Supervised Classification of Diffusion Paths

Benoît CADRE

IRMAR, ENS Cachan Bretagne, CNRS, UEB
Campus de Ker Lann
Avenue Robert Schuman, 35170 Bruz, France
cadre@bretagne.ens-cachan.fr

Abstract

Let $X = (X_t)_{t\in[0,1]}$ be a stochastic process with label $Y \in \{0,1\}$. We assume that X is some Brownian diffusion when $Y = 0$, while X is another Brownian diffusion when $Y = 1$. Based on an explicit computation of the Bayes rule, we construct an empirical classification rule $\hat{g}$ drawn from an i.i.d. sample of copies of (X,Y). In a nonparametric setting, we prove that $\hat{g}$ is a consistent rule, and we derive its rate of convergence under mild assumptions on the model.

Index Terms — Supervised Classification, Brownian Diffusion Modelling, Stochastic Differential Equation.

1 Introduction

Supervised classification deals with predicting the unknown nature Y, called class or label, of an observation X taking values in some metric space $\mathscr{H}$. Assume for simplicity that the label Y only takes two values, say 0 or 1. The statistician creates a classification rule $g : \mathscr{H} \to \{0,1\}$ which represents her guess on the label Y of X. In practical issues, the covariate X usually does not fully determines the label, hence it is certainly possible to misspecify its associated label. An error occurs when $g(X)$ differs from Y, and the probability of error for a particular rule g is denoted by

$$L(g) = \mathbb{P}\big(g(X) \neq Y\big). \tag{1.1}$$

The Bayes rule $g^\star$ defined by

$$g^\star(x) = \begin{cases} 1 \text{ if } \mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x) \\ 0 \text{ otherwise} \end{cases} \tag{1.2}$$

has the smallest probability of error, i.e. Bayes risk $L(g^\star)$ satisfies

$$L(g^\star) \leq L(g),$$

for any rule g (see Theorem 2.1 in the book by Devroye et al, 1996). The supervised classification problem is to construct rules $\hat{g}_n$ based on independent random variables $(X_1,Y_1),\cdots,(X_n,Y_n)$ with the same distribution as (X,Y) and independent of it, whose performance is close to that of Bayes rule, i.e.

$$\lim_{n\to\infty} \mathbb{E}L(\hat{g}_n) = L(g^\star),$$

where $L(\hat{g}_n) = \mathbb{P}\big(\hat{g}_n(X) \neq Y | (X_1,Y_1),\cdots,(X_n,Y_n)\big)$. An empirical rule satisfying this property is said to be consistent.

This paper is dedicated to the functional supervised classification problem. In this setting, the covariate X belongs to an infinite dimensional space of functions $\mathscr{H}$. Direct adaptations of finite-dimensional local averaging methods (e.g. kernel or k-nearest-neighbor methods) to the infinite dimensional case has been studied by Cover and Hart (1967), Kulkarni and Posner (1995), Abraham et al (2003) and Cérou and Guyader (2006). If consistency can be obtained at the price of requiring regularity assumption on the regression function $\mathbb{E}(Y|X=x)$, these direct approaches suffer from the curse of dimensionality, as already pointed out by Ramsay and Silverman (2002, page 129). Taking this fact into account, many other methods has been introduced in the litterature. For instance, Hall et al (2001) and Ramsay and Silverman (1997, 2002) employ a functional data-analytic method for dimension reduction based on Principal Component Analysis; Hastie et al (1995) suggest penalized discriminant analysis; Ferraty and Vieu (2003) propose a single functional index method; Biau et al (2005) introduce a data-driven procedure based on filtrering; Baíllo et al (2011) study the situation where X is some Gaussian process.

The goal is to construct a consistent empirical rule that achieves a fast rate of convergence. A solution for this is to introduce a learning methodology that is specifically adapted to a given nonparametric model. The observation that the law of X may be decomposed into a mixture

$$\mathbb{P}(Y=0)\mu_0 + \mathbb{P}(Y=1)\mu_1,$$

where μ_i is the conditional distribution of X given $Y=i$, lead to consider nonparametric models for the μ_i's. Then, in the spirit of Baíllo et al (2011) that study

