, where

$$\pi_n = 2M^2\exp\left[-\frac{n\boldsymbol{\kappa}^4}{2L^2(1+a_0)^2 s(L^2(1+a_0)^2 s + \boldsymbol{\kappa}^2/3)}\right].$$

Lemma 3.2 assures that the empirical Restricted Eigenvalue condition holds true on an event of large probability, on which we establish a fast non-asymptotic oracle inequality.

Remark 3.3 (Censoring case). *In the particular case of the right censoring (see Remark 1.1), we obtain a better version of Lemma 3.2. Indeed, in this case, $\Lambda_i([0,\tau])$ is exponentially distributed with rate parameter 1 and since $\Lambda_i([0,\tau]) \le \Lambda_i([0,\infty])$ almost surely, its expectation is then just less than 1, so that we obtain (13) with probability larger than*

$$1 - 2M^2\exp\left(-\frac{n\boldsymbol{\kappa}^4}{2L^2(1+a_0)^2 s(L^2(1+a_0)^2 s + \boldsymbol{\kappa}^2)}\right) \quad \text{with } \boldsymbol{\kappa} = (1/\sqrt{2})\boldsymbol{\kappa}_0(s,a_0).$$

Theorem 3.4. *Consider model (2) with known α_0 and for $x > 0$, let ω_j be defined by (5) and $\hat{\boldsymbol{\beta}}_L^{\mu}$ be defined by (11). Let $A_{\varepsilon,v}(x) > 0$ be a numerical positive constant only depending on ε, v and x, $\zeta > 0$ and $s \in \{1,\dots,M\}$ be fixed.*

Let assumptions (A_2), (A_3), (A_5) *and* (RE(s, a_0)) *be satisfied with* $a_0 = (3 + 4/\zeta)$ *and let* $\boldsymbol{\kappa} = (1/\sqrt{2A_0})\boldsymbol{\kappa}_0(s, a_0)$. *Then, with a probability larger than* $1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-x} - \pi_n$, *the following inequality holds*

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L^\mu}) \leq (1+\zeta) \inf_{\substack{\boldsymbol{\beta}\in\Gamma(\mu)\\ |J(\boldsymbol{\beta})|\leq s}} \left\{ \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + C(\zeta,\mu)\frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2}\Big(\max_{1\leq j\leq M}\omega_j\Big)^2 \right\}, \tag{14}$$

where $C(\zeta, \mu) > 0$ *is a constant depending on* ζ *and* μ.

This result allows to compare the prediction error of the estimator and the best sparse approximation of the regression function by an oracle that knows the truth, but is constrained by sparsity. The Lasso estimator approaches the best approximation in the dictionary with a fast error term of order $\log M/n$.

Thanks to Proposition 6.2, which states a connection between the empirical Kullback divergence (3) and the weighted empirical quadratic norm (4), we deduce from Theorem 3.4 a non-asymptotic oracle inequality in weighted empirical quadratic norm.

Corollary 3.5. *Under the assumptions of Theorem* 3.4, *with a probability larger than* $1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-x} - \pi_n$,

$$\|\log\lambda_{\hat{\beta}_L^\mu} - \log\lambda_0\|_{n,\Lambda}^2 \leq (1+\zeta) \inf_{\substack{\beta\in\Gamma(\mu)\\ |J(\beta)|\leq s}} \left\{ \|\log\lambda_{\boldsymbol{\beta}} - \log\lambda_0\|_{n,\Lambda}^2 + \tilde{c}(\zeta,\mu)\frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2}\Big(\max_{1\leq j\leq M}\omega_j\Big)^2 \right\},$$

where $\tilde{c}(\zeta, \mu)$ *is a positive constant depending on* ζ *and* μ.

Note that for α_0 supposed to be known, this oracle inequality is also equivalent to

$$\|f_{\hat{\beta}_L^\mu} - f_0\|_{n,\Lambda}^2 \leq (1+\zeta) \inf_{\substack{\beta\in\Gamma(\mu)\\ |J(\beta)|\leq s}} \left\{ \|f_\beta - f_0\|_{n,\Lambda}^2 + \tilde{c}(\zeta,\mu)\frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2}\Big(\max_{1\leq j\leq M}\omega_j\Big)^2 \right\}.$$

3.3. *Particular case*: *Variable selection in the Cox model*

We now consider the case of variable selection in the Cox model (2) with $f_0(Z_i) = \boldsymbol{\beta}_0^T\mathbf{Z}_i$. In this case, $M = p$ and the functions of the dictionary are such that for $i = 1, \ldots, n$ and $j = 1, \ldots, p$

$$f_j(\mathbf{Z}_i) = Z_{i,j} \quad \text{and} \quad f_\beta(\mathbf{Z}_i) = \sum_{j=1}^p \beta_j Z_{i,j} = \boldsymbol{\beta}_0^T\mathbf{Z}_i.$$

Let $\mathbf{X} = (Z_{i,j})_{1\leq i\leq n, 1\leq j\leq p}$ be the design matrix and for $\hat{\boldsymbol{\beta}}_L$ defined by (9), let

$$\boldsymbol{\Delta}_0 = \mathbf{D}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0), \qquad \mathbf{D} = \big(\mathrm{diag}(\omega_j)\big)_{1\leq j\leq M}, \qquad J_0 = J(\boldsymbol{\beta}_0) \quad \text{and} \quad |J_0| = \mathrm{Card}\{J_0\}.$$

We now state non-asymptotic inequalities for prediction on $\mathbf{X}\boldsymbol{\beta}_0$ and for estimation on $\boldsymbol{\beta}_0$. In this subsection, we do not need to work locally on the set $\Gamma(\mu)$ to obtain Proposition 6.3 and instead of considering assumption (A_5), we only have to introduce the following assumption to connect the empirical Kullback divergence and the weighted empirical quadratic norm:

Let R be a positive constant, such that $\max_{i\in\{1,\ldots,n\}} \|Z_i\|_2 \leq R$. ($A_6$)

We consider the Lasso estimator defined with the regularization parameter $\Gamma_1 > 0$:

$$\hat{\boldsymbol{\beta}}_L = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ C_n(\lambda_{\boldsymbol{\beta}}) + \Gamma_1 \sum_{j=1}^p \omega_j|\beta_j| \right\},$$

Theorem 3.6. *Consider model (1) with known α_0. For $x > 0$, let ω_j be defined by (5) and denote $\boldsymbol{\kappa}' = (1/\sqrt{2A_0})\boldsymbol{\kappa}_0(s,3)$. Let $A_{\varepsilon,\nu}(x)$ be some numerical positive constant depending on ε, ν and x. Under assumptions (A_2), (A_3), (A_6) and (RE(s, a_0)) with $a_0 = 3$, for all Γ_1 such that*

$$\Gamma_1 \leq \frac{1}{48Rs}\frac{\min_{1\leq j\leq M}\omega_j^2}{\max_{1\leq j\leq M}\omega_j^2}\frac{\boldsymbol{\kappa}'^2}{\max_{1\leq j\leq M}\omega_j},$$

with a probability larger than $1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-\Gamma_1 x} - \pi_n$, then

$$\big\|\mathbf{X}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0)\big\|_{n,\Lambda}^2 \leq \frac{4}{\xi^2}\frac{|J_0|}{\boldsymbol{\kappa}'^2}\Gamma_1^2\Big(\max_{1\leq j\leq p}\omega_j\Big)^2 \tag{15}$$

and

$$\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_1 \leq 8\frac{\max_{1\leq j\leq p}\omega_j}{\min_{1\leq j\leq p}\omega_j}\frac{|J_0|}{\xi\boldsymbol{\kappa}'^2}\Gamma_1\max_{1\leq j\leq p}\omega_j. \tag{16}$$

This theorem gives non-asymptotic upper bounds for two types of loss functions. Inequality (15) gives a non-asymptotic bound on prediction loss with a rate of convergence in $\log M/n$, while Inequality (16) states a bound on $\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0$.

4. Oracle inequalities for general intensity

In the previous section, we have assumed α_0 known and have obtained results on the relative risk. Now, we consider a general intensity λ_0 that does not rely on an underlying model. Oracle inequalities are established under different assumptions with slow and fast rates of convergence.

4.1. *A slow oracle inequality*

The slow oracle inequality for a general intensity is obtained under light assumptions that concern only the construction of the two dictionaries $\mathbb{F}_M$ and $\mathbb{G}_N$.

Theorem 4.1. *For $x > 0$ and $y > 0$, let ω_j and δ_k be defined by (5) and (6) respectively and $(\hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\gamma}}_L)$ be defined in Estimation procedure 2.3. Let $A_{\varepsilon,\nu}(x)$ and $B_{\tilde{\varepsilon},\tilde{\nu}}(y) > 0$ be two positive numerical constants depending on ε, ν, x and $\tilde{\varepsilon}$, $\tilde{\nu}$, y respectively and assumptions (A_2), (A_3), (A_4) be satisfied. Then, with probability larger than $1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-x} - B_{\tilde{\varepsilon},\tilde{\nu}}(y)\mathrm{e}^{-y}$*

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) \leq \inf_{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathbb{R}^M\times\mathbb{R}^N}\big\{\widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma}) + 2\,\mathrm{pen}(\boldsymbol{\beta}) + 2\,\mathrm{pen}(\boldsymbol{\gamma})\big\}. \tag{17}$$

We have chosen to estimate the complete intensity, which involves two different parts: the first part is the baseline function $\alpha_{\boldsymbol{\gamma}}:\mathbb{R}\to\mathbb{R}$ and the second part is the function of the covariates $f_{\boldsymbol{\beta}}:\mathbb{R}^p\to\mathbb{R}$. The double ℓ_1-penalization considered here is tuned to concurrently estimate the function f_0 depending on high-dimensional covariates and the non-parametric function α_0. As f_0 and α_0 are estimated at once, the resulting rate of convergence is the sum of the two expected rates in both situations considered separately ($\sim\sqrt{\log M/n} + \sqrt{\log N/n}$). Nevertheless, from Bertin et al. [6], we expect that a choice of N of order n would suitably estimate α_0. As a consequence, in a very high-dimensional setting the leading error term in (17) would be of order $\sqrt{\log M/n}$, which again is the classical slow rate of convergence in a regression setting.

