Cox Process Functional Learning

Gérard Biau
Université Pierre et Marie Curie[1] *& Ecole Normale Supérieure*[2]*, France*
gerard.biau@upmc.fr

Benoît Cadre
IRMAR, ENS Rennes, CNRS, UEB, France[3]
benoit.cadre@ens-rennes.fr

Quentin Paris
ENSAE-ParisTech, France
quentin.paris@ensae.fr

Abstract

This article addresses the problem of functional supervised classification of Cox process trajectories, whose random intensity is driven by some exogenous random covariable. The classification task is achieved through a regularized convex empirical risk minimization procedure, and a nonasymptotic oracle inequality is derived. We show that the algorithm provides a Bayes-risk consistent classifier. Furthermore, it is proved that the classifier converges at a rate which adapts to the unknown regularity of the intensity process. Our results are obtained by taking advantage of martingale and stochastic calculus arguments, which are natural in this context and fully exploit the functional nature of the problem.

Index Terms — Functional data analysis, Cox process, supervised classification, oracle inequality, consistency, regularization, stochastic calculus.

2010 Mathematics Subject Classification: 62G05, 62G20.

[1]Research partially supported by the French National Research Agency (grant ANR-09-BLAN-0051-02 "CLARA") and by the Institut universitaire de France.

[2]Research carried out within the INRIA project "CLASSIC" hosted by Ecole Normale Supérieure and CNRS.

[3]Research sponsored by the French National Research Agency (grant ANR-09-BLAN-0051-02 "CLARA").

1 Introduction

1.1 Functional classification and Cox processes

In supervised classification one considers a random pair (X, Y), where X takes values in some space $\mathcal{X}$ and Y takes only finitely values, say -1 or 1 to simplify. Given a learning sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of i.i.d. copies of (X, Y) observed in the past, the aim is to predict the value of Y associated with a new value of X. In medicine, for example, we specifically want to evaluate patients according to their disease risk, and the typical questions for classification are: "Is this person affected?", "Will this patient respond to the treatment?", or "Will this patient have serious side effects from using the drug?"—in all these cases, a yes/no or $-1/1$ decision has to be made.

The classification task is generally achieved by designing a decision rule (also called classifier) $g_n : \mathcal{X} \to \{-1, 1\}$, which represents our guess on the label Y of X (the subscript n in g_n means that the classifier measurably depends upon the sample). As the pair (X, Y) is random, an error occurs whenever $g_n(X)$ differs from Y, and the probability of error of the rule g_n is

$$L(g_n) = \mathbb{P}\left(g_n(X) \neq Y | \mathcal{D}_n\right).$$

The Bayes rule $g^\star$, defined by

$$g^\star(x) = \begin{cases} 1 \text{ if } \mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = -1|X = x) \\ 0 \text{ otherwise,} \end{cases}$$

has the smallest probability of error, in the sense that $L(g^\star) \leq L(g)$ for any classifier g (see, e.g., Devroye et al., 1996).

In the classical statistical setting, each observation X_i is a collection of numerical measurements represented by a d-dimensional vector. However, in an increasing number of application domains, input data are in the form of random functions rather than standard vectors, thereby turning the classification task into a functional data analysis problem. Here, the vocable "random functions" means that the variables X_i's take values in a space $\mathcal{X}$ of functions rather than $\mathbb{R}^d$, equipped with an appropriate topology. Thus, in this context, the challenge is to design classification rules which exploit the functional nature of the X_i's, and this calls for new methodological concepts. Accordingly, the last few years have witnessed important developments in both the theory and practice of functional data analysis, and numerous procedures have been adapted to handle functional inputs. The books by

Ramsay and Silverman (2002, 2005) and Ferraty (2011) provide a presentation of the area, and the survey of Baíllo et al. (2011) offers some essential references for functional supervised classification.

Curiously, despite a huge research activity in the field, few attempts have been made to connect the area of functional data analysis with the theory of stochastic processes, which also deals with the analysis of time-dependent quantities (interesting ideas towards this direction are included in Illian et al., 2006; Baíllo et al., 2011; Cadre, 2013; Shuang et al., 2013). As advocated in the present paper, stochastic calculus theory can be used efficiently to analyse Cox models and may serve as a starting point for more exchanges between the two fields.

To motivate the use of Cox models for classification, consider for instance a sample of AIDS patients observed until time T. Assume that, for each of them, we know the dates of visits to the hospital, a bunch of personal data (such as gender, distance from home to hospital, etc.), as well as the diagnostic (-1=aggravation, 1=remission, for example). Based on this learning sample, a classification strategy aims at predicting the ± 1 evolution diagnostic of a new patient. In this time-dependent setting, $\mathcal{X}$ is the set of counting paths on $[0, T]$ (that is, right-continuous and piecewise constant paths on $[0, T]$ starting at 0, and with jump size 1), and a relevant model for X is a mixture of two Cox processes (or doubly stochastic Poisson processes) with (random) intensities $\lambda_+ = (\lambda_{+,t})_{t \in [0,T]}$ and $\lambda_- = (\lambda_{-,t})_{t \in [0,T]}$. In other words, conditionally on $Y = 1$ (resp., $Y = -1$), the law of X given λ_+ (resp., λ_-) is the law of a Poisson process with intensity λ_+ (resp., λ_-). (For more information on Cox processes, we refer the reader to the original paper by Cox, 1955; see also the book by Bening and Korolev, 2002, for an overview of the application areas of these processes.) Compared to a Poisson process, the benefit of the random intensity lies in the fact that the statistician can take into account the auxiliary information carried by the personal data of the patients.

As we shall see, because of a martingale property of Cox processes, stochastic calculus proves to be a natural and efficient tool to investigate this classification problem. It is stressed that the originality of our work is that it takes advantage of the theory of stochastic processes to handle a functional data analysis problem—in that sense, it differs from other studies devoted to nonparametric estimation of Cox process intensity (see for instance Hansen et al., 2013, and the references therein).

1.2 Classification strategy

In the sequel, $T > 0$ is fixed and $\mathcal{X}$ stands for the set of counting paths on $[0, T]$. We consider a prototype random triplet (X, Z, Y), where Y is a binary label taking the values ± 1 with respective positive probabilities p_+ and p_- ($p_+ + p_- = 1$). In this model, $Z = (Z_t)_{t\in[0,T]}$ plays the role of a d-dimensional random covariable (process), whereas $X = (X_t)_{t\in[0,T]}$ is a mixture of two Cox processes, both being adapted with respect to the same filtration. More specifically, it is assumed that Z is independent of Y and that, conditionally on $Y = 1$ (resp., $Y = -1$), X is a Cox process with intensity $(\lambda_+(t, Z_t))_{t\in[0,T]}$ (resp., $(\lambda_-(t, Z_t))_{t\in[0,T]}$).

It will be assumed that the observation of the trajectories of X is stopped after its u-th jump, where u is some known, prespecified, positive integer. Thus, formally, we are to replace X and Z by X^τ and Z^τ, where $\tau = \inf\{t \in [0, T] : X_t = u\}$ (stopping time), $X_t^\tau = X_{t\wedge\tau}$ and $Z_t^\tau = Z_{t\wedge\tau}$. (Notation $t_1 \wedge t_2$ means the minimum of t_1 and t_2 and, by convention, $\inf \emptyset = 0$.) Stopping the observation of X after its u-th jump is essentially a technical requirement, with no practical incidence insofar u may be chosen arbitrarily large. However, it should be stressed that with this assumption, X^τ is, with probability one, nicely bounded from above by u. Additionally, to keep things simple, we suppose that each Z_t takes its values in $[0, 1]^d$ and we let $\mathcal{Z}$ be state space of Z.

Our objective is to learn the relation between (X^τ, Z^τ) and Y within the framework of supervised classification. Given a training dataset of n i.i.d. observation/label pairs $\mathcal{D}_n = \{(X_1^{\tau_1}, Z_1^{\tau_1}, Y_1), \ldots, (X_n^{\tau_n}, Z_n^{\tau_n}, Y_n)\}$ (with evident notation for τ_i's), distributed as (and independent of) the prototype triplet (X^τ, Z^τ, Y), the problem is to design a decision rule $g_n : \mathcal{X} \times \mathcal{Z} \to \{-1, 1\}$, based on $\mathcal{D}_n$, whose role is to assign a label to each possible new instance of the observation (X^τ, Z^τ). The classification strategy that we propose is based on empirical convex risk minimization. It is described in the next subsection.

