Nonparametric plug-in classifier for multiclass classification of S.D.E. paths

Christophe Denis, Charlotte Dion-Blanc, Eddy ELLA MINTSA, Chi Tran

Nonparametric plug-in classifier for multiclass classification of S.D.E. paths

Christophe Denis[(1)], Charlotte Dion-Blanc[(2)], Eddy Ella-Mintsa[(1)], Viet Chi Tran[(1,3)]

December 16, 2022

(1) LAMA, Université Gustave Eiffel
(2) Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75013 Paris, France
(3) CRM-CNRS, Université de Montréal.

Abstract

We study the multiclass classification problem where the features come from a mixture of time-homogeneous diffusion. Specifically, the classes are discriminated by their drift functions while the diffusion coefficient is common to all classes and unknown. In this framework, we build a plug-in classifier which relies on nonparamateric estimators of the drift and diffusion functions. We first establish the consistency of our classification procedure under mild assumptions and then provide rates of convergence under different set of assumptions. Finally, a numerical study supports our theoretical findings.

Keywords: Supervised learning; Multiclass classification; Nonparametric estimation; Plug-in classifier; Diffusion process

MSC: 62G05; 62M05; 62H30

1 Introduction

The massive collection of functional data has found many applications in recent years for the modeling of the joint (time)-evolution of agents – individuals, species, particles – that are represented by some sets of features – time-varying variables such as geographical positions, population sizes, portfolio values etc. Examples can be found in mathematical finance (see *e.g.* El Karoui *et al.*, 1997), biology (see *e.g.* Erban & Chapman, 2009), or physics (see *e.g.* Domingo *et al.*, 2020). This gave rise to an abundant literature on statistical methods for functional data, (see *e.g.* Ramsay & Silverman, 2005; Wang *et al.*, 2016, for a review). Within this context, the study of efficient supervised classification procedures that are designed to handle temporal data is a major challenge. Indeed, usual learning algorithms such as random forests, kernel methods or neural networks are not directly tailored to take into account the temporal dependency of the data. Recently, this question has drawn a lot of attention, see Rossi & Villa (2008); Baíllo *et al.* (2011); Wang *et al.* (2020); De Micheaux *et al.* (2021); Kidger *et al.* (2021) any references therein.

In the present paper, we tackle the multiclass classification problem where the features belong to a particular family of functional data, namely trajectories, whose temporal dynamic is modelled by stochastic differential equation. In this framework, we propose a nonparametric plug-in type procedure for such data generated by diffusion processes observed at discrete time. Hence, our work takes place

in the high frequency setup. Let us denote by (X, Y) a random couple built on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The feature $X = (X_t)_{t\in[0,1]}$ is a real-valued diffusion process whose drift and diffusion coefficient depend on its associated label Y taking values in $\mathcal{Y} = \{1, \cdots, K\}$, with $K \geq 2$. More precisely, for each $i \in \mathcal{Y}$, X is a solution of a stochastic differential equation whose drift function, denoted by b_i^*, depends on the class i. The marginal distribution of X is hence a mixture of distributions of time-homogeneous diffusion processes. We assume that a learning sample $\mathcal{D}_N = \{((X_t^i)_{t\in[0,1]}, Y_i), i = 1, \ldots, N\}$ is provided, composed of N *i.i.d.* random couple with distribution $\mathbb{P}_{(X,Y)}$. Additionally, in this paper, the diffusions X_i are observed on a subdivision $\{0, 1/n, \cdots, 1\}$ of the time interval $[0, 1]$, for a positive integer n. Since we deal with multiclass classification setting, the statistical goal is then to build, based on $\mathcal{D}_N$, a classifier $\widehat{g}$, such that $\widehat{g}(X)$ is a prediction of the associated label Y of a new path X. Besides, we expect that the empirical classifier mimics the optimal Bayes classifier g^* characterized as

$$g^*(X) \in \underset{k\in\mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}\left(g^*(X) \neq Y\right).$$

Specifically, we propose a classification procedure based on the plug-in principle. In particular, the construction of our empirical classifier relies on estimators of both drift and diffusion coefficients. The performance of a predictor $\widehat{g}$ is assessed through its excess risk $\mathbb{E}\left[\mathbb{P}(\widehat{g}(X) \neq Y) - \mathbb{P}\left(g^*(X) \neq Y\right)\right]$. In the finite dimensional classification setup (*e.g.* $X \in \mathbb{R}^d$), rates of convergence for plug-in rules are usually obtained under the strong density assumption (X admits a density which is lower bounded) as in Audibert *et al.* (2007); Gadat *et al.* (2016). However, theoretical properties of plug-in rules in supervised classification of trajectories are much less studied.

Related works. Up to our knowledge, the work of Cadre (2013) is the first one that tackles the problem of supervised classification in the stochastic differential equation framework. More precisely, the authors consider the model where $X = (X_t)_{t\in[0,1]}$ is a mixture of two diffusion processes and provide a classifier based on the empirical risk minimization strategy for which they establish rates of convergence. However, the proposed method is not implementable since it involves the minimization of a non-convex criterion. More recently, Gadat *et al.* (2020), and Denis *et al.* (2020) study plug-in classifiers for classification of diffusion paths. In Gadat *et al.* (2020) the authors propose a plug-in rule for the binary classification problem where the trajectories are generated by Gaussian processes, solutions of the white noise model. In this model, the drift function depends on time and on the label Y, also, the diffusion coefficient is supposed to be constant and known. Within this framework, Gadat *et al.* (2020) establish the optimality of their estimation procedure which reaches the minimax rate of convergence of order $N^{-s/(2s+1)}$, where the drift function is assumed to belong to a Sobolev space of regularity $s \geq 1$. Under an additional margin type assumption, they also derive faster rates of convergence. Closest to our framework, Denis *et al.* (2020) also consider the challenging multiclass problem where the drift functions are space-dependent. However, the authors consider drift functions modeled under parametric assumptions, keeping the diffusion coefficient known and constant. They propose a plug-in classifier for which only consistency is established.

In the present work, we consider plug-in classifier that relies on nonparametric estimators of the drift and diffusion coefficients. The literature on this topic is extensive. Usually, the construction of estimators of drift and diffusion functions relies on the observation of a single path. For instance, Hoffmann (1999b) study minimax rate of convergence for the estimation of the diffusion coefficient on a compact interval. For the inference of the drift coefficient, the main references using penalized contrasts can be found for long time observation with high frequency data in Hoffmann (1999a); Comte *et al.* (2007); Comte & Genon-Catalot (2021). However, since we deal with the multiclass classification framework, the construction of estimators of both drift and diffusion coefficients is based on the learning sample $\mathcal{D}_N$ which is composed of repeated observations of the process on the fixed time-interval $[0, 1]$.

Recently, Comte & Genon-Catalot (2020a); Marie & Rosier (2021); Della-Maestra & Hoffmann (2022) consider nonparametric procedures for the estimation of the drift function for continuous observations in the context of *i.i.d.* observations when the horizon time is fixed. Furthermore, towards high-frequency data, Denis *et al.* (2021) study minimum contrast estimator under a l_2 constraint.

Main contributions. In this paper, we extend the results of Denis *et al.* (2020) and Gadat *et al.* (2020) in several directions. In particular, one of the major contribution is to provide, up to our knowledge, the first study of rates of convergence for plug-in classifier in the mixture model of time-homogeneous diffusion. Importantly, we highlight that extending the results of Gadat *et al.* (2020) to diffusion models in which the drift functions are space-dependent and the diffusion coefficient is either unknown or non-constant add many difficulties. Besides, contrary to Denis *et al.* (2020), we consider the nonparametric mixture model where both drift *and* diffusion functions are unknown as well as the weights of the mixture. Specifically, we build a plug-in classifier that relies on the Girsanov's theorem and involves nonparametric estimators of the drift functions $b_i^*, i \in \mathcal{Y}$, and the diffusion coefficient. The construction of our estimators is inspired of the ridge estimators provided in Denis *et al.* (2021), and consists in the minimization of a least-squares type contrast over a finite dimensional subspace under a l_2-constraint. The considered space of approximation is then spanned by the B-spline basis De Boor (1978).

One of the main difficulty of the study of statistical properties of the plug-in classifiers in our context is that it requires deriving rates of convergence for the drift and diffusion coefficients on a non-compact interval. It hence implies that the strong density assumption does not hold, although, we consider assumptions that ensure existence of transition density. Notably, our results embed generalization of the results provided in Denis *et al.* (2021) for the estimation of non-compactly supported drift functions for B-spline based estimators, but also exhibit the first result for the estimation of the diffusion coefficient in the *i.i.d.* framework. A salient point of our theoretical findings is obtained when the diffusion coefficient is constant and known. In this case, by leveraging the results of Comte & Genon-Catalot (2020a), we show that optimal rates for drift estimation can only be achieved on intervals included in $[-C\sqrt{\log(N)}, C\sqrt{\log(N)}]$, with $C > 0$.

To sum up our results, a first part is dedicated to the consistency of our plug-in classifier which is obtained under very mild assumptions. In a second part, convergence rates are established in three particular cases.

(i) When the drift functions are bounded and Lipschitz, and the diffusion coefficient is unknown and possibly non-constant, we obtain a rate of convergence of order $N^{-1/5}$ for the plug-in classifier (up to a factor of order $\exp(\sqrt{c\log(N)}),\ c > 0$).

(ii) When the diffusion coefficient is known and constant, and when the drift functions are bounded and belongs to some Hölder space with regularity β, using some arguments developed in Comte & Genon-Catalot (2020b) and Comte & Genon-Catalot (2021) for the estimation of non-compactly supported drift functions, together with approximations of the transition density of X (as they are intractable), we then prove that the plug-in classifier reaches rate of order $N^{-\beta/(2\beta+1)}$ (up to a factor of order $\exp(\sqrt{c\log(N)}),\ c > 0$).

(iii) When the drifts are unbounded but re-entrant and Hölder continuous with regularity β, we obtain a rate of convergence of order $N^{-3\beta/(4(2\beta+1))}$. Notice that when $\beta = 1$ and $d = 1$, it corresponds to the rate found in Gadat *et al.* (2016).

Outline of the paper. Section 2 is dedicated to presentation of the mathematical framework for the classification task. Then, the construction of the plug-in classifier is described in Section 3 and

its consistency is established in Section 3.3. In Section 4 we provide rates of convergence of our plug-in procedure under different assumptions. We perform a numerical experiment that supports our theoretical results in Section 5. Finally, We provide a discussion in Section 6 and the proofs of our results are postponed to Section 7.

2 Statistical setting

We consider the multiclass classification problem, where the feature X comes from a mixture of Brownian diffusions with drift. More precisely, the generic data-structure is a couple (X, Y) where the label Y takes its values in the set $\mathcal{Y} := \{1, \dots, K\}$ with distribution denoted by $\mathfrak{p}^* = (\mathfrak{p}_1^*, \cdots, \mathfrak{p}_K^*)$, and where the process $X = (X_t)_{t\in[0,1]}$ is defined as the solution of the following stochastic differential equation

$$dX_t = b_Y^*(X_t)dt + \sigma^*(X_t)dW_t, \quad X_0 = 0, \tag{1}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion independent of Y. In the following, we denote by $\mathbf{b}^* = (b_1^*, \dots, b_K^*)$ the vector of drift functions. The real-valued functions $b_i^*(.)$, $i \in \mathcal{Y}$, and the diffusion coefficient $\sigma^*(.)$ are assumed to be unknown. We also assume that $0 < \mathfrak{p}_0^* = \min_{i\in\mathcal{Y}} \mathfrak{p}_i^*$.

In this framework, the objective is to build a classifier g, *i.e.* a measurable function such that the value $g(X)$ is a prediction of the associated label Y of X. The accuracy of such classifier g is then assessed through its misclassification risk, denoted by

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

In the following, the set of all classifiers is denoted by $\mathcal{G}$.

The main assumptions considered throughout the paper are presented in Section 2.1. The definition and characterization of the optimal classifier *w.r.t.* the misclassification risk, namely the *Bayes classifier*, is provided in Section 2.2

2.1 Assumptions

The following assumptions ensure that Equation (1) admits a unique strong solution (see Karatzas & Shreve, 2014, Theorem 2.9), and that the diffusion process X admits a transition density $p : (t, x) \in ([0, 1] \times \mathbb{R}) \mapsto p(t, x)$ (see for example Gobet, 2002).

Assumption 2.1. *(Ellipticity and regularity)*

(i) There exists $L_0 > 0$ such that the functions $b_i^*, i = 1, \dots, K$ and σ^* are L_0-Lipschitz:

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

(ii) There exist real constants σ_0^*, σ_1^* such that

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

(iii) $\sigma^* \in \mathcal{C}^2(\mathbb{R})$ and there exists $\gamma \geq 0$ such that : $|\sigma^{*\prime}(x)| + |\sigma^{*\prime\prime}(x)| \leq \gamma\left(1 + |x|^\gamma\right), \quad \forall x \in \mathbb{R}.$

Assumption 2.1 insures that for any integer $q \geq 1$, there exists $C_q > 0$ such that

$$\mathbb{E}\left[\sup_{t\in[0,1]} |X_t|^q\right] \leq C_q.$$

We also assume that the following Novikov's criterion is fulfilled (Revuz & Yor, 1999, Prop. (1.15) p. 308) .

Assumption 2.2. *(Novikov's condition) For all $i \in \mathcal{Y}$, we have*

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s)ds\right)\right] < +\infty.$$

In particular, this assumption allows to apply Girsanov's theorem that is a key ingredient to derive a characterization of the Bayes classifier in the next section.

2.2 Bayes Classifier

The Bayes classifier g^* is a minimizer of the misclassification risk over $\mathcal{G}$

$$g^* \in \operatorname*{argmin}_{g\in\mathcal{G}} \mathcal{R}(g),$$

and is expressed as

$$g^*(X) \in \operatorname*{argmax}_{i\in\mathcal{Y}} \pi_i^*(X), \quad \text{with} \quad \pi_i^*(X) := \mathbb{P}\left(Y = i|X\right).$$

The following result of Denis *et al.* (2020) provides a closed form of the conditional probabilities π_i^*, $i \in \mathcal{Y}$.

Proposition 2.3. *(Denis* et al.*, 2020) Under Assumptions 2.1, 2.2, for all $i \in \mathcal{Y}$, we define*

$$F_i^*(X) := \int_0^1 \frac{b_i^*}{\sigma^{*2}}(X_s)dX_s - \frac{1}{2}\int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s)ds.$$

Under Assumptions 2.1,2.2, for each $i \in \mathcal{Y}$, the conditional probability π_i^* is given as follows:

$$\pi_i^*(X) = \phi_i\left(\mathbf{F}^*(X)\right),$$

where $\mathbf{F}^* = (F_1^*, \ldots, F_k^*)$, and $\phi_i^* : (x_1, \cdots, x_K) \mapsto \frac{\mathfrak{p}_i^* e^{x_i}}{\sum_{k=1}^K \mathfrak{p}_k^* e^{x_k}}$ are the softmax functions.

The above proposition provides an explicit dependency of the Bayes classifier on the unknown parameters $\mathbf{b}^*$, σ^*, and $\mathfrak{p}^*$. Hence, it naturally suggests to build *plug-in* type estimators $\widehat{g}$ of the Bayes classifier g^*, relying on estimators of the unknown parameters. In this way, we aim at building an empirical classifier whose misclassification risk is closed to the minimum risk which is reached by the Bayes classifier. The following section is devoted to the presentation of the classification procedure.

3 Classification procedure: a plug-in approach

Let $n \geq 1$ be an integer, and $\Delta_n = 1/n$ the time step which defines the regular grid of the observation time interval $[0, 1]$. Let us assume now that an observation is a couple $(\bar{X}, Y)$, with $\bar{X} := (X_{k\Delta_n})_{0\leq k\leq n}$ a high frequency sample path coming from $(X_t)_{t\in[0,1]}$ a solution of Equation (1), and Y its associated label. We also introduce, for $N \geq 1$, a learning dataset $\mathcal{D}_N = \{(\bar{X}^j, Y_j),\ j \in \{1, \ldots N\}\}$ which consist of N independent copies of $(\bar{X}, Y)$. The asymptotic framework is such that N and n tend to infinity.

Based on $\mathcal{D}_N$ we build a classification procedure that relies on the result of Proposition 2.3. Our classifier uses the knowledge of the class Y_j for the path X^j, placing our work in the frame of supervised learning. The procedure is formally described in Section 3.1 and Section 3.2 while its statistical properties are provided in Section 3.3.

3.1 Classifier and excess risk

As suggested by Proposition 2.3, based on $\mathcal{D}_N$, we first build estimators $\widehat{\boldsymbol{b}} = (\widehat{b}_1, \dots, \widehat{b}_K)$, and $\widehat{\sigma}$ of $\mathbf{b}^*$ and σ^* respectively. Besides, we consider the empirical estimators of $\mathfrak{p}_i^*$, $i = 1, \dots, K$

$$\widehat{\mathfrak{p}}_i = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{Y_j = i\}}. \tag{2}$$

Then, in a second step, we introduce the discretized estimator of $\mathbf{F}^*$

$$\widehat{\boldsymbol{F}} = (\widehat{F}_1, \dots, \widehat{F}_K), \text{ with } \widehat{F}_i(X) = \sum_{k=0}^{n-1} \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2}(X_{k\Delta}) \left(X_{(k+1)\Delta} - X_{k\Delta} \right) - \frac{\Delta}{2} \frac{\widehat{b}_i^2}{\widehat{\sigma}^2}(X_{k\Delta}) \right). \tag{3}$$

Finally, considering the functions $\widehat{\phi}_i : (x_1, \cdots, x_K) \mapsto \frac{\widehat{\mathfrak{p}}_i e^{x_i}}{\sum_{k=1}^K \widehat{\mathfrak{p}}_k e^{x_k}}$, we naturally define the resulting plug-in classifier $\widehat{g}$ as

$$\widehat{g}(X) \in \underset{i \in \mathcal{Y}}{\operatorname{argmax}}\, \widehat{\pi}_i(X), \text{ with } \widehat{\pi}_i(X) = \widehat{\phi}_i(\widehat{\boldsymbol{F}}(X)). \tag{4}$$

Hereafter, we establish that the consistency of the plug-in classifier $\widehat{g}$ can be obtained through an empirical distance between estimators $\widehat{\boldsymbol{b}}$, and $\widehat{\sigma}$ and the true functions $\mathbf{b}^*$, and σ^* respectively. This distance relies on the empirical norm $\|.\|_{n,i}$ defined for $h : \mathbb{R} \to \mathbb{R}$ as.

$$\|h\|_{n,i}^2 := \mathbb{E}_{X|Y=i} \left[\frac{1}{n} \sum_{k=0}^{n-1} h^2(X_{k\Delta}) \right].$$

We also introduce the general empirical norm $\|.\|_n$ which for any function h is

$$\|h\|_n^2 := \mathbb{E}_X \left[\frac{1}{n} \sum_{k=0}^{n-1} h^2(X_{k\Delta}) \right].$$

Let us now announce the main result on the excess risk of a plug-in type classifier.

Theorem 3.1. *Assume N and n fixed (and large). Grant Assumptions 2.1, 2.2. Assume that there exists $b_{\max}, \sigma_0^2 > 0$ such that for all $x \in \mathbb{R}$*

$$\max_{i \in \mathcal{Y}} |\widehat{b}_i(x)| \leq b_{\max} \text{ and } \widehat{\sigma}^2(x) \geq \sigma_0^2. \tag{5}$$

Then the classifier $\widehat{g}$ defined in Equation (4) satisfies

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C \left(\sqrt{\Delta_n} + \frac{1}{\mathfrak{p}_0^* \sqrt{N}} + \mathbb{E}\left[b_{\max} \sigma_0^{-2} \sum_{i=1}^{K} \|\widehat{b}_i - b_i^*\|_n \right] + \mathbb{E}\left[\sigma_0^{-2} \|\widehat{\sigma}^2 - \sigma^{2*}\|_n \right] \right),$$

where $C > 0$ is a constant which depends on b^*, σ^*, and K.

Theorem 3.1 highlights that the excess risk of the plug-in classifier depends on the discretization error which is of order $\Delta_n^{-1/2}$, the L_2 error of $\widehat{\mathfrak{p}}$ which is of order $N^{-1/2}$, and the estimation error of $\widehat{\mathbf{b}}$ and $\widehat{\sigma}^2$ assessed through the empirical norm $\|.\|_n$. Therefore, a straightforward consequence of Theorem 3.1 is that consistent estimators of $\mathbf{b}^*$, and σ^{*2} yield the consistency of plug-in classifier $\widehat{g}$. Notice that the additional assumption (5) does not require that the true functions b_i^*'s are bounded, only their estimators should be. For the difference between b_i^* and $\widehat{b}_i$ to remain controlled in the norm $\|\cdot\|_n$, it is necessary that the process X rests with high probability in a compact region of $\mathbb{R}$. The next section is devoted to the construction of consistent estimators of both drift and diffusion coefficient.

3.2 Estimators of drift and diffusion coefficients

In this section, we provide consistent estimators $\widehat{\boldsymbol{b}}$, and $\widehat{\sigma}^2$, implying the consistency of the associated plug-in classifier. These estimators are defined as minimum contrast estimators under an l_2-constraint on a finite dimensional vector space spanned by the B-spline basis, but other families of nonparametric estimators could have been chosen as well. In particular, to ensure statistical guarantees on $\mathbb{R}$, the considered estimators are built on a large interval $(-\log(N), \log(N))$, parameterized by the number N of sample paths, and that tends to the whole real line as N goes to infinity.

3.2.1 Spaces of approximation

Let $K_N > 0$, and $M \geq 1$. Let $\mathbf{u} = (u_{-M}, \ldots, u_{K_N+M})$, a sequence of knots of the compact interval $[-\log(N), \log(N)]$ such that

$$u_{-M} = \cdots = u_{-1} = u_0 = -\log(N), \quad \text{and} \quad u_K = u_{K_N+1} = \cdots = u_{K_N+M} = \log(N).$$
$$\forall \ell \in [\![0, K_N]\!], \quad u_\ell = -\log(N) + \frac{2\ell \log N}{K_N}.$$

Let us consider the B-spline basis $(B_{-M}, \ldots, B_{K_N+M})$ of order M defined by the knots sequence $\mathbf{u}$. For the construction of the B-spline and its properties, we refer for instance to (Györfi *et al.*, 2006). Let us mention that the considered B-spline functions are M-1 continuously differentiable on $(-\log(N), \log(N))$ and are zero outside $[-\log(N), \log(N)]$. Besides, for all $x \in (-\log(N), \log(N))$, we have that $\sum_{i=-M}^{K_N+M} B_i(x) = 1$. Now, we introduce the space of approximation $\mathcal{S}_{K_N,M}$ defined as

$$\mathcal{S}_{K_N,M} := \left\{ \sum_{\ell=-M}^{K_N+M} a_\ell B_\ell, \quad \|\mathbf{a}\|_2 \leq (K_N + M)\log^3(N) \right\}, \tag{6}$$

where $\|\mathbf{a}\|_2 = \sum_{\ell=-M}^{K_N+M} a_\ell^2$ is the usual ℓ_2-norm. The introduction of the constraint space $\mathcal{S}_{K_N,M}$ is motivated by two facts. The first one is the following important property of spline approximations, inspired by the related properties for the Hölder functions (see Györfi *et al.* (2006)):

Proposition 3.2. *Let h be a L-lipschitz function. Then there exists $\tilde{h} \in \mathcal{S}_{K_N,M}$, such that*

$$|\tilde{h}(x) - h(x)| \leq C\frac{\log(N)}{K_N}, \quad \forall x \in (-\log(N), \log(N)),$$

where $C > 0$ depends on L, and M.

