CONSISTENT PROCEDURES FOR MULTICLASS CLASSIFICATION OF DISCRETE DIFFUSION PATHS

Christophe Denis, Charlotte Dion, Miguel Martinez

CONSISTENT PROCEDURES FOR MULTICLASS CLASSIFICATION OF DISCRETE DIFFUSION PATHS

Abstract. The recent advent of modern technology has generated a large number of datasets which can be frequently modeled as functional data. This paper focuses on the problem of multiclass classification for stochastic diffusion paths. In this context we establish a closed formula for the optimal Bayes rule. We provide new statistical procedures which are built either on the *plug-in* principle or on the empirical risk minimization principle. We show the consistency of these procedures under mild conditions. We apply our methodologies to the parametric case and illustrate their accuracy with a simulation study through examples.

Keywords : Multiclass classification, diffusion paths, plug-in estimators, drift estimation, empirical risk minimization.

AMS Subject Classification: 62M05, 62H30, 62F12

Christophe Denis[(1)], Charlotte Dion[(2)], Miguel Martinez[(1)]
[(1)] LAMA, Université Paris Est Marne-la-Vallée
[(2)] LPSM, Sorbonne Université

1. Introduction

In the multiclass classification framework, it is assumed that we have at our disposal a learning sample of observations that consists of N independent realizations of (X, Y) with the feature $X \in \mathcal{X}$ and the label $Y \in \{1, \dots, K\}$ constructed on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For a new observation X the goal is to predict the associated unobserved label Y. This is done through a *classifier* $g : \mathcal{X} \to \{1, \dots, K\}$. The misclassification risk of g is $\mathbb{P}(g(X) \neq Y)$. The accuracy of the classifier is then evaluated by comparison with the Bayes classifier g^*. For $x \in \mathcal{X}$, $g^*(x)$ is defined as the maximizer over $\{1, \dots, K\}$ of the conditional probabilities $\mathbb{P}(Y = k|X = x)$. Moreover, the Bayes classifier minimizes the misclassification error over the set of all classifiers (see *e.g.* Devroye *et al.*, 1996; Vapnik, 1998). Therefore, the performance of an arbitrary classifier g is measured by considering the *excess risk* $\mathbb{P}(g(X) \neq Y) - \mathbb{P}(g^*(X) \neq Y)$. In statistical learning, the joint distribution of (X, Y) is unknown. Consequently, based on the learning sample, the objective is to build an empirical classifier $\widehat{g}$ such that the expectation of its excess risk tends to zero as N tends to infinity (*consistency*).

Within this context, the present work focuses on the case where the feature $X = (X_t)_{t\in[0,T]}$ is a diffusion process solution of some stochastic differential equation (s.d.e.) with an unknown drift function depending on the label Y. This kind of functional random data is widely used to model the behavior of an agent that produces real valued stochastic data features along time. Such type of random data are used in many domains such as medical sciences (see *e.g.* Donnet & Samson, 2013), physics (see *e.g.* Parisi & Sourlas, 1992), financial mathematics (see *e.g.* El Karoui *et al.*, 1997).

We propose statistical classification strategies based on the learning sample and relying on the diffusion model assumption. Naturally, our classification procedures involve drift coefficient estimators. One specificity of the paper is that diffusions are sampled at high frequency (time step Δ) over a fixed time interval $[0, T]$.

1.1. Motivation and state of the art.

The classification problem for diffusion sample paths may be regarded as a particular case of functional data analysis problems. Many methods have been developed to solve such problems in general (see *e.g.* Ramsay & Silverman, 2007; Wang *et al.*, 2015). Among all these methodologies, we may mention k-nearest neighbors in Hilbert spaces (see Biau *et al.*, 2005, 2010) that could be applied to our classification problem. There is also some recent developments for related problems such as functional random forests (Gregorutti *et al.*, 2015), functional principal component

analysis, kernel estimators, just to mention a few of them. Recent works on depth classification for functional data (López-Pintado & Romo, 2006; Cuevas *et al.*, 2007; Lange *et al.*, 2014; Kuelbs & Zinn, 2016) propose various elegant computational solutions. These methods have the strong robustness of not specifying any model on the data, which makes them very interesting for practitioners. However, the counterpart is that the convergence may be difficult to obtain.

Let us now talk about specific methods for our classification problem of diffusion sample paths. Not too far from our problem, we mention Baíllo *et al.* (2011) that studies supervised classification for a family of Gaussian processes and Delattre *et al.* (2015b) that uses mixed stochastic differential equations in order to solve a data clustering issue. Closer to our problem Denis (2014) investigates multiclass classification for Cox-Ingersoll-Ross processes.

To the best of our knowledge, the main theoretical contribution for the classification problem of general diffusion sample paths discriminated by the drift function is Cadre (2013). The obtained results focus on binary classification for continuous observations and rely on the empirical risk minimization strategy. The author provides a consistent empirical classification rule. However, the resulting procedure cannot be implemented in practical situations.

Note that, by itself, the interesting question of providing an estimation of the drift coefficient from the observation of a single trajectory has been thoroughly studied. In the frequentist framework, one can cite for example Bibby & Sørensen (1995) and Kessler *et al.* (1999) for martingale estimation functions, Gobet *et al.* (2004) in the low frequency context. In the case of continuous ergodic diffusions, the LAN properties are treated in Yoshida (1992); Gobet (2002). Nonparametric estimators are proposed for example in Hoffmann (1999); Comte *et al.* (2007).

This work takes place in the high-frequency observation framework. The sampling interval Δ between successive observations is assumed to tend to zero as the number of observations n tends to infinity. The length of the observation time interval $T = n\Delta$ is supposed to be fixed and we do not assume any stationary property for the underlying process. In this context it is well-known that a consistent estimation of the drift from a single trajectory is impossible. However, in our framework we take advantage of the repeated observations of the learning sample to derive consistent estimators of the drift.

1.2. **Main contribution.** We provide a closed formula for the optimal Bayes classifier which yields an explicit representation for the excess risk of a general classifier. Thus, the relation between the conditional probabilities $\mathbb{P}(Y = k|X)$ and the vector b of unknown drift functions is fully explicit. Our strategy relies on the plug-in principle (see *e.g.* Audibert & Tsybakov, 2007). Based on an estimator $\widehat{b}$ of b, we consider an estimator of the conditional probabilities. Then, for each estimator $\widehat{b}$, we consider the empirical classifier $\widehat{g} := g_{\widehat{b}}$ defined as the maximizer of the estimated conditional probabilities. The major part of the paper is then devoted to show that plug-in classification procedures derived from drift coefficient estimators are indeed consistent. In particular we first exhibit a sufficient condition on the estimator $\widehat{b}$ which ensures the consistency of the resulting procedure. Secondly, we construct an estimator based on the minimization of the empirical risk over the learning sample. We show the consistency of this new procedure. Under mild assumptions, we show that the rate of convergence is comparable to the one obtained in Cadre (2013) but in the multiclass context with discrete observations.

A substantial part of the paper is devoted to the study of the parametric case. We study minimum contrast estimators of the parameters that rule the drift and show their consistency and asymptotic normality. The resulting plug-in classification procedure is then shown to be consistent. Furthermore, we propose to use a convex version of the empirical risk minimizer which involves convex surrogates of the misclassification risks (see Zhang, 2004; Bartlett *et al.*, 2006). We present here two new easily implementable classifiers and prove their consistency.

In comparison to Cadre (2013), the present work brings three main extensions. The first one is the generalization of the binary missclassification problem for diffusion paths to the corresponding multiclass classification problem. The second one is the discrete setting of our framework. Closer

to reality, we assume that the data collected are recorded at discrete times. This introduces an additional error term due to the time step and we give the order of this additional error in the rates of convergence. Thirdly, in the parametric setting, we exhibit procedures that are easily implementable. We present convincing numerical results on some classical examples.

1.3. **Plan of the Paper.** In Section 2, we start with the presentation of our general framework and settle the model. Section 3 is devoted to the construction of discrete observations classifiers. We introduce the two classes of classification procedures studied in the sequel, namely the ones that are based on a consistent estimation of the drift and those that rely on the minimization of the empirical risk. General abstract convergence results are then derived for these procedures. In Section 4 the drift of the underlying diffusion is taken out of a regular parametric family. First, we construct a consistent and asymptotic normal minimum contrast estimator. Secondly, we propose a first procedure referred as *the constrained method* and a second one based on a *one versus all* strategy. Finally, we investigate the performances of our predictors in a simulation study presented in Section 5. The results obtained in the paper together with opened issues and future perspectives are discussed in Section 6, whereas the proofs are relegated to Section 7.

2. General framework

2.1. **Model and assumptions.** Let $T > 0$ be a fixed time horizon. Let (X, Y) the generic data-structure taking its values in $\mathcal{X}_T \times \mathcal{Y}$, with $\mathcal{X}_T := (C([0,T]), \mathcal{C})$ the set of real valued continuous functions with its corresponding σ-algebra endowed by the uniform topology, $\mathcal{Y} = \{1, \dots, K\}$, with $K \geq 2$. Assume we are given a vector of K unknown Borel real functions $b^* = (b_1^*, \dots, b_K^*) \in \mathcal{B}$ (with $\mathcal{B}$ a set of functions), a known Borel real function σ and a starting point $x_0 \in \mathbb{R}$. The process $X = (X_t)_{t\in[0,T]}$ is assumed to come from the following diffusion model

$$\begin{cases} X_0 &= x_0 \\ dX_t &= b_Y^*(X_t)dt + \sigma(X_t)dW_t, \end{cases} \tag{2.1}$$

where $(W_t)_{t\geq 0}$ denotes a standard Brownian motion on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and such that the label Y is independent of $(W_t)_{t\geq 0}$ with known distribution under $\mathbb{P}$ given by $(p_i)_{i\in\mathcal{Y}}$. In the sequel, $(\mathcal{F}_t^X)_{t\geq 0} := \{\sigma(X_s : \ s \leq t) \ ; \ t \geq 0\}$ denotes the natural filtration of the process X. Note that the process X is then a mixture of Brownian motions. We make the following assumptions.

Assumption 2.1 (Ellipticity and regularity)**.** *There exist strictly positive constants σ_0, σ_1 such that*

$$0 < \sigma_0 \leq \sigma(x) \leq \sigma_1, \ \ \forall x \in \mathbb{R}.$$

There exists a positive constant L_0 such that for each $b = (b_1, \dots, b_K) \in \mathcal{B}$

$$\sup_{i\in\mathcal{Y}} |b_i(x) - b_i(y)| + |\sigma(x) - \sigma(y)| \leq L_0|x - y|, \quad \forall (x, y) \in \mathbb{R}^2.$$

Assumption 2.1 ensures the existence and uniqueness of a strong solution for Equation (2.1) and that $\mathbb{E}[\sup_{t\in[0,T]} |X_t|^q] < \infty$ for any integer $q \geq 1$. Furthermore, it implies that

$$\sup_{i\in\mathcal{Y}} |b_i(x)| \leq C_0(1 + |x|). \tag{2.2}$$

Finally, b^* satisfies the following condition.

Assumption 2.2 (Novikov condition)**.**

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T \frac{b_i^{*2}}{\sigma^2}(X_s)ds\right)\right] < +\infty, \quad i \in \mathcal{Y}.$$

Here Novikov's condition (Assumption 2.2) is sufficient to apply Girsanov's theorem (see *e.g.* Revuz & Yor, 2013, Chapter VIII), which is the key ingredient for the proof of Proposition 2.3.

2.2. **Online classification rule.** Let $0 \leq t \leq T$. An online classifier at time t is a measurable function g_t mapping $\mathcal{X}_t$ onto $\mathcal{Y}$. The set of online classifiers at time t is denoted by $\mathcal{G}_t$. The performance of an online classifier g_t is assessed through the misclassification risk associated to g_t and defined by

$$R(g_t) = \mathbb{P}\left(g_t(X) \neq Y\right).$$

The minimizer of R over $\mathcal{G}_t$ is called the Bayes classifier and is denoted by g_t^*. The classifier g_t^* is characterized by

$$g_t^* \in \underset{i\in\mathcal{Y}}{\operatorname{argmax}}\, \pi_t(i), \quad \pi_t^*(i) := \mathbb{P}\left(Y = i|\mathcal{F}_t^X\right). \tag{2.3}$$

The following proposition gives an explicit expression of the online Bayes classifier. This result is similar to the one obtained in Cadre (2013) in the context of binary classification.

Proposition 2.3. *For all $t \in (0,T)$ and each $i \in \mathcal{Y}$ we define*

$$F_t^i := \int_0^t \frac{b_i^*}{\sigma^2}(X_s)dX_s - \frac{1}{2}\int_0^t \frac{(b_i^*)^2}{\sigma^2}(X_s)ds. \tag{2.4}$$

The sequence of conditional probabilities satisfies

$$\pi_t^*(i) = \mathbb{P}\left(Y = i|\mathcal{F}_t^X\right) = \varphi_i(F_t) \quad \mathbb{P}-a.s \tag{2.5}$$

where $F_t = (F_t^1, \ldots, F_t^K)$, and $\varphi_i : (x_1, \ldots, x_K) \mapsto \dfrac{p_i e^{x_i}}{\sum_{j=1}^K p_j e^{x_j}}$ are the softmax functions.

The Bayes classifier g_t^* is the best possible online classifier at time t. Unfortunately, the functions b_i^* are unknown and thus it is unreachable. Proposition 2.3 is a key result and is the main ingredient of all the classification procedures presented in this paper. Indeed, this result highlights the dependencies of the optimal Bayes classifier on the unknown functions b_i^*. Therefore, we naturally focus on classification procedures $\widehat{g}_t$ based on estimators of the functions b_i^* and provide a control of the excess risk $R(\widehat{g}_t) - R(g_t^*)$. The following proposition characterizes the excess risk of an arbitrary online classifier $g_t \in \mathcal{G}_t$.

Proposition 2.4. *The online Bayes classifier g_t^* defined by (2.3) satisfies, for any online classifier g_t,*

$$R(g_t) - R(g_t^*) = \mathbb{E}\left[\sum_{i=1}^K \sum_{k\neq i} |\pi_t^*(i) - \pi_t^*(k)| \mathbb{1}_{\{g_t(X)=k\}} \mathbb{1}_{\{g_t^*(X)=i\}}\right]. \tag{2.6}$$

In the sequel, we only consider online classifier at final time T and then remove the notation dependency on parameter T. Nevertheless, the study of the time dynamics of $R(g_t^*)$ is an important feature to address in a future work and is beyond the scope of this paper.

3. Classification procedure

Let $(X_t)_{t\in[0,T]}$ be the solution of (2.1). We assume that the observation consists of a single discretized sample path $\overline{X}(\omega) := (X_{k\Delta}(\omega))_{k\in\{0,\ldots,n\}}$ with $T = n\Delta$. Through the paper, the asymptotic is chosen to be

$$\Delta \to 0, \quad (n \to +\infty).$$

It is important to note that this asymptotic framework is not classical for the estimation of the drift functions. Usually, the classical asymptotic framework in the estimation of the drift function for solutions of stochastic differential equations is $T \to +\infty$ and ergodicity properties of the diffusion are used in force to handle this problem. In the context of supervised learning, it is possible to assume that the horizon T is fixed because we have at hand a learning sample of independent copies of (X, Y).

In Section 3.1, we describe the set of classifiers of interest which are based on the discrete time observations $(X_{k\Delta})_{k\in\{0,\ldots,n\}}$. In particular, we provide a control of the excess risk of these classifiers. Section 3.2 is devoted to the presentation of general estimation procedures with several theoretical results.

3.1. **Discrete observations classifiers.** Let us define the set of *discrete observations classifiers* which are based on the discrete time observations of X. For a trajectory X (driven by drift $b^* \in \mathcal{B}$), any $b \in \mathcal{B}$, we define for $i \in \mathcal{Y}$ the discrete version of F based on $(X_{k\Delta})_{k\in\{0,\dots,n\}}$ and b:

$$\overline{F}_b^i := \sum_{k=0}^{n-1} \left(\frac{b_i}{\sigma^2}(X_{k\Delta})(X_{(k+1)\Delta} - X_{k\Delta}) - \frac{\Delta}{2}\frac{b_i^2}{\sigma^2}(X_{k\Delta}) \right), \ \overline{F}_b := (\overline{F}_b^1, \dots, \overline{F}_b^K). \tag{3.1}$$

Then we set $\overline{\pi}_b(i) := \varphi_i(\overline{F}_b)$, $i = 1, \dots, K$. The corresponding continuous analogs of $\overline{F}_b$ (resp. $\overline{\pi}_b$) is denoted by F_b (resp. π_b) (so that with these notations $F_T = F_{b^*}$ and $\pi_T^* = \pi^*$). Finally, for any function $b \in \mathcal{B}$, we define the discrete observations classifier $\overline{g}_b$ by

$$\overline{g}_b(X) := \operatorname*{argmax}_{i\in\mathcal{Y}} \overline{\pi}_b(i). \tag{3.2}$$

Hereafter, we state a proposition which gives a bound for the excess risk of some discrete observations classifier $\overline{g}_b$, and highlights the link with the discrete observations and a suitable distance between b and b^*. We introduce the norm $\|.\|_T$ defined for a real valued function f and a process X from model (2.1):

$$\|f\|_T^2 := \sup_{t\in[0,T]} \mathbb{E}[|f(X_t)|^2].$$

Moreover, for a function $b \in \mathcal{B}$, we define the $\|.\|_T$ as $\|b\|_T = \max_{i\in\mathcal{Y}} \|b_i\|_T$.

Proposition 3.1. *Let $b \in \mathcal{B}$. The discrete observations classifier $\overline{g}_b$ satisfies*

$$R(\overline{g}_b) - R(g^*) \le C \left(\sqrt{\Delta} + \|b - b^*\|_T \right)$$

where C is a positive constant which depends on T, K, and on the constants in the Assumptions 2.1, 2.2.

3.2. **Classification procedures.** It is always possible to assert the existence of a probability $\widehat{\mathbf{P}}$ supporting an infinite sequence of independent copies of (X, Y) (with its corresponding expectation $\widehat{\mathbf{E}}$).