the supervised classification problem for a given family of stochastic processes (namely, the μ_i's are distribution of Gaussian processes with triangular covariance functions), we propose to focus on the case where the law of X is a mixture of two Brownian diffusions, i.e. the μ_i's are distributions of unknown Brownian diffusions. Many practical issues can be derived from this nonparametric setting, since stochastic modelling of diffusion processes has a wide range of applications. Mention, among many others, population dynamics (Lande et al, 2003), signal processing (Chonavel, 2002), biology (Renshaw, 1991), chemical kinetics (van Kampen, 2007), finance (Lamberton and Lapeyre, 1996)... In biology, for instance, a supervised classification method for diffusion processes can be a powerful tool to classify different types of bacteria with the help of one of its most obvious external manifestation, namely its population dynamics. Derived from a birth and death process, one can model the population density of bacteria X_t at time $t \geq 0$ by a diffusion process $(X_t)_{t\geq 0}$, solution to the stochastic differential equation (see Lande et al, 2003)

$$\mathrm{d}X_t = H(X_t)\,\mathrm{d}t + \sqrt{\gamma X_t}\,\mathrm{d}B_t,$$

where the unknown parameters $\gamma \geq 0$ and $H : \mathbb{R} \to \mathbb{R}$ depend on the type of bacteria and, here and in the following, $(B_t)_{t\geq 0}$ is a standard 1-dimensional Brownian motion. (We refer the reader to the books by Revuz and Yor, 1999, or Kloeden and Platen, 1999, for materials on stochastic calculus.) Observe that the particular case $H(x) = (r - cx)x$ leads to the well-known Feller logistic model. In this situation, the learning sample is composed of n observations of the evolution of the population density and, for each of them, the type of bacteria. Based on the learning sample, the empirical rule introduced in this paper aims at specifying the type of bacteria associated with a population density path.

In the whole paper, the covariate $X = (X_t)_{t\in[0,1]}$ belongs to the space $\mathscr{H} = C([0,1])$ of continuous real-valued functions defined on $[0,1]$. The pair $(X,Y) \in C([0,1]) \times \{0,1\}$ is defined by a nonparametric model in which X is some diffusion process when $Y = 0$, while X is another diffusion process when $Y = 1$. As in (1.1), we associate the probability of error $L(g)$ of the classification rule $g : C([0,1]) \to \{0,1\}$, and Bayes rule $g^\star$ defined by (1.2) has the smallest probability of error. In this paper, we study the performance of an empirical rule $\hat{g}$ based on the explicit representation of Bayes rule; consistency and rate of convergence are established. More precisely, we prove that for some $u > 0$ that depends on the regularity of the

coefficients of the diffusion, we have:

$$\mathbb{E}L(\hat{g}) - L(g^\star) \leq Cn^{-u},$$

with an explicit constant $C > 0$ (under the same functional setting, this rate is the optimal rate of convergence in the nonparametric regression estimation). As far as we know, only Baìllo et al (2011) obtain similar results in supervised classification for stochastic processes, but dealing with some Gaussian processes.

The paper is organized as follows. In Section 2, we fix the nonparametric model and we compute Bayes rule. Section 3 is devoted to the construction of the empirical rule. We also establish consistency. In Section 4, we give the rate of convergence of the empirical rule. Finally, Section 5 contains proofs.

2 Bayes rule for diffusion models

Model – The model (X,Y) is a random pair that takes values in $C([0,1]) \times \{0,1\}$, with $p_0 = \mathbb{P}(Y=0) \in]0,1[$. The covariate $X = (X_t)_{t\in[0,1]}$ and the label Y are related by the equation:

$$(M): \begin{cases} \mathrm{d}X_t = b_0(t,X_t)\mathrm{d}t + \sigma_0(t,X_t)\mathrm{d}B_t \text{ when } Y=0, \text{ while} \\ \mathrm{d}X_t = \big(b_0(t,X_t) + (f_0\sigma_0)(t,X_t)\big)\mathrm{d}t + \sigma_0(t,X_t)\mathrm{d}B_t \text{ when } Y=1, \end{cases}$$

for all $t \in [0,1]$, where $(B_t)_{t\in[0,1]}$ is a standard real Brownian motion independent of Y, the initial value X_0 is independent of Y and $(B_t)_{t\in[0,1]}$, and $b_0, f_0, \sigma_0 : [0,1]\times\mathbb{R} \to \mathbb{R}$ are unknown Borel functions such that each equation in (M) has a strong solution for given initial value X_0. In the sequel, for a function $r : [0,1]\times\mathbb{R}\to\mathbb{R}$, we write the stochastic integral $\int_0^t r(t,X_t)\mathrm{d}B_t$ (resp. the integral $\int_0^t r(t,X_t)\mathrm{d}t$) under the implicit condition

$$\mathbb{E}\int_0^1 r(t,X_t)^2\mathrm{d}t < \infty \text{ (resp. } \mathbb{E}\int_0^1 |r(t,X_t)|\mathrm{d}t < \infty).$$

