Oracle inequalities for the Lasso for the conditional hazard rate in a high-dimensional setting

Sarah Lemler

Oracle inequalities for the Lasso for the conditional hazard rate in a high-dimensional setting

Sarah Lemler

Laboratoire Statistique et Génome UMR CNRS 8071- USC INRA,

Universite d'Evry Val d'Essonne, France

e-mail : sarah.lemler@genopole.cnrs.fr

Abstract

We aim at obtaining a prognostic on the survival time adjusted on covariates in a high-dimensional setting. Towards this end, we consider a conditional hazard rate function that does not rely on an underlying model and we estimate it by the best Cox's 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. Since we are in high-dimension, we consider the Lasso procedure to estimate the unknown parameters of the best Cox's model approximating the conditional hazard rate function. We provide non-asymptotic oracle inequalities for the Lasso estimator of the conditional hazard risk function. Our results are mainly based on an empirical Bernstein's inequalities for martingales with jumps.

Keywords: Survival analysis; Conditional hazard rate function; Cox's proportional hazards model; Right-censored data; Semi-parametric model; Nonparametric model; High-dimensional covariates; Lasso; Non-asymptotic oracle inequalities; Empirical Bernstein inequality

1 Introduction

We consider the problem of determining the prognostic factors among a large number of covariates for the survival time. For example, in Dave et al. [13], 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, the goal is to predict the survival from follicular lymphoma adjusted on the covariates. To adjust on covariates, the most popular semi-parametric regression model is the Cox's proportional hazards model (see Cox [12]) : the conditional hazard rate function of the survival time T given a the vector of covariates $\boldsymbol{Z} = (Z_1, ..., Z_p)^T$ is defined by

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) \exp(\boldsymbol{\beta_0^T Z}), \tag{1}$$

where $\boldsymbol{\beta_0} = (\beta_{0_1}, ..., \beta_{0_p})^T$ is the vector of regression coefficients and $\alpha_0(t)$ is the baseline hazard function. The unknown parameters of the model are $\boldsymbol{\beta_0} \in \mathbb{R}^p$ and the functional parameter α_0. If one is only interested in assessing the effects of the covariates, one would consider the Cox's partial log-likelihood. Cox [12] has introduced this partial log-likelihood to estimate $\boldsymbol{\beta_0}$ without having to know α_0, when the number of covariates p is less than the sample size n ($p < n$). Our objective in this paper is different : we are interested here in obtaining a prognostic on the survival time adjusted on the covariates (see Gourlay [15] and Steyerberg [23]). As a consequence, we want to estimate the complete conditional hazard rate function λ_0 and thus, we will rather consider the total log-likelihood.

An estimator typically used in a high-dimensional setting is the Lasso estimator. It has been introduced by Tibshirani [24] and widely studied since then with consistence results (see Meinshausen and Bühlmann [20]), variable selection (see Zhao and Yu [32], Zhang and Huang [29]) and estimation results (see Bunea et al. [9, 10]). The Lasso has been mainly studied in a high-dimensional additive regression model of the form

$$Y = f(\boldsymbol{Z}) + \varpi, \tag{2}$$

where f is the unknown regression function and ϖ a noise term. In this model, in order to estimate f with the Lasso estimator, a classical method consists in introducing a dictionary $\mathbb{F}_M = \{f_1, ..., f_M\}$ and assuming that f is well estimated by a linear combination of the form

$$f_\beta = \sum_{j=1}^{M} \beta_j f_j, \text{ with } \boldsymbol{\beta} \text{ in } \mathbb{R}^M.$$

The parameter $\boldsymbol{\beta}$ has to be estimated with the Lasso procedure by minimizing an ℓ_1-penalized criterion. In the case of the additive regression model (2), the Lasso estimator of f is obtained by minimizing the ℓ_1-penalized least squares criterion

$$f_{\hat{\beta}_L} = \sum_{j=1}^{M} \hat{\beta}_{L,j} f_j \text{ with } \hat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta} \in \mathbb{R}^M}{\arg\min} \{||Y - f_\beta(\boldsymbol{Z})||_n^2 + \Gamma ||\boldsymbol{\beta}||_1\},$$

where Γ is a tuning parameter and $||.||_n$ the usual empirical quadratic norm. In this paper, we are interested in non-asymptotic oracle inequalities for the Lasso that allow to compare the performances of an estimator $f_{\hat{\beta}_L}$ obtained without a priori knowledge of the true function f, to those of the best approximation f_β of f in the dictionary for all n. Our aim is to state an oracle inequality of the form

$$||f_{\hat{\beta}_L} - f||^2 \leq (1 + \zeta) \inf_{\boldsymbol{\beta} \in \mathbb{R}^M} \{||f_\beta - f||^2 + T_{\zeta,n,M}\},$$

with $\zeta \geq 0$. The quantity $T_{\zeta,n,M}$ is a variance term of order $\sqrt{\log M/n}$ or $\log M/n$ depending on whether the rate of convergence of the estimator to the true function is slow or fast respectively. Several contributions deal with establishing non-asymptotic oracle inequalities for the Lasso in an additive regression model (see Bunea et al. [9], Bickel et al. [6], Massart and Meynet [19] among

others). In this setting and under "the restricted eigenvalues assumption", Bickel et al. [6] have stated a fast non-asymptotic oracle inequality. They provide prediction results, i.e. on $f_{\hat{\beta}_L} - f$, and estimation inequalities, i.e. on $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}$, in the linear case ($f(\boldsymbol{Z}) = \boldsymbol{\beta_0^T Z}$). Massart and Meynet [19] have also obtained non-asymptotic oracle inequalities for the Lasso in an additive regression model via the application of a single general theorem of model selection among a collection of nonlinear models.

In the setting of survival analysis, the Lasso procedure has been first considered by Tibshirani [25]. Nevertheless, few results exist on the Lasso estimator in the Cox's model. Antoniadis et al. [3] have established asymptotic estimation inequalities in the Cox's proportional hazard model for the Dantzig estimator, which is similar to the Lasso estimator (see Bickel et al. [6] for a comparison between these two estimators). In Bradic et al. [7], asymptotic estimation inequalities on $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}$ for the Lasso estimator have been also obtained in the Cox's model. However, in practice, one can not consider that the asymptotic regime has been reached : in Dave et al. [13], for example, the expression levels of 44 929 genes and survival information are measured for only 191 patients. Some non-asymptotic results have also been established in survival analysis. Gaïffas and Guilloux [14] have proved non-asymptotic oracle inequalities for an additive hazards model. Recently, Kong and Nan [17] have established a non-asymptotic oracle inequality for the Lasso in the Cox's model

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) e^{f_{\beta_0}(\boldsymbol{Z})} \text{ with } f_{\beta_0} = \sum_{j=1}^{M} \beta_{0,j} f_j.$$

However, since Kong and Nan [17] have used the Cox's partial log-likelihood to estimate $\boldsymbol{\beta_0}$, the obtained results are on $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}$ and on $f_{\hat{\beta}_L} - f_{\beta_0}$ and the problem of estimating the whole conditional hazard rate function λ_0 is not considered, as needed for the prevision of the survival time.

There are two main motivations in the present paper. First we address the problem of estimating λ_0 regardless of an underlying model : we opt for an agnostic approach, see Rigollet [21]. Secondly we aim at obtaining non-asymptotic oracle inequalities. As we provide oracle inequalities for the conditional hazard rate function, we reach the announced goal of obtaining a prognostic for the survival adjusted on the covariates.

More precisely, we consider two finite families of functions $\mathbb{F}_M = \{f_1, ..., f_M\}$ with $f_j : \mathbb{R}^p \to \mathbb{R}$, $j = 1, ..., M$ and $\mathbb{G}_N = \{\theta_1, ..., \theta_N\}$ with $\theta_k : \mathbb{R}_+ \to \mathbb{R}_+$, $k = 1, ..., N$, called dictionaries. We aim at estimating λ_0 by the best approximating Cox's model constructed with functions of the dictionaries : λ_0 will be estimate by a function of the form

$$\lambda_{\beta,\gamma}(t, \boldsymbol{Z_i}) = \alpha_\gamma(t) e^{f_\beta(\boldsymbol{Z_i})} \text{ for } (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N \tag{3}$$

with

$$\log \alpha_\gamma = \sum_{k=1}^{N} \gamma_k \theta_k \text{ and } f_\beta = \sum_{j=1}^{M} \beta_j f_j.$$

Our goal is not to estimate the parameters of an underlying 'true' model but rather to construct an estimator that mimics the performance of the best Cox's model, whether this model is true or not.

We propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ simultaneously with the Lasso method using the full log-likelihood, with a weighted ℓ_1-penalization for each parameter.

Towards this end, we will proceed in two steps. First we start with assuming that λ_0 verifies $\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t)e^{f_0(\boldsymbol{Z_i})}$, where α_0 is a known baseline function. If we take $f_0(\boldsymbol{Z_i}) = \boldsymbol{\beta_0^T Z_i}$, we obtain the Cox's model. In this particular case, the only nonparametric 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 the first non-asymptotic oracle inequalities for the Cox's model when α_0 is supposed to be known.

In a second time, we consider the general problem of estimating the whole conditional hazard rate function λ_0. We state non-asymptotic oracle inequalities in terms of both empirical Kullback divergence and weighted empirical norm for our Lasso estimators. These results are obtained from an empirical Bernstein inequality. The empirical processes to be controlled involve martingales with jumps, whose predictable variation are not observable. We establish an empirical version of Bernstein involving the optional variation, which is observable. This allows us to define a fully data-driven weighted ℓ_1-penalization.

The paper is organized as follows. In Section 2, we describe the framework and the Lasso procedure for estimating the conditional hazard rate function. We also present in this section the estimation risk we choose to work with and its associated loss function. In Section 3, the first oracle inequalities are obtained in the particular Cox's model with known baseline hazard function. In this section, we give some prediction and estimation inequalities. In Section 4, non-asymptotic oracle inequalities with different convergence rates are given for a general conditional hazard rate function. Section 5 is devoted to the empirical Bernstein's inequality associated to our processes. Proofs are gathered in section 6.

2 Framework and estimation procedure

2.1 Framework

We present our procedure and establish the oracle inequalities in the general setting of counting processes. Towards that end, for $i = 1, ..., n$, let N_i be a marked counting process and Y_i a predictable random process 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

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), 0 \leq s \leq t, Z_i, i = 1, ..., n\}.$$

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.

Assumption 1. *N_i satisfies the Aalen multiplicative intensity model : for all $t \geq 0$,*

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \boldsymbol{Z_i})Y_i(s)ds, \tag{4}$$

where λ_0 is an unknown nonnegative function called intensity and $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ is the $\mathcal{F}_0$-measurable random vector of covariates of individual i.

This general setting, introduced by Aalen, see 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. *In the specific case of right censoring, let $(T_i)_{i=1,...,n}$ be i.i.d. survival times of n individuals and $(C_i)_{i=1,...,n}$ their i.i.d. censoring times. We observe $\{(X_i, \boldsymbol{Z_i}, \delta_i)\}_{i=1,...,n}$ where $X_i = \min(T_i, C_i)$ is the event time, $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T$ 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 the vector of covariates $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ for $i = 1, ..., n$. With these notations, the $\mathcal{F}_t$-adapted processes $Y_i(t)$ and $N_i(t)$ 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.*

The estimation procedure is based on the independent and identically distributed (i.i.d.) data $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$, where $[0, \tau]$ is the time interval between the beginning and the end of the study.