The second one is that the set of functions $\mathcal{S}_{K_N,M}$ is a *totally bounded class*, in the following sense (Devroye *et al.*, 2013, Chapter 28). According to Denis *et al.* (2021), for each $\varepsilon > 0$, there exists an ε-net $\tilde{\mathcal{S}}_\varepsilon$ of $\mathcal{S}_{K_N,M}$ *w.r.t.* to the supremum norm $\|.\|_\infty$ such that

$$\log\left(\mathrm{card}(\tilde{\mathcal{S}}_\varepsilon)\right) \leq C_M \left(\frac{\sqrt{K_N \log^3(N)}}{\varepsilon}\right)^{K_N}.$$

It shows that the complexity of $\mathcal{S}_{K_N,M}$ given in Equation (6) is parametric which is particularly appealing in order to apply concentration inequalities.

3.2.2 Minimum contrast estimators

In this section, we propose two estimators of $\boldsymbol{b}^*$, and σ^{*2} which lead to a plug-in classifier that exhibits appealing properties. The construction of the estimators $\widehat{\boldsymbol{b}}$, and $\widehat{\sigma}^2$ relies on the minimization of a least squares contrast function over the space $\mathcal{S}_{K_N,M}$. They are both based on the observed increments of the process X.

Estimator of the drift functions. Let $i \in \mathcal{Y}$ and $N_i := \sum_{j=1}^{N} \mathbb{1}_{\{Y_j=i\}}$ a random variable of Binomial distribution with parameters $(N, \mathfrak{p}_i^*)$. We define the random set $\mathcal{I}_i := \{j,\ Y_j = i\}$ and consider the dataset $\left\{\bar{X}^j,\ j \in \mathcal{I}_i\right\}$ composed of the observations of the class i. The first estimator $\widetilde{b}_i$ of b_i^* is defined as

$$\tilde{b}_i \in \underset{h \in \mathcal{S}_{K_N,M}}{\operatorname{argmin}} \frac{1}{nN_i} \sum_{j \in \mathcal{I}_i} \sum_{k=0}^{n-1} \left(Z^j_{k\Delta_n} - h(\bar{X}^j_{k\Delta_n}) \right)^2 \mathbb{1}_{N_i>0}, \text{ with } Z^j_{k\Delta_n} := \frac{(\bar{X}^j_{(k+1)\Delta_n} - \bar{X}^j_{k\Delta_n})}{\Delta_n}. \tag{7}$$

Then, to fit the assumption of Theorem 3.1, rather than $\tilde{b}_i$, we consider its thresholded counterpart

$$\widehat{b}_i(x) := \tilde{b}_i(x)\mathbb{1}_{\{|\tilde{b}_i(x)| \le \log^{3/2}(N)\}} + \operatorname{sgn}(\tilde{b}_i(x)) \log^{3/2}(N) \mathbb{1}_{\{|\tilde{b}_i(x)| > \log^{3/2}(N)\}}. \tag{8}$$

Note that the value of the threshold $\log^{3/2}(N)$ corresponds to the bound $b_{\max}$ in (5). Although this bound depends on N, Theorem 3.1 can be applied, but to ensure the consistency of the classifier, we now have to prove that the estimation rate for $\widehat{b}_i$ decreases sufficiently fast.

Estimator of the diffusion coefficient. The construction of the estimator of σ^* follows the same lines. However, since the diffusion coefficient is the same for all classes, we can use the whole dataset $\mathcal{D}_N$ to build its estimator. More precisely, we define

$$\tilde{\sigma}^2 \in \underset{h \in \mathcal{S}_{K_N,M}}{\operatorname{argmin}} \frac{1}{nN} \sum_{j=1}^{N} \sum_{k=0}^{n-1} \left(U^j_{k\Delta_n} - h(\bar{X}^j_{k\Delta_n}) \right)^2, \text{ with } U^j_{k\Delta_n} = \frac{(\bar{X}^j_{(k+1)\Delta_n} - \bar{X}^j_{k\Delta_n})^2}{\Delta_n} \tag{9}$$

Finally, as for the drift estimator we consider the truncated version $\widehat{\sigma}^2$ as

$$\widehat{\sigma}^2(x) := \tilde{\sigma}^2(x)\mathbb{1}_{\{\frac{1}{\log(N)} \le \tilde{\sigma}^2(x) \le \log^{3/2}(N)\}} + \log^{3/2}(N)\mathbb{1}_{\{\tilde{\sigma}^2(x) > \log^{3/2}(N)\}} + \frac{1}{\log(N)}\mathbb{1}_{\{\tilde{\sigma}^2(x) \le \frac{1}{\log(N)}\}}. \tag{10}$$

Although this constraint does not appear in Theorem 3.1, it remains natural in view of Assumption 2.1 (ii). We will impose that $\widehat{\sigma}^2$ is bounded by $\log^{3/2}(N)$ to derive its consistency.

3.3 A general consistency result

In this section, we establish the consistency of the empirical classifier based on the estimators presented in the previous section. We first provide rates of convergence for the estimator of both drift and diffusion coefficient diffusion.

Theorem 3.3. *Let $i \in \mathcal{Y}$. Assume that Assumptions 2.1, 2.2 are satisfied. Considering the estimator $\widehat{b}_i$ of b_i^* (8) and the estimator $\widehat{\sigma}^2$ of σ^{*2} (10), we have, for N then n large enough, such that $\Delta_n = O(1/N)$ and $K_N = (N\log(N))^{1/5}$,*

$$\mathbb{E}\left[\|\widehat{b}_i - b_i^*\|_{n,i}\right] \le C_1 \left(\frac{\log^4(N)}{N}\right)^{1/5}, \text{ and } \mathbb{E}\left[\|\widehat{\sigma}^2 - \sigma^{*2}\|_n\right] \le C_2 \left(\frac{\log^4(N)}{N}\right)^{1/5},$$

where $C_1, C_2 > 0$ are constants which depend on L_0, $\mathfrak{p}_0$, and K.

Several comments can be made. First, we obtain a general rate of convergence for the estimation on $\mathbb{R}$ for both drift and diffusion coefficient functions under mild assumptions. This rate is, up to a logarithmic factor, of order $N^{-1/5}$. Hence, it extends the result of Theorem 3.3 in Denis *et al.* (2021), where only consistency of drift estimators is obtained. In particular, a difficulty in establishing the

convergence rate on $\mathbb{R}$ is to control the exit probabilities from $(-\log(N), \log(N))$, which is provided here by careful estimates for the transition densities following Gobet (2002).

This result together with Theorem 3.1 yields the consistency of the plug-in classifier

$$\widehat{g} := \widehat{g}_{\widehat{\mathbf{p}},\widehat{\mathbf{b}},\widehat{\sigma}^2} \tag{11}$$

where the unknown parameters are replaced by their estimators in Equation (4). However, application of Theorem 3.1 requires the consistency of the estimator $\widehat{b}_i$ in terms of empirical norm $\|.\|_n$ and not in terms of norm $\|.\|_{n,i}$. To circumvent this issue, we can use a change of probability to get rid of the conditioning on $Y = i$. For this purpose, we take advantage of Lemma 7.3 and 7.4 to derive precise control of the transition density of the process X conditioned on $Y = i$, and then to establish the consistency of the plug-in classifier.

Theorem 3.4. *Grant Assumptions 2.1, 2.2. For N large enough such that $\Delta_n = O(1/N)$ and $K_N = (N\log(N))^{1/5}$, the classifier $\widehat{g}$ satisfies*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \underset{N\to\infty}{\longrightarrow} 0.$$

The consistency of our classification procedure is obtained under very mild assumptions. The study of the rates of convergence requires more structural assumptions. In the following section, we obtain rates of convergence of the plug-in classifier under different kind of assumptions.

4 Rates of convergence

In this section, we study the rates of convergence of the proposed method described in Section 3.1. A general rate of convergence is first provided in Section 4.1 under the additional assumption that the drift functions of the considered mixture model are bounded, no additional assumptions being made on the diffusion coefficient. In Section 4.2, we consider the case where the diffusion coefficient is known and assumed to be constant. In this case, the procedure achieves faster rates of convergence.

4.1 General rate of convergence for bounded drift function

Let us consider the following assumption.

Assumption 4.1. *There exists $C_{\mathbf{b}^*}$ such that*

$$\max_{i\in\mathcal{Y}} \|b_i^*\|_\infty \leq C_{\mathbf{b}^*}.$$

Let $i, j \in \mathcal{Y}^2$ with $i \neq j$, The following property allows to upper bound the expectation conditional on $\{Y = i\}$ by the expectation conditional on $\{Y = j\}$. This happens to be the cornerstone to derive rates of convergence for our procedure.

Proposition 4.2. *Let $N > 1$, $\alpha > 0$, and $0 \leq Z \leq \log^\alpha(N)$ a random variable measurable w.r.t. $\{\sigma(X_s, s \leq 1)\}$. Under Assumptions 2.1, 2.2, and 4.1, we have for all $i, j \in \mathcal{Y}^2$ such that $i \neq j$, and N large enough*

$$\mathbb{E}_{X|Y=i}[Z] \leq C\exp\left(\sqrt{c\log(N)}\right)\mathbb{E}_{X|Y=j}[Z] + C\frac{\log^\alpha(N)}{N},$$

where $C, c > 0$ depend on $C_{\mathbf{b}^*}, \sigma_1$, and σ_0.

A crucial consequence of this result is that in particular the empirical norms $\|.\|_{n,i}$, $i \in \mathcal{Y}$, are now equivalent up to a factor of order $\exp\left(\sqrt{c\log(N)}\right)$. Notice that for all $r_1, r_2 > 0$,

$$\log^{r_1}(N) = o\big(\exp\left(\sqrt{c\log(N)}\right)\big), \quad \text{and} \quad \exp\left(\sqrt{c\log(N)}\right) = o\big(N^{r_2}\big). \tag{12}$$

In particular, the factor $\exp\left(\sqrt{c\log(N)}\right)$ is negligible with respect to any power of N. Therefore, combining Theorem 3.1, 3.3, and Proposition 4.2, we are able to give the rate of convergence for our procedure (when the drift coefficients are globally Lipschitz and bounded).

Theorem 4.3. *Grant Assumptions 2.1, 2.2, and 4.1. the plug-in classifier $\widehat{g}$ given in Equation* (11)*, provided that $\Delta_n = O\left(N^{-1}\right)$, $K_N = (N\log(N))^{1/5}$ and N large enough, satisfies*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\exp\left(\sqrt{c\log(N)}\right)N^{-1/5},$$

where $C > 0$ depends on $C_{\mathbf{b}^*}, \sigma_1$, and σ_0.

Leveraging the result of Theorem 3.3 and Proposition 4.2, we obtain a rate of convergence which is of order $N^{-1/5}$ up to the extra factor $\exp\left(\sqrt{c\log(N)}\right)$. Note that the optimal rate of convergence obtained when the estimation of drift function is done over on a compact set is of order $N^{-1/3}$ *w.r.t.* $\|.\|_n$ rather than $N^{-1/5}$ (see Denis *et al.*, 2021; Comte & Genon-Catalot, 2021). Here, this slower rate is mainly due to the fact that our procedure requires a control of the drift estimators over $\mathbb{R}$.

In the next section, we show that when σ^* is constant and assumed to be known, we derive faster rates of convergence. In particular, under Assumption 4.1, we show that our plug-in procedure achieves a rate of convergence of order $N^{-1/3}$. Lastly, note that Theorem 4.3 can be easily extended to higher order of regularity for the drift functions (*e.g.* Hölder with regularity $\beta > 1$). In this case, the obtained rate of convergence is of order $N^{-\beta/(2\beta+3)}$.

4.2 Classifier's rate of convergence with known diffusion coefficient

In this section, we consider that the diffusion coefficient is known and constant. For sake of simplicity, we choose $\sigma^* = 1$. In this case, our plug-in procedure only involves the estimation of the drift function $\widehat{\boldsymbol{b}}$. Hence, the plug-in classifier now writes as $\widehat{g} = \widehat{g}_{\mathfrak{p},\widehat{\boldsymbol{b}},1}$.

In order to derive a general rate of convergence as a function of the drift regularity, we consider the following smoothness assumption (Tsybakov, 2008), which is a subset of Lipschitz functions.

Assumption 4.4. *For all $i \in \mathcal{Y}$, b_i^* is Hölder with regularity parameter $\beta \geq 1$.*

4.2.1 Construction of the drift estimators

Let $i \in \mathcal{Y}$, the construction of the drift estimators $\widehat{b}_i$ is slightly different as the one provided in Section 3.2. We recall that the number of paths in each class $N_i = \sum_{j=1}^{N} \mathbb{1}_{\{Y_j=i\}}$, and $\mathcal{I}_i = \{j,\ Y_j = i\} = \{i_1, \cdots, i_{N_i}\}$. Hereafter, we work *conditional on* $(\mathbb{1}_{\{Y_1=i\}}, \ldots, \mathbb{1}_{\{Y_N=i\}})$, on the event $\{N_i > 1\}$. Hence, N_i is viewed as a deterministic variable such that $N_i > 1$. Let $A_{N_i} > 0$, and $K_{N_i} > 0$, we consider a drift estimator $\widehat{b}_i$ built over the symmetric interval $[-A_{N_i}, A_{N_i}]$.

Precisely, we consider the B-spline basis of order M defined by the knots sequence $\mathbf{u}$

$$u_{-M} = \cdots = u_{-1} = u_0 = -A_{N_i}, \quad \text{and} \quad u_K = u_{K_N+1} = \cdots = u_{K_{N_i}+M} = A_{N_i}.$$
$$\forall \ell \in [\![0, K_{N_i}]\!], \quad u_\ell = -A_{N_i} + \frac{2\ell A_{N_i}}{K_{N_i}}.$$

and the set of functions

$$\mathcal{S}_{K_{N_i},M} := \left\{ \sum_{\ell=-M}^{K_{N_i}+M} a_\ell B_\ell, \;\; \|\mathbf{a}\|_2 \leq (K_{N_i}+M)\log(N_i)A_{N_i}^2 \right\}.$$

Note that this space $\mathcal{S}_{K_{N_i},M}$ is globally the same as the one given in Equation (6) except that it depends on the label i and on the interval through A_{N_i}. Finally, the estimator $\widehat{b}_i$ is defined for all x

$$\widehat{b}_i(x) := \tilde{b}_i(x)\mathbb{1}_{\{|\tilde{b}_i(x)|\leq A_{N_i}\log(N_i)^{1/2}\}} + \mathrm{sgn}(\tilde{b}_i(x))A_{N_i}\log^{1/2}(N_i)\mathbb{1}_{\{|\tilde{b}_i(x)|>A_{N_i}\log^{1/2}(N_i)\}}, \tag{13}$$

where $\tilde{b}_i$ is defined as in Equation (7). In a first step, we focus on the properties of the estimators $\widehat{b}_i$.

4.2.2 Rates of convergence for drift estimators

Let $i \in \mathcal{Y}$. The study of the rates of convergence of the estimator $\widehat{b}_i$ relies on the properties of the matrix $\Psi_{K_{N_i}} \in \mathbb{R}^{(K_{N_i}+M)^2}$ defined by

$$\Psi_{K_{N_i}} := \left(\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}_{X|Y=i}\left[B_\ell(X_{k\Delta}^i) B_{\ell'}(X_{k\Delta}^i) \right] \right)_{\ell,\ell' \in [-M, K_{N_i}-1]}. \tag{14}$$

Note that for $t \in S_{K_{N_i},M}$, $t = \sum_{i=-M}^{K_{N_i}-1} a_i B_{i,M,\mathbf{u}}$, we have the relation

$$\|t\|_{n,i}^2 = \mathbf{a}'\Psi_{K_{N_i}}\mathbf{a}, \quad \text{with} \quad \mathbf{a} = \left(a_{-M}, \cdots, a_{K_{N_i}-1}\right)'.$$

Let us remind the reader that for a matrix A, the operator norm $\|A\|_{\mathrm{op}}$ is defined as the square root of the largest eigenvalue of the matrix $A'A$. Besides, if A is symmetric, its norm is equal to its largest eigenvalue. The matrix $\Psi_{K_{N_i}}$ satisfies the following property.

Lemma 4.5. *Conditional on $(\mathbb{1}_{\{Y_1=i\}}, \ldots, \mathbb{1}_{\{Y_N=i\}})$, on the event $\{N_i > 1\}$, the matrix $\Psi_{K_{N_i}}$ given in Equation (14) satisfies*

(i) if $K_{N_i} \geq 1$, $\Psi_{K_{N_i}}$ is invertible,

(ii) under Assumption 2.1, for N large enough, if $K_{N_i} \leq \sqrt{N_i}$, there exists two constants $C, c > 0$ such that

$$c\frac{K_{N_i}}{A_{N_i}}\exp\left(\frac{A_{N_i}^2}{6}\right) \leq \|\Psi_{K_{N_i}}^{-1}\|_{\mathrm{op}} \leq C\frac{K_{N_i}\log(N_i)}{A_{N_i}}\exp\left(\frac{2}{3}A_{N_i}^2\right).$$

A major consequence of Lemma 4.5 is to give the order of A_{N_i} *w.r.t.* N_i to obtain optimal rates of convergence for the estimation of the drift function b_i^*. Indeed, Comte & Genon-Catalot (2021) show that the rates of convergence for $\widehat{b}_i$ and $\widehat{\sigma}$ in Theorem 3.3 can be established if the following constraint is satisfied

$$\|\Psi_{K_{N_i}}^{-1}\|_{\mathrm{op}} \leq C\frac{N_i}{\log^2(N_i)}, \tag{15}$$

with a constant $C > 0$ depending on β. Notably, conditional on $(\mathbb{1}_{\{Y_1=i\}}, \ldots, \mathbb{1}_{\{Y_N=i\}})$, if K_{N_i} is of order $N_i^{1/(2\beta+1)}$ (up to some extra logarithmic factors), and A_{N_i} is chosen such that Equation (15) is satisfied, then the drift estimator converges as $N_i^{-2\beta/(2\beta+1)}$ *w.r.t.* $\|.\|_{n,i}^2$. Interestingly, this is the

same rate of convergence obtained in (Denis *et al.*, 2021) when the estimation of the drift function is performed over a fixed compact interval.

From this remark, Lemma 4.5 teaches us that if K_{N_i} is of order $\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}$, Equation (15) is satisfied for $A_{N_i} \leq \sqrt{\frac{3\beta}{2\beta+1}\log(N)}$. Furthermore, the lemma shows that the order of A_{N_i} is tight. Indeed, for another choice of A_{N_i} such that

$$\frac{A_{N_i}}{\sqrt{\log(N_i)}} \longrightarrow +\infty \text{ as } N \to +\infty,$$

then the condition (15) is no longer satisfied. Based on this observation, the next result establishes the rates of convergence for our proposed drift estimator on the event $\{N_i > 1\}$.

Theorem 4.6. *Let Assumptions 2.1, 2.2 and 4.4 be satisfied. Let* $b^*_{A_{N_i},i} = b^*_i \mathbb{1}_{[-A_{N_i},A_{N_i}]}$ *defined on the event* $\{N_i > 1\}$. *If* $A_{N_i} \leq \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$, $K_{N_i} \propto \left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right)$, *and* $\Delta_n = O\left(N^{-1}\right)$. *Then for all* $i \in \mathcal{Y}$

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i} \mathbb{1}_{N_i>1}\right] \leq C\log^{6\beta}(N)N^{-2\beta/(2\beta+1)},$$

where C *is a constant which depends on* $\mathbf{b}^*$.

The above result shows that for a proper choice of A_{N_i} the drift estimators $\widehat{b}_i$ achieves, up to a logarithmic factor, the minimax rates of convergence *w.r.t.* $\|.\|_{n,i}$ (see Theorem 4.7 in (Denis *et al.*, 2021)). Notably, Theorem 4.6 extends results obtained in (Denis *et al.*, 2021) to the estimation of the drift function on a interval which depends on N.

In Section 4.2.3 and Section 4.2.4, we exploit this result to derive rates of convergence for the plug-in classifier $\widehat{g}$ defined as follows. On the event $\{\min_{i\in\mathcal{Y}} N_i > 1\}$, we consider the estimators $\widehat{\mathbf{b}}$ presented in Section 4.2.1, and define the plug-in classifier $\widehat{g} = \widehat{g}_{\widehat{\mathfrak{p}},\widehat{\mathbf{b}},1}$. On the complementary event $\{\min_{i\in\mathcal{Y}} N_i \leq 1\}$, we simply set $\widehat{g} = 1$.

4.2.3 Rates of convergence: bounded drift functions

In this section, we assume that additionally to $\sigma^* = 1$, Assumption 4.1 is fulfilled (the drift function is bounded). Hence, we can use Proposition 4.2, and apply Theorem 4.6 to derive rates of convergence for plug-in estimator $\hat{g}$.

Theorem 4.7. *Grant Assumptions 2.1, 2.2, 4.1, 4.4. Assume that for all* $i \in \mathcal{Y}$, *on the event* $\{N_i > 1\}$, $A_{N_i} = \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$ *and* $K_{N_i} \propto \left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right)$, *and* $\Delta_n = O\left(N^{-1}\right)$. *Then the plug-in classifier* $\widehat{g} = \widehat{g}_{\widehat{\mathbf{b}},1}$ *satisfies*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\exp\left(\sqrt{c\log(N)}\right)N^{-\beta/(2\beta+1)}$$

where $C, c > 0$ *are constants depending on* $\mathbf{b}^*$, β, K *and* $\mathfrak{p}_0$.

The above theorem shows that the plug-in classifier $\widehat{g}$ achieves faster rates of convergence than in the case where σ^* is unknown (see Theorem 4.3). Notably, the obtained rate is of the same order, up to a factor of order $\exp\left(\sqrt{c\log(N)}\right)$, than the rates of convergence provided in Gadat *et al.* (2020) in the framework of binary classification of functional data where the observation are assumed to come from a white noise model. In their setting, $\sigma^* = 1$ and the drift functions depend only on the observation time interval, which is also assumed to be $[0, 1]$. Therefore, our specific setup is more challenging

since the drift functions are space-dependent which involves to deal with estimation of function on a non-compact interval. Finally, it is worth noting that, up to the $\exp\left(\sqrt{c\log(N)}\right)$ factor, the rate of convergence provided in Theorem 4.7 is the same as the minimax rates in the classical classification framework where the feature vector X belongs to $\mathbb{R}$ and that X admits a lower bounded density (Yang, 1999; Audibert *et al.*, 2007).

4.2.4 Rates of convergence: when the drift functions are re-entrant

In this section, we study performance of the plug-in classifier when the drift functions are not necessarily bounded. In this context, rates of convergence are obtained under the following assumption.

Assumption 4.8. *(re-entrant drift function) For each label $i \in \mathcal{Y}$, there exists $c_0 > 4$ and $K_0 \in \mathbb{R}$ such that*

$$\forall x \in \mathbb{R}, \quad b_i^*(x)x \leq -c_0 x^2 + K_0.$$

An important consequence of this assumption is that there exists $C > 0$ (see Proposition 1.1 in (Gobet, 2002)) such that

$$\mathbb{E}\left[\exp(4|X_t|^2)\right] \leq C, \tag{16}$$

which yields a better bound on the tail probability $\mathbb{P}\left(|X_t| \geq A\right)$ for $A > 0$. It worth noting that under Assumption 4.8, the drift functions are not bounded. Hence, we can not take advantage of Proposition 4.2 to derive rates of convergence. Nonetheless, we obtain the following result.

