Learning and adaptive estimation for marker-dependent counting processes

Stéphane Gaïffas, Agathe Guilloux

Learning and adaptive estimation for marker-dependent counting processes

Stéphane Gaïffas[1] Agathe Guilloux[1,2]

September 29, 2009

Abstract

We consider the problem of statistical learning for the intensity of a counting process with covariates. In this context, we introduce an empirical risk, and prove risk bounds for the corresponding empirical risk minimizers. Then, we give an oracle inequality for the popular algorithm of aggregation with exponential weights. This provides a way of constructing estimators that are adaptive to the smoothness and to the structure of the intensity. We prove that these estimators are adaptive over anisotropic Besov balls. The probabilistic tools are maximal inequalities using the generic chaining mechanism, which was introduced by Talagrand (2005), together with Bernstein's inequality for the underlying martingales.

Keywords. Counting processes, Statistical learning, Adaptive estimation, Empirical risk minimization, Aggregation with exponential weights, Generic chaining

1 Introduction

Over the last decade, statistical learning theory (initiated by Vapnik, see for instance Vapnik (2000)) has known a tremendous amount of mathematical developments. By mathematical developments, we mean risk bounds for learning algorithms, such as empirical risk minimization, penalization or aggregation. However, in the vast majority of papers, such bounds are derived in the context of regression, density or classification. In the regression model, one observes independent copies of (X, Y), where X is an input, or a covariate, and Y is a real output, or label. The aim is then to infer on $\mathsf{E}(Y|X)$. The aim of this paper is to study the same learning algorithms (such as empirical risk minimization) in a more sophisticated setting, where the output is not a real number, but a stochastic process. Namely, we focus on the situation where, roughly, the output is a counting process, which has an intensity that depends on the covariate X. The aim is then to infer on this intensity. This

[1]Université Pierre et Marie Curie - Paris 6, Laboratoire de Statistique Théorique et Appliquée, 175 rue du Chevaleret, 75013 PARIS. Supported in part by ANR Grant "Prognostic"
Email: stephane.gaiffas@upmc.fr
[2]Université Pierre et Marie Curie - Paris 6, Unité INSERM 762 "Instabilité des Microsatellites et Cancers"
Email: agathe.guilloux@upmc.fr

framework contains many models, that are of importance in practical situations, such as in medicine, actuarial science or econometrics, see Andersen et al. (1993).

In this paper, we give risk bounds for empirical risk minimization and aggregation algorithms. In summary, we try to "find back" the kind of results one usually has in more "standard" models (see below for references). Then, as an application of these results, we construct estimators that have the property to adapt to the smoothness and to the structure of the intensity (in the context of a single-index model). Several papers work in a setting close to ours. Model selection has been first studied in Reynaud-Bouret (2003) for the non-conditional intensity of a Poisson process, see also Reynaud-Bouret (2006), Birge (2007), Baraud and Birgé (2009) and Brunel and Comte (2005). Model selection for the same problem as the one considered here has been studied in Comte et al. (2008).

The agenda of the paper is the following. In this Section, we describe the general setting and the corresponding estimation problem. Section 1.2 is devoted to a presentation of the main examples embedded in this setting. The main objects (such as the empirical risk) and the basic deviation inequalities are described in Section 2. In Section 3, we give risk bounds for the empirical risk minimization (ERM) algorithm. To that end, we provide useful uniform deviation inequalities using the generic chaining mechanism introduced in Talagrand (2005) (see Theorem 1 and Corollary 1), and we give a general risk bound for the ERM in Theorem 3 and its Corollary 2. In Section 4, we adapt a popular aggregation algorithm (aggregation with exponential weights) to our setup, and give an oracle inequality (see Theorem 4). In Section 5, we use the results from Sections 3 and 4 to construct estimators that adapt to the smoothness and to the structure of the intensity. We compute the convergence rates of the estimators, that are minimax optimal over anisotropic Besov balls. Section 6 contains the proofs. Some useful results and tools are recalled in the Appendices.

1.1 The model

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ a filtration satisfying the usual conditions, see Jacod and Shiryaev (1987). Let N be a marked counting process with compensator Λ with respect to $(\mathcal{F}_t)_{t\geq 0}$, so that $M = N - \Lambda$ is a $(\mathcal{F}_t)_{t\geq 0}$-martingale. We assume that N is a marked point process satisfying the *Aalen multiplicative intensity model*. This means that Λ writes

$$\Lambda(t) = \int_0^t \alpha_0(u, X)Y(u)du \tag{1}$$

for all $t \geq 0$, where:

- α_0 is an unknown deterministic and nonnegative function called *intensity*;
- $X \in \mathbb{R}^d$ is a $\mathcal{F}_0$-measurable random vector called *covariates* or *marks*;
- Y is a predictable random process in $[0, 1]$.

With differential notations, this model can be written has

$$dN(t) = \alpha_0(t, X)Y(t)dt + dM(t) \tag{2}$$

for all $t \geq 0$ with the same notations as before, and taking $N(0) = 0$. Now, assume that we observe n i.i.d. copies

$$D_n = \{(X_i, N^i(t), Y^i(t)) : t \in [0,1], 1 \leq i \leq n\} \tag{3}$$

of $\{(X, N(t), Y(t)) : t \in [0,1]\}$. This means that we can write

$$dN^i(t) = \alpha_0(t, X_i)Y^i(t)dt + dM^i(t)$$

for any $i = 1, \ldots, n$ where M^i are independent $(\mathcal{F}_t)_{t\geq 0}$-martingales. In this setting, the random variable $N^i(t)$ is the number of observed failures during the time interval $[0,t]$ of the individual i.

The aim of the paper is to recover the intensity α_0 on $[0,1]$ based on the observation of the sample D_n. This general setting includes several specific problems where the estimation of α_0 is of importance for practical applications, see Section 1.2. In all what follows, we assume that the support of P_X is compact, but in order to simplify the presentation, we shall assume the following.

Assumption 1. *The support of P_X is $[0,1]^d$, and*

$$\|\alpha\|_\infty := \sup_{(t,x)\in[0,1]^{d+1}} |\alpha(t,x)| \tag{4}$$

is finite.

These assumptions on the model are very mild, excepted for the i.i.d assumption of the sample, meaning that the individuals i are independent. Let us give several examples of interest that fit in this general setting.

1.2 Examples

1.2.1 Regression model for right-censored data

Let T be a nonnegative random variable (r.v.) and X a vector of covariates in $\mathbb{R}^d$. In this model, T is not directly observable: what we observe instead is

$$T^C := \min(T, C) \text{ and } \delta := I(T \leq C), \tag{5}$$

where C is a nonnegative random variable called *censoring*. This setting, where the data is *right censored*, is of first importance in applications, especially in medicine, biology and econometrics. In these cases, the r.v. T can represent the lifetime of an individual, the time from the the onset of a disease to the healing, the duration of unemployment, etc. The r.v. C is often the time of last contact or the duration of follow-up. In this model we assume the following mild assumption:

$$T \text{ and } C \text{ are independent conditionally to } X, \tag{6}$$

which allows the censoring to depend on the covariates, see Heuchenne and Van Keilegom (2007). This assumption is weaker than the more common assumption that T and C are independent, see in particular Stute (1996).

In this case, the counting process writes

$$N^i(t) = I(T_i^C \leq t, \delta_i = 1) \text{ and } Y^i : Y^i(t) = I(T_i^C \geq t),$$

see e.g. Andersen et al. (1993). In this setting, the intensity α_0 is the conditional hazard rate of T given $X = x$, which is defined for all $t > 0$ and $x \in \mathbb{R}^d$ by

$$\alpha_0(t,x) = \alpha_{T|X}(t,x) = \frac{f_{T|X}(t,x)}{1 - F_{T|X}(t,x)},$$

where $f_{T|X}$ and $F_{T|X}$ are the conditional probability density function (p.d.f.) and the conditional distribution function (d.f.) of T given X respectively. The available data in this setting becomes

$$D_n := [(X_i, T_i^C, \delta_i) : 1 \leq i \leq n],$$

where (X_i, T_i^C, δ_i) are i.i.d. copies of (X, T^C, δ), where we assumed (6), namely T_i and C_i are independent conditionally to X_i for $1 \leq i \leq n$.

The nonparametric estimation of the hazard rate was initiated by Beran (1981), Stute (1986), Dabrowska (1987), McKeague and Utikal (1990) and Li and Doss (1995) extended his results. Many authors have considered semiparametric estimation of the hazard rate, beginning with Cox (1972), see Andersen et al. (1993) for a review of the enormous literature on semiparemetric models. We refer to Huang (1999) and Linton et al. (2003) for some recent developments. As far as we know, adaptive nonparametric estimation for censored data in presence of covariates has only been considered in Brunel et al. (2007), who constructed an optimal adaptive estimator of the conditional density.

1.2.2 Cox processes

Let $\eta^i, 1 \leq i \leq n$, be n independent Cox processes on $\mathbb{R}_+$, with mean-measure A^i given by :

$$A^i(t) = \int_0^t \alpha(s, X_i) ds,$$

where X_i is a vector of covariates in $\mathbb{R}^d$. This is a particular case of longitudinal data, see e.g. Example VII.2.15 in Andersen et al. (1993). The nonparametric estimation of the intensity of Poisson processes without covariates has been considered in several papers. We refer to Reynaud-Bouret (2003) for the adaptive estimation (using model selection) for the intensity of nonhomogeneous Poisson processes in a general space

1.2.3 Regression model for transition intensities of Markov processes

Consider a n-sample of nonhomogeneous time-continuous Markov processes $P^1, \ldots, P^n$ with finite state space $\{1, \ldots, k\}$ and denote by λ_{jl} the transition intensity from state j to state l. For an individual i with covariate X_i, the r.v. $N_{jl}^i(t)$ counts the number of observed direct transitions from j to l before time t (we allow the possibility of right-censoring for example). Conditionally on the initial state, the counting process N_{jl}^i verifies the following Aalen multiplicative intensity model:

$$N_{jl}^i(t) = \int_0^t \lambda_{jl}(X_i, z) Y_j^i(z) dz + M^i(t) \text{ for all } t \geq 0,$$

where $Y_j^i(t) = I(P^i(t-) = j)$ for all $t \geq 0$, see Andersen et al. (1993) or Jacobsen (1982). This setting is discussed in Andersen et al. (1993), see Example VII.11 on mortality and nephropathy for insulin dependent diabetics.

We finally cite three papers, where the estimation of the intensity of counting processes was considered, gathering as a consequence all the previous examples, but in none of them the presence of covariates was considered. Ramlau-Hansen (1983) proposed a kernel-type estimator, Grégoire (1993) studied least squares cross-validation. More recently, Reynaud-Bouret (2006) considered adaptive estimation by projection and Baraud and Birgé (2009) considered the adaptive estimation of the intensity of a random measure by histogram-type estimators.

1.3 Some notations

From now on, we will denote by L an absolute constant that can vary from place to place (even in the same line), and by c a constant that depends on some parameters, that we shall indicate into subscripts. In all of what follows, D_n is an i.i.d. sample satisfying model (2), and we take $(X,(Y_t),(N_t))$ independent of D_n that satisfies also model (2). Note that we will use both notations $(Z_t)_{t\geq 0}$ and $(Z(t))_{t\geq 0}$ for a stochastic process Z. We denote by $\mathsf{P}^n[\cdot]$ the joint law of D_n and $\mathsf{P}[\cdot]$ the law of $(X,(Y_t),(N_t))$, and by $\mathsf{E}^n[\cdot]$ and $\mathsf{E}[\cdot]$ the corresponding expectations.

2 Main constructions and objects

2.1 An empirical risk

Let $x\in\mathbb{R}^d$ and (y_t), (n_t) be functions $[0,1]\to\mathbb{R}^+$ with bounded variations, and let $\alpha:[0,1]^{d+1}\to\mathbb{R}^+$ be a bounded and predictable function (that can eventually depend on D_n). We define the loss function

$$\ell_\alpha(x,(y_t),(n_t))=\int_0^1\alpha(t,x)^2y(t)dt-2\int_0^1\alpha(t,x)dn(t).$$

We define the least-squares type *empirical risk* of α as:

$$\begin{aligned}P_n(\ell_\alpha)&:=\frac{1}{n}\sum_{i=1}^n\ell_\alpha(X_i,(Y_t^i),(N_t^i))\\&=\frac{1}{n}\sum_{i=1}^n\int_0^1\alpha(t,X_i)^2Y^i(t)dt-\frac{2}{n}\sum_{i=1}^n\int_0^1\alpha(t,X_i)dN^i(t).\end{aligned}\tag{7}$$

This quantity measures the goodness-of-fit of α to the data from D_n. It has been used in Comte et al. (2008) to perform model selection. It is the empirical version of the *theoretical risk*

$$\begin{aligned}P(\ell_\alpha)&:=\mathsf{E}[\ell_\alpha(X,(Y_t),(N_t))\,|\,D_n]\\&=\mathsf{E}\Big[\int_0^1\alpha(t,X)^2Y(t)dt-2\int_0^1\alpha(t,X)dN(t)\,\Big|\,D_n\Big].\end{aligned}$$

This risk is natural in this model. Indeed, if α is independent of D_n, we have in view of (2), since $M(t)$ is centered:

$$P(\ell_\alpha) = \mathsf{E}\Big[\int_0^1 (\alpha(t,X)^2 - 2\alpha(t,X)\alpha_0(t,X))Y(t)dt\Big] - 2\mathsf{E}\Big[\int_0^1 \alpha(t,X)dM(t)\Big]$$
$$= \|\alpha\|^2 - 2\langle\alpha,\alpha_0\rangle \tag{8}$$
$$= \|\alpha-\alpha_0\|^2 - \|\alpha_0\|^2, \tag{9}$$

where:

$$\langle\alpha,\alpha_0\rangle := \int_{\mathbb{R}^d}\int_0^1 \alpha(t,x)\alpha_0(t,x)\mathsf{E}[Y(t)|X=x]dtP_X(dx), \quad \|\alpha\|^2 := \langle\alpha,\alpha\rangle. \tag{10}$$

This is an inner product with respect to the bounded measure (it is smaller than 1)

$$d\mu(t,x) := \mathsf{E}[Y(t)|X=x]dtP_X(dx). \tag{11}$$

We will denote by $\mathbb{L}^2(\mu)$ the corresponding Hilbert space, and define $\mathbb{L}^\infty(\mu)$ as the subset of $\mathbb{L}^2(\mu)$ consisting of functions α such that $\|\alpha\|_\infty < +\infty$.

In view of (8), $P(\ell_\alpha) - P(\ell_{\alpha_0})$ (called *excess risk*) is equal to $\|\alpha-\alpha_0\|^2$. As a consequence, α_0 minimizes $\alpha \mapsto P(\ell_\alpha)$, so a natural way to recover α_0 is to take a minimizer of $\alpha \mapsto P_n(\ell_\alpha)$. This is the basic idea of empirical risk minimization, for which we propose risk bounds in Section 3 below. Let us define the *empirical norm*

$$\|\alpha\|_n^2 := \frac{1}{n}\sum_{i=1}^n \int_0^1 \alpha(t,X_i)^2 Y^i(t)dt, \tag{12}$$

so that we have $\mathsf{E}^n\|\alpha\|_n^2 = \|\alpha\|^2$ if α is deterministic. Note that $\|\alpha\|_n \le \|\alpha\|_\infty$ and $\|\alpha\| \le \|\alpha\|_\infty$. An important fact is that (2) entails

$$P_n(\ell_\alpha) - P_n(\ell_{\alpha_0}) = \|\alpha-\alpha_0\|_n^2 - \frac{2}{\sqrt{n}}Z_n(\alpha-\alpha_0), \tag{13}$$

where $Z_n(\cdot)$ is given by

$$Z_n(\alpha) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^1 \alpha(t,X_i)dM^i(t), \tag{14}$$

where M^i are the independent copies of the martingale innovation from (2). The decomposition (13) will be of importance in the analysis of the problem.

Remark 1 (Regression model for right-censored data). In the problem of censored survival times with covariates, see Section 1.2.1, the semi-norm of estimation becomes

$$\|\alpha\|^2 = \int\int_0^1 \alpha(t,x)^2 \bar{H}_{T^C|X}(t,x)dtP_X(dx),$$

where $\bar{H}_{T^C|X}(t,x) := \mathsf{P}[T^C > t|X=x]$, and where by (5) and (6):

$$\bar{H}_{T^C|X}(t,x) = \mathsf{P}[T>t|X=x]\mathsf{P}[C>t|X=x].$$

This weighting of the norm is natural and, somehow, unavoidable in models with censored data. The same normalization can be found, for instance, in the Dvoretzky-Kiefer-Wolfowitz concentration inequality for the Kaplan-Meier estimator (without covariates), see Theorem 1 in Bitouzé et al. (1999).