Assume we have at our disposal a learning sample of size N denoted by $D_N = (\overline{X}^{(j)}, Y^{(j)})_{j=1,\dots,N}$ which consists of independent copies of $(\overline{X}, Y)$ under $\widehat{\mathbf{P}}$. We define $(N_i)_{i\in\mathcal{Y}}$ by $N_i = \sum_{j=1}^N \mathbb{1}_{\{Y^{(j)}=i\}}$ and D_{N_i} the associated learning sub-sample.

In the sequel, the symbol $\mathbf{P}$ stands for total probability (over the whole sample and a new discrete trajectory of $\overline{X}$, which is assumed to be independent of D_N). We have $\mathbf{P} = \widehat{\mathbf{P}} \otimes \mathbb{P}$.

Based on the observation of D_N, we consider estimators $\widehat{b}_i$ of the drift functions with asymptotic $N \to +\infty, \Delta \to 0$ and T fixed. We denote by $\underset{N,\Delta}{\to}$ this asymptotic. We also write $N \to +\infty$ meaning that $N_i \to \infty$ for all $i \in \mathcal{Y}$. For now, we do not precise anything on the estimator $\widehat{b}$.

In view of (3.2), we naturally consider the classifier

$$\overline{g}_{\widehat{b}}(X) := \operatorname*{argmax}_{i\in\mathcal{Y}} \overline{\pi}_{\widehat{b}}(i). \tag{3.3}$$

3.2.1. *Classification procedures based on consistent drift estimation.* Using the result of Proposition 3.1, we can easily deduce the following result.

Corollary 3.2. *Let $\widehat{b}$ be an estimator of b^* such that $\|\widehat{b}\|_T < \infty$ and $\|\widehat{b} - b^*\|_T \xrightarrow[N,\Delta]{\mathbf{P}} 0$, then*

$$\widehat{\mathbf{E}} \left[R(\overline{g}_{\widehat{b}}) - R(g^*) \right] \underset{N,\Delta}{\to} 0.$$

Corollary 3.2 shows that we may derive consistent classification procedures as soon as the estimator of $\widehat{b}$ remains consistent w.r.t the norm $\|.\|_T$. In this case, the result of Proposition 3.1 ensures that, up to a $\sqrt{\Delta}$ factor, the rate of convergence of such a discrete observations classifier is the same as the rate of convergence of $\widehat{b}$ towards b^* in $\|.\|_T$-norm.

3.2.2. *Empirical risk minimization.* Another way to obtain estimators such that the resulting discrete observations classifiers are consistent, is to consider estimators which rely on the empirical risk minimization principle (see *e.g.* Devroye *et al.*, 1996; Bartlett & Mendelson, 2006; Massart & Nédélec, 2006). In Cadre (2013), the empirical risk minimization procedure is used in the context of binary classification where the features come from continuous diffusion sample paths and are discriminated by their drift.

Following the same idea, we investigate the case where the estimator of b^* is defined as an empirical risk minimizer. To this end, we introduce the empirical risk of a discrete observations classifier $\overline{g}_b$ by

$$\widehat{R}(\overline{g}_b) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{\{\overline{g}_b(\overline{X}^{(j)}) \neq Y^{(j)}\}}.$$

Now, assume that there exists a finite ε-net $\mathcal{B}_\varepsilon \subseteq \mathcal{B}$ with respect to the norm $\|.\|_T$. We define the estimator $\widehat{b}^\varepsilon = (\widehat{b}_1^\varepsilon, \ldots, \widehat{b}_K^\varepsilon)$ as

$$\widehat{b}^\varepsilon \in \operatorname*{argmin}_{b \in \mathcal{B}_\varepsilon} \widehat{R}(\overline{g}_b). \tag{3.4}$$

The following theorem gives the theoretical performances of the classification procedure $\overline{g}_{\widehat{b}^\varepsilon}$ through the excess risk. In particular, we show that $\overline{g}_{\widehat{b}^\varepsilon}$ is consistent and derive its rate of convergence.

Theorem 3.3. *The classifier defined through* (3.4) *satisfies*

$$\widehat{\mathbf{E}}\left[R(\overline{g}_{\widehat{b}^\varepsilon}) - R(g^*)\right] \leq C\left(\sqrt{\frac{\log(\mathrm{Card}\mathcal{B}_\varepsilon)}{N}} + \left(\sqrt{\Delta} + \varepsilon\right)\right)$$

where C is a positive constant which depends on T, K, and on the constants in the Assumptions 2.1, 2.2.

The upper bound of Theorem 3.3 is decomposed into an estimation error which is usual when we deal with empirical risk minimization using ε-nets and an error of order $\sqrt{\Delta}$ coming from the discretization error. This result shows that if $\varepsilon = \varepsilon_N \to 0$ such that $\log(\mathrm{Card}\mathcal{B}_{\varepsilon_N})/N \to 0$, then the classification procedure $\overline{g}_{\widehat{b}^\varepsilon}$ is consistent provided that $\Delta \to 0$.

Along those lines, we provide a corollary of Theorem 3.3 which gives a classical nonparametric rate of convergence depending only on N whenever Δ and ε are well calibrated w.r.t. N.

To this purpose, we make an assumption on the complexity of the class of functions $\mathcal{B}$: we assume that $\mathcal{B} = (\mathcal{H})^K$ and such that there exists an $\varepsilon -$ net $\mathcal{H}_\varepsilon \subseteq \mathcal{H}$, satisfying

$$u > 0, C > 0, \quad \log(\mathrm{Card}\mathcal{H}_\varepsilon) \leq C\varepsilon^{-u}. \tag{3.5}$$

Corollary 3.4. *Under condition (3.5), if $\Delta = O\left(N^{-2/(2+u)}\right)$ and $\varepsilon \propto N^{-1/(2+u)}$ we have*

$$\widehat{\mathbf{E}}\left[R(\overline{g}_{\widehat{b}^\varepsilon}) - R(g^*)\right] \leq O\left(\frac{1}{N^{1/(2+u)}}\right).$$

Note that this nonparametric rate of convergence can be reached here because some complexity assumption on $\mathcal{B}$ is assumed. This rate of convergence is obtained in Cadre (2013) in the context of binary classification with continuous time observations under the same kind of assumptions. In binary classification, Audibert & Tsybakov (2007) shows that this rate is achieved by the plug-in classifiers when the feature $X \in \mathbb{R}^d$ and the regression function belongs to a subset Σ of L_∞ under a similar complexity assumption on Σ.

As an example, we provide an explicit class of functions $\mathcal{B}$ such that the assumption (3.5) holds. Define $\psi_\gamma(x) := C_0(1+|x|)^\gamma$ where $L \geq 1$ (and C_0 given in Equation (2.2)). Let $k \geq 1$ and $\gamma \geq 1$ and consider

$$\mathcal{H}_k = \left\{ b \in \mathcal{C}^k,\ \exists C > 0,\ \forall j \in \mathbb{Z},\ i = 0, \dots, k \ \sup_{[j,j+1[} \left| \frac{d^i b}{dx^i} \right| \leq C(|j|+1)^\gamma,\ |b(x)| \leq \psi_\gamma(x) \right\}$$

Then an application of the result given in Van Der Vaart & Wellner (1996) (see the proof of Theorem 2.7.1) yields the following result.

Proposition 3.5. *Let $k \in \mathbb{N}^*$. The set $\mathcal{B} = (\mathcal{H}_k)^K$ fulfills assumption (3.5) with $u = 1/k$.*

Unfortunately, in general the estimator defined by (3.4) is not computable in practice. However, in Section 4.2, we manage to built an alternative procedure which is still based on the empirical minimization principle.

4. Case of a parametric family of drift functions

In this section, we focus on the case where the set $\mathcal{B}$ is a parametric family of drift functions defined as follows

$$\mathcal{B} = \{(b(\theta_i, .))_{i \in \mathcal{Y}},\ \forall i \in \mathcal{Y},\ \theta_i \in \Theta\},$$

where $\Theta \subset \mathbb{R}^d$ is compact and for each $\theta \in \Theta$, $x \mapsto b(\theta, x)$ is a real valued function which satisfies Assumptions 2.1, 2.2. Moreover, we assume that the function b is known. For each $i \in \mathcal{Y}$, we denote the drift functions by $b_i^*(x) := b(\theta_i^*, x)$, $\theta_i^* \in \Theta$ (and $\pi^* = \pi_{b_{\theta^*}}$). Furthermore, for $\theta = (\theta_1, \dots, \theta_K) \in \Theta^K$, we denote the vector $(b(\theta_i, .))_{i \in \mathcal{Y}}$ by $b_\theta = (b_{\theta_1}, \dots, b_{\theta_K})$. Finally, for $\theta \in \Theta^K$, we also define $\|\theta\| = \max_{i \in \mathcal{Y}} \|\theta_i\|_\infty$. In order to derive consistent classification procedures which rely on the estimation procedure described in Section 3.2, we add two assumptions.

Assumption 4.1. *(Identifiability condition)*

$$\forall \theta, \theta' \in \Theta,\ \ \mathbb{E}\left[\int_0^T (b(\theta, X_s) - b(\theta', X_s))^2 ds \right] = 0 \Leftrightarrow \theta = \theta'.$$

Assumption 4.2. *Function b is Lipschitz-continuous with respect to $\theta \in \Theta$:*

$$|b(\theta, x) - b(\theta', x)| \leq C\,(1 + |x|^\alpha)\,\|\theta - \theta'\|_\infty \ \text{ for some } \ \alpha \geq 1.$$

The first assumption is classical when we deal with parametric estimation. The second one implies that for $\theta, \theta^{'} \in \Theta$, we have

$$\|b(\theta, \cdot) - b(\theta', \cdot)\|_T \leq C\|\theta - \theta^{'}\|_\infty, \tag{4.1}$$

for a constant C depending on T and α.

In Section 4.1 and in the special case where $\sigma \equiv 1$ and $d = 1$, we derive a classification procedure based on a minimum contrast estimation principle for θ^* built on a the learning sample D_N, which is shown to be consistent. Note that the problem of consistent estimators of parameters θ^* is not new, but in our context we drop the assumption of long time observation and take advantage of the learning sample.

In Section 4.2 we derive an implementable classification procedure which relies on the empirical risk minimization principle. This procedure involves a convex surrogate of the minimization problem described in Section 3.2.

4.1. **Contrast estimator.** In this section assume that $\sigma \equiv 1$ and $d = 1$.

We consider estimators of θ^* defined as the minimizer of a contrast function based on the Gaussian log-likelihood approximation from the approximated discrete-time Euler-Maruyama scheme (Kessler, 1997).

For each $i \in \mathcal{Y}$ and $\theta \in \Theta$, let us define $L_\Delta(D_{N_i}; \theta)$ as the Euler-approximation of the likelihood function,

$$L_\Delta(D_{N_i};\theta) \;=\; \prod_{j=1}^{N_i}\prod_{k=0}^{n-1}\sqrt{\frac{\Delta}{2\pi}}\exp\left(-\frac{\Delta}{2}\left(\frac{X^{(j)}_{(k+1)\Delta}-X^{(j)}_{k\Delta}}{\Delta}-b(\theta,X^{(j)}_{k\Delta})\right)^2\right).$$

Then we naturally consider the associated contrast function given by

$$\gamma_{N_i,n}(\theta) := \frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}\left(\frac{\Delta}{2}b^2(\theta,X^{(j)}_{k\Delta}) - b(\theta,X^{(j)}_{k\Delta})(X^{(j)}_{(k+1)\Delta}-X^{(j)}_{k\Delta})\right). \tag{4.2}$$

The minimum contrast estimator is

$$\widehat{\theta}_i \in \underset{\theta\in\Theta}{\operatorname{argmin}}\ \gamma_{N_i,n}(\theta). \tag{4.3}$$

Let us state the asymptotic properties of the estimators $\widehat{\theta}_i$.

Theorem 4.3 (Consistency)**.** *Under Assumptions 4.1, the estimator $\widehat{\theta}_i$ given by Equation* (4.3) *satisfies,*

$$\widehat{\theta}_i \xrightarrow[N,\Delta]{\widehat{\mathbf{P}}} \theta_i^*.$$

The proof relies on consistency results for minimum contrast estimators (see e.g. Dacunha-Castelle & Duflo, 1983) and a crucial lemma given in Yoshida (1990).

The asymptotic normality is given in the following theorem. Denote $\dot{b}$ the partial derivative of b w.r.t the parameter θ.

Theorem 4.4 (Asymptotic normality)**.** *Under Assumptions 4.1, 4.2, assume that there exists $C' > 0$ such that*

$$\forall x, y \in \mathbb{R},\ \sup_{\theta\in\Theta}|\dot{b}(\theta,x)-\dot{b}(\theta,y)| \le C'|x-y|$$

and there exist $\beta \ge 1$, $C'' > 0$ such that for all $\theta, \theta' \in \Theta$, for all $x \in \mathbb{R}$

$$|\dot{b}(\theta,x)-\dot{b}(\theta',x)| \le C''\left(1+|x|^\beta\right)|\theta-\theta'|$$

with a constant C'' depending on T and β.

Then for each $i \in \mathcal{Y}$, under the condition $\Delta = o(1/N)$,

$$\sqrt{N_i}(\widehat{\theta}_i-\theta_i^*) \xrightarrow[N]{\mathcal{L}} \mathcal{N}\left(0, \mathbb{E}\left[\int_0^T \dot{b}^2(\theta_i^*,X_s)ds\right]^{-1}\right)$$

where the convergence takes place under $\widehat{\mathbf{P}}$.

The asymptotic variance is classically the inverse of the Fisher information for continuous diffusion processes at finite time (see *e.g.* Kutoyants, 2004). The constraint $N\Delta = o(1)$ is needed to reach the asymptotic normality in the context of discrete observations. Note that the same condition is required in Delattre *et al.* (2015a) in the mixed diffusion context to estimate a distribution parameter in the diffusion coefficient. Our estimator is also asymptotically Gaussian at the same rate $\sqrt{N_i}$.

We use the asymptotic properties of the estimator $\widehat{\theta}$ to show the consistency of the classification procedure $\overline{g}_{b_{\widehat{\theta}}}$. Applying Corollary 3.2, Theorem 4.3 and Equation (4.1) leads to the following convergence.

Proposition 4.5. *Under Assumption 4.2, the predictor* $\overline{g}_{b_{\widehat{\theta}}}$ (3.3) *satisfies*

$$\widehat{\mathbf{E}}\left[R(\overline{g}_{b_{\widehat{\theta}}}) - R(g^*)\right] \underset{N,\Delta}{\longrightarrow} 0.$$

This result does not provide a rate of convergence. In order to obtain a more refined result one needs to control the moments of the estimator $\widehat{\theta}$. This problem has been investigated in the continuous observation setting for a single trajectory ($T \to \infty, N = 1$) in Kutoyants (2004), or in Dion & Genon-Catalot (2015) for a special drift function $b(x, \theta)$ of the multiplicative form $\theta b(x)$.

In the next section, we focus on the classification procedures based on the empirical risk minimization principle.

4.2. **Empirical risk minimizer.** The diffusion coefficient is no more assumed to be constant and the dimension parameter d may be chosen greater than one.

We can apply the Proposition 3.3 to prove the consistency of the resulting classifier $\overline{g}_{\widehat{\theta}}$ where $\widehat{\theta} \in \underset{\theta\in\Theta_N}{\operatorname{argmin}}\ \widehat{R}(\overline{g}_{b_\theta})$. Nevertheless, the estimator $\widehat{\theta}$ is the solution of a non convex minimization problem and can be computationally intractable. To overcome this difficulty it is classical to propose a convex surrogate of the previous minimization problem (see *e.g.* Zhang, 2004; Bartlett *et al.*, 2006; Biau *et al.*, 2015). In this context, we focus on the following minimization problem

$$\widehat{h} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ \frac{1}{N}\sum_{j=1}^{N} \Psi_{Y^{(j)}}(h(X^{(j)})),$$

where Ψ_Y is a real-valued function that takes a vector of $\mathbb{R}^K$ as its argument and $\mathcal{H}$ is the set of *score functions*. From $\widehat{h} = \left(\widehat{h}^1, \ldots, \widehat{h}^K\right)$, we can define a classifier

$$\widehat{g} = \underset{i\in\mathcal{Y}}{\operatorname{argmax}}\ \widehat{h}^i.$$

The function h returns a score for each label and naturally the chosen label is the one maximizing the score. The consistency of the classification rules obtained in this minimization risk framework with respect to the misclassification risk depends on the several possible choices for the function Ψ_Y.

4.2.1. *Constrained method.* The constrained comparison method is dedicated to the multiclass framework (see *e.g.* Zhang, 2004; Tewari & Bartlett, 2007; Pires *et al.*, 2013). Let us consider the convex set of constrained score functions $\mathcal{H} = \left\{h = (h^1, \ldots, h^K)\ : \mathcal{X}_T \to \mathbb{R}^K, \quad \sum_{i=1}^K h^i = 0\right\}$. For $\phi : \mathbb{R} \to \mathbb{R}^+$ a convex function, the ϕ-risk associated to $h \in \mathcal{H}$ and the minimizer are given by

$$R_\phi(h) = \mathbb{E}\left[\sum_{i=1}^{K} \mathbb{1}_{\{Y\neq i\}}\phi\left(-h^i(X)\right)\right], \quad h^* \in \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ R_\phi(h). \tag{4.4}$$

This risk formulation encourages small scores for $i \neq Y$ and, due to the sum to zero constraint, large score for $i = Y$. Moreover, the empirical counterpart of $R_\phi(h)$ is

$$\widehat{R}_\phi(h) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K} \mathbb{1}_{\{Y^{(j)}\neq i\}}\phi\left(-h^i(X^{(j)})\right). \tag{4.5}$$

An important property required for the function ϕ is the calibration property which implies the consistency of the ϕ-risk.

Definition 4.6. *The function* ϕ *is calibrated if for any sequence of measurable score functions* h_N

$$R_\phi(h_N) \to R_\phi(h^*) \qquad \textit{implies that} \qquad R(g_N) \to R(g^*),$$

where $g_N := \underset{i\in\mathcal{Y}}{\operatorname{argmax}}\ h_N^i$.