2.2 The estimation criterion and the loss function

Let $\mathbb{F}_M = \{f_1, ..., f_M\}$ where $f_j : \mathbb{R}^p \to \mathbb{R}$ for $j = 1, ..., M$, and $\mathbb{G}_N = \{\theta_1, ..., \theta_N\}$ where $\theta_k : \mathbb{R}_+^* \to \mathbb{R}$ for $k = 1, ..., N$, be two finite sets of functions, called dictionaries, where M and N are large (typically $M >> n$ and $N >> n$). The sets $\mathbb{F}_M$ and $\mathbb{G}_N$ can be collections of basis functions such as wavelets, splines, step functions, etc. They can also be collections of several estimators computed using different tuning parameters. We assume that the unknown λ_0 can be well approximated by a function defined for all $\boldsymbol{\beta}$ in $\mathbb{R}^M$ and $\boldsymbol{\gamma}$ in $\mathbb{R}^N$ by

$$\lambda_{\beta,\gamma}(t, \boldsymbol{Z_i}) = \alpha_\gamma(t) e^{f_\beta(\boldsymbol{Z_i})} \tag{5}$$

where

$$\log \alpha_\gamma = \sum_{k=1}^{N} \gamma_k \theta_k \text{ and } f_\beta = \sum_{j=1}^{M} \beta_j f_j.$$

By Jacod Formula (see Andersen et al. [2]), the log-likelihood based on the data $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$ is given by

$$C_n(\lambda_{\beta,\gamma}) = -\frac{1}{n} \sum_{i=1}^{n} \left\{ \int_0^\tau \log(\lambda_{\beta,\gamma}(t, \boldsymbol{Z_i})) dN_i(t) - \int_0^\tau \lambda_{\beta,\gamma}(t, \boldsymbol{Z_i}) Y_i(t) dt \right\}. \tag{6}$$

We propose an estimation procedure based on the minimization of this empirical risk. To this estimation criterion, we associate the empirical Kullback divergence defined for all $\boldsymbol{\beta}$ in $\mathbb{R}^M$ and $\boldsymbol{\gamma}$ in $\mathbb{R}^N$ by

$$\begin{aligned} \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) = & \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\log \left(\lambda_0(t, \boldsymbol{Z_i}) \right) - \log \left(\lambda_{\beta,\gamma}(t, \boldsymbol{Z_i}) \right) \right) \lambda_0(t, \boldsymbol{Z_i}) Y_i(t) dt \\ & - \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\lambda_0(t, \boldsymbol{Z_i}) - \lambda_{\beta,\gamma}(t, \boldsymbol{Z_i}) \right) Y_i(t) dt. \end{aligned} \tag{7}$$

We refer to van de Geer [27] and Senoussi [22] for close definitions.

Remark 2. *The loss function $\widetilde{K}_n$ is similar to the classical Kullback-Leibler information for density. Indeed, the term*

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\frac{\log(\lambda_0(t,\boldsymbol{Z_i}))}{\log(\lambda_{\beta,\gamma}(t,\boldsymbol{Z_i}))}\right)\lambda_0(t,\boldsymbol{Z_i})Y_i(t)dt,$$

would correspond to the usual Kullback for density if $\lambda_0(t,\boldsymbol{Z_i})$ and $\lambda_{\beta,\gamma}(t,\boldsymbol{Z_i})$ were densities. However, since the hazard rate function is not a density, we have a residual term in our Kullback divergence defined by the second term of $\widetilde{K}_n$.

Proposition 1. *The empirical Kullback divergence $\widetilde{K}_n(\lambda_0,\lambda_{\beta,\gamma})$ is nonnegative and equals zero if and only if $\lambda_{\beta,\gamma}=\lambda_0$ almost surely.*

We also introduce :

- the weighted empirical norms defined for all function h on $[0,\tau]\times\mathbb{R}^p$ by

$$||h||_{n,\Lambda}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (h(t,\boldsymbol{Z_i}))^2 d\Lambda_i(t)}, \tag{8}$$

  where Λ_i is defined in (4). 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.

- the empirical sup-norm $||.||_{n,\infty}$ for any h function on $[0,\tau]\times\mathbb{R}^p$

$$||h||_{n,\infty}=\max_{\substack{1\le i\le n\\ t\in[0,\tau]}}|h(t,\boldsymbol{Z_i})|.$$

We assume that dictionaries $\mathbb{F}_M$ and $\mathbb{G}_N$ are chosen such that the two following assumptions are fullfiled.

Assumption 2. *For all j in $\{1,...,M\}$,*

$$||f_j||_{n,\infty}=\max_{1\le i\le n}|f_j(Z_i)|<\infty \tag{9}$$

Assumption 3. *For all k in $\{1,...,N\}$,*

$$||\theta_k||_\infty=\max_{t\in[0,\tau]}|\theta_k(t)|<\infty. \tag{10}$$

To connect the empirical Kullback divergence (7) and the weighted empirical norm (8), we introduce the following assumption :

Assumption 4. *There exists $\mu>0$ a numerical positive constant such that, for all $\boldsymbol{\beta}$ in $\mathbb{R}^M$ and $\boldsymbol{\gamma}$ in $\mathbb{R}^N$*

$$||\log\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}-\log\lambda_0||_{n,\Lambda}\le\mu \tag{11}$$

This assumption is classic (see e.g. van de Geer [28] or Kong and Nan [17]). This assumption means that the candidate functions are in a "neighborhood" of the true function.

Remark 3. *An alternative of Assumption 4 would be*

$$||\log \lambda_{\beta,\gamma}||_{n,\infty} \leq \mu,\ \forall(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N$$

and

$$||\log \lambda_0||_{n,\infty} < \infty.$$

However, in this case, we could not consider conditional hazard rate functions that vanish at a point, for example hazard rates from the Weibull family that vanish when $t = 0$.

Proposition 2. *Under Assumption 4, there exist two positive numerical constants* μ' *and* μ'' *such that for all* $\boldsymbol{\beta} \in \mathbb{R}^M$ *and* $\boldsymbol{\gamma} \in \mathbb{R}^N$

$$\mu'||\log \lambda_{\beta,\gamma} - \log \lambda_0||^2_{n,\Lambda} \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) \leq \mu''||\log \lambda_{\beta,\gamma} - \log \lambda_0||^2_{n,\Lambda}. \tag{12}$$

This proposition will allow to deduce, from an oracle inequality in empirical Kullback divergence, an inequality in weighted empirical norm.

2.3 The Lasso estimation procedure

We consider a weighted Lasso procedure for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The Lasso estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ which minimize the ℓ_1-penalized empirical likelihood are defined by

$$(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathbb{R}^M\times\mathbb{R}^N}{\arg\min} \{C_n(\lambda_{\beta,\gamma}) + \text{pen}(\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\gamma})\}, \tag{13}$$

with

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

The weights ω_j and the δ_k are positive data-driven weights suitably chosen (see Equations (16) and (17)) and are respectively of order

$$\omega_j \approx \sqrt{\frac{\log M}{n}\hat{V}_n(f_j)} \text{ and } \delta_k \approx \sqrt{\frac{\log N}{n}\hat{R}_n(\theta_k)},$$

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

$$\hat{V}_n(f_j) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (f_j(\boldsymbol{Z_i}))^2 dN_i(s) \tag{14}$$

and

$$\hat{R}_n(\theta_k) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\theta_k(s))^2 dN_i(s). \tag{15}$$

The Lasso estimator of λ_0 is then defined by

$$\lambda_{\hat{\beta}_L,\hat{\gamma}_L} = \alpha_{\hat{\gamma}_L}(t)e^{f_{\hat{\beta}_L}(\boldsymbol{Z_i})},$$

with

$$\log \alpha_{\hat{\gamma}_L} = \sum_{k=1}^{N} \hat{\gamma}_{k,L}\theta_k \text{ and } f_{\hat{\beta}_L} = \sum_{j=1}^{M} \hat{\beta}_{j,L} f_j.$$

Usually, the Lasso estimator for $\boldsymbol{\beta}$ is defined by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^M}\{C_n(\lambda_{\beta,\gamma}) + \Gamma\sum_{j=1}^{M}|\beta_j|\},$$

So the Lasso penalization for $\boldsymbol{\beta}$ corresponds to the simple choice $\omega_j = \Gamma$ where $\Gamma > 0$ is a smoothing parameter. 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 [33]) 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 [33] and Zhang [31] for regression analysis and Zhang and Lu [30] for the Cox's 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.

The data-driven weights are defined for $j = 1, ..., M$ and $k = 1, ..., N$ by

$$\omega_j = c_{1,\varepsilon}\sqrt{\frac{x + \log M + \hat{\ell}_{n,x}(f_j)}{n}\hat{V}_n(f_j)} + c_{2,\varepsilon}\frac{x + 1 + \log M + \hat{\ell}_{n,x}(f_j)}{n}||f_j||_{n,\infty} \tag{16}$$

and

$$\delta_k = c'_{1,\varepsilon'}\sqrt{\frac{y + \log N + \hat{\ell}'_{n,x}(\theta_k)}{n}\hat{R}_n(\theta_k)} + c'_{2,\varepsilon'}\frac{y + 1 + \log N + \hat{\ell}'_{n,x}(\theta_k)}{n}||\theta_k||_{\infty}, \tag{17}$$

where

- $x > 0$, $y > 0$, $\varepsilon > 0$ and $\varepsilon' > 0$ are fixed
- $\hat{\ell}_{n,x}(f_j)$ and $\hat{\ell}'_{n,y}(\theta_k)$ are small technical terms coming out of our analysis :

$$\hat{\ell}_{n,x}(f_j) = c_\ell \log\log\left(\frac{2en\hat{V}_n(f_j) + 8e(4/3+\varepsilon)x||f_j||^2_{n,\infty}}{4(ec_0 - 2(4/3+\varepsilon)c_\ell)||f_j||^2_{n,\infty}} \vee e\right)$$

and

$$\hat{\ell}'_{n,y}(\theta_k) = c'_\ell \log\log\left(\frac{2en\hat{R}_n(\theta_k) + 8e(4/3+\varepsilon')y||\theta_k||^2_{n,\infty}}{4(ec'_0 - 2(4/3+\varepsilon')c'_\ell)||\theta_k||^2_{n,\infty}} \vee e\right),$$

where $c_\ell > 1$, $c'_\ell > 1$, $c_0 > 0$ and $c'_0 > 0$ such that $ec_0 \geq 2(4/3+\varepsilon)c_\ell$ and $ec'_0 \geq 2(4/3+\varepsilon')c'_\ell$.

- $c_{1,\varepsilon} = 4\sqrt{1+\varepsilon}$ and $c_{2,\varepsilon} = 2\sqrt{2\max(c_0, 2(1+\varepsilon)(4/3+\varepsilon))} + 2/3$,
- $c'_{1,\varepsilon'} = 4\sqrt{1+\varepsilon'}$ and $c'_{2,\varepsilon'} = 2\sqrt{2\max(c'_0, 2(1+\varepsilon')(4/3+\varepsilon'))} + 2/3$.

We have introduced the main notations that we will use in the following to present and prove our theorems.

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

As a first step, we suppose that the conditional hazard rate function satisfies the following generalization of the Cox's model

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) e^{f_0(\boldsymbol{Z_i})}, \tag{18}$$

where α_0 is the known baseline hazard function and f_0 a regression function. In this context, only f_0 has to be estimated and λ_0 is estimated by

$$\lambda_{\hat{\beta}_L}(t, \boldsymbol{Z_i}) = \alpha_0(t) e^{f_{\hat{\beta}_L}(\boldsymbol{Z_i})}$$

and

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^M}{\arg\min}\{C_n(\lambda_\beta) + \mathrm{pen}(\boldsymbol{\beta})\}.$$

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 theorem below, oracle inequality in the Cox's model with slow rate of convergence is stated. This inequality is obtain under a very light assumption that only concerns the construction of the dictionary $\mathbb{F}_M$.

Theorem 1. *Consider Model (18) with known α_0. Let ω_j be defined by (16) and for all $\boldsymbol{\beta} \in \mathbb{R}^M$,*

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

Let A be some numerical positive constant and $x > 0$ be fixed. Under Assumption 2, with a probability larger than $1 - Ae^{-x}$, then

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) \leq \inf_{\beta\in\mathbb{R}^M} \left(\widetilde{K}_n(\lambda_0, \lambda_\beta) + 2\,\mathrm{pen}(\boldsymbol{\beta})\right). \tag{19}$$

Note that, this is a prediction result on the conditional hazard rate function. This non-asymptotic inequality in prediction is a new result in the case of the Cox's model. On the other hand, 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. [6] for the regression model, Bertin et al. [5] and Bunea et al. [11] for density estimation).