2.2 Deviation inequalities

Let us denote by $\langle Z \rangle$ the predictable variation of a random process Z. Note that we have, using Assumption 1:

$$\langle Z_n(\alpha) \rangle = \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \alpha(t, X_i)^2 \alpha_0(t, X) Y^i(t) dt \leq \|\alpha_0\|_\infty \|\alpha\|_n^2. \tag{15}$$

A useful result is then the following. First, introduce, for $\delta > 0$,

$$\psi_{n,\delta}(h) := \log \mathsf{E}^n [e^{h Z_n(\alpha)} \mathbf{1}_{\langle Z_n(\alpha) \rangle \leq \delta^2}]$$

and the Cramér transform $\psi^*_{n,\delta}(z) := \sup_{h>0}(hz - \psi_{n,\delta}(h))$.

Proposition 1. *For any bounded α and any $z, \delta > 0$, the following inequality holds*:

$$\psi^*_{n,\delta}(z) \geq \frac{n\delta^2}{\|\alpha\|_\infty^2} g\Big(\frac{z\|\alpha\|_\infty}{\delta^2 \sqrt{n}}\Big), \tag{16}$$

where $g(x) := (1+x)\log(1+x) - x$.

This result and the deviation inequalities stated below are related to standard results concerning martingales with jumps, see Liptser and Shiryayev (1989), van de Geer (1995) or Reynaud-Bouret (2006), among others. For the sake of completeness we give a proof of Proposition 1 in Section 6. From the minoration (16), we can derive several deviation inequalities. Using the Cramér-Chernoff bound $\mathsf{P}^n[Z_n(\alpha) > z, \langle Z_n(\alpha) \rangle \leq \delta^2] \leq \exp(-\psi^*_{n,\delta}(z))$, we obtain the following Benett's inequality:

$$\mathsf{P}^n[Z_n(\alpha) > z, \langle Z_n(\alpha) \rangle \leq \delta^2] \leq \exp\Big(- \frac{n\delta^2}{\|\alpha\|_\infty^2} g\Big(\frac{z\|\alpha\|_\infty}{\delta^2\sqrt{n}}\Big)\Big)$$

for any $z > 0$. As a consequence, since $g(x) \geq 3x^2/(2(x+3))$ for any $x \geq 0$, we obtain the following Bernstein's inequality:

$$\mathsf{P}^n[Z_n(\alpha) > z, \langle Z_n(\alpha) \rangle \leq \delta^2] \leq \exp\Big(- \frac{z^2}{2(\delta^2 + z\|\alpha\|_\infty/(3\sqrt{n}))}\Big). \tag{17}$$

Another useful Bernstein's inequality can be derived using the following trick from Birgé and Massart (1998): since $g(x) \geq g_2(x)$ for any $x \geq 0$ where $g_2(x) = x + 1 - \sqrt{1+2x}$, and since $g_2^{-1}(y) = \sqrt{2y} + y$, we have

$$\mathsf{P}^n\Big[Z_n(\alpha) > \delta\sqrt{2x} + \frac{\|\alpha\|_\infty x}{\sqrt{n}},\ \langle Z_n(\alpha) \rangle \leq \delta^2\Big] \leq \exp(-x) \tag{18}$$

for any $x > 0$. Note that from (16), we can derive a uniform deviation inequality. Consider a family $(Z_n(\alpha) : \alpha \in A)$, where A is a set of bounded functions with finite cardinality N. Since $\psi^{*-1}_{n,\delta}(z) \leq z\|\alpha\|_\infty/\sqrt{n} + \delta\sqrt{2z}$ (see above) we have, using Pisier's argument (see Section 2 in Massart (2007)), that

$$\mathsf{P}^n\Big[Z_n(\alpha) > \delta\sqrt{2(\ln N + x)} + \frac{\|\alpha\|_\infty(\ln N + x)}{\sqrt{n}},\ \langle Z_n(\alpha) \rangle \leq \delta^2 \text{ for some } \alpha \in A\Big] \leq \exp(-x). \tag{19}$$

In view of the next Lemma, we can remove the event $\{\langle Z_n(\alpha) \rangle \leq \delta^2\}$ from the previous inequalities. Indeed, a consequence of (18) is the following.

Lemma 1. *If α is bounded, we have for any $x > 0$:*

$$\mathsf{P}^n\Big[Z_n(\alpha) \geq c\|\alpha\|\sqrt{x} + (c+1)\frac{\|\alpha\|_\infty x}{\sqrt{n}}\Big] \leq 2\exp(-x),$$

where $c = c_{\|\alpha_0\|_\infty} := [\sqrt{2}(\sqrt{2}+1)\|\alpha_0\|_\infty]^{1/2}$.

Proof. Since $\mathsf{E}[(\int_0^1 \alpha(t,X)^2 Y(t)dt)^2] \leq \|\alpha^2\|^2$ and $\|\int_0^1 \alpha(t,X)^2 Y(t)dt\|_\infty \leq \|\alpha\|_\infty^2$, Bernstein's inequality for the deviation of the sum of i.i.d. random variables gives:

$$\mathsf{P}^n\Big[\|\alpha\|_n^2 - \|\alpha\|^2 > \frac{\|\alpha^2\|\sqrt{2x}}{\sqrt{n}} + \frac{\|\alpha\|_\infty^2 x}{n}\Big] \leq \exp(-x). \tag{20}$$

Take $\delta_{n,x}^2 := \|\alpha\|^2 + \|\alpha^2\|\sqrt{2x}/\sqrt{n} + \|\alpha\|_\infty^2 x/n$. We have $\mathsf{P}[\|\alpha\|_n^2 > \delta_{n,x}^2] \leq \exp(-x)$ and

$$\delta_{n,x}\sqrt{2\|\alpha_0\|_\infty x} + \frac{\|\alpha\|_\infty x}{\sqrt{n}} \leq c_{\|\alpha_0\|_\infty}\sqrt{x}\|\alpha\| + (c_{\|\alpha_0\|_\infty}+1)\frac{\|\alpha\|_\infty x}{\sqrt{n}}.$$

Now, use (15) and (18) to obtain

$$\mathsf{P}\Big[Z_n(\alpha) \geq \delta_{n,x}\sqrt{2\|\alpha_0\|_\infty x} + \frac{\|\alpha\|_\infty x}{\sqrt{n}}, \|\alpha\|_n^2 \leq \delta_{n,x}^2\Big] \leq \exp(-x)$$

for any $x > 0$. This concludes the proof of the Lemma, by a decomposition over $\{\|\alpha\|_n > \delta_{n,x}\}$ and $\{\|\alpha\|_n \leq \delta_{n,x}\}$. □

These deviation inequalities are the starting point of the proof of risk bounds for the algorithm of empirical risk minimization (ERM). Such a bound is given in Section 3 below, see Theorem 3. It requires a generalization of the bound (19) to a general set A, which is given in Section 3.2.

3 Empirical risk minimization

The very basic idea of empirical risk minimization (ERM) is the following. Since α_0 minimizes the risk $\alpha \mapsto P(\ell_\alpha)$, a natural estimate of α_0 is a minimizer of the empirical risk $\alpha \mapsto P_n(\ell_\alpha)$ over some set of function A, usually called a *sieve*. There is hope that such an empirical minimizer is close to α_0, at least if α_0 is not far from A and if $(P - P_n)(\ell_\alpha)$ is small (more details below). Also known as M-estimation, this algorithm has been studied extensively, see for instance Birgé and Massart (1998), Vapnik (2000), van de Geer (2000), Massart (2007), Bartlett and Mendelson (2006), among many others.

If no minimizer of the empirical risk exists, we can simply consider, as this is usually done in the literature, a ρ-minimizer according to the following definition.

Definition 1 (ρ-ERM). Let $\rho > 0$ be fixed. A ρ-Empirical Risk Minimizer (ρ-ERM) is an estimator $\bar{\alpha}_n \in A$ satisfying

$$P_n(\ell_{\bar{\alpha}_n}) \leq \rho + \inf_{\alpha \in A} P_n(\ell_\alpha),$$

where $P_n(\ell_\alpha)$ is the empirical risk (7).

For what follows, one can take $\rho = 1/n$, since typically, the risk of $\bar{\alpha}_n$ is larger than that. To prove a risk bound for the ERM, one usually needs a deviation inequality for

$$\zeta_n(A) := \sup_{\alpha \in A} (P - P_n)(\ell_\alpha).$$

However, when A is not countable, ζ_n may be not measurable. This is not a problem since we can always consider the outer expectation in the statement of the deviation (see van der Vaart and Wellner (1996)), or simply assume the following.

Assumption 2. *There is a countable subset A' of A such that almost surely,*

$$\sup_{\alpha \in A'} P_n(\ell_\alpha) = \sup_{\alpha \in A} P_n(\ell_\alpha).$$

Moreover, assume that there is $b > 0$ such that $\|\alpha\|_\infty \leq b$ for every $\alpha \in A$.

The map $\alpha \mapsto P_n(\ell_\alpha)$ is continuous over $\mathsf{C}([0,1]^{d+1})$ endowed with the norm $\|\cdot\|_\infty$. So, given that $A \subset \mathsf{C}([0,1]^{d+1})$, the first part of Assumption 2 is met. Note that this embedding holds in the examples considered in Section 5. The second part is rather unpleasant, but mandatory if no extra assumption is made on A, and since an $\mathbb{L}^2$ metric is considered for the estimation of α_0.

From now on, we take α_* such that $P(\ell_{\alpha_*}) = \inf_{\alpha \in A} P(\ell_\alpha)$ (if no such α_* exists, we can simply consider α_* such that $P(\ell_{\alpha_*}) \leq \inf_{\alpha \in A} P(\ell_\alpha) + \rho$). Note that α_* may not be unique at this point, we just pick one of the minimizers. The function α_* is usually called the *target* function, or the *oracle* in learning theory, see Cucker and Smale (2002) for instance.

3.1 Peeling

A common way to prove a risk bound for the ERM uses the idea of *localization* or *peeling* (see for instance Massart (2007), Lemma 4.23 and van de Geer (2000, 2007), among others). The idea presented here is very close to these references. First, do a shift: take $\epsilon > 0$, and use the fact that $\bar{\alpha}_n$ is a ρ-ERM to obtain

$$\begin{aligned} P(\ell_{\bar{\alpha}_n}) - P(\ell_{\alpha_*}) &\leq (1+\epsilon)\rho + P(\ell_{\bar{\alpha}_n}) - P(\ell_{\alpha_*}) - (1+\epsilon)(P_n(\ell_{\bar{\alpha}_n}) - P_n(\ell_{\alpha_*})) \\ &\leq (1+\epsilon)\rho + \xi_{n,\epsilon}(A), \end{aligned}$$

where

$$\xi_{n,\epsilon}(A) := \sup_{\alpha \in A} \Big((1+\epsilon)(P - P_n)(\ell_\alpha - \ell_{\alpha_*}) - \epsilon P(\ell_\alpha - \ell_{\alpha_*}) \Big).$$

Then, for some constants $\delta > 0$ and $q > 1$, decompose the supremum over A into suprema over annuli $A_j(\delta)$, where $A(\delta) = \{\alpha \in A : P(\ell_\alpha) - P(\ell_{\alpha_*}) \leq \delta\}$, and for $j \geq 1$, $A_j(\delta) = \{\alpha \in A : q^j \delta < P(\ell_\alpha) - P(\ell_{\alpha_*}) \leq q^{j+1}\delta\}$. Assume for the moment that there exists an increasing function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ and $\delta_{\min} > 0$ such that for any $x > 0$ and $\delta > \delta_{\min}$, we have with a probability larger than $1 - Le^{-x}$:

$$\sup_{\alpha \in A(\delta)} (P - P_n)(\ell_\alpha - \ell_{\alpha_*}) \leq \frac{\psi(\delta)(1 + \sqrt{x} \vee x)}{\sqrt{n}}. \tag{21}$$

Such an inequality will be proved in Section 3.2 below. It entails that, with a probability larger than $1 - Le^{-x}$:

$$\xi_{n,\epsilon}(A) \leq (1+\epsilon)\frac{\psi(\delta)(1 + \sqrt{x} \vee x)}{\sqrt{n}} + \sup_{j \geq 1} \Big((1+\epsilon)\frac{\psi(q^{j+1}\delta)(1 + \sqrt{x} \vee x)}{\sqrt{n}} - \epsilon q^j \delta \Big). \tag{22}$$

Assume further that ψ is continuous, increasing, such that $\delta \mapsto \psi(\delta)/\delta$ is decreasing and ψ^{-1} is strictly convex. We can define the convex conjugate of ψ^{-1} as

$$\psi^{-1*}(\delta) := \sup_{x>0}\{x\delta - \psi^{-1}(x)\}. \tag{23}$$

The following Lemma comes in handy to choose a parameter δ that kills the second term in the right hand side of (22).

Lemma 2. *Let $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ be a continuous and increasing function and assume that ψ^{-1} is strictly convex. If $\delta := \psi^{-1*}(2x/y)$, we have*

$$x\psi(\delta) \leq y\delta$$

for any $x, y > 0$.

Proof. Simply write

$$x\psi(\delta) = \frac{y}{2}\frac{2x}{y}\psi\Big(\psi^{-1*}\Big(\frac{2x}{y}\Big)\Big) \leq \frac{y}{2}\Big(\psi^{-1*}\Big(\frac{2x}{y}\Big) + \psi^{-1*}\Big(\frac{2x}{y}\Big)\Big),$$

where the trick is to use the fact that $uv \leq \psi^{-1*}(u) + \psi^{-1}(v)$ for any $u, v > 0$. □

Using Lemma 2 and the fact that $\psi(q^{j+1}\delta) \leq q^{j+1}\psi(\delta)$, we obtain that for the choice

$$\delta_{n,\epsilon}(x) := \psi^{-1*}\Big(\frac{2q(1+\epsilon)(1+\sqrt{x}\vee x)}{\epsilon\sqrt{n}}\Big),$$

we have, with a probability larger than $1 - Le^{-x}$:

$$P(\ell_{\bar{\alpha}_n}) - P(\ell_{\alpha_*}) \leq (1+\epsilon)\rho + \epsilon\delta_{n,\epsilon}(x).$$

We have proved the following result.

Proposition 2 (Peeling)**.** *Assume that (21) holds for any $\delta > \delta_{\min}$, where $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is a continuous and increasing function such that ψ^{-1} is strictly convex and $\delta \mapsto \psi(\delta)/\delta$ is decreasing. If $\bar{\alpha}_n$ is a ρ-ERM according to Definition 1, we have for any $x > 0$:*

$$P(\ell_{\bar{\alpha}_n}) \leq P(\ell_{\alpha_*}) + (1+\epsilon)\rho + \epsilon\delta_{n,\epsilon}(x)$$

with probability larger than $1 - Le^{-x}$, where

$$\delta_{n,\epsilon}(x) := \psi^{-1*}\Big(\frac{2(1+\epsilon)q(1+\sqrt{x}\vee x)}{\epsilon\sqrt{n}}\Big) \vee \delta_{\min}.$$

In the next section, we prove Inequality (21) using the generic chaining mechanism, under an assumption on the complexity of A.

3.2 Generic chaining

The *generic chaining* technique, which is introduced in Talagrand (2005) is, in our setting, a nice way to prove (21). It is based on the $\gamma_\nu(A, d)$ functional (see below) which is an alternative to Dudley's entropy integral (see Dudley (1978) for instance). The idea is to decompose A using an approximating sequence of partitions, instead

of nets with decreasing radius, as this is done in the standard chaining method. Let us briefly recall some necessary notions that can be found in details in Talagrand (2005), Chapter 1.

Let (A, d) be a metric space (d can be a semi-distance). Denote by $\Delta(A, d) := \sup_{a,b\in B} d(a, b)$ the diameter of A. An *admissible* sequence of A is an increasing sequence $(\mathcal{A}_j)_{j\geq 0}$ of partitions of A (every set of $\mathcal{A}_{j+1}$ is included in a set of $\mathcal{A}_j$) such that $|\mathcal{A}_j| \leq 2^{2^j}$ and $|\mathcal{A}_0| = 1$. If $a \in A$, we denote by $A_j(a)$ the unique element of $\mathcal{A}_j$ that contains a. For $\nu > 0$, define the function

$$\gamma_\nu(A, d) := \inf \sup_{a\in A} \sum_{j\geq 0} 2^{j/\nu} \Delta(A_j(a), d) \tag{24}$$

where the infimum is taken among all admissible sequence of A. This quantity is an alternative (and an improvement, see Talagrand (2005), in particular Theorem 3.3.2) of the Dudley's entropy integral (see for instance van der Vaart and Wellner (1996)). Indeed, we have:

$$\gamma_\nu(A, d) \leq L \int_0^{\Delta(A,d)} (\log N(A, \varepsilon, d))^{1/\nu} \mathrm{d}\varepsilon, \tag{25}$$

where $N(A, \varepsilon, d)$ is the *covering number* of A, namely the smallest integer N such that there is $B \subset A$ satisfying $|B| \leq N$ and $d(a, B) \leq \varepsilon$ for any $a \in A$.