In order to describe our classification procedure, some more notation is required. The performance of a classifier $g_n : \mathcal{X} \times \mathcal{Z} \to \{-1, 1\}$ is measured by the probability of error

$$L(g_n) = \mathbb{P}\left(g_n(X^\tau, Z^\tau) \neq Y \,|\, \mathcal{D}_n\right),$$

and the minimal possible probability of error is the Bayes risk, denoted by

$$L^\star = \inf_g L(g) = \mathbb{E} \min\left[\eta(X^\tau, Z^\tau), 1 - \eta(X^\tau, Z^\tau)\right].$$

In the identity above, the infimum is taken over all measurable classifiers $g : \mathcal{X} \times \mathcal{Z} \to \{-1, 1\}$, and $\eta(X^\tau, Z^\tau) = \mathbb{P}(Y = 1 \,|\, X^\tau, Z^\tau)$ denotes the posterior probability function. The infimum is achieved by the Bayes classifier

$$g^\star(X^\tau, Z^\tau) = \mathrm{sign}(2\eta(X^\tau, Z^\tau) - 1),$$

where $\mathrm{sign}(t) = 1$ for $t > 0$ and -1 otherwise. Our first result (Theorem 2.1) shows that

$$\eta(X^\tau, Z^\tau) = \frac{p_+}{p_- e^{-\xi} + p_+},$$

where ξ is the random variable defined by

$$\xi = \int_0^{T\wedge\tau} (\lambda_- - \lambda_+)(s, Z_s)\mathrm{d}s + \int_0^{T\wedge\tau} \ln \frac{\lambda_+}{\lambda_-}(s, Z_s)\mathrm{d}X_s.$$

An important consequence is that the Bayes rule associated with our decision problem takes the simple form

$$g^\star(X^\tau, Z^\tau) = \mathrm{sign}\left(\xi - \ln \frac{p_-}{p_+}\right).$$

Next, let $(\varphi_j)_{j\geq 1}$ be a countable dictionary of measurable functions defined on $[0, T] \times [0, 1]^d$. Assuming that both $\lambda_- - \lambda_+$ and $\ln \frac{\lambda_+}{\lambda_-}$ belong to the span of the dictionary, we see that

$$\xi = \sum_{j\geq 1} \left[a_j^\star \int_0^{T\wedge\tau} \varphi_j(s, Z_s)\mathrm{d}s + b_j^\star \int_0^{T\wedge\tau} \varphi_j(s, Z_s)\mathrm{d}X_s \right],$$

where $(a_j^\star)_{j\geq 1}$ and $(b_j^\star)_{j\geq 1}$ are two sequences of unknown real coefficients. Thus, for each positive integer B, it is quite natural to introduce the class $\mathcal{F}_B$ of real-valued functions $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, defined by

$$\mathcal{F}_B = \left\{ f = \sum_{j=1}^{B} [a_j \Phi_j + b_j \Psi_j] + c : \max\left(\sum_{j=1}^{B} |a_j|, \sum_{j=1}^{B} |b_j|, |c| \right) \leq B \right\}, \tag{1.1}$$

where

$$\Phi_j(x, z) = \int_0^{T\wedge\tau(x)} \varphi_j(s, z_s)\mathrm{d}s, \qquad \Psi_j(x, z) = \int_0^{T\wedge\tau(x)} \varphi_j(s, z_s)\mathrm{d}x_s,$$

and, by definition, $\tau(x) = \inf\{t \in [0, T] : x_t = u\}$ for $x \in \mathcal{X}$.

Each $f \in \mathcal{F}_B$ defines a classifier g_f by $g_f = \mathrm{sign}(f)$. To simplify notation, we write $L(f) = L(g_f) = \mathbb{P}(g_f(X^\tau, Z^\tau) \neq Y)$, and note that

$$\mathbb{E}\mathbf{1}_{[-Yf(X^\tau, Z^\tau)>0]} \leq L(f) \leq \mathbb{E}\mathbf{1}_{[-Yf(X^\tau, Z^\tau)\geq 0]}.$$

Therefore, the minimization of the probability of error $L(f)$ over $f \in \mathcal{F}_B$ is approximately equivalent to the minimization of the expected 0-1 loss $\mathbf{1}_{[.\geq 0]}$ of $-Yf(X^\tau, Z^\tau)$. The parameter B may be regarded as an $\mathbb{L}^1$-type smoothing parameter. Large values of B improve the approximation properties of the class $\mathcal{F}_B$ at the price of making the estimation problem more difficult. Now, given the sample $\mathcal{D}_n$, it is reasonable to consider an estimation procedure based on minimizing the sample mean

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[-Y_i f(X_i^{\tau_i}, Z_i^{\tau_i}) \geq 0]},$$

of the 0-1 loss.

It is now well established, however, that such a procedure is computationally intractable as soon as the class $\mathcal{F}_B$ is nontrivial, since the 0-1 loss function $\mathbf{1}_{[.\geq 0]}$ is nonconvex. A genuine attempt to circumvent this difficulty is to base the minimization procedure on a convex surrogate ϕ of the loss $\mathbf{1}_{[.\geq 0]}$. Such convexity-based methods, inspired by the pioneering works on boosting (Freund, 1995; Schapire, 1990; Freund and Schapire, 1997), have now largely displaced earlier nonconvex approaches in the machine learning literature (see, e.g., Blanchard et al., 2003; Lugosi and Vayatis, 2004; Zhang, 2004; Bartlett et al., 2006, and the references therein).

It turns out that in our Cox process context, the choice of the logit surrogate loss $\phi(t) = \ln_2(1+e^t)$ is the most natural one. This will be clarified in Section 2 by connecting the empirical risk minimization procedure and the maximum likelihood principle. Thus, with this choice, the corresponding risk functional and empirical risk functional are defined by

$$A(f) = \mathbb{E}\phi\left(-Yf(X^\tau, Z^\tau)\right) \quad \text{and} \quad A_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(-Y_i f(X_i^{\tau_i}, Z_i^{\tau_i})\right).$$

Given a nondecreasing sequence $(B_k)_{k\geq 1}$ of integer-valued smoothing parameters, the primal estimates we consider take the form

$$\hat{f}_k \in \arg\min_{f \in \mathcal{F}_{B_k}} A_n(f).$$

Remark 1.1 *Note that the minimum may not be achieved in $\mathcal{F}_{B_k}$. However, to simplify the arguments, we implicitly assume that the minimum indeed exists. All proofs may be adjusted, in a straightforward way, to handle approximate minimizers of the empirical risk functional.*

Remark 1.2 *The minimization of the functional A_n over the class $\mathcal{F}_B$ is indeed a convex problem in the a_j's, b_j's and c introduced in (1.1), which makes our method computationally tractable. An alternative approach is to consider functional classes based on the development of the intensity functions λ_- and λ_+ instead of $\lambda_- - \lambda_+$ and $\ln \frac{\lambda_+}{\lambda_-}$. However, such a procedure induces a non convex optimization problem.*

Starting from the collection $(\hat{f}_k)_{k\geq 1}$, the final estimate uses a value of k chosen empirically, by minimizing a penalized version of the empirical risk $A_n(\hat{f}_k)$. To achieve this goal, consider a penalty (or regularization) function pen : $\mathbb{N}^\star \to \mathbb{R}_+$ to be specified later on. Then the resulting penalized estimate $\hat{f}_n = \hat{f}_{\hat{k}}$ has

$$\hat{k} \in \arg\min_{k\geq 1} \left[A_n(\hat{f}_k) + \text{pen}(k) \right].$$

The role of the penalty is to compensate for overfitting and helps finding an adequate value of k. For larger values of k, the class $\mathcal{F}_{B_k}$ is larger, and therefore pen(k) should be larger as well.

By a careful choice of the regularization term, specified in Theorem 2.2, one may find a close-to-optimal balance between estimation and approximation errors and investigate the probability of error $L(\hat{f}_n)$ of the classifier $g_{\hat{f}_n}$ induced by the penalized estimate. Our conclusion asserts that $\hat{f}_n$ adapts nicely to the unknown smoothness of the problem, in the sense that with probability at least $1 - 1/n^2$,

$$L(\hat{f}_n) - L^\star = \mathrm{O}\left(\frac{\ln n}{n}\right)^{\frac{\beta}{2\beta+16}},$$

where β is some Sobolev-type regularity measure pertaining to λ_+ and λ_-. For the sake of clarity, proofs are postponed to Section 3. An appendix at the end of the paper recalls some important results by Blanchard et al. (2008) and Koltchinskii (2011) on model selection and suprema of Rademacher processes, together with more technical stochastic calculus material.

2 Results

As outlined in the introduction, our first result shows that the posterior probabilities $\mathbb{P}(Y = \pm 1 | X^\tau, Z^\tau)$ have a simple form. The crucial result that is needed here is Lemma A.1 which uses stochastic calculus arguments. For more clarity, this lemma has been postponed to the Appendix section. Recall that both p_+ and p_- are (strictly) positive and satisfy $p_+ + p_- = 1$.