3.2 A fast oracle inequality

To obtain a fast oracle inequality, an additional assumption is required. We present here the result obtained with the restricted eigenvalue condition, introduced in Bickel et al. [6]. First, let us introduce further notations :

$$\boldsymbol{\Delta} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta}) \text{ with } \boldsymbol{D} = (\text{diag}(\omega_j))_{1\leq j\leq M},$$

$$\boldsymbol{X} = (f_j(\boldsymbol{Z_i}))_{i,j}, \text{ with } i \in \{1, ..., n\} \text{ and } j \in \{1, ..., M\},$$

$$\boldsymbol{G_n} = \frac{1}{n}\boldsymbol{X^T C X} \text{ with } \boldsymbol{C} = (\text{diag}(\Lambda_i(\tau)))_{1\leq i\leq n}.$$

In $\boldsymbol{G_n}$, the covariates of individual i is re-weighted by its cumulative risk $\Lambda_i(\tau)$, which is consistent with the definition of the empirical norm in (8).

Let also $J(\boldsymbol{\beta})$ and $J(\boldsymbol{\gamma})$ be the sparsity sets of vectors $\boldsymbol{\beta} \in \mathbb{R}^M$ and $\boldsymbol{\gamma} \in \mathbb{R}^N$ respectively defined by

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

and the sparsity indexes are then given by

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

For $J \subset \{1, ..., M\}$, we also introduce the notation $\boldsymbol{\beta}_J$ to define the vector $\boldsymbol{\beta}$ restricted to the space J : $(\beta_J)_j = \beta_j$ if $j \in J$ and $(\beta_J)_j = 0$ if $j \in J^c$ where $J^c = \{1, ..., M\} \setminus J$.

Assumption 5 (Restricted eigenvalue condition $\mathbf{RE}(s, c_0)$)**.** *For some integer $s \in \{1, ..., M\}$ and a constant $c_0 > 0$, we assume that $\boldsymbol{G_n}$ satisfies :*

$$0 < \boldsymbol{\kappa}(s, c_0) = \min_{\substack{J\subset\{1,...,M\},\\ |J|\leq s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^M\setminus\{0\},\\ ||\boldsymbol{b}_{J^c}||_1\leq c_0||\boldsymbol{b}_J||_1}} \frac{(\boldsymbol{b^T G_n b})^{1/2}}{||\boldsymbol{b}_J||_2}.$$

This assumption is a key hypothesis on the weighted Gram matrix $\boldsymbol{G_n}$. The restricted eigenvalue assumption has been introduced in Bickel et al. [6] for the additive regression model. It ensures that the smallest eigenvalue restricted to the sparse set is strictly positive, namely $\boldsymbol{G_n}$ verifies a kind of "restricted" positive definiteness, which is only required for the vectors $\boldsymbol{b}$ satisfying

$$||\boldsymbol{b}_{J^c}||_1 \leq c_0||\boldsymbol{b}_J||_1. \tag{20}$$

It is one of the weakest assumption on the design matrix. See Bühlmann and van de Geer [8] and Bickel et al. [6] for further details on assumptions required for oracle inequalities.

Theorem 2. *Consider Model (18) with known α_0 and let ω_j be defined by (16). Let $A > 0$ be a numerical positive constant, $x > 0$, $\zeta > 0$ and $s \in \{1, ..., M\}$ be fixed and denote*

$$a_0 = 3 + \frac{4}{\zeta} \text{ and } \boldsymbol{\kappa} = \boldsymbol{\kappa}(s, a_0). \tag{21}$$

Under Assumptions 2, 4 and Assumption $\mathbf{RE}(s, a_0)$*, with a probability larger than $1 - Ae^{-x}$, the following inequality holds*

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) \leq (1+\zeta) \inf_{\substack{\boldsymbol{\beta} \in \mathbb{R}^M \\ |J(\boldsymbol{\beta})| \leq s}} \left\{ \widetilde{K}_n(\lambda_0, \lambda_\beta) + C(\zeta, \mu') \frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2} (\max_{1 \leq j \leq M} \omega_j)^2 \right\}, \tag{22}$$

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

This oracle inequality is the first non-asymptotic oracle inequality in prediction for the conditional hazard rate function with a fast rate of convergence of order $\log M/n$ in the Cox's model.

Thanks to the relation (12) between the empirical Kullback divergence (7) and the weighted empirical norm (8), we can deduce the following corollary :

Corollary 1. *Under the assumptions of Theorem 2, with a probability larger than $1 - Ae^{-x}$,*

$$|| \log \lambda_{\hat{\beta}_L} - \log \lambda_0 ||^2_{n,\Lambda} \leq (1+\zeta) \inf_{\substack{\beta \in \mathbb{R}^M \\ |J(\beta)| \leq s}} \left\{ || \log \lambda_\beta - \log \lambda_0 ||^2_{n,\Lambda} + C'(\zeta, \mu') \frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2} (\max_{1 \leq j \leq M} \omega_j)^2 \right\}, \tag{23}$$

where $C'(\zeta, \mu')$ 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

$$||f_{\hat{\beta}_L} - f_0||^2_{n,\Lambda} \leq (1+\zeta) \inf_{\substack{\beta \in \mathbb{R}^M \\ |J(\beta)| \leq s}} \left\{ ||f_\beta - f_0||^2_{n,\Lambda} + C'(\zeta, \mu') \frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2} (\max_{1 \leq j \leq M} \omega_j)^2 \right\}. \tag{24}$$

We get a non-asymptotic oracle inequality in weighted empirical norm, which 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. This inequality is comparable to the one obtained in Bickel et al. [6] in an additive regression model under a similar restricted eigenvalue assumption.

3.3 Particular case : variable selection in the Cox's model

We now consider the case of variable selection in the Cox's model (18) with $f_0(Z_i) = \boldsymbol{\beta_0^T Z_i}$. In this case, the functions of the dictionary are such that for $i = 1, ..., n$ and $j = 1, ..., p$

$$f_j(\boldsymbol{Z_i}) = Z_{i,j},$$

and then

$$f_\beta(\boldsymbol{Z_i}) = \sum_{i=1}^{p} \beta_j Z_{i,j} = \boldsymbol{\beta^T Z_i}.$$

Let $\boldsymbol{X} = (Z_{i,j})_{\substack{1 \le i \le n \\ 1 \le j \le p}}$ be the design matrix and

$$\boldsymbol{\Delta_0} = \hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0},\ J_0 = J(\boldsymbol{\beta_0}) \text{ and } |J_0| = \text{Card}\{J_0\}.$$

Our goal here is to obtain non-asymptotic inequalities for prediction on $\boldsymbol{X\beta_0}$ and for estimation on $\boldsymbol{\beta_0}$.

Theorem 3. *Consider Model (1) with known α_0. Let ω_j be defined by (16) and denote*

$$b_0 = 4\frac{\max\limits_{1 \le j \le p} \omega_j}{\min\limits_{1 \le j \le p} \omega_j} - 1 \text{ and } \boldsymbol{\kappa'} = \boldsymbol{\kappa}(s, b_0).$$

Let A be some numerical positive constant and $x > 0$ be fixed. Under Assumptions 2, 4 and $\mathbf{RE}(s, b_0)$, with a probability larger than $1 - Ae^{-x}$, then

$$||\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})||_{n,\Lambda}^2 \le \frac{4}{\mu'^2} \frac{|J_0|}{\boldsymbol{\kappa'}^2} (\max_{1 \le j \le p} \omega_j)^2 \tag{25}$$

and

$$||\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}||_1 \le \frac{1 + b_0}{\mu'} \frac{|J_0|}{\boldsymbol{\kappa'}^2} (\max_{1 \le j \le p} \omega_j). \tag{26}$$

This theorem gives non-asymptotic upper bounds on the loss. The first inequality of this theorem is an inequality for prediction with a rate of convergence in $\log M/n$, while the second one is a result in estimation on $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}$.

4 Oracle inequalities for general conditional hazard rate

In this section, we consider a general conditional hazard rate function λ_0. 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 conditional hazard rate is obtained under light assumptions that concern only the construction of the two dictionaries $\mathbb{F}_M$ and $\mathbb{G}_N$.

Theorem 4. *Let $B > 0$ be a positive numerical constant and $z > 0$ be fixed and Assumptions 2, 3 be satisfied.Then, with probability larger than $1 - Be^{-z}$*

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) \leq \inf_{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N} \{\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 2\,\mathrm{pen}(\boldsymbol{\beta}) + 2\,\mathrm{pen}(\boldsymbol{\gamma})\}. \tag{27}$$

In this inequality, $\mathrm{pen}(\beta)$ is the order of $||\boldsymbol{\beta}||_1\sqrt{\log M/n}$ and $\mathrm{pen}(\boldsymbol{\gamma})$ the order of $||\gamma||_1\sqrt{\log N/n}$. In Bertin et al. [5], for estimating a density function, different dictionaries of size of order n are proposed. We expect that the choice of N of order n would be suited for estimating the baseline hazard function. For such a choice, the leading term in Inequality (27) in high-dimension ($M >> n$) is of order $||\boldsymbol{\beta}||_1\sqrt{\log M/n}$. In this case, we obtain a non-asymptotic oracle inequality with a slow rate of convergence of the order of $\sqrt{\log M/n}$.

4.2 A fast oracle inequality

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

$$\tilde{\boldsymbol{\Delta}} = \tilde{\boldsymbol{D}} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma} \end{pmatrix} \in \mathbb{R}^{M+N} \text{ with } \tilde{\boldsymbol{D}} = \mathrm{diag}(\omega_1, ..., \omega_M, \delta_1, ..., \delta_N).$$

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

$$\tilde{\boldsymbol{X}}(t) = \begin{bmatrix} (f_j(\boldsymbol{Z_i}))_{\substack{1\leq i\leq n \\ 1\leq j\leq M}} & \mathbf{1}_{N\times N}(\mathrm{diag}(\theta_k(t)))_{1\leq k\leq N} \end{bmatrix}$$

$$= \left[\boldsymbol{X}\middle| \begin{matrix} \theta_1(t)...\theta_N(t) \\ ... \\ \theta_1(t)...\theta_N(t) \end{matrix}\right] \in \mathbb{R}^{n\times(M+N)}$$

and

$$\tilde{\boldsymbol{G}}_{\boldsymbol{n}} = \frac{1}{n}\int_0^\tau \tilde{\boldsymbol{X}}(t)^T \tilde{\boldsymbol{C}}(t) \tilde{\boldsymbol{X}}(t) dt \text{ with } \tilde{\boldsymbol{C}} = (\mathrm{diag}(\lambda_0(t, \boldsymbol{Z_i})Y_i(t)))_{1\leq i\leq n}.$$

Assumption 6 (Restricted eigenvalue condition $\mathbf{RE}(s, c_0)$ for the matrix $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$)**.** *For some integer $s \in \{1, ..., M + N\}$ and a constant $c_0 > 0$, we assume that $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$ satisfies*

$$0 < \tilde{\boldsymbol{\kappa}}(s, c_0) = \min_{\substack{J\subset\{1,...,M+N\}, \\ |J|\leq s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^{M+N}\setminus\{0\}, \\ ||\boldsymbol{b}_{J^c}||_1 \leq c_0||\boldsymbol{b}_J||_1}} \frac{(\boldsymbol{b}^T\tilde{\boldsymbol{G}}_{\boldsymbol{n}}\boldsymbol{b})^{1/2}}{||\boldsymbol{b}_J||_2}.$$

The **RE** condition on the matrix $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$ is quite 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 conditional hazard rate function. 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.