Introduce $d_2(a, b) := \|a - b\|$ where $\|\cdot\|$ is the semi-norm given by (10) and $d_\infty(a, b) = \|a - b\|_\infty$, where $\|\cdot\|_\infty$ is the uniform norm (4). Using the generic chaining argument, we obtain the following deviation inequality.

Theorem 1. *Grant Assumptions 1 and 2. For any $x > 0$, we have*

$$\sup_{\alpha\in A} \sqrt{n}(P - P_n)(\ell_\alpha - \ell_{\alpha_*}) \leq c\Big(\gamma_2(A, d_2)(1 + \sqrt{x}) + \gamma_1(A, d_\infty)\frac{1 + x}{\sqrt{n}}\Big) \tag{26}$$

with a probability larger than $1 - Le^{-x}$, where $c = c_{b,\|\alpha_0\|_\infty} = 4(b + \|\alpha_0\|_\infty) + 2([\sqrt{2}(\sqrt{2} + 1)\|\alpha_0\|_\infty]^{1/2} + 1)$ (and $L \approx 1.545433$).

The proof of Theorem 1 is given in Section 6 below. In (26), the function γ_2 is related to the subgaussian term of the Bernstein inequality (18), while γ_1 is related to the subexponential term. However, if we have an extra condition on the complexity of A, it is possible to "remove" the γ_1 term from (26). This is called the *adaptive truncation argument*, which is related to the use of brackets (instead of balls) to construct a covering of A.

3.3 Brackets

Entropy with bracketing has been introduced by Dudley (1978). The adaptive truncation argument was introduced by Bass (1985) for partial sum process and Ossiander (1987) for the empirical process. We refer to van de Geer (2000) (in particular the proof of Theorem 8.13) herein and Massart (2007) (see the proof of Theorem 6.8) for the use of this technique with statistical applications in mind. In the context of generic chaining, bracketing can defined as follows. Following Talagrand (2005) (see in particular Theorem 2.7.10), we consider

$$\gamma^{[]}(A) := \inf \sup_{a\in A} \sum_{j\geq 0} 2^{j/2} \|\mathscr{B}_{A_j(a)}\|, \tag{27}$$

where the infimum is taken among all admissible sequences of A, where we recall that $\|\cdot\|$ is defined by (10), and where

$$\mathscr{B}_A(z) := \sup_{a,a' \in A} |a(z) - a'(z)|$$

for any $z \in [0,1]^{d+1}$. If $a^L, a^U \in A$, the *bracket* $[a^L, a^U]$ is the band

$$[a^L, a^U] := \{a \in A : a^L \leq a \leq a^U \text{ pointwise}\}.$$

The quantity $\|a^U - a^L\|$ is the *diameter* of the bracket. We denote by $N^{[]}(A, \epsilon)$ the minimal number of brackets with diameter not larger than ϵ necessary to cover A. Analogously to (25), one has

$$\gamma^{[]}(A) \leq L \int_0^{\Delta(A,d_\infty)} (\log N^{[]}(A,\epsilon))^{1/2} \mathrm{d}\epsilon. \tag{28}$$

Entropy with bracketing is a refinement of $\mathbb{L}^\infty$-entropy, that can be suitable for some class of functions, for instance functions with uniformly bounded variation, see for instance van de Geer (1993) and Bitouzé et al. (1999). In our setting, it is useful to "remove" the γ_1 term from (26), thanks to the following result, which is Talagrand's version of the adaptive truncation argument.

Theorem 2 (Talagrand (2005), Theorem 2.7.11)**.** *Let A be a countable set of measurable functions, and let $u > 0$. If $\gamma^{[]}(A) \leq \Gamma$, we can find two sets A_1, A_2 with the following properties*:

- $\gamma_2(A_1, d_2) \leq L\Gamma, \quad \gamma_1(A_1, d_\infty) \leq Lu\Gamma,$
- $\gamma_2(A_2, d_2) \leq L\Gamma, \quad \gamma_1(A_2, d_\infty) \leq Lu\Gamma,$
- *for any $a \in A_2$, we have $a \geq 0$ and $\|a\|_1 \leq L\Gamma/u$, and*

$$A \subset A_1 + A_2', \text{ where } A_2' = \{a' : \exists a \in A_2, |a'| \leq a\}.$$

Indeed, an immediate consequence of Proposition 1 and Theorem 2 (simply take $u = \sqrt{n}$ in Theorem 2) is the following.

Corollary 1. *Grant Assumptions 1 and 2. For any $x > 0$, we have*

$$\sup_{\alpha \in A} \sqrt{n}(P - P_n)(\ell_\alpha - \ell_{\alpha_*}) \leq c\gamma^{[]}(A)(1 + \sqrt{x} \vee x)$$

with a probability larger than $1 - Le^{-x}$, where $c = c_{b,\|\alpha_0\|_\infty}$ is the same as in Theorem 1.

3.4 A risk bound for the ERM

Corollary 1 is close to the concentration inequality (21) required in the peeling argument, see Proposition 2 above. However, note that the peeling was done using sets $A(\delta) = \{\alpha \in A : P(\ell_\alpha) - P(\ell_{\alpha_*}) \leq \delta\}$ for $\delta > 0$, while we can bound from above the entropy (and consequently the functionals γ and $\gamma^{[]}$) of balls $B(\delta) = \{\alpha \in A : \|\alpha - \alpha_*\| \leq \delta\}$ using a standard result (see Section 3.5). Hence, it will be convenient to work under the following assumption.

Assumption 3. *Assume that* $\|\alpha - \alpha_*\|^2 \leq P(\ell_\alpha) - P(\ell_{\alpha_*})$ *for every* $\alpha \in A$.

This assumption is a bit stronger than the standard *margin assumption*, see Mammen and Tsybakov (1999); Tsybakov (2004), or the β-*Bernstein condition*, see Bartlett and Mendelson (2006) [Note that here $\beta = 1$, as in most statistical models, see Lecué (2007).] Indeed, let us prove that Assumption 3 entails, together with Assumptions 1 and 2:

$$P((\ell_\alpha - \ell_{\alpha_*})^2) \leq cP(\ell_\alpha - \ell_{\alpha_*}) \text{ for every } \alpha \in A, \tag{29}$$

where $c = c_{b,\|\alpha_0\|_\infty} := 8((b + \|\alpha_0\|_\infty)^2 + \|\alpha_0\|_\infty)$, which is the $(1, c)-$Bernstein condition from Bartlett and Mendelson (2006). We have using (2):

$$\ell_\alpha(X, (Y_t), (N_t)) = \ell'_\alpha(X, (Y_t)) - 2\int_0^1 \alpha(t, X)dM(t),$$

where ℓ'_α is the loss function

$$\ell'_\alpha(x, (y_t)) := \int_0^1 \alpha(t, x)^2 y(t)dt - 2\int_0^1 \alpha(t, x)\alpha_0(t, x)y(t)dt,$$

so the following decomposition holds:

$$\begin{aligned}
&\ell_\alpha(X, (Y_t),(N_t)) - \ell_{\alpha_*}(X, (Y_t), (N_t)) \\
&\qquad = \ell'_\alpha(X, (Y_t)) - \ell'_{\alpha_*}(X, (Y_t)) + 2\int_0^1 (\alpha_*(t, X) - \alpha(t, X))dM(t) \\
&\qquad = \int_0^1 (\alpha(t, X) - \alpha_*(t, X))(\alpha(t, X) + \alpha_*(t, X) - 2\alpha_0(t, X))Y(t)dt \\
&\qquad + 2\int_0^1 (\alpha_*(t, X) - \alpha(t, X))dM(t).
\end{aligned}$$

Hence, using Assumptions 1 and 2, we have:

$$\begin{aligned}
P((\ell_\alpha - \ell_{\alpha_*})^2) &\leq 8(b + \|\alpha_0\|_\infty)^2\|\alpha - \alpha_*\|^2 \\
&\quad + 8\mathsf{E}\Big[\int_0^1 (\alpha_*(t, X) - \alpha(t, X))^2\alpha_0(t, X)Y(t)dt\Big] \\
&\leq 8((b + \|\alpha_0\|_\infty)^2 + \|\alpha_0\|_\infty)\|\alpha - \alpha_*\|^2,
\end{aligned}$$

and (29) follows using Assumption 3. Now, let us show that Assumption 3 is mild: it is met when A is convex, for instance. The fact that convexity entails the margin assumption is true in most statistical models, such a in regression, see for instance Lee et al. (1998).

Lemma 3. *Grant Assumption 1 and let* A *be a convex class of functions bounded by* $b > 0$. *Then, Assumption 3 is met.*

Proof. Since A is convex and $P(\alpha_*) = \inf_{\alpha \in A} P(\ell_\alpha)$, we have $\langle \alpha_* - \alpha_0, \alpha_*, \alpha\rangle \leq 0$ for any $\alpha \in A$, where we recall that the inner product is given by (10). This entails

$$\begin{aligned}
P(\ell_\alpha - \ell_{\alpha_*}) &= \|\alpha\|^2 - 2\langle \alpha, \alpha_0\rangle - \|\alpha_*\|^2 + 2\langle \alpha_*, \alpha_0\rangle \\
&= \|\alpha - \alpha_*\|^2 - 2\langle \alpha_* - \alpha_0, \alpha_* - \alpha\rangle \\
&\geq \|\alpha - \alpha_*\|^2. \square
\end{aligned}$$

We are now in position to state the following risk bound for the ERM, under a condition on the complexity of A.

Theorem 3. *Grant Assumptions 1, 2 and 3. Assume that there is $\delta_{\min} > 0$ and a continuous and increasing function $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ such that for any $\delta > \delta_{\min}$, any $\alpha' \in A$ and any $\sqrt{\delta}$-ball $B(\sqrt{\delta}) = \{\alpha \in A : \|\alpha - \alpha'\|^2 \leq \delta\}$, we have either*:

$$\varphi(\delta) \geq \gamma_2(B(\sqrt{\delta}), d_2) + \frac{1}{\sqrt{n}}\gamma_1(B(\sqrt{\delta}), d_\infty) \quad \textit{for any} \quad \delta > \delta_{\min},$$

or:

$$\varphi(\delta) \geq \gamma^{[]}(B(\sqrt{\delta})) \quad \textit{for any} \quad \delta > \delta_{\min}.$$

Assume further that φ^{-1} is strictly convex and that $\delta \mapsto \varphi(\delta)/\delta$ is decreasing. Then, if $\bar{\alpha}_n$ is a ρ-ERM according to Definition 1, we have for any $\epsilon > 0$, $x > 0$:

$$P(\ell_{\bar{\alpha}_n}) \leq P(\ell_{\alpha_*}) + (1+\epsilon)\rho + \epsilon\delta_{n,\epsilon}(x)$$

with a probability larger than $1 - Le^{-x}$, where

$$\delta_{n,\epsilon}(x) := \varphi^{-1*}\Big(\frac{c(1+\epsilon)(1+\sqrt{x} \vee x)}{\epsilon\sqrt{n}}\Big) \vee \delta_{\min},$$

and $c = c_{b,\|\alpha_0\|_\infty}$ is the same as in Theorem 1.

Proof. Because of Assumption 3, we have $A(\delta) \subset B(\sqrt{\delta})$, so Inequality (21) is satisfied under the assumptions of the theorem with $\psi(\delta) = c\varphi(\delta)$, using Theorem 1 or Corollary 1. Hence, we can apply Proposition 2, which entails the Theorem since $\psi^{-1*}(x) = \varphi^{-1*}(cx)$. □

Remark 2 (Comparison). This bound for the ERM is of the same nature as previous bounds for the ERM in more "standard" models, such as density, regression or classification. The rate given in Theorem 3 gives, on examples, the same rate (up to constants) as the one given in Massart (2007) (see Theorem 8.3), for instance. Consider the situation where $\varphi(\delta) = c\delta^\alpha$ for $c > 0$ and $\alpha \in (0,1)$ ($\varphi(\delta)$ is of order $\sqrt{D\delta}$ when A has a finite dimension, see Section 3.5 below). In this case, we have $\varphi^{-1*}(x) = (1-\alpha)\alpha^{\alpha/(1-\alpha)}(cx)^{1/(1-\alpha)}$, so $\delta_{n,\epsilon}(x)$ is of order $(c/\sqrt{n})^{1/(1-\alpha)}$. The rate ε_*^2 in the bound by Massart is solution to the equation $\sqrt{n}\varepsilon^2 = \varphi(\varepsilon^2)$, hence $\varepsilon_*^2 = (c/\sqrt{n})^{1/(1-\alpha)}$, and both rates have the same order.

Remark 3 (Talagrand's inequality). Usually, the complexity of the sieve A is measured by the functional $\phi(B) = \sqrt{n}\,\mathsf{E}^n[\sup_{\alpha\in B}(P - P_n)(\ell_\alpha - \ell_{\alpha_*})]$ where B are balls in A, like in Massart (2007) or spheres in A, see Bartlett and Mendelson (2006). The rate is then the solution of a fixed point problem involving these functional, such as, roughly, the equation $\phi(B(\varepsilon)) = \sqrt{n}\varepsilon^2$ from Massart (2007). Note that the main tool in the proof of these results is Talagrand's deviation inequality, see Massart (2000), Rio (2001) or Bousquet (2002). In Theorem 3, we were not able to state the bound with a rate defined in such a way. Indeed, we needed a "stronger" control on the complexity, given by the γ functionals, to define $\delta_{n,\epsilon}$. This is related to the fact that we cannot use a Talagrand's type deviation inequality in the general model (2) for

$$\sup_{\alpha\in A}(P - P_n)(\ell_\alpha - \ell_{\alpha_*}) - \mathsf{E}^n[\sup_{\alpha\in A}(P - P_n)(\ell_\alpha - \ell_{\alpha_*})].$$

Indeed, $\int_0^1 \alpha(t,X)dM(t)$ is not, in general, bounded (think of the Poisson process for instance, which is a particular case of Section 1.2.2).

However, the story is different when $N(t)$ is bounded, such as in the models of regression for right-censored data, and of transition intensities of Markov processes (see Sections 1.2.1 and 1.2.3). It is then possible to use the strength of Talagrand's inequality, following the arguments from Bartlett and Mendelson (2006) (up to significant modifications, since the analysis is conducted in the regression model).

A case of importance (particularly in practice) is when A is included in a linear space $\bar{A}$ with a finite dimension D (see Birgé and Massart (1998) and Massart (2007) for instance). Using the version of Massart (2007) of a classical result concerning $\mathbb{L}^\infty$-coverings of a ball in such a space (see below), we can show that $\delta_{n,\epsilon}(x)$ is smaller than a quantity of order D/n. This will be useful to compute rates of convergence in Section 5 below.

3.5 When A is finite dimensional

Let us now consider the case where A is a subset of some linear space $\bar{A} \subset \mathbb{L}^2 \cap \mathbb{L}^\infty[0,1]^{d+1}$ with finite dimension D. Following Birgé and Massart (1998) and Barron et al. (1999), we can consider the $\mathbb{L}^\infty$-*index*

$$r(\bar{A}) := \frac{1}{\sqrt{D}} \inf_{\psi} \sup_{\beta \neq 0} \frac{\| \sum_{\lambda \in \Lambda} \beta_\lambda \psi_\lambda \|_\infty}{|\beta|_\infty},$$

where $|\beta|_\infty = \max_{\lambda \in \Lambda} |\beta_\lambda|$ and where the infimum is taken over all orthonormal basis $\{\psi_\lambda : \lambda \in \Lambda\}$ of $\bar{A}$.

This index can be estimated for all the linear spaces usually chosen as approximation spaces for adaptive estimation, see Birgé and Massart (1998) and Barron et al. (1999). In particular, if $\bar{A}$ is spanned by a localized basis, then $r(\bar{A})$ can be bounded independently of D (think of a wavelet basis for instance, more on that in Section 5 below).

Using this index, we can derive a bound for $\gamma^{[]}(B(\sqrt{\delta}))$. For any $\varepsilon \in (0, \delta]$, the following holds (see Massart (2007), Lemma 7.14):

$$N(B(\delta), \varepsilon, d_\infty) \leq \Big(\frac{Lr(\bar{A})\delta}{\varepsilon} \Big)^D, \tag{30}$$

where L can be $\sqrt{3\pi e/2}$. But, using (28) together with the fact that $N^{[]}(A, \epsilon/2) \leq N(A, \epsilon, d_\infty)$, we obtain

$$\gamma^{[]}(B(\delta)) \leq \sqrt{D} \int_0^\delta \sqrt{\ln\Big(\frac{2Lr(\bar{A})\delta}{\varepsilon} \Big)} \mathrm{d}\epsilon \leq L\delta \sqrt{D(\ln r(\bar{A}) + 1)}. \tag{31}$$

So, we have the control required in Theorem 3: $\gamma^{[]}(B(\sqrt{\delta})) \leq L\varphi(\delta)$, with

$$\varphi(\delta) = \sqrt{\delta} \sqrt{D(\ln r(\bar{A}) + 1)},$$

which is a function that satisfies the assumptions of Theorem 3. Note that

$$\varphi^{-1*}(x) = \frac{x^2 D \ln(r(\bar{A}) + 1)}{4}, \tag{32}$$

so we have the following.