Theorem 2.1 *Let ξ be the random variable defined by*

$$\xi = \int_0^{T\wedge\tau} (\lambda_- - \lambda_+)(s, Z_s)\mathrm{d}s + \int_0^{T\wedge\tau} \ln\frac{\lambda_+}{\lambda_-}(s, Z_s)\mathrm{d}X_s.$$

Then

$$\mathbb{P}(Y = 1|X^\tau, Z^\tau) = \frac{p_+}{p_- e^{-\xi} + p_+} \quad \textit{and} \quad \mathbb{P}(Y = -1|X^\tau, Z^\tau) = \frac{p_-}{p_+ e^{\xi} + p_-}.$$

This result, which is interesting by itself, sheds an interesting light on the Cox process classification problem. To see this, fix $Y_1 = y_1, \dots, Y_n = y_n$, and observe that the conditional likelihood of the model is

$$\begin{aligned}\mathcal{L}_n &= \prod_{i=1}^n \mathbb{P}(Y_i = y_i|X_i^{\tau_i}, Z_i^{\tau_i}) \\ &= \prod_{i=1}^n \left(\frac{p_+}{p_- e^{-y_i\xi_i} + p_+}\right)^{\mathbf{1}_{[y_i=1]}} \left(\frac{p_-}{p_+ e^{-y_i\xi_i} + p_-}\right)^{\mathbf{1}_{[y_i=-1]}},\end{aligned}$$

where of course

$$\xi_i = \int_0^{T\wedge\tau_i} (\lambda_- - \lambda_+)(s, Z_{i,s})\mathrm{d}s + \int_0^{T\wedge\tau_i} \ln\frac{\lambda_+}{\lambda_-}(s, Z_{i,s})\mathrm{d}X_{i,s}.$$

Therefore, the log-likelihood takes the form

$$\begin{aligned}\ln\mathcal{L}_n &= \sum_{i=1}^n \left[\ln\left(\frac{p_+}{p_- e^{-y_i\xi_i} + p_+}\right)\mathbf{1}_{[y_i=1]} + \ln\left(\frac{p_-}{p_+ e^{-y_i\xi_i} + p_-}\right)\mathbf{1}_{[y_i=-1]}\right] \\ &= -\sum_{i=1}^n \left[\ln\left(1 + \frac{p_-}{p_+}e^{-y_i\xi_i}\right)\mathbf{1}_{[y_i=1]} + \ln\left(1 + \frac{p_+}{p_-}e^{-y_i\xi_i}\right)\mathbf{1}_{[y_i=-1]}\right] \\ &= -\sum_{i=1}^n \ln\left(1 + \left(\frac{p_-}{p_+}\right)^{y_i} e^{-y_i\xi_i}\right) \\ &= -\sum_{i=1}^n \ln\left(1 + \exp\left[-y_i\left(\xi_i - \ln\frac{p_-}{p_+}\right)\right]\right).\end{aligned}$$

Thus, letting $\phi(t) = \ln_2(1 + e^t)$, we obtain

$$\ln\mathcal{L}_n = -\ln 2\sum_{i=1}^n \phi\left(-y_i\left(\xi_i - \ln\frac{p_-}{p_+}\right)\right). \tag{2.1}$$

Since the ξ_i's, p_+ and p_- are unknown, the natural idea, already alluded to in the introduction, is to expand $\lambda_- - \lambda_+$ and $\ln \frac{\lambda_+}{\lambda_-}$ on the dictionary $(\varphi_j)_{j \geq 1}$. To this end, we introduce the class $\mathcal{F}_B$ of real-valued functions

$$\mathcal{F}_B = \left\{ f = \sum_{j=1}^{B} [a_j \Phi_j + b_j \Psi_j] + c : \max \left(\sum_{j=1}^{B} |a_j|, \sum_{j=1}^{B} |b_j|, |c| \right) \leq B \right\},$$

where B is a positive integer,

$$\Phi_j(x, z) = \int_0^{T \wedge \tau(x)} \varphi_j(s, z_s) \mathrm{d}s, \quad \text{and} \quad \Psi_j(x, z) = \int_0^{T \wedge \tau(x)} \varphi_j(s, z_s) \mathrm{d}x_s.$$

For a nondecreasing sequence $(B_k)_{k \geq 1}$ of integer-valued smoothing parameters and for each $k \geq 1$, we finally select $\hat{f}_k \in \mathcal{F}_{B_k}$ for which the log-likelihood (2.1) is maximal. Clearly, such a maximization strategy is strictly equivalent to minimizing over $f \in \mathcal{F}_{B_k}$ the empirical risk

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} \phi\left(-Y_i f(X_i^{\tau_i}, Z_i^{\tau_i})\right).$$

This remark reveals the deep connection between our Cox process learning model and the maximum likelihood principle. In turn, it justifies the logit loss $\phi(t) = \ln_2(1 + e^t)$ as the natural surrogate candidate to the nonconvex 0-1 classification loss. (Note that the $\ln 2$ term is introduced for technical reasons only and plays no role in the analysis). Finally, we stress the fact that by convexity, this approach is computationally tractable.

As for now, denoting by $\|.\|_\infty$ the functional supremum norm, we assume that there exists a positive constant L such that, for each $j \geq 1$, $\|\varphi_j\|_\infty \leq L$. It immediately follows that for all integers $B \geq 1$, the class $\mathcal{F}_B$ is uniformly bounded by UB, where $U = 1 + (T + u)L$. We are now ready to state our main theorem, which offers a bound on the difference $A(\hat{f}_n) - A(f^\star)$.

Theorem 2.2 *Let $(B_k)_{k \geq 1}$ be a nondecreasing sequence of positive integers such that $\sum_{k \geq 1} B_k^{-\alpha} \leq 1$ for some $\alpha > 0$. For all $k \geq 1$, let*

$$R_k = A_k^2 B_k C_k + \frac{\sqrt{A_k}}{C_k},$$

where

$$A_k = UB_k \phi'(UB_k) \quad \text{and} \quad C_k = 2(\phi(UB_k) + 1 - \ln 2).$$

Then there exists a universal constant $C > 0$ *such that if the penalty* pen : $\mathbb{N}^\star \to \mathbb{R}_+$ *satisfies*

$$\mathrm{pen}(k) \geq C\left[R_k \frac{\ln n}{n} + \frac{C_k(\alpha \ln B_k + \delta + \ln 2)}{n}\right]$$

for some $\delta > 0$*, one has, with probability at least* $1 - e^{-\delta}$*,*

$$A(\hat{f}_n) - A(f^\star) \leq 2 \inf_{k \geq 1} \left\{ \inf_{f \in \mathcal{F}_{B_k}} (A(f) - A(f^\star)) + \mathrm{pen}(k) \right\}. \tag{2.2}$$

Some remarks are in order. At first, we note that Theorem 2.2 provides us with an oracle inequality which shows that, for each B_k, the penalized estimate does almost as well as the best possible classifier in the class $\mathcal{F}_{B_k}$, up to a term of the order $\ln n/n$. It is stressed that this remainder term tends to 0 at a much faster rate than the standard $(1/\sqrt{n})$-term suggested by a standard uniform convergence argument (see, e.g., Lugosi and Vayatis, 2004). This is a regularization effect which is due to the convex loss ϕ. In fact, proof of Theorem 2.2 relies on the powerful model selection machinery presented in Blanchard et al. (2008) coupled with modern empirical process theory arguments developed in Koltchinskii (2011). We also emphasize that a concrete but suboptimal value of the constant C may be deduced from the proof, but that no attempt has been made to optimize this constant. Next, observing that, for the logit loss,

$$\phi'(t) = \frac{1}{\ln 2(e^{-t}+1)},$$

we notice that a penalty behaving as B_k^4 is sufficient for the oracle inequality of Theorem 2.2 to hold. This corresponds to a regularization function proportional to the fourth power of the $\mathbb{L}^1$-norm of the collection of coefficients defining the base class functions. Such regularizations have been explored by a number of authors in recent years, specifically in the context of sparsity and variable selection (see, e.g., Tibshirani, 1996; Candès and Tao, 2005; Bunea et al., 2007; Bickel et al., 2009). With this respect, our approach is close to the view of Massart and Meynet (2011), who provide information about the Lasso as an $\mathbb{L}^1$-regularization procedure per se, together with sharp $\mathbb{L}^1$-oracle inequalities. Let us finally mention that the result of Theorem 2.2 can be generalized, with more technicalities, to other convex loss functions by following, for example, the arguments presented in Bartlett et al. (2006).

If we are able to control the approximation term $\inf_{f \in \mathcal{F}_{B_k}}(A(f) - A(f^\star))$ in inequality (2.2), then it is possible to give an explicit rate of convergence to 0 for the quantity $A(\hat{f}_n) - A(f^\star)$. This can be easily achieved by assuming, for example, that $(\varphi_j)_{j\geq 1}$ is an orthonormal basis and that both combinations $\lambda_- - \lambda_+$ and $\ln \frac{\lambda_+}{\lambda_-}$ enjoy some Sobolev-type regularity with respect to this basis. Also, the following additional assumption will be needed:

Assumption A. There exists a measure μ on $[0,1]^d$ and a constant $D > 0$ such that, for all $t \in [0, T]$, the distribution of Z_t has a density h_t with respect to μ which is uniformly bounded by D. In addition, λ_- and λ_+ are both $[\varepsilon, D]$-valued for some $\varepsilon > 0$.