4.2. *A fast oracle inequality*

We are now interested in obtaining the fast non-asymptotic oracle inequality and as usual, we need to introduce further notations and assumptions. In this subsection, we shall again work locally for $\rho > 0$ on the set $\widetilde{\Gamma}_{M,N}(\rho) = \{(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N\colon \|\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0\|_{n,\infty} \leq \rho\}$, simply denoted $\widetilde{\Gamma}(\rho)$ and we consider the following assumption:

There exists $\rho > 0$, such that $\widetilde{\Gamma}(\rho)$ contains a non-empty open set of $\mathbb{R}^M \times \mathbb{R}^N$. ($A_7$)

On $\widetilde{\Gamma}(\rho)$, we define the weighted Lasso estimator as

$$\big(\hat{\boldsymbol{\beta}}_L^{\rho}, \hat{\boldsymbol{\gamma}}_L^{\rho}\big) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)}{\arg\min} \big\{C_n(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\gamma})\big\}.$$

Let us give the additional notations. Set $\tilde{\boldsymbol{\Delta}}$ be

$$\tilde{\boldsymbol{\Delta}} = \tilde{\mathbf{D}} \begin{pmatrix} \hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}}_L - \boldsymbol{\gamma} \end{pmatrix} \in \mathbb{R}^{M+N} \quad \text{with } (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho) \quad \text{and} \quad \tilde{\mathbf{D}} = \mathrm{diag}(\omega_1, \ldots, \omega_M, \delta_1, \ldots, \delta_N).$$

Let $\mathbf{1}_{n\times N}$ be the matrix $n \times N$ with all coefficients equal to one,

$$\tilde{\mathbf{X}}(t) = [\,(f_j(\mathbf{Z}_i))_{1\leq i\leq n, 1\leq j\leq M} \quad \mathbf{1}_{n\times N}(\mathrm{diag}(\theta_k(t)))_{1\leq k\leq N}\,] = \left[\begin{array}{c|c} \mathbf{X} & \begin{matrix} \theta_1(t) & \cdots & \theta_N(t) \\ \vdots & & \vdots \\ \theta_1(t) & \cdots & \theta_N(t) \end{matrix} \end{array}\right] \in \mathbb{R}^{n\times(M+N)}$$

and

$$\tilde{\mathbf{G}}_n = \frac{1}{n}\int_0^{\tau} \tilde{\mathbf{X}}(t)^T \tilde{\mathbf{C}}(t) \tilde{\mathbf{X}}(t)\,\mathrm{d}t \quad \text{with } \tilde{\mathbf{C}}(t) = \big(\mathrm{diag}\big(\lambda_0(t, \mathbf{Z}_i) Y_i(t)\big)\big)_{1\leq i\leq n}, \quad \forall t \geq 0.$$

Let also $J(\boldsymbol{\beta})$ and $J(\boldsymbol{\gamma})$ be the sparsity sets of vectors $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$ respectively defined by

$$J(\boldsymbol{\beta}) = \big\{j \in \{1, \ldots, M\}\colon \beta_j \neq 0\big\} \quad \text{and} \quad J(\boldsymbol{\gamma}) = \big\{k \in \{1, \ldots, N\}\colon \gamma_k \neq 0\big\},$$

and the sparsity indexes are then given by

$$\big|J(\boldsymbol{\beta})\big| = \sum_{j=1}^{M} \mathbb{1}_{\{\beta_j \neq 0\}} = \mathrm{Card}\big\{J(\boldsymbol{\beta})\big\} \quad \text{and} \quad \big|J(\boldsymbol{\gamma})\big| = \sum_{k=1}^{N} \mathbb{1}_{\{\gamma_k \neq 0\}} = \mathrm{Card}\big\{J(\boldsymbol{\gamma})\big\}.$$

To obtain the fast non-asymptotic oracle inequality, we consider the Restricted Eigenvalue condition applied to the matrix $\mathbb{E}(\tilde{\mathbf{G}}_n)$.

For some integer $s \in \{1, \ldots, M+N\}$ and a constant $r_0 > 0$, we assume that $\tilde{\mathbf{G}}_n$ satisfies:

$$0 < \tilde{\boldsymbol{\kappa}}_0(s, r_0) = \min_{\substack{J\subset\{1,\ldots,M+N\}, \\ |J|\leq s}} \min_{\substack{\mathbf{b}\in\mathbb{R}^{M+N}\setminus\{0\}, \\ \|\mathbf{b}_{J^c}\|_1 \leq r_0\|\mathbf{b}_J\|_1}} \frac{(\mathbf{b}^T \mathbb{E}(\tilde{\mathbf{G}}_n)\mathbf{b})^{1/2}}{\|\mathbf{b}_J\|_2}. \tag{$\widetilde{\mathrm{RE}}(s, r_0)$}$$

The condition on the matrix $\mathbb{E}(\tilde{\mathbf{G}}_n)$ is rather strong because the block matrix involves both functions of the covariates of $\mathbb{F}_M$ and functions of time which belong to $\mathbb{G}_N$. This is the price to pay for an oracle inequality on the full intensity. If we had instead considered two restricted eigenvalue assumptions on each block, we would have established an oracle inequality on the sum of the two unknown parameters α_0 and f_0 and not on λ_0. As in Lemma 3.2, we can show that under assumption ($\widetilde{\mathrm{RE}}(s, r_0)$), we have an empirical Restricted Eigenvalue condition on the matrix $\tilde{\mathbf{G}}_n$.

Lemma 4.2. *Let L defined as in Lemma 3.2. Under assumptions* (A_2) *and* ($\widetilde{\mathrm{RE}}(s, r_0)$), *we have*

$$0 < \tilde{\boldsymbol{\kappa}} = \min_{\substack{J \subset \{1,\dots,M\}, \\ |J| \le s}} \min_{\substack{\mathbf{b} \in \mathbb{R}^M \setminus \{0\}, \\ \|\mathbf{b}_{J^c}\|_1 \le r_0 \|\mathbf{b}_J\|_1}} \frac{(\mathbf{b}^T \tilde{\mathbf{G}}_n \mathbf{b})^{1/2}}{\|\mathbf{b}_J\|_2} \quad \text{and} \quad \tilde{\boldsymbol{\kappa}} = (1/\sqrt{2A_0})\tilde{\boldsymbol{\kappa}}_0(s, r_0), \tag{18}$$

with probability larger than $1 - \tilde{\pi}_n$, where

$$\tilde{\pi}_n = 2M^2 \exp\left[-\frac{n\tilde{\boldsymbol{\kappa}}^4}{2L^2(1+r_0)^2 s(L^2(1+r_0)^2 s + \tilde{\boldsymbol{\kappa}}^2/3)} \right].$$

Theorem 4.3. *For $x > 0$ and $y > 0$, let ω_j and δ_k be defined by* (5) *and* (6) *respectively. Let $A_{\varepsilon,\nu}(x) > 0$ and $B_{\tilde{\varepsilon},\tilde{\nu}}(y) > 0$ be two numerical positive constants depending on ε, ν, x and $\tilde{\varepsilon}, \tilde{\nu}, y$ respectively, $\zeta > 0$ and $s \in \{1, \dots, M+N\}$ be fixed. Let assumptions* (A_2), (A_3), (A_4), (A_7) *and* ($\widetilde{\mathrm{RE}}(s, r_0)$) *be satisfied with*

$$r_0 = \big(3 + 8\max\big(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\big)/\zeta\big),$$

and let $\tilde{\boldsymbol{\kappa}} = (1/\sqrt{2A_0})\tilde{\boldsymbol{\kappa}}_0(s, r_0)$. Then, with probability larger than $1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-x} - B_{\tilde{\varepsilon},\tilde{\nu}}(y)\mathrm{e}^{-y} - \tilde{\pi}_n$

$$\begin{aligned} &\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L^\rho, \hat{\gamma}_L^\rho}) \\ &\quad \le (1+\zeta) \inf_{\substack{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho) \\ \max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|) \le s}} \left\{ \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + \widetilde{C}(\zeta, \rho) \frac{\max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|)}{\tilde{\boldsymbol{\kappa}}^2} \max_{\substack{1 \le j \le M \\ 1 \le k \le N}} \{\omega_j^2, \delta_k^2\} \right\}, \end{aligned} \tag{19}$$

and

$$\begin{aligned} &\|\log\lambda_0 - \log\lambda_{\hat{\beta}_L^\rho, \hat{\gamma}_L^\rho}\|_{n,\Lambda}^2 \\ &\quad \le (1+\zeta) \inf_{\substack{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho) \\ \max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|) \le s}} \left\{ \|\log\lambda_0 - \log\lambda_{\beta,\gamma}\|_{n,\Lambda}^2 + \widetilde{C}'(\zeta, \rho) \frac{\max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|)}{\tilde{\boldsymbol{\kappa}}^2} \max_{\substack{1 \le j \le M \\ 1 \le k \le N}} \{\omega_j^2, \delta_k^2\} \right\}, \end{aligned} \tag{20}$$

where $\widetilde{C}(\zeta, \rho) > 0$ and $\widetilde{C}'(\zeta, \rho) > 0$ are constants depending only on ζ and ρ.

We obtain a non-asymptotic fast oracle inequality in prediction. Indeed, the rate of convergence of this oracle inequality is of order

$$\Big(\max_{\substack{1 \le j \le M \\ 1 \le k \le N}} \{\omega_j, \delta_k\} \Big)^2 \approx \max\left\{ \frac{\log M}{n}, \frac{\log N}{n} \right\},$$

namely, if we choose $\mathbb{G}_N$ of size n, the rate of convergence of this oracle inequality is then of order $\log M/n$ (see Section 4.1 for more details). While Estimation procedure 2.3 allows to derive a prediction for the survival time through the conditional intensity, Theorem 4.3 measures the accuracy of this prediction. In that sense, the clinical problem of establishing a prognosis has been addressed at this point. To our best knowledge, this oracle inequality is the first non-asymptotic oracle inequality in prediction for the whole intensity with a fast rate of convergence of order $\log M/n$.

For the part depending on the covariates, recent results establish non-asymptotic oracle inequalities for the Lasso estimator of f_0 in the usual Cox model (see Bradic and Song [9] and Kong and Nan [24]). We cannot compare our results to theirs, since we estimate the whole intensity with the total empirical log-likelihood whereas both of them consider the partial log-likelihood.

The remaining part of the paper is devoted to the technical results and proofs.

5. An empirical Bernstein's inequality

The main ingredient of Proposition 3.1 and Theorems 3.4, 4.1 and 4.3 are Bernstein's concentration inequalities that we present in this section. To clarify the relation between the stated oracle inequalities and the Bernstein's inequality, we sketch here the proof of Theorem 4.1. Using the Doob–Meyer decomposition $N_i = M_i + \Lambda_i$, we can easily show that for all $\boldsymbol{\beta} \in \mathbb{R}^M$ and for all $\boldsymbol{\gamma} \in \mathbb{R}^N$

$$C_n(\lambda_{\hat{\beta}_L,\hat{\gamma}_L}) - C_n(\lambda_{\beta,\gamma}) = \widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) - \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + (\hat{\boldsymbol{\gamma}}_L - \boldsymbol{\gamma})^T \boldsymbol{\nu}_{n,\tau} + (\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})^T \boldsymbol{\eta}_{n,\tau}, \tag{21}$$

where

$$\boldsymbol{\eta}_{n,\tau} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \vec{\mathbf{f}}(\mathbf{Z}_i)\,\mathrm{d}M_i(t) \quad \text{and} \quad \boldsymbol{\nu}_{n,\tau} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \vec{\boldsymbol{\theta}}(t)\,\mathrm{d}M_i(t), \tag{22}$$

with $\vec{\mathbf{f}} = (f_1, \ldots, f_M)^T$ and $\vec{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_N)^T$. By definition of the Lasso estimator, we have for all $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in $\mathbb{R}^M \times \mathbb{R}^N$

$$C_n(\lambda_{\hat{\beta}_L,\hat{\gamma}_L}) + \mathrm{pen}(\hat{\boldsymbol{\beta}}_L) + \mathrm{pen}(\hat{\boldsymbol{\gamma}}_L) \le C_n(\lambda_{\beta,\gamma}) + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\gamma}),$$

and we finally obtain

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) \le \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + (\hat{\boldsymbol{\gamma}}_L - \boldsymbol{\gamma})^T \boldsymbol{\nu}_{n,\tau} + (\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})^T \boldsymbol{\eta}_{n,\tau} + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\hat{\boldsymbol{\beta}}_L) + \mathrm{pen}(\boldsymbol{\gamma}) - \mathrm{pen}(\hat{\boldsymbol{\gamma}}_L).$$

Consequently, $\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L})$ is bounded by

$$\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + \sum_{j=1}^{M} (\hat{\beta}_{L,j} - \beta_j)\eta_{n,\tau}(f_j) + \sum_{j=1}^{M} \omega_j \big(|\beta_j| - |\hat{\beta}_{L,j}|\big)$$
$$+ \sum_{k=1}^{N} (\hat{\gamma}_{L,k} - \gamma_k)^T \nu_{n,\tau}(\theta_k) + \sum_{k=1}^{N} \delta_k \big(|\gamma_k| - |\hat{\gamma}_{L,k}|\big),$$

with

$$\eta_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t f_j(\mathbf{Z}_i)\,\mathrm{d}M_i(s) \quad \text{and} \quad \nu_{n,t}(\theta_k) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t \theta_k(s)\,\mathrm{d}M_i(s).$$