Theorem 5. *Let ω_j and δ_k be defined by (16) and (17) respectively. Let $B > 0$ be a numerical positive constant, $z > 0$, $\zeta > 0$ and $s \in \{1, ..., M+N\}$ be fixed, and denote*

$$r_0 = \left(3 + \frac{8}{\zeta}\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)\right) \frac{\max\limits_{\substack{1\le j\le M\\ 1\le k\le N}} \{\omega_j, \delta_k\}}{\min\limits_{\substack{1\le j\le M\\ 1\le k\le N}} \{\omega_j, \delta_k\}} \text{ and } \tilde{\boldsymbol{\kappa}} = \tilde{\boldsymbol{\kappa}}(s, r_0). \tag{28}$$

Under Assumptions 2, 3, 4 and Assumption $\mathbf{RE}(s, r_0)$*, then with probability larger than* $1 - Be^{-z}$

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) \le (1+\zeta) \inf_{\substack{\boldsymbol{\beta}\in\mathbb{R}^M, \boldsymbol{\gamma}\in\mathbb{R}^N\\ \max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|)\le s}} \left\{\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + \widetilde{C}(\zeta, \mu')\frac{\max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|)}{\tilde{\boldsymbol{\kappa}}^2} \max_{\substack{1\le j\le M\\ 1\le k\le N}} \{\omega_j^2, \delta_k^2\}\right\}, \tag{29}$$

and

$$\begin{aligned} &||\log\lambda_0 - \log\lambda_{\hat{\beta}_L, \hat{\gamma}_L}||^2_{n,\Lambda} \\ &\quad \le (1+\zeta) \inf_{\substack{\boldsymbol{\beta}\in\mathbb{R}^M, \boldsymbol{\gamma}\in\mathbb{R}^N\\ \max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|)\le s}} \left\{||\log\lambda_0 - \log\lambda_{\beta,\gamma}||^2_{n,\Lambda} + \widetilde{C}'(\zeta, \mu')\frac{\max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|)}{\tilde{\boldsymbol{\kappa}}^2} \max_{\substack{1\le j\le M\\ 1\le k\le N}} \{\omega_j^2, \delta_k^2\}\right\}, \end{aligned} \tag{30}$$

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

We obtain a non-asymptotic fast oracle inequality in prediction. Indeed,

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

namely, if we choose $\mathbb{G}_N$ of size n, the rate of convergence of this oracle inequality is then of order $\log M/n$. This inequality compares the unknown conditional hazard risk to the best Cox's model obtained by estimating the baseline hazard function and the vector of parameters by Lasso estimators. Such inequality allows to predict the survival time throughout the conditional hazard rate in a high dimensional setting. This is a new approach of the problem.

5 An empirical Bernstein's inequality

In this section, we present two empirical Bernstein's inequalities that are the key results in proving our oracle inequalities.

Using the Doob-Meier decomposition $N_i = M_i + \Lambda_i$, we can easily show that for all $\boldsymbol{\beta} \in \mathbb{R}^M$ and for all $\boldsymbol{\gamma} \in \mathbb{R}^N$

$$C_n(\lambda_{\hat{\beta}_L,\hat{\gamma}_L}) - C_n(\lambda_{\beta,\gamma}) = \widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) - \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + (\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma})^T \boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}} + (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}}, \tag{31}$$

where

$$\boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \vec{\boldsymbol{f}}(\boldsymbol{Z_i}) dM_i(t), \text{ with } \vec{\boldsymbol{f}} = (f_1, ..., f_M)^T \tag{32}$$

$$\boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \vec{\boldsymbol{\theta}}(t) dM_i(t), \text{ with } \vec{\boldsymbol{\theta}} = (\theta_1, ..., \theta_N)^T. \tag{33}$$

The main part of the proofs of the theorems relies on the control of the centered empirical processes $\boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}}$ and $\boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}}$. We then introduce the jth coordinate (respectively the kth coordinate) of the processes $\boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}}$ (respectively $\boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}}$) in t :

$$\eta_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t f_j(\boldsymbol{Z_i}) dM_i(s),$$
$$\nu_{n,t}(\theta_k) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t \theta_k(s) dM_i(s).$$

We define the predictable variations of $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ by

$$V_{n,t}(f_j) = n < \eta_n(f_j) >_t = \frac{1}{n}\sum_{i=1}^{n} \int_0^t (f_j(\boldsymbol{Z_i}))^2 \lambda_0(t, \boldsymbol{Z_i}) Y_i(s) ds,$$
$$R_{n,t}(\theta_k) = n < \nu_n(\theta_k) >_t = \frac{1}{n}\sum_{i=1}^{n} \int_0^t (\theta_k(t))^2 \lambda_0(t, \boldsymbol{Z_i}) Y_i(s) ds,$$

and the optional variations of $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ by

$$\hat{V}_{n,t}(f_j) = n[\eta_n(f_j)]_t = \frac{1}{n}\sum_{i=1}^{n} \int_0^t (f_j(\boldsymbol{Z_i}))^2 dN_i(s),$$
$$\hat{R}_{n,t}(\theta_k) = n[\nu_n(\theta_k)]_t = \frac{1}{n}\sum_{i=1}^{n} \int_0^t (\theta_k(t))^2 dN_i(s).$$

The optional variations can be seen as estimators of $V_{n,t}(f_j)$ and $R_{n,t}(\theta_k)$ respectively. The following theorem is close to Theorem 3 in Gaïffas and Guilloux [14] proved for the Aalen additive model. See also Hansen et al. [16].

Theorem 6. *For any numerical constant $c_\ell > 1$, $c'_\ell > 1$, $\varepsilon > 0$, $\varepsilon' > 0$ and $c_0 > 0$, $c'_0 > 0$ such that $ec_0 > 2(4/3+\varepsilon)c_\ell$ and $ec'_0 > 2(4/3+\varepsilon')c'_\ell$, the following holds for any $x > 0$, $y > 0$:*

$$\mathbb{P}\Big[|\eta_{n,t}(f_j)| \geq c_{1,\varepsilon}\sqrt{\frac{x+\hat{\ell}_{n,x}(f_j)}{n}\hat{V}_{n,t}(f_j)} + c_{2,\varepsilon}\frac{x+1+\hat{\ell}_{n,x}(f_j)}{n}||f_j||_{n,\infty}\Big] \leq c_{3,\varepsilon,c_\ell}e^{-x}, \tag{34}$$

$$\mathbb{P}\Big[|\nu_{n,t}(\theta_k)| \geq c'_{1,\varepsilon'}\sqrt{\frac{y+\hat{\ell}'_{n,y}(\theta_k)}{n}\hat{R}_{n,t}(\theta_k)} + c'_{2,\varepsilon'}\frac{y+1+\hat{\ell}'_{n,y}(\theta_k)}{n}||\theta_k||_{n,\infty}\Big] \leq c'_{3,\varepsilon',c'_\ell}e^{-y}, \tag{35}$$

where

$$\hat{\ell}_{n,x}(f_j) = c_\ell \log\log\left(\frac{2en\hat{V}_{n,t}(f_j) + 8e(4/3+\varepsilon)x||f_j||^2_{n,\infty}}{4(ec_0 - 2(4/3+\varepsilon)c_\ell)||f_j||^2_{n,\infty}} \vee e\right), \quad ||f_j||_{n,\infty} = \max_{i=1,\dots,n}|f_j(\boldsymbol{Z_i})|,$$

$$\hat{\ell}'_{n,y}(\theta_k) = c'_\ell \log\log\left(\frac{2en\hat{R}_{n,t}(\theta_k) + 8e(4/3+\varepsilon')y||\theta_k||^2_{n,\infty}}{4(ec'_0 - 2(4/3+\varepsilon')c'_\ell)||\theta_k||^2_{n,\infty}} \vee e\right), \quad ||\theta_k||_{n,\infty} = \max_{t\in[0,\tau]}|\theta_k(t)|,$$

and where

$$c_{1,\varepsilon} = 2\sqrt{1+\varepsilon}, \qquad c_{2,\varepsilon} = 2\sqrt{2\max(c_0, 2(1+\varepsilon)(4/3+\varepsilon))} + 2/3,$$

$$c_{3,\varepsilon,c_\ell} = 8 + 6(\log(1+\varepsilon))^{-c_\ell}\sum_{k\geq 1}k^{-c_\ell},$$

$$c'_{1,\varepsilon} = 2\sqrt{1+\varepsilon}, \qquad c'_{2,\varepsilon} = 2\sqrt{2\max(c_0, 2(1+\varepsilon)(4/3+\varepsilon))} + 2/3,$$

$$c'_{3,\varepsilon,c'_\ell} = 8 + 6(\log(1+\varepsilon))^{-c'_\ell}\sum_{k\geq 1}k^{-c'_\ell}.$$

This empirical Berntein's inequality hold true for martingales with jumps, when the predictable variation is not observable.

Acknowledgements

All my thanks go to my two Phd Thesis supervisors Agathe Guilloux and Marie-Luce Taupin for their help, their availability and their advices. I also thank Marius Kwemou for helpful discussions.

6 Proofs

6.1 Proof of Proposition 1

Following the proof of Theorem 1 in Senoussi [22], we rewrite the empirical Kullback divergence (7) as

$$\begin{aligned}\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[\log\lambda_0(t,\boldsymbol{Z_i}) - \log\lambda_{\beta,\gamma}(t,\boldsymbol{Z_i}) - \Big(1 - \frac{\lambda_{\beta,\gamma}(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})}\Big)\Big]\lambda_0(t,\boldsymbol{Z_i})Y_i(t)dt \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big[e^{\log\frac{\lambda_{\beta,\gamma}(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})}} - \log\frac{\lambda_{\beta,\gamma}(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})} - 1\Big]\lambda_0(t,\boldsymbol{Z_i})Y_i(t)dt.\end{aligned}$$

Since the map $t \to e^t - t - 1$ is a positive function on $\mathbb{R}$, we deduce that

$$e^{\log\frac{\lambda_{\beta,\gamma}(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})}} - \log\frac{\lambda_{\beta,\gamma}(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})} - 1 > 0$$

except for $\lambda_{\beta,\gamma} = \lambda_0$. Thus $\widetilde{K}_n(f_0, f_\beta)$ is positive and vanishes only if $(\log\lambda_0 - \log\lambda_{\beta,\gamma})(t,\boldsymbol{Z_i}) = 0$ almost surely, namely if $\lambda_0 = \lambda_{\beta,\gamma}$ almost surely. □

6.2 Proof of Proposition 2

To compare the empirical Kullback divergence (7) and the weighted empirical norm (8), we use Lemma 1 in Bach [4].

Lemma 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 $t \ge 0$:*

$$\frac{g''(0)}{S^2}(e^{-St} + St - 1) \le g(t) - g(0) - g'(0)t \le \frac{g''(0)}{S^2}(e^{St} - St - 1). \tag{36}$$

This Lemma gives upper and lower Taylor expansions for some convex and three times differentiable function. It has been introduced by Bach 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 to the logistic loss.