Hence, ϕ is calibrated if any consistent procedure for the ϕ-risk remains consistent for the misclassification risk. A characterization of the calibration property may be found in Zhang (2004) :

Proposition 4.7. *(Zhang (2004)) The function ϕ is calibrated if ϕ is non negative, $\phi^{'}(0)$ exists and $\phi^{'}(0) < 0$.*

Let us consider now $\phi : x \mapsto \exp(-x)$ the exponential calibrated loss. In this case h^* given by Equation (4.4) is

$$h^{*i}(X) = \frac{1}{K}\sum_{\ell=1}^{K} \log\left(\frac{1-\pi^*(\ell)}{1-\pi^*(i)}\right).$$

Therefore, for $\theta \in \Theta^K$ we should consider the score functions $\overline{h}_\theta$ defined for $i \in \mathcal{Y}$ by

$$\overline{h}^i_\theta(\overline{X}) = \frac{1}{K}\sum_{\ell=1}^{K} \log\left(\frac{1-\overline{\pi}_{b_\theta}(\ell)}{1-\overline{\pi}_{b_\theta}(i)}\right).$$

As shown in Definition 4.6, the consistency of a procedure based on the empirical minimization of this ϕ-risk involves a control of the excess ϕ-risk over the set of all possible score functions $\overline{h}_\theta$. Unfortunately, it does not seem possible to control the ϕ-risk if one of the $\overline{\pi}_{b_\theta}(i)$ is too close too 1. In order to circumvent this difficulty, let us fix some threshold $0 < \varepsilon < 1/3$. For an observed vector $\overline{\pi}_{b_\theta}$ and

$$i_0 = \operatorname*{argmax}_{i\in\mathcal{Y}} \overline{\pi}_{b_\theta}(i)$$

we define the vector $\overline{\pi}^\varepsilon_{b_\theta}$ as follows

$$\overline{\pi}^\varepsilon_{b_\theta}(i) := \overline{\pi}_{b_\theta}(i)\mathbb{1}_{\{\overline{\pi}_{b_\theta}(i_0)<1-\varepsilon\}} + \mathbb{1}_{\{\overline{\pi}_{b_\theta}(i_0)\geq 1-\varepsilon\}}\left(\left(\overline{\pi}_{b_\theta}(i) + \frac{\overline{\pi}_{b_\theta}(i_0)-(1-\varepsilon)}{K-1}\right)\mathbb{1}_{\{i\neq i_0\}} + (1-\varepsilon)\mathbb{1}_{\{i=i_0\}}\right),$$

and we denote by $\pi^\varepsilon_{b_\theta}$ its continuous counterpart (depending h^i_θ with natural notations). One may easily check that for each θ, $\sum_{i=1}^K \overline{\pi}^\varepsilon_{b_\theta}(i) = 1$ by construction. Then, the condition $\varepsilon < 1/3$ ensures that for each $i \in \mathcal{Y}$, $\overline{\pi}^\varepsilon_{b_\theta}(i) \leq 1-\varepsilon$. Finally, for each $\theta \in \Theta$, we consider the score functions

$$\overline{h}^{\varepsilon,i}_\theta(\overline{X}) = \frac{1}{K}\sum_{\ell=1}^{K} \log\left(\frac{1-\overline{\pi}^\varepsilon_{b_\theta}(\ell)}{1-\overline{\pi}^\varepsilon_{b_\theta}(i)}\right). \tag{4.6}$$

The resulting sets of score functions are bounded. Indeed, for each $\theta \in \Theta$ and $i \in \mathcal{Y}$,

$$|\overline{h}^{\varepsilon,i}_\theta(\overline{X})| \leq \log\left(\frac{1}{\varepsilon}\right).$$

We are now able to define the empirical risk threshold minimizer

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta^K} \widehat{R}_\phi(\overline{h}^\varepsilon_\theta), \tag{4.7}$$

with $\widehat{R}_\phi$ given in Equation (4.5). Note that $\widehat{\theta}$ depends on ε but for sake of simplicity the dependency will only appears on the notation of the scores $h, \overline{h}$. The following proposition establishes the consistency of the corresponding classification strategy with respect to the ϕ-risk of $\overline{h}^\varepsilon_{\widehat{\theta}}$.

Proposition 4.8. *Assume that $\Theta = [0,1]^d$ and that there exists $\alpha > 0$ such that $\Delta = O(N^{-\alpha})$. Under Assumption 4.2, if $\varepsilon = O\left(N^{-\beta}\right)$ with $0 < \beta < \min(1/2, \alpha/4)$ then the classification procedure $\overline{h}^\varepsilon_{\widehat{\theta}}$ given by* (4.6) *and* (4.7) *satisfies,*

$$\widehat{\mathbf{E}}\left[R_\phi(\overline{h}^\varepsilon_{\widehat{\theta}}) - R_\phi(h^*)\right] \underset{N}{\longrightarrow} 0.$$

Therefore, the results of Proposition 4.8 and the calibration property ensure the consistency of the classification procedure $\overline{g}_{b_{\widehat{\theta}}}$ with respect to the misclassification risk.

Nevertheless, this result does not specify the rate of convergence of $\overline{g}_{b_{\widehat{\theta}}}$ since, up to our knowledge, there is no explicit link between the convergence rates of the excess misclassification risk and those of the excess ϕ-risk.

In the Section 4.2.2, we consider another formulation of the ϕ-risk for which we manage to derive a rate of convergence.

4.2.2. *One-Versus-All method.* Let us now study another approach based on the one-versus-all principle of Zhang (2004). Recently, this risk has also been considered in Denis & Hebiri (2017) for the confidence sets aggregation. We consider the set of score functions $\mathcal{H} = \{h = (h^1, \ldots, h^K) \, : \mathcal{X}_T \to \mathbb{R}^K\}$. For ϕ a convex function and h_θ, $\theta \in \Theta^K$ a vector of score functions, we define the following risk function,

$$R_\phi\left(h\right) = \mathbb{E}\left[\sum_{i=1}^{K} \phi(Z_i h_\theta^i(X))\right], \; Z_i = 2\,\mathbb{1}_{\{Y=i\}} - 1. \tag{4.8}$$

We can note that there is no sum-of-zero constraint on the vector of score function h_θ here. Equation (4.8) can be rewritten as

$$R_\phi(h) = \mathbb{E}\left[\sum_{i=1}^{K} \pi^*(i)\phi(h_\theta^i(X)) + (1-\pi^*(i)))\phi(-h_\theta^i(X))\right].$$

Therefore, this formulation can be viewed as K separate binary classification problems where for each $i \in \mathcal{Y}$ we focus on the classification problem $Y = i$ against $Y \neq i$. Now we consider $\phi : x \mapsto (1-x)^2$ the least squares loss. In this case, we easily deduce that

$$h^{*i}(X) = 2\pi^*(i) - 1, \;\; i \in \mathcal{Y}.$$

As for the constrained comparison method, we consider the following estimator $\widehat{\theta}$ of θ^*

$$\widehat{\theta} \in \underset{\theta \in \Theta^K}{\operatorname{argmin}} \, \widehat{R}_\phi(\overline{h}_\theta), \; \overline{h}_\theta^i = 2\overline{\pi}_{b_\theta}(i) - 1, i \in \mathcal{Y}, \tag{4.9}$$

(with $\widehat{R}_\phi$ given in (4.5)). There are two advantages of considering the least squares loss with the one-versus-all approach. The first one is that the score functions $\overline{h}_\theta$ are bounded and there is no need to define a set of thresholded score functions. The second one is the Zhang's Lemma (see Zhang, 2004) that we state here for the least squares loss.

Lemma 4.9. *For ϕ the least squares loss, the resulting classifier $g_{b_{\widehat{\theta}}}$ with $\widehat{\theta}$ given in Equation* (4.9)*, satisfies*

$$\widehat{\mathbf{E}}\left[R(\overline{g}_{b_{\widehat{\theta}}}) - R(g^*)\right] \leq \frac{1}{\sqrt{2}}\left(\widehat{\mathbf{E}}\left[R_\phi(\overline{h}_{\widehat{\theta}}) - R_\phi(h^*)\right]\right)^{1/2}.$$

As a consequence, using similar arguments as in proof of Proposition 4.8, one can show the consistency of $\overline{g}_{b_{\widehat{\theta}}}$ with respect to the misclassification excess risk provided that $\Delta = O(N^{-\alpha})$ for $\alpha > 0$. Furthermore if $\alpha \geq 2$, we derive the following rate of convergence

Theorem 4.10. *Assume that $\Theta = [0,1]^d$ and that there exists $\alpha \geq 2$ such that $\Delta = O(N^{-\alpha})$. Under Assumption 4.2, the classification procedure $\overline{g}_{b_{\widehat{\theta}}}$ given by* (4.9) *satisfies,*

$$\widehat{\mathbf{E}}\left[R(\overline{g}_{b_{\widehat{\theta}}}) - R(g^*)\right] \leq O\left(\sqrt{\frac{\alpha d \log(N)}{N}}\right).$$

We can note that up to the logarithmic factor, we obtain a rate of convergence of order of $N^{-1/2}$. Comparing to the rate provided in Corollary 3.4, this rate is better due to the lower complexity of

the parametric model. Interestingly, if we consider $\widehat{\theta} \in \underset{\theta\in\Theta_N}{\operatorname{argmin}}\, \widehat{R}(\overline{g}_{b_\theta})$ with Θ_N a $1/N$-net of Θ^K, from Theorem 3.3, one can show that the rate of convergence is also of order $N^{-1/2}$. Hence, from a theoretical point of view, the use of convex surrogate does not degrade the performances of the classification procedure when $\alpha \geq 2$.

5. Simulation study

In this section, we investigate the numerical performances of the proposed classification procedures through a simulation study. The simulation scheme is presented in Section 5.1. In Section 5.2, we evaluate the quality of the classification procedures described in Section 4 and illustrate the statistical properties of the contrast estimator.

5.1. **Description and examples.**

Let us describe the models under consideration for our numerical experiments. We fix $K = 3$, $p_i = 1/K$ and $\sigma = 1$. We consider the following examples:

(1) *Additive OU* $b(\theta, x) = -(x - \theta)$, $x_0 = 4$;
(2) *Multiplicative OU* $b(\theta, x) = -\theta x$, $x_0 = 4$;
(3) *Polynomial* $b(\theta, x) = -(x - \theta)^3 - (x + \theta)^3$, $x_0 = 4$.
(4) *Hyperbolic* $b(\theta, x) = -\theta x/\sqrt{1 + x^2}$, $x_0 = 4$.

We investigate $K = 3$. Hence in our study the number of class K is not suppose to grow and be large. We compare the results on the design: $\theta^* = \{1, 2, 4\}$ for model 1, 2, 4, and $\theta^* = \{1/4, 1/2, 1\}$ for model 3. Models 1 and 2 are widely used in practical applications, and they satisfy all the assumptions required for our theoretical results, while the model 3 does not fulfill the Assumption 2.1, illustrating the robustness of the classification procedures. Model 4 is known as the hyperbolic model in mathematical finance (see *e.g.* Eberlein *et al.*, 1995), and it is used to model log-returns of assets prices in stock markets.

The trajectories are simulated from the Euler scheme. Note that the two first models are simulated from the exact solution of the equation. In Figure 1, we plot some trajectories generated according to the model 1 (Additive). At first sight, without the knowledge of the labels, it seems to be difficult to assign a class to each trajectory. It illustrates the difficulty of this classification problem (see Table 1).

To apply the first procedure, let us give the minimum contrast estimators (4.3) in the two first examples. For each $i \in \mathcal{Y}$, for additive OU:

$$\widehat{\theta}_i = \frac{\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}(X^{(j)}_{(k+1)\Delta} - X^{(j)}_{k\Delta} + \Delta X^{(j)}_{k\Delta})}{N_i n\Delta},$$

and for multiplicative OU:

$$\widehat{\theta}_i = -\frac{\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}(X^{(j)}_{(k+1)\Delta} - X^{(j)}_{k\Delta})X^{(j)}_{k\Delta}/\Delta}{\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}(X^{(j)}_{k\Delta})^2}.$$

In order to illustrate our convergence results, for each model we provide an evaluation of the misclassification risk of the Bayes rule. At this end, we repeat B times the following steps:

i) simulate a data set $\mathcal{D}_M$ with $M = 10000$ and 2500 points for each trajectory,
ii) based on $\mathcal{D}_M$, compute the misclassification error rate of the classifier $\overline{g}_{b_{\theta^*}}$.

Finally, we compute the mean and standard deviation of the misclassification risk, the results are reported in Table 1 with $B = 100$. One can see that model 1 and model 4 seem to be more tricky for the misclassification risk. This is due to the fact that the classes generated by θ^*_1 and θ^*_2 are much overlapped. On the contrary, the classification problem involved by model 2 is more easier although the considered design is the same.

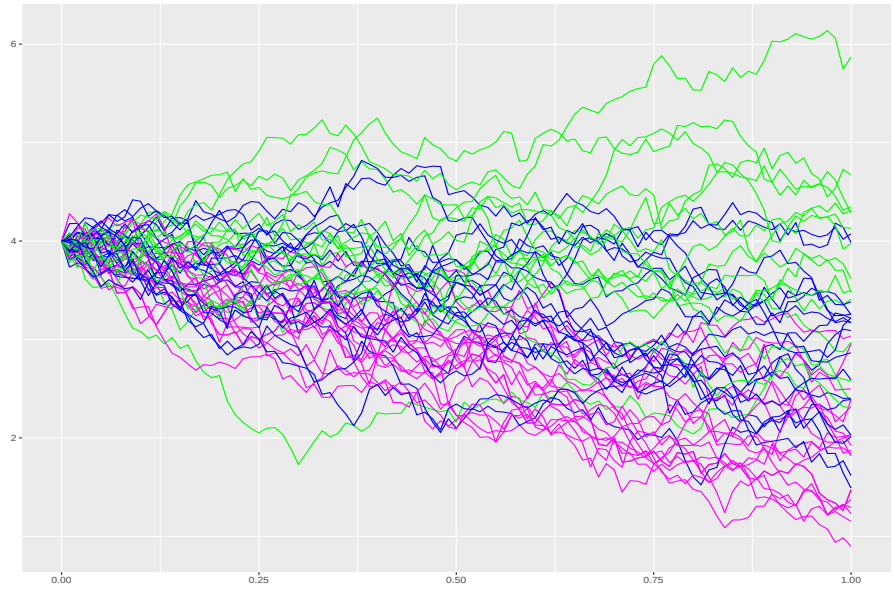

FIGURE 1. Illustration of the classification problem with 3 classes (blue, green and magenta) for the three different values of θ^*) for model *Additive* with $n = 100,\ N = 100$.

5.2. **Numerical performances of the classification procedures.** Now, for each model we evaluate the misclassifcation risk of the three classification procedures presented in Section 4. The procedure based on the contrast estimation is referred as `MLE`: $\overline{g}_{\widehat{\theta}}$ with $\widehat{\theta}$ from Equation (4.3); the procedure which relies on the constrained method is referred as `CM` $\overline{g}_{\widehat{\theta}}$ with $\widehat{\theta}$ given in Equation (4.7) (with $\varepsilon = 0.01$); while the procedure based on the one-versus-all strategy is referred as `OVA` $\overline{g}_{\widehat{\theta}}$ with $\widehat{\theta}$ given in Equation (4.9). The three procedures rely on an optimization function (in Python or R languages `optim` is used with argument method `"BFGS"`). In the case where the `MLE` is an explicit estimator the procedure is naturally fast. Other optimization functions could be used to reduce the computational cost of the procedures in other cases.

In order to stress the robustness w.r.t the theoretical conditions between Δ, N, we consider the following asymptotics: $n \in \{50, 250\},\ \Delta = 1/n,\ N \in \{50, 500\}$.

For each classification procedure and each model, we repeat independently B times the following steps:

i) simulate two datasets $\mathcal{D}_N$ and $\mathcal{D}_M$ with $M = 1000$. For each trajectory of the datasets, simulate first a trajectory with 2500 points and then consider the subsampled trajectory with n equidistant points,
ii) from $\mathcal{D}_N$, compute the considered classification rule $\overline{g}_{\widehat{\theta}}$,
iii) evaluate the misclassification error rate of $\overline{g}_{\widehat{\theta}}$ from $\mathcal{D}_M$.

Table 2 and 3 provide the mean and standard deviation of the results. Our main observation is that, except for model 3 with $n = 50$, all the classification procedures perform well. Indeed, the evaluation of the misclassifcation risk are closed to the Bayes risk with small variances. In particular, for $N = 500$, the classification procedures have similar performances. Furthermore, we can see the influence of the sample size for the procedures `CM` and `OVA`. For instance for model 1 with $n = 50$, the risk of the procedure `CM` is evaluated at 0.34 (with standard deviation equal to 0.04) for $N = 50$, while it is evaluated at 0.31 (with standard deviation equal to 0.01) for $N = 500$. Interestingly, this is not the case for `MLE`. Hence, it seems to be preferable to use the classification procedure `MLE` when

	`Bayes rule`
Model 1	0.31 (0.002)
Model 2	0.12 (0.003)
Model 3	0.22 (0.003)
Model 4	0.33 (0.004)

TABLE 1. Average and standard deviation of the misclassification error rate for the Bayes classifier with $n = 2500$.

	`MLE`		`CM`		`OVA`	
	$N = 50$	$N = 500$	$N = 50$	$N = 500$	$N = 50$	$N = 500$
Model 1	0.32 (0.02)	0.31 (0.01)	0.34 (0.04)	0.31 (0.01)	0.33 (0.05)	0.31 (0.01)
Model 2	0.12 (0.01)	0.12 (0.01)	0.13 (0.02)	0.12 (0.01)	0.13 (0.02)	0.12 (0.01)
Model 3	0.67 (0.02)	0.67 (0.01)	0.67 (0.01)	0.67 (0.01)	0.66 (0.02)	0.66 (0.01)
Model 4	0.34 (0.01)	0.33 (0.01)	0.36 (0.04)	0.33 (0.01)	0.35 (0.03)	0.33 (0.01)

TABLE 2. Average and standard deviation of the misclassification error rate for the three procedures with $n = 50$.