Proposition 2.1 *Assume that Assumption* **A** *holds. Assume, in addition, that $(\varphi_j)_{j\geq 1}$ is an orthonormal basis of $\mathbb{L}^2(\mathrm{d}s \otimes \mu)$, where $\mathrm{d}s$ stands for the Lebesgue measure on $[0,T]$, and that both $\lambda_- - \lambda_+$ and $\ln \frac{\lambda_+}{\lambda_-}$ belong to the ellipsoïd*

$$\mathcal{W}(\beta, M) = \left\{ f = \sum_{j=1}^{\infty} a_j \varphi_j : \sum_{j=1}^{\infty} j^{2\beta} a_j^2 \leq M^2 \right\},$$

for some fixed $\beta \in \mathbb{N}^\star$ and $M > 0$. Then, letting

$$\lambda_- - \lambda_+ = \sum_{j=1}^{\infty} a_j^\star \varphi_j \quad \text{and} \quad \ln \frac{\lambda_+}{\lambda_-} = \sum_{j=1}^{\infty} b_j^\star \varphi_j,$$

we have, for all $B \geq \max(M^2, \ln \frac{p_+}{p_-})$,

$$\inf_{f \in \mathcal{F}_B} \left(A(f) - A(f^\star) \right) \leq \frac{2D\sqrt{T\mu\left([0,1]^d\right) M}\|a^\star\|_2}{B^{\beta/2}} + \frac{2D(1 + D\sqrt{T\mu([0,1]^d)})\sqrt{M\|b^\star\|_2}}{B^{\beta/2}},$$

where $\|a^\star\|_2^2 = \sum_{j=1}^{\infty} a_j^{\star 2}$ and $\|b^\star\|_2^2 = \sum_{j=1}^{\infty} b_j^{\star 2}$.

A careful inspection of Theorem 2.2 and Proposition 2.1 reveals that for the choice $B_k = \lceil (\pi k)^{2/\alpha}/6^{1/\alpha} \rceil$ and $\delta = 2\ln n$, there exists a universal constant $C > 0$ such that

$$A(\hat{f}_n) - A(f^\star) \leq C\mathbf{L} \left(\sqrt{\|a^\star\|_2} + \sqrt{\|b^\star\|_2} \right)^{\frac{8}{\beta+8}} \left(\frac{\ln n}{n} \right)^{\frac{\beta}{\beta+8}},$$

with probability at least $1 - 1/n^2$, where

$$\mathbf{L} = U^{\frac{3\beta}{\beta+8}} \left[2D\sqrt{M}(1 + D\sqrt{T\mu([0,1]^d)}) \right]^{\frac{8}{\beta+8}} \max\left(\left(\frac{\beta}{8}\right)^{\frac{8}{\beta+8}}, \left(\frac{8}{\beta}\right)^{\frac{\beta}{\beta+8}} \right).$$

Observe that, due to the specific form of the ellipsoïd $\mathcal{W}(\beta, M)$, the rate of convergence does not depend upon the dimension d.

Of course, our main concern is not the behavior of the expected risk $A(\hat{f}_n)$ but the probability of error $L(\hat{f}_n)$ of the corresponding classifier. Fortunately, the difference $L(\hat{f}_n) - L^\star$ may directly be related to $A(\hat{f}_n) - A(f^\star)$. Applying for example Lemma 2.1 in Zhang (2004), we conclude that with probability at least $1 - 1/n^2$,

$$L(\hat{f}_n) - L^\star \leq 2\sqrt{C\mathbf{L}} \left(\sqrt{\|a^\star\|_2} + \sqrt{\|b^\star\|_2}\right)^{\frac{4}{\beta+8}} \left(\frac{\ln n}{n}\right)^{\frac{\beta}{2\beta+16}}.$$

To understand the significance of this inequality, just recall that what we are after in this article is the supervised classification of (infinite-dimensional) stochastic processes. As enlightened in the proofs, this makes the analysis different from the standard context, where one seeks to learn finite-dimensional quantities. The bridge between the two worlds is crossed via stochastic calculus arguments. Lastly, it should be noted that the regularity parameter β is assumed to be unknown, so that our results are adaptive as well.

3 Proofs

Throughout this section, if P is a probability measure and f a function, the notation Pf stands for the integral of f with respect to P. By $\mathbb{L}^2(P)$ we mean the space of square integrable real functions with respect to P. Also, for a class $\mathcal{F}$ of functions in $\mathbb{L}^2(P)$ and $\varepsilon > 0$, we denote by $N(\varepsilon, \mathcal{F}, \mathbb{L}^2(P))$ the ε-covering number of $\mathcal{F}$ in $\mathbb{L}^2(P)$, i.e., the minimal number of metric balls of radius ε in $\mathbb{L}^2(P)$ that are needed to cover $\mathcal{F}$ (see, e.g., Definition 2.1.5 in van der Vaart and Wellner, 1996).

3.1 Proof of Theorem 2.1

For any stochastic processes M_1 and M_2, the notation $\mathbb{Q}_{M_2|M_1}$ and $\mathbb{Q}_{M_2}$ respectively mean the distribution under $\mathbb{Q}$ of M_2 given M_1, and the distribution under $\mathbb{Q}$ of M_2.

We start the proof by observing that

$$\mathbb{P}(Y = 1 \,|\, X^\tau = x, Z = z) = p_+ \frac{\mathrm{d}\mathbb{P}_{X^\tau, Z|Y=1}}{\mathrm{d}\mathbb{P}_{X^\tau, Z}}(x, z). \tag{3.1}$$

Thus, to prove the theorem, we need to evaluate the above Radon-Nikodym density. To this aim, we introduce the conditional probabilities $\mathbb{P}^\pm = \mathbb{P}(.|Y =$

± 1). For any path z of Z, the conditional distributions $\mathbb{P}^+_{X|Z=z}$ and $\mathbb{P}^-_{X|Z=z}$ are those of Poisson processes with intensity $\lambda_+(.,z)$ and $\lambda_-(.,z)$, respectively. Consequently, according to Lemma A.1, the stopped process X^τ satisfies

$$D_+(x,z)\,\mathbb{P}^+_{X^\tau|Z=z}(\mathrm{d}x) = D_-(x,z)\,\mathbb{P}^-_{X^\tau|Z=z}(\mathrm{d}x),$$

where

$$D_\pm(x,z) = \exp\left(-\int_0^{T\wedge\tau}(1-\lambda_\pm(s,z_s))\,\mathrm{d}s - \int_0^{T\wedge\tau}\ln\lambda_\pm(s,z_s)\mathrm{d}x_s\right).$$

Therefore,

$$D_+(x,z)\,\mathbb{P}^+_{X^\tau|Z=z}\otimes\mathbb{P}_Z(\mathrm{d}x,\mathrm{d}z) = D_-(x,z)\,\mathbb{P}^-_{X^\tau|Z=z}\otimes\mathbb{P}_Z(\mathrm{d}x,\mathrm{d}z).$$

But, by independence of Y and Z, one has $\mathbb{P}_Z = \mathbb{P}^+_Z = \mathbb{P}^-_Z$. Thus,

$$\mathbb{P}_{X^\tau,Z|Y=\pm1}(\mathrm{d}x,\mathrm{d}z) = \mathbb{P}^\pm_{X^\tau|Z=z}\otimes\mathbb{P}_Z(\mathrm{d}x,\mathrm{d}z),$$

whence

$$D_+(x,z)\mathbb{P}_{X^\tau,Z|Y=1}(\mathrm{d}x,\mathrm{d}z) = D_-(x,z)\mathbb{P}_{X^\tau,Z|Y=-1}(\mathrm{d}x,\mathrm{d}z).$$

On the other hand,

$$\mathbb{P}_{X^\tau,Z}(x,z) = p_+\mathbb{P}_{X^\tau,Z|Y=1}(x,z) + p_-\mathbb{P}_{X^\tau,Z|Y=-1}(x,z),$$

so that

$$\frac{\mathrm{d}\mathbb{P}_{X^\tau,Z|Y=1}}{\mathrm{d}\mathbb{P}_{X^\tau,Z}}(x,z) = \frac{1}{p_-\frac{D_+(x,z)}{D_-(x,z)}+p_+}.$$

Using identity (3.1), we obtain

$$\mathbb{P}(Y=1\,|\,X^\tau,Z) = \frac{p_+}{p_-e^{-\xi}+p_+},$$

where

$$\xi = \int_0^{T\wedge\tau}(\lambda_- - \lambda_+)(s,Z_s)\mathrm{d}s + \int_0^{T\wedge\tau}\ln\frac{\lambda_+}{\lambda_-}(s,Z_s)\mathrm{d}X_s.$$

Observing now that $\sigma(\tau)\subset\sigma(X^\tau_t, t\leq T)$ and

$$\xi = \int_0^{T\wedge\tau}(\lambda_- - \lambda_+)(s,Z^\tau_s)\mathrm{d}s + \int_0^{T\wedge\tau}\ln\frac{\lambda_+}{\lambda_-}(s,Z^\tau_s)\mathrm{d}X^\tau_s$$

give

$$\mathbb{P}(Y=1\,|\,X^\tau,Z^\tau) = \mathbb{P}(Y=1\,|\,X^\tau,Z).$$

This shows the desired result.□

3.2 Proof of Theorem 2.2

Theorem 2.2 is mainly a consequence of a general model selection result due to Blanchard et al. (2008), which is recalled in the Appendix for the sake of completeness (Theorem A.1). Throughout the proof, the letter C denotes a generic universal positive constant, whose value may change from line to line. We let $\ell(f)$ be a shorthand notation for the function

$$(x, z, y) \in \mathcal{X} \times \mathcal{Z} \times \{-1, 1\} \mapsto \phi(-yf(x, z)),$$

and let P be the distribution of the prototype triplet (X^τ, Z^τ, Y).