We will control $\eta_{n,\tau}(f_j)$ and $\nu_{n,\tau}(\theta_k)$ respectively by ω_j and δ_k. More precisely, the weights ω_j (respectively δ_k) will be chosen such that $|\eta_{n,\tau}(f_j)| \le \omega_j$ (respectively $|\nu_{n,\tau}(\theta_k)| \le \delta_k$) and $\mathbb{P}(|\eta_{n,\tau}(f_j)| > \omega_j)$ (respectively $\mathbb{P}(|\nu_{n,\tau}(\theta_k)| > \delta_k)$ large. As $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ involve martingales, we could directly apply classical Bernstein's inequalities for martingales with $x > 0$ and $y > 0$

$$\mathbb{P}\left[\eta_{n,t}(f_j) \ge \sqrt{\frac{2V_{n,t}(f_j)x}{n}} + \frac{x}{3n}\right] \le \mathrm{e}^{-x} \quad \text{and} \quad \mathbb{P}\left[\nu_{n,t}(\theta_k) \ge \sqrt{\frac{2R_{n,t}(\theta_k)y}{n}} + \frac{y}{3n}\right] \le \mathrm{e}^{-y},$$

where the predictable variations $V_{n,t}(f_j)$ and $R_{n,t}(\theta_k)$ of $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ are respectively defined by

$$V_{n,t}(f_j) = n\big\langle \eta_n(f_j) \big\rangle_t = \frac{1}{n}\sum_{i=1}^{n} \int_0^t \big(f_j(\mathbf{Z}_i)\big)^2 \lambda_0(t, \mathbf{Z}_i) Y_i(s)\,\mathrm{d}s,$$

$$R_{n,t}(\theta_k) = n\big\langle \nu_n(\theta_k) \big\rangle_t = \frac{1}{n}\sum_{i=1}^{n} \int_0^t \big(\theta_k(t)\big)^2 \lambda_0(t, \mathbf{Z}_i) Y_i(s)\,\mathrm{d}s,$$

see, e.g., van de Geer [37]. Applying these inequalities, the weights of Estimation procedure 2.3 would have the forms $\omega_j = \sqrt{2V_{n,t}(f_j)x/n} + x/3n$ and $\delta_k = \sqrt{2R_{n,t}(\theta_k)y/n} + y/3n$. As $V_{n,t}(f_j)$ and $R_{n,t}(\theta_k)$ both depend on λ_0, this would not result a statistical procedure. We propose to replace in the Bernstein's inequality the predictable variations by the optional variations of the processes $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ defined by

$$\hat{V}_{n,t}(f_j) = n\big[\eta_n(f_j)\big]_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \big(f_j(\mathbf{Z}_i)\big)^2\,\mathrm{d}N_i(s) \quad \text{and} \quad \hat{R}_{n,t}(\theta_k) = n\big[\nu_n(\theta_k)\big]_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \big(\theta_k(t)\big)^2\,\mathrm{d}N_i(s).$$

This ensures that the weights ω_j and δ_k will depends on $\hat{V}_{n,t}(f_j)$ and $\hat{R}_{n,t}(\theta_k)$ respectively. Equivalent strategies in different models have been considered in Gaïffas and Guilloux [18] or Hansen et al. [21]. The following theorem states the resulting Bernstein's inequalities.

Theorem 5.1. *Let assumption* (A_2) *be satisfied. For any numerical constant* $\varepsilon > 0$, $\tilde{\varepsilon} > 0$, $c = \sqrt{2(1+\varepsilon)}$ *and* $\tilde{c} = \sqrt{2(1+\tilde{\varepsilon})}$, *the following holds for any* $x > 0$, $y > 0$:

$$\mathbb{P}\left[\big|\eta_{n,t}(f_j)\big| \geq c\sqrt{\frac{\hat{W}_n^{\nu}(f_j)x}{n}} + \frac{x}{3n}\|f_j\|_{n,\infty}\right] \leq \left(\frac{2}{\log(1+\varepsilon)}\log\left(2 + \frac{A_0(\nu/n + \Phi(\nu/n))}{x/n}\right) + 1\right)\mathrm{e}^{-x}, \tag{23}$$

$$\mathbb{P}\left[\big|\nu_{n,t}(\theta_k)\big| \geq \tilde{c}\sqrt{\frac{\hat{T}_n^{\tilde{\nu}}(\theta_k)y}{n}} + \frac{y}{3n}\|\theta_k\|_{\infty}\right] \leq \left(\frac{2}{\log(1+\tilde{\varepsilon})}\log\left(2 + \frac{A_0(\tilde{\nu}/n + \Phi(\tilde{\nu}/n))}{y/n}\right) + 1\right)\mathrm{e}^{-y}, \tag{24}$$

where

$$W_n^{\nu}(f_j) = \frac{\nu/n}{\nu/n - \Phi(\nu/n)}\hat{V}_n(f_j) + \frac{x/n}{\nu/n - \Phi(\nu/n)}\|f_j\|_{n,\infty}^2, \tag{25}$$

$$T_n^{\tilde{\nu}}(\theta_k) = \frac{\tilde{\nu}/n}{\tilde{\nu}/n - \Phi(\tilde{\nu}/n)}\hat{R}_n(\theta_k) + \frac{y/n}{\tilde{\nu}/n - \Phi(\tilde{\nu}/n)}\|\theta_k\|_{\infty}^2, \tag{26}$$

for real numbers $(\nu, \tilde{\nu}) \in (0,3)^2$ *such that* $\nu > \Phi(\nu)$ *and* $\tilde{\nu} > \Phi(\tilde{\nu})$, *where* $\Phi(u) = \exp(u) - u - 1$.

We deduce the weights ω_j and δ_k defined in (5) and (6) respectively, from Theorem 5.1. These empirical Bernstein's inequalities hold true for martingales with jumps, when the predictable variation is not observable.

Remark 5.2. *Theorem 5.1 is closed to Theorem* 3 *in Hansen et al.* [21], *although in our version the event bounding* $\hat{W}_n^{\nu}(f_j)$ *and* $\hat{T}_n^{\tilde{\nu}}(\theta_k)$ *has been removed from the probability* (*see the proof of Theorem 5.1*).

Other weights can also be obtained from empirical Bernstein's inequalities that are closer to those obtained by Gaïffas and Guilloux [18] *in Theorem* 3. *We refer to an other version of the paper* (*see* [26]), *in which these weights appear. Their forms are less simple than those defined in* (5) *and* (6), *but they do not depend on tuning parameters* ν *and* $\tilde{\nu}$ *to determine for the applications. An interesting perspective would be to determine which one of those two forms of weights gives the best results in the applications.*

Remark 5.3 (Censoring case). *In the specific case of right censoring, since* $\max_{1\leq i\leq n}|N_i(\tau)| \leq 1$, *we can directly apply the Bernstein type inequality for martingales of Hansen et al.* [21] *to get quite simpler right term in Inequality* (23). *Indeed in this case, for real numbers* $(u, v) \in (0,3)^2$ *such that* $u > \phi(u)$ *and* $v > \phi(v)$, *where* $\phi(z) = \exp(z) - z - 1$, *and* $c_{1,\varepsilon} = \sqrt{2(1+\varepsilon)}$, *we would get*

$$\mathbb{P}\left[\eta_{n,t}(f_j) \geq c\sqrt{\frac{\hat{W}_n^{u}(f_j)x}{n}} + \frac{x}{3n}\|f_j\|_{n,\infty}\right] \leq 4\left(\frac{\log(1+u/x)}{\log(1+\varepsilon)} + 1\right)\mathrm{e}^{-x}. \tag{27}$$

6. Technical results

In this section, we present the technical results, that are not useful for a first reading of the paper but useful for a better understanding of the theory and used in the proofs of Section 7. Associated proofs are in Appendices A and B.

6.1. *Bernstein concentration inequality*

We recall here the classical Bernstein concentration inequality (see Proposition 2.9 in Massart [29]).

Theorem 6.1. *Let $\zeta_1, \dots, \zeta_n$ be n independent real valued random variables. Assume that there exist some positive numbers v and c such that for all integers $m \geq 2$*

$$\sum_{i=1}^{n} \mathbb{E}\big[|\zeta_i|^m\big] \leq m! v c^{m-2}. \tag{28}$$

For any positive x we have

$$\mathbb{P}\left(\sum_{i}^{n}\big(\zeta_i - \mathbb{E}(\zeta_i)\big) \geq x\right) \leq \exp\left(-\frac{x^2}{2(v+cx)}\right). \tag{29}$$

Note that if the variables ζ_i are bounded, $|\zeta_i| \leq b$ for all i in $\{1, \dots, n\}$, then assumption (28) is satisfied with

$$v = \sum_{i=1}^{n} \mathbb{E}\big[\zeta_i^2\big] \quad \text{and} \quad c = b/3.$$

6.2. *Connection between the weighted empirical norm and the empirical Kullback divergence*

The following propositions connect the empirical Kullback divergence (3) to the weighted empirical norm (4) in the different cases considered in the paper.

Proposition 6.2 holds true when the intensity verifies model (2) with a known baseline hazard function α_0.

Proposition 6.2. *Under assumption* (A_5), *for all* $\boldsymbol{\beta} \in \Gamma(\mu)$,

$$\mu' \| \log \lambda_{\boldsymbol{\beta}} - \log \lambda_0 \|_{n,\Lambda}^2 \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) \leq \mu'' \| \log \lambda_{\boldsymbol{\beta}} - \log \lambda_0 \|_{n,\Lambda}^2,$$

where $\mu' = \phi(\mu)/\mu^2$, $\mu'' = \phi(-\mu)/\mu^2$ *and* $\phi(t) = \mathrm{e}^{-t} + t - 1$.

Proposition 6.2 can be rewriten in the particular case of variable selection in the Cox model as follows:

Proposition 6.3. *Under assumptions* (A_6) *and* (RE(s, a_0)) *with* $a_0 = 3$, *there exist two positive numerical constants* ξ *and* ξ' *such that*

$$\xi \left\| (\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0)^T \mathbf{X} \right\|_{n,\Lambda}^2 \leq \widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_L}) \leq \xi' \left\| (\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0)^T \mathbf{X} \right\|_{n,\Lambda}^2.$$

In the general case, when the intensity does not rely on an underlying model, the connection between the weighted empirical norm (4) and the empirical Kullback divergence (3) is given by the following proposition.

Proposition 6.4. *Under assumption* (A_7), *for all* $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$,

$$\rho' \| \log \lambda_{\beta,\gamma} - \log \lambda_0 \|_{n,\Lambda}^2 \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) \leq \rho'' \| \log \lambda_{\beta,\gamma} - \log \lambda_0 \|_{n,\Lambda}^2,$$

where $\rho' = \phi(\rho)/\rho^2$, $\rho'' = \phi(-\rho)/\rho^2$ *and* $\phi(t) = \mathrm{e}^{-t} + t - 1$.

7. Proofs

7.1. *Proof of Proposition 2.1*

Following the proof of Theorem 1 in Senoussi [32], we rewrite the empirical Kullback divergence (3) as

$$\widetilde{K}_n(\lambda_0,\lambda) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\log\lambda_0(t,\mathbf{Z}_i) - \log\lambda(t,\mathbf{Z}_i) - \left(1-\frac{\lambda(t,\mathbf{Z}_i)}{\lambda_0(t,\mathbf{Z}_i)}\right)\right]\lambda_0(t,\mathbf{Z}_i)Y_i(t)\,\mathrm{d}t$$
$$= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\exp\left(\log\frac{\lambda(t,\mathbf{Z}_i)}{\lambda_0(t,\mathbf{Z}_i)}\right) - \log\frac{\lambda(t,\mathbf{Z}_i)}{\lambda_0(t,\mathbf{Z}_i)} - 1\right]\lambda_0(t,\mathbf{Z}_i)Y_i(t)\,\mathrm{d}t.$$

Since $t\to \mathrm{e}^t - t - 1 > 0$, except for $t=0$, we deduce that, except for $\lambda=\lambda_0$,

$$\exp\left(\log\frac{\lambda(t,\mathbf{Z}_i)}{\lambda_0(t,\mathbf{Z}_i)}\right) - \log\frac{\lambda(t,\mathbf{Z}_i)}{\lambda_0(t,\mathbf{Z}_i)} - 1 > 0.$$

Thus $\widetilde{K}_n(\lambda_0,\lambda)$ is positive and vanishes only if $(\log\lambda_0 - \log\lambda)(t,\mathbf{Z}_i) = 0$ almost surely, namely if $\lambda_0=\lambda$ almost surely.