Theorem 4.9. *Grant Assumptions 2.1, 2.2, 4.4, 4.8. Assume that for all $i \in \mathcal{Y}$, on the event $\{N_i > 1\}$, $A_{N_i} = \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$ and $K_{N_i} \propto \left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right)$, and $\Delta_n = O\left(N^{-1}\right)$. Then, the plug-in classifier $\widehat{g}$ satisfies*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\log^{3\beta+1}(N)N^{-3\beta/4(2\beta+1)}.$$

The above theorem shows that the rate of convergence of the plug-in classifier is, up to a logarithmic factor, of order $N^{-3\beta/4(2\beta+1)}$. Therefore, this rate of convergence is slightly slower than the one provided in Theorem 4.7. It is mainly due to the fact that under Assumption 4.8, Proposition 4.2 does not apply and then, in view of considered assumptions in Theorem 4.9, we only manage to obtain the following bound,

$$\forall i, j \in \mathcal{Y} : \ i \neq j, \ \mathbb{E}_{X|Y=i}[Z] \leq CN^{\beta/4(2\beta+1)}\mathbb{E}_{X|Y=j}[Z],$$

which is clearly worst than the one obtain in Proposition 4.2. Interestingly, for $\beta = 1$, we can note that the rates obtained in Theorem 4.9 are of the same order than the rates of convergence established in Gadat *et al.* (2016) in the classification setup where the input vector lies in $\mathbb{R}$ under the assumption that X does not fulfil the strong density assumption (*e.g.* the density of X is not lower bounded).

5 Simulation study

This section is devoted to numerical experiments that support our theoretical findings. A first part is dedicated to the study of the performance of the plug-in classifier in a setting which meets the assumptions of Section 4.1. The considered model is presented in Section 5.1. The implementation of the proposed procedure is discussed in Section 5.2 while the performances of the plug-in classifier are given in Section 5.3. Finally, several features of the problem are investigated in Section 5.4. In particular, we consider the classical Ornstein-Uhlenbeck model, for which assumptions of Section 4.1 are not fulfilled.

5.1 Models and simulation setting

We fix $K = 3$ classes in the following. Note that, we do not consider larger value of K since the evaluation of the impact of K on the procedure is beyond the scope of this paper. To illustrate the accuracy of the presented plug-in classifier, we investigate the model described in Table 1. This toy

$b_1^*(x)$	$1/4 + (3/4)\cos^2 x$
$b_2^*(x)$	$\theta[1/4 + (3/4)\cos^2 x]$
$b_3^*(x)$	$-\theta[1/4 + (3/4)\cos^2 x]$
$\sigma^*(x)$	$0.1 + 0.9/\sqrt{1+x^2}$

Table 1: *Drift and diffusion coefficients, depending on* $\theta \in \Theta = \{1/2, 3/4, (4+\alpha)/4, \alpha \in [\![1, 12]\!]$.

model, described in Table 1, fulfills the assumptions of Section 4.1. Interestingly, this model allows evaluating the influence of the distance between the drift functions of each of the three classes, on the classification problem, through the parameter θ. Indeed,

$$\min_{i,j=1,2,3} \|b_i^* - b_j^*\|_\infty = \theta, \quad \text{where} \quad \theta \in \Theta = \{1/2, 3/4, (4+\alpha)/4, \alpha \in [\![1, 12]\!]\}.$$

We investigate the consistency of the empirical classifier using learning samples of size $N \in \{100, 1000\}$ with $n \in \{100, 500\}$ (and thus with $\Delta_n = 1/n$). We use the `R`-package `sde` (see Iacus, 2009) to simulate the solution of the stochastic differential equation corresponding to the chosen model.

Figure 1 displays simulated trajectories from the proposed model. On the left panel (right panel respectively) the observed learning sample comes from the model with parameter $\theta = 1/2$ ($\theta = 4$ respectively) and each class is represented by one color. We can see from Figure 1 that the distance between the drift functions strongly impacts the dispersion of the trajectories and leads to a more difficult classification task.

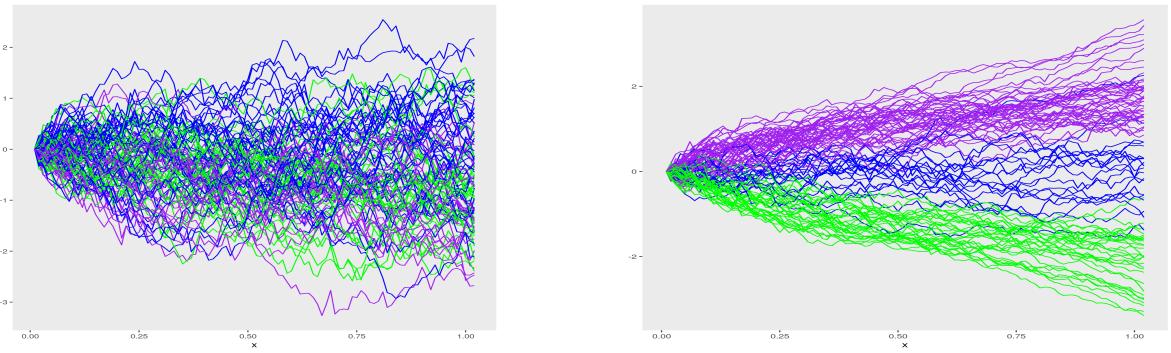

Figure 1: *Dispersion of diffusion paths from model given in Tabel 1. Left:* $\theta = 1/2$*, right:* $\theta = 4$ *(blue lines* $K = 1$*, purple lines* $K = 2$*, green lines* $K = 3$*); with* $N = 100$ *and* $n = 100$*.*

Performance of the Bayes classifier. We evaluate the performance of the Bayes classifier g^* with respect to four values of parameter θ ($\theta \in \{1/2, 3/2, 5/2, 4\}$). To this end, we compute its average error rate over 100 repetitions of the following steps

(i) simulate $\mathcal{D}_M$ of size $M = 4000$ with $n = 500$;

(ii) based on $\mathcal{D}_M$ compute the misclassification error rate of the discrete counterpart of g^*.

Table 2 provides the mean and standard deviation of the misclassification risk. The obtained results highlight the significant impact of the minimum distance θ, between the drift functions of each class, on the performance of g^*. Indeed, as expected, the Bayes classifier is more accurate on our model when parameter θ is large, especially in the case of separable data ($\theta = 4$). On the contrary, the worst case corresponds to $\theta = 0.5$. In this model, the data are highly ambiguous.

	$\theta = 1/2$	$\theta = 3/2$	$\theta = 5/2$	$\theta = 4$
$\widehat{\mathcal{R}}(g^*)$	0.49 (0.01)	0.36 (0.01)	0.22 (0.01)	0.11 (0.01)

Table 2: *Classification risks of the Bayes classifier g^* w.r.t parameter θ from learning samples of size $N = 4000$ with $n = 500$.*

5.2 Implementation of the plug-in classifier

Hereafter, we briefly describe the implementation of the proposed plug-in classifier. We first estimate the drift functions b_i^*, $i = 1, 2, 3$. For each $i \in \{1, 2, 3\}$, the estimator $\widehat{b}_i$ is built on the interval $[-A_{N_i}, A_{N_i}]$. Since the drifts (and the diffusion) coefficients are bounded, we can use the construction considered in Section 3. Therefore, we fix $A_{N_i} = \log(N)$, $M = 3$, and divide the learning sample $\mathcal{D}_N$ into sub-samples $\mathcal{D}_N^i$ of size N_i that contains all diffusion paths belonging to the class i. From the sub-sample $\mathcal{D}_N^i$, we build estimators $\widehat{b}_i$, $i = 1, 2, 3$.

For the construction of the estimator $\widehat{b}_i$, we have to choose the dimension parameter K_{N_i}. We follow Denis *et al.* (2021), and consider an adaptive choice denoted by $\widehat{K}_{N_i}$.

Let us remind the reader that in Denis *et al.* (2021), the adaptive dimension $\widehat{K}_{N_i}$ is selected such that $\widehat{K}_{N_i}$ is the minimizer of the following penalized contrast

$$\widehat{K}_{N_i} := \underset{K \in \mathcal{K}}{\operatorname{argmin}} \left\{ \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} (\widehat{b}_{i,K} - Z^j_{k\Delta_n})^2 + \operatorname{pen}_b(K) \right\}, \tag{17}$$

where $\mathcal{K} \in \{2^q,\ q \in [\![0, 5]\!]\}$, and $\widehat{b}_{i,K}$ is the drift estimator built on the approximation subspace $\mathcal{S}_{K,M}$. Besides, $\operatorname{pen}_b(K) = \kappa(K + M)\log^3(N)/N$ is the penalty function with $\kappa > 0$. We fix the parameter $\kappa = 0.1$ as recommended in Denis *et al.* (2021).

For the estimation of σ^2, we consider the whole sample $\mathcal{D}_N$ and apply the methodology described in Section 3 with $M = 3$. We follow the same lines to build an adaptive estimator of σ^{2*}, and choose $\widehat{K}_N$ as the minimizer over $\mathcal{K}$ of the following penalized contrast

$$\widehat{K}_N := \underset{K \in \mathcal{K}}{\operatorname{argmin}} \left\{ \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} (\widehat{\sigma}^2_K - U^j_{k\Delta_n})^2 + \operatorname{pen}_\sigma(K) \right\}, \tag{18}$$

where $\hat{\sigma}^2_K$ is the estimator built on $\mathcal{S}_{K,M}$, and $\operatorname{pen}_\sigma(K) := \kappa(K + M)\log^3(N)/Nn$ is the penalty function, with $\kappa > 0$. The value of the tuning parameter κ is calibrated through an intensive simulation study and chosen equal to $\kappa = 5$.

5.3 Simulation results

The performance of the plug-in classifier $\widehat{g}$ is evaluated by repeating 100 times the following steps

1. Simulate learning samples $\mathcal{D}_N$ and $\mathcal{D}_{N'}$ with $N \in \{100, 1000\}$, $N' = 1000$, and $n \in \{100, 500\}$;

2. for each $i \in \{1,2,3\}$, from the sub-sample $D_N^i = \{\bar{X}^j, j \in \mathcal{I}_i\}$, select $\widehat{K}_N$ minimizing (17) and compute the estimator $\widehat{b}_{i,\widehat{K}_N}$ of b_i^* given in Equation (8);

3. from $\mathcal{D}_N$ select $\widehat{K}_N$ using Equation (18) and compute the estimator $\widehat{\sigma}^2_{\widehat{K}_N}$ of σ^{*2} given in (10);

4. based on $\mathcal{D}_N$ compute $\widehat{\mathfrak{p}} = \left(\frac{1}{N}\sum_{j=1}^N \mathbb{1}_{Y_j=1}, \frac{1}{N}\sum_{j=1}^N \mathbb{1}_{Y_j=2}, \frac{1}{N}\sum_{j=1}^N \mathbb{1}_{Y_j=3}\right)$;

5. based on $\mathcal{D}_{N'}$, compute the error rate of the plug-in classifier $\widehat{g}$ where $\widehat{\mathbf{b}} = \left(\widehat{b}_{1,\widehat{K}_N}, \widehat{b}_{2,\widehat{K}_N}, \widehat{b}_{3,\widehat{K}_N}\right)$ and $\widehat{\sigma}^2 = \widehat{\sigma}^2_{\widehat{K}_N}$, and $\widehat{\mathfrak{p}}$.

From these repetitions, we compute the empirical mean and standard deviation of the error rate of $\widehat{g}$. The results are given in Table 3 and Figure 2. As expected, from Table 3 and Table 2, we can see that the error rate of the plug-in classifier $\widehat{g}$ is closed to the error rate of the Bayes classifier. In particular, for $N = 1000$, it performs as well as the Bayes classifier. Note that the length of the paths n does not significantly impact the performance of $\widehat{g}$. Moreover, from Figure 2, we can make similar comments as for the Bayes classifier (see Table 2), in particular, the accuracy of $\widehat{g}$ decreases as parameter θ increases.

$\widehat{\mathcal{R}}(\widehat{g})$	$n = 100$		$n = 500$	
	$N = 100$	$N = 1000$	$N = 100$	$N = 1000$
$\theta = 1/2$	0.53 (0.05)	0.50 (0.05)	0.53 (0.05)	0.49 (0.05)
$\theta = 3/2$	0.39 (0.06)	0.37 (0.05)	0.39 (0.05)	0.36 (0.05)
$\theta = 5/2$	0.24 (0.05)	0.22 (0.04)	0.25 (0.04)	0.22 (0.04)
$\theta = 4$	0.12 (0.03)	0.10 (0.03)	0.11 (0.03)	0.10 (0.03)

Table 3: *Risks of the plug-in classifier $\widehat{g}$ w.r.t. values of parameter θ*

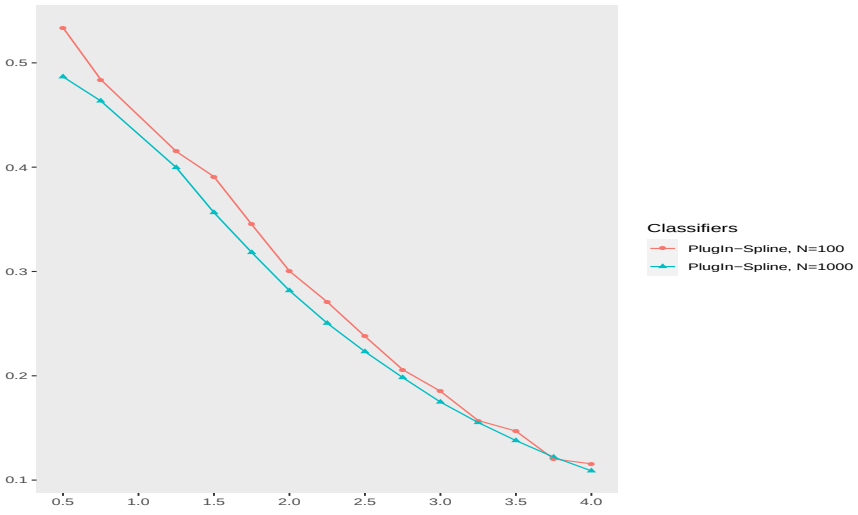

Figure 2: *Risks of the plug-in classifier w.r.t values of the minimum gap θ between the drift functions*

5.4 Ornstein-Uhlenbeck model

In this section, we focus on the influence of the diffusion coefficient σ^* on the performance of our plug-in procedure. To this end, we consider the Ornstein-Uhlenbeck diffusion model given in Table 5.4 where the diffusion coefficient σ^* is constant. Let us notice also that in this model the drift functions are unbounded. We investigate the performance of the plug-in classifier $\widehat{g}$ *w.r.t.* the level of noise σ^*.

$b_1^*(x)$	$1-x$
$b_2^*(x)$	$-1-x$
$b_3^*(x)$	$-x$
$\sigma^*(x)$	σ

Table 4: *Ornstein-Uhlenbeck mixture model with* $K = 3$

This study is motivated by the fact that, inherently, the diffusion coefficient impacts the dispersion of the trajectories. Therefore, it can lead to separable data when σ^* is close to zero, and ambiguous data for large values of σ^*. Thus, we evaluate the performance of $\widehat{g}$ for $\sigma^* = 1/2$ which is close enough to zero, and for larger value $\sigma^* \in \{1, 3/2\}$. We first consider the case where σ^* is unknown. The results are given in Table 5 and confirm our intuition. The error rate of the plug-in classifiers decreases as σ^* decreases.

In a second step, we investigate the influence of estimating the coefficient σ^* in the procedure. To evaluate this point, we assess the error rate of the plug-in classifier when $\sigma^* = 1$ is known. In this case, we only estimate the drift functions and the weights of mixture $\mathfrak{p}p^*$ to build our predictor. The results are given in Table 6. First, we can notice that by comparison with results provided in Table 5, there is almost no impact on the performance of the plug-in classifier when we assume the diffusion coefficient σ^* in the Ornstein-Uhlenbeck model to be known or not.

Finally, we also study the influence of parameter A_N on the estimation procedure. Indeed, our theoretical results indicates that A_N should be of order $\sqrt{\log(N)}$ when σ^* is constant and known, while $A_N = \log(N)$ is recommended when σ^* is unknown. To this end, we evaluate the error rate of our procedure for these choices. The results are also provided in Table 6 and show that the performance are almost the same in the two cases.

	$\widehat{\mathcal{R}}(\widehat{g})$	$\widehat{\mathcal{R}}(g^*)$
$\sigma^* = 1/2$	0.23 (0.04)	0.21 (0.01)
$\sigma^* = 1$	0.44 (0.05)	0.41 (0.01)
$\sigma^* = 3/2$	0.52 (0.05)	0.49 (0.01)

Table 5: *Evolution of the performance of the plug-in classifier* $\widehat{g}$ *and of* g^* *w.r.t values of the constant diffusion coefficient* σ^* *for* $N = 100$ *and* $n = 100$.

6 Conclusion and discussion

In this paper, we propose a plug-in classifier for the multiclass classification of trajectories generated by a mixture of diffusion processes whose drift functions $b_i^*,\ i \in \mathcal{Y}$ and diffusion coefficient σ^* are assumed to be unknown. In the considered model, each class i is characterized by a drift function, b_i^* whereas

	$N = 100$	$N = 1000$
$A_N = \sqrt{\log(N)}$	0.44 (0.05)	0.41 (0.05)
$A_N = \log(N)$	0.43 (0.05)	0.43 (0.05)

Table 6: *Risk classification of $\widehat{g}$ when the diffusion $\sigma^* = 1$ is known, and $n = 100$.*

the diffusion coefficient σ^* is common for all classes. This work extends to the nonparametric case, the multiclass classification procedure provided in Denis *et al.* (2020) where $\sigma^* = 1$ and the drift functions depend on an unknown parameter $\theta \in \mathbb{R}^d$. Our proposed procedure relies on consistent projection estimators $\widehat{b}_i, i \in \mathcal{Y}$ and $\widehat{\sigma}^2$ of the drift and diffusion coefficients on a constrained approximation subspace spanned by the spline basis. We establish the consistency, *w.r.t.* the excess risk, of our procedure and then studied its rate of convergence under different kind of assumptions. In particular, we show that the proposed plug-in classifier reaches a rate of convergence of order $N^{-1/5}$ (up to a factor of order $\exp(\sqrt{c\log(N)})$) when $\mathbf{b}^*$, σ^*, and $\mathfrak{p}^*$ are unknown. Besides, a numerical study illustrates the performance of our classification procedure.

In the case where $\sigma^* = 1$, we manage to derive faster rates of convergence. In particular, when the drift functions are bounded and Hölder with regularity $\beta \geq 1$, we obtained a rate of order $N^{-\beta/(2\beta+1)}$ (up to a factor of order $\exp(\sqrt{c\log(N)})$). Interestingly, this result can be viewed as an extension of the one obtained in Gadat *et al.* (2020) to the multiclass mixture model, where the drift functions are time-dependent. Furthermore, up to $\exp(\sqrt{c\log(N)})$ factor, our rate of convergence matches the optimal rates of convergence obtained in the univariate setting (*e.g.* $X \in \mathbb{R}$), in Audibert *et al.* (2007). Finally, for the case of unbounded drift functions, we assume that the drift functions are the re-entrant. Taking advantage of this property, we establish that our plug-in classifier achieves a rate of convergence of order $N^{-3\beta/4(2\beta+1)}$. For $\beta = 1$, this rate of convergence is of the same order as the one obtained in Gadat *et al.* (2016) for plug-in classifier in the univariate classification setting, when the feature X does not satisfy the strong density assumption.

A question that can be tackled for future research is the study of the optimality in the minimax sense of our plug-in procedure. In particular, the adaptivity of estimators of the drift and diffusion coefficients should be investigated. Furthermore, it might be interesting to consider the margin type assumption as in Gadat *et al.* (2020) to derive faster rates of convergence. Also, following Denis *et al.* (2020), it is natural to derive theoretical properties for empirical risk minimization procedure based on convex losses. Finally, the extension to the high-dimensional setting would require further work. In particular, the control of the transition densities is different in this setting.

7 Proofs

The section is devoted to the proofs of our main results. In order to simplify the notation, we write $\Delta_n = \Delta$. Besides, $C > 0$ is a constant which may change from one line to another. When the dependency on a parameter θ needs to be highlighted, we write C_θ.

7.1 Technical results on the process X

Lemma 7.1. *For all integer $q \geq 1$, there exists $C^* > 0$ depending on q such that for all $0 \leq s < t \leq 1$,*

$$\mathbb{E}\,|X_t - X_s|^{2q} \leq C^*(t-s)^q.$$

For each $t \in [0,1]$ and $x \in \mathbb{R}$, we denote by $p(t,x)$ the transition density of the underlying process X_t given the starting point $X_0 = 0$. We also denote by $p_i(t,\cdot)$ the transition density of the process driven

by the drift function b_i^*. Note that Assumption 2.1 ensures the existence of the transition densities. The rest of this section is dedicated to some results on the transition densities p_i for $i = 1, \dots, K$. Nonetheless, since the transition p of the process X writes as

$$p = \sum_{i=1}^{K} \mathfrak{p}_i^* p_i,$$

all these results apply also for p. The following proposition is provided in (Gobet, 2002) (Proposition 1.2).

Proposition 7.2. *Under Assumptions 2.1 and 2.2, there exist constants $c > 1$, $K > 1$ such that for all $t \in (0, 1]$, $x \in \mathbb{R}$, and $i = 1, \dots, K$*

$$\frac{1}{K\sqrt{t}} \exp\left(-c\frac{x^2}{t}\right) \leq p_i(t, x) \leq \frac{K}{\sqrt{t}} \exp\left(-\frac{x^2}{ct}\right).$$

From this result, we can deduce an evaluation of the probability of the process to exit a compact set. This is the purpose of the next result.