Corollary 2. *Grant Assumptions 1 and 2, and assume that $A \subset \bar{A}$, where $\bar{A}$ is a linear space with finite dimension D. Then, if $\bar{\alpha}_n$ is a ρ-ERM according to Definition 1, we have for any $\epsilon > 0$ and $x > 0$:*

$$P(\ell_{\bar{\alpha}_n}) \leq P(\ell_{\alpha_*}) + (1+\epsilon)\rho + \frac{c(1+\epsilon)^2 \ln(r(\bar{A})+1)}{\epsilon} \frac{D}{n}(1 + x \vee \sqrt{x})^2$$

with a probability larger than $1 - Le^{-x}$, where $c = c_{b,\|\alpha_0\|_\infty}$. In particular, we obtain

$$\mathsf{E}^n \|\bar{\alpha}_n - \alpha_0\|^2 \leq 2\rho + \inf_{\alpha \in A} \|\alpha - \alpha_0\|^2 + c \ln(r(\bar{A})+1) \frac{D}{n}. \tag{33}$$

Proof. Note that Assumption 3 is met since $\bar{A}$ is linear. So, Theorem 3 together with (32) gives the first inequality. The second inequality follows by choosing $\epsilon = 1$, by subtracting $P(\ell_{\alpha_0})$ at both sides of the inequality, and by integration with respect to x. □

The next step is, usually, to have a control on the *approximation* or *bias* term $\inf_{\alpha \in A} \|\alpha - \alpha_0\|^2$, and to choose a sieve with a dimension that equilibrates the bias term with the "variance" term D/n, hence the name *bias-variance* problem, see Cucker and Smale (2002) for instance. Usually, this is done using the assumption that α_0 belongs to some smoothness class of functions, together with some results from approximation theory. This is where the problem of adaptive estimation arises: the choice of the optimal D depends on the parameters of the smoothness class itself, which is unknown in practice. So, one has to find a procedure with the capability to select automatically a sieve or a *model* among a collection $\{A_m : m \in \mathcal{M}\}$. This is usually done using model-selection, see the seminal paper Barron et al. (1999). Model selection in the setup considered here has been studied in Comte et al. (2008). In Section 4 below, we consider an alternative approach, based on a popular aggregation procedure. It will allow the construction of smoothness and structure adaptive estimators, see Section 5.

4 Agnostic learning, aggregation

Let $A = A(\Lambda) := \{\alpha_\lambda : \lambda \in \Lambda\}$ be a set of arbitrary functions called *dictionary* with cardinality M. For instance, this can be a set of so-called *weak* estimators, computed based on a set of observations independent of the sample D_n. We consider the problem of agnostic learning: without any assumption on α_0, excepted for some boundedness assumption, we want to construct (from the data) a procedure $\hat{\alpha}_n$ with a risk as close as possible to the smallest risk over A. Namely, we want to obtain an oracle inequality of the form

$$\mathsf{E}^n \|\hat{\alpha}_n - \alpha_0\|^2 \leq c \min_{\alpha \in A} \|\alpha - \alpha_0\|^2 + \phi(n, M),$$

where $c \geq 1$ and $\phi(n, M)$ is called the *residue* or *rate of aggregation*, which is a quantity that we want to be small as n increases. An oracle inequality that holds with $c = 1$ is called *sharp*.

This problem has been considered in several statistical models, mainly in regression, density and classification, see among others Nemirovski (2000); Catoni (2001); Juditsky et al. (2006); Leung and Barron (2006); Dalalyan and Tsybakov (2007); Yang

(2000); Audibert (2009). For instance, we know from Tsybakov (2003) that the optimal rate of aggregation in the Gaussian regression model is $\phi(n, M) = (\log M)/n$ (in the sharp oracle inequality context). This rate is achieved by the algorithms of aggregation with cumulative exponential weights, see Juditsky et al. (2006); Audibert (2009) and aggregation with exponential weights, see Dalalyan and Tsybakov (2007) (when the error of estimation is measured by the empirical norm, a similar result for the integrated norm is, as far as we know, still a conjecture).

Aggregation with exponential weights is a popular algorithm. It is of importance in machine learning, for estimation, prediction using expert advice, in PAC-Bayesian learning and other settings, see Cesa-Bianchi and Lugosi (2006), Audibert (2009) and Catoni (2001), among others. However, there is no result for this algorithm in the general model (2), nor for any of the particular cases given in Section (1.2). In this Section, we construct this algorithm for model (2), and give in Theorem 4 below an oracle inequality.

The idea of aggregation is to mix the elements from $A(\Lambda)$: using the data, compute weights $\theta(\alpha) \in [0,1]$ for each $\alpha \in A(\Lambda)$ satisfying $\sum_{\lambda\in\Lambda} \theta(\alpha_\lambda) = 1$. These weights give a level of significance to α. The aggregate is the convex combination

$$\hat{\alpha}_n := \sum_{\lambda\in\Lambda} \theta(\alpha_\lambda)\alpha_\lambda, \tag{34}$$

where the weight of $\alpha \in A(\Lambda)$ is given by

$$\theta(\alpha) := \frac{\exp\big(-nP_n(\ell_\alpha)/T\big)}{\sum_{\lambda\in\Lambda}\exp\big(-nP_n(\ell_{\alpha_\lambda})/T\big)}, \tag{35}$$

where $T > 0$ is the so-called *temperature* parameter and where we recall that

$$P_n(\ell_\alpha) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \alpha(t, X_i)^2 Y^i(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^1 \alpha(t, X_i)dN^i(t)$$

is the empirical risk of α. The shape of this mixing estimator is easily explained. Indeed, the weighting scheme (35) is the only minimizer of

$$\mathsf{R}_n(\theta) + \frac{T}{n}\sum_{\lambda\in\Lambda}\theta_\lambda\log\theta_\lambda \tag{36}$$

among all $\theta \in \Theta$ (we use the convention $0\log 0 = 0$) where

$$\Theta := \big\{\theta \in \mathbb{R}^M : \theta_\lambda \geq 0,\ \sum_{\lambda\in\Lambda}\theta_\lambda = 1\big\},$$

and where $\mathsf{R}_n(\theta)$ is the *linearized empirical risk*

$$\mathsf{R}_n(\theta) := \sum_{\lambda\in\Lambda}\theta_\lambda P_n(\ell_{\alpha_\lambda}).$$

Equation (36) is the linearized risk of $\theta \in \Theta$, which is penalized by a quantity proportional to the Shannon's entropy of θ. The resulting aggregated estimator $\hat{\alpha}_n$ is then something between the ERM among the elements of $A(\Lambda)$ (when T is small), and the mean of the elements of $A(\Lambda)$ (when T is large).

Theorem 4. *Assume that* $\|\alpha_0\|_\infty < +\infty$*, and that there is* $b > 0$ *such that* $\|\alpha\|_\infty \leq b$ *for any* $\alpha \in A(\Lambda)$*. Then, for any* $\epsilon > 0$*, the mixing estimator* $\hat{\alpha}_n$ *defined by* (34) *satisfies*

$$\mathsf{E}^n \|\hat{\alpha}_n - \alpha_0\|^2 \leq (1+\epsilon) \inf_{\lambda \in \Lambda} \|\alpha_\lambda - \alpha_0\|^2 + \frac{c \log M}{n}$$

for any $n \geq 1$*, where* $c = c_{b,\|\alpha_0\|_\infty,T,\epsilon}$*.*

Theorem 4 is a model-selection type oracle inequality for the aggregation procedure given by (34). The residual term in the oracle inequality is of order $(\log M)/n$, which is the correct rate of convex aggregation, see Tsybakov (2003) (in the Gaussian regression setup, and for other models with margin parameter equal to 1, see Lecué (2007)).

Remark 4. The main criticism one can make about Theorem 4 is that it is not sharp: the leading constant is $1+\epsilon$ instead of 1 in front of $\inf_{\lambda \in \Lambda} \|\alpha_\lambda - \alpha_0\|^2$, and the constant c in front of the residue is far from being optimal. The consequence is that we are not able in this setting to give a theoretically optimal value for T. Sharp oracle inequalities are available for aggregation with exponential weights or cumulative weights, see Dalalyan and Tsybakov (2007), Juditsky et al. (2006) and Audibert (2009), see also references mentioned above. However, in the setup considered here, the proof of a sharp oracle inequality seems quite challenging, and will be the subject of further investigations.

5 Structure and smoothness adaptive estimation

In this Section, we propose an application of the results obtained in Sections 3 and 4. We construct an estimator that adapts to the smoothness of α_0 in a purely non-parametric setting, see Section 5.1, and to its structure in a single-index setup, see Section 5.2. The steps of the construction of the estimator are given in Definition 2 below. As usual with algorithms coming from statistical learning theory, we need to split the sample (a very particular exception can be found in Leung and Barron (2006)). To simplify, we shall assume that the sample size is $2n$, see (3), so D_{2n} is the full sample.

Definition 2. The steps for the computation of an aggregated estimator $\hat{\alpha}_n$ are the following:

1. split the whole sample D_{2n} (see (3)) into a training sample $D_{n,1}$ of size n and a *learning sample* $D_{n,2}$ of size n;
2. choose a collection of sieves $\{A_m : m \in \mathcal{M}_n\}$ and compute, using $D_{n,1}$, the corresponding empirical risk minimizers $\{\bar{\alpha}_m : m \in \mathcal{M}_n\}$ (see Definition 1);
3. using the learning sample $D_{n,2}$, compute the aggregated estimator $\hat{\alpha}_n$ based on the dictionary $\{\bar{\alpha}_m : m \in \mathcal{M}_n\}$, see (34) and (35).

Examples of collections $\{A_m : m \in \mathcal{M}_n\}$ are given in Appendix A.1, together with the necessary control of the $\mathbb{L}^\infty$-index (see Section 3.5), and a useful approximation result.

Remark 5 (Jackknife). The behavior of the aggregate $\hat{\alpha}_n$ typically depends on the split selected in Step 1, in particular when the number of observations is small. Hence, a good strategy is to jackknife: repeat, say, J times Steps 1–3 to obtain aggregates $\{\hat{\alpha}_n^{(1)}, \dots, \hat{\alpha}_n^{(J)}\}$, and compute the mean:

$$\hat{\alpha}_n := \frac{1}{J} \sum_{j=1}^{J} \hat{\alpha}_n^{(j)}.$$

This jackknifed estimator should provide more stable results than a single aggregate. Moreover, by convexity of the risk $\alpha \mapsto P(\ell_\alpha)$, the jackknifed estimator satisfies the same risk bounds as a single aggregate.

5.1 Adaptive estimation in the purely nonparametric setting

In model (2), the behaviour of $\alpha_0(t, x)$ with respect to time t and with respect to the covariates x have no statistical reason to be linked. So, in a purely nonparametric setting, it is mandatory to consider anisotropic smoothness for α_0. We shall assume in the statement of the upper bound, see Theorem 5 below, that $\alpha_0 \in B^{\boldsymbol{s}}_{2,\infty}$, where $\alpha_0 \in B^{\boldsymbol{s}}_{2,\infty}$ is an anisotropic Besov space (see Appendix A.1) and $\boldsymbol{s} = (s_1, \dots, s_{d+1})$ is a vector of smoothness, where s_i is the smoothness in the ith coordinate. For the construction of the adaptive estimator, see Step 2 above, we need a collection of sieves $\{A_m : m \in \mathcal{M}_n\}$.

Definition 3 (Collection). We take $\{A'_m : m \in \mathcal{M}_n\}$ as:

- a collection of linear spaces spanned by piecewise polynomials (see Section A.1.1), with degrees not larger than l_i in the ith coordinate, or
- a collection of linear spaces spanned by wavelets (see Section A.1.2) with l_i vanishing moments in the ith coordinate.

In both cases, we say that $(l_1, \dots, l_{d+1})$ is the *smoothness* of the collection, and we take

$$\mathcal{M}_n := \{(m_1, \dots, m_{d+1}) \in \mathbb{N}^{d+1} : 2^{m_i} \leq n^{1/(d+1)} \text{ for } i = 1, \dots, d+1\}.$$

Finally, we fix a constant $b > 0$ and take $A_m := \{\alpha \in A'_m : \|\alpha\|_\infty \leq b\}$ for every $m \in \mathcal{M}_n$.

For the statement of the adaptive upper bound, we need the following assumption, which is a stronger version of the previous Assumption 1.

Assumption 4. *Assume that P_X has a density f_X with respect to the Lebesgue measure, which is bounded and with support $[0,1]^{d+1}$. Moreover, we assume that $\|\alpha_0\|_\infty \leq b$, where b is known (it is used in the definition of the sieves, see Definition 3).*

Now, we can use together Corollary 2 (see Section 3.5), Theorem 4 (see Section 4) and Lemma 5 (see Appendix A.1) to derive an adaptive upper bound. Take $\rho = 1/n$ in Corollary 2 and, say, $\epsilon = 1$ in Theorem 4, to obtain

$$\mathsf{E}^{2n}\|\hat{\alpha}_n - \alpha_0\|^2 \leq 2 \inf_{m \in \mathcal{M}_n} \left(\inf_{\alpha \in A_m} \|\alpha - \alpha_0\|^2 + \frac{c_b D_m}{n} \right) + c_{b,T} \frac{\log |\mathcal{M}_n|}{n},$$

where for $m = (m_1, \ldots, m_{d+1}) \in \mathcal{M}_n$,

$$D_m := \prod_{j=1}^{d+1} D_{m_j} = \prod_{j=1}^{d+1} 2^{m_j}.$$

Let us assume that $\alpha_0 \in B^{\boldsymbol{s}}_{2,\infty}$, where $\boldsymbol{s} = (s_1, \ldots, s_{d+1})$ satisfies $s_i > (d+1)/2$ for each $i = 1, \ldots, d+1$. Assumption 4 entails $\|\alpha\|^2 \leq \|f_X\|_\infty \|\alpha\|_2^2$ for any $\alpha \in \mathbb{L}^2[0,1]^{d+1}$, where $\|\alpha\|_2^2 = \int_{[0,1]} \int_{[0,1]^d} \alpha(t,x)^2 dt dx$. So, using Lemma 5, we have when $\alpha_0 \in B^{\boldsymbol{s}}_{2,\infty}$:

$$\mathsf{E}^{2n} \|\hat{\alpha}_n - \alpha_0\|^2 \leq c\Big(\sum_{j=1}^{d+1} D_{m_j}^{-2s_j} + \frac{\prod_{j=1}^{d+1} D_{m_j}}{n} + \frac{\log |\mathcal{M}_n|}{n} \Big),$$

where $c = c_{b,T,\boldsymbol{s},d,\|f_X\|_\infty,|\alpha_0|_{B^{\boldsymbol{s}}_{2,\infty}}}$. Note that $(\log |\mathcal{M}_n|)/n \leq c_d (\log n)/n$, so the rate of convergence is given by the optimal tradeoff between the bias and the variance terms. Since $s_i > (d+1)/2$ for any $i = 1, \ldots, d+1$, we have $n^{\bar{\boldsymbol{s}}/s_i(2\bar{\boldsymbol{s}}+d+1)} \leq n^{1/(d+1)}$, so we can choose $m = (m_1, \ldots, m_{d+1}) \in \mathcal{M}_n$ such that

$$2^{m_i - 1} \leq n^{\frac{\bar{\boldsymbol{s}}/s_i}{2\bar{\boldsymbol{s}}+d+1}} \leq 2^{m_i} \text{ for } i = 1, \ldots, d+1. \tag{37}$$

This gives

$$\sum_{j=1}^{d+1} D_{m_j}^{-2s_j} + \frac{\prod_{j=1}^{d+1} D_{m_j}}{n} = c_d n^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+d+1)},$$

so we proved the following theorem.

Theorem 5. *Grant Assumption 4, and consider a collection $\{A_m : m \in \mathcal{M}_n\}$ given by Definition 3 with smoothness $(l_1, \ldots, l_{d+1})$. Assume that $\alpha_0 \in B^{\boldsymbol{s}}_{2,\infty}$, where $\boldsymbol{s} = (s_1, \ldots, s_{d+1})$ satisfies $(d+1)/2 < s_i \leq l_i$ for each $i = 1, \ldots, d+1$. Then, if $\hat{\alpha}_n$ is the aggregated estimator given by Steps 1-3, we have*

$$\mathsf{E}^{2n} \|\hat{\alpha}_n - \alpha_0\|^2 \leq c n^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+d+1)},$$

where $c = c_{b,T,\boldsymbol{s},d,\|f_X\|_\infty,|\alpha_0|_{B^{\boldsymbol{s}}_{2,\infty}}}$.