7.2. *Proof of Proposition 3.1*

According to the definition (9) of $\hat{\boldsymbol{\beta}}_L$, for all $\boldsymbol{\beta}$ in $\mathbb{R}^M$, we have

$$C_n(\lambda_{\hat{\beta}_L}) + \mathrm{pen}(\hat{\boldsymbol{\beta}}_L) \le C_n(\lambda_{\boldsymbol{\beta}}) + \mathrm{pen}(\boldsymbol{\beta}).$$

Here α_0 is assumed to be known. Hence applying (21), we obtain

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L}) \le \widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})^T\boldsymbol{\eta}_{n,\tau} + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\hat{\boldsymbol{\beta}}_L). \tag{30}$$

It remains to control the term $(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})^T\boldsymbol{\eta}_{n,\tau}$. For ω_j defined in (5), set

$$\mathcal{A} = \bigcap_{j=1}^{M}\left\{\left|\eta_{n,\tau}(f_j)\right| \le \frac{\omega_j}{2}\right\}. \tag{31}$$

On $\mathcal{A}$, we have

$$\left|(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})^T\boldsymbol{\eta}_{n,\tau}\right| \le \sum_{j=1}^{M}\omega_j\left|(\hat{\beta}_L - \beta)_j\right|. \tag{32}$$

The result (10) follows since $\mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M}\omega_j|\beta_j|$. It remains to bound up $\mathbb{P}(\mathcal{A}^c)$. By applying Theorem 5.1

$$\mathbb{P}\big(\mathcal{A}^c\big) \le \sum_{j=1}^{M}\mathbb{P}\left(\left|\eta_{n,\tau}(f_j)\right| > \frac{\omega_j}{2}\right) \le A_{\varepsilon,\nu}(x)\mathrm{e}^{-x},$$

with

$$A_{\varepsilon,\nu}(x) = \frac{2}{\log(1+\varepsilon)}\log\left(2 + \frac{A_0(\nu/n + \Phi(\nu/n))}{x/n}\right) + 1. \tag{33}$$

We conclude that $\mathbb{P}(\mathcal{A}) \ge 1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-x}$, which ends up the proof of Theorem 3.1.

7.3. *Proof of Lemma 3.2*

We show with high probability, that under (RE(s, a_0)), for all $J \subset \{1, \dots, M\}$ such that $|J| \leq s$ and for all $b \in \mathbb{R}^M \setminus \{0\}$ such that $\|b_{J^c}\|_1 \leq a_0 \|b_J\|_1$,

$$\frac{\mathbf{b}^T \mathbf{G}_n \mathbf{b}}{\|\mathbf{b}_J\|_2^2} > \kappa^2, \quad \text{with } \kappa = (1/\sqrt{2A_0})\kappa_0(s, a_0) \text{ and } A_0 \text{ defined in assumption } (A_2).$$

Let consider the set $\Omega_{G_n,t} = \{|(\mathbf{G}_n - \mathbb{E}(\mathbf{G}_n))_{j,k}| \leq t, \forall (j,k) \in \{1, \dots, M\}^2\}$, with a fixed $t \geq 0$. Under (RE(s, a_0)), on $\Omega_{G_n,t}$, for all $J \subset \{1, \dots, M\}$ such that $|J| \leq s$ and for all $b \in \mathbb{R}^M \setminus \{0\}$ such that $\|b_{J^c}\|_1 \leq a_0 \|b_J\|_1$, we have

$$\mathbf{b}^T \mathbf{G}_n \mathbf{b} = \mathbf{b}^T \big(\mathbf{G}_n - \mathbb{E}(\mathbf{G}_n)\big)\mathbf{b} + \mathbf{b}^T \mathbb{E}(\mathbf{G}_n)\mathbf{b} \geq \mathbf{b}^T \big(\mathbf{G}_n - \mathbb{E}(\mathbf{G}_n)\big)\mathbf{b} + \kappa_0^2 \|\mathbf{b}_J\|_2^2.$$

On $\Omega_{G_n,t}$, under (RE(s, a_0)) we deduce that

$$\mathbf{b}^T \mathbf{G}_n \mathbf{b} \geq -\sum_{i,j} t|b_i||b_j| + \kappa_0^2 \|\mathbf{b}_J\|_2^2 \geq \big(-2t(1+a_0)^2 s + \kappa_0^2\big)\|\mathbf{b}_J\|_2^2.$$

We choose $t = A_0\kappa^2/(1+a_0)^2 s$ with $\kappa = \kappa_0/\sqrt{2A_0}$ to get $\mathbf{b}^T \mathbf{G}_n \mathbf{b} \geq \kappa_0^2 \|\mathbf{b}_J\|_2^2$.

It remains to calculate $\mathbb{P}(\Omega_{G_n,t})$. The coefficient (j,k) of the matrix $\mathbf{G}_n - \mathbb{E}(\mathbf{G}_n)$ is given by

$$\frac{1}{n}\sum_{i=1}^n \big(\Lambda_i - \mathbb{E}(\Lambda_i)\big) f_j(\mathbf{Z}_i) f_k(\mathbf{Z}_i).$$

For sake of simplicity, we put $\zeta_i^{j,k} = \Lambda_i f_j(\mathbf{Z}_i) f_k(\mathbf{Z}_i)$ for $i = 1, \dots, n$ and $(j,k) \in \{1, \dots, M\}^2$ fixed. Under assumptions (A_2) and (A_3), we can apply Bernstein's inequality (29) to get

$$\mathbb{P}\left(\big|\big(\mathbf{G}_n - \mathbb{E}(\mathbf{G}_n)\big)_{i,j}\big| > \frac{A_0\kappa^2}{(1+a_0)^2 s}\right) \leq 2\exp\left(-\frac{n\kappa^4}{2(1+a_0)^2 s L^2 (L^2(1+a_0)^2 s + \kappa^2/3)}\right).$$

So the probability of $\Omega_{G_n,t}^c$ with $t = A_0\kappa^2/(1+a_0)^2 s$ is given by

$$\mathbb{P}\big(\Omega_{G_n,t}^c\big) \leq 2M^2 \exp\left(-\frac{n\kappa^4}{2(1+a_0)^2 s L^2 (L^2(1+a_0)^2 s + \kappa^2/3)}\right),$$

via an union bound and by denoting

$$\pi_n = 2M^2 \exp\left(-\frac{n\kappa^4}{2(1+a_0)^2 s L^2 (L^2(1+a_0)^2 s + \kappa^2/3)}\right),$$

we finally get (13) with probability larger than $1 - \pi_n$.

7.4. *Proof of Theorem 3.4*

Let us introduce the event

$$\boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\kappa) = \left\{0 < \kappa = \min_{\substack{J \subset \{1,\dots,M\}, \\ |J| \leq s}} \min_{\substack{\mathbf{b} \in \mathbb{R}^M \setminus \{0\}, \\ \|\mathbf{b}_{J^c}\|_1 \leq a_0 \|\mathbf{b}_J\|_1}} \frac{(\mathbf{b}^T \mathbf{G}_n \mathbf{b})^{1/2}}{\|\mathbf{b}_J\|_2}\right\}. \tag{34}$$

From Inequality (30), on $\mathcal{A}$ defined by (31), for $\boldsymbol{\beta} \in \Gamma(\mu)$, it follows that

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L^\mu}) + \sum_{j=1}^M \frac{\omega_j}{2}\big|\big(\hat{\beta}_L^\mu - \beta\big)_j\big| \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + \sum_{j=1}^M \omega_j\big(\big|\big(\hat{\beta}_L^\mu - \beta\big)_j\big| + |\beta_j| - \big|\big(\hat{\beta}_L^\mu\big)_j\big|\big).$$

On $J(\boldsymbol{\beta})^c$, $|(\hat{\beta}_L^{\mu}-\beta)_j|+|\beta_j|-|(\hat{\beta}_L^{\mu})_j|=0$, so on $\mathcal{A}$ we obtain

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})+\sum_{j=1}^{M}\frac{\omega_j}{2}\big|\big(\hat{\beta}_L^{\mu}-\beta\big)_j\big|\leq\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})+2\sum_{j\in J(\beta)}\omega_j\big|\big(\hat{\beta}_L^{\mu}-\beta\big)_j\big|. \tag{35}$$

We apply Cauchy–Schwarz inequality to the second right-hand side of (35) to get

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})+\sum_{j=1}^{M}\frac{\omega_j}{2}\big|\big(\hat{\beta}_L^{\mu}-\beta\big)_j\big|\leq\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})+2\sqrt{\big|J(\boldsymbol{\beta})\big|}\sqrt{\sum_{j\in J(\beta)}\omega_j^2\big|\hat{\beta}_L^{\mu}-\beta\big|_j^2}. \tag{36}$$

With the notations $\boldsymbol{\Delta}=\mathbf{D}(\hat{\boldsymbol{\beta}}_L^{\mu}-\boldsymbol{\beta})$ and $\mathbf{D}=(\operatorname{diag}(\omega_j))_{1\leq j\leq M}$ introduced in Section 3.2, Inequalities (35) and (36) become respectively

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})+\frac{1}{2}\|\boldsymbol{\Delta}\|_1\leq\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})+2\|\boldsymbol{\Delta}_{J(\beta)}\|_1, \tag{37}$$

and

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})\leq\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})+2\sqrt{\big|J(\boldsymbol{\beta})\big|}\|\boldsymbol{\Delta}_{J(\beta)}\|_2. \tag{38}$$

We fix some $\zeta>0$ and we consider the following set

$$\mathcal{A}_1=\big\{\zeta\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})\leq 2\|\boldsymbol{\Delta}_{J(\beta)}\|_1\big\}. \tag{39}$$

Here, we could take $\zeta=1$, but this parameter ζ allows to have more freedom. The smaller ζ is, the higher $\mathbb{P}(\mathcal{A}_1)$ is, but the smaller $\mathbb{P}(\boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa}))$ is. So ζ realizes a compromise between these two probabilities. On $\mathcal{A}\cap\mathcal{A}_1^c$, the result of the theorem follows immediately from (37). As soon as, $\|\boldsymbol{\Delta}_{J(\beta)^c}\|_1\leq(3+4/\zeta)\|\boldsymbol{\Delta}_{J(\beta)}\|_1$, on $\boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa})$, with $a_0=(3+4/\zeta)$ and $\boldsymbol{\kappa}=(1/\sqrt{2A_0})\boldsymbol{\kappa}_0(s,a_0)$ we get

$$\boldsymbol{\kappa}^2\|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}\|_2^2\leq\boldsymbol{\Delta}^T\mathbf{G}_n\boldsymbol{\Delta}.$$

So, initially we will assume that $\|\boldsymbol{\Delta}_{J(\beta)^c}\|_1\leq(3+4/\zeta)\|\boldsymbol{\Delta}_{J(\beta)}\|_1$, and we will verify later that this inequality holds. Since

$$\boldsymbol{\Delta}^T\mathbf{G}_n\boldsymbol{\Delta}\leq\Big(\max_{1\leq j\leq M}\omega_j\Big)^2\|\log\lambda_{\hat{\beta}_L^{\mu}}-\log\lambda_{\boldsymbol{\beta}}\|_{n,\Lambda}^2.$$

Inequality (38) becomes on $\mathcal{A}\cap\boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa})$