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

$$G(h) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log(h(s,\boldsymbol{Z_i}))d\Lambda_i(s) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau h(s,\boldsymbol{Z_i})Y_i(s)ds.$$

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 :

$$g'(t) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau k(s,\boldsymbol{Z_i})d\Lambda_i(s) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau k(s,\boldsymbol{Z_i})e^{h(s,\boldsymbol{Z_i})+tk(s,\boldsymbol{Z_i})}Y_i(s)ds,$$
$$g''(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (k(s,\boldsymbol{Z_i}))^2 e^{h(s,\boldsymbol{Z_i})+tk(s,\boldsymbol{Z_i})}Y_i(s)ds,$$
$$g'''(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (k(s,\boldsymbol{Z_i}))^3 e^{h(s,\boldsymbol{Z_i})+tk(s,\boldsymbol{Z_i})}Y_i(s)ds.$$

It follows that

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

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

$$\frac{g''(0)}{||k||^2_{n,\infty}}(e^{-t||k||_{n,\infty}} + t||k||_{n,\infty} - 1) \leq g(t) - g(0) - g'(0)t \leq \frac{g''(0)}{||k||^2_{n,\infty}}(e^{t||k||_{n,\infty}} - t||k||_{n,\infty} - 1).$$

Take $t = 1$ and $h(s,\boldsymbol{Z_i}) = \log\lambda_0(s,\boldsymbol{Z_i})$ and $k(s,\boldsymbol{Z_i}) = \log\lambda_{\beta,\gamma}(s,\boldsymbol{Z_i}) - \log\lambda_0(s,\boldsymbol{Z_i})$, and introduce the two following functions

$$\Phi : t \to \frac{e^{-t}+t-1}{t^2} \text{ and } \Psi : t \to \frac{e^t - t - 1}{t^2}.$$

We obtain

$$G(\log\lambda_{\beta,\gamma}) - G(\log\lambda_0) - g'(0) \geq g''(0)\Phi(||\log\lambda_{\beta,\gamma} - \log\lambda_0||_{n,\infty}) \tag{37}$$

and

$$G(\log\lambda_{\beta,\gamma}) - G(\log\lambda_0) - g'(0) \leq g''(0)\Psi(||\log\lambda_{\beta,\gamma} - \log\lambda_0||_{n,\infty}). \tag{38}$$

Now straightforward calculations show that g'(0)=0 and

$$\begin{aligned} g''(0) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau ((\log\lambda_{\beta,\gamma} - \log\lambda_0)(s,\boldsymbol{Z_i}))^2 d\Lambda_i(s) \\ &= ||\log\lambda_{\beta,\gamma} - \log\lambda_0||^2_{n,\Lambda}. \end{aligned}$$

Replacing $g'(0)$ and $g''(0)$ by their expressions in (37) and (38) and noting that

$$G(\log\lambda_{\beta,\gamma}) - G(\log\lambda_0) = \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}),$$

we get

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

and

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

According to Assumption 4,

$$||\log\lambda_{\beta,\gamma} - \log\lambda_0||_{n,\infty} \leq \mu.$$

Since Φ (respectively Ψ) is decreasing (respectively increasing) and bounded below by 0, we can deduce that

$$\Phi(||\log\lambda_{\beta,\gamma} - \log\lambda_0||_{n,\infty}) \geq \Phi(\mu)$$

and

$$\Psi(||\log\lambda_{\beta,\gamma} - \log\lambda_0||_{n,\infty}) \leq \Psi(\mu).$$

Take $\mu' := \Phi(\mu) > 0$ and $\mu'' := \Psi(\mu) > 0$ and conclude that

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

□

6.3 Proof of Theorem 6

The proofs of (34) and (35) are quite similar, so we only present the one of (34). To prove (35), it suffices to replace $\eta_{n,t}(f_j)$ by the process $\nu_{n,t}(\theta_k)$ throughout the following. Let us denote by $U_{n,t}$ and $H_i(f_j)$ the equations

$$U_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(f_j)dM_i(s) \text{ and } H_i(f_j) := \frac{f_j(\boldsymbol{Z_i})}{\max\limits_{1\leq i\leq n}|f_j(\boldsymbol{Z_i})|}.$$

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

$$\vartheta_{n,t}(f_j) = n < U_{n,}(f_j) >_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (f_j(\boldsymbol{Z_i}))^2 d\Lambda_i(s)$$

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

$$\hat{\vartheta}_{n,t}(f_j) = n[U_n(f_j)]_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (f_j(\boldsymbol{Z_i}))^2 dN_i(s).$$

The proof relies on the three following steps :

Step1 : We prove first that

$$\mathbb{P}\Big[U_{n,t}(f_j) \geq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}, v < \vartheta_{n,t}(f_j) \leq \omega\Big] \leq e^{-x}.$$

Step 2 : Step 2 consists in replacing $\vartheta_{n,t}(f_j)$ by the observable $\hat{\vartheta}_{n,t}(f_j)$ in Step 1. It follows that

$$\mathbb{P}\Big[U_{n,t}(f_j) \geq 2\sqrt{\frac{\omega x}{vn}\hat{\vartheta}_{n,t}(f_j)} + \Big(2\sqrt{\frac{\omega}{v}\Big(\frac{\omega}{v}+\frac{1}{3}\Big)} + \frac{1}{3}\Big)\frac{x}{n}, v \leq \vartheta_{n,t}(f_j) < \omega\Big] \leq 3e^{-x}. \tag{39}$$

Step 3 : Finally, Step 3 is devoting to remove the event $\{v \leq \vartheta_{n,t}(f_j) < \omega\}$ from Inequality (39) to finish the proof.

Step 1 : Let

$$S_{\lambda,t}(f_j) = \sum_{i=1}^{n} \int_0^t \phi\Big(\frac{\lambda}{n} H_i(f_j)\Big) d\Lambda_i(s) \text{ and } \phi(x) = e^x - x - 1.$$

From van de Geer [27], we know that

$$W_{n,\lambda,t}(f_j) = \exp(\lambda U_{n,t}(f_j) - S_{\lambda,t}(f_j)) \tag{40}$$

is a supermartingale, so that, from Markov Inequality, for any λ, $x > 0$, we obtain

$$\mathbb{P}\left[U_{n,t}(f_j) \geq \frac{S_{\lambda,t}(f_j)}{\lambda} + \frac{x}{\lambda}\right] \leq e^{-x}. \tag{41}$$

We then introduce three properties :

1. $\phi(xh) \leq h^2\phi(x)$ for any $0 \leq h \leq 1$ and $x > 0$;
2. $\phi(\lambda) \leq \dfrac{\lambda^2}{2(1-\lambda/3)}$ for any $\lambda \in (0,3)$;
3. $\min\limits_{\lambda\in(0,1/b)} \left(\dfrac{a\lambda}{1-b\lambda} + \dfrac{x}{\lambda}\right) = 2\sqrt{ax} + bx$, for any $a, b, x > 0$.

With $a = \dfrac{\omega}{2n}$ and $b = \dfrac{1}{3n}$, let λ_ω be defined by

$$\lambda_\omega = \arg\min_{\lambda\in(0,1/b)} \left(\frac{a\lambda}{1-b\lambda} + \frac{x}{\lambda}\right).$$

These three properties entail the following embeddings

$$\begin{aligned}
\left\{U_{n,t}(f_j) \geq \sqrt{\frac{2\omega x}{n}} + \frac{x}{3n}, \vartheta_{n,t}(f_j) \leq \omega\right\} &= \left\{U_{n,t}(f_j) \geq \frac{\lambda_\omega}{2(n-\lambda_\omega/3)}\omega + \frac{x}{\lambda_\omega}, \vartheta_{n,t}(f_j) \leq \omega\right\} \\
&\subset \left\{U_{n,t}(f_j) \geq \frac{\phi(\lambda_\omega/n)}{\lambda_\omega} n\vartheta_{n,t}(f_j) + \frac{x}{\lambda_\omega}, \vartheta_{n,t}(f_j) \leq \omega\right\} \\
&\subset \left\{U_{n,t}(f_j) \geq \frac{S_{\lambda_\omega,t}(f_j)}{\lambda_\omega} + \frac{x}{\lambda_\omega}, \vartheta_{n,t}(f_j) \leq \omega\right\}.
\end{aligned} \tag{42}$$

This leads to the standard Bernstein's inequality (see Uspensky [26] or Massart [18] for the classical Bernstein's inequality and van de Geer [27] for the Bernstein's inequality for martingales)

$$\mathbb{P}\left[U_{n,t}(f_j) \geq \sqrt{\frac{2\omega x}{n}} + \frac{x}{3n}, \vartheta_{n,t}(f_j) \leq \omega\right] \leq e^{-x}.$$

By choosing $\omega = c_0(x+1)/n$ for some constant $c_0 > 0$, we obtain

$$\mathbb{P}\left[U_{n,t}(f_j) \geq \left(\sqrt{2c_0} + \frac{1}{3}\right)\frac{x+1}{n}, \vartheta_{n,t}(f_j) \leq \frac{c_0(x+1)}{n}\right] \leq e^{-x}. \tag{43}$$

This inequality says that when the variance term $\vartheta_{n,t}(f_j)$ is small, the sub-exponential term is dominating in Bernstein's inequality. For any $0 < v < \omega < +\infty$, we have

$$\begin{aligned}\left\{U_{n,t}(f_j) \geq &\sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}\right\} \cap \{v < \vartheta_{n,t}(f_j) \leq \omega\} \\ &\subset \left\{U_{n,t}(f_j) \geq \sqrt{\frac{2\omega x}{n}} + \frac{x}{3n}\right\} \cap \{v < \vartheta_{n,t}(f_j) \leq \omega\}.\end{aligned}$$

It follows

$$\mathbb{P}\Big[U_{n,t}(f_j) \geq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}, v < \vartheta_{n,t}(f_j) \leq \omega\Big] \leq e^{-x}, \tag{44}$$

which ends up the proof of Step 1.

Step 2 : We aim at replacing $\vartheta_{n,t}(f_j)$ which is non observable by the observable $\hat{\vartheta}_{n,t}(f_j)$ in Equation (44). Let us denote by $\widetilde{U}_{n,t}(f_j)$ the quantity

$$\begin{aligned}\widetilde{U}_{n,t}(f_j) = \hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t (H_i(f_j))^2\Big(dN_i(s) - d\Lambda_i(s)\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t (H_i(f_j))^2 dM_i(s).\end{aligned}$$

The process $\widetilde{U}_{n,t}(f_j)$ is a martingale and hence, again from van de Geer [27], $\exp(\lambda\widetilde{U}_{n,t}(f_j) - \widetilde{S}_{\lambda,t}(f_j))$ is a supermartingale with

$$\widetilde{S}_{\lambda,t}(f_j) = \sum_{i=1}^{n}\int_0^t \phi\Big(\frac{\lambda}{n}H_i(f_j)^2\Big)d\Lambda_i(s).$$

Now, writing again (42) for $\widetilde{U}_{n,t}(f_j)$ and using the same argument as in Step 1, we obtain

$$\mathbb{P}\Big[|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \geq \frac{\phi(\lambda/n)}{\lambda}n\vartheta_{n,t}(f_j) + \frac{x}{\lambda}\Big] \leq 2e^{-x} \tag{45}$$

and

$$\mathbb{P}\Big[|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \geq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}, v < \vartheta_{n,t}(f_j) \leq \omega\Big] \leq 2e^{-x}. \tag{46}$$

If $\vartheta_{n,t}(f_j)$ satisfies

$$|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \leq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}, \tag{47}$$

then, it satisfies

$$\vartheta_{n,t}(f_j) \leq \hat{\vartheta}_{n,t}(f_j) + \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}.$$

Thanks to the fact that $A \leq b + \sqrt{aA}$ entails $A \leq a + 2b$ for any $a, A, b > 0$, taking $A = \vartheta_{n,t}(f_j)$, $a = \dfrac{2\omega x}{vn}$ and $b = \hat{\vartheta}_{n,t}(f_j) + \dfrac{x}{3n}$, we obtain

$$\vartheta_{n,t}(f_j) \leq 2\hat{\vartheta}_{n,t}(f_j) + 2\Big(\frac{\omega}{v} + \frac{1}{3}\Big)\frac{x}{n}. \tag{48}$$

If $\vartheta_{n,t}(f_j)$ satisfies Inequality (47), we also have

$$\hat{\vartheta}_{n,t}(f_j) \leq \vartheta_{n,j} + \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}.$$

Applying Inequality (48) in the previous inequality, we get

$$\begin{aligned}\hat{\vartheta}_{n,t}(f_j) &\leq \vartheta_{n,t}(f_j) + \sqrt{\frac{2\omega x}{vn}\Big(2\hat{\vartheta}_{n,t}(f_j) + 2\Big(\frac{\omega}{v} + \frac{1}{3}\Big)\frac{x}{n}\Big)} + \frac{x}{3n} \\ &\leq \vartheta_{n,t}(f_j) + \sqrt{\frac{4\omega x}{vn}\hat{\vartheta}_{n,t}(f_j)} + \sqrt{\frac{4\omega x}{vn}\Big(\frac{\omega}{v} + \frac{1}{3}\Big)\frac{x}{n}} + \frac{x}{3n},\end{aligned}$$

once again using that $A \leq b + \sqrt{aA}$ entails $A \leq a + 2b$ with $A = \hat{\vartheta}_{n,t}(f_j)$, $a = \dfrac{4\omega x}{vn}$ and $b = \vartheta_{n,t}(f_j)\sqrt{\dfrac{4\omega x}{vn}\Big(\dfrac{\omega}{v} + \dfrac{1}{3}\Big)\dfrac{x}{n}} + \dfrac{x}{3n}$, we obtain

$$\hat{\vartheta}_{n,t}(f_j) \leq 2\vartheta_{n,t}(f_j) + 2\Big(\frac{1}{3} + 2\sqrt{\frac{\omega}{v}\Big(\frac{\omega}{v} + \frac{1}{3}\Big)} + \frac{2\omega}{v}\Big)\frac{x}{n}. \tag{49}$$