	`MLE`		`CM`		`OVA`	
	$N = 50$	$N = 500$	$N = 50$	$N = 500$	$N = 50$	$N = 500$
Model 1	0.32 (0.02)	0.31 (0.01)	0.33 (0.03)	0.31 (0.01)	0.32 (0.03)	0.31 (0.01)
Model 2	0.12 (0.01)	0.12 (0.01)	0.13 (0.02)	0.12 (0.01)	0.13 (0.02)	0.12 (0.01)
Model 3	0.23 (0.01)	0.23 (0.01)	0.24 (0.03)	0.23 (0.01)	0.22 (0.02)	0.22 (0.01)
Model 4	0.33 (0.02)	0.33 (0.01)	0.36 (0.05)	0.33 (0.01)	0.34 (0.03)	0.33 (0.01)

TABLE 3. Average and standard deviation of the misclassification error rate for the three procedures with $n = 250$.

the sample size is moderate. Then, we can see that the parameter n plays a crucial role for model 3. Indeed, for $n = 50$ all procedures have poor performances while for $n = 250$ the empirical risks are all close to the Bayes classifier.

Finally, we recall that the procedure `MLE` relies on estimators of the design θ^* for which we provide consistency and asymptotic normality in Section 4. Hereafter, we briefly evaluate these properties. Considering $\theta^* = 1$ for model 1, $\theta^* = 3$ for model 2, $\theta^* = 1/2$ for model 3, and $\theta^* = 2$ for model 4, we evaluate the empirical quadratic risks of the estimator on $B = 300$ repetitions. A dataset consists of $N \in \{50, 100, 1000\}$ trajectories composed of $n = 250$ points. The results are shown in Table 4. As expected, the evaluation of the quadratic risk (given $\times 10^2$) are increasingly closed to 0 with respect to N. At the same time, one can see that the estimates have relatively poor performances for small N especially for model 1. We illustrate in Figure 2 the constraint $N/n = o(1)$ required for the asymptotic normality. The variance decreases with N and the bias decreases with n. As expected, the best situation is to have both parameters large enough. Lastly, Figure 3 gives an illustration of Theorem 4.4. One can see that the empirical distribution functions of $\sqrt{N}\left(\widehat{\theta} - \theta\right)$ are closed to the Gaussian distribution function of the theoretical limit.

6. Discussion

In this paper, we provide an explicit formula for the Bayes classifier in the special setting of multiclass discretized diffusion paths discriminated by their drift functions. Section 3 is devoted to

		$N = 50$	$N = 100$	$N = 1000$
Model 1	$\theta = 1$	2.26 (3.13)	0.95 (1.35)	0.10 (0.13)
Model 2	$\theta = 3$	0.81 (1.14)	0.38 (0.50)	0.07 (0.08)
Model 3	$\theta = \frac{1}{2}$	3.80 (1.32)	3.71 (0.88)	3.67 (0.28)
Model 4	$\theta = 2$	2.40 (3.44)	1.23 (1.73)	0.11 (0.15)

TABLE 4. Average and standard deviation of $10^2\times$ quadratic error for the MLE estimator with $n = 250$.

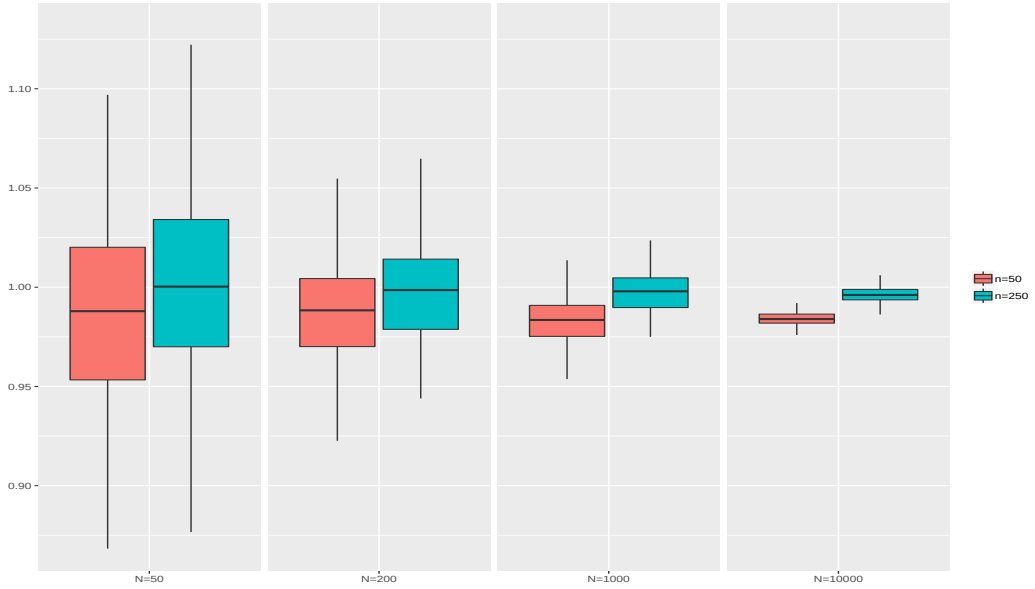

FIGURE 2. Boxplots for the MLE estimators of $\theta^* = 1$ in the additive OU model 1 as a function of parameters n and N.

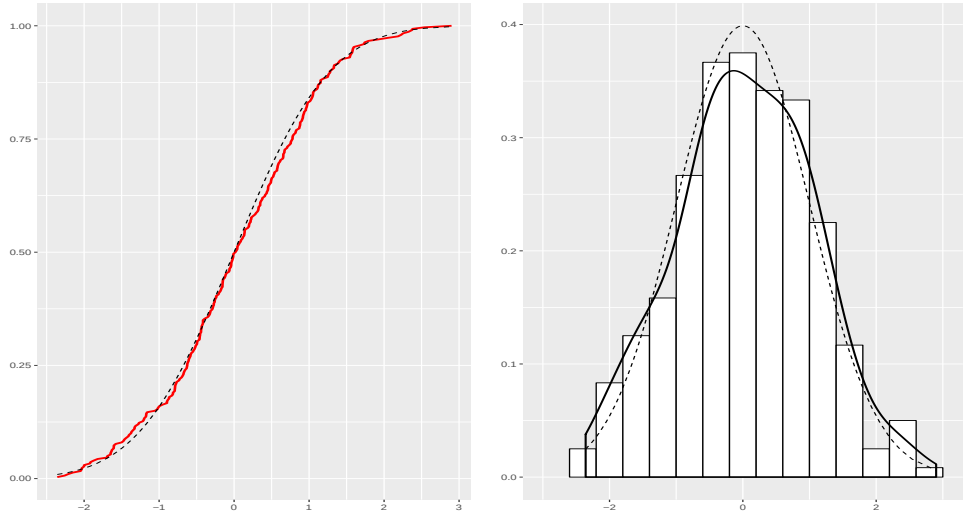

FIGURE 3. Illustration of the asymptotic normality of $\widehat{\theta}$. Left: estimated cumulative distribution function of $Z := \sqrt{N}(\widehat{\theta} - \theta^*)$ (solid curve) and the theoretical one (dotted curve). Right: histogram of Z_i's with the estimated density (solid curve) and the theoretical one (dotted curve), with $N = 1000$, $n = 1000$.

theoretical guarantees for two types of classification procedures. The first one relies on some consistent estimator of the drift function w.r.t. the $\|\|\|_T$-norm. The second one involves the empirical risk minimization principle. Section 4 focuses on the case of parametric drift functions. In this setting, we investigate the two methods and prove their consistency. Moreover, we derive three easily implementable algorithms: `MLE`, `CM` and `OVA`.

Let us make a few comments on these algorithms. Contrary to `CM` and `OVA`, `MLE` is studied only in the case $\sigma \equiv 1$ and $d = 1$. However, note that if we assume that $\mathcal{D}_N$ is observed until some time horizon T' and a new datum X is observed until another time T with $T' \neq T$, then from Corollary 3.2 and Equation (4.1) the consistency of the `MLE` procedure still holds, whereas this is no more the case for the other procedures. There are other many remaining questions of interest. We wish also to investigate nonparametric estimators of the drift function, based on the learning sample, that are consistent w.r.t. the $\|\|\|_T$-norm. To our knowledge, this is an open issue. Furthermore and from the statistical learning point of view, the case where K is very large is a crucial question.

Let us now comment some of our framework assumptions. In Equation (2.1) we made two strong assumptions: the law of Y and the diffusion coefficient σ are assumed to be known. Several strategies may be investigated to overcome these restrictions leading to an extension of the results presented in this paper.

The case where the distribution $(p_i)_{i\in\mathcal{Y}}$ is unknown is not really an issue : it is always possible to plug an empirical estimator of the distribution $(p_i)_{i\in\mathcal{Y}}$ for the estimation of the softmax functions $(\varphi_i)_{i\in\mathcal{Y}}$. We can then extend our results to obtain the consistency of these procedures (with possibly different rates of convergence). Another idea would be to include the possible distributions of Y in the empirical risk minimization as done in (Cadre, 2013).

Regarding the case where the diffusion coefficient is unknown, we believe that things are more intricate. It is well known that the coefficient σ can be estimated in the high-frequency scheme of observations. For example estimators studied in Genon-Catalot & Jacod (1993); Gloter (2000); Jakobsen & Sørensen (2017) may be used. For example a strategy where an estimator of σ is plugged in the expressions of $\overline{F}$ (3.1) could be investigated and one may hope to extend the results of this paper in this new framework. However, the contrast function of the plug-in classification procedure described in Section 4.1 seems more tricky to study. Note that for the empirical risk minimization procedure, one may adapt the strategy in Cadre (2013) and circumvent this difficulty by including $(b_i/\sigma^2)_{i\in\mathcal{Y}}$ and $(b_i^2/\sigma^2)_{i\in\mathcal{Y}}$ in the minimization procedure. In the context where σ is unknown, another challenging issue would be to consider the case where classes are discriminated by both the drift and the diffusion coefficients. Indeed, it seems difficult to adapt directly the previous strategies that rely heavily on Girsanov's formula.

One may also think of generalizing the initial model. In order to cover a broader class of possible applications, a first generalization would be to extend our results to the case of diffusions with inhomogeneous coefficients. We believe our results extend easily to a framework where the classes are discriminated by inhomogeneous drift functions as long as suitable assumptions are made (for e.g. the drift functions are Lipschitz for the space variable uniformly w.r.t the time variable). With such assumptions all the procedures of the paper remain consistent and we chose not to present them for the sake of clarity and conciseness. However, relaxing the Lipschitz assumptions on the drift coefficients or even more, allowing an inhomogeneous diffusion coefficient are important questions that should be addressed for further investigations. Other interesting stochastic process models could be also imagined, for example the case of jump processes discriminated either by their drift or their jump component would be a very interesting development of this work. Indeed, in some cases the likelihood function is available. Our work should only be considered as a first step towards the construction of similar classification procedures for trajectories of much richer classes of stochastic processes.

Acknowledgments

The authors would like to thank Valentine Genon-Catalot for the many fruitful discussions and advice.

7. Proofs

This section is devoted to the proofs of the announced results.

7.1. **Technical results.** We give here some useful results needed in the following proofs, nevertheless they may have an interest *per se*.

Lemma 7.1. *Let $x \mapsto f(x)$ be a continuous, real-valued function, then*

$$\Delta \sum_{k=0}^{n-1} f(X_{k\Delta}) \xrightarrow{\mathbb{P}} \int_0^T f(X_s)ds.$$

This is a consequence of the convergence of Riemann sums.

Lemma 7.2 (Gloter A.(2000))**.** *If X is a diffusion process from model (2.1), for $k \in \mathbb{N}^*$,*

$$\forall t, t+h \in [0,T],\ \mathbb{E}\left[\sup_{s\in[t,t+h]} |X_s - X_t|^k | \mathcal{F}_t\right] \leq c(k)(1+|X_t|^k)h^{k/2}.$$

This result comes from Gronwall's Lemma and the Burkholder-Davis-Gundy inequality and is given in Gloter (2000).

Lemma 7.3. *For any $b_\theta \in \mathcal{B}$ (defined in Section 4),*

$$\mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta} b(\theta, X_{k\Delta}) - b(\theta, X_s)ds\right)^p\right] = O(\Delta^{3p/2})$$

for any integer $p \geq 2$.

Proof of Lemma 7.3. According to Cauchy-Schwarz and Jensen's inequality,

$$\begin{aligned}
\left(\int_{k\Delta}^{(k+1)\Delta} b(\theta, X_{k\Delta}) - b(\theta, X_s)ds\right)^p &= \Delta^p \left(\frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} b(\theta, X_{k\Delta}) - b(\theta, X_s)ds\right)^p \\
&\leq \Delta^p \left(\frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} (b(\theta, X_{k\Delta}) - b(\theta, X_s))^2 ds\right)^{p/2} \\
&\leq \Delta^{p-1} \int_{k\Delta}^{(k+1)\Delta} (b(\theta, X_{k\Delta}) - b(\theta, X_s))^p ds.
\end{aligned}$$

Applying Lemma 7.2, gives

$$\begin{aligned}
\mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta} b(\theta, X_{k\Delta}) - b(\theta, X_s)ds\right)^p\right] &\leq \Delta^{p-1}\mathbb{E}\left[\int_{k\Delta}^{(k+1)\Delta} (b(\theta, X_{k\Delta}) - b(\theta, X_s))^p ds\right] \\
&\leq O\left(\Delta^{3p/2}\right).
\end{aligned}$$

□

Lemma 7.4. *Let $b \in \mathcal{B}$. For all $i \in \mathcal{Y}$,*

$$\mathbb{E}\left[|\pi_b(i) - \overline{\pi}_b(i)|^2\right] \leq C\Delta,$$

where C is a positive constant which depends on T and the constants in Assumption 2.1.

Proof of Lemma 7.4. Let $i \in \mathcal{Y}$. It is sufficient to prove that for any $i_0 \in \mathcal{Y}$,

$$\mathbb{E}_{i_0}\left[|F_b^i - \overline{F}_b^i|^2\right] \leq C_{i_0,i}\Delta, \tag{7.1}$$

for a constant $C_{i_0,i}$ depending only on T and the constants in Assumption 2.1. In order to prove (7.1), one may use the same kind of arguments as in the proof of Theorem 7.11 p. 174 in Graham & Talay (2013) (in the much more difficult context of the Euler scheme). Similar arguments will be detailed on the following proof of Lemma 7.5.

Observe that for $i, j \in \{1, \ldots, K\}$, $|\partial_j \varphi_i| \leq 1$ and consequently φ_i is a Lipschitz function. Hence, using (7.1) simultaneously for all $i_0 \in \mathcal{Y}$ and the definitions of $\pi_b(i) = \varphi_i(F_b)$ and $\overline{\pi}_b(i) = \varphi_i(\overline{F}_b)$, we deduce directly the announced result. □

Lemma 7.5. *Let $b, \tilde{b} \in \mathcal{B}$. For each $i \in \mathcal{Y}$, the following holds*

$$\mathbb{E}\left[|\overline{\pi}_b(i) - \overline{\pi}_{\tilde{b}}(i)|\right] \leq \frac{C\ K}{\sigma_0^2}\|b - \tilde{b}\|_T,$$

where C is a positive constant which depends on T and C_0.

Proof of Lemma 7.5. Since for each $i, j \in \mathcal{Y}$, $|\partial_j \varphi_i| \leq 1$, we have

$$|\overline{\pi}_b(i) - \overline{\pi}_{\tilde{b}}(i)| \quad = \quad |\varphi_i(\overline{F}_b) - \varphi_i(\overline{F}_{\tilde{b}})| \leq \sum_{j=1}^{K} |\overline{F}_b^j - \overline{F}_{\tilde{b}}^j|.$$

Thus, let us look at the following difference for any trajectory X from (2.1) with drift b_Y^*,

$$|\overline{F}_b^i - \overline{F}_{\tilde{b}}^i| \quad = \quad \left|\sum_{k=0}^{n-1} \frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{k\Delta})(X_{(k+1)\Delta} - X_{k\Delta}) - \frac{\Delta}{2}\frac{({b_i}^2 - \tilde{b}_i^2)}{\sigma^2}(X_{k\Delta})\right|. \tag{7.2}$$

First, we use the relation

$$\sum_{k=0}^{n-1} f(X_{k\Delta})(X_{(k+1)\Delta} - X_{k\Delta}) = \int_0^T f(X_{\xi(s)})dX_s,\ \xi(s) = k\Delta,\ k\Delta \leq s < (k+1)\Delta. \tag{7.3}$$

Then we obtain:

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{k=0}^{n-1} \frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{k\Delta})(X_{(k+1)\Delta} - X_{k\Delta})\right|\right] \quad &\leq \quad \mathbb{E}\left[\left|\int_0^T \frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{\xi(s)})b_Y^*(X_{\xi(s)})ds\right|\right] \\ &\quad + \mathbb{E}\left[\left|\int_0^T \frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{\xi(s)})\sigma(X_{\xi(s)})dW_s\right|\right].\end{aligned}$$

The second term of the right hand side is bounded as follows,

$$\begin{aligned}\mathbb{E}\left[\left|\int_0^T \frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{\xi(s)})\sigma(X_{\xi(s)})dW_s\right|\right] \quad &\leq \quad \mathbb{E}\left[\left(\int_0^T \frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{\xi(s)})\sigma(X_{\xi(s)})dW_s\right)^2\right]^{1/2} \\ &= \quad \mathbb{E}\left[\int_0^T \frac{(b_i - \tilde{b}_i)^2}{\sigma^4}(X_{\xi(s)})\sigma^2(X_{\xi(s)})ds\right]^{1/2} \\ &\leq \quad \frac{\sqrt{T}}{\sigma_0}\|b_i - \tilde{b}_i\|_T.\end{aligned}$$

Besides,

$$\mathbb{E}\left[\left|\int_0^T \frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{\xi(s)})b_Y^*(X_{\xi(s)})ds\right|\right] \quad \leq \quad \int_0^T \mathbb{E}\left[\left(\frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{\xi(s)})\right)^2\right]^{1/2} \mathbb{E}\left[(b_Y^*(X_{\xi(s)})^2\right]^{1/2} ds$$

But by Assumption 2.1, $|b_Y^*(x)| = \sum_{i=1}^K |b_i^*(x)| \mathbb{1}_{\{Y=i\}} \leq C_0(1+|x|) \sum_{i=1}^K \mathbb{1}_{\{Y=i\}} = C_0(1+|x|)$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left|\int_0^T \frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{\xi(s)}) b_Y^*(X_{\xi(s)}) ds\right|\right] \\
&\qquad \leq \int_0^T \mathbb{E}\left[\left(\frac{(b_i - \tilde{b}_i)}{\sigma^2}(X_{\xi(s)})\right)^2\right]^{1/2} ds\ \mathbb{E}\left[C_0^2\left(1 + \sup_{s\in[0,T]} X_s\right)^2\right]^{1/2} \\
&\qquad \leq \frac{C_1\, T}{\sigma_0^2}\|b_i - \tilde{b}_i\|_T.
\end{aligned}$$