The rate $n^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+d+1)}$ is the optimal rate of convergence (in the minimax sense) in this model, under the extra assumption that f_X is bounded away from zero on $[0,1]^d$, see Theorem 3 in Comte et al. (2008). Hence, Theorem 5 shows that $\hat{\alpha}_n$ adapts to the smoothness of α_0 over a range of Besov spaces $B^{\boldsymbol{s}}_{2,\infty}$, for $(d+1)/2 < s_i \leq l_i$.

5.2 Dimension reduction, single-index

The mark X is d-dimensional so the intensity α_0 takes $d+1$ variables. As with any other nonparametric estimation model, we know that when d gets large the dimension has a significant impact on the accuracy of estimation. This the so-called *curse of dimensionality* phenomenon, which is reflected by the rate $n^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+d+1)}$, see Theorem 5 above. This rate is slow if d is large compared to $\bar{\boldsymbol{s}}$. In this Section, we propose a way to "get back" the rate $n^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+2)}$, using single-index modelling. Thanks to

our approach based on aggregation, we are able to construct an estimator that automatically takes advantage (without any prior testing) of the single-index structure when possible: the rate is then $n^{-2\bar{s}/(2\bar{s}+2)}$, otherwise it is the the purely nonparametric rate $n^{-2\bar{s}/(2\bar{s}+d+1)}$. This idea of mixing nonparametric and semiparametric estimators can be also found in Yang (2000) for density estimation.

Dimension reduction techniques usually involves an assumption on the structure of the object to be estimated. Main examples are the additive and the single-index models. Additive modelling was proposed by Linton et al. (2003) in the same context as the one considered here, with very different techniques (kernel estimation and backfitting). In this paper, we focus on single-index modelling (see Remark 6 below). On single-index models (mainly in regression) and the corresponding estimation problems (estimation of the link function, estimation of the index), see Hristache et al. (2001), Delecroix et al. (2003), Xia and Härdle (2006), Delecroix et al. (2006), Geenens and Delecroix (2005), Gaiffas and Lecue (2007), Dalalyan et al. (2008) among many others. The single-index structure is as follows: assume that there is an unknown function $\beta_0 : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}_+$ (called *link function*, with has unknown smoothness here) and an unknown vector $v_0 \in \mathbb{R}^d$ (called *index*) such that

$$\alpha_0(t,x) = \beta_0(t, v_0^\top x). \tag{38}$$

In order to make the representation (38) unique (identifiability), we shall assume (see Assumption 5 below) that $v_0 \in S_+^{d-1}$, where S_+^{d-1} is the half-unit sphere defined by

$$S_+^{d-1} = \big\{v \in \mathbb{R}^d : |v|_2 = 1 \text{ and } v_d \geq 0\big\}, \tag{39}$$

where $|\cdot|_2$ is the Euclidean norm over $\mathbb{R}^d$;

The steps of the construction of the adaptive estimator in this context follows the ones from Definition 2, but the dictionary is enlarged by a set $\{\bar{\alpha}_{m,v}^{\mathrm{SIM}} : m \in \mathcal{M}_n^{\mathrm{SIM}}, v \in S_\Delta^{d-1}\}$, of empirical risk minimizers, where S_Δ^{d-1} is a Δ-net of S_+^{d-1}. So, compared to Section 5.1, the idea is simply to add estimators that works under the single-index assumption in the dictionary.

Definition 4. The steps for the computation of the aggregated estimator $\hat{\alpha}_n$ are the following:

1. split the whole sample D_{2n} (see (3)) into a training sample $D_{n,1}$ of size n and a *learning sample* $D_{n,2}$ of size n;
2. Compute a $\Delta = (n \log n)^{-1/2}$-net of the half-unit sphere S_+^{d-1} denoted by S_Δ^{d-1} and for each $v \in S_\Delta^{d-1}$ compute the pseudo-training samples

$$D_{n,1}(v) := \big[(v^\top X_i, N^i(t), Y^i(t)) : t \in [0,1], 1 \leq i \leq n\big], \tag{40}$$

   where the d-dimensional marks X_i are simply replaced univariate marks $v^\top X_i$.
3. Fix a collection of 2-dimensional sieves ($d = 1$) $\{A_m^{\mathrm{SIM}} : m \in \mathcal{M}_n^{\mathrm{SIM}}\}$ given by Definition 3. Compute, for every $m \in \mathcal{M}_n^{\mathrm{SIM}}$ and $v \in S_\Delta^{d-1}$, empirical risk minimizers $\bar{\beta}_{m,v}^{\mathrm{SIM}}$, over A_m^{SIM}, of the empirical risks

$$P_{n,1}^{(v)}(\ell_\alpha) = \frac{1}{n}\sum_{i=1}^n \int_0^1 \alpha(t, v^\top X_i)^2 Y^i(t) dt - \frac{2}{n}\sum_{i=1}^n \int_0^1 \alpha(t, v^\top X_i) dN^i(t),$$

and define

$$\bar{\alpha}^{\mathrm{SIM}}_{m,v}(\cdot,\cdot) := \bar{\beta}^{\mathrm{SIM}}_{m,v}(\cdot, v^\top \cdot).$$

(so that each $\bar{\alpha}^{\mathrm{SIM}}_{m,v}$ works as if v were the true index).

4. follow Steps 2 and 3 from Definition 2, where we add the estimators $\{\bar{\beta}^{\mathrm{SIM}}_{m,v} : m \in \mathcal{M}^{\mathrm{SIM}}_n, v \in S^{d-1}_\Delta\}$ to the set of purely nonparametric estimators $\{\bar{\alpha}_m : m \in \mathcal{M}_n\}$ in the aggregation step.

An important point of this algorithm is that we do not estimate the index directly: we mix estimators in order to *adapt* to the unknown v_0 and to the unknown smoothness of β_0. This approach was previously adopted in Gaiffas and Lecue (2007) for the estimation of the regression function. Note that the size of S^+_Δ increases strongly with n and d, so this method is restricted to a reasonably small d. High dimensional covariates cannot be handled in such a semiparametric approach, this problem will be the subject of another work. About high dimension, see Tibshirani (1997), where the LASSO has been studied in the Cox model.

The following set of assumptions gives the identifiability of model (38) (see for instance the survey paper by Geenens and Delecroix (2005), or Chapter 2 in Horowitz (1998)), excepted for (41) and (42) which are technical assumptions.

Assumption 5. *Assume that (38) holds, and that*

- $x \mapsto \beta_0(t,x)$ *is not constant over the support of* $v_0^\top X$*;*
- X *admits at least one continuously distributed coordinate (w.r.t. the Lebesgue measure);*
- *the support of* X *is not contained in any linear subspace of* $\mathbb{R}^d$*;*
- $v_0 \in S^{d-1}_+$*;*
- *there is* $c_0 > 0$ *such that for any* $x, y \in [0,1]^d$*, any* $t \geq 0$*:*

$$|\beta_0(t,x) - \beta_0(t,y)| \leq c_0 |x - y|; \tag{41}$$

- *there is* $b_0 > 0$ *such that*

$$\inf_{(t,x)\in[0,1]^{d+1}} \beta_0(t,x) \geq b_0. \tag{42}$$

Remark 6. In the problem of estimating the intensity of a counting process in presence of covariates, two of the most popular models are special cases of the single-index model, as described in Equation (38):

- the Cox model (see Cox (1972)), where there exists an unknown function β_0 such that:

$$\alpha_0(t,x) = \beta_0(t)\exp(v_0^\top x). \tag{43}$$

and

- the Aalen model (see Aalen (1980)), which can be written as:

$$\alpha_0(t,x) = \beta_0(t) + v_0^\top x. \tag{44}$$

This emphasizes the relevance of considering single-index models in this context, and the use of anisotropic smoothness. This paper is only a first step in this direction, for the expected rate of convergence in these two models would be $n^{-2s/(2s+1)}$ when the link function has smoothness s in some sense. Adaptive estimation by aggregation, including the Cox and Aalen models, will be addressed in a forthcoming paper.

Theorem 6. *Grant the same assumptions as in Theorem 5 and let $\hat{\alpha}_n$ be the aggregated estimator from Definition 4.*

- *If Assumption 5 holds (single-index) with $\beta_0 \in B^{\boldsymbol{s}}_{2,\infty}$, where $\boldsymbol{s} = (s_1, s_2)$ satisfies $1 < s_i \leq l_i$ for $i = 1, 2$, we have*

$$\mathsf{E}^{2n}\|\hat{\alpha}_n - \alpha_0\|^2 \leq cn^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+2)}$$

  for n large enough, where $c = c_{b,T,\boldsymbol{s},d,\|f_X\|_\infty,|\beta_0|_{B^{\boldsymbol{s}}_{2,\infty}},v_0,b_0,c_0}$.

- *Otherwise, we have, when $\alpha_0 \in B^{\boldsymbol{s}}_{2,\infty}$, where $\boldsymbol{s} = (s_1, \ldots, s_{d+1})$ satisfies $(d+1)/2 < s_i \leq l_i$ for each $i = 1, \ldots, d+1$, that*

$$\mathsf{E}^{2n}\|\hat{\alpha}_n - \alpha_0\|^2 \leq cn^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+d+1)},$$

  for n large enough, where $c = c_{b,T,\boldsymbol{s},d,\|f_X\|_\infty,|\alpha_0|_{B^{\boldsymbol{s}}_{2,\infty}}}$.

The proof of this theorem is given in Section 6. This theorem proves that $\hat{\alpha}_n$ adapts to the smoothness of the intensity, and to its structure: if the single-index model (38) holds, then the rate is $n^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+2)}$, which is the optimal rate when X is one-dimensional. Otherwise, the rate of convergence is $n^{-2\bar{\boldsymbol{s}}/(2\bar{\boldsymbol{s}}+d+1)}$ when the covariate is d-dimensional. Of course, this result is not surprising, since any kind of estimator can be used in the dictionary to be aggregated. However, note that the proof of Theorem 6 involves a technical tool concerning counting processes, namely a concentration inequality for the likelihood ratio between two indexes in S_+^{d-1}, see Lemma 4 in Section 6.

6 Proofs

Proof of Proposition 1

Proof of Proposition 1. Let us define the process

$$Z_n(\alpha, t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \alpha(u, X_i) dM^i(u) := \sum_{i=1}^{n} Z_n^i(\alpha, t),$$

so that $Z_n(\alpha) = Z_n(\alpha, 1)$. The predictable variation of M^i is given by $\langle M^i(t) \rangle = \int_0^t \alpha_0(u, X_i) Y^i(u) du$, so we have

$$\langle Z_n^i(\alpha, t) \rangle = \frac{1}{n} \int_0^t \alpha(u, X_i)^2 \alpha_0(u, X_i) Y^i(u) du$$

for any $t \in [0, 1]$. Moreover, we have $\Delta M^i(t) \in \{0, 1\}$ for any $i = 1, \ldots, n$ since the counting processes N^i have an intensity. We can write $Z_n^i = Z_n^{i,c} + Z_n^{i,d}$ where $Z_n^{i,c}$

is a continuous martingale and where $Z_n^{i,d}$ is a purely discrete martingale (see e.g. Liptser and Shiryayev (1989)). Let $h > 0$ be fixed and define $U_h^i(t) := hZ_n^i(\alpha, t) - S_h^i(t)$, where $S_h^i(t)$ is the compensator of

$$\frac{1}{2}h^2\langle Z_n^{i,c}(\alpha, t)\rangle + \sum_{s\le t}\left(\exp(h|\Delta Z_n^i(\alpha, s)|) - 1 - h|\Delta Z_n^i(\alpha, s)|\right). \tag{45}$$

We know from the proof of Lemma 2.2 and Corollary 2.3 of van de Geer (1995), see also Liptser and Shiryayev (1989), that $\exp(U_h^i(t))$ is a super-martingale. Then, if $S_h := \sum_{i=1}^n S_h^i$, $U_h := \sum_{i=1}^n U_h^i$, we have

$$\begin{aligned}\mathsf{E}^n[e^{hZ_n(\alpha)}\mathbf{1}_{\langle Z_n(\alpha)\rangle\le\delta^2}] &\le \left(\mathsf{E}^n[e^{2U_h(1)}]\right)^{1/2}\left(\mathsf{E}^n[e^{2S_h(1)}\mathbf{1}_{\langle Z_n(\alpha)\rangle\le\delta^2}]\right)^{1/2}\\ &\le \left(\mathsf{E}^n[e^{2S_h(1)}\mathbf{1}_{\langle Z_n(\alpha)\rangle\le\delta^2}]\right)^{1/2}.\end{aligned} \tag{46}$$

The last inequality holds since $\exp(U_h^i(t)) = \exp(hZ_n^i(\alpha, t) - S_h^i(t))$ are independent super-martingales with $U_h^i(0) = 0$, so that $\mathsf{E}[\exp(2U_h^i(t))] \le 1$, for $i = 1, \dots, n$. Let us decompose $M^i = M^{i,c} + M^{i,d}$, with $M^{i,c}$ a continuous martingale and $M^{i,d}$ a purely discrete martingale. The process $V_2^i(t) := \langle M^i(t)\rangle$ is the compensator of the quadratic variation process $[M^i(t)] = \langle M^{i,c}(t)\rangle + \sum_{s\le t}|\Delta M^i(t)|^2$. If $k \ge 3$, we define $V_k^i(t)$ as the compensator of the k-variation process $\sum_{s\le t}|\Delta M^i(t)|^k$ of $M^i(t)$. Since $\Delta M^i(t) \in \{0, 1\}$ for all $0 \le t \le 1$, the V_k^i are all equal for $k \ge 3$ and such that $V_k^i(t) \le V_2^i(t)$, for all $k \ge 3$. The process $S_h^i(t)$ has been defined as the compensator of (45). As a consequence, we have:

$$\begin{aligned}S_h^i(t) &= \sum_{k\ge2}\frac{1}{k!}\Big(\frac{h}{\sqrt{n}}\Big)^k\int_0^t|\alpha(u, X_i)|^k dV_k^i(u)\\ &\le \int_0^t\alpha(u, X_i)^2 dV_2^i(u)\times\sum_{k\ge2}\frac{\|\alpha\|_\infty^{k-2}}{k!}\Big(\frac{h}{\sqrt{n}}\Big)^k\end{aligned}$$

and if $\langle Z_n(\alpha)\rangle \le \delta^2$

$$S_h(1) \le \frac{n\delta^2}{\|\alpha\|_\infty^2}\Big(\exp\Big(\frac{h\|\alpha\|_\infty}{\sqrt{n}}\Big) - 1 - \frac{h\|\alpha\|_\infty}{\sqrt{n}}\Big).$$

Thus, plugging this in (46) gives

$$\psi_{n,\delta}(h) \le \frac{n\delta^2}{\|\alpha\|_\infty^2}\Big(\exp\Big(\frac{h\|\alpha\|_\infty}{\sqrt{n}}\Big) - 1 - \frac{h\|\alpha\|_\infty}{\sqrt{n}}\Big)$$

for any $h > 0$. Now, choosing

$$h := \frac{\sqrt{n}}{\|\alpha\|_\infty}\log\Big(\frac{z\|\alpha\|_\infty}{\delta^2\sqrt{n}} + 1\Big)$$

entails (16). □

Proof of Theorem 1

Proof of Theorem 1. First, note that (2) entails

$$\ell_\alpha(X,(Y_t),(N_t)) = \ell'_\alpha(X,(Y_t)) - 2\int_0^1 \alpha(t,X)dM(t),$$

where ℓ'_α is the loss function

$$\ell'_\alpha(x,(y_t)) := \int_0^1 \alpha(t,x)^2 y(t)dt - 2\int_0^1 \alpha(t,x)\alpha_0(t,x)y(t)dt.$$