To frame our problem in the vocabulary of Theorem A.1, we consider the family of models $(\mathcal{F}_{B_k})_{k\geq 1}$ and start by verifying that assumptions (i) to (iv) are satisfied. If we define

$$\mathbf{d}^2(f, f') = P\left(\ell(f) - \ell(f')\right)^2,$$

then assumption (i) is immediately satisfied. A minor modification of the proof of Lemma 19 in Blanchard et al. (2003) reveals that, for all integers $B > 0$ and all $f \in \mathcal{F}_B$,

$$P\left(\ell(f) - \ell(f^\star)\right)^2 \leq (\phi(UB) + \phi(-UB) + 2 - 2\ln 2)\, P\left(\ell(f) - \ell(f^\star)\right).$$

This shows that assumption (ii) is satisfied with $C_k = 2(\phi(UB_k) + 1 - \ln 2)$. Moreover, it can be easily verified that assumption (iii) holds with $b_k = \phi(UB_k)$.

The rest of the proof is devoted to the verification of assumption (iv). To this aim, for all $B > 0$ and all $f_0 \in \mathcal{F}_B$, we need to bound the expression

$$F_B(r) = \mathbb{E}\sup\left\{|(P_n - P)\left(\ell(f) - \ell(f_0)\right)| : f \in \mathcal{F}_B, \mathbf{d}^2(f, f_0) \leq r\right\},$$

where

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\left(X_i^{\tau_i}, Z_i^{\tau_i}, Y_i\right)}$$

is the empirical distribution associated to the sample. Let

$$\mathcal{G}_{B,f_0} = \left\{\ell(f) - \ell(f_0) : f \in \mathcal{F}_B\right\}.$$

Then

$$F_B(r) = \mathbb{E}\sup\left\{|(P_n - P)g| : g \in \mathcal{G}_{B,f_0}, Pg^2 \leq r\right\}.$$

Using the symmetrization inequality presented in Theorem 2.1 of Koltchinskii (2011), it is easy to see that

$$F_B(r) \leq 2\,\mathbb{E}\sup\left\{\frac{1}{n}\sum_{i=1}^{n}\sigma_i g\left(X_i^{\tau_i}, Z_i^{\tau_i}, Y_i\right) : g \in \mathcal{G}_{B,f_0}, Pg^2 \leq r\right\}, \tag{3.2}$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables (that is, $\mathbb{P}(\sigma_i = \pm 1) = 1/2$), independent from the $(X_i^{\tau_i}, Z_i^{\tau_i}, Y_i)$'s. Now, since the functions in $\mathcal{F}_B$ take their values in $[-UB, UB]$, and since ϕ is Lipschitz on this interval with constant $\phi'(UB)$, we have, for all $f, f' \in \mathcal{F}_B$,

$$\sqrt{P_n\left(\ell(f) - \ell(f')\right)^2} \leq \phi'(UB)\sqrt{P_n\left(f - f'\right)^2}.$$

Consequently, for all $\varepsilon > 0$,

$$N\left(2\varepsilon UB\phi'(UB), \mathcal{G}_{B,f_0}, \mathbb{L}^2\left(P_n\right)\right) \leq N\left(2\varepsilon UB, \mathcal{F}_B, \mathbb{L}^2\left(P_n\right)\right).$$

Since $\mathcal{F}_B$ is included in a linear space of dimension at most $2B+1$, Lemma 2.6.15 in van der Vaart and Wellner (1996) indicates that it is a VC-subgraph class of VC-dimension at most $2B+3$. Observing that the function constantly equal to $2UB$ is a measurable envelope for $\mathcal{F}_B$, we conclude from Theorem 9.3 in Kosorok (2008) that, for all $\varepsilon > 0$,

$$N\left(2\varepsilon UB, \mathcal{F}_B, \mathbb{L}^2\left(P_n\right)\right) \leq C\left(2B+3\right)\left(4e\right)^{2B+3}\left(\frac{1}{\varepsilon}\right)^{4(B+1)}.$$

Therefore,

$$N\left(2\varepsilon UB\phi'(UB), \mathcal{G}_{B,f_0}, \mathbb{L}^2\left(P_n\right)\right) \leq C\left(2B+3\right)\left(4e\right)^{2B+3}\left(\frac{1}{\varepsilon}\right)^{4(B+1)}.$$

Now, notice that the constant function equal to $2UB\phi'(UB)$ is a measurable envelope for $\mathcal{G}_{B,f_0}$. Thus, applying Lemma A.2 yields

$$F_B(r) \leq \psi_B(r),$$

where ψ_B is defined for all $r > 0$ by

$$\psi_B(r) = \frac{C\sqrt{r}}{\sqrt{n}}\sqrt{B\ln\left(\frac{A'_B}{\sqrt{r}}\right)} \vee \frac{CBA_B}{n}\ln\left(\frac{A'_B}{\sqrt{r}}\right) \vee \frac{CA_B}{n}\sqrt{B\ln\left(\frac{A'_B}{\sqrt{r}}\right)},$$

with $A_B = UB\phi'(UB)$ and $A'_B = A_B((2B+3)(4e)^{2B+3})^{1/4(B+1)}$. (Notation $t_1 \vee t_2$ means the maximum of t_1 and t_2.)

Attention shows that ψ_B is a sub-root function and assumption (iv) is therefore satisfied. It is routine to verify that the solution $r_k^\star$ of $\psi_{B_k}(r) = r/C_k$ satisfies, for all $k \geq 1$ and all $n \geq 1$,

$$r_k^\star \leq C\left(A_{B_k}^2 B_k C_k^2 + \sqrt{A'_{B_k}}\right)\frac{\ln n}{n}.$$

Furthermore, observing that the function $B \mapsto ((2B+3)(4e)^{2B+3})^{1/4(B+1)}$ is bounded from above, we obtain

$$r_k^\star \leq C\left(A_{B_k}^2 B_k C_k^2 + \sqrt{A_{B_k}}\right)\frac{\ln n}{n}.$$

Hence, taking $x_k = \alpha \ln \lambda_k$ and $K = 11/5$ in Theorem A.1, and letting

$$R_k = A_{B_k}^2 B_k C_k + \frac{\sqrt{A_{B_k}}}{C_k},$$

we conclude that there exists a universal constant $C > 0$ such that, if the penalty $\mathrm{pen} : \mathbb{N}^\star \to \mathbb{R}_+$ satisfies

$$\mathrm{pen}(k) \geq C\left\{R_k\frac{\ln n}{n} + \frac{C_k\left(\alpha \ln B_k + \delta + \ln 2\right)}{n}\right\}$$

for some $\delta > 0$, then, with probability at least $1 - e^{-\delta}$,

$$A(\hat{f}_n) - A(f^\star) \leq 2\inf_{k\geq 1}\left\{\inf_{f\in\mathcal{F}_{B_k}}\left(A(f) - A(f^\star)\right) + \mathrm{pen}(k)\right\}.$$

This completes the proof. □

3.3 Proof of Proposition 2.1

Proof of Proposition 2.1 relies on the following intermediary lemma, which is proved in the next subsection.

Lemma 3.1 *Assume that Assumption* **A** *holds. Then, for all positive integers $B \geq 1$,*

$$\begin{aligned}\inf_{f\in\mathcal{F}_B}\left(A(f) - A(f^\star)\right) \leq 2D\sqrt{T\mu([0,1]^d)}\min\left\|\sum_{j=1}^B \alpha_j\varphi_j - (\lambda_- - \lambda_+)\right\| \\ + 2D(1 + D\sqrt{T\mu([0,1]^d)})\min\left\|\sum_{j=1}^B \alpha_j\varphi_j - \ln\frac{\lambda_+}{\lambda_-}\right\| \\ + 2\min_{|x|\leq B}\left|x - \ln\frac{p_+}{p_-}\right|,\end{aligned}$$

where the first two minima are taken over all $\alpha = (\alpha_1, \dots, \alpha_B) \in \mathbb{R}^B$ with $\sum_{j=1}^{B} |\alpha_j| \leq B$ and where we have denoted by $\|.\|$ the $\mathbb{L}^2(\mathrm{d}s \otimes \mu)$-norm.