Lemma 7.3. *Under Assumption 2.1 and 2.2, there exist $C_1, C_2 > 0$ such that for all $A > 0$*

$$\sup_{t\in[0,1]} \mathbb{P}\left(|X_t| \geq A\right) \leq \frac{C_1}{A} \exp(-C_2 A^2).$$

Proof. Let $A > 0$, we have for $t \in (0, 1]$,

$$\mathbb{P}\left(|X_t| \geq A\right) = 2\int_A^{+\infty} p(t, x)\mathrm{d}x.$$

From Proposition 7.2, we then deduce that

$$\mathbb{P}\left(|X_t| \geq A\right) \leq C\frac{\sqrt{t}}{A}\int_A^{+\infty} \frac{2c}{t} \exp\left(-c\frac{x^2}{t}\right) \mathrm{d}x \leq \frac{C\sqrt{t}}{A} \exp\left(-\frac{cA^2}{t}\right).$$

From the above inequality, and using that $t \in (0, 1]$, we deduce the result. □

Lemma 7.4. *There exist C_0, C_1, and C_2, such that for $i = 1, \dots, K$, for $x \in [-A, A]$, we have*

$$C_1 \exp\left(-C_2 x^2\right) \leq \frac{1}{n}\sum_{k=1}^{n-1} p_i(k\Delta, x) \leq C_0.$$

Proof of Lemma 7.4. For $i \in \{1, \dots, K\}$, for all $x \in \mathbb{R}$, we have from Proposition 7.2,

$$\frac{1}{n}\sum_{k=1}^{n} p_i(k\Delta, x) \leq \frac{C}{n}\sum_{k=1}^{n} \frac{1}{\sqrt{k\Delta}} = \frac{C}{\sqrt{n}}\sum_{k=1}^{n} \frac{1}{\sqrt{k}} \leq \frac{2C}{\sqrt{n}}\sum_{k=1}^{n} \frac{1}{\sqrt{k+1}}. \tag{19}$$

Since the function $x \mapsto \frac{1}{\sqrt{x}}$ is decreasing over $[1, +\infty[$, we deduce from Equation (19) that

$$\frac{1}{n}\sum_{k=1}^{n} p_i(k\Delta, x) \leq \frac{4C\sqrt{n+1}}{\sqrt{n}} \leq C_0,$$

which gives the upper bound. For the lower bound, we observe from Proposition 7.2 that for $k \in [\![1, n-1]\!]$, and $x \in \mathbb{R}$,

$$C \exp\left(-\frac{cx^2}{k\Delta}\right) \leq \frac{C}{\sqrt{k\Delta}} \exp\left(-\frac{cx^2}{k\Delta}\right) \leq p_i(k\Delta, x). \tag{20}$$

Since $g : (s, x) \mapsto \exp\left(-\frac{cx^2}{s}\right)$ is strictly increasing in s over $(0, 1]$, we obtain for $k \in [\![1, n-1]\!]$,

$$\int_{\frac{1}{n}}^{\frac{n-1}{n}} g(s,x)ds \leq \sum_{k=2}^{n-1} \int_{(k-1)\Delta}^{k\Delta} \left(g(k\Delta, x) + g(s,x) - g(k\Delta, x)\right) ds \leq \frac{1}{n} \sum_{k=1}^{n-1} \exp\left(-\frac{cx^2}{k\Delta}\right).$$

Hence, we deduce that for $n \geq 3$, and $x \in [-A, A]$

$$\frac{1}{6} \exp(-2cA^2) \leq \int_{\frac{1}{2}}^{\frac{n-1}{n}} g(s,x)\mathrm{d}s \leq \frac{1}{n} \sum_{k=1}^{n-1} \exp\left(-\frac{cx^2}{k\Delta}\right).$$

For the first lower bound, we use that $g(s, x) \geq e^{-2cA^2}$ for $x \in [-A, A]$ and $s \geq 1/2$, and that the length of $[1/2, (n-1)/n]$ is larger than $1/6$ for $n \geq 3$. Finally, gathering this bound with Equation (20), leads to

$$\frac{1}{6} \exp(-2cA^2) \leq \frac{1}{n} \sum_{k=1}^{n-1} \exp\left(-\frac{cx^2}{k\Delta}\right) \leq \frac{1}{n} \sum_{k=1}^{n-1} p_i(k\Delta, x).$$

□

Lemma 7.5. *Suppose that σ^* is a constant. For all $q > 1$, there exists $K_q > 1$ such that for all $(t, x) \in (0, 1] \times [-A, A]$,*

$$\frac{1}{K_q\sqrt{t}} \exp\left(-\frac{2q-1}{2q\sigma^{*2}t}x^2\right) \leq p(t,x) \leq \frac{K_q}{\sqrt{t}} \exp\left(-\frac{x^2}{2q\sigma^{*2}t}\right).$$

Proof of Lemma 7.5. The transition density p^0 of the process $(0 + \sigma^* W_t)_{t\in]0,1]}$ (with a constant diffusion coefficient σ^*) is given by

$$p^0(t,x) := \frac{1}{\sqrt{2\pi\sigma^{*2}t}} \exp\left(-\frac{1}{2\sigma^{*2}t} |0-x|^2\right). \tag{21}$$

We are going to demonstrate the inequality for p_i, which is the transition density of X in class number i. Indeed, then it will be true for all $i \in \mathcal{Y}$ and thus for $p = p_Y$. We follow here the arguments given in the *proof of (1.6)* in Gobet (2002). Let us denote,

$$Z_{i,t} = \exp\left(\int_0^t \frac{b_i^*(X_s)}{\sigma^*} dW_s - \int_0^t \frac{b_i^{*2}(X_s)}{\sigma^{*2}} ds\right).$$

We have $\forall (t, x) \in]0, 1] \times \mathbb{R}$,

$$p_i(t,x) = p^0(t,x)\mathbb{E}^0\left[Z_{i,t} | X_t = x\right],$$

and

$$\frac{1}{p_i(t,x)} \leq \frac{1}{p^0(t,x)} \mathbb{E}^0\left[Z_{i,t}^{-1} | X_t = x\right]. \tag{22}$$

Then,

$$\mathbb{E}^0\left[Z_{i,t}|X_t=x\right]=1+\frac{1}{p^0(t,x)}\int_0^t\mathbb{E}^0\left[Z_{i,t}b_i^*(X_s)\frac{X_s-x}{\sigma^{*2}(t-s)}p^0(t-s,x)\right]ds$$

and

$$\mathbb{E}^0\left[Z_{i,t}^{-1}|X_t=x\right]=1+\frac{1}{p^0(t,x)}\int_0^t\mathbb{E}^0\left[Z_{i,s}^{-1}b_i^*(X_s)\frac{X_s-x}{\sigma^{*2}(t-s)}p^0(t-s,x)\right]ds.$$

For all $(t,x)\in]0,1]\times\mathbb{R}$, one has :

$$\begin{aligned}\mathbb{E}^0\left[Z_{i,t}|X_t=x\right]&=1+\frac{1}{p^0(t,x)}\int_0^t\mathbb{E}^0\left[Z_{i,s}b_i^*(X_s)\frac{X_s-x}{\sigma^{*2}(t-s)}p^0(t-s,x)\right]ds\\&\leq 1+\frac{C}{p^0(t,x)}\int_0^t\mathbb{E}^0\left[Z_{i,s}\left|b_i^*(X_s)\right|\frac{|X_s-x|}{(t-s)^{3/2}}\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]ds\\&\leq 1+\frac{C}{p^0(t,x)}\int_0^t\mathbb{E}^0\left[Z_{i,s}\left|b_i^*(X_s)\right|\frac{1}{\varepsilon(t-s)}\exp\left(-\frac{(1-\varepsilon)(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]ds\end{aligned}$$

using that $y\varepsilon\exp(-\varepsilon y^2/2)\leq 1$ for $0<\varepsilon<1$. Let $q,q'>1$ be two real numbers such that $\frac{1}{q}+\frac{1}{q'}=1$. Using Hölder's inequality, and the Lipschitz property of b^*, one has:

$$\mathbb{E}^0\left[Z_{i,t}|X_t=x\right]\leq 1+\frac{C\varepsilon^{-1}}{p^0(t,x)}\int_0^t\left(\mathbb{E}^0\left[\frac{Z_{i,s}q\left(1+|X_s|\right)^q}{(t-s)^q}\right]\right)^{\frac{1}{q}}\left(\mathbb{E}^0\left[\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]\right)^{\frac{1}{q'}}ds \quad (23)$$

with $\varepsilon=1-1/q'$. According to *Lemma A.1 in Gobet (2002)*, one has:

$$\forall q>1,\ \ \mathbb{E}^0\left[Z_{i,s}^q\left(1+|X_s|\right)^q\right]+\mathbb{E}^0\left[Z_{i,s}^{-q}\left(1+|X_s|\right)^q\right]\leq C_1$$

where $C_1>0$ is a constant. Thus, it remains to upper bound $\mathbb{E}^0\left[\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]$ and then deduce an upper bound of $\mathbb{E}^0\left[Z_{i,t}|X_t=x\right]$. For all $s<t$, we have:

$$\sqrt{2\pi\sigma^{*2}s}\mathbb{E}^0\left[\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]=\int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^{*2}(t-s)}\left(z-x\right)^2\right)\exp\left(-\frac{1}{2\sigma^{*2}s}z^2\right)dz$$

It follows that,

$$\mathbb{E}^0\left[\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]=\sqrt{\frac{t-s}{t}}\exp\left(-\frac{x^2}{2\sigma^{*2}t}\right).$$

Thus, from Equation (23), we obtain:

$$\begin{aligned}\mathbb{E}^0\left[Z_{i,t}|X_t=x\right]&\leq 1+\frac{C\varepsilon^{-1}}{p^0(t,x)}\int_0^t\frac{(t-s)^{\frac{1}{2q'}-1}}{t^{\frac{1}{2q'}}}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right)ds\\&\leq 1+\frac{C\varepsilon^{-1}}{p^0(t,x)\sqrt{t}}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right),\end{aligned}$$

by noticing that the integral of $(t-s)^{\frac{1}{2q'}-1}$ is smaller than 1 (since $0<s<t\leq 1$) and that $t^{-\frac{1}{2q'}}\geq 1/\sqrt{t}$. From the definition of function p^0 given in Equation (21) together with relation (22), we obtain that

$$p_i(t,x)\leq p^0(t,x)\left(1+\frac{C\varepsilon^{-1}}{p^0(t,x)\sqrt{t}}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right)\right).$$

Thus, there exists a constant $K_q > 1$ (as $\varepsilon = 1 - 1/q'$ and $1/q + 1/q' = 1$) such that,

$$\forall (t,x) \in]0,1] \times \mathbb{R}, \quad p_i(t,x) \leq \frac{K_q}{\sqrt{t}} \exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right), \quad \forall q' > 1. \tag{24}$$

Following the same lines, one has

$$\mathbb{E}^0\left[Z_{i,t}^{-1} | X_t = x\right] \leq 1 + \frac{C^{\text{te}}}{p^0(t,x)\sqrt{t}} \exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right).$$

Also, there exists a constant $K_q > 1$, such that,

$$\forall (t,x) \in]0,1] \times \mathbb{R} \quad p_i(t,x) \geq \frac{1}{K_q\sqrt{t}} \exp\left(-\frac{2q'-1}{2q'\sigma^{*2}t}x^2\right), \quad \forall q' > 1. \tag{25}$$

The final result is deduced from (24) and (25). □

7.2 Proofs of Section 3

Let us begin this section with a proposition which establishes a closed formula of the excess risk in multiclass classification.

Proposition 7.6. *Let g a classifier. The following holds*

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}\left[\sum_{i=1}^{K}\sum_{j\neq i} \left|\pi_i^*(X) - \pi_j^*(X)\right| \mathbb{1}_{\{g(X)=j, g^*(X)=i\}}\right]$$

The proof of this result is omitted and can be found instance in Denis *et al.* (2020). Now we provide the proof of Theorem 3.1 that relies in part on Proposition 7.6.

Proof of Theorem 3.1. From Proposition 7.6, we have the following inequality

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq 2\sum_{i=1}^{K} \mathbb{E}\left[|\widehat{\pi}_i(X) - \pi_i^*(X)|\right]. \tag{26}$$

We define $\bar{\mathbf{F}}$ the discretized version of $\mathbf{F}^*$,

$$\bar{\mathbf{F}} = (\bar{F}_1, \ldots, \bar{F}_K), \text{ with } \bar{F}_i(X) = \sum_{k=0}^{n-1}\left(\frac{b_i^*}{\sigma^{*2}}(X_{k\Delta})\left(X_{(k+1)\Delta} - X_{k\Delta}\right) - \frac{\Delta}{2}\frac{b_i^{*2}}{\sigma^{*2}}(X_{k\Delta})\right),$$

and for each $i \in \mathcal{Y}$, $\bar{\pi}_i^* = \phi_i\left(\bar{\mathbf{F}}\right)$ the discretized version of π_i^*, and $\bar{\pi}_i = \phi_i(\widehat{\boldsymbol{F}})$. From Equation (26), we deduce

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] &\leq 2\left(\sum_{i=1}^{K}\mathbb{E}\left[|\widehat{\pi}_i(X) - \bar{\pi}_i(X)|\right] + \mathbb{E}\left[|\bar{\pi}_i(X) - \bar{\pi}_i^*(X)|\right]\right. \\
&\quad \left. + \sum_{i=1}^{K}\mathbb{E}\left[|\bar{\pi}_i^*(X) - \pi_i^*(X)|\right]\right) \\
&\leq 2\sum_{i=1}^{K}\mathbb{E}\left[\left|\widehat{\phi}_i(\widehat{\boldsymbol{F}}(X)) - \phi_i(\widehat{\boldsymbol{F}}(X))\right|\right] + 2\sum_{i=1}^{K}\mathbb{E}\left[\left|\phi_i(\widehat{\boldsymbol{F}}(X)) - \phi_i(\bar{\mathbf{F}}(X))\right|\right] \\
&\quad + 2\sum_{i=1}^{K}\mathbb{E}\left[\left|\phi_i(\bar{\mathbf{F}}(X)) - \phi_i(\mathbf{F}^*(X))\right|\right].
\end{aligned} \tag{27}$$

For the first term of the *r.h.s.* of the above inequality, we observe that for $(x_1, \dots, x_K) \in \mathbb{R}^K$, and $(i,j) \in \mathcal{Y}^2$ we have

$$\left| \frac{\partial}{\partial_{\mathfrak{p}_j^*}} \frac{\mathfrak{p}_i^* \exp(x_i)}{\sum_{k=1}^K \mathfrak{p}_k^* \exp(x_k)} \right| \leq \frac{1}{\mathfrak{p}_0^*}.$$

Therefore,

$$\sum_{i=1}^K \mathbb{E}\left[\left| \widehat{\phi}_i(\widehat{\boldsymbol{F}}(X)) - \phi_i(\widehat{\boldsymbol{F}}(X)) \right| \right] \leq C_{K,\mathfrak{p}_0^*} \sum_{k=1}^K \mathbb{E}\left[\left| \widehat{\mathfrak{p}}_k - \mathfrak{p}_k \right| \right] \leq \frac{C_{K,\mathfrak{p}_0^*}}{\sqrt{N}}. \tag{28}$$

For the second term of Equation (27), since the softmax function is 1-Lipschitz, we have for $j \in \mathcal{Y}$

$$\mathbb{E}\left| \phi_j(\widehat{\boldsymbol{F}}(X)) - \phi_j(\bar{\mathbf{F}}(X)) \right| \leq \sum_{i=1}^K \mathbb{E}\left[\left| \widehat{F}_i(X) - \bar{F}_i(X) \right| \right].$$

We set $\xi(s) := k\Delta$, if $s \in [k\Delta, (k+1)\Delta)$, for $k \in [\![0, n-1]\!]$. We then deduce that

$$\begin{aligned} \left| \widehat{F}_i(X) - \bar{F}_i(X) \right| \leq \int_0^1 \left| \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{b_i^*}{\sigma^{*2}} \right) (X_{\xi(s)})\, b_Y^*(X_s) \right| ds + \frac{1}{2} \int_0^1 \left| \left(\frac{\widehat{b}_i^2}{\widehat{\sigma}^2} - \frac{b_i^{*2}}{\sigma^{*2}} \right) (X_{\xi(s)}) \right| ds \\ + \left| \int_0^1 \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{b_i^*}{\sigma^{*2}} \right) (X_{\xi(s)})\, \sigma^*(X_s) dW_s \right|, \end{aligned}$$

which implies

$$\begin{aligned} \mathbb{E}\left[\left| \widehat{F}_i(X) - \bar{F}_i(X) \right| \right] \;\leq\; & \mathbb{E}\left[\int_0^1 \left| \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{b_i^*}{\sigma^{*2}} \right) (X_{\xi(s)})\, b_Y^*(X_s) \right| ds \right] + \frac{1}{2} \mathbb{E}\left[\int_0^1 \left| \left(\frac{\widehat{b}_i^2}{\widehat{\sigma}^2} - \frac{b_i^{*2}}{\sigma^{*2}} \right) (X_{\xi(s)}) \right| ds \right] \\ & + \mathbb{E}\left[\int_0^1 \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{b_i^*}{\sigma^{*2}} \right)^2 (X_{\xi(s)})\, \sigma^{*2}(X_s) ds \right]. \end{aligned}$$

Since for all x, $\sigma^*(x) \geq \sigma_0^*$, and $\widehat{\sigma} \geq \sigma_0$, we get

$$\begin{cases} \left| \frac{\widehat{b}_i}{\widehat{\sigma}^2}(x) - \frac{b_i^*}{\sigma^{*2}}(x) \right| \leq \sigma_0^{-2} \left| \widehat{b}_i(x) - b_i^*(x) \right| + \sigma_0^{-2} \sigma_0^{*-2} \left| b_i^*(x) \right| \left| \widehat{\sigma}^2(x) - \sigma^{*2}(x) \right|, \\ \left| \frac{\widehat{b}_i^2}{\widehat{\sigma}^2}(x) - \frac{b_i^{*2}}{\sigma^{*2}}(x) \right| \leq \sigma_0^{-2} \left| \widehat{b}_i(x) + b_i^*(x) \right| \left| \widehat{b}_i(x) - b_i^*(x) \right| + \sigma_0^{-2} \sigma_0^{*-2} \left| b_i^*(x) \right|^2 \left| \widehat{\sigma}^2(x) - \sigma^{*2}(x) \right|. \end{cases} \tag{29}$$

Hence, as $\widehat{b_i}(x) \leq b_{\max}$, and $\mathbb{E}\left[\sup_{t \in [0,1]} |b_i^*(X_t)| \right] \leq C_1$, the above inequalities and the Cauchy-Schwarz inequality yield

$$\mathbb{E}\left| \widehat{F}_i(X) - \bar{F}_i(X) \right| \leq C_{\sigma_0^*} \sigma_0^{-2} \left(b_{\max} \mathbb{E}\left\| \widehat{b}_i - b_i^* \right\|_n + \mathbb{E}\left\| \widehat{\sigma}^2 - \sigma^2 \right\|_n \right).$$

Therefore, we have,

$$\sum_{i=1}^K \mathbb{E}\left[\left| \phi_i(\widehat{\boldsymbol{F}}(X)) - \phi_i(\bar{\mathbf{F}}(X)) \right| \right] \leq C_{K,\sigma_0^*} \sigma_0^{-2} \sum_{i=1}^K \left(b_{\max} \mathbb{E}\left\| \widehat{b}_i - b_i^* \right\|_n + \mathbb{E}\left\| \widehat{\sigma}^2 - \sigma^2 \right\|_n \right). \tag{30}$$

Finally, the last term is bounded as follows. We first observe that for all $i \in \mathcal{Y}$

$$\begin{aligned} \mathbb{E}\left[\left| \bar{F}_i(X) - F_i^*(X) \right|^2 \right] \leq\; & 3\mathbb{E} \int_0^1 \left(\frac{b_i^*\left(X_{\xi(s)}\right)}{\sigma^{*2}\left(X_{\xi(s)}\right)} - \frac{b_i^*\left(X_s\right)}{\sigma^{*2}(X_s)} \right)^2 b_Y^{*2}(X_s) ds \\ & + 3\mathbb{E} \int_0^1 \left(\frac{b_i^{*2}\left(X_{\xi(s)}\right)}{\sigma^{*2}\left(X_{\xi(s)}\right)} - \frac{b_i^{*2}\left(X_s\right)}{\sigma^{*2}(X_s)} \right)^2 ds + 3\mathbb{E} \int_0^1 \left(\frac{b_i^*\left(X_{\xi(s)}\right)}{\sigma^{*2}\left(X_{\xi(s)}\right)} - \frac{b_i^*\left(X_s\right)}{\sigma^{*2}(X_s)} \right)^2 \sigma^{*2}(X_s) ds. \end{aligned}$$

Using again that $\sigma^*(\cdot) \geq \sigma_0^*$, and $\mathbb{E}\left[\sup_{t\in[0,1]} |b_i^*(X_t)|^q\right] \leq C$ for $q \geq 1$ (by Assumption 2.1), the Cauchy-Schwarz inequality implies

$$\mathbb{E}\left[\left|\bar{F}_i(X) - F_i^*(X)\right|^2\right] \leq C_{\sigma_0^*}\left(\int_0^1 \mathbb{E}\left[\left|b_i^*\left(X_{\xi(s)}\right) - b_i^*(X_s)\right|^2\right] ds \right.$$
$$\left. + \int_0^1 \sqrt{\mathbb{E}\left[\left|b_i^*\left(X_{\xi(s)}\right) - b_i^*(X_s)\right|^4\right]} ds + \int_0^1 \sqrt{\mathbb{E}\left[\left|\sigma^{*2}\left(X_{\xi(s)}\right) - \sigma^{*2}(X_s)\right|^4\right]} ds\right).$$

Finally, since the functions b_i^*, and σ^* are Lipschitz, we deduce from Lemma 7.1 that

$$\mathbb{E}\left[\left|\bar{F}_i(X) - F_i^*(X)\right|^2\right] \leq C_{\sigma_0^*}\Delta,$$

which implies together with the fact that the sofmax function is 1-Lipschitz and the Jensen inequality that

$$\sum_{i=1}^K \mathbb{E}\left[\left|\phi_i(\bar{\mathbf{F}}(X)) - \phi_i(\mathbf{F}^*(X))\right|\right] \leq C_{K,\sigma_0^*}\sqrt{\Delta}. \tag{31}$$

In view of Equation 27, the combination of Equations (28), and (30), and (31) yields the desired result. □

Proof of Proposition 3.2. We consider h a L-Lipschitz function. We define the spline-approximation $\tilde{h}$ of h by

$$\tilde{h}(x) := \sum_{\ell=-M}^{K_N-1} h(u_\ell) B_\ell(x), \quad \forall x \in \mathbb{R}.$$

First, we note that $\tilde{h} \in \mathcal{S}_{K_N,M}$. Indeed, since h is L-Lipschitz, there exists $C_L > 0$ such that

$$|h(x)| \leq C_L|x| \leq C\log(N), \quad \forall x \in (-\log(N), \log(N)).$$

Therefore, for N large enough, we have

$$|h(x)| \leq \log(N)^{3/2}.$$

Then, we deduce

$$\sum_{\ell=-M}^{K_N-1} h^2(u_\ell) \leq (K_N + M)\log^3(N).$$

For $x \in (-\log(N), \log(N))$, there exists $0 \leq \ell_0 \leq K_N - 1$ such that $x \in [u_{\ell_0}, u_{\ell_0+1})$. We use the following property of the B-spline basis

$$B_\ell(x) = 0, \quad \text{if } \; x \notin [u_\ell, u_{\ell+M+1}), \quad \ell = -M, \ldots, K_N + M.$$

Hence, for $x \in [u_{\ell_0}, u_{\ell_0+1})$, we have $B_\ell(x) = 0$ for $\ell \leq \ell_0 - M - 1$, and $\ell \geq \ell_0 + M$. Thus,

$$\begin{aligned}
\left|\tilde{h}(x) - h(x)\right| &\leq \sum_{\ell=-M}^{K_N-1} |h(u_\ell) - h(x)|\, B_\ell(x) \\
&= \sum_{\ell=\ell_0-M}^{\ell_0} |h(u_\ell) - h(x)|\, B_l(x) \\
&\leq \max_{\ell=\ell_0-M,\ldots,\ell_0} |h(u_\ell) - h(x)| \\
&\leq L(u_{\ell_0+1} - u_{\ell_0-M}) \leq \frac{2L(M+1)\log(N)}{K_N},
\end{aligned}$$

which concludes the proof. □

Proof of Theorem 3.3. The proof is divided in two parts. The first part establishes the rates of convergence of the drift estimators, and the second part is devoted to the study of the rates of convergence of the diffusion coefficient estimator.