Let us now deal with the second term of Equation (7.2) with the same arguments,

$$\mathbb{E}\left[\left|\Delta \sum_{k=0}^{n-1} \frac{({b_i}^2 - \tilde{b}_i^2)}{\sigma^2}(X_{k\Delta})\right|\right] \quad \leq \quad \frac{1}{\sigma_0^2}\mathbb{E}\left[\left|\int_0^T (b_i - \tilde{b}_i)(b_i + \tilde{b}_i)(X_{\xi(s)}) ds\right|\right].$$

Then, Cauchy-Schwarz inequality leads to

$$\begin{aligned}
\mathbb{E}\left[\left|\Delta \sum_{k=0}^{n-1} \frac{({b_i}^2 - \tilde{b}_i^2)}{\sigma^2}(X_{k\Delta})\right|\right] \quad &\leq \quad \frac{1}{\sigma_0^2}\int_0^T \mathbb{E}\left[(b_i - \tilde{b}_i)^2(X_{\xi(s)})\right]^{1/2} \mathbb{E}\left[(b_i + \tilde{b}_i)^2(X_{\xi(s)})\right]^{1/2} ds \\
&\leq \quad \frac{C_2 T}{\sigma_0^2}\|b_i - \tilde{b}_i\|_T,
\end{aligned}$$

which concludes the proof. □

7.2. **Proofs of Section 2.**

7.2.1. *Proof of Proposition 2.3.* The probability measure $\mathbb{P}$ can be decomposed in $\mathbb{P} = \sum_{i=1}^K p_i \mathbb{P}_i$, with $\mathbb{P}_i := \mathbb{P}(.|Y = i)$. Denote $\mathbb{P}_0$ the probability measure under which $(X_t)_{t\geq 0}$ is solution of $dX_t = \sigma(X_t)d\widetilde{W}_t$ where $\widetilde{W}$ is a Brownian motion under $\mathbb{P}_0$. Then Girsanov's Theorem (see *e.g.* Jacod & Shiryaev, 2003) implies

$$\Phi_t^i := \frac{d\mathbb{P}_i|_{\mathcal{F}_t^X}}{d\mathbb{P}_0|_{\mathcal{F}_t^X}} = \exp\left(\int_0^t \frac{b_i}{\sigma^2}(X_s) dX_s - \int_0^t \frac{b_i^2}{2\sigma^2}(X_s) ds\right).$$

Thus, for $t \geq 0$, we have that

$$d\mathbb{P}|_{\mathcal{F}_t^X} = \sum_{i=1}^K p_i d\mathbb{P}_i|_{\mathcal{F}_t^X} = \sum_{i=1}^K p_i \Phi_t^i d\mathbb{P}_0|_{\mathcal{F}_t^X}. \tag{7.4}$$

Denote

$$\Psi_t^i := \frac{d\mathbb{P}_i|_{\mathcal{F}_t^X}}{d\mathbb{P}|_{\mathcal{F}_t^X}} = \frac{\Phi_t^i\ d\mathbb{P}_0|_{\mathcal{F}_t^X}}{\sum_{j=1}^K p_j \Phi_t^j\ d\mathbb{P}_0|_{\mathcal{F}_t^X}} = \frac{\Phi_t^i}{\sum_{j=1}^K p_j \Phi_t^j}. \tag{7.5}$$

Moreover, let $h : \{1, \ldots, K\} \to \mathbb{R}$ a bounded function and Z an $\mathcal{F}_t^X$-measurable bounded random variable. We have that

$$\begin{aligned}
\mathbb{E}\left[h(Y)Z\right] &= \mathbb{E}\left[h(Y)\mathbb{E}\left[Z|\sigma(Y)\right]\right] = \mathbb{E}\left[\sum_{i=1}^K h(i)\mathbb{1}_{\{Y=i\}}\mathbb{E}_i\left[Z\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^K h(i)\mathbb{1}_{\{Y=i\}}\mathbb{E}\left[\Psi_t^i Z\right]\right] = \sum_{i=1}^K p_i h(i)\mathbb{E}\left[\Psi_t^i Z\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^K h(i) p_i \Psi_t^i\right) Z\right]
\end{aligned}$$

giving

$$\mathbb{E}\left[h(Y)|\mathcal{F}_t^X\right] = \sum_{i=1}^{K} p_i h(i) \Psi_t^i \quad \mathbb{P}-a.s.$$

Using definition (7.5) we obtain

$$\mathbb{E}\left[h(Y)|\mathcal{F}_t^X\right] = \frac{\sum_{i=1}^{K} h(i) p_i \Phi_t^i}{\sum_{i=1}^{K} p_i \Phi_t^i}.$$

Thus, for $h(x) = \mathbb{1}_{\{i\}}(x)$ we obtain the expected result. □

7.2.2. *Proof of Proposition 2.4.* Let $0 < t \leq T$. For $g_t \in \mathcal{G}_t$, we have

$$\begin{aligned}
R(g_t) - R(g_t^*) &= \mathbb{E}\left[\mathbb{1}_{\{g_t(X)\neq Y\}} - \mathbb{1}_{\{g_t^*(X)\neq Y\}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K} \pi_t^*(i)\left(\mathbb{1}_{\{g_t(X)\neq i\}} - \mathbb{1}_{\{g_t^*(X)\neq i\}}\right)\mathbb{1}_{\{g_t(X)=k\}}\mathbb{1}_{\{g_t^*(X)=j\}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{k\neq i}^{K} \pi_t^*(i)\mathbb{1}_{\{g_t(X)=k\}}\mathbb{1}_{\{g_t^*(X)=i\}} - \sum_{k=1}^{K}\sum_{i\neq k}^{K} \pi_t^*(k)\mathbb{1}_{\{g_t(X)=k\}}\mathbb{1}_{\{g_t^*(X)=i\}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{k\neq i}^{K} (\pi_t^*(i) - \pi_t^*(k))\mathbb{1}_{\{g_t(X)=k\}}\mathbb{1}_{\{g_t^*(X)=i\}}\right]
\end{aligned}$$

and $(\pi_t^*(i) - \pi_t^*(k)) = |\pi_t^*(i) - \pi_t^*(k)|$ on the event $\{g_t^*(X) = i\}$. □

7.3. **Proofs of Section 3.**

7.3.1. *Proof of Proposition 3.1.* According to Proposition 2.4 we get

$$\begin{aligned}
R(\overline{g}_b) - R(g^*) &= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{k\neq i}^{K} |\pi^*(i) - \pi^*(k)|\mathbb{1}_{\{\overline{g}_b(X)=k\}}\mathbb{1}_{\{g^*(X)=i\}}\right] \\
&\leq 2\mathbb{E}\left[\max_{i\in\mathcal{Y}} |\pi^*(i) - \overline{\pi}_b(i)|\mathbb{1}_{\{\overline{g}_b(X)\neq g^*(X)\}}\right] \leq 2\sum_{i=1}^{K}\mathbb{E}\left[|\pi^*(i) - \overline{\pi}_b(i)|\right]
\end{aligned}$$

Since,

$$|\overline{\pi}_b(i) - \pi^*(i)| \leq |\overline{\pi}_b(i) - \overline{\pi}_{b^*}(i)| + |\overline{\pi}_{b^*}(i) - \pi^*(i)|.$$

(with $\overline{\pi}_{b^*}(i) := \varphi_i(\overline{F}_{b^*})$ given by (3.1)), then

$$R(\overline{g}_b) - R(g^*) \leq 2\sum_{i=1}^{K}\mathbb{E}\left[|\overline{\pi}_{b^*}(i) - \pi^*(i)|\right] + 2\sum_{i=1}^{K}\mathbb{E}\left[|\overline{\pi}_b(i) - \overline{\pi}_{b^*}(i)|\right].$$

Lemma 7.4 and Lemma 7.5 lead to the following bound

$$R(\overline{g}_b) - R(g^*) \leq C_1\sqrt{\Delta} + \frac{C_2 K^2}{\sigma_0^2}\|b - b^*\|_T$$

with C_1, C_2 two positive constants depending on T, C_0.
□

7.3.2. *Proof of Proposition 3.5.* Denote $I_j = [j, j+1[$. Following the proof of Van Der Vaart & Wellner (1996) Corollary 2.7.4, we know that there exists $\mathcal{H}_{k,\varepsilon_j}$ an ε_j-net of $\{h_{|I_j}, h \in \mathcal{H}_k\}$ with respect to $\|\cdot\|_\infty$:

$$\mathcal{H}_{k,\varepsilon_j} = \{b_{j,k_j}, k_j = 1, \dots, N_{\varepsilon_j}\}$$

with cardinal N_{ε_j} satisfying

$$\log(N_{\varepsilon_j}) \leq C_k \left(\frac{M_j}{\varepsilon_j}\right)^{1/k}, \quad M_j = C_0(1+|j|)^\gamma.$$

Let $\varepsilon > 0$. Set $\varepsilon_j = \varepsilon \max(1, |j|)^{1+\gamma+k} = \varepsilon_{|j|}$ (with $k \geq 1$ and $\gamma \geq 1$), and

$$\mathcal{H}_\varepsilon := \left\{ \sum_{j\in\mathbb{Z}} b_{j,k_j} \mathbb{1}_{I_j},\ k_j \in \{1, \dots N_{\varepsilon_j}\} \right\}$$

Let $b \in \mathcal{H}_k$, there exist $b_{j,k_j} \in \mathcal{H}_{k,\varepsilon_j}$ such that $\sup_{I_j} |b - b_{j,k_j}| < \varepsilon_j$, then

$$\mathbb{E}\left[\left(b - \sum_{j\in\mathbb{Z}} b_{j,k_j}\mathbb{1}_{I_j}\right)^2 (X_t)\right] \ \leq \ \sum_{j\in\mathbb{Z}} \varepsilon_j^2\, \mathbb{P}(X_t \in [j, j+1[).$$

Besides for $q = 4 + 2\gamma + 2k$, and $j \neq \{-1, 0\}$,

$$\mathbb{P}(X_t \in [j, j+1[) = \mathbb{E}\left[\frac{1}{|X_t|^q}|X_t|^q \mathbb{1}_{\{X_t \in [j,j+1[\}}\right] \leq \frac{\mathbb{E}[\sup_{t\in[0,T]} |X_t|^q]}{|j|^q \wedge |j+1|^q}.$$

Thus as $\varepsilon_0 = \varepsilon_{-1} = \varepsilon$, it comes

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left(b - \sum_{j\in\mathbb{Z}} b_{j,k_j}\mathbb{1}_{I_j}\right)^2 (X_t)\right] \ \leq \ \sum_{j\geq 1} \frac{\varepsilon_j^2\ \mathbb{E}[\sup_{t\in[0,T]} |X_t|^q]}{|j|^q \wedge |j+1|^q} + 2\varepsilon \leq C\varepsilon$$

where C is a positive constant.

Note that there exists j_0 (depending on ε) such that for all $|j| \geq j_0$, $\varepsilon_j > M_j$ and consequently, for all $|j| \geq j_0$, one may take $N_{\varepsilon_j} = 1$. Hence with this choice,

$$\mathrm{Card}\mathcal{H}_\varepsilon = \prod_{j\in\mathbb{Z}} N_{\varepsilon_j},$$

and

$$\log(\mathrm{Card}\mathcal{H}_\varepsilon) \leq C_k \varepsilon^{-1/k} \sum_{j\in\mathbb{Z}} \frac{|j|^{\gamma/k}}{|j|^{(k+\gamma+1)/k}} \leq C\varepsilon^{-1/k}$$

where C is a positive constant (depending on k). □

7.3.3. *Proof of Proposition 3.3.* Consider b^ε such that $\max_i \|b_i^* - b_i^\varepsilon\|_T \leq \varepsilon$. Similarly as in Proof 7.3.1 we obtain:

$$\begin{aligned}
\widehat{\mathbf{E}}\left[R(\overline{g}_{b^\varepsilon}) - R(g^*)\right] &\leq 2\sum_{i=1}^K \mathbb{E}\left[\overline{\pi}_{b^\varepsilon}(i) - \overline{\pi}_{b^*}(i)\right] + 2\sum_{i=1}^K \mathbb{E}\left[\overline{\pi}_{b^*}(i) - \pi^*(i)\right] \\
&\leq C(\sqrt{\Delta} + \varepsilon)
\end{aligned}$$

where C is a positive constant depending on T, C_0, σ_0. Then, it comes, by definition of estimator $\widehat{b}^{\varepsilon}$,

$$\begin{aligned} R(\overline{g}_{\widehat{b}^{\varepsilon}}) - R(g^*) &\leq R(\overline{g}_{\widehat{b}^{\varepsilon}}) - R(\overline{g}_{b^{\varepsilon}}) + C(\Delta + \varepsilon) \\ &\leq R(\overline{g}_{\widehat{b}^{\varepsilon}}) - \widehat{R}(\overline{g}_{\widehat{b}^{\varepsilon}}) + \widehat{R}(\overline{g}_{\widehat{b}^{\varepsilon}}) - R(\overline{g}_{b^{\varepsilon}}) + C(\sqrt{\Delta} + \varepsilon) \\ &\leq R(\overline{g}_{\widehat{b}^{\varepsilon}}) - \widehat{R}(\overline{g}_{\widehat{b}^{\varepsilon}}) + \widehat{R}(\overline{g}_{b^{\varepsilon}}) - R(\overline{g}_{b^{\varepsilon}}) + C(\sqrt{\Delta} + \varepsilon) \\ &\leq 2 \max_{b \in \mathcal{B}_{\varepsilon}} |\widehat{R}(\overline{g}_b) - R(\overline{g}_b)| + C(\sqrt{\Delta} + \varepsilon). \end{aligned}$$

Moreover, according to Hoeffding's inequality it comes

$$\widehat{\mathbf{P}}\left(\max_{b \in \mathcal{B}_{\varepsilon}} |\widehat{R}(\overline{g}_b) - R(\overline{g}_b)| \geq t \right) \leq \min(1, 2\mathrm{Card}\mathcal{B}_{\varepsilon} \exp(-2Nt^2)).$$

Finally, integrating the last equation it comes:

$$\begin{aligned} &\widehat{\mathbf{E}}\left[\max_{b \in \mathcal{B}_{\varepsilon}} |\widehat{R}(\overline{g}_b) - R(\overline{g}_b)| \right] \\ &\qquad \leq \int_0^{\infty} \min(1, \exp(\log(2\mathrm{Card}\mathcal{B}_{\varepsilon})) - 2Nt^2) dt \\ &\qquad \leq \int_0^{\infty} \exp\left(-(2Nt^2 - \log(2\mathrm{Card}\mathcal{B}_{\varepsilon}))_+ \right) dt \\ &\qquad = \sqrt{\frac{\log(2\mathrm{Card}\mathcal{B}_{\varepsilon})}{2N}} + \int_{t \geq \sqrt{\frac{\log(2\mathrm{Card}\mathcal{B}_{\varepsilon})}{2N}}} \exp\left(-(2Nt^2 - \log(2\mathrm{Card}\mathcal{B}_{\varepsilon})) \right) dt \\ &\qquad \leq \sqrt{\frac{\log(2\mathrm{Card}\mathcal{B}_{\varepsilon})}{2N}} + \frac{\sqrt{\pi}}{2\sqrt{2N}} \end{aligned}$$

using $\int_0^{\infty} e^{-u^2} du = \sqrt{\pi}/2$. Gathering the results we obtain

$$\mathbf{E}\left[R(\overline{g}_{\widehat{b}_{\varepsilon}}) - R(g^*) \right] \leq \sqrt{\frac{\log(2\mathrm{Card}\mathcal{B}_{\varepsilon})}{2N}} + \frac{\sqrt{\pi}}{2\sqrt{2N}} + C(\sqrt{\Delta} + \varepsilon) \leq 2\sqrt{\frac{\log(2\mathrm{Card}\mathcal{B}_{\varepsilon})}{2N}} + C(\sqrt{\Delta} + \varepsilon)$$

(since there are at least two elements in $\mathcal{B}_{\varepsilon}$). □

7.4. **Proofs of Section 4.**

7.4.1. *Proof of Theorem 4.3.* Let us remind a generalised version of a key result, from Dacunha-Castelle & Duflo (1983) for example or Guyon (1995).

Lemma 7.6. *Consider* $\widehat{\theta}_{N,n} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, \gamma_{N,n}(\theta)$, *with* $\gamma_{N,n}$ *a contrast process. If*

(1) $\Theta \subset \mathbb{R}$ *compact,* $\theta \in int(\Theta)$
(2) *The contrast function* $\theta \to \gamma_{N,n}(\theta)$ *is continuous* $\widehat{\mathbf{P}}-a.s.$, *and* $\theta \to \gamma(\theta, \theta^*)$ *is a continuous contrast function which has a unique minimum in* θ^*.
(3) $\gamma_{N,n}(\theta) + c(\theta^*) \xrightarrow[N,n]{\widehat{\mathbf{P}}} \gamma(\theta, \theta^*)$.
(4) $\exists \varepsilon_k \to 0,\ \forall k,\ \lim_{N,n} \widehat{\mathbf{P}}\left(\sup_{|\alpha - \beta| \leq k^{-1}} |\gamma_{N,n}(\alpha) - \gamma_{N,n}(\beta)| \geq \varepsilon_k \right) = 0$

then

$$\widehat{\theta}_{N,n} \xrightarrow{\widehat{\mathbf{P}}} \theta^*.$$

Proof. Denote

$$\omega_{N,n}(1/k) = \sup_{|\theta - \theta'| \leq k^{-1}} \{ |\gamma_{N,n}(\theta') - \gamma_{N,n}(\theta)| \}.$$

Let $a > 0$. The function γ satisfies $\gamma(\theta^*, \theta^*) = 0$ and by continuity, there exists $\varepsilon > 0$ such that on $\Theta \setminus I$ the function $\gamma(\theta^*, \cdot)$ is bounded from below by 2ε where $I = [\theta^* - a, \theta^* + a]$.