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})\leq\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})+2\sqrt{\big|J(\boldsymbol{\beta})\big|}\Big(\max_{1\leq j\leq M}\omega_j\Big)\boldsymbol{\kappa}^{-1}\big(\|\log\lambda_{\hat{\beta}_L^{\mu}}-\log\lambda_0\|_{n,\Lambda}+\|\log\lambda_0-\log\lambda_{\boldsymbol{\beta}}\|_{n,\Lambda}\big).$$

Now, applying Proposition 6.2 to connect the weighted empirical norm to the empirical Kullback divergence, it follows that

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})\leq\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})+2\sqrt{\big|J(\boldsymbol{\beta})\big|}\Big(\max_{1\leq j\leq M}\omega_j\Big)\frac{\boldsymbol{\kappa}^{-1}}{\sqrt{\mu'}}\big(\sqrt{\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})}+\sqrt{\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})}\big).$$

We now use the elementary inequality $2uv\leq bu^2+\frac{v^2}{b}$ with $b>0$, $u=\sqrt{|J(\boldsymbol{\beta})|}(\max_{1\leq j\leq M}\omega_j)\boldsymbol{\kappa}^{-1}$ and v being either $\sqrt{\frac{1}{\mu'}\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})}$ or $\sqrt{\frac{1}{\mu'}\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})}$. Consequently

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})\leq\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}})+2b\big|J(\boldsymbol{\beta})\big|\Big(\max_{1\leq j\leq M}\omega_j\Big)^2\boldsymbol{\kappa}^{-2}+\frac{1}{b\mu'}\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^{\mu}})+\frac{1}{b\mu'}\widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}}).$$

Hence,

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L^{\mu}}) \leq \frac{b\mu'+1}{b\mu'-1}\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2\frac{b^2\mu'}{b\mu'-1}\big|J(\beta)\big|\Big(\max_{1\leq j\leq M}\omega_j\Big)^2\boldsymbol{\kappa}^{-2}.$$

We take $\frac{b\mu'+1}{b\mu'-1} = 1+\zeta$ and $C(\zeta,\mu) = 2\frac{b^2\mu'}{b\mu'+1}$ a constant depending on ζ and μ. It follows that for any $\boldsymbol{\beta} \in \Gamma(\mu)$:

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L^{\mu}}) \leq (1+\zeta)\Big\{\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + C(\zeta,\mu)\big|J(\boldsymbol{\beta})\big|\Big(\max_{1\leq j\leq M}\omega_j\Big)^2\boldsymbol{\kappa}^{-2}\Big\}.$$

Finally, taking the infimum over all $\boldsymbol{\beta} \in \Gamma(\mu)$ such that $|J(\boldsymbol{\beta})| \leq s$, we obtain (14).

We have now to verify that $\|\boldsymbol{\Delta}_{J(\beta)^c}\|_1 \leq (3+4/\zeta)\|\boldsymbol{\Delta}_{J(\beta)}\|_1$. On $\mathcal{A}\cap\mathcal{A}_1$, applying (37) we get that

$$\|\boldsymbol{\Delta}\|_1 \leq 4\left(1+\frac{1}{\zeta}\right)\|\boldsymbol{\Delta}_{J(\beta)}\|_1,$$

so by splitting $\boldsymbol{\Delta} = \boldsymbol{\Delta}_{J(\boldsymbol{\beta})} + \boldsymbol{\Delta}_{J(\boldsymbol{\beta})^c}$, we finally obtain

$$\|\boldsymbol{\Delta}_{J(\beta)^c}\|_1 \leq \left(3+\frac{4}{\zeta}\right)\|\boldsymbol{\Delta}_{J(\beta)}\|_1.$$

Finally, Lemma 3.2 ensures that $\mathbb{P}(\mathcal{A}^c \cup \boldsymbol{\Omega}^c_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa})) \leq A_{\varepsilon,\nu}(x)\mathrm{e}^{-x} + \pi_n$, which achieves the proof of Theorem 3.4.

7.5. *Proof of Theorem 3.6*

To prove Inequality (15) of Theorem 3.6, we start from (35) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_L$ defined by (9). Consequently $\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) = 0$. Applying Proposition 6.3 with $\lambda_0(t,\mathbf{Z}_i) = \alpha_0(t)\mathrm{e}^{\boldsymbol{\beta}_0^T\mathbf{Z}_i}$ and $\lambda_{\hat{\beta}_L}(t,\mathbf{Z}_i) = \alpha_0(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}_L^T\mathbf{Z}_i}$, we obtain that, on $\mathcal{A}_{\Gamma_1} = \bigcap_{j=1}^p\{|\eta_{n,\tau}(f_j)| \leq \Gamma_1\frac{\omega_j}{2}\}$

$$\xi\left\|(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0)^T\mathbf{X}\right\|^2_{n,\Lambda} + \Gamma_1\sum_{j=1}^p\frac{\omega_j}{2}|\hat{\beta}_L - \beta_0|_j \leq 2\Gamma_1\sum_{j\in J_0}\omega_j|\hat{\beta}_L - \beta_0|_j. \tag{40}$$

From this inequality, we deduce

$$\xi\left\|\mathbf{X}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0)\right\|^2_{n,\Lambda} \leq 2\Gamma_1\sum_{j\in J_0}\omega_j|\hat{\beta}_L - \beta_0|_j \leq 2\sqrt{|J_0|}\Gamma_1\|\boldsymbol{\Delta}_{0,J_0}\|_2. \tag{41}$$

From (40), we also have

$$\sum_{j=1}^p\omega_j|\hat{\beta}_L - \beta_0|_j \leq 4\sum_{j\in J_0}\omega_j|\hat{\beta}_L - \beta_0|_j$$

and we obtain $\|\boldsymbol{\Delta}_0\|_1 \leq 4\|\boldsymbol{\Delta}_{0J_0}\|_1$. We then split $\|\boldsymbol{\Delta}_0\|_1 = \|\boldsymbol{\Delta}_{0J_0}\|_1 + \|\boldsymbol{\Delta}_{0J_0^c}\|_1$ to get

$$\|\boldsymbol{\Delta}_{0J_0^c}\|_1 \leq 3\|\boldsymbol{\Delta}_{0J_0}\|_1. \tag{42}$$

On $\boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa}')$, defined by (34), with $a_0 = 3$ and $\boldsymbol{\kappa}' = (1/\sqrt{2A_0})\boldsymbol{\kappa}_0(s,3)$ we get

$$\|\mathbf{X}\boldsymbol{\Delta}_0\|^2_{n,\Lambda} \geq \boldsymbol{\kappa}'^2\|\boldsymbol{\Delta}_{0,J_0}\|_2^2. \tag{43}$$

According to (41), we conclude that on $\mathcal{A}_{\Gamma_1} \cap \boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa}')$

$$\xi \left\| \mathbf{X}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0) \right\|_{n,\Lambda}^2 \leq 2\sqrt{|J_0|}\Gamma_1 \max_{1\leq j\leq p} \omega_j \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0)\|_{n,\Lambda}}{\boldsymbol{\kappa}'},$$

which entails that

$$\left\| \mathbf{X}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0) \right\|_{n,\Lambda}^2 \leq \frac{4|J_0|}{\xi^2 \boldsymbol{\kappa}'^2} \Gamma_1^2 \Big(\max_{1\leq j\leq p} \omega_j \Big)^2,$$

with $\mathbb{P}(\mathcal{A}_{\Gamma_1} \cap \boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa}')) \geq 1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-\Gamma_1 x} - \pi_n$.

Let us come to the proof of Inequality (16) in Theorem 3.6. On $\mathcal{A}_{\Gamma_1} \cap \boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa}')$, with $a_0 = 3$, Inequality (41) becomes

$$\xi \frac{\boldsymbol{\kappa}'^2}{\max_{1\leq j\leq M} \omega_j^2} \|\boldsymbol{\Delta}_{0,J_0}\|_2^2 \leq 2\sqrt{|J_0|}\Gamma_1 \|\boldsymbol{\Delta}_{0,J_0}\|_2. \tag{44}$$

According to (42) and thanks to Cauchy–Schwarz inequality, we have

$$\|\boldsymbol{\Delta}_0\|_1 = \|\boldsymbol{\Delta}_{0J_0}\|_1 + \|\boldsymbol{\Delta}_{0J_0^c}\|_1 \leq 4\|\boldsymbol{\Delta}_{0J_0}\|_1 \leq 4\sqrt{|J_0|}\|\boldsymbol{\Delta}_{0J_0}\|_2.$$

From (44), we get

$$\frac{\|\boldsymbol{\Delta}_0\|_1}{4\sqrt{|J_0|}} \leq \frac{2\sqrt{|J_0|}}{\xi \boldsymbol{\kappa}'^2} \Gamma_1 \max_{1\leq j\leq p} \omega_j^2,$$

and finally

$$\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_1 \leq 8 \frac{|J_0|}{\xi \boldsymbol{\kappa}'^2} \Gamma_1 \frac{\max_{1\leq j\leq p} \omega_j^2}{\min_{1\leq j\leq p} \omega_j},$$

with $\mathbb{P}(\mathcal{A}_{\Gamma_1} \cap \boldsymbol{\Omega}_{\mathbf{RE}_n(s,a_0)}(\boldsymbol{\kappa}')) \geq 1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-\Gamma_1 x} - \pi_n$.

7.6. *Proof of Theorem 4.1*

The proof is very similar to the one of Theorem 3.1. We start from (21) and (22), and write

$$\begin{aligned} \widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) &+ (\hat{\boldsymbol{\gamma}}_L - \boldsymbol{\gamma})^T \boldsymbol{\nu}_{n,\tau} + \mathrm{pen}(\boldsymbol{\gamma}) - \mathrm{pen}(\hat{\boldsymbol{\gamma}}_L) \\ &+ (\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta})^T \boldsymbol{\eta}_{n,\tau} + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\hat{\boldsymbol{\beta}}_L). \end{aligned} \tag{45}$$

Consider the set $\mathcal{A}$ defined by (31) and let define similarly the set $\mathcal{B}$ such that

$$\mathcal{B} = \bigcap_{k=1}^{N} \left\{ |\nu_{n,\tau}(\theta_k)| \leq \frac{\delta_k}{2} \right\}. \tag{46}$$

Applying Theorem 5.1, we obtain that $\mathbb{P}(\mathcal{A}^c) \leq A_{\varepsilon,\nu}(x)\mathrm{e}^{-x}$ and $\mathbb{P}(\mathcal{B}^c) \leq B_{\tilde{\varepsilon},\tilde{\nu}}(y)\mathrm{e}^{-y}$, with $A_{\varepsilon,\nu}(x)$ defined by (33) and

$$B_{\tilde{\varepsilon},\tilde{\nu}}(y) = \frac{2}{\log(1+\tilde{\varepsilon})} \log\left(2 + \frac{A_0(\tilde{\nu}/n + \Phi(\tilde{\nu}/n))}{y/n} \right) + 1,$$

and thus

$$\mathbb{P}\big[(\mathcal{A} \cap \mathcal{B})^c\big] = \mathbb{P}\big(\mathcal{A}^c \cup \mathcal{B}^c\big) \leq \mathbb{P}\big(\mathcal{A}^c\big) + \mathbb{P}\big(\mathcal{B}^c\big) \leq A_{\varepsilon,\nu}(x)\mathrm{e}^{-x} + B_{\tilde{\varepsilon},\tilde{\nu}}(y)\mathrm{e}^{-y}. \tag{47}$$

On $\mathcal{A} \cap \mathcal{B}$ arguing as in the proof of Theorem 3.1, with probability larger than $1 - A_{\varepsilon,\nu}(x)\mathrm{e}^{-x} - B_{\tilde{\varepsilon},\tilde{\nu}}(y)\mathrm{e}^{-y}$, we finish the proof by writing (17).