Rates of convergence for drift estimator. Let $i \in \{1, \dots, K\}$. We introduce the function,

$$\bar{b}_i := b_i^* \mathbb{1}_{(-\log(N), \log(N))}.$$

We recall that the random number of paths in the class number i is $N_i = \sum_{j=1}^{N} \mathbb{1}_{\{Y_j = i\}}$. For a function h, we introduce the empirical norm of class i for $N_i > 0$ as

$$\|h\|_{n,N_i}^2 := \frac{1}{nN_i} \sum_{j \in \mathcal{I}_j} \sum_{k=0}^{n-1} h^2(X_{k\Delta}^j)$$

We first observe that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n\right] = \mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n \mathbb{1}_{\{N_i > 0\}}\right] + \mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n \mathbb{1}_{\{N_i = 0\}}\right].$$

Let us work at first on the event $\{N_i > 0\}$. For all $i \in \mathcal{Y}$, we define the following conditional expectation

$$\mathbb{E}_i[.] = \mathbb{E}[.|\mathbb{1}_{\{Y_1 = i\}}, \dots, \mathbb{1}_{\{Y_N = i\}}].$$

We apply Proposition 3.2, and Proposition 3.2 of Denis *et al.* (2021) on the event $\{N_i > 0\}$ and deduce that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,N_i}^2\right] \leq C\left(\frac{\log^2(N)}{K_N^2} + \sqrt{\frac{K_N \log^3(N)}{N_i}} + \Delta\right). \tag{32}$$

Now, for all $i \in \mathcal{Y}$, let us write

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] = \mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,N_i}^2\right] + 2\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,N_i}^2\right]. \tag{33}$$

For $h \in \mathcal{S}_{K_N,M}$, we denote by $\bar{h}$ its thresholded counterpart

$$\bar{h}(\cdot) := h(\cdot)\mathbb{1}_{\{|h(\cdot)| \leq \log^{3/2}(N)\}} + \operatorname{sgn}(h(\cdot)) \log^{3/2}(N) \mathbb{1}_{\{|h(\cdot)| > \log^{3/2}(N)\}}.$$

We also denote $\mathcal{H}_{K_N,M} := \{\bar{h}, h \in \mathcal{S}_{K_N,M}\}$. Then, we have that

$$\begin{aligned}
\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,N_i}^2\right] &\leq \mathbb{E}_i\left[\sup_{\bar{h} \in \mathcal{H}_{K_N,M}} \left\|\bar{h} - \bar{b}_i\right\|_{n,i}^2 - 2\left\|\bar{h} - \bar{b}_i\right\|_{n,N_i}^2\right] \\
&\leq \mathbb{E}_i\left[\sup_{g \in \mathcal{G}_{K_N,M}} \mathbb{E}_{X|Y=i}\left[g(\bar{X}) - \frac{2}{N_i}\sum_{i \in \mathcal{I}} g(\bar{X}^i)\right]\right],
\end{aligned}$$

with $\mathcal{G}_{K_N,M} = \{(x_1, \dots, x_n) \mapsto \frac{1}{n}\sum_{k=1}^{n} \left|\bar{h}(x_k) - \bar{b}_i(x_k)\right|^2, \bar{h} \in \mathcal{H}_{K_N,M}\}$. For each $g \in \mathcal{G}_{K_N,M}$ and $x \in \mathbb{R}$, we have

$$0 \leq g(x) \leq 4\log^3(N).$$

Furthermore, we have that (see Denis *et al.*, 2021)

$$\mathcal{N}_\infty(\varepsilon, \mathcal{G}_{K_N,M}) \leq \left(\frac{12(K_N + M)\log^3(N)}{\varepsilon}\right)^{K_N+M}.$$

Therefore, we deduce from Lemma A.2 in Denis *et al.* (2021) with $\varepsilon = \dfrac{12(K_N + M)\log^3(N)}{N_i}$, Equation (32), and Equation (33), that on the event $\{N_i > 0\}$

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] \leq C\left(\frac{\log^2(N)}{K_N^2} + \sqrt{\frac{K_N \log^3(N)}{N_i}} + \frac{\log^4(N) K_N}{N_i} + \Delta\right). \tag{34}$$

Using Jensen's inequality, we have

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{\{N_i>0\}}\right] \leq \sqrt{\mathbb{E}\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2 \mathbb{1}_{\{N_i>0\}}\right] + \mathbb{E}\left[\left\|\bar{b}_i - b_i^*\right\|_{n,i}^2 \mathbb{1}_{\{N_i>0\}}\right]}.$$

Finally, let us study then the error $\left\|\bar{b}_i - b_i^*\right\|_{n,i}^2$. We observe with the Cauchy-Schwarz inequality

$$\begin{aligned}\left\|\bar{b}_i - b_i^*\right\|_{n,i}^2 &= \mathbb{E}_{X|Y=i}\left[\frac{1}{n}\sum_{k=1}^{n}\left(b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}|>\log(N)\}}\right] \\ &\leq C\sqrt{\sup_{t\in[0,1]} \mathbb{P}_{X|Y=i}\left(|X_t| \geq \log(N)\right)},\end{aligned}$$

since $\sup_{t\in[0,1]} \mathbb{E}\left[b_i^*(X_t)^4\right] \leq C$. From Lemma 7.3, we obtain

$$\left\|\bar{b}_i - b_i^*\right\|_{n,i}^2 \leq C\exp\left(-\frac{C_2}{2}\log^2(N)\right),$$

which for N large enough yields

$$\left\|\bar{b}_i - b_i^*\right\|_{n,i} \leq CN^{-1/2}.$$

This result leads us to obtain, from Equation (34), that

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{\{N_i>0\}}\right] \leq C\left(\frac{\log(N)}{K_N} + \sqrt{\frac{1}{N}} + \sqrt{\Delta}\right) \\ + C\left(\mathbb{E}\left[\left(\left(\frac{K_N\log^3(N)}{N_i}\right)^{1/4} + \sqrt{\frac{K_N\log^4(N)}{N_i}}\right)\mathbb{1}_{\{N_i>0\}}\right]\right).\end{aligned}$$

Using Jensen's inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{\{N_i>0\}}\right] \leq C\left(\frac{\log(N)}{K_N} + \sqrt{\frac{1}{N}} + \sqrt{\Delta}\right) \\ C\left(\left(K_N\log^3(N)\mathbb{E}\left[\frac{\mathbb{1}_{N_i>0}}{N_i}\right]\right)^{1/4} + \sqrt{K_N\log^4(N)}\sqrt{\mathbb{E}\left[\frac{\mathbb{1}_{N_i>0}}{N_i}\right]}\right).\end{aligned}$$

To finish the proof, since for all $i \in \mathcal{Y}$, $N_i \sim \mathcal{B}(N, \mathfrak{p}_i^*)$ we use Lemma 4.1 in (Györfi *et al.*, 2006) to deduce that

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_i>0}}{N_i}\right] \leq \frac{2}{\mathfrak{p}_i^* N} \leq \frac{2}{\mathfrak{p}_0^* N}$$

and finally, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{\{N_i>0\}}\right] \leq C\left(\frac{\log(N)}{K_N} + \left(\frac{K_N \log^3(N)}{N\mathfrak{p}_0^*}\right)^{1/4} + \sqrt{\Delta}\right). \tag{35}$$

To conclude the proof for the rates of convergence of the drift coefficient, we observe that since $\hat{b}_i$ is bounded by $\log^{3/2}(N)$ and $\sup_{t\in[0,1]} \mathbb{E}\left[b_i^*(X_t)^2\right] < +\infty$, we have

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{\{N_i=0\}}\right] \leq \log^{3/2}(N)\mathbb{P}\left(N_i = 0\right) + C\mathbb{P}\left(N_i = 0\right). \tag{36}$$

Since N_i is distributed according to a Binomial distribution with parameters $(N, \mathfrak{p}_i^*)$. We deduce that

$$\mathbb{P}\left(N_i = 0\right) = \exp\left(N\log(1-\mathfrak{p}_i^*)\right).$$

Hence, gathering Equation (35) and Equation (36), and choosing $\Delta = O(1/N)$ and $K_N = (N\log(N))^{1/5}$, we get the desired result.

Diffusion coefficient: rates of convergence. We estimate the square σ^{*2} of the diffusion coefficient as solution of the following regression model

$$\frac{(X^j_{(k+1)\Delta} - X^j_{k\Delta})^2}{\Delta} = \sigma^{*2}(X^j_{k\Delta}) + \zeta^j_{k\Delta} + R^j_{k\Delta} \tag{37}$$

where $\zeta^j_{k\Delta} := \zeta^{j,1}_{k\Delta} + \zeta^{j,2}_{k\Delta} + \zeta^{j,3}_{k\Delta}$ with

$$\begin{aligned}
\zeta^{j,1}_{k\Delta} &:= \frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta} \sigma^*\left(X^j_s\right) dW^j_s\right)^2 - \int_{k\Delta}^{(k+1)\Delta} \sigma^{*2}\left(X^j_s\right) ds\right] \\
\zeta^{j,2}_{k\Delta} &:= \frac{2}{\Delta}\int_{k\Delta}^{(k+1)\Delta} ((k+1)\Delta - s)\sigma^{*\prime}\left(X^j_s\right)\sigma^{*2}\left(X^j_s\right) dW^j_s \\
\zeta^{j,3}_{k\Delta} &:= 2b^*_Y\left(X^j_{k\Delta}\right)\int_{k\Delta}^{(k+1)\Delta} \sigma^*\left(X^j_s\right) dW^j_s,
\end{aligned}$$

and $R^j_{k\Delta} := R^{j,1}_{k\Delta} + R^{j,2}_{k\Delta} + R^{j,3}_{k\Delta}$ with,

$$R^{j,1}_{k\Delta} := \frac{1}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta} b^*_Y\left(X^j_s\right) ds\right)^2, \quad R^{j,2}_{k\Delta} := \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} ((k+1)\Delta - s)\phi_Y\left(X^j_s\right) ds \tag{38}$$

$$R^{j,3}_{k\Delta} := \frac{2}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta} \left(b^*_Y\left(X^j_s\right) - b^*_Y\left(X_{k\Delta}\right)\right) ds\right)\left(\int_{k\Delta}^{(k+1)\Delta} \sigma^*\left(X^j_s\right) dW^j_s\right) \tag{39}$$

where $\phi_Y := b^*_Y\sigma^{*\prime}\sigma^* + \left[\sigma^{*\prime\prime}\sigma^* + (\sigma^{*\prime})^2\right]\sigma^{*2}$. We prove in the sequel that $\zeta^{j,1}_{k\Delta}$ is the error term, and all the other terms are negligible residuals. We remind the reader that the estimator $\widehat{\sigma}^2$ of σ^{*2} is given in (10). We rely on the following result:

Lemma 7.7. *Under Assumption 2.1, the following holds*

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|^2_{n,N} \leq 3\inf_{h\in\mathcal{S}_{K_N,M}} \|h - \sigma^{*2}\|^2_n + C\left(\sqrt{\frac{K_N\log^3(N)}{Nn}} + \Delta_n^2\right)$$

where $C > 0$ is a constant depending on σ_1, and where

$$\|h\|_{n,N}^2 = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h^2(X_{k\Delta}^j).$$

The empirical error of the estimator $\widehat{\sigma}^2$ is given by

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n^2 = 2\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 + \left[\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n^2 - 2\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2\right]$$

Let us define $\mathcal{H}^\sigma$ as the set of functions $\bar{h}$ such that there exists a function $h \in \mathcal{S}_{K_N,M}$ satisfying

$$\bar{h} = h(x)\mathbb{1}_{\{\frac{1}{\log(N)} \le h(x) \le \log^{3/2}(N)\}} + \log^{3/2}(N)\mathbb{1}_{h(x)>\log^{3/2}(N)\}} + \frac{1}{\log(N)}\mathbb{1}_{\{h(x)\le\frac{1}{\log(N)}\}}.$$

Using then an $\varepsilon-$net $\mathcal{H}^{\sigma,\varepsilon}$ of $\mathcal{H}^\sigma$ with $\varepsilon = \frac{12(K_N+M)\log^3(N)}{N}$, we finally obtain (see Denis *et al.* (2021), Lemma A.2)

$$\begin{aligned}\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n^2 - 2\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 &\le \mathbb{E}\left[\sup_{\bar{h}\in\mathcal{H}^\sigma}\left\{\mathbb{E}\left\|\bar{h} - \sigma^{*2}\right\|_n^2 - 2\mathbb{E}\left\|\bar{h} - \sigma^{*2}\right\|_{n,N}^2\right\}\right] \\ &\le C\frac{K_N\log^4(N)}{N}.\end{aligned}$$

Thus, as $\Delta_n = \mathrm{O}(1/N)$,

$$\begin{aligned}\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n^2 &\le 3\inf_{h\in\mathcal{S}_{K_N,M}}\|h - \sigma^{*2}\|_n^2 + C\left(\frac{\sqrt{K_N\log^3(N)}}{N} + \frac{K_N\log^4(N)}{N} + \frac{1}{N^2}\right) \\ &\le 3\inf_{h\in\mathcal{S}_{K_N,M}}\|h - \sigma^{*2}\|_n^2 + C\frac{K_N\log^4(N)}{N},\end{aligned}$$

for N large enough. According to Proposition 3.2, the bias term satisfies

$$\inf_{h\in\mathcal{S}_{K_N,M}}\|h - \sigma^{*2}\|_n^2 \le C\frac{\log^2(N)}{K_N^2}.$$

Taking $K_N = (N\log(N))^{1/5}$ leads to

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n \le C_2\left(\frac{\log^4(N)}{N}\right)^{1/5}.$$

This concludes the proof of Theorem 3.3. □

Proof of Lemma 7.7 . Denote by

$$\gamma_{N,n}(h) = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(U_{k\Delta}^j - h(X_{k\Delta}^j)\right)^2,$$

the least square contrast appearing in (9). For all $h \in \mathcal{S}_{K_N,M}$, we deduce that

$$\gamma_{n,N}(\widehat{\sigma}^2) - \gamma_{n,N}(\sigma^{*2}) \le \gamma_{n,N}(h) - \gamma_{n,N}(\sigma^{*2}). \tag{40}$$

Using (37), we have for all $h \in \mathcal{S}_{K_N,M}$,

$$\gamma_{n,N}(h) - \gamma_{n,N}(\sigma^{*2}) = \left\|h - \sigma^{*2}\right\|_{n,N}^2 + 2\nu_1(\sigma^{*2} - h) + 2\nu_2(\sigma^{*2} - h) + 2\nu_3(\sigma^{*2} - h) + 2\mu(\sigma^{*2} - h) \quad (41)$$

where

$$\nu_i(h) = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h(X_{k\Delta}^j)\zeta_{k\Delta}^{j,i}, \quad i \in \{1,2,3\}, \quad \mu(h) = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h(X_{k\Delta}^j)R_{k\Delta}^j, \quad (42)$$

we derive from Equations (40) and (41) that for all $h \in \mathcal{S}_{K_N,M}$,

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 \leq \inf_{h\in\mathcal{S}_{K_N,M}} \|h - \sigma^{*2}\|_n^2 + 2\sum_{i=1}^{3}\mathbb{E}\left[\nu_i(\widehat{\sigma}^2 - h)\right] + 2\mathbb{E}\left[\mu(\widehat{\sigma}^2 - h)\right]. \quad (43)$$

For all $i \in \{1,2,3\}$ and for all $h \in \mathcal{S}_{K_N,M}$, taking the constraints (6) into account, one has

$$\mathbb{E}\left[\nu_i\left(\widehat{\sigma}^2 - h\right)\right] \leq \sqrt{2(K_N + M)\log^3(N)}\sqrt{\sum_{\ell=-M}^{K_N-1}\mathbb{E}\left[\nu_i^2(B_{\ell,M,\mathbf{u}})\right]}. \quad (44)$$

1. Upper bound of $\sum_{\ell=-M}^{K_N-1}\mathbb{E}\left[\nu_1^2(B_{\ell,M,\mathbf{u}})\right]$. According to Equation (42), we have

$$\forall \ell \in [\![-M, K_N - 1]\!],\ \nu_1(B_{\ell,M,\mathbf{u}}) = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1} B_{\ell,M,\mathbf{u}}(X_{k\Delta}^j)\zeta_{k\Delta}^{j,1}$$

where $\zeta_{k\Delta}^{j,1} = \frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X_s^j)dW_s^j\right)^2 - \int_{k\Delta}^{(k+1)\Delta}\sigma^{*2}(X_s^j)ds\right]$ is a martingale satisfying

$$\mathbb{E}\left[\zeta_{k\Delta}^{1,1}|\mathcal{F}_{k\Delta}^1\right] = 0 \ \text{ and } \ \mathbb{E}\left[\left(\zeta_{k\Delta}^{1,1}\right)^2|\mathcal{F}_{k\Delta}^1\right] \leq \frac{1}{\Delta^2}\mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^{*2}(X_s^1)ds\right)^2\right] \leq C\sigma_1^{*4}$$

with $(\mathcal{F}_t^1)_{t\geq 0}$ the natural filtration associated with the Brownian motion W^1. We derive that

$$\begin{aligned}\sum_{\ell=-M}^{K_N-1}\mathbb{E}\left[\nu_1^2(B_{\ell,M,\mathbf{u}})\right] =& \frac{1}{Nn^2}\sum_{\ell=-M}^{K_N-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1} B_{\ell,M,\mathbf{u}}(X_{k\Delta}^j)\zeta_{k\Delta}^{1,1}\right)^2\right] \\ =& \frac{1}{Nn^2}\mathbb{E}\left[\sum_{k=0}^{n-1}\sum_{\ell=-M}^{K_N-1} B_{\ell,M,\mathbf{u}}^2(X_{k\Delta}^1)\left(\zeta_{k\Delta}^{1,1}\right)^2\right] \\ \leq& \frac{C}{Nn}\end{aligned}$$

where C is a constant depending on σ^*, for each $k \in [\![0, n-1]\!]$, $\sum_{\ell=-M}^{K_N-1} B_{\ell,M,\mathbf{u}}^2(X_{k\Delta}^1) \leq 1$ since $\sum_{\ell=-M}^{K_N-1} B_{\ell,M,\mathbf{u}}(X_{k\Delta}^1) = 1$ and $B_{\ell,M,\mathbf{u}}(X_{k\Delta}^1) \leq 1$ for all $\ell = -M, \cdots, K_N - 1$.

2. Upper bound of $\sum_{\ell=-M}^{K_N-1}\mathbb{E}\left[\nu_2^2(B_{\ell,M,\mathbf{u}})\right]$. For all $k \in [\![0, n-1]\!]$ and for all $s \in [0,1]$, set $\xi(s) = k\Delta$

if $s \in [k\Delta, (k+1)\Delta)$. We have:

$$\begin{aligned}
&\sum_{\ell=-M}^{K_N-1} \mathbb{E}\left[\nu_2^2(B_{\ell,M,\mathbf{u}})\right] \\
&\qquad = \frac{4}{Nn^2} \sum_{\ell=-M}^{K_N-1} \mathbb{E}\left[\left(\sum_{k=0}^{n-1} \int_{k\Delta}^{(k+1)\Delta} B_{\ell,M,\mathbf{u}}(X^1_{k\Delta})((k+1)\Delta - s)\sigma^{*\prime}(X^1_s)\sigma^{*2}(X^1_s)dW_s\right)^2\right] \\
&\qquad = \frac{4}{Nn^2} \sum_{\ell=-M}^{K_N-1} \mathbb{E}\left[\left(\int_0^1 B_{\ell,M,\mathbf{u}}(X^1_{\xi(s)})(\xi(s) + \Delta - s)\sigma^{*\prime}(X^1_s)\sigma^{*2}(X^1_s)dW_s\right)^2\right] \\
&\qquad \leq \frac{C}{Nn^2}
\end{aligned}$$

where the constant $C > 0$ depends on the diffusion coefficient.

3. Upper bound of $\sum_{\ell=-M}^{K_N-1} \mathbb{E}\left[\nu_3^2(B_{\ell,M,\mathbf{u}})\right]$. We have:

$$\begin{aligned}
\sum_{\ell=-M}^{K_N-1} \mathbb{E}\left[\nu_3^2(B_{\ell,M,\mathbf{u}})\right] &= \frac{4}{Nn^2} \sum_{\ell=-M}^{K_N-1} \mathbb{E}\left[\left(\sum_{k=0}^{n-1} \int_{k\Delta}^{(k+1)\Delta} B_{\ell,M,\mathbf{u}}(X^1_{k\Delta}) b^*_Y(X^1_{k\Delta})\sigma^*(X^1_s)dW_s\right)^2\right] \\
&= \frac{4}{Nn^2} \sum_{\ell=-M}^{K_N-1} \mathbb{E}\left[\left(\int_0^1 B_{\ell,M,\mathbf{u}}(X^1_{\eta(s)}) b^*_Y(X^1_{\eta(s)})\sigma^*(X^1_s)dW_s\right)^2\right] \\
&\leq \frac{4}{Nn^2} \mathbb{E}\left[\int_0^1 \sum_{\ell=-M}^{K_N-1} B^2_{\ell,M,\mathbf{u}}(X^1_{\eta(s)}) b^{*2}_Y(X^1_{\eta(s)})\sigma^{*2}(X^1_s)ds\right].
\end{aligned}$$

Since for all $x \in \mathbb{R}$, $b^{*2}_Y(x) \leq C_0(1+x^2)$, $\sigma^{*2}(x) \leq \sigma^{*2}_1$ and $\sup_{t\in[0,1]} \mathbb{E}\left(|X_t|^2\right) < \infty$, there exists a constant $C > 0$ depending on the upper bound σ^*_1 of the diffusion coefficient such that

$$\sum_{\ell=-M}^{K_N-1} \mathbb{E}\left[\nu_3^2(B_{\ell,M,\mathbf{u}})\right] \leq \frac{C}{Nn^2}.$$

We finally deduce from Equations (43) and (44) that for all $h \in \mathcal{S}_{K_N,M}$,

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|^2_{n,N} \leq \inf_{h\in\mathcal{S}_{K_N,M}} \|h - \sigma^{*2}\|^2_n + C\sqrt{\frac{(K_N+M)\log^3(N)}{Nn}} + 2\mathbb{E}\left[\mu(\widehat{\sigma}^2 - h)\right]. \tag{45}$$

It remains to obtain an upper bound of the term $\mu(\widehat{\sigma}^2 - h)$. Notice that for $a > 0$, x and $y \in \mathbb{R}$,

$$2xy = 2\frac{x}{\sqrt{a}} \times \sqrt{a}y \leq \frac{x^2}{a} + ay^2.$$

Then, for all $h \in \mathcal{S}_{K_N,M}$ and $a > 0$,

$$2\mu\left(\widehat{\sigma}^2 - h\right) \leq \frac{2}{a}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|^2_{n,N} + \frac{2}{a}\left\|h - \sigma^{*2}\right\|^2_{n,N} + \frac{a}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(R^j_{k\Delta}\right)^2.$$

We set $a = 4$ and from Equation (45) we deduce that,

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 \leq 3 \inf_{h \in \mathcal{S}_{K_N,M}} \|h - \sigma^{*2}\|_n^2 + C\sqrt{\frac{(K_N + M)\log^3(N)}{Nn}} + \frac{4}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(R_{k\Delta}^j\right)^2\right]. \quad (46)$$

We have

$$\mathbb{E}\left[\left(R_{k\Delta}^j\right)^2\right] \leq 3\left(\mathbb{E}\left[\left(R_{k\Delta}^{j,1}\right)^2\right] + \mathbb{E}\left[\left(R_{k\Delta}^{j,2}\right)^2\right] + \mathbb{E}\left[\left(R_{k\Delta}^{j,3}\right)^2\right]\right)$$

where for all $j \in [\![1, N]\!]$ and $k \in [\![0, n-1]\!]$, $R_{k\Delta}^{j,1}, R_{k\Delta}^{j,2}$ and $R_{k\Delta}^{j,3}$ are given in Equations (38) and (39). There exist constants $C_1, C_2, C_3 > 0$ such that

$$\begin{aligned}
\mathbb{E}\left[\left(R_{k\Delta}^{j,1}\right)^2\right] &\leq \mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta} b_Y^{*2}\left(X_{k\Delta}^j\right) ds\right)^2\right] \leq \Delta\mathbb{E}\left[\int_{k\Delta}^{(k+1)\Delta} b_Y^{*4}\left(X_{k\Delta}^j\right) ds\right] \leq C_1\Delta^2 \\
\mathbb{E}\left[\left(R_{k\Delta}^{j,2}\right)^2\right] &\leq \frac{1}{\Delta^2}\int_{k\Delta}^{(k+1)\Delta}((k+1)\Delta - s)^2 ds \int_{k\Delta}^{(k+1)\Delta}\mathbb{E}\left[\phi_Y^2\left(X_s^j\right)\right] ds \leq C_2\Delta^2 \\
\mathbb{E}\left[\left(R_{k\Delta}^{j,3}\right)^2\right] &\leq \frac{4}{\Delta^2}\mathbb{E}\left[\Delta\int_{k\Delta}^{(k+1)\Delta} L_0^2\left|X_s^j - X_{k\Delta}^j\right|^2 ds\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X_s^j)dW_s\right)^2\right] \leq C_3\Delta^2.
\end{aligned}$$