Let k large enough such that $\varepsilon_k < \varepsilon$. Since Θ is a compact set, we may consider a finite recovering of $\Theta \setminus I$ by M_k intervals $I_i = [\theta_i - k^{-1}, \theta_i + k^{-1}]$. When $\theta \in I_i$:

$$\gamma_{N,n}(\theta) \geq \gamma_{N,n}(\theta_i) - |\gamma_{N,n}(\theta_i) - \gamma_{N,n}(\theta)|,$$
$$\inf_{\theta \in \Theta \setminus I} \gamma_{N,n}(\theta) \geq \inf_{1 \leq i \leq M_k} \gamma_{N,n}(\theta_i) - w_{N,n}(1/k)$$

and thus

$$\begin{aligned}\{\widehat{\theta}_{N,n} \notin [\theta^* - a, \theta^* + a]\} &\subseteq \left\{ \min_{\theta \in \Theta \setminus I} \gamma_{N,n}(\theta) < \gamma_{N,n}(\theta^*) \right\} \\ &\subseteq \left\{ \inf_{1 \leq i \leq M_k} \gamma_{N,n}(\theta_i) - \gamma_{N,n}(\theta^*) - \omega_{N,n}(1/k) < 0 \right\}.\end{aligned}$$

Finally, for all $\eta > 0$,

$$\widehat{\mathbf{P}}(\widehat{\theta}_{N,n} \notin I) \leq \widehat{\mathbf{P}}(w_{N,n}(1/k) > \eta) + \widehat{\mathbf{P}}\left(\inf_{1 \leq i \leq M_k} (\gamma_{N,n}(\theta_i) - \gamma_{N,n}(\theta^*)) \leq \eta \right).$$

But we know that the first term of the right hand side goes to 0, and as previously

$$\inf_{1 \leq i \leq M_k} (\gamma_{N,n}(\theta_i) - \gamma_{N,n}(\theta^*)) \underset{N,n \to \infty}{\overset{\widehat{\mathbf{P}}}{\to}} \gamma(\theta^*, \theta_i) \geq 2\varepsilon$$

thus, choosing $\eta = \varepsilon$ in the previous

$$\widehat{\mathbf{P}}(|\widehat{\theta}_{N,n} - \theta^*| \geq a) \underset{N,n \to \infty}{\to} 0.$$

□

In the following we consider that we are in the class number i estimating θ_i^* from the N_i trajectories, and the index i is omitted (and $\widehat{\mathbf{P}}, \widehat{\mathbf{E}}$ refer to the class i of the learning sample).

Let us remind that

$$\gamma_{N,n}(\theta) := \frac{1}{N} \sum_{j=1}^{N} \sum_{k=0}^{n-1} \frac{\Delta}{2} b^2(\theta, X_{k\Delta}^{(j)}) - b(\theta, X_{k\Delta}^{(j)})(X_{(k+1)\Delta}^{(j)} - X_{k\Delta}^{(j)}).$$

For $\theta \in \Theta$ we denote

$$\gamma_N(\theta) := \frac{1}{N} \sum_{j=1}^{N} \left\{ \frac{1}{2} \int_0^T b^2(\theta, X_s^{(j)}) ds - \int_0^T b(\theta, X_s^{(j)}) dX_s^{(j)} \right\}$$

thus

$$\gamma_{N,n}(\theta) = \gamma_N(\theta) + R_{1,N,n} - R_{2,N,n} \tag{7.6}$$

with

$$\begin{aligned} R_{1,N,n} &:= \frac{1}{N} \sum_{j=1}^{N} \left\{ \int_0^T b(\theta, X_s^{(j)}) dX_s^{(j)} - \left(\sum_{k=0}^{n-1} b(\theta, X_{k\Delta}^{(j)})(X_{(k+1)\Delta}^{(j)} - X_{k\Delta}^{(j)}) \right) \right\} =: \frac{1}{N} \sum_{j=1}^{N} R_{1,n}^{(j)} \\ R_{2,N,n} &:= \frac{1}{N} \sum_{j=1}^{N} \left\{ \frac{1}{2} \int_0^T b^2(\theta, X_s^{(j)}) ds - \left(\frac{1}{2} \sum_{k=0}^{n-1} \Delta b^2(\theta, X_{k\Delta}^{(j)}) \right) \right\} =: \frac{1}{2N} \sum_{j=1}^{N} R_{2,n}^{(j)}. \end{aligned}$$

Moreover,

$$\gamma_N(\theta) = \frac{1}{N} \sum_{j=1}^{N} \left\{ \frac{1}{2} \int_0^T b^2(\theta, X_s^{(j)}) ds - \int_0^T b(\theta, X_s^{(j)}) [b(\theta^*, X_s^{(j)}) ds + dW_s^{(j)}] \right\}$$

and under our assumptions we have

$$\widehat{\mathbf{E}}\left[\left|\int_0^T b(\theta, X_s^{(j)})b(\theta^*, X_s^{(j)})ds\right|\right] < \infty \quad \widehat{\mathbf{E}}\left[\left|\int_0^T b^2(\theta, X_s^{(j)})ds\right|\right] < \infty,$$

thus the Law of Large Number implies that

$$\gamma_N(\theta) \underset{N\to\infty}{\overset{\widehat{\mathbf{P}}}{\to}} \frac{1}{2}\widehat{\mathbf{E}}\left[\int_0^T b^2(\theta, X_s)ds\right] - \widehat{\mathbf{E}}\left[\int_0^T b(\theta, X_s)b(\theta^*, X_s)ds\right] =: \Gamma(\theta, \theta^*). \tag{7.7}$$

Now,

$$\begin{aligned}
&R_{1,n}^{(j)} \\
&= \int_0^T b(\theta, X_s^{(j)})dX_s^{(j)} - \sum_{k=0}^{n-1} b(\theta, X_{k\Delta}^{(j)})(X_{(k+1)\Delta}^{(j)} - X_{k\Delta}^{(j)}) \\
&= \int_0^T b(\theta, X_s^{(j)})\left(b(\theta^*, X_s^{(j)})ds + dW_s^{(j)}\right) - \sum_{k=0}^{n-1} b(\theta, X_{k\Delta}^{(j)})\left[\int_{k\Delta}^{(k+1)\Delta} b(\theta^*, X_s^{(j)})ds + W_{(k+1)\Delta}^{(j)} - W_{k\Delta}^{(j)}\right] \\
&=: \int_0^T H_{s,n}^{(j)} dW_s^{(j)} + \rho_n^{(j)}
\end{aligned}$$

with

$$H_{s,n}^{(j)} := \sum_{k=0}^{n-1} \mathbb{1}_{]k\Delta,(k+1)\Delta]}(s)(b(\theta, X_{k\Delta}^{(j)}) - b(\theta, X_s^{(j)})).$$

The first term of the right hand side is

$$\begin{aligned}
\rho_n^{(j)} &= \sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta} b(\theta, X_{k\Delta}^{(j)})b(\theta^*, X_s^{(j)})ds - \int_0^T b(\theta, X_s^{(j)})b(\theta^*, X_s^{(j)})ds \\
&= \sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta} b(\theta^*, X_s^{(j)})(b(\theta, X_{k\Delta}^{(j)}) - b(\theta, X_s^{(j)}))ds
\end{aligned}$$

then according to the assumptions on b we get

$$\begin{aligned}
\widehat{\mathbf{E}}[|\rho_n^{(j)}|] &\leq \widehat{\mathbf{E}}\left[\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta} C\left(1 + \sup_{s\in[0,T]}|X_s^{(j)}|\right)|X_{k\Delta}^{(j)} - X_s^{(j)}|ds\right] \\
&\leq \widehat{\mathbf{E}}\left[C(1 + \sup_{s\in[0,T]}|X_s^{(j)}|)^2\right]^{1/2} \widehat{\mathbf{E}}\left[\left(\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta}|X_{k\Delta}^{(j)} - X_s^{(j)}|ds\right)^2\right]^{1/2}
\end{aligned}$$

with

$$\begin{aligned}
\widehat{\mathbf{E}}\left[\left(\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta}|X_{k\Delta}^{(j)} - X_s^{(j)}|ds\right)^2\right] &\leq \widehat{\mathbf{E}}\left[\left(\frac{1}{T}\int_0^T |X_{\xi(s)}^{(j)} - X_s^{(j)}|ds\right)^2\right] \\
&\leq T\widehat{\mathbf{E}}\left[\frac{1}{T}\int_0^T |X_{\xi(s)}^{(j)} - X_s^{(j)}|^2 ds\right] \leq C_T\Delta
\end{aligned}$$

with ξ is defined by Equation (7.3). We have obtained

$$\widehat{\mathbf{E}}|\rho_n^{(j)}| \leq C_T\sqrt{\Delta}.$$

(where the constants C can change from a line to another). Then similarly, the martingale term is controlled

$$\widehat{\mathbf{E}}\left[\left|\int_0^T H_{s,n}^{(j)} dW_s^{(j)}\right|\right] \leq \left(\widehat{\mathbf{E}}\left[\int_0^T \left(H_{s,n}^{(j)}\right)^2 ds\right]\right)^{1/2} \leq C_T\sqrt{\Delta}.$$

Then, $R_{2,n}^{(j)} = \sum_{k=0}^{n-1} \int_{k\Delta}^{(k+1)\Delta} \left(b^2(\theta, X_{k\Delta}^{(j)}) - b^2(\theta, X_s^{(j)})\right) ds$ and we obtain

$$\widehat{\mathbf{E}}\left[\left|R_{2,n}^{(j)}\right|\right] \leq \sum_{k=0}^{n-1} \int_{k\Delta}^{(k+1)\Delta} C\sqrt{\Delta} = C_T\sqrt{\Delta}$$

using assumptions on b, which do not depend on N. Thus, we have shown that

$$\begin{aligned}
\gamma_{N,n}(\theta) + \frac{1}{2}\mathbb{E}_{\theta^*}\left[\int_0^T b^2(\theta^*, X_s) ds\right] &\underset{N\to\infty, n\to\infty}{\to} \Gamma(\theta, \theta^*) + \frac{1}{2}\mathbb{E}_{\theta^*}\left[\int_0^T b^2(\theta^*, X_s) ds\right] \\
&= \frac{1}{2}\mathbb{E}_{\theta^*}\left[\int_0^T (b(\theta^*, X_s) - b(\theta, X_s))^2 ds\right] \\
&=: \gamma(\theta, \theta^*). \qquad (7.8)
\end{aligned}$$

To end the proof let us look at the following random variable

$$\omega_{N,n}(\varepsilon) = \sup_{|\alpha-\beta|<\varepsilon} \{|\gamma_{N,n}(\alpha) - \gamma_{N,n}(\beta)|\}.$$

Let $\eta > 0$, we have from the triangular inequality

$$\{\omega_{N,n}(\varepsilon) > \eta\} \subseteq \{\sup_{\alpha\in\Theta} |\gamma_{N,n}(\alpha)| > \eta/4\} \cup \{\sup_{|\alpha-\beta|<\varepsilon} |\gamma(\alpha, \theta^*) - \gamma(\beta, \theta^*)| > \eta/2\}$$

thus, it is enough to study the convergence in probability of the random variable $\sup_{\alpha\in\Theta} |\gamma_{N,n}(\alpha)|$ in order to verify the remaining condition 4) of Lemma 7.6 since $\theta \to \gamma(\theta, \theta^*)$ is continuous on Θ.

The verification of the remaining condition 4) will follow from the application of Lemma 3.1. in Yoshida (1990) (with $k = 1$, $p = \ell = 2$ and $T = \{N, n\}$) considering

$$f_{N,n}(\theta) := \gamma_{N,n}(\theta) + c(\theta^*) - \gamma(\theta, \theta^*).$$

Indeed, using assumption (4.2) on function b and standard arguments as previously, one can check that $f_{N,n}$ fulfills the assumptions of this lemma, namely :

- $\widehat{\mathbf{E}}[(f_{N,n}(\alpha) - f_{N,n}(\beta))^2] \leq C|\alpha - \beta|^2$
- $\widehat{\mathbf{E}}[f_{N,n}(\theta)^2] \leq C$
- $f_{N,n}(\theta) \overset{\widehat{\mathbf{P}}_{\theta^*}}{\to} 0$

Using the uniform continuity of $\gamma(\cdot, \theta^*)$, we thus verify the condition 4) of Lemma 7.6 by taking a sequence $\varepsilon_k \to 0$ such that $\{\sup_{|\alpha-\beta|<k^{-1}} |\gamma(\alpha, \theta^*) - \gamma(\beta, \theta^*)| > \varepsilon_k\} = \emptyset$. The statement of Theorem 4.3 follows then by application of Lemma 7.6. □

7.4.2. *Proof of Theorem 4.4.* We follow the scheme of proof of Dacunha-Castelle & Duflo (1983) Chapter 3.
Here the derivative are the derivative according to θ^*.

$$\dot{\gamma}_{N,n}(\widehat{\theta}) = 0 = \dot{\gamma}_{N,n}(\theta^*) + (\widehat{\theta} - \theta^*)\ddot{\gamma}_{N,n}(\widetilde{\theta})$$

with $\widetilde{\theta}$ some point between $\widehat{\theta}$ and θ^*. This yields to

$$\sqrt{N}(\widehat{\theta} - \theta^*) = -\frac{\dot{\gamma}_{N,n}(\theta^*)}{\ddot{\gamma}_{N,n}(\widetilde{\theta})}\sqrt{N}. \qquad (7.9)$$

Let us study first the denominator. We have

$$\sqrt{N}\dot{\gamma}_{N,n}(\theta^*) = \sqrt{N}\dot{\gamma}_N(\theta^*) + \sqrt{N}\dot{R}_{1,N,n} + \sqrt{N}\dot{R}_{2,N,n}$$

where $R_{1,N,n}, R_{2,N,n}$ are given by (7.6). Precisely,

$$\dot{\gamma}_N(\theta^*) = \frac{1}{N}\sum_{j=1}^{N}\int_0^T \dot{b}(\theta^*, X_s^{(j)})\left[b(\theta^*, X_s^{(j)})ds - dX_s^{(j)}\right] = -\frac{1}{N}\sum_{j=1}^{N}\int_0^T \dot{b}(\theta^*, X_s^{(j)})dW_s^{(j)}$$

then we obtain

$$-\sqrt{N}\dot{\gamma}_N(\theta^*) \underset{N\to\infty}{\to} \mathcal{N}\left(0, \mathrm{Var}\left(\int_0^T \dot{b}(\theta^*, X_s)dW_s\right)\right) = \mathcal{N}\left(0, \mathbb{E}\left[\int_0^T \dot{b}^2(\theta^*, X_s)ds\right]\right) \tag{7.10}$$

where the convergence in distribution takes places under $\widehat{\mathbf{P}}$. Moreover, as previously, (see proof Section 7.4.1),

$$\widehat{\mathbf{E}}[\dot{R}_{1,N,n}] = \frac{1}{N}\sum_{j=1}^{N}\widehat{\mathbf{E}}\left(\left[\sum_{k=0}^{n-1}\dot{b}(\theta^*, X_{k\Delta}^{(j)})(X_{(k+1)\Delta}^{(j)} - X_{k\Delta}^{(j)})\right] - \int_0^T \dot{b}(\theta^*, X_s^{(j)})dX_s^{(j)}\right) \le C\sqrt{\Delta}$$

and the same for $\dot{R}_{2,N,n}$, thus $\sqrt{N}\widehat{\mathbf{E}}(\dot{R}_{1,N,n} + \dot{R}_{2,N,n})$ goes to zero if $N\Delta \to 0$ as soon as $x \mapsto \dot{b}(\theta^*, x)$ is Lipschitz.