7.7. *Proof of Theorem 4.3*

Let introduce the event $\boldsymbol{\Omega}_{\widetilde{\mathbf{RE}}_n(s,r_0)}(\tilde{\boldsymbol{\kappa}}) = \{0 < \tilde{\boldsymbol{\kappa}} = \min_{J\subset\{1,\dots,M\},|J|\le s} \min_{\mathbf{b}\in\mathbb{R}^M\setminus\{0\},\|\mathbf{b}_{J^c}\|_1\le r_0\|\mathbf{b}_J\|_1} \frac{(\mathbf{b}^T\tilde{\mathbf{G}}_n\mathbf{b})^{1/2}}{\|\mathbf{b}_J\|_2}\}$. From Inequality (45), on $\mathcal{A}\cap\mathcal{B}$ defined in (31) and (46), for $(\boldsymbol{\beta},\boldsymbol{\gamma})\in\widetilde{\Gamma}(\rho)$, we obtain

$$\begin{aligned}&\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^\rho,\hat{\gamma}_L^\rho}) + \sum_{j=1}^M \frac{\omega_j}{2}\big|\big(\hat{\beta}_L^\rho-\beta\big)_j\big| + \sum_{k=1}^N \frac{\delta_k}{2}\big|\big(\hat{\gamma}_L^\rho-\gamma\big)_k\big| \\ &\quad\le \widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma}) + 2\sum_{j\in J(\boldsymbol{\beta})}\omega_j\big|\big(\hat{\beta}_L^\rho-\beta\big)_j\big| + 2\sum_{k\in J(\boldsymbol{\gamma})}\delta_k\big|\big(\hat{\gamma}_L^\rho-\gamma\big)_k\big|. \end{aligned} \tag{48}$$

We then apply Cauchy–Schwarz inequality to the second right-term of (48) and obtain

$$\begin{aligned}&\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^\rho,\hat{\gamma}_L^\rho}) + \sum_{j=1}^M \frac{\omega_j}{2}\big|\big(\hat{\beta}_L^\rho-\beta\big)_j\big| + \sum_{k=1}^N \frac{\delta_k}{2}\big|\big(\hat{\gamma}_L^\rho-\gamma\big)_k\big| \\ &\quad\le \widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma}) + 2\sqrt{\big|J(\boldsymbol{\beta})\big|}\sqrt{\sum_{j\in J(\beta)}\omega_j^2\big|\hat{\beta}_L^\rho-\beta\big|_j^2} + 2\sqrt{\big|J(\boldsymbol{\gamma})\big|}\sqrt{\sum_{k\in J(\boldsymbol{\gamma})}\delta_k^2\big|\hat{\gamma}_L^\rho-\gamma\big|_k^2}. \end{aligned} \tag{49}$$

With the notations of Section 4.2, Inequality (48) is rewritten as:

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^\rho,\hat{\gamma}_L^\rho}) + \frac{1}{2}\|\tilde{\boldsymbol{\Delta}}\|_1 \le \widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma}) + 2\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}\|_1, \tag{50}$$

where $\tilde{\boldsymbol{\Delta}}_{J(\beta),J(\gamma)} = \tilde{\mathbf{D}}\binom{(\hat{\boldsymbol{\beta}}_L^\rho-\boldsymbol{\beta})_{J(\boldsymbol{\beta})}}{(\hat{\boldsymbol{\gamma}}_L^\rho-\boldsymbol{\gamma})_{J(\boldsymbol{\gamma})}}$. In the same way, Inequality (49) becomes:

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^\rho,\hat{\gamma}_L^\rho}) \le \widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma}) + 4\max\Big(\sqrt{\big|J(\boldsymbol{\beta})\big|},\sqrt{\big|J(\boldsymbol{\gamma})\big|}\Big)\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}\|_2. \tag{51}$$

Consider

$$\tilde{\mathcal{A}}_1 = \big\{\zeta\,\widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma}) \le 2\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}\|_1\big\}. \tag{52}$$

On $\mathcal{A}\cap\mathcal{B}\cap\tilde{\mathcal{A}}_1$, Inequality (19) in Theorem 4.3 follows immediately from (50). As soon as, $\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c,J(\boldsymbol{\gamma})^c}\|_1 \le (3+8\max(\sqrt{|J(\boldsymbol{\beta})|},\sqrt{|J(\boldsymbol{\gamma})|})/\zeta)\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}\|_1$, on $\boldsymbol{\Omega}_{\widetilde{\mathbf{RE}}_n(s,r_0)}(\tilde{\boldsymbol{\kappa}})$, with

$$\tilde{\boldsymbol{\kappa}} = (1/\sqrt{2})\tilde{\boldsymbol{\kappa}}_0(s,r_0) \quad\text{and}\quad r_0 = \Big(3+8\max\Big(\sqrt{\big|J(\boldsymbol{\beta})\big|},\sqrt{\big|J(\boldsymbol{\gamma})\big|}\Big)/\zeta\Big),$$

we get that

$$\tilde{\boldsymbol{\kappa}}^2\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}\|_2^2 \le \tilde{\boldsymbol{\Delta}}^T\tilde{\mathbf{G}}_n\tilde{\boldsymbol{\Delta}} \quad\text{with } \tilde{\boldsymbol{\Delta}}^T\tilde{\mathbf{G}}_n\tilde{\boldsymbol{\Delta}} \le \max_{\substack{1\le j\le M\\ 1\le k\le N}}\{\omega_j,\delta_k\}\|\log\lambda_{\hat{\beta}_L^\rho,\hat{\gamma}_L^\rho} - \log\lambda_{\beta,\gamma}\|_{n,\Lambda}^2.$$

On $\mathcal{A}\cap\mathcal{B}\cap\boldsymbol{\Omega}_{\widetilde{\mathbf{RE}}_n(s,r_0)}(\tilde{\boldsymbol{\kappa}})$, from Equation (51) and applying Proposition 6.4 to connect the weighted empirical norm to the empirical Kullback divergence, we obtain that $\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^\rho,\hat{\gamma}_L^\rho})$ is less than

$$\widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma}) + 4\max\Big(\sqrt{\big|J(\boldsymbol{\beta})\big|},\sqrt{\big|J(\boldsymbol{\gamma})\big|}\Big)\max_{\substack{1\le j\le M\\ 1\le k\le N}}\{\omega_j,\delta_k\}\frac{\tilde{\boldsymbol{\kappa}}^{-1}}{\sqrt{\rho'}}\Big(\sqrt{\widetilde{K}_n(\lambda_0,\lambda_{\hat{\beta}_L^\rho,\hat{\gamma}_L^\rho})} + \sqrt{\widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma})}\Big).$$

Using again $2uv \le bu^2 + \frac{v^2}{b}$ with $b > 0$, $u = 2\max(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}) \max_{1\le j\le M, 1\le k\le N}\{\omega_j, \delta_k\}\tilde{\boldsymbol{\kappa}}^{-1}$ and v being either $\sqrt{\frac{1}{\rho'}\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L^\rho, \hat{\gamma}_L^\rho})}$ or $\sqrt{\frac{1}{\rho'}\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma})}$, we obtain

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L^\rho, \hat{\gamma}_L^\rho}) \le \frac{b\rho'+1}{b\rho'-1}\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 8\frac{b^2\rho'}{b\rho'-1}\max\big(\big|J(\boldsymbol{\beta})\big|, \big|J(\boldsymbol{\gamma})\big|\big)\Big(\max_{\substack{1\le j\le M\\ 1\le k\le N}}\{\omega_j, \delta_k\}\Big)^2\frac{\tilde{\boldsymbol{\kappa}}^{-2}}{\rho'}. \tag{53}$$

Finally, taking $\frac{b\rho'+1}{b\rho'-1} = 1+\zeta$ and $\widetilde{C}(\zeta,\rho) = 8\frac{b^2\rho'}{b\rho'+1}$ a constant depending on ζ and ρ, and taking the infimum over all $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$ such that $\max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|) \le s$, we obtain Inequality (19).

Inequality (20) follows by applying Proposition 6.2 with $b = \frac{(1+\zeta)\rho'+\rho''}{(1+\zeta)\rho'-\rho''}$ in (53).

We have now to verify that $\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c, J(\boldsymbol{\gamma})^c}\|_1 \le (3 + 8\max(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|})/\zeta)\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}), J(\boldsymbol{\gamma})}\|_1$. We deduce from (50) that, on $\mathcal{A}\cap\mathcal{B}\cap\tilde{\mathcal{A}}_1$, by splitting $\tilde{\boldsymbol{\Delta}} = \tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}), J(\boldsymbol{\gamma})} + \tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c, J(\boldsymbol{\gamma})^c}$

$$\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c, J(\boldsymbol{\gamma})^c}\|_1 \le \left(3 + \frac{8}{\zeta}\max\Big(\sqrt{\big|J(\boldsymbol{\beta})\big|}, \sqrt{\big|J(\boldsymbol{\gamma})\big|}\Big)\right)\|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}), J(\boldsymbol{\gamma})}\|_1.$$

To achieve the proof of Theorem 4.3, we combine Equation (47) with Lemma 4.2 to conclude

$$\mathbb{P}\Big[\big(\mathcal{A}\cap\mathcal{B}\cap\boldsymbol{\Omega}_{\widetilde{\mathbf{RE}}_n(s,r_0)}(\tilde{\boldsymbol{\kappa}})\big)^c\Big] \le A_{\varepsilon,\nu}(x)\mathrm{e}^{-x} + B_{\tilde{\varepsilon},\tilde{\nu}}(y)\mathrm{e}^{-y} + \tilde{\pi}_n.$$

7.8. *Proof of Theorem 5.1*

The proofs of (23) and (24) are quite similar, so we only present the one of (23). To prove (24), it suffices to replace $\eta_{n,t}(f_j)$ by the process $\nu_{n,t}(\theta_k)$ throughout the following. Denote by $U_{n,t}$ and $H_i(f_j)$ the quantities

$$U_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^n\int_0^t H_i(f_j)\,\mathrm{d}M_i(s) \quad \text{and} \quad H_i(f_j) := \frac{f_j(\mathbf{Z}_i)}{\max_{1\le i\le n}|f_j(\mathbf{Z}_i)|}.$$

Since $H_i(f_j)$ is a bounded predictable process with respect to $\mathcal{F}_t$, $U_{n,t}(f_j)$ is a square integrable martingale. Its predictable variation is given by

$$\vartheta_{n,t}(f_j) = n\big\langle U_n(f_j)\big\rangle_t = \frac{1}{n}\sum_{i=1}^n\int_0^t \big(H_i(f_j)\big)^2\,\mathrm{d}\Lambda_i(s)$$

and the optional variation of $U_{n,t}(f_j)$ is

$$\hat{\vartheta}_{n,t}(f_j) = n\big[U_n(f_j)\big]_t = \frac{1}{n}\sum_{i=1}^n\int_0^t \big(H_i(f_j)\big)^2\,\mathrm{d}N_i(s).$$

We also define

$$\hat{\mathcal{W}}_n^\nu(f_j) = \frac{\nu/n}{\nu/n - \Phi(\nu/n)}\hat{\vartheta}_{n,t}(f_j) + \frac{x/n}{\nu/n - \Phi(\nu/n)}, \tag{54}$$

for $\nu \in (0,3)$ such that $\nu > \Phi(\nu)$ with $\Phi(u) = \mathrm{e}^u - u - 1$.