Proof of Proposition 2.1 – For ease of notation, we will denote by $\|.\|$ the $\mathbb{L}^2(\mathrm{d}s \otimes \mu)$-norm throughout the proof. For all $B \geq 1$,

$$\min\left\{\left\|\sum_{j=1}^{B} \alpha_j \varphi_j - (\lambda_- - \lambda_+)\right\| : \sum_{j=1}^{B} |\alpha_j| \leq B\right\} \\ \leq \min\left\{\left\|\sum_{j=1}^{B} \alpha_j \varphi_j - (\lambda_- - \lambda_+)\right\| : \sum_{j=1}^{B} \alpha_j^2 \leq B\right\}. \tag{3.3}$$

Since $\lambda_- - \lambda_+ \in \mathcal{W}(\beta, M)$ and $B \geq M^2$, we have

$$\sum_{j=1}^{B} a_j^{\star 2} \leq \sum_{j=1}^{\infty} j^{2\beta} a_j^{\star 2} \leq M^2 \leq B. \tag{3.4}$$

Thus, combining (3.3) and (3.4) yields, for $B \geq M^2$,

$$\min\left\{\left\|\sum_{j=1}^{B} \alpha_j \varphi_j - (\lambda_- - \lambda_+)\right\| : \sum_{j=1}^{B} |\alpha_j| \leq B\right\} \leq \left\|\sum_{j=1}^{B} a_j^{\star} \varphi_j - (\lambda_- - \lambda_+)\right\| \\ = \left\|\sum_{j=B+1}^{\infty} a_j^{\star} \varphi_j\right\|. \tag{3.5}$$

It follows from the properties of an orthonormal basis and the definition of $\mathcal{W}(\beta, M)$ that

$$\begin{aligned} \left\|\sum_{j=B+1}^{\infty} a_j^{\star} \varphi_j\right\|^2 &= \sum_{j=B+1}^{\infty} a_j^{\star 2} \\ &\leq \sqrt{\sum_{j=B+1}^{\infty} j^{2\beta} a_j^{\star 2}} \sqrt{\sum_{j=B+1}^{\infty} \frac{a_j^{\star 2}}{j^{2\beta}}} \\ &\leq M \sqrt{\sum_{j=B+1}^{\infty} \frac{a_j^{\star 2}}{j^{2\beta}}} \\ &\leq \frac{M \|a^{\star}\|_2}{B^{\beta}}. \end{aligned} \tag{3.6}$$

Inequalities (3.5) and (3.6) show that, for all $B \geq M^2$,

$$\min\left\{\left\|\sum_{j=1}^{B}\alpha_j\varphi_j - (\lambda_- - \lambda_+)\right\| : \sum_{j=1}^{B}|\alpha_j| \leq B\right\} \leq \sqrt{\frac{M\|a^\star\|_2}{B^\beta}}.$$

Similarly, it may be proved that, for all $B \geq M^2$,

$$\min\left\{\left\|\sum_{j=1}^{B}\alpha_j\varphi_j - \ln\frac{\lambda_+}{\lambda_-}\right\| : \sum_{j=1}^{B}|\alpha_j| \leq B\right\} \leq \sqrt{\frac{M\|b^\star\|_2}{B^\beta}}.$$

Applying Lemma 3.1 we conclude that, whenever $B \geq \max(M^2, \ln\frac{p_+}{p_-})$,

$$\begin{aligned}\inf_{f\in\mathcal{F}_B}(A(f) - A(f^\star)) \leq& \frac{2D\sqrt{T\mu\left([0,1]^d\right)M\|a^\star\|_2}}{B^{\beta/2}}\\ &+\frac{2D(1+D\sqrt{T\mu([0,1]^d)})\sqrt{M\|b^\star\|_2}}{B^{\beta/2}},\end{aligned}$$

which ends the proof. □

3.4 Proof of Lemma 3.1

We start with a technical lemma.

Lemma 3.2 *Let $\phi(t) = \ln_2(1+e^t)$ be the logit loss. Then*

$$\arg\min_f \mathbb{E}\phi\left(-Yf(X^\tau, Z^\tau)\,|\,X^\tau, Z^\tau\right) = \xi - \ln\frac{p_-}{p_+},$$

where the minimum is taken over all measurable functions $f : \mathcal{X}\times\mathcal{Z}\to\mathbb{R}$.

Proof – According to the results of Section 2.2 in Bartlett et al. (2006), one has

$$\arg\min_f \mathbb{E}\phi\left(-Yf(X^\tau, Z^\tau)\,|\,X^\tau, Z^\tau\right) = \alpha^\star\left(\eta(X^\tau, Z^\tau)\right),$$

where, for all $0 \leq \eta \leq 1$,

$$\alpha^\star(\eta) = \arg\min_{\alpha\in\mathbb{R}}\left(\eta\phi(-\alpha) + (1-\eta)\phi(\alpha)\right).$$

With our choice for ϕ, it is straightforward to check that, for all $0 \leq \eta < 1$,

$$\alpha^\star(\eta) = \ln\left(\frac{\eta}{1-\eta}\right).$$

Since, by assumption, $p_- > 0$, we have

$$\eta(X^\tau, Z^\tau) = \frac{p_+}{p_- e^{-\xi} + p_+} < 1.$$

Thus

$$\alpha^\star\left(\eta(X^\tau, Z^\tau)\right) = \xi - \ln \frac{p_-}{p_+},$$

which is the desired result. $\square$

Proof of Lemma 3.1 – Let $B > 0$ be fixed. Let $a_1, \ldots, a_B$ and $b_1, \ldots, b_B$ be real numbers such that

$$\left\| \sum_{j=1}^{B} a_j \varphi_j - (\lambda_- - \lambda_+) \right\|_{\mathbb{L}^2(\mathrm{d}s \otimes \mu)} = \min \left\| \sum_{j=1}^{B} \alpha_j \varphi_j - (\lambda_- - \lambda_+) \right\|_{\mathbb{L}^2(\mathrm{d}s \otimes \mu)}$$

and

$$\left\| \sum_{j=1}^{B} b_j \varphi_j - \ln \frac{\lambda_+}{\lambda_-} \right\|_{\mathbb{L}^2(\mathrm{d}s \otimes \mu)} = \min \left\| \sum_{j=1}^{B} \alpha_j \varphi_j - \ln \frac{\lambda_+}{\lambda_-} \right\|_{\mathbb{L}^2(\mathrm{d}s \otimes \mu)},$$

where, in each case, the minimum is taken over all $\alpha = (\alpha_1, \ldots, \alpha_B) \in \mathbb{R}^B$ with $\sum_{j=1}^{B} |\alpha_j| \leq B$. Let also $c \in \mathbb{R}$ be such that

$$\left| c - \ln \frac{p_+}{p_-} \right| = \min_{|x| \leq B} \left| x - \ln \frac{p_+}{p_-} \right|.$$

Introduce f_B, the function in $\mathcal{F}_B$ defined by

$$\begin{aligned} f_B &= \sum_{j=1}^{B} \left[a_j \Phi_j + b_j \Psi_j \right] + c \\ &= \int_0^{T \wedge \tau} \sum_{j=1}^{B} a_j \varphi_j(s, Z_s) \mathrm{d}s + \int_0^{T \wedge \tau} \sum_{j=1}^{B} b_j \varphi_j(s, Z_s) \mathrm{d}X_s + c. \end{aligned}$$

Clearly,

$$\inf_{f \in \mathcal{F}_B} \left(A(f) - A(f^\star) \right) \leq A(f_B) - A(f^\star). \tag{3.7}$$

Since ϕ is Lipschitz with constant $\phi'(UB) = (\ln 2(1 + e^{-UB}))^{-1} \leq 2$ on the interval $[-UB, UB]$, we have

$$|A(f_B) - A(f^\star)| \leq 2\mathbb{E} \left| f_B(X^\tau, Z^\tau) - f^\star(X^\tau, Z^\tau) \right|. \tag{3.8}$$

But, by Lemma 3.2,

$$f^\star(X^\tau, Z^\tau) = \int_0^{T\wedge\tau} (\lambda_- - \lambda_+)(s, Z_s)\mathrm{d}s + \int_0^{T\wedge\tau} \ln\frac{\lambda_+}{\lambda_-}(s, Z_s)\mathrm{d}X_s + \ln\frac{p_+}{p_-}.$$

Thus, letting,

$$\vartheta_1 = \sum_{j=1}^{B} a_j\varphi_j - (\lambda_- - \lambda_+) \quad \text{and} \quad \vartheta_2 = \sum_{j=1}^{B} b_j\varphi_j - \ln\frac{\lambda_+}{\lambda_-},$$

it follows

$$\begin{aligned}\mathbb{E}\left|f_B(X^\tau, Z^\tau) - f^\star(X^\tau, Z^\tau)\right| &\leq \mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_1(s, Z_s)\mathrm{d}s\right| \\ &\quad + \mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}X_s\right| \\ &\quad + \left|c - \ln\frac{p_+}{p_-}\right|. \end{aligned} \tag{3.9}$$

Using Assumption **A** and Cauchy-Schwarz's Inequality, we obtain

$$\begin{aligned}\mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_1(s, Z_s)\mathrm{d}s\right| &\leq \int_0^T \int_{[0,1]^d} \left|\vartheta_1(s,z)\right| \mathbb{P}_{Z_s}(\mathrm{d}z)\mathrm{d}s \\ &= \int_0^T \int_{[0,1]^d} \left|\vartheta_1(s,z)\right| h_s(z)\mu(\mathrm{d}z)\mathrm{d}s \\ &\leq D\|\vartheta_1\|_{\mathbb{L}^1(\mathrm{d}s\otimes\mu)} \\ &\leq D\sqrt{T\mu([0,1]^d)}\|\vartheta_1\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}. \end{aligned} \tag{3.10}$$

With a slight abuse of notation, set $\lambda_Y = \lambda_\pm$, depending on whether $Y = \pm 1$, and

$$\Lambda_{Y,Z}(t) = \int_0^t \lambda_Y(s, Z_s)\mathrm{d}s, \quad t \in [0, T].$$