We now deduce from Inequality (48) that

$$\begin{aligned}\Big\{U_{n,t}(f_j) \leq &\sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}\Big\} \cap \Big\{|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \leq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}\Big\} \\ &\subset \Big\{U_{n,t}(f_j) \leq 2\sqrt{\frac{\omega x}{vn}\hat{\vartheta}_{n,t}(f_j)} + \Big(2\sqrt{\frac{\omega}{v}\Big(\frac{\omega}{v} + \frac{1}{3}\Big)} + \frac{1}{3}\Big)\frac{x}{n}\Big\}.\end{aligned} \tag{50}$$

Using (44) and (46), we finally obtain

$$\mathbb{P}\Big[U_{n,t}(f_j) \geq 2\sqrt{\frac{\omega x}{vn}\hat{\vartheta}_{n,t}(f_j)} + \Big(2\sqrt{\frac{\omega}{v}\Big(\frac{\omega}{v} + \frac{1}{3}\Big)} + \frac{1}{3}\Big)\frac{x}{n}, v < \vartheta_{n,t}(f_j) \leq \omega\Big] \leq 3e^{-x}. \tag{51}$$

Step 3 : It remains to remove the event $\{v \le \vartheta_{n,t}(f_j) < \omega\}$ in (51). For $k \ge 0$, set

$$v_k = c_0 \frac{x+1}{n}(1+\varepsilon)^k,$$

and use the following decomposition into disjoint sets :

$$\{\vartheta_{n,t}(f_j) > c_0 x/n\} = \bigcup_{k\ge 0} \{v_k < \vartheta_{n,t}(f_j) \le v_{k+1}\}. \tag{52}$$

Instead of considering the event $\{v < \vartheta_{n,t}(f_j) \le \omega\}$, we calculate the probabilities on $\{\vartheta_{n,t}(f_j) > v_0\}$ and on its complementary to finally get the expected probability. According to (51)

$$\mathbb{P}\Big[U_{n,t}(f_j) \ge c_{1,\epsilon}\sqrt{\frac{x}{n}\hat{\vartheta}_{n,t}(f_j)} + c_{2,\varepsilon}\frac{x}{n}, v_k < \vartheta_{n,t}(f_j) \le v_{k+1}\Big] \le 3e^{-x}, \tag{53}$$

with

$$c_{1,\varepsilon} = 2\sqrt{1+\varepsilon} \text{ and } c_{2,\varepsilon} = 2\sqrt{(1+\varepsilon)(4/3+\varepsilon)} + 1/3.$$

Set for some constant $c_\ell > 1$,

$$\ell = c_\ell \log\log\Big(\frac{\vartheta_{n,t}(f_j)}{v_0} \vee e\Big).$$

On the event

$$D_{n,\ell,\varepsilon} = \Big\{|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \le \sqrt{\frac{2(1+\varepsilon)\vartheta_{n,t}(f_j)(x+\ell)}{n}} + \frac{x+\ell}{3n}\Big\} \tag{54}$$

applying (48) with $\dfrac{\omega}{v} = 1+\varepsilon$ and replacing x by $x+\ell$, we have

$$\vartheta_{n,t}(f_j) \le 2\hat{\vartheta}_{n,t}(f_j) + 2(4/3+\varepsilon)\frac{x}{n} + \frac{2(4/3+\varepsilon)c_\ell}{n}\log\log\Big(\frac{\vartheta_{n,t}(f_j)}{v_0} \vee e\Big).$$

We now use the fact that $\log\log(x) \le x/e - 1$ for any $x \ge e$, and since $ec_0 > 2(4/3+\varepsilon)c_\ell$ we get

$$\vartheta_{n,t}(f_j) \le \frac{ec_0}{ec_0 - 2(4/3+\varepsilon)c_\ell}\Big(2\hat{\vartheta}_{n,t}(f_j) + 2(4/3+\varepsilon)\frac{x}{n}\Big).$$

Combining the last inequality with (50), we obtain the following embeddings :

$$\begin{aligned}\Big\{U_{n,t}(f_j) \le \sqrt{\frac{2(1+\varepsilon)\vartheta_{n,t}(f_j)(x+\ell)}{n}} + \frac{x+\ell}{3n}\Big\} \cap D_{n,\ell,\varepsilon} \\ \subset \Big\{U_{n,t}(f_j) \le c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}\Big\},\end{aligned} \tag{55}$$

where

$$\hat{\ell} = c_\ell \log\log\Big(\frac{2en\hat{\vartheta}_{n,t}(f_j) + 2e(4/3+\varepsilon)x}{ec_0 - 2(4/3+\varepsilon)c_\ell} \vee e\Big).$$

From (52), we have

$$\begin{aligned}
&\mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, \vartheta_{n,t}(f_j) > v_0\Big] \\
&\leq \sum_{k\geq 0} \mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\Big].
\end{aligned}$$

Then, we write

$$\begin{aligned}
&\sum_{k\geq 0} \mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\Big] \\
&= \sum_{k\geq 0} \mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, D_{n,\ell,\varepsilon}, v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\Big] \\
&+ \sum_{k\geq 0} \mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, D^c_{n,\ell,\varepsilon}, v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\Big],
\end{aligned}$$

where $D^c_{n,\ell,\varepsilon}$ is the complementary of $D_{n,\ell,\varepsilon}$. Applying (55), we get

$$\begin{aligned}
&\sum_{k\geq 0} \mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\Big] \\
&\leq \sum_{k\geq 0} \mathbb{P}\Big[U_{n,t}(f_j) \geq \sqrt{\frac{2(1+\varepsilon)\vartheta_{n,t}(f_j)(x+\ell)}{n}} + \frac{x+\ell}{3n}, v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\Big] \\
&+ \sum_{k\geq 0} \mathbb{P}\Big[D^c_{n,\ell,\varepsilon}, v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\Big].
\end{aligned}$$

Gathering (44) and (46), we finally obtain

$$\begin{aligned}
&\mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, \vartheta_{n,t}(f_j) > v_0\Big] \\
&\leq 3\Big(e^{-x} + \sum_{j\geq 1} e^{-(x+c_\ell \log\log(v_j/v_0))}\Big) \\
&= 3(1 + (\log(1+\varepsilon))^{-c_\ell}\sum_{j\geq 1} j^{-c_\ell})e^{-x}. \qquad (56)
\end{aligned}$$

According to (43), we get

$$\mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{3,\varepsilon}\frac{x+1+\hat{\ell}}{n}\Big] \leq \Big(4 + 3(\log(1+\varepsilon))^{-c_\ell}\sum_{j\geq 1} j^{-c_\ell}\Big)e^{-x},$$

where $c_{3,\varepsilon} = \sqrt{2\max(c_0, 2(1+\varepsilon)(4/3+\varepsilon))} + 1/3$.

Now it suffices to multiply both sides of the inequality

$$U_{n,t}(f_j) \geq c_{1,\varepsilon}\sqrt{\frac{x+\hat{\ell}}{n}\hat{\vartheta}_{n,t}(f_j)} + c_{3,\varepsilon}\frac{x+1+\hat{\ell}}{n}$$

by $||f_j||_{n,\infty} = \max_{i=1,\ldots,n} |f_j(\boldsymbol{Z_i})|$ to end up the proof of Theorem 6.

□

6.4 Proof of Theorem 1

By definition of $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$, for all β in $\mathbb{R}^M$, we have

$$C_n(\lambda_{\hat{\beta}_L}) + \text{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}) \leq C_n(\lambda_\beta) + \text{pen}(\boldsymbol{\beta}) \qquad \forall \boldsymbol{\beta} \in \mathbb{R}^M.$$

Applying (31) when α_0 is known, we obtain

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) \leq \widetilde{K}_n(\lambda_0, \lambda_\beta) + (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta_{n,\tau}} + \text{pen}(\boldsymbol{\beta}) - \text{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}). \tag{57}$$

It remains to control the term $(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta_{n,\tau}}$. Set

$$\mathcal{A} = \bigcap_{j=1}^{M} \left\{ |\eta_{n,\tau}(f_j)| \leq \frac{\omega_j}{2} \right\}, \tag{58}$$

where the weights ω_j are given by (16). On $\mathcal{A}$, we have

$$|(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta_{n,\tau}}| \leq \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}_L - \beta)_j| \leq \sum_{j=1}^{M} \omega_j |(\hat{\beta}_L - \beta)_j|.$$

Since $\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j |\beta_j|$, for any $\boldsymbol{\beta}$ in $\mathbb{R}^M$ we get

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) \leq \inf_{\boldsymbol{\beta} \in \mathbb{R}^M} \left(\widetilde{K}_n(\lambda_0, \lambda_\beta) + 2\,\text{pen}(\boldsymbol{\beta}) \right).$$

It remains to bound up $\mathbb{P}(\mathcal{A})$ by applying Theorem 6

$$\begin{aligned} \mathbb{P}(\mathcal{A}^c) &\leq \sum_{j=1}^{M} \mathbb{P}\Big(|\eta_n(f_j)| > \frac{\omega_j}{2} \Big) \\ &\leq c_{3,\varepsilon,c_\ell} e^{-x}. \end{aligned}$$

Consequently, by taking $A = c_{3,\varepsilon,c_\ell}$, we conclude

$$\mathbb{P}(\mathcal{A}) \geq 1 - Ae^{-x},$$

which ends up the proof of Theorem 1. □

6.5 Proof of Theorem 2

We start from Inequality (57) and the fact that on $\mathcal{A}$

$$|(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}}| \leq \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}_L - \beta)_j|.$$

It follows that

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

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

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) + \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}_L - \beta)_j| \leq \widetilde{K}_n(\lambda_0, \lambda_\beta) + 2 \sum_{j \in J(\beta)} \omega_j |(\hat{\beta}_L - \beta)_j|. \tag{59}$$

We apply Cauchy-Schwarz Inequality to the second right hand side of this inequality to get

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) + \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}_L - \beta)_j| \leq \widetilde{K}_n(\lambda_0, \lambda_\beta) + 2\sqrt{|J(\boldsymbol{\beta})|} \sqrt{\sum_{j \in J(\beta)} \omega_j^2 |\hat{\beta}_L - \beta|_j^2}. \tag{60}$$

With the notations $\boldsymbol{\Delta} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})$ and $\boldsymbol{D} = (\mathrm{diag}(\omega_j))_{1 \leq j \leq M}$ introduced in Subsection 3.2 , we can rewrite Inequality (59) as

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) + \frac{1}{2} ||\boldsymbol{\Delta}||_1 \leq \widetilde{K}_n(\lambda_0, \lambda_\beta) + 2||\boldsymbol{\Delta}_{J(\beta)}||_1, \tag{61}$$

and Inequality (60) as

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L}) \leq \widetilde{K}_n(\lambda_0, \lambda_\beta) + 2\sqrt{|J(\beta)|} ||\boldsymbol{\Delta}_{J(\beta)}||_2. \tag{62}$$

We then consider two cases

$$2||\boldsymbol{\Delta}_{J(\beta)}||_1 \leq \zeta \widetilde{K}_n(\lambda_0, \lambda_\beta) \tag{63}$$

and

$$\zeta \widetilde{K}_n(\lambda_0, \lambda_\beta) \leq 2||\boldsymbol{\Delta}_{J(\beta)}||_1. \tag{64}$$

In Case (63), the result of the theorem follows immediately from (61). So we focus on Case (64). We introduce the event

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

On $\mathcal{A} \bigcap \mathcal{A}_1$, applying (61) 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