From Inequality (7.12) in Hansen et al. [21], for any $0 < v < \omega < +\infty$, we have

$$\mathbb{P}\left(U_{n,t}(f_j) \ge \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}} + \frac{x}{3n}, v \le \hat{\mathcal{W}}_n^\nu(f_j) \le \omega\right) \le 2\left(\frac{\log(\omega/v)}{\log(1+\varepsilon)} + 1\right)\mathrm{e}^{-x}. \tag{55}$$

We focus now on removing the event $\{v \le \hat{\mathcal{W}}_n^v(f_j) \le \omega\}$ in (55). Let us consider the martingale given $\mathcal{F}_t$

$$\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \big(H_i(f_j)\big)^2\big(\mathrm{d}N_i(s) - \mathrm{d}\Lambda_i(s)\big) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \big(H_i(f_j)\big)^2\,\mathrm{d}M_i(s),$$

and let

$$S_{v,t}(f_j) = \sum_{i=1}^{n}\int_0^t \Phi\left(\frac{v}{n}H_i^2(f_j)\right)\mathrm{d}\Lambda_i(s).$$

From van de Geer [37], we know that

$$\exp\big(v\big(\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)\big) - S_{v,t}(f_j)\big)$$

is a supermartingale. Now from Markov inequality, for any $v, x > 0$, we obtain that

$$\mathbb{P}\left[\big|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)\big| \ge \frac{S_{v,t}(f_j)}{v} + \frac{x}{n}\right] \le 2\mathrm{e}^{-x}. \tag{56}$$

For any $0 < h < 1$ and $x > 0$, $\Phi(xh) \le h^2\Phi(x)$. This combined with the fact that $0 < H_i^2(f_j) < 1$, we get

$$S_{v,t}(f_j) \le \Phi(v/n)\sum_{i=1}^{n}\int_0^t H_i^4(f_j)\,\mathrm{d}M_i(s) \le \Phi(v/n)n\vartheta_{n,t}(f_j). \tag{57}$$

Combining (56) and (57), we deduce that

$$\mathbb{P}\left[\big|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)\big| \ge \frac{\Phi(v/n)}{v/n}\vartheta_{n,t}(f_j) + \frac{x}{v}\right] \le 2\mathrm{e}^{-x}. \tag{58}$$

Now, under assumption (A_2), we have $\vartheta_{n,t}(f_j) \le A_0$, so the events

$$\Omega_n^v = \left\{\frac{x/n}{v/n - \Phi(v/n)} \le \hat{\mathcal{W}}_n^v(f_j)\right\} \cap \big\{\vartheta_{n,t}(f_j) \le A_0\big\}$$

is of probability one and thus

$$\mathbb{P}\left(U_{n,t}(f_j) \ge \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^v(f_j)x}{n}} + \frac{x}{3n}\right) \le \mathbb{P}\left(\left\{U_{n,t}(f_j) \ge \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^v(f_j)x}{n}} + \frac{x}{3n}\right\} \cap \Omega_n^v\right). \tag{59}$$

From (58), we have

$$\mathbb{P}\left[\hat{\vartheta}_{n,t}(f_j) \ge \vartheta_{n,t}(f_j)\left(1 + \frac{\Phi(v/n)}{v/n}\right) + \frac{x}{v}\right] \le \mathrm{e}^{-x},$$

and if we denote E_n^v the event

$$E_n^v = \left\{\hat{\vartheta}_{n,t}(f_j) \le \vartheta_{n,t}(f_j)\left(1 + \frac{\Phi(v/n)}{v/n}\right) + \frac{x}{v}\right\},$$

we get

$$\begin{aligned}&\mathbb{P}\left[U_{n,t}(f_j) \ge \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^v(f_j)x}{n}} + \frac{x}{3n}\right]\\ &\quad\le \mathrm{e}^{-x} + \mathbb{P}\left[\left\{U_{n,t}(f_j) \ge \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^v(f_j)x}{n}} + \frac{x}{3n}\right\} \cap \Omega_n^v \cap E_n^v\right].\end{aligned}$$

On the event $E_n^v \cap \Omega_n^v$, from the definition of $\hat{\mathcal{W}}_n^v(f_j)$ given by (54), we have

$$\begin{aligned}\hat{\mathcal{W}}_n^v(f_j) &\le \frac{v/n}{v/n-\Phi(v/n)}\left(\vartheta_{n,t}(f_j)\left(1+\frac{\Phi(v/n)}{v/n}\right)+\frac{x}{v}\right)+\frac{x/n}{v/n-\Phi(v/n)}\\ &\le A_0\frac{v/n+\Phi(v/n)}{v/n-\Phi(v/n)}+2\frac{x/n}{v/n-\Phi(v/n)}. \end{aligned} \tag{60}$$

From (60), we obtain

$$\begin{aligned}&\mathbb{P}\left[\left\{U_{n,t}(f_j)\ge\sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^v(f_j)x}{n}}+\frac{x}{3n}\right\}\cap\Omega_n^v\cap E_n^v\right]\\ &\quad\le\mathbb{P}\left[U_{n,t}(f_j)\ge\sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^v(f_j)x}{n}}+\frac{x}{3n},\right.\\ &\qquad\left.\frac{x/n}{v/n-\Phi(v/n)}\le\hat{\mathcal{W}}_n^v(f_j)\le A_0\frac{v/n+\Phi(v/n)}{v/n-\Phi(v/n)}+2\frac{x/n}{v/n-\Phi(v/n)}\right].\end{aligned}$$

We now apply Inequality (55) with $v=\frac{x/n}{v/n-\Phi(v/n)}$ and $\omega=A_0\frac{v/n+\Phi(v/n)}{v/n-\Phi(v/n)}+2\frac{x/n}{v/n-\Phi(v/n)}$,

$$\begin{aligned}&\mathbb{P}\left[U_{n,t}(f_j)\ge\sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^v(f_j)x}{n}}+\frac{x}{3n}\right]\\ &\quad\le\left(\frac{2}{\log(1+\varepsilon)}\log\left(2+\frac{A_0(v/n+\Phi(v/n))}{x/n}\right)+1\right)\mathrm{e}^{-x}.\end{aligned} \tag{61}$$

Now it suffices to multiply both sides of the inequality inside the probability by $\|f_j\|_{n,\infty}=\max_{1\le i\le n}|f_j(\mathbf{Z}_i)|$ to end up the proof of Theorem 5.1.

Appendix A: Proof of Proposition 6.4

The proofs of Propositions 6.2 and 6.4 are similar. So we only prove Proposition 6.4 which corresponds to the general case. To compare the empirical Kullback divergence (3) and the weighted empirical norm (4), we use Lemma 1 in Bach [4], that we recall here:

Lemma A.1. *Let g be a convex three times differentiable function $g:\mathbb{R}\to\mathbb{R}$ such that for all $t\in\mathbb{R}$, $|g'''(t)|\le Sg''(t)$, for some $S\ge 0$. Then, for all $t\ge 0$:*

$$\frac{g''(0)}{S^2}\phi(St)\le g(t)-g(0)-g'(0)t\le\frac{g''(0)}{S^2}\phi(-St)\quad\text{with }\phi(u)=\mathrm{e}^{-u}+u-1.$$

This lemma gives upper and lower Taylor expansions for some convex and three times differentiable function. It has been introduced to extend tools from self-concordant functions (i.e., which verify $|g'''(t)|\le 2g''(t)^{3/2}$) and provide simple extensions of theoretical results for the square loss for logistic regression.

Let h be a function on $[0,\tau]\times\mathbb{R}^p$ and define

$$G(h)=-\frac{1}{n}\sum_{i=1}^n\int_0^\tau h(s,\mathbf{Z}_i)\,\mathrm{d}\Lambda_i(s)+\frac{1}{n}\sum_{i=1}^n\int_0^\tau\mathrm{e}^{h(s,\mathbf{Z}_i)}Y_i(s)\,\mathrm{d}s.$$

Consider the function $g:\mathbb{R}\to\mathbb{R}$ defined by $g(t)=G(h+tk)$, where h and k are two functions defined on $\mathbb{R}^p$. By differentiating G with respect to t we get:

$$g'(t)=-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau k(s,\mathbf{Z}_i)\,\mathrm{d}\Lambda_i(s)+\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau k(s,\mathbf{Z}_i)\mathrm{e}^{h(s,\mathbf{Z}_i)+tk(s,\mathbf{Z}_i)}Y_i(s)\,\mathrm{d}s,$$

$$g''(t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(k(s,\mathbf{Z}_i)\big)^2\mathrm{e}^{h(s,\mathbf{Z}_i)+tk(s,\mathbf{Z}_i)}Y_i(s)\,\mathrm{d}s,$$

$$g'''(t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(k(s,\mathbf{Z}_i)\big)^3\mathrm{e}^{h(s,\mathbf{Z}_i)+tk(s,\mathbf{Z}_i)}Y_i(s)\,\mathrm{d}s.$$

It follows that

$$\big|g'''(t)\big|\le\|k\|_{n,\infty}g''(t).$$

Applying Lemma A.1 with $S=\|k\|_{n,\infty}$, we obtain for all $t\ge 0$,

$$\frac{g''(0)}{\|k\|_{n,\infty}^2}\phi\big(t\|k\|_{n,\infty}\big)\le g(t)-g(0)-g'(0)t\le\frac{g''(0)}{\|k\|_{n,\infty}^2}\phi\big(-t\|k\|_{n,\infty}\big).$$

Take $t=1$, $h(s,\mathbf{Z}_i)=\log\lambda_0(s,\mathbf{Z}_i)$ and for $(\boldsymbol{\beta},\boldsymbol{\gamma})\in\widetilde{\Gamma}(\rho)$, $k(s,\mathbf{Z}_i)=\log\lambda_{\beta,\gamma}(s,\mathbf{Z}_i)-\log\lambda_0(s,\mathbf{Z}_i)$. We obtain

$$\begin{aligned}g''(0)\frac{\phi(\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty})}{\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty}^2}&\le G(\log\lambda_{\beta,\gamma})-G(\log\lambda_0)-g'(0)\\&\le g''(0)\frac{\phi(-\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty})}{\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty}^2}.\end{aligned}\tag{A.1}$$

Now straightforward calculations show that $g'(0)=0$ and $g''(0)=\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\Lambda}^2$. Replacing $g'(0)$ and $g''(0)$ by their expressions in (A.1) and noting that $G(\log\lambda_{\beta,\gamma})-G(\log\lambda_0)=\widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma})$, we get

$$\frac{\phi(\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty})}{\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty}^2}\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\Lambda}^2\le\widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma})$$

and

$$\widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma})\le\frac{\phi(-\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty})}{\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty}^2}\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\Lambda}^2.$$

According to assumption (A_5) for $(\boldsymbol{\beta},\boldsymbol{\gamma})\in\widetilde{\Gamma}(\rho)$, $\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty}\le\rho$. Since $\phi(t)/t^2$ is decreasing and bounded below by 0, we can deduce that

$$\frac{\phi(\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty})}{\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty}^2}\ge\frac{\phi(\rho)}{\rho^2}$$

and

$$\frac{\phi(-\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty})}{\|\log\lambda_{\beta,\gamma}-\log\lambda_0\|_{n,\infty}^2}\le\frac{\phi(-\rho)}{\rho^2}.$$

Take $\rho':=\phi(\rho)/\rho^2>0$ and $\rho'':=\phi(-\rho)/\rho^2>0$ to finish the proof.