With this notation,

$$\begin{aligned}\mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}X_s\right| &\leq \mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}\left(X_s - \Lambda_{Y,Z}(s)\right)\right| \\ &\quad + \mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}\Lambda_{Y,Z}(s)\right| \\ &= \mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}\left(X_s - \Lambda_{Y,Z}(s)\right)\right| \\ &\quad + \mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\lambda_Y(s, Z_s)\mathrm{d}s\right|. \end{aligned} \tag{3.11}$$

Therefore, applying Assumption **A** and Cauchy-Schwarz's Inequality,

$$\mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}X_s\right| \leq \mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}\left(X_s - \Lambda_{Y,Z}(s)\right)\right| + D^2\sqrt{T\mu([0,1]^d)}\|\vartheta_2\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}. \tag{3.12}$$

Since $X - \Lambda_{Y,Z}$ is a martingale conditionally to Y and Z, the Ito isometry (see Theorem I.4.40 in Jacod and Shiryaev, 2003) yields

$$\mathbb{E}\left[\left(\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}\left(X_s - \Lambda_{Y,Z}(s)\right)\right)^2 \middle| Y, Z\right] = \mathbb{E}\left[\int_0^{T\wedge\tau} \vartheta_2^2(s, Z_s)\mathrm{d}\langle X - \Lambda_{Y,Z}\rangle_s \middle| Y, Z\right], \tag{3.13}$$

where $\langle M\rangle$ stands for the predictable compensator of the martingale M. Observing that, conditionally on Y, Z, X is a Poisson process with compensator $\Lambda_{Y,Z}$, we deduce that $\langle X - \Lambda_{Y,Z}\rangle = \langle X\rangle = \Lambda_{Y,Z}$. As a result,

$$\mathbb{E}\left[\int_0^{T\wedge\tau} \vartheta_2^2(s, Z_s)\mathrm{d}\langle X - \Lambda_{Y,Z}\rangle_s \middle| Y, Z\right] = \mathbb{E}\left[\int_0^{T\wedge\tau} \vartheta_2^2(s, Z_s)\lambda_Y(s, Z_s)\mathrm{d}s \middle| Y, Z\right]. \tag{3.14}$$

Hence,

$$\mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}\left(X_s - \Lambda_{Y,Z}(s)\right)\right| \leq D\|\vartheta_2\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}. \tag{3.15}$$

Combining (3.12) and (3.15) we deduce that

$$\mathbb{E}\left|\int_0^{T\wedge\tau} \vartheta_2(s, Z_s)\mathrm{d}X_s\right| \leq D(1 + D\sqrt{T\mu([0,1]^d)})\|\vartheta_2\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}. \tag{3.16}$$

Putting together identities (3.7)-(3.10) and (3.16) yields

$$\begin{aligned}\inf_{f\in\mathcal{F}_B} (A(f) - A(f^\star)) &\leq 2D\sqrt{T\mu([0,1]^d)}\|\vartheta_1\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)} + 2D(1 + D\sqrt{T\mu([0,1]^d)})\|\vartheta_2\|_{\mathbb{L}^2(\mathrm{d}s\otimes\mu)}\\ &\quad + 2\min_{|x|\leq B}\left|x - \ln\frac{p_+}{p_-}\right|,\end{aligned}$$

which concludes the proof by definition of ϑ_1 and ϑ_2. □

A Appendix

A.1 A general theorem for model selection

The objective of this section is to recall a general model selection result due to Blanchard et al. (2008).

Let $\mathcal{X}$ be a measurable space and let $\ell : \mathbb{R} \times \{-1, 1\} \to \mathbb{R}$ be a loss function. Given a function $g : \mathcal{X} \to \mathbb{R}$, we let $\ell(g)$ be a shorthand notation for the function $(x, y) \in \mathbb{R} \times \{-1, 1\} \mapsto \ell(g(x), y)$. Let P be a probability distribution on $\mathcal{X} \times \{-1, 1\}$ and let $\mathfrak{G}$ be a set of extended-real valued functions on $\mathcal{X}$ such that, for all $g \in \mathfrak{G}$, $\ell(g) \in \mathbb{L}^2(P)$. The target function $g^\star$ is defined as

$$g^\star \in \arg\min_{g \in \mathfrak{G}} P\ell(g).$$

Let $(\mathcal{G}_k)_{k \geq 1}$ be a countable family of models such that, for all $k \geq 1$, $\mathcal{G}_k \subset \mathfrak{G}$. For each $k \geq 1$, we define the empirical risk minimizer $\hat{g}_k$ as

$$\hat{g}_k \in \arg\min_{g \in \mathcal{G}_k} P_n \ell(g).$$

If pen denotes a real-valued function on $\mathbb{N}^\star$, we let the penalized empirical risk minimizer $\hat{g}$ be defined by $\hat{g}_{\hat{k}}$, where

$$\hat{k} \in \arg\min_{k \geq 1} \left[P_n \ell(\hat{g}_k) + \mathrm{pen}(k) \right].$$

Recall that a function $\mathbf{d} : \mathfrak{G} \times \mathfrak{G} \to \mathbb{R}_+$ is a pseudo-distance if (i) $\mathbf{d}(g, g) = 0$, (ii) $\mathbf{d}(g, g') = \mathbf{d}(g', g)$, and (iii) $\mathbf{d}(g, g') \leq \mathbf{d}(g, g'') + \mathbf{d}(g'', g')$ for all g, g', g'' in $\mathfrak{G}$. Also, a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be a sub-root function if (i) it is nondecreasing and (ii) the function $r \in \mathbb{R}_+ \mapsto \psi(r)/\sqrt{r}$ is nonincreasing.

Theorem A.1 (Blanchard et al., 2008) *Assume that there exist a pseudo-distance $\mathbf{d}$ on $\mathfrak{G}$, a sequence of sub-root functions $(\psi_k)_{k \geq 1}$, and two nondecreasing sequences $(b_k)_{k \geq 1}$ and $(C_k)_{k \geq 1}$ of real numbers such that*

(i) $\forall g, g' \in \mathfrak{G} : P(\ell(g) - \ell(g'))^2 \leq \mathbf{d}^2(g, g')$;

(ii) $\forall k \geq 1, \forall g \in \mathcal{G}_k : \mathbf{d}^2(g, g^\star) \leq C_k P(\ell(g) - \ell(g^\star))$;

(iii) $\forall k \geq 1, \forall g \in \mathcal{G}_k, \forall (x, y) \in \mathcal{X} \times \{-1, 1\} : |\ell(g(x), y)| \leq b_k$;

and, if $r_k^\star$ denotes the solution of $\psi_k(r) = r/C_k$, (*iv*) $\forall k \geq 1$, $\forall g_0 \in \mathcal{G}_k$, $\forall r \geq r_k^\star$:

$$\mathbb{E} \sup \left\{ \left| (P_n - P) \left(\ell(g) - \ell(g_0) \right) \right| : g \in \mathcal{G}_k, \mathbf{d}^2(g, g_0) \leq r \right\} \leq \psi_k(r).$$

Let $(x_k)_{k \geq 1}$ be a nonincreasing sequence such that $\sum_{k \geq 1} e^{-x_k} \leq 1$. Let $\delta > 0$ and $K > 1$ be two fixed real numbers. If $\mathrm{pen}(k)$ *denotes a penalty term satisfying*

$$\forall k \geq 1, \quad \mathrm{pen}(k) \geq 250K \frac{r_k^\star}{C_k} + \frac{(65KC_k + 56b_k)(x_k + \delta + \ln 2)}{3n},$$

then, with probability at least $1 - e^{-\delta}$, one has

$$P\left(\ell(\hat{g}) - \ell(g^\star)\right) \leq \frac{K + \frac{1}{5}}{K - 1} \inf_{k \geq 1} \left\{ \inf_{g \in \mathcal{G}_k} P\left(\ell(g) - \ell(g^\star)\right) + 2\mathrm{pen}(k) \right\}.$$

A.2 Expected supremum of Rademacher processes

Let S be a measurable space and let P be a probability measure on S. Let $\mathcal{G}$ be a class of functions $g : S \to \mathbb{R}$. The Rademacher process $(R_n(g))_{g \in \mathcal{G}}$ associated with P and indexed by $\mathcal{G}$ is defined by

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(Z_i),$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables, and $Z_1, \ldots, Z_n$ is a sequence of i.i.d. random variables, with distribution P and independent of the σ_i's.