So, the following decomposition holds:

$$(P-P_n)(\ell_\alpha - \ell_{\alpha_*}) = (P-P_n)(\ell'_\alpha - \ell'_{\alpha_*}) + \frac{2}{\sqrt{n}} Z_n(\alpha_* - \alpha),$$

where we recall that $Z_n(\cdot)$ is given by (14), and where $P(\ell'_\alpha) := \mathsf{E}[\ell'_\alpha(X,(Y_t))]$ and $P_n(\ell'_\alpha) := \frac{1}{n}\sum_{i=1}^n \ell'_\alpha(X_i,(Y_t^i))$. First, let us prove the concentration inequality for $\sup_{\alpha\in A}(Z_n(\alpha_*) - Z_n(\alpha))$. The proof follows the lines of the proof of Theorem 1.2.7 in Talagrand (2005). Consider admissible sequences $(\mathcal{B}_j)_{j\geq 0}$ and $(\mathcal{C}_j)_{j\geq 0}$ such that

$$\sum_{j\geq 0} 2^j \Delta(B_j(\alpha), d_\infty) \leq 2\gamma_1(A,d_\infty) \quad \text{and} \quad \sum_{j\geq 0} 2^{j/2}\Delta(C_j(\alpha),d_2) \leq 2\gamma_2(A,d_2)$$

for any $\alpha \in A$. We construct partitions $\mathcal{A}_j$ of A as follows. Set $\mathcal{A}_0 = \{A\}$ and for $j \geq 1$, $\mathcal{A}_j$ is the partition generated by $\mathcal{B}_{j-1}$ and $\mathcal{C}_{j-1}$, namely the partition consisting of every set $B \cap C$ where $B \in \mathcal{B}_{j-1}$ and $C \in \mathcal{C}_{j-1}$. Note that $|\mathcal{A}_j| \leq (2^{2^{j-1}})^2 = 2^{2^j}$ so that $(\mathcal{A}_j)$ is admissible. Define a sequence $(A_j)_{j\geq 0}$ of increasing subsets of A by taking exactly one element in each set of $\mathcal{A}_j$. Such a set A_j is then used as an approximation of A, and is such that $|A_j| \leq 2^{2^j}$. Define $\pi_j(\alpha)$ by the relation

$$A_j \cap A_j(\alpha) = \{\pi_j(\alpha)\},$$

and take $\pi_0(\alpha) = \alpha_*$. In view of Lemma 1, we have with a probability larger than $1 - 2\exp(-(x+2^{j+1}))$:

$$\begin{aligned} Z_n(\pi_{j-1}(\alpha)) - Z_n(\pi_j(\alpha)) \leq{}& C_0 d_2(\pi_j(a),\pi_{j-1}(\alpha))\sqrt{x+2^{j+1}} \\ &+ (C_0+1)\frac{d_\infty(\pi_j(\alpha),\pi_{j-1}(\alpha))(x+2^{j+1})}{\sqrt{n}}. \end{aligned}$$

Now, for a fixed $\alpha \in A$, decompose the increment $Z_n(\alpha_*) - Z_n(\alpha)$ along the *chain* $(\pi_j(\alpha))_{j\geq 0}$:

$$Z_n(\alpha_*) - Z_n(\alpha) = \sum_{j\geq 1} \big(Z_n(\pi_{j-1}(\alpha)) - Z_n(\pi_j(\alpha))\big),$$

and note that the number of pairs $\{\pi_j(\alpha),\pi_{j-1}(\alpha)\}$ is at most $2^{2^j} \times 2^{2^{j-1}} \leq 2^{2^{j+1}}$. This gives, together with union bounds for each term of the chain:

$$\begin{aligned} \sup_{\alpha\in A}(Z_n(\alpha_*) - Z_n(\alpha)) \leq \sup_{\alpha\in A}\sum_{j\geq 1}\Big(& C_0\sqrt{x+2^{j+1}}\, d_2(\pi_j(\alpha),\pi_{j-1}(\alpha)) \\ & + \frac{C_0+1}{\sqrt{n}}(x+2^{j+1}) d_\infty(\pi_j(\alpha),\pi_{j-1}(\alpha))\Big) \end{aligned}$$

with a probability larger than $1 - 2\sum_{j\geq 1} 2^{2^{j+1}} \exp(-(x + 2^{j+1})) \geq 1 - L\exp(-x)$ (with $L \approx 0.773$). But, for any $j \geq 2$, $\pi_j(\alpha), \pi_{j-1}(\alpha) \in A_{j-1}(\alpha) \subset B_{j-2}(\alpha)$, so $d_\infty(\pi_j(\alpha), \pi_{j-1}(\alpha)) \leq \Delta(B_{j-2}(\alpha), d_\infty)$ and $d_\infty(\pi_1(\alpha), \pi_0(\alpha)) \leq \Delta(B_0(\alpha), d_\infty) = \Delta(A, d_\infty)$. Doing the same for d_2, we obtain that, with probability $\geq 1 - L\exp(-x)$:

$$\sup_{\alpha\in A}(Z_n(\alpha_*) - Z_n(\alpha)) \leq 2C_0(1+\sqrt{x})\gamma_2(A, d_2) + \frac{2(C_0+1)}{\sqrt{n}}(1+x)\gamma_1(A, d_\infty). \quad (47)$$

We can do the same job for $\sup_{\alpha\in A}(P - P_n)(\ell'_\alpha - \ell'_{\alpha_*})$. Note that

$$\begin{aligned}
&\ell_\alpha(X, (Y_t)) - \ell_{\alpha_*}(X, (Y_t)) \\
&= \int_0^1 (\alpha(t, X) - \alpha_*(t, X))(\alpha(t, X) + \alpha_*(t, X) - 2\alpha_0(t, X))Y(t)dt,
\end{aligned}$$

so using Assumptions 1 and 2, we have $|\ell_\alpha(X, (Y_t)) - \ell_{\alpha_*}(X, (Y_t))| \leq 2(b + \|\alpha_0\|_\infty)\|\alpha - \alpha_*\|_\infty$ and

$$\mathsf{E}[\ell_\alpha(X, (Y_t)) - \ell_{\alpha_*}(X, (Y_t)))^2] \leq 4(b + \|\alpha_0\|_\infty)^2\|\alpha - \alpha_*\|^2.$$

Therefore, the Bernstein's inequality (for the sum of i.i.d. random variables) entails that

$$(P - P_n)(\ell'_\alpha - \ell'_{\alpha_*}) \leq 2(b + \|\alpha_0\|_\infty)\Big(\frac{\|\alpha - \alpha_*\|\sqrt{2x}}{\sqrt{n}} + \frac{\|\alpha - \alpha_*\|_\infty x}{n}\Big)$$

holds with a probability larger than $1 - e^{-x}$. Then, we can apply again the generic chaining argument to prove that with a probability larger than $1 - Le^{-x}$:

$$\sup_{\alpha\in A}(P - P_n)(\ell'_\alpha - \ell'_{\alpha_*}) \leq 4(b + \|\alpha_0\|_\infty)\Big(\frac{\gamma_2(A, d_2)(1+\sqrt{x})}{\sqrt{n}} + \frac{\gamma_1(A, d_\infty)(1+x)}{n}\Big).$$

This concludes the proof of the Theorem. □

Proof of Theorem 4

Proof of Theorem 4. Recall that the *linearized risk* over $A(\Lambda)$ is given by

$$\mathsf{R}(\theta) := \sum_{\lambda\in\Lambda} \theta_\lambda P(\ell_{\alpha_\lambda})$$

for $\theta \in \Theta$, where we recall that

$$\Theta = \{\theta \in \mathbb{R}^M : \theta_\lambda \geq 0, \sum_{\lambda\in\Lambda} \theta_\lambda = 1\},$$

and the *linearized empirical risk* is given by

$$\mathsf{R}_n(\theta) = \sum_{\lambda\in\Lambda} \theta_\lambda P_n(\ell_{\alpha_\lambda}).$$

We recall that the mixing estimator $\hat{\alpha}$ is given by

$$\hat{\alpha} := \sum_{\lambda\in\Lambda} \hat{\theta}_\lambda \alpha_\lambda,$$

where the Gibbs weights $\hat{\theta} = (\hat{\theta}_\lambda)_{\lambda\in\Lambda} := (\theta(\alpha_\lambda))_{\lambda\in\Lambda}$ are given by (35) and are the unique solution of the minimization problem (36). By convexity of the risk, we have for any $\epsilon > 0$:

$$P(\ell_{\hat{\alpha}} - \ell_{\alpha_0}) \le (1+\epsilon)(\mathsf{R}_n(\hat{\theta}) - P_n(\ell_{\alpha_0})) + \mathcal{R}_n,$$

where we introduced the residual term

$$\begin{aligned}\mathcal{R}_n &:= \mathsf{R}(\hat{\theta}) - P(\ell_{\alpha_0}) - (1+\epsilon)(\mathsf{R}_n(\hat{\theta}) - P_n(\ell_{\alpha_0})) \\ &= \sum_{\lambda\in\Lambda} \hat{\theta}_\lambda \Big(P(\ell_{\alpha_\lambda} - \ell_{\alpha_0}) - (1+\epsilon) P_n(\ell_{\alpha_\lambda} - \ell_{\alpha_0}) \Big).\end{aligned}$$

Let $\hat{\lambda}$ be such that $\alpha_{\hat{\lambda}}$ is the empirical risk minimizer in $A(\Lambda)$, namely

$$P_n(\ell_{\alpha_{\hat{\lambda}}}) = \min_{\lambda\in\Lambda} P_n(\ell_{\alpha_\lambda}).$$

Since

$$\sum_{\lambda\in\Lambda} \hat{\theta}_\lambda \log\Big(\frac{\hat{\theta}_\lambda}{1/M}\Big) = K(\hat{\theta}, u) \ge 0,$$

where $K(\hat{\theta}, u)$ denotes the Kullback-Leibler divergence between the weights $\hat{\theta}$ and the uniform weights $u := (1/M)_{\lambda\in\Lambda}$, we have

$$\begin{aligned}\mathsf{R}_n(\hat{\theta}) &\le \mathsf{R}_n(\hat{\theta}) + \frac{T}{n} K(\hat{\theta}, u) \\ &= \mathsf{R}_n(\hat{\theta}) + \frac{T}{n} \sum_{\lambda\in\Lambda} \hat{\theta}_\lambda \log \hat{\theta}_\lambda + \frac{T \log M}{n} \\ &\le \mathsf{R}_n(e_{\hat{\lambda}}) + \frac{T \log M}{n} \\ &= P_n(\ell_{\alpha_{\hat{\lambda}}}) + \frac{T \log M}{n},\end{aligned}$$

where $e_\lambda \in \Theta$ is the vector with all its coordinates equal to 0 excepted for the λ-th which is equal to 1. This gives

$$P(\ell_{\hat{\alpha}} - \ell_{\alpha_0}) \le (1+\epsilon) \min_{\lambda\in\Lambda} P_n(\ell_{\alpha_\lambda} - \ell_{\alpha_0}) + \mathcal{R}_n,$$

and consequently

$$\mathsf{E}^n \|\hat{\alpha} - \alpha_0\|^2 \le (1+\epsilon) \min_{\lambda\in\Lambda} \|\alpha_\lambda - \alpha_0\|^2 + (1+\epsilon) \frac{T \log M}{n} + \mathsf{E}^n[\mathcal{R}_n].$$

Hence, it remains to prove that for some constant $C = C_{\epsilon,b,\|\alpha\|_\infty}$, we have

$$\mathsf{E}^n[\mathcal{R}_n] \le \frac{C \log M}{n}. \tag{48}$$

Since $\mathsf{R}(\cdot)$ and $\mathsf{R}_n(\cdot)$ are linear on Θ, we have

$$\mathcal{R}_n \le \max_{\alpha\in A(\Lambda)} \Big((1+\epsilon)\big(P(\ell_\alpha - \ell_{\alpha_0}) - P_n(\ell_\alpha - \ell_{\alpha_0})\big) - \epsilon P(\ell_\alpha - \ell_{\alpha_0}) \Big).$$

The following decomposition holds (see Section 3.4):

$$(P-P_n)(\ell_\alpha-\ell_{\alpha_0})=(P-P_n)(\ell'_\alpha-\ell'_{\alpha_0})+\frac{2}{\sqrt{n}}Z_n(\alpha_0-\alpha).$$

The Bernstein's inequality for the sum of i.i.d. variables (see the proof of Theorem 1) gives

$$(P-P_n)(\ell'_\alpha-\ell'_{\alpha_0})\le(b+\|\alpha_0\|_\infty)\Big(\frac{\|\alpha-\alpha_0\|\sqrt{2x}}{\sqrt{n}}+\frac{\|\alpha-\alpha_0\|_\infty x}{n}\Big),$$

so together with Lemma 1, and since $P(\ell_\alpha-\ell_{\alpha_0})=\|\alpha-\alpha_0\|^2$, we obtain that

$$(P-P_n)(\ell_\alpha-\ell_{\alpha_0})\le\frac{C^1_{\|\alpha_0\|_\infty,b}\sqrt{2xP(\ell_\alpha-\ell_{\alpha_0})}}{\sqrt{n}}+\frac{C^2_{\|\alpha_0\|_\infty,b}x}{n}$$

with probability larger than $1-3e^{-x}$, where $C^1_{\|\alpha_0\|_\infty,b}:=C_{\|\alpha_0\|_\infty}/\sqrt{2}+b+\|\alpha_0\|_\infty$ and $C^2_{\|\alpha_0\|_\infty,b}:=(C_{\|\alpha_0\|_\infty}+1+b+\|\alpha\|_\infty)(b+\|\alpha_0\|_\infty)$, with $C_{\|\alpha_0\|_\infty}$ given in Lemma 1. Now, using the fact that

$$\frac{C^1_{\|\alpha_0\|_\infty,b}\sqrt{2xP(\ell_\alpha-\ell_{\alpha_0})}}{\sqrt{n}}\le\frac{\epsilon}{1+\epsilon}P(\ell_\alpha-\ell_{\alpha_0})+\frac{(1+\epsilon)(C^1_{\|\alpha_0\|_\infty,b})^2}{\epsilon}\frac{x}{n},$$

we obtain that with a probability larger than $1-3e^{-x}$:

$$(1+\epsilon)\big(P(\ell_\alpha-\ell_{\alpha_0})-P_n(\ell_\alpha-\ell_{\alpha_0})\big)-\epsilon P(\ell_\alpha-\ell_{\alpha_0})\le C_{\epsilon,\|\alpha_0\|_\infty,b}\frac{x}{n},$$

where $C_{\epsilon,\|\alpha_0\|_\infty,b}:=(C^1_{\|\alpha_0\|_\infty,b})^2(1+\epsilon)^2/\epsilon+(1+\epsilon)C^2_{\|\alpha_0\|_\infty,b}$. This subexponential deviation entails that for any $x>0$:

$$\mathsf{E}^n\big[\mathcal{R}_n\big]\le 2x+\frac{3MC\exp(-nx/C)}{n},$$

where $C=C_{\epsilon,\|\alpha_0\|_\infty,b}$. If we denote by $x(y)$ the unique solution of $x=y\exp(-x)$, where $y>0$, we obtain

$$\mathsf{E}^n\big[\mathcal{R}_n\big]\le\frac{5C\log M}{n}$$

for the choice $x=Cx(M)/n$, since we have $x(M)\le\log M$. This concludes the proof of Theorem 4. □

Proof of Theorem 6. Assume for now that (38) holds. Take $v_\Delta\in S^+_\Delta$ such that $|v_\Delta-v_0|_2\le\Delta$, and let $m^*=(m^*_1,m^*_2)$ be the oracle dimension of the sieve for the link function, that satisfies (37) with $d=1$. Denote for short the oracle estimator

$$\bar\alpha_*=\bar\beta_{m^*,v_\Delta}(\cdot,v_\Delta^\top\cdot),$$

that is, the element of A_{m^*} that minimizes the empirical risk computed using the training sample $D_{n,1}(v_\Delta)$.

Note that the cardinality of S^+_Δ is smaller than c/Δ^{d-1}, where $\Delta=(n\log n)^{-1/2}$, so the cardinality of the whole dictionary $\{\bar\alpha_m:m\in\mathcal{M}_n\}\cup\{\bar\alpha^{\mathrm{SIM}}_{m,v}:m\in\mathcal{M}^{\mathrm{SIM}}_n,v\in$

$S_\Delta^{d-1}\}$ is of order $cn^{(d-1)/2}(\log n)^{2+1/2} + (\log n)^{d+1}$. As a consequence, Theorem 4 gives

$$\mathsf{E}^{2n}\|\hat{\alpha}_n - \alpha_0\|^2 \leq 2\mathsf{E}^n\|\bar{\alpha}_* - \alpha_0\|^2 + c\frac{\log n}{n}.$$

Note that (41) entails $\|\beta_0(\cdot, v_\Delta^\top \cdot) - \beta_0(\cdot, v_0^\top \cdot)\|^2 \leq c\Delta^2 = c/(n\log n)$. Hence,

$$\|\bar{\alpha}_* - \alpha_0\|^2 \leq 2\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 + \frac{2c}{n\log n}.$$

We shall denote in what follows by E_v^n the expectation wrt P_v^n, the joint law of the observations when the intensity writes $\beta_0(\cdot, v^\top \cdot)$ (the true index is v). For two indexes $v, v_0 \in S_+^{d-1}$, we introduce the following likelihood ratio:

$$L_n(v_0, v) = \frac{\mathrm{d}\mathsf{P}^n_{\beta_0(\cdot, v_0^\top \cdot)}}{\mathrm{d}\mathsf{P}^n_{\beta_0(\cdot, v^\top \cdot)}},$$

which is the likelihood ratio of the training data $D_{n,1}$ "between" the two indexes v and v_0. It can be explicitly computed using Jacod's formula, see Appendix A.2 below. Of course, when v and v_0 are close to each other, we expect $L_n(v_0, v)$ to be small. This is the statement of the next Lemma.