We deduce from Equation (46) that there exists a constant $C > 0$ depending on σ_1^* and M such that,

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 \leq 3 \inf_{h \in \mathcal{S}_{K_N,M}} \|h - \sigma^{*2}\|_n^2 + C\left(\sqrt{\frac{K_N\log^3(N)}{Nn}} + \Delta_n^2\right).$$

This is the announced result. □

Proof of Theorem 3.4. Let $A > 0$ and $i \in \mathcal{Y}$, we have that

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n^2\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}|\leq A\}}\right] \\
&\quad + \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}|>A\}}\right]. \quad (47)
\end{aligned}$$

We bound each term of the *r.h.s.* of the above inequality. From Lemma 7.3, and Cauchy-Schwarz Inequality, under Assumption 2.1, we have for the second term of (47),

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}|>A\}}\right] \leq C\sqrt{\exp\left(-CA^2\right)}. \quad (48)$$

For the first term of (47), we observe that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}|\leq A\}}|\mathcal{D}_N\right] &= \int_{-A}^{A}\left(\widehat{b}_i(x) - b_i^*(x)\right)^2\left(\frac{1}{n}\sum_{k=1}^{n-1}p(k\Delta, x)\right)\mathrm{d}x \\
&\quad + \frac{1}{n}\left(\widehat{b}_i(0) - b_i^*(0)\right)^2.
\end{aligned}$$

From Lemma 7.4, we then deduce that

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta})-b_i^*(X_{k\Delta})\right)^2\mathbb{1}_{\{|X_{k\Delta}|\leq A\}}|\mathcal{D}_N\right]$$
$$\leq C_1 e^{C_2A^2}\mathbb{E}_{X|Y=i}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta})-b_i^*(X_{k\Delta})\right)^2\mathbb{1}_{\{|X_{k\Delta}|\leq A\}}|\mathcal{D}_N\right]$$
$$\leq C_1 e^{C_2A^2}\left\|\widehat{b}_i-b_i^*\right\|_{n,i}^2.$$

From the above key equation, Equation (47), Equation (48), and Theorem 3.3, we deduce, for $A>0$

$$\mathbb{E}\left[\left\|\widehat{b}_i-b_i^*\right\|_n^2\right]\leq C\left(\exp\left(C_2A^2\right)\left(\frac{\log(N)^4}{N}\right)^{1/5}+\exp(-C_2A^2)\right).$$

Besides from Theorem 3.3, we also have

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2-\sigma^{2*}\right\|_n\right]\leq\left(\frac{\log(N)^4}{N}\right)^{1/5}.$$

Therefore, applying Theorem 3.1 with $b_{\max}=\log(N)^{3/2}$, $\sigma_0^{-2}=\log(N)$, and $A=(\log(N))^{1/4}$, we get the desired result. □

7.3 Proofs of Section 4

Proof of Proposition 4.2 . For all $i\in\mathcal{Y}$, let $\mathbb{P}_i=\mathbb{P}(.|Y=i)$ and denote by $\mathbb{P}_0$ the probability measure under which the diffusion process $X=(X_t)_{t\geq0}$ is solution of $dX_t=d\widetilde{W}_t$ where $\widetilde{W}$ is a Brownian motion under $\mathbb{P}_0$. We deduce from the Girsanov's Theorem (see e.g. Jacod & Shiryaev (2013), Chapter III) that

$$\forall i\in\mathcal{Y},\ \forall t\in[0,1],\quad \frac{d\mathbb{P}_i}{d\mathbb{P}_0}|_{\mathcal{F}_t^X}=\exp\left(\int_0^t b_i^*(X_s)dX_s-\frac{1}{2}\int_0^t b_i^{*2}(X_s)ds\right),$$

where $(\mathcal{F}_t^X)_{t\in[0,1]}$ is the natural filtration of X. Then, for all $i,j\in\mathcal{Y}$ such that $i\neq j$,

$$\forall t\in[0,1],\quad \frac{d\mathbb{P}_i}{d\mathbb{P}_j}|_{\mathcal{F}_t^X}=\exp\left(\int_0^t(b_i^*-b_j^*)(X_s)dX_s-\frac{1}{2}\int_0^t(b_i^{*2}-b_j^{*2})(X_s)ds\right)\leq C\exp\left(M_t^{i,j}\right)\quad(49)$$

where the constant C depends on C_{b^*} given in Assumption 4.1 and

$$\forall i,j\in\mathcal{Y}:i\neq j,\quad M_t^{i,j}=\int_0^t(b_i^*-b_j^*)(X_s)dW_s,\quad t\in[0,1].$$

Recall that $Z\in[0,\log^\alpha(N)]$ is a random variable measurable with respect to the natural filtration of the diffusion process $X=(X_t)_{t\leq1}$. Then, for all $i,j\in\mathcal{Y}$ such that $i\neq j$ and for all $a>0$, using Equation (49) we have

$$\mathbb{E}_{X|Y=i}[Z]=\mathbb{E}_{X|Y=j}\left[Z\frac{d\mathbb{P}_i}{d\mathbb{P}_j}|_{\mathcal{F}_t^X}\right]\leq C\mathbb{E}_{X|Y=j}\left[Z\exp\left(M_t^{i,j}\right)\right]$$
$$=C\mathbb{E}_{X|Y=j}\left[Z\exp\left(M_t^{i,j}\right)\mathbb{1}_{M_t^{i,j}\leq a}\right]+C\mathbb{E}_{X|Y=j}\left[Z\exp\left(M_t^{i,j}\right)\mathbb{1}_{M_t^{i,j}>a}\right]$$
$$\leq C\exp(a)\,\mathbb{E}_{X|Y=j}[Z]+C\log^\alpha(N)\mathbb{E}_{X|Y=j}\left[\exp\left(M_t^{i,j}\right)\mathbb{1}_{M_t^{i,j}>a}\right].$$

Using the Cauchy-Schwarz inequality and *Lemma 2.1 in Van-de Geer (1995)*, there exist constants $C > 0$ and $c > 0$ depending on $C_{\mathbf{b}^*}$ such that,

$$\mathbb{E}\left[\exp\left(M_t^{i,j}\right)\mathbb{1}_{M_t^{i,j}>a}\right] \leq \sqrt{\mathbb{P}\left(M_t^{i,j} > a\right)}\sqrt{\mathbb{E}\left[\exp\left(2M_t^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\exp\left(2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\right]}$$
$$\leq C\exp(-a^2/c)\sqrt{\mathbb{E}\left[\exp\left(2M_t^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\right]}$$

where $\mathbb{P}\left(M_t^{i,j} > a\right) \leq \exp(-a^2/c)$ (Van-de Geer (1995)) and $\exp\left(2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right) < \infty$ $a.s$ since the drift functions are bounded. Moreover, since $(M_t^{i,j})_{t\leq 1}$ is a martingale and

$$\mathbb{E}\left[\exp\left(\left\langle M^{i,j}, M^{i,j}\right\rangle_1\right)\right] < \infty,$$

according to the Novikov assumption, thus $\mathcal{E}(M^{i,j}) := \left\{\exp\left(2M_t^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\right\}_{t\leq 1}$ is a martingale with respect to the natural filtration $\mathcal{F}^M$ of $M^{i,j}$ (see Le Gall (2013), Proposition 5.8 and Theorem 5.9). We deduce that for all $t \in [0,1]$,

$$\mathbb{E}\left[\exp\left(2M_t^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\right] = \mathbb{E}\left[\mathbb{E}\left(\mathcal{E}(M^{i,j})_t|\mathcal{F}_0^M\right)\right] = \mathbb{E}\left[\exp\left(2M_0^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_0\right)\right] = 1.$$

Thus, for all $a > 0$, we obtain $\mathbb{E}\left[\exp\left(M_t^{i,j}\right)\mathbb{1}_{M_t^{i,j}>a}\right] \leq C\exp(-a^2/c)$. Finally, set $a = \sqrt{c\log(N)}$, it follows that for all $i, j \in \mathcal{Y}$ such that $i \neq j$, there exists a constant $C > 0$ such that

$$\mathbb{E}_{X|Y=i}[Z] \leq C\exp\left(\sqrt{c\log(N)}\right)\mathbb{E}_{X|Y=j}[Z] + C\frac{\log^\alpha(N)}{N}.$$

□

Proof of Theorem 4.3 . From Theorem 3.1, and its assumptions, we have

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\sqrt{\Delta} + \frac{1}{\mathfrak{p}_0^*\sqrt{N}} + \mathbb{E}\left[b_{\max}\sigma_0^{-2}\sum_{i=1}^K \|\widehat{b}_i - b_i^*\|_n\right] + \mathbb{E}\left[\sigma_0^{-2}\|\widehat{\sigma}^2 - \sigma^{2*}\|_n\right]\right).$$

For all $i \in \mathcal{Y}$ we obtain from Proposition 4.2 with $\alpha = 1/2$ that there exist constants $C_1, c > 0$ such that

$$\mathbb{E}\left\|\widehat{b}_i - b_i^*\right\|_n = \sum_{j=1}^K \mathfrak{p}_j^*\mathbb{E}\left\|\widehat{b}_i - b_i^*\right\|_{n,j} \leq C_1\exp\left(\sqrt{c\log(N)}\right)\mathbb{E}\left\|\widehat{b}_i - b_i^*\right\|_{n,i} + C_1\frac{\sqrt{\log(N)}}{N}.$$

Then, from Theorem 3.3, for $\Delta_n = \mathrm{O}(N^{-1})$ and $K_N = (N\log(N))^{1/5}$, there exist constants $C_2, C_3 > 0$ such that

$$\forall i \in \mathcal{Y},\ \mathbb{E}\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \leq C_2\left(\frac{\log^4(N)}{N}\right)^{1/5},\ \text{ and }\ \mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n \leq C_3\left(\frac{\log^4(N)}{N}\right)^{1/5}.$$

Finally, by (12), we deduce that there exist constants $C, c > 0$ such that

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\exp\left(\sqrt{c\log(N)}\right)N^{-1/5}.$$

□

Let us now turn to the proof of Theorem 4.6. For fixed n and N_i in $\mathbb{N}^*$, let us denote

$$\Omega_{n,N_i,K_{N_i}} := \bigcap_{h\in\mathcal{S}_{K_{N_i},M}\setminus\{0\}} \left\{ \left| \frac{\|h\|^2_{n,N_i}}{\|h\|^2_{n,i}} - 1 \right| \leq \frac{1}{2} \right\}.$$

As we can see, the empirical norms $\|h\|_{n,N_i}$ and $\|h\|_{n,i}$ of any function $h \in \mathcal{S}_{K_{N_i},M} \setminus \{0\}$ are equivalent on $\Omega_{n,N_i,K_{N_i}}$. More precisely, on the set $\Omega_{n,N_i,K_{N_i}}$, for all $h \in \mathcal{S}_{K_{N_i},M} \setminus \{0\}$, we have

$$\frac{1}{2}\|h\|^2_{n,i} \leq \|h\|^2_{n,N_i} \leq \frac{3}{2}\|h\|^2_{n,i}.$$

We have the following lemma.

Lemma 7.8. *Let $\beta \geq 1$ be a real number and suppose that $K_{N_i} = \mathrm{O}\left(\log^{-5/2}(N_i) N_i^{1/(2\beta+1)}\right)$ with N_i a.s large enough, and $A_{N_i} = \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$. Under Assumption 2.1, the following holds:*

$$\mathbb{P}_i\left(\Omega^c_{n,N_i,K_{N_i}}\right) \leq c\frac{K_{N_i}}{N_i}$$

where $c > 0$ is a constant.

Proof of Theorem 4.6. Note that throughout the proof we work conditional on the random variables $\mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i}$ and on the event $\{N_i > 1\}$, so that N_i can be viewed as a deterministic variable. Then, to alleviate the notations, let use denote

$$\mathbb{P}_i := \mathbb{P}(.|\mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i}) \text{ and } \mathbb{E}_i = \mathbb{E}[.|\mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i}].$$

For each class $i \in \mathcal{Y}$, the drift function b^*_i is the solution of the following regression model

$$Z^j_{k\Delta} = b^*_i(X^j_{k\Delta}) + \xi^j_{k\Delta} + R^j_{k\Delta}, \quad j \in \mathcal{I}_i, \quad k \in [\![0, n-1]\!]$$

where we recall that $\mathcal{I}_i$ is the set of indices j such that $Y_j = i$, and

$$\xi^j_{k\Delta} := \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} \sigma^*(X^j_s)dW^j_s, \qquad R^j_{k\Delta} := \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} (b^*_i(X^j_s) - b^*_i(X^j_{k\Delta}))ds. \tag{50}$$

We first focus on the error $\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right]$ for each label $i \in \mathcal{Y}$. Therefore, we consider the following decomposition:

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] = \mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\mathbb{1}_{\Lambda_i}\right] + \mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\mathbb{1}_{\Lambda'_i}\right] \tag{51}$$

where

$$\Lambda_i = \Omega_{n,N_i,K_{N_i}} \text{ and } \Lambda'_i = \Omega^c_{n,N_i,K_{N_i}}.$$

Upper bound of $\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\mathbb{1}_{\Lambda_i}\right]$. From the proof of Proposition 4.4 in Denis *et al.* (2021), Equation (D.5), we see that for all $h \in \mathcal{S}_{K_{N_i},M}$ and for all $a, d > 0$, we have on the event $\Lambda_i = \Omega_{n,N_i,K_{N_i}}$,

$$\left(1 - \frac{2}{a} - \frac{4}{d}\right)\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i} \leq \left(1 + \frac{2}{a} + \frac{4}{d}\right)\left\|h - b^*_{A_{N_i},i}\right\|^2_{n,N_i} + d\sup_{\left\{h\in\mathcal{S}_{K_{N_i},M},\|h\|_{n,i}=1\right\}} \nu^2(t) + aC\Delta$$

where $C > 0$ is a constant and where for all $h \in \mathcal{S}_{K_{N_i},M}$,

$$\nu(h) = \frac{1}{N_i n} \sum_{j \in I_i} \sum_{k=0}^{n-1} h(X_{k\Delta}^j) \xi_{k\Delta}^j. \tag{52}$$

We set $a = d = 8$, and we obtain,

$$\mathbb{E}_i \left[\left\| \widehat{b}_i - b^*_{A_{N_i},i} \right\|^2_{n,N_i} \mathbb{1}_{\Lambda_i} \right] \leq 7 \inf_{h \in \mathcal{S}_{K_{N_i},M}} \left\| h - b^*_{A_{N_i},i} \right\|^2_{n,i} + 32 \mathbb{E}_i \left[\sup_{\left\{ h \in \mathcal{S}_{K_{N_i},M}, \|h\|_{n,i}=1 \right\}} \nu^2(h) \right] + 32 C \Delta.$$

For $h \in \mathcal{S}_{K_{N_i},M}$, $h = \sum_{\ell=-M}^{K_{N_i}-1} w_\ell B_{\ell,M,\mathbf{u}}$ and $\|h\|^2_{n,i} = w' \Psi^i_{K_{N_i}} w$ equals to one here, then $w = \Psi^{-1/2}_{K_{N_i}} u$ where the vector u satisfies $\|u\|_{2,K_{N_i}+M} = 1$. Finally, one obtains,

$$h = \sum_{\ell=-M}^{K_{N_i}-1} w_\ell B_{\ell,M,\mathbf{u}} = \sum_{\ell=-M}^{K_{N_i}-1} u_\ell \left(\sum_{\ell'=-M}^{K_{N_i}-1} \left[\Psi^{-1/2}_{K_{N_i}} \right]_{\ell',\ell} B_{\ell',M,\mathbf{u}} \right). \tag{53}$$

For all $h \in \mathcal{S}_{K_{N_i},M}$ such that $\|h\|_{n,i} = 1$, using Equation (52) and (53), gives

$$\nu^2(h) = \left(\sum_{\ell=-M}^{K_{N_i}-1} u_\ell \frac{1}{N_i n} \sum_{j=1}^{N_i} \sum_{k=0}^{n-1} \sum_{\ell'=-M}^{K_{N_i}-1} \left[\Psi^{-1/2}_{K_{N_i}} \right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}(X^{i_j}_{k\Delta}) \xi^{i_j}_{k\Delta} \right)^2.$$

Cauchy-Schwarz inequality together with $\|u\|_2 = 1$, produce

$$\nu^2(h) \leq \sum_{\ell=-M}^{K_{N_i}-1} \left(\frac{1}{N_i n} \sum_{j=1}^{N_i} \sum_{k=0}^{n-1} \sum_{\ell'=-M}^{K_{N_i}-1} \left[\Psi^{-1/2}_{K_{N_i}} \right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}(X^{i_j}_{k\Delta}) \xi^{i_j}_{k\Delta} \right)^2.$$

Finally we obtain,

$$\begin{aligned} \mathbb{E}_i \left[\sup_{h \in \mathcal{S}_{K_{N_i},M}, \|h\|_{n,i}=1} \nu^2(h) \right] &\leq \frac{1}{N_i} \mathbb{E}_i \left[\frac{1}{n^2} \sum_{\ell=-M}^{K_{N_i}-1} \left(\sum_{k=0}^{n-1} \sum_{\ell'=-M}^{K_{N_i}-1} \left[\Psi^{-1/2}_{K_{N_i}} \right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}(X^{i_1}_{k\Delta}) \xi^{i_1}_{k\Delta} \right)^2 \right] \\ &= \frac{1}{N_i} \mathbb{E}_i \left[\frac{1}{n^2} \sum_{\ell=-M}^{K_{N_i}-1} \sum_{k=0}^{n-1} \left(\sum_{\ell'=-M}^{K_{N_i}-1} \left[\Psi^{-1/2}_{K_{N_i}} \right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}(X^{i_1}_{k\Delta}) \right)^2 \left(\xi^{i_1}_{k\Delta} \right)^2 \right]. \end{aligned}$$

According to Equation (50) and considering the natural filtration $(\mathcal{F}_t)_{t \geq 0}$ of the Brownian motion, for all $k \in [\![0, n-1]\!]$, we have $\mathbb{E}_i \left(\xi^{i_1}_{k\Delta} | \mathcal{F}_{k\Delta} \right) = 0$ and

$$\mathbb{E}_i \left[\left(\xi^{i_1}_{k\Delta} \right)^2 | \mathcal{F}_{k\Delta} \right] = \frac{1}{\Delta^2} \mathbb{E} \left[\sigma^{*2} \left(X^{i_1}_{k\Delta} \right) \mathbb{E} \left(\left(\int_{k\Delta}^{(k+1)\Delta} \sigma^*(X^{i_1}_s) \right)^2 | \mathcal{F}_{k\Delta} \right) \right] \leq \frac{\sigma^{*2}_1}{\Delta}.$$

By definition of the Gram matrix $\Psi_{K_{N_i}}$, we deduce that

$$\begin{aligned}\mathbb{E}_i\left[\sup_{h\in\mathcal{S}_{K_{N_i},M},\|h\|_{n,i}=1}\nu^2(h)\right] \leq & \frac{\sigma_1^{*2}}{N_i}\mathbb{E}_i\left[\frac{1}{n}\sum_{\ell=-M}^{K_{N_i}-1}\sum_{k=0}^{n-1}\left(\sum_{\ell'=-M}^{K_{N_i}-1}\left[\Psi_{K_{N_i}}^{-1/2}\right]_{\ell',\ell}B_{\ell',M,\mathbf{u}}(X_{k\Delta}^{1,i})\right)^2\right]\\ \leq & \frac{\sigma_1^{*2}}{N_i}\mathbb{E}_i\left(\sum_{\ell,\ell',\ell''=-M}^{K_{N_i}-1}\left[\Psi_{K_{N_i}}^{-1/2}\right]_{\ell',\ell}\left[\Psi_{K_{N_i}}^{-1/2}\right]_{\ell'',\ell}\left[\Psi_{K_{N_i}}\right]_{\ell',\ell''}\right)\\ = & \frac{\sigma_1^{*2}}{N_i}\mathbb{E}_i\left(\mathrm{Tr}\left(\Psi_{K_{N_i}}^{-1}\Psi_{K_{N_i}}\right)\right).\end{aligned}$$

Besides,

$$\mathrm{Tr}\left(\Psi_{K_{N_i}}^{-1}\Psi_{K_{N_i}}\right)=K_{N_i}+M.$$

Thus, finally, there exists a constant $C_1>0$ depending on σ_1^* and M such that

$$\mathbb{E}_i\left[\sup_{h\in\mathcal{S}_{K_{N_i},M},\|h\|_{n,i}=1}\nu^2(h)\right]\leq C_1\frac{K_{N_i}}{N_i}.$$

Thus, there exists a constant $C>0$ such that,

$$\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\mathbb{1}_{\Lambda_i}\right]\leq\ 7\inf_{h\in\mathcal{S}_{K_{N_i},M}}\left\|h-b^*_{A_{N_i},i}\right\|_{n,i}^2+C\left(\frac{K_{N_i}}{N_i}+\Delta\right). \tag{54}$$

Upper bound of $\mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\mathbb{1}_{\Lambda_i'}\right]$. Using the Cauchy-Schwarz inequality, we have

$$\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\mathbb{1}_{\Lambda_i'}\right]\leq C_0\log^2(N_i)\mathbb{P}_i\left(\Omega^c_{n,N_i,K_{N_i}}\right)$$

since for N large enough, using (13), we have,

$$\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\leq 2\|\widehat{b}_i\|_\infty^2+2\|b^*_{A_{N_i},i}\|_\infty^2\leq 4A_{N_i}^2\log(N_i)\leq C_0\log^2(N_i)$$

where $C_0>0$ is a constant. Using Lemma 7.8, we have

$$\mathbb{P}_i(\Lambda_i')=\mathbb{P}_i\left(\Omega^c_{n,N_i,K_{N_i}}\right)\leq c\frac{K_{N_i}}{N_i}. \tag{55}$$

Then, from Equation (55), there exists a constant $C>0$ such that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_N,i}\right\|_{n,N_i}^2\mathbb{1}_{\Lambda_i'}\right]\leq C\log^2(N_i)\frac{K_{N_i}}{N_i}. \tag{56}$$

Upper bound of $\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\right]$. From Equations (51), (54) and (56), there exists a constant $C>0$ such that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\right]\leq 7\inf_{h\in\mathcal{S}_{K_{N_i},M}}\left\|h-b^*_{A_{N_i},i}\right\|_{n,i}^2+C\left(\frac{\log^2(N_i)K_{N_i}}{N_i}+\Delta\right). \tag{57}$$

Upper bound of $\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right]$. Using Equation (57), we have

$$\begin{aligned}\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] &= \mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] \\ &\quad + 2\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] \\ &\leq \mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] \\ &\quad + 7\inf_{h\in\mathcal{S}_{K_{N_i},M}}\left\|h - b^*_{A_{N_i},i}\right\|^2_{n,i} + C\left(\frac{\log^2(N_i)K_{N_i}}{N_i} + \Delta\right).\end{aligned}$$