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

For $\boldsymbol{\Delta}' = \boldsymbol{D}^{-1}\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta}$, we have

$$||\boldsymbol{\Delta}'_{J(\boldsymbol{\beta})^c}||_1 \leq \left(3 + \frac{4}{\zeta}\right)\frac{\max\limits_{1\leq j\leq M}\omega_j}{\min\limits_{1\leq j\leq M}\omega_j}||\boldsymbol{\Delta}'_{J(\boldsymbol{\beta})}||_1.$$

Thus, under Assumption $\mathbf{RE}(s, a_0)$, with

$$a_0 = \left(3 + \frac{4}{\zeta}\right)\frac{\max\limits_{1\leq j\leq M}\omega_j}{\min\limits_{1\leq j\leq M}\omega_j} \text{ and } \boldsymbol{\kappa} = \boldsymbol{\kappa}(s, a_0),$$

we infer that

$$\boldsymbol{\kappa}^2||\boldsymbol{\Delta}'_{J(\boldsymbol{\beta})}||_2^2 \leq \boldsymbol{\Delta}'^T\boldsymbol{G_n}\boldsymbol{\Delta}'$$

with

$$\begin{aligned}
\boldsymbol{\Delta}'^T\boldsymbol{G_n}\boldsymbol{\Delta}' &= \frac{1}{n}\sum_{i=1}^{n}((f_{\hat{\beta}_L} - f_\beta)(\boldsymbol{Z_i}))^2\Lambda_i(\tau) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\log(\alpha_0(t)e^{f_{\hat{\beta}_L}(\boldsymbol{Z_i})}) - \log(\alpha_0(t)e^{f_{\hat{\beta}}(\boldsymbol{Z_i})})\right)^2 d\Lambda_i(t) \\
&= ||\log\lambda_{\hat{\beta}_L} - \log\lambda_\beta||_{n,\Lambda}
\end{aligned}$$

Using the fact that

$$||\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}||_2 \leq \max_{1\leq j\leq M}\omega_j||\boldsymbol{\Delta}'_{J(\boldsymbol{\beta})}||_2,$$

Inequality (62) becomes

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

Now we apply Proposition 2 which compares the empirical Kullback divergence and the weighted empirical norm. It follows that

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

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

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

Hence,

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

and

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

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

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

Finally, taking the infimum over all $\boldsymbol{\beta} \in \mathbb{R}^M$ such that $|J(\boldsymbol{\beta})| \leq s$, we obtain

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

□

6.6 Proof of Corollary 1

To prove the corollary, it suffices to use Proposition 2 and to rewrite the previous proof with

$$b = \frac{\mu'(1+\zeta) + \mu''}{\mu'(1+\zeta) - \mu''}.$$

□

6.7 Proof of Theorem 3

We start to prove Inequality (25) of Theorem 3. In (60), we take $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Consequently $\widetilde{K}_n(\lambda_0, \lambda_\beta) = 0$ and by applying Proposition 2 with $\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t)e^{\boldsymbol{\beta_0^T Z_i}}$ and $\lambda_{\hat{\beta}_L}(t, \boldsymbol{Z_i}) = \alpha_0(t)e^{\boldsymbol{\hat{\beta}_L^T Z_i}}$, we obtain that, on $\mathcal{A}$

$$\mu' ||\boldsymbol{\hat{\beta}_L^T Z_i} - \boldsymbol{\beta_0^T Z_i}||_{n,\Lambda}^2 + \sum_{j=1}^{p} \frac{\omega_j}{2} |\hat{\beta}_L - \beta_0|_j \leq 2 \sum_{j \in J_0} \omega_j |\hat{\beta}_L - \beta_0|_j. \tag{65}$$

From this inequality, we deduce two other inequalities. The first one is obtained by noting that $||\boldsymbol{\hat{\beta}_L^T Z_i} - \boldsymbol{\beta_0^T Z_i}||_{n,\Lambda}^2 = ||\boldsymbol{X\Delta_0}||_{n,\Lambda}^2$, with $\boldsymbol{X} = (Z_{i,j})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ and $\boldsymbol{\Delta_0} = \boldsymbol{\hat{\beta}_L} - \boldsymbol{\beta_0}$,

$$\mu' ||\boldsymbol{X\Delta_0}||_{n,\Lambda}^2 \leq 2 \sum_{j \in J_0} \omega_j |\hat{\beta}_L - \beta_0|_j \leq 2\sqrt{|J_0|} \max_{1 \leq j \leq p} \omega_j ||\boldsymbol{\Delta_{0J_0}}||_2, \tag{66}$$

where $J_0 = J(\boldsymbol{\beta_0})$. From (65), 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

$$\min_{1 \leq j \leq p} \omega_j ||\boldsymbol{\Delta_0}||_1 \leq 4 \max_{1 \leq j \leq p} \omega_j ||\boldsymbol{\Delta_{0J_0}}||_1.$$

We then split $||\boldsymbol{\Delta_0}||_1 = ||\boldsymbol{\Delta_{0J_0}}||_1 + ||\boldsymbol{\Delta_{0J_0^c}}||_1$ to get

$$||\boldsymbol{\Delta_{0J_0^c}}||_1 \leq \Big(4 \frac{\max_{1 \leq j \leq p} \omega_j}{\min_{1 \leq j \leq p} \omega_j} - 1 \Big) ||\boldsymbol{\Delta_{0J_0}}||_1. \tag{67}$$

Set $b_0 := 4 \dfrac{\max_{1 \leq j \leq p} \omega_j}{\min_{1 \leq j \leq p} \omega_j} - 1$, and apply Assumption $\mathbf{RE}(s, a_0)$ to write

$$||\boldsymbol{X\Delta_0}||_{n,\Lambda}^2 \geq \boldsymbol{\kappa'}^2 ||\boldsymbol{\Delta_{0J_0}}||_2^2, \tag{68}$$

where $\boldsymbol{\kappa'} = \boldsymbol{\kappa'}(s, b_0)$. According to (66), we conclude that

$$\mu' ||\boldsymbol{X\Delta_0}||_{n,\Lambda}^2 \leq 2\sqrt{|J_0|} \max_{1 \leq j \leq p} \omega_j \frac{||\boldsymbol{X\Delta_0}||_{n,\Lambda}}{\boldsymbol{\kappa'}},$$

which entails that

$$||\boldsymbol{X}(\boldsymbol{\hat{\beta}_L} - \boldsymbol{\beta_0})||_{n,\Lambda}^2 \leq \frac{4|J_0|}{\mu'^2 \boldsymbol{\kappa'}^2} (\max_{1 \leq j \leq p} \omega_j)^2.$$

Let us come to the proof of Inequality (26) in Theorem 3. Combine Inequality (66) and Assumption $\mathbf{RE}(s, b_0)$ to write

$$\mu' \boldsymbol{\kappa'}^2 ||\boldsymbol{\Delta_{0J_0}}||_2^2 \leq 2\sqrt{|J_0|} \max_{1 \leq j \leq p} \omega_j ||\boldsymbol{\Delta_{0J_0}}||_2,$$

and hence

$$||\boldsymbol{\Delta_0}_{J_0}||_2 \leq \frac{2\sqrt{|J_0|}}{\mu' {\boldsymbol{\kappa}'}^2} \max_{1\leq j\leq p} \omega_j. \tag{69}$$

According to (67) and thanks to Cauchy-Schwarz Inequality, we have

$$||\boldsymbol{\Delta_0}||_1 = ||\boldsymbol{\Delta_0}_{J_0}||_1 + ||\boldsymbol{\Delta_0}_{J_0^c}||_1 \leq (1+b_0)||\boldsymbol{\Delta_0}_{J_0}||_1 \leq (1+b_0)\sqrt{|J_0|}||\boldsymbol{\Delta_0}_{J_0}||_2,$$

where

$$b_0 = 4\frac{\max\limits_{1\leq j\leq p} \omega_j}{\min\limits_{1\leq j\leq p} \omega_j} - 1.$$

From (69), we get

$$\frac{||\boldsymbol{\Delta_0}||_1}{(1+b_0)\sqrt{|J_0|}} \leq \frac{2\sqrt{|J_0|}}{\mu' {\boldsymbol{\kappa}'}^2} \max_{1\leq j\leq p} \omega_j,$$

and finally

$$||\boldsymbol{\Delta_0}||_1 \leq 8\frac{\max\limits_{1\leq j\leq p} \omega_j}{\min\limits_{1\leq j\leq p} \omega_j} \frac{|J_0|}{\mu' {\boldsymbol{\kappa}'}^2} \max_{1\leq j\leq p} \omega_j.$$

□

6.8 Proof of Theorem 4

The proof is very similar to the one of Theorem 1. We start from (31), (32) and (33), and write

$$\begin{aligned}\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) &+ (\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma})^T \boldsymbol{\nu_{n,\tau}} + \mathrm{pen}(\boldsymbol{\gamma}) - \mathrm{pen}(\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) \\ &+ (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta_{n,\tau}} + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}).\end{aligned} \tag{70}$$

We apply Theorem 6 to events $\mathcal{A}$ and $\mathcal{B}$ defined by

$$\mathcal{A} = \bigcap_{j=1}^{M} \left\{ |\eta_{n,\tau}(f_j)| \leq \frac{\omega_j}{2} \right\} \text{ and } \mathcal{B} = \bigcap_{k=1}^{N} \left\{ |\nu_{n,\tau}(\theta_k)| \leq \frac{\delta_k}{2} \right\},$$

to control the terms $(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta_{n,\tau}}$ and $(\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma})^T \boldsymbol{\nu_{n,\tau}}$. We have respectively from (34) and (35) that

$$\mathbb{P}(\mathcal{A}^c) \leq c_{3,\varepsilon,c_\ell} e^{-x} \text{ and } \mathbb{P}(\mathcal{B}^c) \leq c'_{3,\varepsilon',c'_\ell} e^{-y}.$$

Hence

$$\mathbb{P}((\mathcal{A}\cap\mathcal{B})^c) = \mathbb{P}(\mathcal{A}^c \cup \mathcal{B}^c) \leq \mathbb{P}(\mathcal{A}^c) + \mathbb{P}(\mathcal{B}^c) \leq c_{3,\varepsilon,c_\ell} e^{-x} + c'_{3,\varepsilon',c_\ell} e^{-y} \leq Be^{-z},$$

with $B = c_{3,\varepsilon,c_\ell} + c'_{3,\varepsilon',c_\ell}$ and $z = \min\{x, y\} > 0$ fixed. On $\mathcal{A}\cap\mathcal{B}$ arguing as in the proof of Theorem 1, with probability larger than $1 - Be^{-z}$, we have

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 2\,\mathrm{pen}(\boldsymbol{\beta}) + 2\,\mathrm{pen}(\boldsymbol{\gamma}).$$

Theorem 4 is thus proved. □

6.9 Proof of Theorem 5

We start from Inequality (70). On $\mathcal{A} \cap \mathcal{B}$

$$|(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}}| \leq \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}_L - \beta)_j| \text{ and } |(\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma})^T \boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}}| \leq \sum_{k=1}^{N} \frac{\delta_k}{2} |(\hat{\gamma}_L - \gamma)_k|,$$

and therefore

$$\begin{aligned} \widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) + \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}_L - \beta)_j| + \sum_{k=1}^{N} \frac{\delta_k}{2} |(\hat{\gamma}_L - \gamma)_k| \\ \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 2 \sum_{j \in J(\boldsymbol{\beta})} \omega_j |(\hat{\beta}_L - \beta)_j| + 2 \sum_{k \in J(\boldsymbol{\gamma})} \delta_k |(\hat{\gamma}_L - \gamma)_k|. \end{aligned} \tag{71}$$

We then apply Cauchy-Schwarz inequality to the second right-term of this inequality and obtain

$$\begin{aligned} \widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) + \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}_L - \beta)_j| + \sum_{k=1}^{N} \frac{\delta_k}{2} |(\hat{\gamma}_L - \gamma)_k| \\ \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 2\sqrt{|J(\boldsymbol{\beta})|} \sqrt{\sum_{j \in J(\beta)} \omega_j^2 |\hat{\beta}_L - \beta|_j^2} + 2\sqrt{|J(\boldsymbol{\gamma})|} \sqrt{\sum_{k \in J(\boldsymbol{\gamma})} \delta_k^2 |\hat{\gamma}_L - \gamma|_k^2}. \end{aligned} \tag{72}$$