Lemma 4. *Grant Assumption 5, and let $v, v_0 \in S_+^{d-1}$ be such that $\|v - v_0\|_2 \leq \Delta_n$, where $\Delta_n = (n\log n)^{-1/2}$. Then, if n is large enough, one has for any $x > 0$:*

$$\mathsf{P}_{v_0}^n[L_n(v_0, v) \geq x] \leq \sqrt{x} n^{-c(\log x)^2},$$

where $c = b_0/(2dc_0^2)$.

The proof of this Lemma can be found below. It uses the same kind of arguments as the proof of Proposition 1. Let $x > 0$ to be chosen later on, and decompose the expectation over $\{L_n(v_0, v_\Delta) > x\}$ and $\{L_n(v_0, v_\Delta) \leq x\}$ to get

$$\begin{aligned}\mathsf{E}_{v_0}^n\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 = \mathsf{E}_{v_\Delta}^n[\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 \mathbf{1}_{L_n(v_0, v_\Delta) \leq x} L_n(v_0, v_\Delta)] \\ + \mathsf{E}_{v_0}^n[\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 \mathbf{1}_{L_n(v_0, v_\Delta) > x}]\end{aligned}$$

so using Assumption 3 and Lemma 4, we obtain

$$\mathsf{E}_{v_0}^n\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 \leq x\mathsf{E}_{v_\Delta}^n\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 + 4b^2\sqrt{x}n^{-c(\log x)^2},$$

so for $x = e^{1/\sqrt{c}}$, we have

$$\mathsf{E}_{v_0}^n\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 \leq c\mathsf{E}_{v_\Delta}^n\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 + \frac{c}{n}.$$

But, $\mathsf{E}_{v_\Delta}^n\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2$ is nothing but the risk of the minimizer $\bar{\beta}_{m^*}$ of the empirical risk $R_{n,1}^{(v_\Delta)}$ over the sieve A_{m^*}: in this risk, the "true covariate" is now $v_\Delta^\top X$. Indeed,

$$\begin{aligned}&\mathsf{E}_{v_\Delta}^n\|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 \\ &\quad = \mathsf{E}_{v_\Delta}^n\Big[\int_0^1 \int (\bar{\beta}_{m^*}(t, v_\Delta^\top x) - \beta_0(t, v_\Delta^\top x))^2 \mathsf{E}[Y(t)|X = x]dtP_X(dx)\Big] \\ &\quad = \mathsf{E}_{v_\Delta}^n\Big[\int_0^1 \int (\bar{\beta}_{m^*}(t, x') - \beta_0(t, x'))^2 \mathsf{E}[Y(t)|v_\Delta^\top X = x']dtP_{v_\Delta^\top X}(dx')\Big],\end{aligned}$$

so conducting the same analysis as in Section 5.1, we can prove that the choice of m^* entails that

$$\mathsf{E}^n_{v_\Delta} \|\bar{\alpha}_* - \beta_0(\cdot, v_\Delta^\top \cdot)\|^2 \leq c n^{-2\bar{s}/(2\bar{s}+2)}.$$

This concludes the proof of Theorem 6 in the single-index case. If (38) does not hold, then in the oracle inequality we take the oracle purely nonparametric element, using the same analysis as in Section 5.1. □

Proof of Lemma 4. In view of Equation (54), see Appendix A.2, we can write, using (2):

$$\begin{aligned}
\log L_n(v_0, v) &= \sum_{i=1}^n \int_0^1 \Big(\mathcal{L}_{v_0,v}(t, X_i) dN^i(t) - \Upsilon_{v_0,v}(t, X_i) Y^i(t) dt \Big) \\
&= \sum_{i=1}^n \int_0^1 \mathcal{L}_{v_0,v}(t, X_i) dM^i(t) \\
&+ \sum_{i=1}^n \int_0^1 \Big\{ \mathcal{L}_{v_0,v}(t, X_i) \beta_0(t, v_0^\top X_i) - \Upsilon_{v_0,v}(t, X_i) \Big\} Y^i(t) dt,
\end{aligned}$$

where we shall use the notations

$$\begin{aligned}
\Upsilon_{v_0,v}(t, X_i) &:= \beta_0(t, v_0^\top X_i) - \beta_0(t, v^\top X_i) \\
\mathcal{L}_{v_0,v}(t, X_i) &:= \log \beta_0(t, v_0^\top X_i) - \log \beta_0(t, v^\top X_i)
\end{aligned}$$

throughout the proof the Lemma. Now, fix some $h > 0$ (to be chosen later on) and write

$$\begin{aligned}
&\mathsf{P}^n_{v_0}[L_n(v_0, v) \geq x] \\
&\leq \mathsf{E}^n_{v_0}[L_n(v_0, v)^h] e^{-h \log x} \\
&= \mathsf{E}^n_{v_0} \Big[\exp\Big(\sum_{i=1}^n h \int_0^1 \mathcal{L}_{v_0,v}(t, X_i) dM^i(t) \\
&+ h \sum_{i=1}^n \int_0^1 \Big\{ \mathcal{L}_{v_0,v}(t, X_i) \beta_0(t, v_0^\top X_i) - \Upsilon_{v_0,v}(t, X_i) \Big\} Y^i(t) dt - h \log x \Big) \Big].
\end{aligned}$$

We follow the main steps of the proof of Proposition 1. Define

$$\tilde{U}^i_h(t) := h \int_0^t \mathcal{L}_{v_0,v}(s, X_i) dM^i(s) - \tilde{S}^i_h(t) := h O^i(t) - \tilde{S}^i_h(t),$$

where $\tilde{S}^i_h(t)$ is the compensator of

$$\frac{1}{2} h^2 \langle O^{i,c}(t) \rangle + \sum_{s \leq t} \left(\exp(h|\Delta O^i(s)|) - 1 - h|\Delta O^i(s)| \right), \tag{49}$$

where $O^{i,c}$ is the continuous part of the process O^i. We know from the proof of Lemma 2.2 and Corollary 2.3 of van de Geer (1995), see also Liptser and Shiryayev

(1989), that $\exp(\tilde{U}_h^i) = \exp(hO^i - \tilde{S}_h^i)$, for $i = 1, \ldots, n$ are i.i.d. super-martingales. As a consequence, we get:

$$\mathsf{P}_{v_0}[L_n(v_0, v) \geq x] \leq \mathsf{E}_{v_0}^{1/2}\Big[\exp\Big(2\sum_{i=1}^n \tilde{U}_h^i(t)\Big)\Big]\, \mathsf{E}_{v_0}^{1/2}[\mathbb{L}_n(1)] \leq \mathsf{E}_{v_0}^{1/2}[\mathbb{L}_n(1)],$$

where

$$\mathbb{L}_n(1) := \exp\Big(2\sum_{i=1}^n \Big\{\tilde{S}_h^i(1) + h\int_0^1 \Big\{\mathcal{L}_{v_0,v}(t, X_i)\beta_0(t, v_0^\top X_i) - \Upsilon_{v_0,v}(t, X_i)\Big\}Y^i(t)dt\Big\} - 2h\log C\Big).$$

We are now establishing an upper bound for $\mathsf{E}_{v_0}^{1/2}[\mathbb{L}_n(1)]$. Looking closer to the process $\tilde{S}_h^i(t)$, we can write:

$$\tilde{S}_h^i(t) = \sum_{k\geq 2} \frac{h^k}{k!}\int_0^t |\mathcal{L}_{v_0,v}(s, X_i)|^k dV_k^i(s),$$

where the processes V_k^i have been defined in the proof of Proposition 1. Assumption 5 and the fact that $\|v - v_0\|_2 \leq \Delta$ gives

$$|\Upsilon_{v_0,v}(t, x)| \leq c_0\sqrt{d}\Delta := \epsilon \tag{50}$$

for any $t \geq 0$ and $x \in [0, 1]^d$. In particular, we have $|\Upsilon_{v_0,v}(t, x)| \leq b_0/2$ when n is large enough. This allows to write:

$$|\mathcal{L}_{v_0,v}(t, X_i)| \leq \Psi_{1/\beta_0(t,v_0^\top X_i)}(\Upsilon_{v_0,v}(t, X_i)) \times (1/\beta_0(t, v_0^\top X_i))$$

where $\Psi_a(x) := -\log(1 - ax)/a$ for $a > 0$ and $x < 1/a$. Since $\Psi_a(x) = -\log(1 - ax)/a \leq x + ax^2$ for any $x \in [0, 1/(2a)]$, we obtain

$$\begin{aligned}|\mathcal{L}_{v_0,v}(t, X_i)| &\leq \frac{|\Upsilon_{v_0,v}(t, X_i)|}{\beta_0(t, v_0^\top X_i) \wedge \beta_0(t, v^\top X_i)}\Big(1 + \frac{|\Upsilon_{v_0,v}(t, X_i)|}{\beta_0(t, v_0^\top X_i) \wedge \beta_0(t, v^\top X_i)}\Big) \\ &\leq \Big(\frac{\epsilon}{b_0}\Big)\Big(1 + \frac{\epsilon}{b_0}\Big).\end{aligned}$$

We can write, as a consequence:

$$\begin{aligned}\tilde{S}_h^i(t) &= \sum_{k\geq 2} \frac{h^k}{k!}\int_0^t |\mathcal{L}_{v_0,v}(s, X_i)|^k dV_k^i(s) \\ &\leq \int_0^t |\mathcal{L}_{v_0,v}(s, X_i)|^2\beta_0(s, v_0^\top X_i)Y^i(s)ds \times \sum_{k\geq 2}\frac{h^k}{k!}\Big(\frac{\epsilon}{b_0}\Big)^{k-2}\Big(1 + \frac{\epsilon}{b_0}\Big)^{k-2} \\ &\leq \int_0^t |\mathcal{L}_{v_0,v}(s, X_i)|^2\beta_0(s, v_0^\top X_i)Y^i(s)ds \times \frac{h^2}{2}(1 + c_h),\end{aligned}$$

where

$$c_h := 2\sum_{k\geq 1}\frac{h^k}{(k+2)!}\Big(\frac{\epsilon}{b_0}\Big)^k\Big(1 + \frac{\epsilon}{b_0}\Big)^k.$$

Note that $c_h \leq 1$ for $h\epsilon$ and ϵ small enough. We obtain:

$$\mathsf{E}^n_{v_0}[\mathbb{L}_n(1)] \leq \mathsf{E}_{v_0}\Big[\exp\Big(n\Big\{\int_0^1 |\mathcal{L}_{v_0,v}(t,X)|^2\beta_0(t,v_0^\top X)Y(t)dt \times h^2(1+c_h) \\ + 2h\int_0^1 \Big\{\mathcal{L}_{v_0,v}(t,X)\beta_0(t,v_0^\top X) - \Upsilon_{v_0,v}(t,X)\Big\}Y(t)dt\Big\} \\ - 2h\log x\Big)\Big].$$

Using again the above trick involving the function Ψ_a, we obtain:

$$\mathcal{L}_{v_0,v}(t,X_i)\beta_0(t,v_0^\top X_i) - \Upsilon_{v_0,v}(t,X_i) \leq \frac{\Upsilon_{v_0,v}(t,X_i)^2}{\beta_0(t,v^\top X_i)} \leq \frac{\epsilon^2}{b_0}.$$

Using the fact that $\log(x/y)^2 x \leq 2\epsilon^2/(x\wedge y)$ for any $x,y>0$ such that $|x-y| \leq \epsilon \leq (x\wedge y)/2$ and $\epsilon>0$ small enough [decompose over $\{x\leq y\}$ and $\{x>y\}$ and use again the previous majoration of $\Psi_a(x)$], we have in view of (50):

$$\mathcal{L}_{v_0,v}(t,x)^2\beta_0(t,v_0^\top x) \leq \frac{2\epsilon^2}{b_0}$$

for any $t\geq 0$ and $x\in[0,1]^d$ and ϵ small enough. In fine, we get, using the fact that $Y^i \leq 1$,

$$\mathsf{E}^n_{v_0}[\mathbb{L}_n(1)] \leq \mathsf{E}_{v_0}\Big[\exp\Big(n\int_0^1\Big\{\frac{2\epsilon^2h^2}{b_0}+\frac{2h\epsilon^2}{b_0}\Big\}Y^i(t)dt - 2h\log x\Big)\Big] \\ \leq \exp\Big(\frac{2n\epsilon^2h}{b_0}(1+h) - 2h\log x\Big)$$

for any $h>0$, so for the choice $h = b_0\log x/(2n\epsilon^2)$, we obtain

$$\mathsf{P}^n_{v_0}[L_n(v_0,v)\geq x] \leq \sqrt{x}\exp\Big(-\frac{b_0(\log x)^2}{2n\epsilon^2}\Big),$$

and the conclusion follows, since $\Delta = 1/\sqrt{n\log n}$ and $n\epsilon^2 = dc_0^2/\log n$. □

A Appendix

A.1 Some tools from approximation theory

Let us give two examples of sieves, that are spanned by localized basis. In each case, we give the control on $\bar{r}(A)$ and we give a standard but useful approximation result below. Note that other examples of sieves are available, see Barron et al. (1999) for instance.

A.1.1 Piecewise polynomials

Fix $l_1,\ldots,l_{d+1} \in \mathbb{N}$ and $m_1,\ldots,m_{d+1}\in\mathbb{N}$, and define the set $\mathscr{R}$ of rectangles $\prod_{i=1}^{d+1}[(j_i-1)2^{-m_i}, j_i2^{-m_i}[$ for $0\leq j_i\leq 2^{m_i}$, $i=0,\ldots,d+1$. So, $\mathscr{R}$ is a regular partition of $[0,1]^{d+1}$. Take $m=(m_1,\ldots,m_{d+1})$ and define A_m as the set of functions

$f : [0,1]^{d+1} \to \mathbb{R}$ such that for any $R \in \mathscr{R}$, the restriction of f to R coincides with a polynomial of degree not larger than l_i in the ith coordinate, for $i = 1, \ldots, d+1$. The dimension of A_m is then

$$D_m := \prod_{i=1}^{d+1} 2^{m_i}(l_i + 1),$$

and using Barron et al. (1999), see Section 3.2.1, we have, since $\mathscr{R}$ is a regular partition,

$$\bar{r}(A_m) \leq c_{l_1,\ldots,l_{d+1},d},$$

where $c_{l_1,\ldots,l_{d+1},d} = (\prod_{i=1}^{d+1}(l_i+1)(2l_i+1))^{1/2}$.