From the proof of Theorem 3.3, we deduce that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] \leq C\log^3(N_i)K_{N_i}/N_i$$

with $C>0$ a constant depending on $\mathfrak{p}_0 = \min_{i\in\mathcal{Y}}\mathfrak{p}^*_i$. Besides, since $b^*_i \in \Sigma(\beta,R)$, we have

$$\inf_{h\in\mathcal{S}_{K_{N_i},M}}\|h - b^*_{A_{N_i},i}\|^2_{n,i} \leq C\left(\frac{A_{N_i}}{K_{N_i}}\right)^{2\beta}$$

where $C>0$ is a constant (see Denis *et al.* (2021), Lemma D.2). Then it comes that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] \leq C\left(\left(\frac{A_{N_i}}{K_{N_i}}\right)^{2\beta} + \frac{K_{N_i}\log^3(N_i)}{N_i} + \Delta\right)$$

where $C>0$ is a constant depending on β, $\Delta = \mathrm{O}(1/N)$. Since

$$K_{N_i} = \mathrm{O}\left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right),$$

we obtain

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] \leq C\log^{6\beta}(N_i)N_i^{-\frac{2\beta}{2\beta+1}} \leq C\log^{6\beta}(N)N_i^{-\frac{2\beta}{2\beta+1}}.$$

Using the Jensen's inequality,

$$\mathbb{E}\left[\mathbb{1}_{N_i>1}\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] \leq C\log^{6\beta}(N)\mathbb{E}\left[\mathbb{1}_{N_i>1}N_i^{-\frac{2\beta}{2\beta+1}}\right] \leq C\log^{6\beta}(N)\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{N_i}\right]\right)^{\frac{2\beta}{2\beta+1}}.$$

Using again *Lemma 4.1* from Györfi *et al.* (2006), we obtain

$$\mathbb{E}\left[\mathbb{1}_{N_i>1}\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] \leq C\log^{6\beta}(N)\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{N_i}\right]\right)^{\frac{2\beta}{2\beta+1}} \leq C\log^{6\beta}(N)N^{-\frac{2\beta}{2\beta+1}}.$$

□

Proof of Theorem 4.7 . For all $i\in\mathcal{Y}$, recall that $b^*_{A_{N_i},i} = b^*_i\mathbb{1}_{[-A_{N_i},A_{N_i}]}$. Furthermore, set

$$N_0 := \min_{i\in\mathcal{Y}} N_i, \quad \text{then} \quad A_{N_0} := \min_{i\in\mathcal{Y}} A_{N_i}. \tag{58}$$

We have

$$\mathbb{E}\left[\mathcal{R}(\widehat{g})-\mathcal{R}(g^*)\right]=\mathbb{E}\left[(1-\mathcal{R}(g^*))\,\mathbb{1}_{N_0\leq 1}\right]+\mathbb{E}\left[(\mathcal{R}(\widehat{g})-\mathcal{R}(g^*))\mathbb{1}_{N_0>1}\right].$$

Then, from Proposition 7.6, we deduce that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{g})-\mathcal{R}(g^*)\right] &\leq \sum_{i=1}^{K}\mathbb{P}(N_i\leq 1)+2\sum_{i=1}^{K}\mathbb{E}\left[|\widehat{\pi}_i(X)-\pi_i^*(X)|\,\mathbb{1}_{N_0>1}\right]\\ &\leq 2KN(1-\mathfrak{p}_0^*)^{N-1}+2\sum_{i=1}^{K}\mathbb{E}\left[|\widehat{\pi}_i(X)-\pi_i^*(X)|\,\mathbb{1}_{N_0>1}\right]\end{aligned}$$

since $\widehat{g}=1$ on the event $\{N_0\leq 1\}$. For all $i\in\mathcal{Y}$ and on the event $\{N_0>1\}$,

$$|\widehat{\pi}_i(X)-\pi_i^*(X)|\leq\left|\widehat{\pi}_i(X)-\bar{\pi}_i^{A_{N_0}}(X)\right|+\left|\bar{\pi}_i^{A_{N_0}}(X)-\bar{\pi}_i^*(X)\right|+|\bar{\pi}_i^*(X)-\pi_i^*(X)|$$

where $\bar{\pi}_i^{A_{N_0}}(X):=\phi_i\left(\bar{\mathbf{F}}^{A_{N_0}}\right)$ and $\bar{\mathbf{F}}^{A_{N_0}}=\left(\bar{F}_1^{A_{N_0}},\cdots,\bar{F}_K^{A_{N_0}}\right)$ with

$$\forall i\in\mathcal{Y},\quad \bar{F}_i^{A_{N_0}}=\sum_{k=0}^{n-1}b^*_{A_{N_0},i}(X_{k\Delta})(X_{(k+1)\Delta}-X_{k\Delta})-\frac{\Delta}{2}b^{*2}_{A_{N_0},i}(X_{k\Delta}).$$

Then, there exists a constant $c>0$ such that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{g})-\mathcal{R}(g^*)\right] &\leq 2\left(\sum_{i=1}^{K}\mathbb{E}\left(\left|\widehat{\pi}_i(X)-\bar{\pi}_i^{A_{N_0}}(X)\right|\mathbb{1}_{N_0>1}\right)+\sum_{i=1}^{K}\mathbb{E}\left(\left|\bar{\pi}_i^{A_{N_0}}(X)-\bar{\pi}_i^*(X)\right|\mathbb{1}_{N_0>1}\right)\right)\\ &\quad +c(1-\mathfrak{p}_0^*)^{N/2}+2\sum_{i=1}^{K}\mathbb{E}\,|\bar{\pi}_i^*(X)-\pi_i^*(X)|.\end{aligned}$$

From the proof of Theorem 3.1, there exists a constant $C_1>0$ depending on $K,\mathfrak{p}_0^*$ and $C_{\mathbf{b}^*}$ and a constant $C_2>0$ depending on K such that

$$\begin{aligned}&\sum_{i=1}^{K}\mathbb{E}\left|\widehat{\pi}_i(X)-\bar{\pi}_i^{A_{N_0}}(X)\right|\leq C_1\left(\frac{1}{\sqrt{N}}+\sum_{i=1}^{K}\mathbb{E}\left[\mathbb{1}_{N_0>1}\left\|\widehat{b}_i-b^*_{A_{N_0},i}\right\|_n\right]\right),\\ &\sum_{i=1}^{K}\mathbb{E}\,|\bar{\pi}_i^*(X)-\pi_i^*(X)|\leq C_2\sqrt{\Delta}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{g})-\mathcal{R}(g^*)\right] &\leq 2C_1\left(\frac{1}{\sqrt{N}}+\sum_{i=1}^{K}\mathbb{E}\left[\mathbb{1}_{N_0>1}\left\|\widehat{b}_i-b^*_{A_{N_0},i}\right\|_n\right]\right)+2C_2\sqrt{\Delta}+c(1-\mathfrak{p}_0^*)^{N/2}\\ &\quad +2K\sum_{i=1}^{K}\mathbb{E}\left[\left|\bar{F}_i^{A_{N_0}}(X)-\bar{F}_i(X)\right|\mathbb{1}_{N_0>1}\right].\end{aligned}$$

For all $i\in\mathcal{Y}$,

$$\begin{aligned}\mathbb{E}\left[\left|\bar{F}_i^{A_{N_0}}(X)-\bar{F}_i(X)\right|\mathbb{1}_{N_0>1}\right] &\leq \mathbb{E}\left[\left|\sum_{k=0}^{n-1}b_i^*(X_{k\Delta})\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\int_{k\Delta}^{(k+1)\Delta}b_i^*(X_s)ds\right|\mathbb{1}_{N_0>1}\right]\\ &\quad +\frac{\Delta}{2}\sum_{k=0}^{n-1}\mathbb{E}\left[\mathbb{1}_{N_0>1}b_i^{*2}(X_{k\Delta})\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\right]\\ &\quad +\mathbb{E}\left|\sum_{k=0}^{n-1}b_i^*(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}(W_{(k+1)\Delta}-W_{k\Delta})\right|.\end{aligned}$$

Under Assumption 4.1, we easily obtain that

$$\mathbb{E}\left|\sum_{k=0}^{n-1} b_i^*(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\int_{k\Delta}^{(k+1)\Delta} b_i^*(X_s)ds\right| \leq C_{\mathbf{b}^*}^2 \sup_{t\in[0,1]}\mathbb{P}\left(\{N_0>1\}\cap\{|X_t|>A_{N_0}\}\right),$$

and

$$\frac{\Delta}{2}\sum_{k=0}^{n-1}\mathbb{E}\left[b_i^{*2}(X_{k\Delta})\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\right] \leq \frac{C_{\mathbf{b}^*}^2}{2}\sup_{t\in[0,1]}\mathbb{P}\left(\{N_0>1\}\cap\{|X_t|>A_{N_0}\}\right).$$

For the last term, consider the natural filtration $\mathcal{F}_t=\sigma(W_s,s\leq t)$ of the Brownian motion $(W_t)_{t\geq 0}$. For all $k\in[\![0,n-1]\!]$, $X_{k\Delta}$ is measurable with respect to $\mathcal{F}_{k\Delta}$ and $W_{(k+1)\Delta}-W_{k\Delta}$ is independent of $\mathcal{F}_{k\Delta}$ since the Brownian motion is an independently increasing process. Consequently, we obtain

$$\begin{aligned}
&\mathbb{E}\left|\sum_{k=0}^{n-1} b_i^*(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}(W_{(k+1)\Delta}-W_{k\Delta})\right| \\
&\qquad\leq C_{\mathbf{b}^*}\sup_{t\in[0,1]}\mathbb{P}\left(\{N_0>1\}\cap\{|X_t|>A_{N_0}\}\right)\mathbb{E}\left[\sum_{k=0}^{n-1}|W_{(k+1)\Delta}-W_{k\Delta}|\right], \\
&\qquad\leq C_{\mathbf{b}^*}\sup_{t\in[0,1]}\mathbb{P}\left(\{N_0>1\}\cap\{|X_t|>A_{N_0}\}\right)
\end{aligned}$$

since,

$$\mathbb{E}\left[\sum_{k=0}^{n-1}|W_{(k+1)\Delta}-W_{k\Delta}|\right]\leq 1$$

Finally, there exists a constant $C>0$ such that

$$\mathbb{E}\left[\mathcal{R}(\widehat{g})-\mathcal{R}(g^*)\right]\leq C\left(\frac{1}{\sqrt{N}}+\sum_{i=1}^{K}\sum_{j=1}^{K}\mathfrak{p}_j^*\mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_0},i}\right\|_{n,j}\mathbb{1}_{N_0>1}\right]+\sup_{t\in[0,1]}\mathbb{P}\left(N_0>1,\ |X_t|>A_{N_0}\right)\right). \tag{59}$$

From Proposition 4.2 with $\alpha=1$, for all $i,j\in\mathcal{Y}$ such that $i\neq j$, we have

$$\mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_0},i}\right\|_{n,j}\mathbb{1}_{N_0>1}\right]\leq\ C\exp(\sqrt{c\log(N)})\mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_0},i}\right\|_{n,i}\mathbb{1}_{N_0>1}\right]+C\frac{\log(N)}{N}. \tag{60}$$

Furthermore, for all $i\in\mathcal{Y}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_0},i}\right\|_{n,i}\mathbb{1}_{N_0>1}\right] &\leq \mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,i}\mathbb{1}_{N_i>1}\right]+\mathbb{E}\left[\left\|b^*_{A_{N_i},i}-b^*_{A_{N_0},i}\right\|_{n,i}\mathbb{1}_{N_0>1}\right] \\
&\leq \mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,i}\mathbb{1}_{N_i>1}\right]+\|b_i^*\|_\infty\sup_{t\in[0,1]}\mathbb{P}\left(\{A_{N_i}\geq|X_t|>A_{N_0}\}\cap\{N_0>1\}\right) \\
&\leq \mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,i}\mathbb{1}_{N_i>1}\right]+C_{\mathbf{b}^*}\sup_{t\in[0,1]}\sum_{j\neq i}\mathbb{P}\left(\{|X_t|>A_{N_j}\}\cap\{N_j>1\}\right).
\end{aligned}$$

We deduce from Equations (59) and (60) that there exists a constant $C>0$ depending on $C_{\mathbf{b}^*},K$ and p_0 such that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\widehat{g})-\mathcal{R}(g^*)\right]\leq C\Bigg(&\frac{1}{\sqrt{N}}+\exp\left(\sqrt{c\log(N)}\right)\sum_{i=1}^{K}\mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,i}\mathbb{1}_{N_i>1}\right]\Bigg) \\
&+C\exp\left(\sqrt{c\log(N)}\right)\sup_{t\in[0,1]}\sum_{i=1}^{K}\mathbb{P}\left(\{|X_t|>A_{N_i}\}\cap\{N_i>1\}\right).
\end{aligned}$$

Under the Assumptions of the Proposition and according to Theorem 4.6, there exist two constants $C_1, C_2 > 0$ such that $\forall i \in \mathcal{Y}$,

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_{n,i} \mathbb{1}_{N_i>1}\right] \leq C_1 \log^{3\beta}(N) N^{-\beta/(2\beta+1)}$$

and we deduce from Lemma 7.5 with $q = 3/2$, for all $i \in \mathcal{Y}$, and for all $t \in [0,1]$,

$$\begin{aligned}\mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_0 > 1\}) &= \mathbb{E}\left[\mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\} | \mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i})\right] \\ &\leq C_2 \mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{A_{N_i}} \exp\left(-\frac{A^2_{N_i}}{3}\right)\right].\end{aligned}$$

Thus, we obtain

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\exp\left(2\sqrt{c\log(N)}\right) N^{-\beta/(2\beta+1)} + \sum_{i=1}^{K} \mathbb{E}\left[\mathbb{1}_{N_i>1} \exp\left(-\frac{A^2_{N_i}}{3}\right)\right]\right)$$

where $C > 0$ is a constant depending on $\beta, C_{\mathbf{b}^*}, K, \mathfrak{p}^*_0$. Finally, choosing $A_{N_i} = \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$ for each $i \in \mathcal{Y}$ leads to the attended result applying the Jensen's inequality together with Lemma 4.1 in Györfi *et al.* (2006). □

Proof of Theorem 4.9 . From Theorem 3.1, as we assumed $\sigma^*(.) = 1$, the excess risk of $\widehat{g}$ satisfies

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\sqrt{\Delta} + \frac{1}{\mathfrak{p}^*_0\sqrt{N}} + \sum_{i=1}^{K} \mathbb{E}\left[\left\|\widehat{b}_i - b^*_i\right\|_n \mathbb{1}_{N_i>1}\right] + \sum_{i=1}^{K} \mathbb{P}(N_i \leq 1)\right) \tag{61}$$

where the constant $C > 0$ depends on $b^* = (b^*_1, \cdots, b^*_K)$ and K. For each $i \in \mathcal{Y}$, we have

$$\mathbb{P}(N_i \leq 1) \leq 2N(1 - \mathfrak{p}^*_0)^{N-1}$$

and

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_i\right\|_n \mathbb{1}_{N_i>1}\right] \leq \sqrt{\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_n \mathbb{1}_{N_i>1}\right] + \mathbb{E}\left[\left\|b^*_i \mathbb{1}_{[-A_{N_i},A_{N_i}]^c}\right\|^2_n \mathbb{1}_{N_i>1}\right]}$$

Using the Cauchy-Schwarz inequality and Assumption 2.1, there exists a constant $C' > 0$ such that

$$\mathbb{E}\left[\left\|b^*_i \mathbb{1}_{[-A_{N_i},A_{N_i}]^c}\right\|^2_n \mathbb{1}_{N_i>1}\right] \leq C' \sqrt{\sup_{t\in[0,1]} \mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\})}.$$

Thus, for all $i \in \mathcal{Y}$, we obtain

$$\mathbb{E}\left\|\widehat{b}_i - b^*_i\right\|_n \leq \sqrt{\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_n \mathbb{1}_{N_i>1}\right] + C' \sqrt{\sup_{t\in[0,1]} \mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\})}}. \tag{62}$$

For each label $i \in \mathcal{Y}$,

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_n \mathbb{1}_{N_i>1}\right] = E\left(\mathbb{1}_{N_i>1} \int_{-A_{N_i}}^{A_{N_i}} \left(\widehat{b}_i - b^*_{A_{N_i},i}\right)^2 (x) f_{n,Y}(x) dx\right) + \frac{2\log^3(N)}{n}$$

where

$$f_{n,Y}(x) := \frac{1}{n}\sum_{k=1}^{n-1} p_Y(k\Delta, x).$$

From the proof of Lemma 4.5, under Assumption 2.1, there exist constants $C_1, C_2 > 0$ such that on the event $\{N_i > 1\}$,

$$\forall x \in [-A_{N_i}, A_{N_i}],\ f_{n,Y}(x) \geq \frac{C_1}{\log(N)} \exp\left(-\frac{2A_{N_i}^2}{3(1-\log^{-1}(N))}\right) \geq \frac{C_2}{\log(N)} \exp\left(-\frac{2}{3}A_{N_i}^2\right)\ a.s$$

and from Lemma 7.4 there exists another constant $C_0 > 0$ such that $f_{n,Y}(x) \leq C_0$ for all $x \in \mathbb{R}$. Then we have

$$\forall i \in \mathcal{Y},\ \forall x \in [-A_{N_i}, A_{N_i}],\ \frac{f_{n,Y}(x)}{f_{n,i}(x)} \leq \frac{C_0}{C_2}\log(N)\exp\left(\frac{2}{3}A_N^2\right).$$

Then, for all $i \in \mathcal{Y}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_n^2 \mathbb{1}_{N_i>1}\right] &\leq \mathbb{E}\left[\mathbb{1}_{N_i>1}\int_{-A_{N_i}}^{A_{N_i}}\left(\widehat{b}_i - b^*_{A_{N_i},i}\right)^2(x) f_{n,i}(x)\frac{f_{n,Y}(x)}{f_{n,i}(x)}\right] + \frac{2\log^3(N)}{n} \\ &\leq \frac{C_0}{C_2}\log(N)\exp\left(\frac{2}{3}A_N^2\right)\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_{n,i}^2 \mathbb{1}_{N_i>1}\right] + \frac{2\log^3(N)}{n}.\end{aligned}$$

From Theorem 4.6, Equation (62) and for $n \propto N$, there exists a constant $C_3 > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n \mathbb{1}_{N_i>1}\right] \leq C_3\sqrt{\exp\left(\frac{2}{3}A_N^2\right)\log^{6\beta+1}(N)N^{-\frac{2\beta}{2\beta+1}} + \sqrt{\sup_{t\in[0,1]}\mathbb{P}(\{|X_t| > A_{N_i}\}\cap\{N_i>1\})}}.$$

Using the Markov inequality, for all $t \in [0,1]$, we have

$$\begin{aligned}\mathbb{P}(\{|X_t| > A_{N_i}\}\cap\{N_i>1\}) =& \mathbb{E}\left[\mathbb{P}\left(\{\exp(4|X_t|^2) > \exp(4A_{N_i}^2)\}\cap\{N_i>1\}|\mathbb{1}_{Y_1=i},\cdots,\mathbb{1}_{Y_N=i}\right)\right] \\ \leq& \mathbb{E}\left[\exp(4|X_t|^2)\right]\mathbb{E}\left[\exp(-4A_{N_i}^2)\mathbb{1}_{N_i>1}\right]\end{aligned}$$

and since $\sigma^*(.) = 1$ and under Assumption 4.8, there exists a constant $C_* > 0$ such that $\mathbb{E}\left[\exp(4|X_t|^2)\right] \leq C_*$ (according to Gobet (2002), Proposition 1.1). Thus, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n \mathbb{1}_{N_i>1}\right] \leq C\left(\exp\left(\frac{1}{3}A_N^2\right)\log^{3\beta+1}(N)N^{-\beta/(2\beta+1)}\right) + C\mathbb{E}\left[\exp(-4A_{N_i}^2)\mathbb{1}_{N_i>1}\right]. \quad (63)$$

From Equations (63) and (61), we finally obtain

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\log^{3\beta+1}(N)N^{-3\beta/4(2\beta+1)}$$

with $A_{N_i} = \sqrt{\frac{3\beta}{4(2\beta+1)}\log(N_i)}$ and $C > 1$ a new constant. □

References

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

Baíllo, A., Cuevas, A. & Fraiman, R. (2011). Classification methods for functional data. *The Oxford handbook of functional data analysis* .

Cadre, B. (2013). Supervised classification of diffusion paths. *Mathematical Methods of Statistics* **22**, 213–225.

Cohen, A., Davenport, M. & Leviatan, D. (2013). On the stability and accuracy of least squares approximations. *Foundations of computational mathematics* **13**, 819–834.

Comte, F. & Genon-Catalot, V. (2020a). Nonparametric drift estimation for i.i.d. paths of stochastic differential equations. *The Annals of Statistics* **48**, 3336–3365.

Comte, F. & Genon-Catalot, V. (2020b). Regression function estimation as a partly inverse problem. *Annals of the Institute of Statistical Mathematics* **72**, 1023–1054.

Comte, F. & Genon-Catalot, V. (2020c). Regression function estimation as a partly inverse problem. *Annals of the Institute of Statistical Mathematics* **72**, 1023–1054.

Comte, F. & Genon-Catalot, V. (2021). Drift estimation on non compact support for diffusion models. *Stochastic Processes and their Applications* **134**, 174–207.

Comte, F., Genon-Catalot, V., Rozenholc, Y. *et al.* (2007). Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* .

De Boor, C. (1978). *A practical guide to splines*, vol. 27. springer-verlag New York.

De Micheaux, P.-L., Mozharovskyi, P. & Vimond, M. (2021). Depth for curve data and applications. *Journal of the American Statistical Association* **116**, 1881–1897.

Della-Maestra, L. & Hoffmann, M. (2022). Nonparametric estimation for interacting particle systems: Mckean–vlasov models. *Probability Theory and Related Fields* **182**, 551–613.

Denis, C., Dion-Blanc, C. & Martinez, M. (2020). Consistent procedures for multiclass classification of discrete diffusion paths. *Scandinavian Journal of Statistics* **47**, 516–554.

Denis, C., Dion-Blanc, C. & Martinez, M. (2021). A ridge estimator of the drift from discrete repeated observations of the solutions of a stochastic differential equation. *Bernoulli* .

Devroye, L., Györfi, L. & Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, vol. 31. Springer Science & Business Media.

Domingo, D., d'Onofrio, A. & Flandoli, F. (2020). Properties of bounded stochastic processes employed in biophysics. *Stochastic Analysis and Applications* **38**, 277–306.

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

Erban, R. & Chapman, S. J. (2009). Stochastic modelling of reaction–diffusion processes: algorithms for bimolecular reactions. *Physical biology* **6**, 046001.

Gadat, S., Gerchinovitz, S. & Marteau, C. (2020). Optimal functional supervised classification with separation condition. *Bernoulli* **26**, 1797–1831.

Gadat, S., Klein, T. & Marteau, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics* **44**, 982–1009.

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

Györfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

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

Hoffmann, M. (1999b). Lp estimation of the diffusion coefficient. *Bernoulli* pp. 447–481.

Iacus, S.-M. (2009). *Simulation and inference for stochastic differential equations: with R examples*. Springer Science & Business Media.

Jacod, J. & Shiryaev, A. (2013). *Limit theorems for stochastic processes*, vol. 288. Springer Science & Business Media.

Karatzas, I. & Shreve, S. (2014). *Brownian motion and stochastic calculus*, vol. 113. springer.

Kidger, P., Foster, J., Li, X. & Lyons, T. (2021). Neural sdes as infinite-dimensional gans. In *International Conference on Machine Learning*, pp. 5453–5463. PMLR.

Le Gall, J.-F. (2013). *Mouvement brownien, martingales et calcul stochastique*. Springer.