We recall in this subsection a bound for the supremum of the Rademacher process defined by

$$\|R_n\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} |R_n(g)|,$$

which follows from the results of Giné and Koltchinskii (2006). Let G be a measurable envelope for $\mathcal{G}$, i.e., a measurable function $G : S \to \mathbb{R}_+$ such that, for all $x \in S$,

$$\sup_{g \in \mathcal{G}} |g(x)| \leq G(x).$$

Define $\|G\| = \sqrt{PG^2}$ and $\|G\|_n = \sqrt{P_n G^2}$, where $P_n = n^{-1} \sum_{i=1}^{n} \delta_{Z_i}$ stands for the empirical measure associated to $Z_1, \ldots, Z_n$. Finally, let $\sigma^2 > 0$ be a real number satisfying

$$\sup_{g \in \mathcal{G}} Pg^2 \leq \sigma^2 \leq \|G\|^2.$$

Theorem A.2 (Giné and Koltchinskii, 2006) *Assume that the functions in $\mathcal{G}$ are uniformly bounded by a constant $U > 0$. Assume, in addition, that there exist two constants C and $V > 0$ such that, for all $n \geq 1$ and all $0 < \epsilon \leq 2$,*

$$N\left(\varepsilon \|G\|_n, \mathcal{G}, \mathbb{L}^2\left(P_n\right)\right) \leq \left(\frac{C}{\varepsilon}\right)^V .$$

Then, for all $n \geq 1$,

$$\mathbb{E}\|R_n\|_{\mathcal{G}} \leq \frac{c\sigma}{\sqrt{n}}\sqrt{V \ln\left(\frac{c'\|G\|}{\sigma}\right)} \vee \frac{8c^2UV}{n} \ln\left(\frac{c'\|G\|}{\sigma}\right) \vee \frac{cU}{9n}\sqrt{V \ln\left(\frac{c'\|G\|}{\sigma}\right)},$$

where $c = 432$ and $c' = 2e \vee C$.

A.3 Some stochastic calculus results

Up to the stopped part, the following result is a classical consequence of the Girsanov theorem. We give its proof for convenience of the reader.

Lemma A.1 *Let μ (resp., ν) be the distribution of a Poisson process on $[0, T]$ with intensity $\lambda : [0, T] \to \mathbb{R}_+^\star$ (resp., with intensity 1) stopped after its u-th jump. Then, μ and ν are equivalent. Moreover,*

$$\nu(\mathrm{d}x) = \exp\left(-\int_0^{T\wedge\tau(x)} (1 - \lambda(s))\,\mathrm{d}s - \int_0^{T\wedge\tau(x)} \ln \lambda(s)\mathrm{d}x_s\right) \mu(\mathrm{d}x),$$

where, for all $x \in \mathcal{X}$, $\tau(x) = \inf\{t \in [0, T] : x_t = u\}$.

Proof. Consider the canonical Poisson process $N = (N_t)_{t\in[0,T]}$ with intensity λ on the filtered space $(\mathcal{X}, (\mathcal{A}_t)_{t\in[0,T]}, \mathbb{P})$, where $\mathcal{A}_t = \sigma(N_s : s \in [0, t])$, and let, for all $t \in [0, T]$,

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s \quad \text{and} \quad h(t) = \frac{1}{\lambda(t)} - 1.$$

Recall that the process $M = (M_t)_{t\in[0,T]}$ defined by $M_t = N_t - \Lambda(t)$ is a martingale. The Doléans-Dade exponential $\mathcal{E} = (\mathcal{E}_t)_{t\in[0,T]}$ of the martingale $h.M$ (see, e.g., Theorem I.4.61 in Jacod and Shiryaev, 2003) is defined for all $t \in [0, T]$ by

$$\begin{aligned}
\mathcal{E}_t &= e^{h.M_t} \prod_{s\leq t}(1 + \Delta h.M_s)e^{-\Delta h.M_s} \\
&= \exp\left(-\int_0^t h(s)\lambda(s)\mathrm{d}s + \int_0^t \ln\left(1 + h(s)\right)\mathrm{d}N_s\right) \\
&= \exp\left(-\int_0^t (1 - \lambda(s))\,\mathrm{d}s - \int_0^t \ln \lambda(s)\mathrm{d}N_s\right), \qquad \text{(A.1)}
\end{aligned}$$

where $\Delta h.M_s = h.M_s - h.M_{s^-} = h.N_s - h.N_{s^-}$. Equivalently, $\mathcal{E}$ is the solution to the stochastic equation

$$\mathcal{E} = 1 + \mathcal{E}^-.(h.M) = 1 + (\mathcal{E}^- h).M,$$

where $\mathcal{E}^-$ stands for the process defined by $\mathcal{E}_t^- = \mathcal{E}_{t^-}$. In particular, $\mathcal{E}$ is a martingale. Observe also, since N is a counting process, that the quadratic covariation between M and $\mathcal{E}$ is

$$[M, \mathcal{E}] = (\mathcal{E}^- h).[N, N] = (\mathcal{E}^- h).N.$$

Consequently,

$$[M, \mathcal{E}] - (\mathcal{E}^- h).\Lambda = (\mathcal{E}^- h).M$$

is a martingale. Since $(\mathcal{E}^- h).\Lambda$ is a continuous and adapted process, it is a predictable process and the predictable compensator of $[M, \mathcal{E}]$ takes the form $\langle M, \mathcal{E}\rangle = (\mathcal{E}^- h).\Lambda$. Now let $\mathbb{Q}$ be the measure defined by

$$\mathrm{d}\mathbb{Q} = \mathcal{E}_T \mathrm{d}\mathbb{P}.$$

Since the process $\mathcal{E}$ is a martingale, $\mathbb{Q}$ is a probability and in addition, for all $t \in [0, T]$,

$$\mathrm{d}\mathbb{Q}_t = \mathcal{E}_t \mathrm{d}\mathbb{P}_t, \tag{A.2}$$

where $\mathbb{Q}_t$ and $\mathbb{P}_t$ are the respective restrictions of $\mathbb{Q}$ and $\mathbb{P}$ to $\mathcal{A}_t$. It follows, by the Girsanov theorem (see, e.g., Theorem III.3.11 in Jacod and Shiryaev, 2003), that the stochastic process $M - (\mathcal{E}^-)^{-1}.\langle M, \mathcal{E}\rangle$ is a $\mathbb{Q}$-martingale. But, for all $t \in [0, T]$,

$$\begin{aligned} M_t - (\mathcal{E}^-)^{-1}.\langle M, \mathcal{E}\rangle_t &= N_t - \Lambda(t) - h.\Lambda(t) \\ &= N_t - (1 + h).\Lambda(t) \\ &= N_t - t. \end{aligned}$$

Therefore, the counting process N is such that $(N_t - t)_{t \in [0,T]}$ is a $\mathbb{Q}$-martingale. In consequence, by the Watanabe theorem (see, e.g., Theorem IV.4.5 in Jacod and Shiryaev, 2003), this implies that the distribution of N under $\mathbb{Q}$ is that of a Poisson process with unit intensity. So, $\nu = \mathbb{Q}_{T\wedge\tau}$, where $\mathbb{Q}_{T\wedge\tau}$ is the restriction of $\mathbb{Q}$ to the stopped σ-field $\mathcal{A}_{T\wedge\tau}$. Moreover, by Theorem III.3.4 in Jacod and Shiryaev (2003) and identity (A.2), we have

$$\mathrm{d}\mathbb{Q}_{T\wedge\tau} = \mathcal{E}_{T\wedge\tau} \mathrm{d}\mathbb{P}_{T\wedge\tau},$$

where the definition of $\mathbb{P}_{T\wedge\tau}$ is clear. Since $\mu = \mathbb{P}_{T\wedge\tau}$, the result is a consequence of identity (A.1). □

References

A. Baíllo, J. Cuestas-Alberto, and A. Cuevas. Supervised classification for a family of Gaussian functional models. *Scandinavian Journal of Statistics*, 38:480–498, 2011.

A. Baíllo, A. Cuevas, and R. Fraiman. Classification methods for functional data. In F. Ferraty and Y. Romain, editors, *The Oxford Handbook of Functional Data Analysis*, pages 259–297, Oxford, 2011. Oxford University Press.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

V.E. Bening and V.Y. Korolev. *Generalized Poisson Models and their Applications in Insurance and Finance*. VSP, Utrecht, 2002.

P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting methods. *Journal of Machine Learning Research*, 4: 861–894, 2003.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36:489–531, 2008.

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

B. Cadre. Supervised classification of diffusion paths. *Mathematical Methods of Statistics*, 3:213–225, 2013.

E.J. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2005.

D.R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B*, 17:129–164, 1955.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

F. Ferraty. *Proceedings IWFOS'2011: Recent Advances on Functional Data Analysis and Related Fields (Ed.)*. Physica-Verlag, Heidelberg, 2011.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34: 1143–1216, 2006.

N.R. Hansen, P. Raynaud-Bourret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli, in press*, 2013.

J. Illian, E. Benson, J. Crawford, and H. Staines. Principal component analysis for spatial point processes – assessing the appropriateness of the approach in an ecological context. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, and D. Stoyan, editors, *Case Studies in Spatial Point Process Modeling*, pages 135–150, New York, 2006. Springer.

J. Jacod and A.N. Shiryaev. *Limits Theorems for Stochastic Processes. Second Edition*. Springer, New York, 2003.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, Berlin, 2011.

M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32:30–55, 2004.

P. Massart and C. Meynet. An ℓ_1-oracle inequality for the Lasso. *Electronic Journal of Statistics*, 5:669–687, 2011.

J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York, 2002.

J.O. Ramsay and B.W. Silverman. *Functional Data Analysis. Second Edition*. Springer, New York, 2005.

R.E. Schapire. The strength of weak learnability. *Machine Learning*, 5: 197–227, 1990.

W. Shuang, H.-G. Müller, and Z. Zhang. Functional data analysis for point processes with rare events. *Statistica Sinica*, 23:1–23, 2013.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.