We finally assume throughout the whole paper that a Novikov criterion is fullfiled, i.e.

$$\mathbb{E}\exp\Big(\frac{1}{2}\int_0^1 f_0(t,X_t)^2\mathrm{d}t\Big) < \infty. \tag{2.1}$$

Bayes rule – For $h = (r_1, r_2) : [0,1]\times\mathbb{R}\to\mathbb{R}^2$, let

$$F(h,X) = \int_0^1 r_1(t,X_t)\mathrm{d}X_t - \int_0^1 r_2(t,X_t)\mathrm{d}t. \tag{2.2}$$

Lemma 2.1 above is a consequence of Corollary on page 292 in the book by Lipster and Shiryaev (2001). In the sequel, μ_i stands for the conditional distribution of X given by $Y = i$, $i = 0, 1$.

Lemma 2.1. *The probability μ_1 is absolutely continuous with respect to μ_0 with density $\exp F(h_0, .)$, where*

$$h_0 = \Big(\frac{f_0}{\sigma_0}, \frac{f_0 b_0}{\sigma_0} + \frac{f_0^2}{2}\Big).$$

Based on Lemma 2.1, we can now derive the explicit form of Bayes rules $g^\star$. Recall that (see Theorem 2.1 in the book by Devroye et al, 1996):

$$g^\star(x) = \mathbf{1}\big\{\mathbb{E}(Y|X = x) > 1/2\big\},$$

hence the task is to compute $\mathbb{E}(Y|X)$. Observe that

$$\mathbb{E}(Y|X) = \mathbb{P}(Y = 1|X) = (1 - p_0)\psi(X),$$

where $p_0 = \mathbb{P}(Y = 0) \in]0, 1[$ and ψ is the density of μ_1 with respect to the distribution μ of X. Since by Lemma 2.1,

$$\mathrm{d}\mu = p_0 \mathrm{d}\mu_0 + (1 - p_0)\mathrm{d}\mu_1 = \big(p_0 + (1 - p_0)\exp F(h_0, .)\big)\mathrm{d}\mu_0,$$

we deduce that

$$\psi = \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu} = \frac{1}{p_0 + (1 - p_0)\exp F(h_0, .)} \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_0} = \frac{\exp F(h_0, .)}{p_0 + (1 - p_0)\exp F(h_0, .)}.$$

Consequently,

$$\mathbb{E}(Y|X) = \frac{(1 - p_0)\exp F(h_0, X)}{p_0 + (1 - p_0)\exp F(h_0, X)}, \tag{2.3}$$

so that Bayes rule writes as

$$g^\star = \mathbf{1}\Big\{F(h_0, .) > \ln \frac{p_0}{1 - p_0}\Big\}. \tag{2.4}$$

Remark 2.2. *Consider the particular case where coefficient f_0 of model (M) satisfy $f_0(t, x) \equiv f_0(t)$ for all $(t, x) \in [0, 1] \times \mathbb{R}$, with $f_0 \in \mathbb{L}^2([0, 1])$ a non null function. Bayes risk has the expression (see the proof in the Appendix):*

$$L(g^\star) = p_0 \bar{\Phi}\Big(-\frac{\|f_0\|}{2} + \frac{1}{\|f_0\|} \ln \frac{p_0}{1 - p_0}\Big) + (1 - p_0)\Phi\Big(-\frac{\|f_0\|}{2} + \frac{1}{\|f_0\|} \ln \frac{p_0}{1 - p_0}\Big),$$

where Φ is the distribution function of the standard normal law, $\bar{\Phi} = 1 - \Phi$ and $\|.\|$ is the $\mathbb{L}^2([0, 1])$-norm. Note that it does not depend on σ_0 and b_0.