Leon, S.-J., Björck, A. & Gander, W. (2013). Gram-schmidt orthogonalization: 100 years and more. *Numerical Linear Algebra with Applications* **20**, 492–532.

Marie, N. & Rosier, A. (2021). Nadaraya-watson estimator for iid paths of diffusion processes. *arXiv preprint arXiv:2105.06884* .

Ramsay, J.-O. & Silverman, B.-W. (2005). *Fitting differential equations to functional data: Principal differential analysis*. Springer.

Revuz, D. & Yor, M. (1999). *Continuous martingales and Brownian motion*, vol. 293 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edn. ISBN 3-540-64325-7. doi:10.1007/978-3-662-06400-9. URL `https://doi.org/10.1007/978-3-662-06400-9`.

Rossi, F. & Villa, N. (2008). Recent advances in the use of svm for functional data classification. In *Functional and Operatorial Statistics*, pp. 273–280. Physica-Verlag HD, Heidelberg.

Tsybakov, A.-B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Van-de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics* pp. 1779–1801.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and its application* **3**, 257–295.

Wang, S., Cao, J. & Yu, P. (2020). Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering* .

Yang, Y. (1999). Minimax nonparametric classification: Rates of convergence. *IEEE Transactions on Information Theory* **45**, 2271–2284.

Appendix

Proof of Lemma 7.1. Let $s,t\in[0,1]$ with $s<t$, and $q\geq 1$. By convexity of $x\mapsto|x|^{2q}$, we have

$$|X_t-X_s|^{2q}\leq 2^{2q-1}\left(\left|\int_s^t b_Y^*(X_u)du\right|^{2q}+\left|\int_s^t\sigma(X_u)dW_u\right|^{2q}\right)$$

Then, from Jensen's inequality, we have

$$\left|\int_s^t b_Y^*(X_u)du\right|^{2q}\leq(t-s)^{2q-1}\int_s^t|b_Y^*(X_u)|^{2q}\,du,$$

Hence, under Assumption 2.1 on function b_Y^*, we deduce that

$$\mathbb{E}\left[\left|\int_s^t b_Y^*(X_u)du\right|^{2q}\right]\leq C_q(t-s)^{2q}\left(1+\mathbb{E}\left[\sup_{t\in[0,1]}|X_s|^{2q}\right]\right),$$

and using Burkholder-Davis-Gundy inequality, we obtain

$$\forall m>0,\quad\mathbb{E}\left[\left(\int_s^t\sigma(X_u)dW_u\right)^{2m}\right]\leq C_m\mathbb{E}\left[\left(\int_s^t\sigma^2(X_u)du\right)^m\right]\leq C_m\sigma_1^{2m}(t-s)^m.$$

From the above equalities, we get

Finally, as the process has finite moments, we obtain that

$$\mathbb{E}\,|X_t-X_s|^{2q}\leq C(t-s)^q$$

where C is a constant depending on q,L_0, and σ_1. □

Proof of Lemma 4.5 . For all $i\in\mathcal{Y}$ and on the event $\{N_i>1\}$, let us consider a vector $\left(x_{-M},\cdots,x_{K_{N_i}-1}\right)\in\mathbb{R}^{K_{N_i}+M}$ such that $x_j\in[u_{j+M},u_{j+M+1})$ and $B_{j,M,\mathbf{u}}(x_j)\neq 0$. Since $[u_{j+M},u_{j+M+1})\cap[u_{j'+M},u_{j'+M+1})=\emptyset$ for all $j,j'\in\{-M,\cdots,K_{N_i}-1\}$ such that $j\neq j'$, then for all $j,j'\in\{-M,\cdots,K_{N_i}-1\}$ such that $j\neq j',B_{j,M,(u)}(x_{j'})=0$. Consequently, we obtain:

$$\begin{aligned}\det\left(\left(B_{\ell,M,\mathbf{u}}(x_{\ell'})\right)_{-M\leq\ell,\ell'\leq K_{N_i}-1}\right)&=\det\left(\operatorname{diag}\left(B_{-M,M,\mathbf{u}}(x_M),\cdots,B_{K_{N_i}-1,M,\mathbf{u}}(x_{K_{N_i}-1})\right)\right)\\&=\prod_{\ell=-M}^{K_{N_i}-1}B_{\ell,M,\mathbf{u}}(x_\ell)\neq 0.\end{aligned}$$

Then, we deduce from Comte & Genon-Catalot (2020a), *Lemma 1* that the matrix $\Psi_{K_{N_i}}$ is invertible for all $K_{N_i}\in\mathcal{K}_{N_i}$, where the interval $[-A_{N_i},A_{N_i}]$ and the function f_T is replaced by $f_n:x\mapsto\frac{1}{n}\sum_{k=0}^{n-1}p(k\Delta,x)$ with $\lambda([-A_{N_i},A_{N_i}]\cap\operatorname{supp}(f_n))>0$, λ being the Lebesgue measure.

For all $w\in\mathbb{R}^{K_{N_i}+M}$ such that $\|w\|_{2,K_{N_i}+M}=1$, we have:

$$w'\Psi_{K_{N_i}}w=\|h_w\|_n^2=\int_{-A_{N_i}}^{A_{N_i}}h_w^2(x)f_n(x)dx+\frac{h_w^2(x_0)}{n}\quad\text{with}\quad h_w=\sum_{\ell=-M}^{K_{N_i}-1}w_\ell B_{\ell,M,\mathbf{u}}.$$

Since $\sigma^*=1$, according to Lemma 7.5, under Assumption 2.1, the transition density satisfies:

$$\forall(t,x)\in(0,1]\times\mathbb{R},\ \frac{1}{K_q\sqrt{t}}\exp\left(-\frac{(2q-1)x^2}{2qt}\right)\leq p(t,x)\leq\frac{K_q}{\sqrt{t}}\exp\left(-\frac{x^2}{2qt}\right)\ \text{where}\ K_q>1\ \text{and}\ q>1.$$

We set $q = 3/2$, thus, since $s \mapsto \exp\left(-(2q-1)x^2/2qs\right)$ is an increasing function, we have on the event $\{N_i > 1\}$ and for all $x \in [-A_{N_i}, A_{N_i}]$,

$$\begin{aligned} f_n(x) &\geq \frac{1}{Cn}\sum_{k=1}^{n-1}\exp\left(-\frac{2x^2}{3k\Delta}\right) \geq \frac{1}{C}\int_0^{(n-1)\Delta}\exp\left(-\frac{2x^2}{3s}\right)ds \\ &\geq \frac{1}{C}\int_{1-\log^{-1}(N_i)}^{1-2^{-1}\log^{-1}(N_i)}\exp\left(-\frac{2x^2}{3s}\right)ds \\ &\geq \frac{1}{2C\log(N_i)}\exp\left(-\frac{2A_{N_i}^2}{3(1-\log^{-1}(N_i))}\right). \end{aligned}$$

Finally, since there exists a constant $C_1 > 0$ such that $\|h_w\|^2 \geq C_1 A_{N_i} K_{N_i}^{-1}$ (see Denis *et al.* (2021), Lemma 2.6), for all $w \in \mathbb{R}^{K_{N_i}+M}$ such that $\|w\|_{2,K_{N_i}+M} = 1$, there exists constants $C', C > 0$ such that,

$$w'\Psi_{K_{N_i}}w \geq \frac{C'A_{N_i}}{K_{N_i}\log(N_i)}\exp\left(-\frac{2A_{N_i}^2}{3(1-\log^{-1}(N_i))}\right) \geq \frac{CA_{N_i}}{K_{N_i}\log(N_i)}\exp\left(-\frac{2}{3}A_{N_i}^2\right).$$

Furthermore, we set $w_0 = e_{K_{N_i}-1} \in \mathbb{R}^{K_{N_i}+M}$ where for all $\ell \in [\![-M, K_{N_i}-1]\!]$,

$$\left[e_{K_{N_i}-1}\right]_\ell := \delta_{\ell,K_{N_i}-1} = \begin{cases} 0 & \text{if } \ell \neq K_{N_i}-1 \\ 1 & \text{else.} \end{cases}$$

We have,

$$\begin{aligned} w_0'\Psi_{K_{N_i}}w_0 &= \int_{-A_{N_i}}^{A_{N_i}} B_{K_{N_i}-1,M,\mathbf{u}}^2(x)f_n(x) + \frac{B_{K_{N_i}-1,M,\mathbf{u}}(0)}{n} \\ &\leq \frac{C}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp\left(-\frac{u_{K_{N_i}-1}^2}{3k\Delta}\right)\left\|B_{K_{N_i}-1,M,\mathbf{u}}\right\|^2 + \frac{1}{n} \\ &\leq \frac{CC_1A_{N_i}K_{N_i}^{-1}}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp\left(-\frac{\alpha_{N_i}^2}{3k\Delta}\right) + \frac{1}{n} \end{aligned}$$

where $\alpha_{N_i} = A_{N_i}(K_{N_i}-2)/K_{N_i}$, $\left\|B_{K_{N_i}-1,M,\mathbf{u}}\right\|^2 \leq C_1 A_{N_i}K_{N_i}^{-1}$ (see Denis *et al.* (2021), Lemma 2.6) and $C_1 > 0$ is a constant. Since the function $s \mapsto \exp\left(-\alpha_{N_i}^2/3s\right)/\sqrt{s}$ is increasing, we deduce that

$$n^{-1}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp\left(-\alpha_{N_i}^2/3k\Delta\right) \leq n^{-1}\sum_{k=1}^{n-1}\exp\left(-\alpha_{N_i}^2/3\right),$$

and for N large enough,

$$w_0'\Psi_{K_{N_i}}w_0 \leq \frac{CA_{N_i}}{K_{N_i}}\exp\left(-\frac{A_{N_i}^2}{3}\left(\frac{K_{N_i}-2}{K_{N_i}}\right)^2\right) + \frac{1}{n} \leq \frac{C'A_{N_i}}{K_{N_i}}\exp\left(-\frac{A_{N_i}^2}{3}\left(\frac{K_{N_i}-2}{K_{N_i}}\right)^2\right)$$

where $C' > 0$ is a constant and $n \geq N \geq N_i$. □

Proof of Lemma 7.8 . Let us remind the reader of the Gram matrix $\Psi_{K_{N_i}}$ given in Equation (14) for $i \in \mathcal{Y}$,

$$\Psi_{K_{N_i}} = \mathbb{E}\left[\frac{1}{N_i n}\mathbf{B}'_{K_{N_i}}\mathbf{B}_{K_{N_i}}\right] = \mathbb{E}\left(\widehat{\Psi}_{K_{N_i}}\right)$$

where, on the event $\{N_i > 1\}$, and denoting by $\mathcal{I}_i := \{i_1, \ldots, i_{N_i}\}$ the indices j such that $Y_j = i$,

$$\mathbf{B}_{K_{N_i}} := \begin{pmatrix} B_{-M}\left(X_0^{i_1}\right) & \ldots & \ldots & B_{K_{N_i}-1}\left(X_0^{i_1}\right) \\ \vdots & & & \vdots \\ B_{-M}\left(X_{(n-1)\Delta}^{i_1}\right) & \ldots & \ldots & B_{K_{N_i}-1}\left(X_{(n-1)\Delta}^{i_1}\right) \\ \vdots & & & \vdots \\ B_{-M}\left(X_0^{i_{N_i}}\right) & \ldots & \ldots & B_{K_{N_i}-1}\left(X_0^{i_{N_i}}\right) \\ \vdots & & & \vdots \\ B_{-M}\left(X_{(n-1)\Delta}^{i_{N_i}}\right) & \ldots & \ldots & B_{K_{N_i}-1}\left(X_{(n-1)\Delta}^{i_{N_i}}\right) \end{pmatrix} \in \mathbb{R}^{N_i n \times (K_{N_i}+M)}. \tag{64}$$

The empirical counterpart $\widehat{\Psi}$ is the random matrix given by $\widehat{\Psi}_{K_{N_i}}$ of size $(K_{N_i}+M)\times(K_{N_i}+M)$ is given by

$$\widehat{\Psi}_{K_{N_i}} := \frac{1}{N_i n}\mathbf{B}'_{K_{N_i}}\mathbf{B}_{K_{N_i}} = \left(\frac{1}{N_i n}\sum_{j=1}^{N_i}\sum_{k=0}^{n-1} B_\ell(X_{k\Delta}^{i_j})B_{\ell'}(X_{k\Delta}^{i_j})\right)_{\ell,\ell' \in [-M, K_{N_i}-1]}. \tag{65}$$

We build an orthonormal basis $\theta = (\theta_{-M}, \cdots, \theta_{K_{N_i}-1})$ of the subspace $\mathcal{S}_{K_{N_i},M}$ with respect to the $\mathbb{L}^2$ inner product $\langle ., . \rangle$ through the Gram-Schmidt orthogonalization of the spline basis $(B_{-M}, \cdots, B_{K_{N_i}-1})$. Then, we have

$$\mathrm{Span}(B_{-M}, \cdots, B_{K_{N_i}-1}) = \mathrm{Span}(\theta_{-M}, \cdots, \theta_{K_{N_i}-1}) = \mathcal{S}_{K_{N_i},M}$$

and the matrix given in Equation (64) is factorized as follows

$$\mathbf{B}_{K_{N_i}} = \mathbf{\Theta}_{K_{N_i}}\mathbf{R}_{K_{N_i}} \tag{66}$$

where

$$\mathbf{\Theta}_{K_{N_i}} = \left(\left(\theta_\ell(X_0^{i_j}), \theta_\ell(X_\Delta^{i_j}), \cdots, \theta_\ell(X_{n\Delta}^{i_j})\right)'\right)_{\substack{1\le j\le N_i \\ -M\le \ell \le K_{N_i}-1}} \in \mathbb{R}^{N_i n \times (K_{N_i}+M)}$$

and $\mathbf{R}_{K_{N_i}}$ is an upper triangular matrix of size $(K_{N_i}+M)\times(K_{N_i}+M)$ see Leon *et al.* (2013)). Let $\Phi_{K_{N_i}}$ be the Gram matrix under the orthonormal basis $\theta = \left(\theta_{-M}, \cdots, \theta_{K_{N_i}-1}\right)$ and given by

$$\Phi_{K_{N_i}} = \mathbb{E}\left[\frac{1}{N_i n}\mathbf{\Theta}'_{K_{N_i}}\mathbf{\Theta}_{\mathbf{K_{N_i}}}\right] = \mathbb{E}\left(\widehat{\Phi}_{K_{N_i}}\right)$$

where,

$$\widehat{\Phi}_{K_{N_i}} := \frac{1}{N_i n}\mathbf{\Theta}'_{K_{N_i}}\mathbf{\Theta}_{K_{N_i}} = \left(\frac{1}{N_i n}\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}\theta_\ell(X_{k\Delta}^{i_j})\theta_{\ell'}(X_{k\Delta}^{i_j})\right)_{\ell,\ell' \in [-M, K_{N_i}-1]}. \tag{67}$$

The matrices $\Psi_{K_{N_i}}$ and $\widehat{\Psi}_{K_{N_i}}$ are respectively linked to the matrices $\Phi_{K_{N_i}}$ and $\widehat{\Phi}_{K_{N_i}}$ through the following relations

$$\Psi_{K_{N_i}} = \mathbf{R}'_{K_{N_i}} \Phi_{K_{N_i}} \mathbf{R}_{K_{N_i}} \quad \text{and} \quad \widehat{\Psi}_{K_{N_i}} = \mathbf{R}'_{K_{N_i}} \widehat{\Phi}_{K_{N_i}} \mathbf{R}_{K_{N_i}}$$

Since for all $h = \sum_{i=-M}^{K_{N_i}-1} a_i B_{i,M,\mathbf{u}} \in S_{K_{N_i},M}$ one has

$$\|h\|^2_{n,N_i} = a' \widehat{\Psi}_{K_{N_i}} a \quad \text{and} \quad \|h\|^2_{n,i} = a' \Psi_{K_{N_i}} a, \quad \text{with} \quad a = \left(a_{-M}, \cdots, a_{K_{N_i}-1}\right)',$$

we deduce that

$$\|h\|^2_{n,N_i} = w' \widehat{\Phi}_{K_{N_i}} w \quad \text{and} \quad \|h\|^2_{n,i} = w' \Phi_{K_{N_i}} w, \quad \text{with} \quad w = \mathbf{R}_{K_{N_i}} a.$$

Under Assumption 2.1, we follow the lines of Comte & Genon-Catalot (2020c) *Proposition 2.3* and *Lemma 6.2*. Then,

$$\begin{aligned}
\sup_{h \in S_{K_{N_i},M}, \|h\|_{n,i}=1} \left| \|h\|^2_{n,N_i} - \|h\|^2_{n,i} \right| &= \sup_{w \in \mathbb{R}^{K_{N_i}+M}, \left\| \Phi^{1/2}_{K_{N_i}} w \right\|_{2,K_{N_i}+M}=1} \left| w' \left(\widehat{\Phi}_{K_{N_i}} - \Phi_{K_{N_i}} \right) w \right| \\
&= \sup_{u \in \mathbb{R}^{K_{N_i}+M}, \|u\|_{2,K_{N_i}+M}=1} \left| u' \Phi^{-1/2}_{K_{N_i}} \left(\widehat{\Phi}_{K_{N_i}} - \Phi_{K_{N_i}} \right) \Phi^{-1/2}_{K_{N_i}} u \right| \\
&= \left\| \Phi^{-1/2}_{K_{N_i}} \widehat{\Phi}_{K_{N_i}} \Phi^{-1/2}_{K_{N_i}} - \mathrm{Id}_{K_{N_i}+M} \right\|_{\mathrm{op}}.
\end{aligned}$$

Therefore,

$$\Omega^c_{n,N_i,K_{N_i}} = \left\{ \left\| \Phi^{-1/2}_{K_{N_i}} \widehat{\Phi}_{K_{N_i}} \Phi^{-1/2}_{K_{N_i}} - \mathrm{Id}_{K_{N_i}+M} \right\|_{\mathrm{op}} > 1/2 \right\}.$$

Then, we apply here Theorem 1 of Cohen *et al.* (2013), it yields

$$\mathbb{P}_i \left(\Omega^c_{n,N_i,K_{N_i}} \right) \leq 2(K_{N_i}+M) \exp\left(-c_{1/2} \frac{N_i}{\mathcal{L}(K_{N_i}+M)(\|\Phi^{-1}_{K_{N_i}}\|_{\mathrm{op}} \vee 1)} \right) \tag{68}$$

with $c_{1/2} = (3\log(3/2) - 1)/2$ and $\mathcal{L}(K_{N_i}+M) := \sup_{x \in [-A_{N_i}, A_{N_i}]} \sum_{\ell=-M}^{K_{N_i}-1} \theta^2_\ell(x)$ (from application of Lemma 6.2 from Comte & Genon-Catalot (2020b)). For all $h = \sum_{\ell=-M}^{K_{N_i}-1} w_\ell \theta_\ell \in \mathrm{Span}\left(\theta_{-M}, \cdots, \theta_{K_{N_i}-1}\right) = \mathcal{S}_{K_{N_i},M}$, we have

$$\|h\|^2 = \|w\|^2_{2,K_{N_i}+M} \quad \text{and} \quad \|h\|^2_{n,i} = 1 \quad \text{implies} \quad w = \Phi^{-1/2}_{K_{N_i}} u \quad \text{where} \quad u \in \mathbb{R}^{K_{N_i}+M} : \|u\|_{2,K_{N_i}+M} = 1.$$

We deduce that

$$\sup_{h \in \mathcal{S}_{K_{N_i}+M},\ \|h\|^2_{n,i}=1} \|h\|^2 = \sup_{u \in \mathbb{R}^{K_{N_i}+M}, \|u\|_{2,K_{N_i}+M}=1} u' \Phi^{-1}_{K_{N_i}} u = \left\| \Phi^{-1}_{K_{N_i}} \right\|_{\mathrm{op}}.$$

Furthermore, for all $h = \sum_{\ell=-M}^{K_{N_i}-1} a_\ell B_\ell \in \mathrm{Span}\left(B_{-M}, \cdots, B_{K_{N_i}-1}\right) = \mathcal{S}_{K_{N_i},M}$, we have on one side

$$\|h\|^2_{n,i} = 1 \quad \text{implies} \quad a = \Psi^{-1/2}_{K_{N_i}} u \quad \text{where} \quad u \in \mathbb{R}^{K_{N_i}+M} : \|u\|_{2,K_{N_i}+M} = 1$$

and on the other side, for all $h \in \mathcal{S}_{K_{N_i}+M}$ such that $\|h\|_{n,i}^2 = 1$, from Denis *et al.* (2021) *Lemma 2.6*, there exists a constant $C > 0$ such that,

$$\|h\|^2 \leq C A_{N_i} K_{N_i}^{-1} \|a\|_{2,K_{N_i}+M}^2 = C A_{N_i} K_{N_i}^{-1} u' \Psi_{K_{N_i}}^{-1} u.$$

Then we have $a.s$

$$\left\| \Phi_{K_{N_i}}^{-1} \right\|_{\text{op}} = \sup_{h \in \mathcal{S}_{K_{N_i}+M},\ \|h\|_{n,i}^2=1} \|h\|^2 \leq \frac{C A_{N_i}}{K_{N_i}} \sup_{u \in \mathbb{R}^{K_{N_i}+M}, \|u\|_{2,K_{N_i}+M}=1} u' \Psi_{K_{N_i}}^{-1} u = \frac{C A_{N_i}}{K_{N_i}} \left\| \Psi_{K_{N_i}}^{-1} \right\|_{\text{op}}. \tag{69}$$

From Equations (68), (15) and (69), there exists a constant $C > 0$ such that

$$\mathbb{P}_i \left(\Omega_{n,N_i,K_{N_i}}^c \right) \leq 2(K_{N_i} + M) \exp \left(-C \frac{K_{N_i} \log^2(N_i)}{A_{N_i} \mathcal{L}(K_{N_i} + M)} \right). \tag{70}$$

Then, as $\mathcal{L}(K_{N_i} + M) := \sup_{x \in [-A_{N_i}, A_{N_i}]} \sum_{\ell=-M}^{K_{N_i}-1} \theta_\ell^2(x) \leq \sup_{x \in \mathbb{R}} \sum_{\ell=-M}^{K_{N_i}-1} \theta_\ell^2(x)$ $a.s.$ and the functions θ_ℓ bounded, there exists a constant C_θ depending on the orthonormal basis $\theta = \left(\theta_{-M}, \cdots, \theta_{K_{N_i}-1} \right)$ such that $\mathcal{L}(K_{N_i} + M) \leq C_\theta K_{N_i}$. Furthermore, since $A_{N_i} \leq \sqrt{\frac{3\beta}{2\beta+1} \log(N_i)}$, we obtain from Equation (70),

$$\mathbb{P}_i \left(\Omega_{n,N_i,K_{N_i}}^c \right) \leq 2(K_{N_i} + M) \exp \left(-C \log^{3/2}(N_i) \right) \tag{71}$$

where $C > 0$ is a new constant depending on C_θ, β and M. Since $N_i \longrightarrow \infty$ $a.s.$ as $N \longrightarrow \infty$, one has

$$\exp \left(\log(N_i) - C \log^{3/2}(N_i) \right) \longrightarrow 0 \ a.s. \ \text{ as } \ N \longrightarrow \infty.$$

Then, for N large enough, $\exp \left(\log(N_i) - C \log^{3/2}(N_i) \right) \leq 1$ $a.s.$ and from Equation (71),

$$\mathbb{P}_i \left(\Omega_{n,N_i,K_{N_i}}^c \right) \leq \frac{2(K_{N_i} + M)}{N_i} \leq c \frac{K_{N_i}}{N_i}$$

where the constant $c > 0$ depends on M. □