Finally, let us study $\ddot{\gamma}_{N,n}(\widetilde{\theta})$ given by

$$\ddot{\gamma}_{N,n}(\widetilde{\theta}) = \ddot{\gamma}_N(\widetilde{\theta}) + \ddot{R}_{1,N,n}(\widetilde{\theta}) + \ddot{R}_{2,N,n}(\widetilde{\theta})$$

with

$$\ddot{\gamma}_N(\widetilde{\theta}) = -\frac{1}{N}\sum_{j=1}^{N}\int_0^T \ddot{b}(\widetilde{\theta}, X_s^{(j)})dW_s^{(j)} + \frac{1}{N}\sum_{j=1}^{N}\int_0^T \dot{b}^2(\widetilde{\theta}, X_s^{(j)})ds.$$

According to the Lipschitz-assumption of function $\dot{b}(\cdot, x)$, have :

$$\begin{aligned}\frac{1}{N}\sum_{j=1}^{N}\int_0^T \dot{b}^2(\widetilde{\theta}, X_s^{(j)})ds &= \frac{1}{N}\sum_{j=1}^{N}\int_0^T \dot{b}^2(\widetilde{\theta}, X_s^{(j)}) - \dot{b}^2(\theta^*, X_s^{(j)})ds + \frac{1}{N}\sum_{j=1}^{N}\int_0^T \dot{b}^2(\theta^*, X_s^{(j)})ds \\ &\le C\frac{1}{N}\sum_{j=1}^{N}\int_0^T (1 + |X_s^{(j)}|^\beta)ds\ |\widehat{\theta} - \theta^*| + \frac{1}{N}\sum_{j=1}^{N}\int_0^T \dot{b}^2(\theta^*, X_s^{(j)})ds\end{aligned}$$

(where C depends on the Lipschitz constant of $\theta \mapsto \dot{b}(\theta, \cdot)$: C''). Then as $\widehat{\mathbf{E}}[|\widehat{\theta} - \theta^*|] \to 0$ (from Theorem 4.3) and as the expectation of the second term goes to $\mathbb{E}\left[\int_0^T \dot{b}^2(\theta^*, X_s)ds\right]$, then we have

$$\widehat{\mathbf{E}}[\ddot{\gamma}_N(\widetilde{\theta})] \to \mathbb{E}\left[\int_0^T \dot{b}^2(\theta^*, X_s)ds\right].$$

This proves the attended result. □

7.4.3. *Proof of Proposition 4.8.* In order to prove the consistency of the procedure we establish the following results. In the sequel, we denote

$$h^* := h_{\theta^*}.$$

The following notations are recalled:

$$h_{\theta^*} = (h_{\theta^*}^1, \ldots, h_{\theta^*}^K),\ h_{\theta^*}^i(X) = \frac{1}{K}\sum_{\ell=1}^{K}\log\left(\frac{1 - \pi_{b_{\theta^*}}(\ell)}{1 - \pi_{b_{\theta^*}}(i)}\right),$$

also

$$h_\theta^{\varepsilon,i}(X) = \frac{1}{K}\sum_{\ell=1}^{K}\log\left(\frac{1-\pi_{b_\theta}^{\varepsilon}(\ell)}{1-\pi_{b_\theta}^{\varepsilon}(i)}\right),$$

with

$$\pi_{b_\theta}^{\varepsilon}(i) := \pi_{b_\theta}(i)\mathbb{1}_{\{\pi_{b_\theta}(i_0)<1-\varepsilon\}} + \mathbb{1}_{\{\pi_{b_\theta}(i_0)\geq 1-\varepsilon\}}\left(\left(\pi_{b_\theta}(i) + \frac{\pi_{b_\theta}(i_0)-(1-\varepsilon)}{K-1}\right)\mathbb{1}_{\{i\neq i_0\}} + (1-\varepsilon)\mathbb{1}_{\{i=i_0\}}\right).$$

Lemma 7.7. *For $0<\varepsilon<\frac{1}{3}$, and let $i_0 = \operatorname*{argmax}_{i\in\mathcal{Y}} \pi_{\theta^*}(i)$ one has*

$$R_\phi(h_{\theta^*}^{\varepsilon}) - R_\phi(h_{\theta^*}) \leq \frac{3}{2}K\mathbb{P}\left(\pi_{\theta^*}(i_0)\geq 1-\varepsilon\right).$$

Proof of Lemma 7.7. By definition of the ϕ-risk R_ϕ given in Equation (4.4), we have

$$\begin{aligned}
R_\phi(h_{\theta^*}^{\varepsilon}) - R_\phi(h_{\theta^*}) &= \sum_{i=1}^{K}\mathbb{E}\left[(1-\pi_{\theta^*}(i))\left(\exp(h_{\theta^*}^{\varepsilon,i}(X)) - \exp(h_{\theta^*}^{i}(X))\right)\right] \\
&= \sum_{i=1}^{K}\mathbb{E}\left[(1-\pi_{\theta^*}(i))\left(\exp(h_{\theta^*}^{\varepsilon,i}(X)) - \exp(h_{\theta^*}^{i}(X))\right)\mathbb{1}_{\{\pi_{\theta^*}(i_0)\geq 1-\varepsilon\}}\right] \\
&\leq \sum_{i=1}^{K}\mathbb{E}\left[(1-\pi_{\theta^*}(i))\exp(h_{\theta^*}^{\varepsilon,i}(X))\mathbb{1}_{\{\pi_{\theta^*}(i_0)\geq 1-\varepsilon\}}\right].
\end{aligned}$$

On the set $\{\pi_{\theta^*}(i_0)\geq 1-\varepsilon\}$, by definition we have $\pi_{\theta^*}^{\varepsilon}(i_0) = 1-\varepsilon$ then $\pi_{\theta^*}^{\varepsilon}(i)\leq\varepsilon$ for $i\in\mathcal{Y}$. Then, we have

$$h_{\theta^*}^{\varepsilon,i_0}(X) = \frac{1}{K}\sum_{i\in\mathcal{Y}}\log\left(\frac{1-\pi_{\theta^*}^{\varepsilon}(i)}{\varepsilon}\right)\leq\log\left(\frac{1}{\varepsilon}\right).$$

Then, for $i\neq i_0$, $\pi_{\theta^*}^{\varepsilon}(i)\leq\varepsilon$ and for $\ell\neq i$, $1-\pi_{\theta^*}^{\varepsilon}(\ell)\leq 0$, thus

$$h_{\theta^*}^{\varepsilon,i}(X)\leq\log\left(\frac{1}{1-\pi_{\theta^*}^{\varepsilon}(i)}\right)\leq\log\left(\frac{1}{1-\varepsilon}\right).$$

Hence, we get

$$(1-\pi_{\theta^*}(i_0))\exp(h_{\theta^*}^{\varepsilon,i_0}(X))\mathbb{1}_{\{\pi_{\theta^*}(i_0)\geq 1-\varepsilon\}}\leq 1$$

and for $i\neq i_0$

$$(1-\pi_{\theta^*}(i))\exp(h_{\theta^*}^{\varepsilon,i}(X))\mathbb{1}_{\{\pi_{\theta^*}(i_0)\geq 1-\varepsilon\}}\leq 3/2.$$

Therefore combined the previous inequalities, we obtain that

$$R_\phi(h_{\theta^*}^{\varepsilon}) - R_\phi(h_{\theta^*}) \leq \frac{3}{2}K\mathbb{P}\left(\pi_{\theta^*}(i_0)\geq 1-\varepsilon\right).$$

□

Moreover, the following holds for the empirical score function given in Equation (4.6).

Lemma 7.8. *Let $0<\varepsilon<\frac{1}{3}$. The following holds*

$$|R_\phi(\overline{h}_{\theta^*}^{\varepsilon}) - R_\phi(h_{\theta^*}^{\varepsilon})| \leq \frac{CK\sqrt{\Delta}}{\varepsilon^2}.$$

where C is a positive constant depending on T and C_0.

Proof of Lemma 7.8. From the definition of $\overline{h}_\theta^\varepsilon$ given by Equation (4.6) we have for $i \in \mathcal{Y}$, $|h_\theta^{\varepsilon,i}(X)| \leq \log\left(\frac{1}{\varepsilon}\right)$, $|\overline{h}_\theta^{\varepsilon,i}(\overline{X})| \leq \log\left(\frac{1}{\varepsilon}\right)$. Let us study the difference:

$$|R_\phi(\overline{h}_{\theta^*}^\varepsilon) - R_\phi(h_{\theta^*}^\varepsilon)| = \mathbb{E}\left[\sum_{i=1}^K \mathbb{1}_{\{Y\neq i\}}\left(\phi(-\overline{h}_\theta^{\varepsilon,i}(\overline{X})) - \phi(-h_\theta^{\varepsilon,i}(X))\right)\right]$$

The function ϕ is $L-$Lipschitz with constant $L \leq \frac{1}{\varepsilon}$, then it comes

$$|R_\phi(\overline{h}_{\theta^*}^\varepsilon) - R_\phi(h_{\theta^*}^\varepsilon)| \leq \frac{1}{\varepsilon}\mathbb{E}\left[\sum_{i=1}^K \mathbb{1}_{\{Y\neq i\}}\left|\overline{h}_\theta^{\varepsilon,i}(\overline{X}) - h_\theta^{\varepsilon,i}(X)\right|\right],$$

and then the Mean Value Theorem ensures that

$$|R_\phi(\overline{h}_{\theta^*}^\varepsilon) - R_\phi(h_{\theta^*}^\varepsilon)| \leq \frac{2}{\varepsilon^2}\sum_{i=1}^K \mathbb{E}|\overline{\pi}_{b_{\theta^*}}^\varepsilon(i) - \pi_{b_{\theta^*}}^\varepsilon(i)|.$$

Now denote $i_0 = \underset{i\in\mathcal{Y}}{\operatorname{argmax}}\ \pi_{\theta^*}(i)$ and $i_0^{'} = \underset{i\in\mathcal{Y}}{\operatorname{argmax}}\ \overline{\pi}_{\theta^*}(i)$. We observe that if $\overline{\pi}_{b_{\theta^*}}^\varepsilon(i_0^{'}) < 1-\varepsilon$ and $\pi_{b_{\theta^*}}^\varepsilon(i_0) \geq 1-\varepsilon$ then:

$$|\overline{\pi}_{b_{\theta^*}}^\varepsilon(i_0) - \pi_{b_{\theta^*}}^\varepsilon(i_0)| \leq |\overline{\pi}_{b_{\theta^*}}(i_0) - (1-\varepsilon)| \leq |\overline{\pi}_{b_{\theta^*}}(i_0) - \pi_{b_{\theta^*}}(i_0)|$$

and if $i \neq i_0$

$$\begin{aligned}
|\overline{\pi}_{b_{\theta^*}}^\varepsilon(i) - \pi_{b_{\theta^*}}^\varepsilon(i)| &\leq \left|\overline{\pi}_{b_{\theta^*}}(i) - \left(\pi_{b_{\theta^*}}(i) + \frac{\pi_{b_{\theta^*}}(i_0) - (1-\varepsilon)}{K-1}\right)\right| \\
&\leq |\overline{\pi}_{b_{\theta^*}}(i) - \pi_{b_{\theta^*}}(i)| + \frac{1}{K-1}|\pi_{b_{\theta^*}}(i_0) - (1-\varepsilon)| \\
&\leq |\overline{\pi}_{b_{\theta^*}}(i) - \pi_{b_{\theta^*}}(i)| + \frac{1}{K-1}|\overline{\pi}_{b_{\theta^*}}(i_0) - \pi_{b_{\theta^*}}(i_0)|.
\end{aligned}$$

Using similar arguments, we obtain the same bound in the other configurations. Therefore, we get

$$\sum_{i=1}^K \mathbb{E}[|\overline{\pi}_{b_{\theta^*}}^\varepsilon(i) - \pi_{b_{\theta^*}}^\varepsilon(i)|] \leq 3\sum_{i=1}^K \mathbb{E}[|\overline{\pi}_{b_{\theta^*}}(i) - \pi_{b_{\theta^*}}(i)|],$$

and thus

$$|R_\phi(\overline{h}_{\theta^*}^\varepsilon) - R_\phi(h_{\theta^*}^\varepsilon)| \leq \frac{6}{\varepsilon^2}\sum_{i=1}^K \mathbb{E}|\overline{\pi}_{b_{\theta^*}}(i) - \pi_{b_{\theta^*}}(i)|. \tag{7.11}$$

Hence, applying Lemma 7.4, in Inequality (7.11), we obtain the desired result. □

Now, we establish the consistency of our procedure. Denote by $\mathcal{B}_N$ the $\frac{1}{N^{1+(\alpha/2)}}$-net of Θ^K w.r.t the L_∞ norm $\|.\|$. Since $\Theta = [0,1]^d$, we have $\text{Card}\mathcal{B}_N \leq CN^{Kd(1+(\alpha/2))}$.

Let $\tilde{\theta} = \underset{\theta\in\Theta^K}{\operatorname{argmin}}\ R_\phi\left(\overline{h}_\theta^\varepsilon\right)$. From the definition of $\tilde{\theta}$, $R_\phi(\overline{h}_{\tilde{\theta}}^\varepsilon) - R_\phi(\overline{h}_{\theta^*}^\varepsilon) < 0$, then

$$R_\phi(\overline{h}_{\widehat{\theta}}^\varepsilon) - R_\phi(h_{\theta^*}) \leq R_\phi(\overline{h}_{\widehat{\theta}}^\varepsilon) - R_\phi(\overline{h}_{\tilde{\theta}}^\varepsilon) + R_\phi(\overline{h}_{\theta^*}^\varepsilon) - R_\phi(h_{\theta^*}^\varepsilon) + R_\phi(h_{\theta^*}^\varepsilon) - R_\phi(h_{\theta^*}).$$

Applying Lemma 7.7 and 7.8 leads to

$$R_\phi(\overline{h}_{\theta^*}^\varepsilon) - R_\phi(h_{\theta^*}^\varepsilon) + R_\phi(h_{\theta^*}^\varepsilon) - R_\phi(h_{\theta^*}) \leq \frac{CK\sqrt{\Delta}}{\varepsilon^2} + \frac{3}{2}\mathbb{P}\left(\pi_{\theta^*}(i_0) \geq 1-\varepsilon\right).$$

Therefore, it remains to control $R_\phi(\overline{h}_{\widehat{\theta}}^\varepsilon) - R_\phi(\overline{h}_{\tilde{\theta}}^\varepsilon)$. Let $\theta_N \in \mathcal{B}_N$ such that $\|\widehat{\theta} - \theta_N\| \leq \frac{1}{N^{1+(\alpha/2)}}$. Denote

$$D_\theta := R_\phi(\overline{h}_\theta^\varepsilon) - R_\phi(\overline{h}_{\tilde{\theta}}^\varepsilon), \quad \widehat{D}_\theta := \widehat{R}_\phi(\overline{h}_\theta^\varepsilon) - \widehat{R}_\phi(\overline{h}_{\tilde{\theta}}^\varepsilon).$$

$$0 \leq R_\phi(\overline{h}^\varepsilon_{\widehat{\theta}}) - R_\phi(\overline{h}^\varepsilon_{\widetilde{\theta}}) \leq \max_{\theta\in\mathcal{B}_N} |D_\theta - \widehat{D}_\theta| + |R_\phi(\overline{h}^\varepsilon_{\widehat{\theta}}) - R_\phi(\overline{h}^\varepsilon_{\theta_N})| + |\widehat{R}_\phi(\overline{h}^\varepsilon_{\widehat{\theta}}) - \widehat{R}_\phi(\overline{h}^\varepsilon_{\theta_N})|. \tag{7.12}$$

Now, we bound each term of the previous inequality. Since, for each $i \in \mathcal{Y}$, $|\overline{h}^{\varepsilon,i}_\theta(X)| \leq \log\left(\frac{1}{\varepsilon}\right)$ and since the function ϕ is $L-$Lipschitz with constant $L \leq \frac{1}{\varepsilon}$, we have that for each $j \in \{1,\dots,N\}$ and $\theta \in \mathcal{B}_N$

$$\left|\sum_{i=1}^K \mathbb{1}_{\{Y^{(j)}\neq i\}}\phi\left(-\overline{h}^{\varepsilon,i}_\theta\left(\overline{X}^{(j)}\right)\right) - \sum_{i=1}^K \mathbb{1}_{\{Y^{(j)}\neq i\}}\phi\left(-\overline{h}^{\varepsilon,i}_{\widetilde{\theta}}\left(\overline{X}^{(j)}\right)\right)\right| \leq \frac{2K\log(1/\varepsilon)}{\varepsilon}.$$

Therefore, from Hoeffding's inequality, we deduce (as in Section 7.3.3), the following bound

$$\widehat{\mathbf{E}}\left[\max_{\theta\in\mathcal{B}_N}|D_\theta - \widehat{D}_\theta|\right] \leq 8K\sqrt{\frac{\log(\mathrm{Card}\mathcal{B}_N)}{\varepsilon^2 N}}\log\left(\frac{1}{\varepsilon}\right). \tag{7.13}$$

Let us study the term $|R_\phi(\overline{h}^\varepsilon_{\widehat{\theta}}) - R_\phi(\overline{h}^\varepsilon_{\theta_N})|$. Using same arguments as in Lemma 7.8 (see Equation (7.11)), we have

$$\widehat{\mathbf{E}}\left[|R_\phi(\overline{h}^\varepsilon_{\widehat{\theta}}) - R_\phi(\overline{h}^\varepsilon_{\theta_N})|\right] \leq \frac{6}{\varepsilon^2}\sum_{i=1}^K \mathbf{E}\left[|\overline{\pi}_{b_{\widehat{\theta}}}(i) - \overline{\pi}_{b_{\theta_N}}(i)|\right].$$

From Lemma 7.5 and by definition of θ_N, we obtain

$$\widehat{\mathbf{E}}\left[|R_\phi(\overline{h}^\varepsilon_{\widehat{\theta}}) - R_\phi(\overline{h}^\varepsilon_{\theta_N})|\right] \leq \frac{CK^2}{\varepsilon^2 N^{1+(\alpha/2)}}, \tag{7.14}$$

where C is a positive constant depending on T, σ_0, and C_0.

To conclude the proof it remains to control the term $|\widehat{R}_\phi(\overline{h}^\varepsilon_{\widehat{\theta}}) - \widehat{R}_\phi(\overline{h}^\varepsilon_{\theta_N})|$. The empirical risk $\widehat{R}_\phi$ and $\widehat{\theta}$ depend on $\mathcal{D}_N$ then this control requires a different approach. Let $j \in \{1,\dots,N\}$, we denote by $\widehat{\mathbf{E}}_j$ the expectation integrated with respect to $(\overline{X}^{(j)}, Y^{(j)})$. Conditional on $(X^{(1)},Y^{(1)}),\dots,(X^{(j-1)},Y^{(j-1)})$, $(X^{(j+1)},Y^{(j+1)}),\dots,(X^{(N)},Y^{(N)})$, using similar arguments as in Lemma 7.8, we get

$$\widehat{\mathbf{E}}_j\left[\left|\sum_{i=1}^K \mathbb{1}_{\{Y^{(j)}\neq i\}}\phi\left(-\overline{h}^{\varepsilon,i}_{\widehat{\theta}}\left(\overline{X}^{(j)}\right)\right) - \sum_{i=1}^K \mathbb{1}_{\{Y^{(j)}\neq i\}}\phi\left(-\overline{h}^{\varepsilon,i}_{\theta_N}\left(\overline{X}^{(j)}\right)\right)\right|\right] \leq \frac{6}{\varepsilon^2}\sum_{i=1}^K \widehat{\mathbf{E}}_j\left[|\overline{\pi}^j_{b_{\widehat{\theta}}}(i) - \overline{\pi}^j_{b_{\theta_N}}(i)|\right].$$

Denote for each i, and $\theta \in \Theta^K$, $\Delta^{(j)}_k X := X^{(j)}_{(k+1)\Delta} - X^{(j)}_{k\Delta}$,

$$\overline{\pi}^j_{b_\theta}(i) = \phi_i\left(\overline{F}_\theta\left(\overline{X}^{(j)}\right)\right), \quad \overline{F}^i_\theta(\overline{X}^{(j)}) = \sum_{k=0}^{n-1}\frac{b\left(\theta_i, X^{(j)}_{k\Delta}\right)}{\sigma^2(X^{(j)}_{k\Delta})}\left(\Delta^{(j)}_k X\right) - \frac{\Delta}{2}\frac{b^2(\theta_i, X^{(j)}_{k\Delta})}{\sigma^2(X^{(j)}_{k\Delta})}.$$