3 Bayes risk estimation

We introduce the sets $\mathscr{F}$ and $\mathscr{F}(\varepsilon)$, for all $\varepsilon > 0$, of functions $h : [0,1] \times \mathbb{R} \to \mathbb{R}^2$. We only assume that $\mathscr{F}(\varepsilon)$ is a finite set, and we let $N(\varepsilon) = |\mathscr{F}(\varepsilon)|$. One may think of $\mathscr{F}(\varepsilon)$ as an ε-net for $\mathscr{F}$ (see Corollary 3.3 and next section), while the unknown coefficients f_0, b_0 and σ_0 of model (M) satisfy

$$h_0 = \Big(\frac{f_0}{\sigma_0}, \frac{f_0 b_0}{\sigma_0} + \frac{f_0^2}{2}\Big) \in \mathscr{F}.$$

Let $(X_1,Y_1), \cdots, (X_n,Y_n)$ be the learning sample, that is i.i.d. copies of the pair (X,Y). For $h \in \mathscr{F}$ or in $\mathscr{F}(\varepsilon)$ and $p \in]0,1[$, we let $g(h,p)$ be the classification rule defined by:

$$g(h,p)(.) = \mathbf{1}\Big\{F(h,.) > \ln\frac{p}{1-p}\Big\}.$$

Observe that by (2.4),

$$g^\star = g(h_0, p_0),$$

where $p_0 = \mathbb{P}(Y = 0)$. The empirical loss of rule $g(h,p)$ is

$$\hat{L}\big(g(h,p)\big) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\big\{g(h,p)(X_i) \neq Y_i\big\}.$$

Letting

$$I(\varepsilon) = \big\{[k\varepsilon]; k = 1, \cdots, [1/\varepsilon]\big\},$$

we then consider the pair $(\hat{h}_\varepsilon, \hat{p}_\varepsilon) \in \mathscr{F}(\varepsilon) \times I(\varepsilon)$ such that

$$\hat{L}\big(g(\hat{h}_\varepsilon, \hat{p}_\varepsilon)\big) \leq \hat{L}\big(g(h,p)\big) \ \forall (h,p) \in \mathscr{F}(\varepsilon) \times I(\varepsilon). \tag{3.1}$$

By (2.4), a natural estimator for Bayes rule constructed over the finite set $\mathscr{F}(\varepsilon)$ is then

$$\hat{g}_\varepsilon = g(\hat{h}_\varepsilon, \hat{p}_\varepsilon) = \mathbf{1}\Big\{F(\hat{h}_\varepsilon, .) > \ln\frac{\hat{p}_\varepsilon}{1-\hat{p}_\varepsilon}\Big\}.$$

Now we describe the data-driven choice for ε. Let $\lambda_\varepsilon > 0$ be a penalty term such that

$$\Delta = \sum_{\varepsilon \in \mathscr{E}} \frac{N(\varepsilon)}{\varepsilon} e^{-2\lambda_\varepsilon^2} < \infty, \text{ where } \mathscr{E} = \{1/\ell; \ell \geq 1\}.$$

For instance, the λ_ε's can be chosen so that

$$\lambda_\varepsilon^2 \geq 2\ln\frac{1}{\varepsilon} + \ln N(\varepsilon), \tag{3.2}$$

whence $\Delta \leq 2$. Then, the data-driven choice $\hat{\varepsilon}$ of ε is as follows:

$$\hat{L}\big(\hat{g}_{\hat{\varepsilon}}\big) + \frac{\lambda_{\hat{\varepsilon}}}{\sqrt{n}} \leq \hat{L}\big(\hat{g}_{\varepsilon}\big) + \frac{\lambda_{\varepsilon}}{\sqrt{n}} \ \forall \varepsilon \in \mathscr{E}, \tag{3.3}$$

and the empirical rule is then defined by

$$\hat{g} = \hat{g}_{\hat{\varepsilon}} = \mathbf{1}\Big\{F(\hat{h}_{\hat{\varepsilon}}, .) > \ln \frac{\hat{p}_{\hat{\varepsilon}}}{1 - \hat{p}_{\hat{\varepsilon}}}\Big\}. \tag{3.4}$$

In the sequel, we let $\mathscr{G}(\varepsilon) = \mathscr{F}(\varepsilon) \times I(\varepsilon)$. Next theorem aims at evaluating the loss between the performance of $\hat{g}$ and the best performance among the $g(h,p)$'s, provided $(h,p) \in \mathscr{G}(\varepsilon)$ and $\varepsilon \in \mathscr{E}$.