Appendix B: Proof of Proposition 6.3

Similarly to the proof of Proposition 6.4, under assumption (A_6), when considering

$$G(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\bigl(\alpha_0(s)\mathrm{e}^{\boldsymbol{\beta}^T\mathbf{Z}_i}\bigr)\,\mathrm{d}\Lambda_i(s) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha_0(s)\mathrm{e}^{\boldsymbol{\beta}^T\mathbf{Z}_i}Y_i(s)\,\mathrm{d}s \quad \text{and} \quad g(t) = G(\boldsymbol{\beta}+t\boldsymbol{\eta}),$$

we obtain

$$\bigl\|(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0)^T\mathbf{X}\bigr\|_{n,\Lambda}^2\frac{\phi(R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2)}{R^2\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2^2} \le \widetilde{K}_n(\lambda_0,\lambda_{\hat{\boldsymbol{\beta}}_L}) \le \bigl\|(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0)^T\mathbf{X}\bigr\|_{n,\Lambda}^2\frac{\phi(-R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2)}{R^2\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2^2}. \tag{B.1}$$

Now, we will show that $R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2$ is bounded. From Equation (37) with $\hat{\boldsymbol{\beta}}_L^{\mu}=\hat{\boldsymbol{\beta}}_L$ and $\boldsymbol{\beta}=\boldsymbol{\beta}_0$, we can deduce that

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\boldsymbol{\beta}}_L}) \le \frac{3}{2}\Gamma_1\|\boldsymbol{\Delta}_0\|_1,$$

where $\boldsymbol{\Delta}_0=\mathbf{D}(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0)$ and $\mathbf{D}=(\mathrm{diag}(\omega_j))_{1\le j\le M}$. From (B.1), we have

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\boldsymbol{\beta}}_L}) \ge \frac{\|(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0)^T\mathbf{X}\|_{n,\Lambda}^2}{R^2\|(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0)\|_2^2}\phi\bigl(R\bigl\|(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0)\bigr\|_2\bigr).$$

We apply assumption (RE(s, a_0)) with $a_0=3$ and $\boldsymbol{\kappa}'=\boldsymbol{\kappa}'(s,3)$ and we infer that

$$\boldsymbol{\kappa}'^2\|\boldsymbol{\Delta}_{0,J_0}\|_2^2 \le \bigl\|\boldsymbol{\Delta}_0^T\mathbf{X}\bigr\|_{n,\Lambda}^2.$$

So we have,

$$\frac{\boldsymbol{\kappa}'^2\|\boldsymbol{\Delta}_{0,J_0}\|_2^2}{\max_{1\le j\le M}\omega_j^2}\frac{\phi(R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2)}{R^2\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2^2} \le \frac{3}{2}\Gamma_1\|\boldsymbol{\Delta}_0\|_1.$$

We can now use, with $s=|J_0|$, $\|\boldsymbol{\Delta}_0\|_2\le\|\boldsymbol{\Delta}_0\|_1\le 4\|\boldsymbol{\Delta}_{0,J_0}\|_1\le 4\sqrt{s}\|\boldsymbol{\Delta}_{0,J_0}\|_2$ to get

$$\begin{aligned}\boldsymbol{\kappa}'^2\phi\bigl(R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2\bigr) &\le \frac{3}{2}\Gamma_1\frac{\max_{1\le j\le M}\omega_j^2}{\min_{1\le j\le M}\omega_j^2}\max_{1\le j\le M}\omega_j\frac{(4\sqrt{s}\|(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0)_{J_0}\|_2)^2R^2\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2}{\|(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0)_{J_0}\|_2^2}\\ &\le 24\Gamma_1\frac{\max_{1\le j\le M}\omega_j^2}{\min_{1\le j\le M}\omega_j^2}\max_{1\le j\le M}\omega_j s R^2\|\boldsymbol{\Delta}_0\|_2.\end{aligned}$$

A short calculation shows that for all $k\in(0,1]$:

$$\mathrm{e}^{-2k(1-k)^{-1}}+(1-k)2k(1-k)^{-1}-1\ge 0,$$

(see Bach [4] for more details). So by taking $2k(1-k)^{-1}=R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2$, we have

$$\mathrm{e}^{-R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2}+R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2-1\ge\frac{R^2\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2^2}{2+R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2}$$

and we deduce that

$$\frac{\boldsymbol{\kappa}'^2R^2\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2^2}{2+R\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2} \le 24\Gamma_1\frac{\max_{1\le j\le M}\omega_j^2}{\min_{1\le j\le M}\omega_j^2}\max_{1\le j\le M}\omega_j s R^2\|\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_0\|_2.$$

This implies that

$$R\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_2 \leq \frac{(48\Gamma_1 Rs)/\boldsymbol{\kappa}'^2(\max_{1\leq j\leq M}\omega_j^2/\min_{1\leq j\leq M}\omega_j^2)\max_{1\leq j\leq M}\omega_j^2}{1-(24\Gamma_1 Rs)/\boldsymbol{\kappa}'^2(\max_{1\leq j\leq M}\omega_j^2/\min_{1\leq j\leq M}\omega_j^2)\max_{1\leq j\leq M}\omega_j^2} \leq 2$$

$$\text{as soon a } \Gamma_1 \leq \frac{1}{48Rs}\frac{\min_{1\leq j\leq M}\omega_j^2}{\max_{1\leq j\leq M}\omega_j^2}\frac{\boldsymbol{\kappa}'^2}{\max_{1\leq j\leq M}\omega_j}.$$

Since $\phi(t)/t^2$ is decreasing and bounded below by 0, we can deduce that

$$\frac{\phi(R\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_2)}{R^2\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_2^2} \geq \frac{\phi(2)}{4}$$

and

$$\frac{\phi(-R\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_2)}{R^2\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_2^2} \leq \frac{\phi(-2)}{4}.$$

Take $\xi := \phi(2)/4 > 0$ and $\xi' := \phi(-2)/4 > 0$ and conclude that

$$\xi\left\|(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0)^T\mathbf{X}\right\|_{n,\Lambda}^2 \leq \widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_L}) \leq \xi'\left\|(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0)^T\mathbf{X}\right\|_{n,\Lambda}^2.$$

Acknowledgements

My first thanks go to my two Ph.D. thesis supervisors Agathe Guilloux and Marie-Luce Taupin for their help, their availability and their advices. I also thank Marius Kwemou for helpful discussions. I am very grateful to Patricia Reynaud-Boured for her helpful advises and suggestions. Finally, I gratefully acknowledge the referees for carefully reading the manuscript and for numerous suggestions that improved the paper.

References

[1] O. Aalen. A model for nonparametric regression analysis of counting processes. In *Mathematical Statistics and Probability Theory (Proc. Sixth Internat. Conf., Wisła, 1978)* 1–25. *Lecture Notes in Statist.* **2**. Springer, New York, 1980. MR0577267

[2] P. K. Andersen, O. Borgan and R. D. Gill. *Statistical Models Based on Counting Processes*. *Springer Series in Statistics*. Springer, New York, 1993. MR1198884

[3] A. Antoniadis, P. Fryzlewicz and F. Letué. The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **37** (2010) 531–552. MR2779635

[4] F. Bach. Self-concordant analysis for logistic regression. *Electron. J. Stat.* **4** (2010) 384–414. MR2645490

[5] P. L. Bartlett, S. Mendelson and J. Neeman. ℓ_1-regularized linear regression: Persistence and oracle inequalities. *Probab. Theory Related Fields* **154** (2012) 193–224. MR2981422

[6] K. Bertin, E. Le Pennec and V. Rivoirard. Adaptive Dantzig density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.* **47** (2011) 43–74. MR2779396

[7] P. J. Bickel, Y. Ritov and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** (2009) 1705–1732. MR2533469

[8] J. Bradic, J. Fan and J. Jiang. Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39** (2012) 3092–3120. MR3012402

[9] J. Bradic and R. Song. Structured estimation for the nonparametric Cox model. *Electron. J. Stat.* **9** (2015) 492–534. MR3326133

[10] P. Bühlmann and S. van de Geer. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** (2009) 1360–1392. MR2576316

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

[12] F. Bunea, A. B. Tsybakov and M. H. Wegkamp. Aggregation and sparsity via l_1 penalized least squares. In *Learning Theory* 379–391. *Lecture Notes in Comput. Sci.* **4005**. Springer, Berlin, 2006. MR2280619

[13] F. Bunea, A. B. Tsybakov, M. H. Wegkamp and A. Barbu. Spades and mixture models. *Ann. Statist.* **38** (2010) 2525–2558. MR2676897

[14] F. Comte, S. Gaïffas and A. Guilloux. Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **47** (2011) 1171–1196. MR2884230

[15] D. R. Cox. Regression models and life-tables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34** (1972) 187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox. MR0341758
[16] S. S. Dave, G. Wright, B. Tan, A. Rosenwald, R. D. Gascoyne, W. C. Chan, R. I. Fisher, R. M. Braziel, L. M. Rimsza, T. M. Grogan, T. P. Miller, M. LeBlanc, T. C. Greiner, D. D. Weisenburger, J. C. Lynch, J. Vose, J. O. Armitage, E. B. Smeland, S. Kvaloy, H. Holte, J. Delabie, J. M. Connors, P. M. Lansdorp, Q. Ouyang, T. A. Lister, A. J. Davies, A. J. Norton, H. K. Muller-Hermelink, G. Ott, E. Campo, E. Montserrat, W. H. Wilson, E. S. Jaffe, R. Simon, L. Yang, J. Powell, H. Zhao, N. Goldschmidt, M. Chiorazzi and L. M. Staudt. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.* **351** (2004) 2159–2169.
[17] J. Fan and R. Li. Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** (2002) 74–99. MR1892656
[18] S. Gaïffas and A. Guilloux. High-dimensional additive hazard models and the Lasso. *Electron. J. Stat.* **6** (2011) 522–546. MR2988418
[19] R. Gill. Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.* **11** (1983) 49–58. MR0684862
[20] M. L. Gourlay, J. P. Fine, J. S. Preisser, R. C. May, C. Li, L. Y. Lui, D. F. Ransohoff, J. A. Cauley and K. E. Ensrud. Bone-density testing interval and transition to osteoporosis in older women. *N. Engl. J. Med.* **366** (2012) 225–233.
[21] N. R. Hansen, P. Reynaud-Bouret and V. Rivoirard. Lasso and probabilistic inequalities for the multivariate point processes. *Bernoulli* **21** (2015) 83–143. MR3322314
[22] M. J. Kearns, R. E. Schapire and L. M. Sellie. Toward efficient agnostic learning. *Mach. Learn.* **17** (1994) 115–141.
[23] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Ecole d'Eté de Probabilités de Saint-Flour XXXVIII*. *Lecure Notes in Mathematics* **2033**. Springer, Heidelberg, 2011. MR2829871
[24] S. Kong and B. Nan. Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Statist. Sinica* **1** (2014) 25–42. MR3184591
[25] E. Le Pennec and S. X. Cohen. Partition-based conditional density estimation. *ESAIM Probab. Stat.* **1** (2013) 672–697. MR3126157
[26] S. Lemler Oracle inequalities for the Lasso in the high-dimensional multiplicative Aalen intensity model. Preprint, 2012. Available at arXiv:1206.5628.
[27] F. Letué. Modèle de Cox: Estimation par sélection de modele et modèle de chocs bivarié. Ph.D. thesis, 2000.
[28] T. Martinussen and T. H. Scheike. Covariate selection for the semiparametric additive risk model. *Scand. J. Stat.* **36** (2009) 602–619. MR2572578
[29] P. Massart. *Concentration Inequalities and Model Selection*. *Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003*. *Lecture Notes in Mathematics* **1896**. Springer, Berlin, 2007. With a foreword by Jean Picard. MR2319879
[30] P. Massart and C. Meynet. The Lasso as an l_1-ball model selection procedure. *Electron. J. Stat.* **5** (2011) 669–687. MR2820635
[31] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** (2006) 1436–1462. MR2278363
[32] R. Senoussi. Problème d'identification dans le modèle de Cox. *Ann. Inst. Henri Poincaré Probab. Stat.* **26** (1988) 45–64. MR1075438
[33] E. W. Steyerberg, M. Y. V. Homs, A. Stokvis, M. L. Essink-Bot and P. D. Siersema. Stent placement or brachytherapy for palliation of dysphagia from esophageal cancer: A prognostic model to guide treatment selection. *Gastroint. Endosc.* **62** (2005) 333–340.
[34] C. J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** (1994) 118–184. With discussion by Andreas Buja and Trevor Hastie and a rejoinder by the author. MR1272079
[35] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** (1996) 267–288. MR1379242
[36] R. Tibshirani. The Lasso method for variable selection in the Cox model. *Stat. Med.* **16** (1997) 385–395.
[37] S. van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.* **23** (1995) 1779–1801. MR1370307
[38] S. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** (2008) 614–645. MR2396809
[39] C. H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** (2008) 1567–1594. MR2435448
[40] H. H. Zhang and W. Lu. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** (2007) 691–703. MR2410017
[41] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** (2010) 1081–1107. MR2629825
[42] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** (2007) 2541. MR2274449
[43] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** (2006) 1418–1429. MR2279469
[44] H. Zou. A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* **95** (2008) 241–247. MR2409726