If we set $\tilde{\boldsymbol{\Delta}} = \tilde{\boldsymbol{D}} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma} \end{pmatrix}$ and $\tilde{\boldsymbol{D}} = (\text{diag}(\omega_1, ..., \omega_M, \delta_1, ..., \delta_N))$, Inequality (71) is rewritten as :

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) + \frac{1}{2} ||\tilde{\boldsymbol{\Delta}}||_1 \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 2||\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_1, \tag{73}$$

where $\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})} = \tilde{\boldsymbol{D}} \begin{pmatrix} (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})_{J(\boldsymbol{\beta})} \\ (\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma})_{J(\boldsymbol{\gamma})} \end{pmatrix}$. In the same way, Inequality (72) becomes :

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 4 \max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right) ||\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_2. \tag{74}$$

We then consider two cases

$$2||\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_1 \leq \zeta \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) \tag{75}$$

and

$$\zeta \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) \leq 2||\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_1. \tag{76}$$

In Case (75), Inequality (29) in Theorem 5 follows immediately from (73). In Case (76), let us denote by $\mathcal{A}_1$ the event

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

On $\mathcal{A} \cap \mathcal{B} \cap \mathcal{A}_1$, we deduce from (73) that

$$||\tilde{\boldsymbol{\Delta}}||_1 \leq 4 \left(1 + \frac{2}{\zeta} \max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)\right) ||\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_1.$$

By splitting $\tilde{\boldsymbol{\Delta}} = \tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})} + \tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c,J(\boldsymbol{\gamma})^c}$, we infer that

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

If $\tilde{\boldsymbol{\Delta}}' = \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{\Delta}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma}\end{pmatrix}$, then

$$||\tilde{\boldsymbol{\Delta}}'_{J(\boldsymbol{\beta})^c,J(\boldsymbol{\gamma})^c}||_1 \leq \left(3 + \frac{8}{\zeta}\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)\right)\frac{\max\limits_{1\leq j\leq M, 1\leq k\leq N}\{\omega_j, \delta_k\}}{\min\limits_{1\leq j\leq M, 1\leq k\leq N}\{\omega_j, \delta_k\}}||\tilde{\boldsymbol{\Delta}}'_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_1.$$

Now, we apply $\mathbf{RE}(s, r_0)$, with

$$r_0 = \left(3 + \frac{8}{\zeta}\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)\right)\frac{\max\limits_{1\leq j\leq M, 1\leq k\leq N}\{\omega_j, \delta_k\}}{\min\limits_{1\leq j\leq M, 1\leq k\leq N}\{\omega_j, \delta_k\}},$$

to get that

$$\tilde{\boldsymbol{\kappa}}^2||\tilde{\boldsymbol{\Delta}}'_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_2^2 \leq \tilde{\boldsymbol{\Delta}}'^{\boldsymbol{T}}\tilde{\boldsymbol{G}}_{\boldsymbol{n}}\tilde{\boldsymbol{\Delta}}'$$

with

$$\tilde{\boldsymbol{\Delta}}'^{\boldsymbol{T}}\tilde{\boldsymbol{G}}_{\boldsymbol{n}}\tilde{\boldsymbol{\Delta}}' = ||\log\lambda_{\hat{\beta}_L,\hat{\gamma}_L} - \log\lambda_{\beta,\gamma}||_{n,\Lambda}^2$$

and $\tilde{\boldsymbol{\kappa}} = \tilde{\boldsymbol{\kappa}}(s, r_0)$. Since

$$||\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_2 \leq \max_{\substack{1\leq j\leq M\\1\leq k\leq N}}\{\omega_j, \delta_k\}||\tilde{\boldsymbol{\Delta}}'_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_2,$$

Equation (74) becomes

$$\begin{aligned}\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) &\leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 4\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\gamma)|}\right)||\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}||_2 \\ &\leq \widetilde{K}_n(\lambda_0, \lambda_\beta) + 4\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)\max_{\substack{1\leq j\leq M\\1\leq k\leq N}}\{\omega_j, \delta_k\}\tilde{\boldsymbol{\kappa}}^{-1}||\log\lambda_{\hat{\beta}_L,\hat{\gamma}_L} - \log\lambda_{\beta,\gamma}||_{n,\Lambda}.\end{aligned}$$

Since $||\log\lambda_{\hat{\beta}_L,\hat{\gamma}_L} - \log\lambda_{\beta,\gamma}||_{n,\Lambda} \leq ||\log\lambda_{\hat{\beta}_L,\hat{\gamma}_L} - \log\lambda_0||_{n,\Lambda} + ||\log\lambda_0 - \log\lambda_{\beta,\gamma}||_{n,\Lambda}$, we obtain that

$$\begin{aligned}&\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) \\ &\quad\leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) \\ &\quad+ 4\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)\max_{\substack{1\leq j\leq M\\1\leq k\leq N}}\{\omega_j, \delta_k\}\tilde{\boldsymbol{\kappa}}^{-1}(||\log\lambda_{\hat{\beta}_L,\hat{\gamma}_L} - \log\lambda_0||_{n,\Lambda} + ||\log\lambda_0 - \log\lambda_{\beta,\gamma}||_{n,\Lambda}).\end{aligned}$$

We now apply Proposition 2 and write

$$\begin{aligned}&\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L})\\ &\leq \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 4\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right) \max_{\substack{1\leq j\leq M\\ 1\leq k\leq N}} \{\omega_j, \delta_k\} \frac{\tilde{\boldsymbol{\kappa}}^{-1}}{\sqrt{\mu'}} \left(\sqrt{\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L})} + \sqrt{\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma})}\right).\end{aligned}$$

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

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

Hence,

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

and

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) \leq \frac{b+1}{b-1}\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + 8\frac{b^2}{b-1}\max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|) \Big(\max_{\substack{1\leq j\leq M\\ 1\leq k\leq N}} \{\omega_j, \delta_k\}\Big)^2 \frac{\tilde{\boldsymbol{\kappa}}^{-2}}{\mu'}. \tag{77}$$

We take $b = 1 + \dfrac{2}{\zeta}$ and we introduce $\widetilde{C}(\zeta, \mu') = 8\dfrac{b^2}{\mu'(b-1)}$ a constant depending on ζ and μ'. For all $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in $\mathbb{R}^M \times \mathbb{R}^N$, we obtain

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) \leq (1+\zeta)\Big\{\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + \widetilde{C}(\zeta, \mu')\max(|J(\boldsymbol{\beta})|, |J(\gamma)|) \Big(\max_{\substack{1\leq j\leq M\\ 1\leq k\leq N}} \{\omega_j, \delta_k\}\Big)^2 \tilde{\boldsymbol{\kappa}}^{-2}\Big\}.$$

Finally, taking the infimum over all $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N$ such that $\max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|) \leq s$, we obtain Inequality (29)

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L,\hat{\gamma}_L}) \leq (1+\zeta) \inf_{\substack{\boldsymbol{\beta}\in\mathbb{R}^M, \boldsymbol{\gamma}\in\mathbb{R}^N\\ \max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|)\leq s}} \Big\{\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + \widetilde{C}(\zeta, \mu')\max(|J(\boldsymbol{\beta})|, |J(\gamma)|) \Big(\max_{\substack{1\leq j\leq M\\ 1\leq k\leq N}} \{\omega_j, \delta_k\}\Big)^2 \tilde{\boldsymbol{\kappa}}^{-2}\Big\}.$$

To prove Inequality (30), it suffices to apply Proposition 2 and to take $b = \dfrac{(1+\zeta)\mu' + \mu''}{(1+\zeta)\mu' - \mu''}$ in (77). □

References

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

[2] Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, Niels. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993.

[3] Antoniadis, A., Fryzlewicz, P., and Letué, F. The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics*, 37(4):pp. 531–552, 2010.

[4] Bach, F. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:pp. 384–414, 2010.

[5] Bertin, K., Le Pennec, E., and Rivoirard, V. Adaptive Dantzig density estimation. *Annales de l'IHP, Probabilités et Statistiques*, 47(1):pp. 43–74, 2011.

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

[7] Bradic, J., Fan, J., and Jiang, J. Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics*, 39(6):pp. 3092–3120, 2012.

[8] Bühlmann, P. and van de Geer, S. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:pp. 1360–1392, 2009.

[9] Bunea, F., Tsybakov, A. B., and Wegkamp, M. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:pp. 169–194, 2007.

[10] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. Aggregation and sparsity via l1 penalized least squares. In *Proceedings of the 19th annual conference on Learning Theory*, COLT'06, pages 379–391, Berlin, Heidelberg, 2006. Springer-Verlag.

[11] Bunea, F., Tsybakov, A.B., Wegkamp, M.H., and Barbu, A. Spades and mixture models. *The Annals of Statistics*, 38(4):pp. 2525–2558, 2010.

[12] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B. (Methodological)*, 34:pp. 187–220, 1972.

[13] Dave, S. S., Wright, G., Tan, B., Rosenwald, A., Gascoyne, R. D., Chan, W. C., Fisher, R. I., Braziel, R. M., Rimsza, L. M., Grogan, T. M., Miller, T. P., LeBlanc, M., Greiner, T. C., Weisenburger, D. D., Lynch, J. C., Vose, J., Armitage, J. O., Smeland, E. B., Kvaloy, S., Holte, H., Delabie, J., Connors, J. M., Lansdorp, P. M., Ouyang, Q., Lister, T. A., Davies, A. J., Norton, A. J., Muller-Hermelink, H. K., Ott, G., Campo, E., Montserrat, E., Wilson, W.

H., Jaffe, E. S., Simon, R., Yang, L., Powell, J., Zhao, H., Goldschmidt, N., Chiorazzi, M., and Staudt, L. M. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New England Journal of Medicine*, 351(21):pp. 2159–2169, 2004.

[14] Gaïffas, S. and Guilloux, A. High-dimensional additive hazard models and the Lasso. *Electronic Journal of Statistics*, 6:pp. 522–546, 2011.

[15] Gourlay, M. L., Fine, J. P., Preisser, J. S., May, R. C., Li, C., Lui, LY., Ransohoff, D. F., Cauley, J. A., and Ensrud, K. E. Bone-density testing interval and transition to osteoporosis in older women. *New England Journal of Medicine*, 366(3):pp. 225–233, 2012.

[16] Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. Lasso and probabilistic inequalities for multivariate point processes. Work in progress, personnal communication.

[17] Kong, S. and Nan, B. Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Arxiv preprint arXiv:1204.1992*, 2012.

[18] Massart, P. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[19] Massart, P. and Meynet, C. The Lasso as an l1-ball model selection procedure. *Electronic Journal of Statistics*, 5:pp. 669–687, 2011.

[20] Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):pp. 1436–1462, 2006.

[21] Rigollet, P. Kullback-Leibler aggregation and misspecified generalized linear models. *Arxiv preprint arXiv:0911.2919*, 2009.

[22] Senoussi, R. Problème d'identification dans le modèle de Cox. *Annales de l'Institut Henri Poincaré*, 26:pp. 45–64, 1988.

[23] Steyerberg, E. W., Homs, M. Y. V., Stokvis, A., Essink-Bot, ML., and Siersema, P. D. Stent placement or brachytherapy for palliation of dysphagia from esophageal cancer: a prognostic model to guide treatment selection. *Gastrointestinal Endoscopy*, 62(3):pp. 333–340, 2005.

[24] Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996.

[25] Tibshirani, R. The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):pp. 385–395, 1997.

[26] Uspensky, J. V. *Introduction to mathematical probability.* New York: McGraw-Hill , 1937.

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

[28] van de Geer, S. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):pp. 614–645, 2008.

[29] Zhang, C.H. and Huang, J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):pp. 1567–1594, 2008.

[30] Zhang, H. H. and Lu, W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):pp. 691–703, 2007.

[31] Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:pp. 1081–1107, 2010.

[32] Zhao, P. and Yu, B. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7(2):pp. 2541, 2007.

[33] Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):pp. 1418–1429, 2006.