Theorem 3.1. *We have:*

$$\mathbb{E}L(\hat{g}) - L(g^\star) \leq \inf\Big\{L\big(g(h,p)\big) - L(g^\star) + \frac{\lambda_\varepsilon}{\sqrt{n}}\Big\} + \sqrt{\frac{\ln \Delta}{n}} + \frac{1}{\Delta\sqrt{n \ln \Delta}},$$

the infimum being taken over all $\varepsilon \in \mathscr{E}$ and $(h,p) \in \mathscr{G}(\varepsilon)$.

Observe that the error in Theorem 3.1 is decomposed into an estimation error term and a bias term, namely the term

$$\inf\Big\{L\big(g(h,p)\big) - L(g^\star) + \frac{\lambda_\varepsilon}{\sqrt{n}}\Big\}.$$

Remark 3.2. *For the statistical methodology, the main difficulty of this approach lies in the fact that the calculation of $\hat{h}_\varepsilon$ defined by* (3.1) *requires evaluations of stochastic integrals of the type*

$$\int_0^1 r(t, X_t) \mathrm{d}X_t,$$

for some function r. In the literature, many numerical methods have been developed for this purpose, see for instance the book by Kloeden and Platen (1999). They are often based on discretizations of the stochastic integrals, e.g. applying a Euler type scheme.

Under weak assumptions on class $\mathscr{F}$, the empirical rule $\hat{g}$ is consistent (see Introduction for definition), as shown by next result.

Corollary 3.3. *Assume that, on each compact set $K \subset [0,1]\times\mathbb{R}$, the restriction of $\mathscr{F}$ to K is relatively compact for the supremum norm. For all $\varepsilon > 0$, we then let $\mathscr{F}(\varepsilon)$ a finite ε-net for the restriction of $\mathscr{F}$ to $[0,1]\times[-1/\varepsilon, 1/\varepsilon]$, such that the support of each function in $\mathscr{F}(\varepsilon)$ is contained in $[0,1]\times[-1/\varepsilon, 1/\varepsilon]$. Then, the empirical rule $\hat{g}$ is consistent.*

However, such a simple approach can not give a fast estimator because the size of $\mathscr{F}(\varepsilon)$ considerably increases as ε goes to 0. Hence in the next section, which is devoted to the rate of convergence of the empirical rule $\hat{g}$, we need to adapt the construction of $\mathscr{F}(\varepsilon)$.

4 Rates for the Bayes risk

Assumptions on model (M) – In the sequel, we fix $L \geq 1$ and we let for all $a \geq 0$, ψ_a the function defined by

$$\psi_a(x) = L(1+|x|)^a \ \forall x \in \mathbb{R}.$$

We suppose that the coefficients of model (M) have polynomial grows, namely there exist $\alpha, \beta \geq 0$ such that for all $(t,x) \in [0,1]\times\mathbb{R}$:

$$|b_0(t,x)| + |(f_0\sigma_0)(t,x)| \leq \psi_\alpha(x) \text{ and } |\sigma_0(t,x)| \leq \psi_\beta(x). \tag{4.1}$$

Finally, $s \geq 0$ is such that

$$B = \mathbb{E}\exp\left(16L^4 \int_0^1 (1+|X_t|)^s \mathrm{d}t\right) < \infty. \tag{4.2}$$

Huang (2009) proved that this assumption is satisfied by a wide class of diffusion processes, such as the Black-Scholes model for which $s = 2$ and $\alpha = \beta = 1$, even in the case where the drift and the volatility depend on the time variable.

Sets $\mathscr{F}$ and $\mathscr{F}(\varepsilon)$ – For all $a \geq 0$, introduce the set $\mathscr{F}_a$ of real-valued functions on $[0,1]\times\mathbb{R}$ such that:

- ▷ for all compact $K \subset [0,1]\times\mathbb{R}$, the restriction of $\mathscr{F}_a$ to K, equipped with the supremum norm, is relatively compact;
- ▷ for all $\varphi \in \mathscr{F}_a$, $|\varphi(t,x)| \leq \psi_a(x)\ \forall (t,x) \in [0,1]\times\mathbb{R}$.