A.1.2 Wavelets

Consider a pair $\{\phi, \psi\}$ of scaling function and wavelet, where ψ has K vanishing moments. Then ϕ and ψ have a support width of at least $2K-1$, and there is a pair with minimal support, see Daubechies (1988). This is the starting point of the construction of an orthonormal wavelet basis of $\mathbb{L}^2[0,1]$, as proposed in Cohen et al. (1993). Roughly, the idea is to retain the interior scaling functions (those "far" from the edges 0 and 1), and to add adapted edge scaling functions, see Section 4 and Theorem 4.4 in Cohen et al. (1993). This construction allows to keep the orthonormality of the system and the number of vanishing moment unchanged, as well as the number 2^j of scaling function at each resolution j. More precisely, if l is such that $2^l \geq 2K$, consider for $j \geq l-1$:

$$\Psi_{j,k} := \begin{cases} \psi^0_{j,k} & \text{if } j \geq l \text{ and } k = 0, \ldots, K-1 \\ \psi_{j,k} & \text{if } j \geq l \text{ and } k = K, \ldots, 2^j - K - 1 \\ \psi^1_{j,k} & \text{if } j \geq l \text{ and } k = 2^j - K, \ldots, 2^j - 1 \\ \phi^0_{l,k} & \text{if } j = l-1 \text{ and } k = 0, \ldots, K-1 \\ \phi_{l,k} & \text{if } j = l-1 \text{ and } k = K, \ldots, 2^l - K - 1 \\ \phi^1_{l,k} & \text{if } j = l-1 \text{ and } k = 2^l - K, \ldots, 2^l - 1 \end{cases}$$

where $\phi_{j,k} = 2^{j/2}\phi(2^j \cdot - x)$ and $\psi_{j,k} = 2^{j/2}\psi(2^j \cdot - x)$ are the "interior" dilatations and translations of $\{\phi, \psi\}$, and $\phi^0_{j,k}, \psi^0_{j,k}, \phi^1_{j,k}, \psi^1_{j,k}$ are, at each resolution j, dilatations of $2K$ edge scaling functions and wavelets (K for each edge). We know from Cohen et al. (1993) that the collection

$$W := \{\Psi_{j,k} : j \geq l-1, k = 0, \ldots, 2^j - 1\}$$

is an orthonormal basis of $\mathbb{L}^2[0,1]$, and the interior and edge wavelets have K vanishing moments. Let $W^{(i)}, i = 1, \ldots, d+1$ be several collections W based on pairs $\{\phi^{(i)}, \psi^{(i)}\}$ (possibly with different numbers of vanishing moments). Then, the collection

$$\{\otimes_{i=1}^{d+1} \Psi^{(i)}_{j_i,k_i} : j_i \geq l_i - 1, k_i = 0, \ldots, 2^{j_i} - 1, i = 1, \ldots, d+1\},$$

where $\otimes_{i=1}^{d+1} \Psi^{(i)}_{j_i,k_i}(x_1, \ldots, x_{d+1}) = \prod_{i=1}^{d+1} \Psi^{(i)}_{j_i,k_i}(x_i)$, is an orthonormal basis of $\mathbb{L}^2[0,1]^{d+1}$ that has suitable approximation properties for a function with an anisotropic smoothness, see below. Let $m = (m_1, \ldots, m_{d+1}) \in \mathbb{N}^{d+1}$ be fixed, where $m_i \geq l_i$ for any

$i \in \{1, \dots, d+1\}$, and define the sieve

$$A_m := \text{span}\{\Psi_\lambda : \lambda \in \Lambda(m)\}, \tag{51}$$

where for $\lambda = (j_1, k_1, \dots, j_{d+1}, k_{d+1})$,

$$\Psi_\lambda := \otimes_{i=1}^{d+1} \Psi_{j_i,k_i}^{(i)},$$

and where

$$\Lambda(m) = \{(j_1, k_1, \dots, j_{d+1}, k_{d+1}) : l_i - 1 \leq j_i \leq m_i, \\ k_i = 0, \dots, 2^{j_i} - 1, i = 1, \dots, d+1\}$$

The dimension of A_m is $\prod_{i=1}^{d+1} D_{m_i}$, where $D_{m_i} = 2^{m_i} - 2^{l_i} \leq 2^{m_i}$. The control of $r(A_m)$ easily follows from the fact that if the resolution levels $j_i \geq l_i$ are fixed for any $i = 1, \dots, d+1$, the tensor products Ψ_λ have disjoint supports, excepted for a finite number of indexes k_i, that depends only on the support of the scaling and mother wavelet functions used in the construction of W. As a consequence, we have

$$r(A_m) \leq \frac{1}{\sqrt{D_m}} \sup_{\beta \neq 0} \frac{\| \sum_{\lambda \in \Lambda(m)} \beta_\lambda \psi_\lambda \|_\infty}{|\beta|_\infty} \leq c_\Psi,$$

where $D_m = \prod_{i=1}^{d+1} D_{m_i}$, $|\beta|_\infty = \sup_{\lambda \in \Lambda(m)} |\beta_\lambda|$ and where c_Ψ is a constant that depends only the scaling and mother wavelet functions used in the construction of the basis, and not on the resolution level m.

In the next section, we give the definition of the anisotropic Besov space, and recall a useful approximation result. The definitions and results presented here can be found in Triebel (2006), in particular in Chapter 5 which is about anisotropic spaces.

A.1.3 Anisotropic Besov space, approximation

Let $\{e_1, \dots, e_{d+1}\}$ be the canonical basis of $\mathbb{R}^{d+1}$ and $\boldsymbol{s} = (s_1, \dots, s_{d+1})$ with $s_i > 0$ be a vector of directional smoothness, where s_i corresponds to the smoothness in direction e_i. If $k \in \mathbb{N}$ and $x \in \mathbb{R}^{d+1}$, define

$$\mathscr{D}_e^k := \{x \in \mathbb{R}^{d+1} : x + je \in [0,1]^{d+1} \text{ for } j = 0, \dots, k\}.$$

If $f : [0,1]^{d+1} \to \mathbb{R}$, we define $\Delta_e^k f$ as the difference of order $k \geq 1$ and step $e \in [0,1]^{d+1}$, given by $\Delta_e^1 f(x) = f(x+e) - f(x)$ and $\Delta_e^k f(x) = \Delta_e^1(\Delta_e^{k-1} f)(x)$ for any $x \in \mathscr{D}_e^k$. We say that $f \in \mathbb{L}^2[0,1]^{d+1}$ belongs to the anisotropic Besov space $B_{2,\infty}^{\boldsymbol{s}}([0,1]^{d+1})$ if the semi-norm

$$|f|_{B_{2,\infty}^{\boldsymbol{s}}} := \sup_{t>0} \Big(\sum_{i=1}^{d+1} t^{-s_i} \sup_{h : |h| \leq t} \Big(\int_{\mathscr{D}_{he_i}^{k_i}} (\Delta_{he_i}^{k_i} f(x))^2 dx \Big)^{1/2} \Big) \tag{52}$$

is finite. We know that the norms

$$\|f\|_{B_{2,\infty}^{\boldsymbol{s}}} := \|f\|_2 + |f|_{B_{2,\infty}^{\boldsymbol{s}}}$$

are equivalent for any choice of $k_i > s_i$. Note that if $\boldsymbol{s} = (s, \dots, s)$ for some $s > 0$, then $B^{\boldsymbol{s}}_{2,\infty}$ is the standard isotropic Besov space. Moreover, the embedding $B^{\boldsymbol{s}}_{2,2} \subset B^{\boldsymbol{s}}_{2,\infty}$ holds. When $s = (s_1, \dots, s_{d+1})$ has integer coordinates, $B^{\boldsymbol{s}}_{2,2}$ is the anisotropic Sobolev space

$$B^{\boldsymbol{s}}_{2,2} = W^{\boldsymbol{s}}_2 = \Big\{ f \in \mathbb{L}^2 : \sum_{i=1}^{d+1} \Big\| \frac{\partial^{s_i} f}{\partial x_i^{s_i}} \Big\|_2 < \infty \Big\}.$$

If $\boldsymbol{s}$ has non-integer coordinates, then $B^{\boldsymbol{s}}_{2,2}$ is the anisotropic Bessel-potential space

$$H^{\boldsymbol{s}} = \Big\{ f \in \mathbb{L}^2 : \sum_{i=1}^{d+1} \Big\| (1 + |\xi_i|^2)^{s_i/2} \hat{f}(\xi) \Big\|_2 < \infty \Big\},$$

where $\hat{f}$ is the Fourier transform of f. If $f \in B^{\boldsymbol{s}}_{2,\infty}$, we can give a control on the approximation term $\inf_{\alpha \in A} \|\alpha - \alpha_0\|$, when A is spanned by piecewise polynomials or wavelets (see above). Indeed, the next Lemma is a direct consequence of the Jackson's estimate given in Hochmuth (2002), together with definition (52) of the Besov space. Note that this Lemma can be also found in Comte et al. (2008) and Lacour (2007).

Lemma 5. *Assume that $\alpha_0 \in B^{\boldsymbol{s}}_{2,\infty}$ where $\boldsymbol{s} = (s_1, \dots, s_{d+1})$ and let $l_i \geq s_i$ for $i = 1, \dots, d+1$. Let A_m be either*:

- *the piecewise polynomial sieve (see Section A.1.1) with degrees l_i in the ith coordinate, based on a partition with rectangles of sidelengthes 2^{-m_i}, or*
- *the wavelet sieve (see Section A.1.2), where the wavelets have l_i vanishing moments in the ith coordinate.*

Then, there is a constant $c = c_{\boldsymbol{s},d} > 0$ such that

$$\inf_{\alpha \in A_m} \|\alpha - \alpha_0\|_2 \leq c |\alpha_0|_{B^{\boldsymbol{s}}_{2,\infty}} \sum_{i=1}^{d+1} 2^{-s_i m_i}.$$

A.2 Some tools from the theory of counting processes and stochastic calculus

Let P_{α_0} be the joint law of $\{(X, N(t), Y(t)) : t \in [0,1]\}$ when (2) holds (the intensity is α_0). We want to explain why the log-likelihood ratio $\ell(\alpha, \alpha_0) := \log(dP_\alpha / dP_{\alpha_0})$ writes, when both α and α_0 are assumed to be positive on $[0,1]^{d+1}$:

$$\ell(\alpha, \alpha_0) = \int_0^1 \log\Big(\frac{\alpha(t,X)}{\alpha_0(t,X)}\Big) dN(t) - \int_0^1 (\alpha(t,X) - \alpha_0(t,X)) Y(t) dt. \tag{53}$$

This will entail that the log-likelihood ratio $\ell_n(\alpha, \alpha_0) := \log(dP^n_\alpha / dP^n_{\alpha_0})$ of the independent sample (3) satisfies

$$\ell_n(\alpha, \alpha_0) = \sum_{i=1}^n \Big(\int_0^1 \log\Big(\frac{\alpha(t,X_i)}{\alpha_0(t,X_i)}\Big) dN^i(t) - \int_0^1 (\alpha(t,X_i) - \alpha_0(t,X_i)) Y^i(t) dt \Big). \tag{54}$$

Equation (54) is useful in several parts of the paper (dimension reduction and lower bounds).

First, we recall Jacod's formula (see Andersen et al. (1993)) for the likelihood of a counting process. It writes, for the likelihood of N:

$$\prod_{t\in[0,1]}\Big\{(\alpha_0(t,X)Y(t))^{\Delta N(t)}(1-\alpha_0(t,X)Y(t))^{1-\Delta N(t)}\Big\}dt,$$

where $\Delta N(t) = N(t)-N(t_-)$ and where $\prod$ is the product-integral, see Andersen et al. (1993) for a definition. But N has a finite number of jumps on $[0,1]$ and $\Delta N(t) \in \{0,1\}$ for any $t \in [0,1]$, thus $1-\Delta N(t) = 1$ for any $t \in [0,1]$ excepted a finite number of times. Consequently the likelihood of N reduces to

$$\prod_{t\in[0,1]} (\alpha_0(t,X)Y(t))^{\Delta N(t)} \exp\Big(-\int_0^1 \alpha_0(t,X)Y(t)dt\Big)$$

where the first product is actually finite, and where we used the fact that $\prod_{t\in[0,1]}(1-f(t)) = \exp(-\int_0^1 f(t)dt)$ for a continuous function f on $[0,1]$. Thus, the likelihood ratio $L(\alpha,\alpha_0) = dP_\alpha/dP_{\alpha_0}$ writes

$$L(\alpha,\alpha_0) = \prod_{t\in[0,1]} \Big(\frac{\alpha(t,X)}{\alpha_0(t,X)}\Big)^{\Delta N(t)} \exp\Big(-\int_0^1 (\alpha(t,X)-\alpha_0(t,X)Y(t))dt\Big),$$

which entails (53) since $\sum_{t\in[0,1]} f(t)\Delta N^i(t) = \int_0^1 f(t)dN(t)$. Equation (54) is a consequence of (53), together with the fact since $N^1,\dots,N^n$ are independent, they cannot jump at the same time, so that $\sum_{i=1}^n \Delta N^i(t) \in \{0,1\}$ a.s.

References

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

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

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *ANNALS OF STATISTICS*, **37** 1591. URL doi:10.1214/08-AOS623.

Baraud, Y. and Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields*, **143** 239–284.

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113** 301–413.

Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probab. Theory Related Fields*, **135** 311–334.

Bass, R. F. (1985). Law of the iterated logarithm for set-indexed partial sum processes with finite variance. *Z. Wahrsch. Verw. Gebiete*, **70** 591–608.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Tech. rep., University of California, Berkeley.

Birge, L. (2007). Model selection for poisson processes. *IMS LECTURE NOTES MONOGRAPH SERIES*, **55** 32. URL doi:10.1214/074921707000000265.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, **4** 329–375.

Bitouzé, D., Laurent, B. and Massart, P. (1999). A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. H. Poincaré Probab. Statist.*, **35** 735–763.

Bousquet, O. (2002). *Concentration inequalities and empirical process theory applied to the analysis of learning algorithms*. Ph.D. thesis, Ecole Polytechnique.

Brunel, E. and Comte, F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā*, **67** 441–475.

Brunel, E., Comte, F. and Lacour (2007). Adaptive estimation of the conditional density in presence of censoring. *Sankhya*, **69** 734–763.

Catoni, O. (2001). *Statistical Learning Theory and Stochastic Optimization*. Ecole d'été de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics, Springer, N.Y.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press, Cambridge.

Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelets transforms. *Appl. Comput. Harmon. Anal.*, **1** 54–81.

Comte, F., Gaïffas, S. and Guilloux, A. (2008). Adaptive estimation of the conditional intensity of marker-dependent counting processes. Available at http://arxiv.org/abs/0810.4263.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, **34** 187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, **39** 1–49 (electronic).

Dabrowska, D. M. (1987). Nonparametric regression with censored survival time data. *Scand. J. Statist.*, **14** 181–197.

Dalalyan, A. S., Juditsky, A. and Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. *J. Mach. Learn. Res.*, **9** 1648–1678.

Dalalyan, A. S. and Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *COLT*. 97–111.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, **41** 909–996.

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.*, **86** 213–226.

Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference*, **136** 730–769.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.*, **6** 899–929 (1979).

Gaiffas, S. and Lecue, G. (2007). Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, **1** 538. URL doi:10.1214/07-EJS077.

Geenens, G. and Delecroix, M. (2005). A survey about single-index models theory. URL http://www.stat.ucl.ac.be/ISpub/dp/2005/dp0508.pdf.

Grégoire, G. (1993). Least squares cross-validation for counting process intensities. *Scand. J. Statist.*, **20** 343–360.

Heuchenne, C. and Van Keilegom, I. (2007). Location estimation in nonparametric regression with censored data. *J. Multivariate Anal.*, **98** 1558–1582.

Hochmuth, R. (2002). n-term approximation in anisotropic function spaces. *Math. Nachr.*, **244** 131–149.

Horowitz, J. L. (1998). *Semiparametric methods in econometrics*, vol. 131 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, **29** 595–623.

Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.*, **27** 1536–1563.

Jacobsen, M. (1982). *Statistical analysis of counting processes*, vol. 12 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Jacod, J. and Shiryaev, A. N. (1987). *Limit theorems for stochastic processes*, vol. 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

Juditsky, A. B., Rigollet, P. and Tsybakov, A. B. (2006). Learning by mirror averaging. To appear in the *Ann. Statist.*. Available at http://www.imstat.org/aos/future_papers.html.

Lacour, C. (2007). Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, **43** 571–597.

Lecué, G. (2007). *??* Ph.D. thesis, Université Pierre et Marie Curie – Paris 6.

Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1998). The importance of convexity in learning with squared loss. *IEEE Trans. Inform. Theory*, **44** 1974–1980.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, **52** 3396–3410.

Li, G. and Doss, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.*, **23** 787–823.

Linton, O. B., Nielsen, J. P. and van de Geer, S. (2003). Estimating multiplicative and additive hazard functions by kernel methods. *Ann. Statist.*, **31** 464–492. Dedicated to the memory of Herbert E. Robbins.

Liptser, R. S. and Shiryayev, A. N. (1989). *Theory of martingales*, vol. 49 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht. Translated from the Russian by K. Dzjaparidze [Kacha Dzhaparidze].

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.*, **27** 1808–1829.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, **28** 863–884.

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

McKeague, I. W. and Utikal, K. J. (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.*, **18** 1172–1187.

Nemirovski, A. (2000). *Topics in Non-Parametric Statistics*. Ecole d'été de probabilités de Saint-Flour XXVIII - 1998. Lecture Notes in Mathematics, no. 1738, Springer, New York.

Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.*, **15** 897–919.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11** 453–466.

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields*, **126** 103–153.

Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, **12** 633–661.

Rio, E. (2001). Inégalités de concentration pour les processus empiriques de classes de parties. *Probab. Theory Related Fields*, **119** 163–175.

Stute, W. (1986). Conditional empirical processes. *Ann. Statist.*, **14** 638–647.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, **23** 461–471.

Talagrand, M. (2005). *The generic chaining*. Springer Monographs in Mathematics, Springer-Verlag, Berlin. Upper and lower bounds of stochastic processes.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statist. in Med.*, **16** 385–395.

Triebel, H. (2006). *Theory of function spaces. III*, vol. 100 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel.

Tsybakov, A. B. (2003). Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence*, **2777** 303–313. Springer, Heidelberg.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32** 135–166.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, **21** 14–44.

van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, **23** 1779–1801.

van de Geer, S. (2007). Oracle inequalities and regularization. In *Lectures on empirical processes*. EMS Ser. Lect. Math., Eur. Math. Soc., Zürich, 191–252.

van de Geer, S. A. (2000). *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics, Springer-Verlag, New York. With applications to statistics.

Vapnik, V. N. (2000). *The nature of statistical learning theory*. 2nd ed. Statistics for Engineering and Information Science, Springer-Verlag, New York.

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models 1162–1184.

Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.*, **28** 75–87.