Applying Cauchy-Schwartz Inequality, we obtain

$$\widehat{\mathbf{E}}_j\left|\sum_{k=0}^{n-1}\frac{b(\widehat{\theta}_i, X^{(j)}_{k\Delta}) - b(\theta_{N,i}, X^{(j)}_{k\Delta})}{\sigma^2(X^{(j)}_{k\Delta})}\left(\Delta^{(j)}_k X\right)\right| \leq \sum_{k=0}^{n-1}\sqrt{\widehat{\mathbf{E}}_j\left[\frac{(b(\widehat{\theta}_i, X^{(j)}_{k\Delta}) - b(\theta_{N,i}, X^{(j)}_{k\Delta}))^2}{\sigma^4(X^{(j)}_{k\Delta})}\right]\widehat{\mathbf{E}}_j\left[\left(\Delta^{(j)}_k X\right)^2\right]}.$$

Since as $\|\widehat{\theta} - \theta_N\| \leq \frac{1}{N^{1+(\alpha/2)}}$ the first expectation in the r.h.s. is bounded, then Lemma 7.2 controls the second term. Finally, we get

$$\widehat{\mathbf{E}}_j\left[\left|\sum_{k=0}^{n-1}\frac{b(\widehat{\theta}_i, X^{(j)}_{k\Delta}) - b(\theta_{N,i}, X^{(j)}_{k\Delta})}{\sigma^2(X^{(j)}_{k\Delta})}\left(\Delta^{(j)}_k X\right)\right|\right] \leq \frac{1}{\sigma_0^2}\frac{C_T}{N^{1+(\alpha/2)}\sqrt{\Delta}}.$$

Furthermore,

$$\widehat{\mathbf{E}}_j\left[\left|\sum_{k=0}^{n-1}\Delta\frac{b^2(\widehat{\theta}_i,X_{k\Delta}^{(j)})-b^2(\theta_{N,i},X_{k\Delta}^{(j)})}{\sigma^2(X_{k\Delta}^{(j)})}\right|\right]\leq\frac{C_T}{\sigma_0^2N^{1+(\alpha/2)}}.$$

Combining the previous inequalities, we obtain

$$\widehat{\mathbf{E}}\left[|\widehat{R}_\phi(\overline{h}^\varepsilon_{\widehat{\theta}})-\widehat{R}_\phi(\overline{h}^\varepsilon_{\theta_N})|\right]\leq C_T\frac{K^2}{\sqrt{\Delta}\varepsilon^2N^{1+(\alpha/2)}}. \tag{7.15}$$

With the bounds from Lemma 7.7, 7.8, and Equations (7.13), (7.14), (7.15) we obtain an upper bound for $\widehat{\mathbf{E}}\left[R_\phi(\overline{h}^\varepsilon_{\widehat{\theta}})-R_\phi(h_{\theta^*})\right]$. Now to ensure the convergence it is sufficient to choose $\varepsilon=O(N^{-\beta})$ with $0<\beta<\min(1/2,\alpha/4)$ (as $\Delta=O(N^{-\alpha})$). This concludes the proof. □

7.4.4. *Proof of Theorem 4.10.* For each $h\in\mathcal{H}=\{h=(h^1,\dots,h^K)\ :\mathcal{X}_T\to\mathbb{R}^K\}$ being a vector of score functions, we define $l_h(Z,X)=\sum_{i=1}^K\phi(Z_ih^i(X))$ with $\phi:x\mapsto(1-x)^2$ the least squares loss and $Z_i=2\,\mathbb{1}_{\{Y=i\}}-1$ for each $i\in\mathcal{Y}$. First, we prove the following Lemma

Lemma 7.9. *Let $h\in\mathcal{H}$, such that $\|h\|_\infty\leq1$. The following holds*

$$\mathbb{E}\left[\left(l_h(Z,X)-l_{h_{\theta^*}}(Z,X)\right)^2\right]\leq16K\left(R_\phi(h)-R_\phi(h_{\theta^*})\right).$$

Proof. Note that for each $i\in\mathcal{Y}$, we have $|Z_i|=1$ and

$$\phi(Z_ih^i(X))-\phi(Z_ih^i_{\theta^*}(X)) \;=\; (h^i(X)-h^i_{\theta^*}(X))(h^i(X)+h^i_{\theta^*}(X)-2Z_i). \tag{7.16}$$

Hence, from Cauchy-Schwarz Inequality, as $|Z_i|=1$ and $|h^i(X)|\leq1$, we have

$$\begin{aligned}\mathbb{E}\left[\left(l_h(Z,X)-l_{h_{\theta^*}}(Z,X)\right)^2\right] &\leq K\sum_{i=1}^K\mathbb{E}\left[\left(\phi(Z_ih^i(X))-\phi(Z_ih^i_{\theta^*}(X)\right)^2\right]\\ &\leq 16K\sum_{i=1}^K\mathbb{E}\left[(h^i(X)-h^i_{\theta^*}(X))^2\right].\end{aligned}$$

Now since $\mathbb{E}\left[Z_i|X\right]=2\pi_{\theta^*}(i)-1=h^i_{\theta^*}(X)$, we deduce from (4.8) and (7.16) that

$$R_\phi(h)-R_\phi(h_{\theta^*})=\sum_{i=1}^K\mathbb{E}\left[(h^i(X)-h^i_{\theta^*}(X))^2\right],$$

which gives the result. □

Now, we consider $\widetilde{\theta}:=\underset{\theta\in\Theta^K}{\operatorname{argmin}}\,R_\phi\left(\overline{h}_\theta\right)$ and we introduce

$$D_\theta:=R_\phi(\overline{h}_\theta)-R_\phi(\overline{h}_{\widetilde{\theta}})),\quad \widehat{D}_\theta:=\widehat{R}_\phi(\overline{h}_\theta)-\widehat{R}_\phi(\overline{h}_{\widetilde{\theta}})). \tag{7.17}$$

Then

$$R_\phi\left(\overline{h}_{\widehat{\theta}}\right)-R_\phi(h_{\theta^*})\leq D_{\widehat{\theta}}+R_\phi\left(\overline{h}_{\widetilde{\theta}}\right)-R_\phi(h_{\theta^*}).$$

Observing that

$$R_\phi\left(\overline{h}_{\widetilde{\theta}}\right)-R_\phi(h_{\theta^*})\leq R_\phi\left(\overline{h}_{\theta^*}\right)-R_\phi(h_{\theta^*})\leq C_TK\sqrt{\Delta}, \tag{7.18}$$

we deduce

$$R_\phi\left(\overline{h}_{\widehat{\theta}}\right)-R_\phi(h_{\theta^*})\leq D_{\widehat{\theta}}+C_TK\sqrt{\Delta}. \tag{7.19}$$

We now proceed to bound $D_{\widehat{\theta}}$. Since $\widehat{\theta}=\underset{\theta\in\Theta^K}{\operatorname{argmin}}\,\widehat{R}_\phi(\overline{h}_\theta)$, we see that $\widehat{D}_{\widehat{\theta}}\leq0$. In particular, we have the decomposition

$$D_{\widehat{\theta}}\;\leq\;D_{\widehat{\theta}}-2\widehat{D}_{\widehat{\theta}}\;=\;(D_{\widehat{\theta}}-D_{\theta_N})+(2\widehat{D}_{\theta_N}-2\widehat{D}_{\widehat{\theta}})+(D_{\theta_N}-2\widehat{D}_{\theta_N}). \tag{7.20}$$

Now, for $\alpha \geq 2$, we denote by $\mathcal{B}_N$ the $1/N^{(1+\alpha/2)}$-net of Θ^K w.r.t $\|.\|$. Similarly to the proof of Proposition 4.8 given in Section 7.4.3, we have

$$\mathbf{E}\left[|R_\phi(\overline{h}_{\widehat{\theta}}) - R_\phi(\overline{h}_{\theta_N})|\right] \leq \frac{C_T K^2}{N^{1+(\alpha/2)}} \text{ and } \mathbf{E}\left[|\widehat{R}_\phi(\overline{h}_{\widehat{\theta}}) - \widehat{R}_\phi(\overline{h}_{\theta_N})|\right] \leq \frac{C_T K^2}{N^{1+(\alpha/2)}\sqrt{\Delta}}.$$

Therefore, from Equation (7.19) and the decomposition (7.20),

$$\widehat{\mathbf{E}}\left[R_\phi\left(\overline{h}_{\widehat{\theta}}\right) - R_\phi(h_{\theta^*})\right] \leq \frac{C_T K^2}{N^{1+(\alpha/2)}} + \frac{C_T K^2}{N^{1+(\alpha/2)}\sqrt{\Delta}} + C_T K\sqrt{\Delta} + \widehat{\mathbf{E}}\left[\max_{\theta\in\mathcal{B}_N}\left(D_\theta - 2\widehat{D}_\theta\right)\right].$$

To conclude the proof, it remains to control the last term in the r.h.s. We have that for $u \geq 0$

$$\widehat{\mathbf{E}}\left[\max_{\theta\in\mathcal{B}_N}\left(D_\theta - 2\widehat{D}_\theta\right)\right] \leq u + \int_u^{+\infty} \widehat{\mathbf{P}}\left(\max_{\theta\in\mathcal{B}_N}\left(D_\theta - 2\widehat{D}_\theta\right) \geq t\right) dt.$$

In order to control the deviations of $\max_{\theta\in\mathcal{B}_N}\left(D_\theta - 2\widehat{D}_\theta\right)$, we observe that from Lemma 7.9, for each $\theta \in \Theta^K$,

$$\begin{aligned}
\mathbb{E}\left[\left(l_{\overline{h}_\theta}(Z,X) - l_{\overline{h}_{\widetilde{\theta}}}(Z,X)\right)^2\right] &\leq 2\mathbb{E}\left[\left(l_{\overline{h}_\theta}(Z,X) - l_{h_{\theta^*}}(Z,X)\right)^2 + \left(l_{\overline{h}_{\widetilde{\theta}}}(Z,X) - l_{h_{\theta^*}}(Z,X)\right)^2\right] \\
&\leq 32K\left(R_\phi(\overline{h}_\theta) - R_\phi(h_{\theta^*})\right) + 32K\left(R_\phi(\overline{h}_{\widetilde{\theta}}) - R_\phi(h_{\theta^*})\right).
\end{aligned}$$

Hence, from Equation (7.18) and the definition of D_θ, we deduce for $t \geq \sqrt{\Delta}$,

$$\mathbb{E}\left[\left(l_{\overline{h}_\theta}(Z,X) - l_{\overline{h}_{\widetilde{\theta}}}(Z,X)\right)^2\right] \leq C_T K\left(D_\theta + \sqrt{\Delta}\right) \leq C_T K\left(D_\theta + t\right). \tag{7.21}$$

Using the fact that for each θ, $\|\overline{h}_\theta\|_\infty \leq 1$ and that ϕ is $L-$Lipschitz with constant $L = 4$, from Bernstein's Inequality, we get for $t \geq 0$,

$$\begin{aligned}
\widehat{\mathbf{P}}\left(\max_{\theta\in\mathcal{B}_N}\left(D_\theta - 2\widehat{D}_\theta\right) \geq t\right) &\leq \sum_{\theta\in\mathcal{B}_N} \widehat{\mathbf{P}}\left(2D_\theta - 2\widehat{D}_\theta \geq t + D_\theta\right) \\
&\leq \sum_{\theta\in\mathcal{B}_N} \exp\left(\frac{-N(t+D_\theta)^2/8}{\mathbb{E}\left[\left(l_{\overline{h}_\theta}(Z,X) - l_{\overline{h}_{\widetilde{\theta}}}(Z,X)\right)^2\right] + (t+D_\theta)(4K/3)}\right).
\end{aligned}$$

Thus, from Equation (7.21), for $t \geq \sqrt{\Delta}$, we get

$$\widehat{\mathbf{P}}\left(\max_{\theta\in\mathcal{B}_N}\left(D_\theta - 2\widehat{D}_\theta\right) \geq t\right) \leq \mathrm{Card}\left(\mathcal{B}_N\right)\exp\left(-\frac{NC_T t}{K}\right).$$

Finally, since $\Delta = O\left(N^{-\alpha}\right)$ with $\alpha \geq 2$, for N large enough we have that $\sqrt{\Delta} \leq \dfrac{K\log\left(\mathrm{Card}\left(\mathcal{B}_N\right)\right)}{NC_T}$. Hence, with $u = \dfrac{K\log\left(\mathrm{Card}\left(\mathcal{B}_N\right)\right)}{NC_T}$, we obtain for N large enough

$$\widehat{\mathbf{E}}\left[\max_{\theta\in\mathcal{B}_N}\left(D_\theta - 2\widehat{D}_\theta\right)\right] \leq C\frac{K\log\left(\mathrm{Card}\left(\mathcal{B}_N\right)\right)}{N},$$

which concludes the proof. □

References

Audibert, J.-Y. & Tsybakov, A. (2007). Fast learning rates for plug-in classifiers. *The Annals of statistics* **35**, 608–633.

Baíllo, A., Cuevas, A. & Cuesta-Albertos, J. A. (2011). Supervised classification for a family of gaussian functional models. *Scandinavian Journal of Statistics* **38**, 480–498.

Bartlett, P., Jordan, M. & McAuliffe, J. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association* **101**, 138–156.

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

Biau, G., Bunea, F. & Wegkamp, M. (2005). Functional classification in hilbert spaces. *IEEE Transactions on Information Theory* **51**, 2163–2172.

Biau, G., Cadre, B. & Paris, Q. (2015). Cox process functional learning. *Statistical Inference for Stochastic Processes* **18**, 257–277.

Biau, G., Cérou, F. & Guyader, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory* **56**, 2034–2040.

Bibby, B. M. & Sørensen, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* pp. 17–39.

Cadre, B. (2013). Supervised classification of diffusion paths. *Math. Methods Statist.* **22**, 213–225. ISSN 1066-5307.

Comte, F., Genon-Catalot, V. & Rozenholc, Y. (2007). Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13**, 514–543.

Cuevas, A., Febrero, M. & Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics* **22**, 481–496.

Dacunha-Castelle, D. & Duflo, M. (1983). *Probabilité et Statistiques 2 Problémes á temps mobile.* Masson, Paris.

Delattre, M., Genon-Catalot, V. & Samson, A. (2015a). Estimation of population parameters in stochastic differential equations with random effects in the diffusion coefficient. *ESAIM: Probability and Statistics* **19**, 671–688.

Delattre, M., Genon-Catalot, V. & Samson, A. (2015b). Mixtures of stochastic differential equations with random effects: application to data clustering. *Journal of Statistical Planning and Inference, to appear* .

Denis, C. (2014). Classification in postural style based on stochastic process modeling. *The international journal of biostatistics* **10**, 251–260.

Denis, C. & Hebiri, M. (2017). Confidence sets with expected sizes for multiclass classification. *J. Mach. Learn. Res.* **18**, Paper No. 102, 28.

Devroye, L., Györfi, L. & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer.

Dion, C. & Genon-Catalot, V. (2015). Bidimensional random effect estimation in mixed stochastic differential model. *Stochastic Inference for Stochastic Processes* **18**, 1–28.

Donnet, S. & Samson, A. (2013). A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. *Advanced Drug Delivery Reviews* **65**, 929–939.

Eberlein, E., Keller, U. *et al.* (1995). Hyperbolic distributions in finance. *Bernoulli* **1**, 281–299.

El Karoui, N., Peng, S. & Quenez, M. C. (1997). Backward stochastic differential equations in finance. *Mathematical finance* **7**, 1–71.

Genon-Catalot, V. & Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l'IHP Probabilités et statistiques* **29**, 119–151.

Gloter, A. (2000). Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient. *ESAIM: Probability and Statistics* **4**, 205–227.

Gobet, E. (2002). Lan property for ergodic diffusions with discrete observations. *Annales de l'Institut Henri Poincare (B) Probability and Statistics* **38**, 711–737.

Gobet, E., Hoffmann, M., Reiß, M. *et al.* (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics* **32**, 2223–2253.

Graham, C. & Talay, D. (2013). *Stochastic simulation and Monte Carlo methods*, vol. 68 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg.

Gregorutti, B., Michel, B. & Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis* **90**, 15–35.

Guyon, X. (1995). *Random fields on a network: modeling, statistics, and applications*. Springer Science & Business Media.

Hoffmann, M. (1999). Adaptive estimation in diffusion processes. *Stochastic processes and their Applications* **79**, 135–163.

Jacod, J. & Shiryaev, A. (2003). *Limit Theorems for Stochastic Processes, ser. A Series of Comprehensive Studies in Mathematics*. Berlin: Springer-Verlag.

Jakobsen, N., Munkholt & Sørensen, M. (2017). Efficient estimation for diffusions sampled at high frequency over a fixed time interval. *Bernoulli* **23**, 1874–1910.

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics* **24**, 211–229.

Kessler, M., Sørensen, M. *et al.* (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli* **5**, 299–314.

Kuelbs, J. & Zinn, J. (2016). Limit theorems for quantile and depth regions for stochastic processes. In *High Dimensional Probability VII*, pp. 255–280. Springer.

Kutoyants, Y. (2004). *Statistical Inference for Ergodic Diffusion Processes*. Springer, London.

Lange, T., Mosler, K. & Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. *Statistical Papers* **55**, 49–69.

López-Pintado, S. & Romo, J. (2006). Depth-based classification for functional data. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **72**, 103.

Massart, P. & Nédélec, E. (2006). Risk bounds for statistical learning. *The Annals of Statistics* **34**, 2326–2366.

Parisi, G. & Sourlas, N. (1992). *Supersymmetric field theories and stochastic differential equations*. World Scientific.

Pires, B., Szepesvari, C. & Ghavamzadeh, M. (2013). Cost-sensitive multiclass classification risk bounds. In *International Conference on Machine Learning*, pp. 1391–1399.

Ramsay, J. O. & Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Revuz, D. & Yor, M. (2013). *Continuous martingales and Brownian motion*, vol. 293. Springer Science & Business Media.

Tewari, A. & Bartlett, P. (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research* **8**, 1007–1025.

Van Der Vaart, A. W. & Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer.

Vapnik, V. (1998). *Statistical learning theory*, vol. 1. Wiley New York.

Wang, J.-L., Chiou, J.-M. & Mueller, H.-G. (2015). Review of functional data analysis. *arXiv preprint arXiv:1507.05135* .

Yoshida, N. (1990). Asymptotic behavior of m-estimator and related random field for diffusion process. *Annals of the Institute of Statistical Mathematics* **42**, 221–251.

Yoshida, N. (1992). Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis* **41**, 220 – 242.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research* **5**, 1225–1251.