For all $\varepsilon > 0$ and $j \in \mathbb{Z}$, we then let $\mathscr{F}_{a,j}(\varepsilon)$ be a finite ε-net for $\mathscr{F}_{a,j}$, the restriction of $\mathscr{F}_a$ to $[0,1] \times [j, j+1[$ equipped with the supremum norm. Moreover, $N_{a,j}(\varepsilon) = |\mathscr{F}_{a,j}(\varepsilon)|$.

Now introduce the measure Q on $[0,1] \times \mathbb{R}$ defined for all borel sets $I \subset [0,1]$ and $A \subset \mathbb{R}$ by

$$Q(I \times A) = L^4 \mathbb{E} \int_I \mathbf{1}\{X_t \in A\} \big(1 + |X_t|\big)^q \mathrm{d}t.$$

In the sequel, we fix $q \geq 1$ large enough so that $2\max(\alpha, \beta) \leq q$, and we denote:

$$M = \sup_{t \in [0,1]} \mathbb{E}\big(1 + |X_t|\big)^{2q} < \infty.$$

The following result is proved in Section 5.3.

Proposition 4.1. *Let $a \geq 0$, $\varepsilon > 0$ and $2\ell \leq q - 2$. With the sole knowledge of M, one can construct an ε-net $\mathscr{F}_a(\varepsilon)$ for $\mathscr{F}_a$ equipped with the $\mathbb{L}^2(Q)$-norm so that*

$$|\mathscr{F}_a(\varepsilon)| = \prod_{j \in \mathbb{Z}} N_{a,j}\Big(\frac{\varepsilon \max(|j|,1)^\ell}{\sqrt{6ML^4}}\Big),$$

and, for all $\varphi \in \mathscr{F}_a(\varepsilon)$, $|\varphi(t,x)| \leq \psi_a(x)$ $\forall (t,x) \in [0,1] \times \mathbb{R}$.

Remark 4.2. **(1)** *An explicit construction of $\mathscr{F}_a(\varepsilon)$ is given in the proof of Proposition 4.1. In the usual cases (see the examples below), the set $\mathscr{F}_a(\varepsilon)$ can be described from the classical ε-nets.* **(2)** *Note that M is unknown; however, it can be consistently estimated from an i.i.d. sample drawn from X. Hence, in the sequel, we assume that M, hence the $\mathscr{F}_a(\varepsilon)$'s, are at hand.*

Now fix $a, b \geq 0$. The sets $\mathscr{F}$ and $\mathscr{F}(\varepsilon)$ of the previous section are defined by:

$$\mathscr{F} = \mathscr{F}_a \times \mathscr{F}_b \text{ and } \mathscr{F}(\varepsilon) = \mathscr{F}_a(\varepsilon) \times \mathscr{F}_b(\varepsilon).$$

Recall that we assumed that $h_0 = (f_0/\sigma_0, f_0 b_0/\sigma_0 + f_0^2/2) \in \mathscr{F}$.

Combining the observations of Remark 3.2 and 4.2, we then have developped a statistical methodology for nonparametric supervised classification of diffusion paths. However, we realize that the algorithm is time-consuming, so that a further study is needed to improve this drawback.

Rate of convergence – In the setting described above, one may specify the bias term in the inequality of Theorem 3.1, as shown by next result.

Theorem 4.3. *Assume that* (4.1) *and* (4.2) *hold, and that b,* $2(a+\beta)$ *and* $a+\alpha$ *do not exceed s. Then,*

$$\mathbb{E}L(\hat{g}) - L(g^\star) \leq \inf_{\varepsilon\in\mathscr{E}} \Big(K\varepsilon + \frac{\lambda_\varepsilon}{\sqrt{n}} \Big) + \sqrt{\frac{\ln\Delta}{n}} + \frac{1}{\Delta\sqrt{n\ln\Delta}},$$

where $K = 8(1+2B^{1/4})$.

As an example, even in the case where the drift and the volatility depend on the time variable, Black-Scholes model has parameters $a = b = 0$ and $\alpha = \beta = 1$, so that conditions of Theorem 4.3 hold.

We now give an explicit bound in the important case where $\ln N(\varepsilon) \leq C\varepsilon^{-u}$ for some $u, C > 0$. Using a penalty term λ_ε as in (3.2), namely

$$\lambda_\varepsilon^2 = 2\ln\frac{1}{\varepsilon} + \ln N(\varepsilon),$$

it is an easy exercise to prove that provided $C \geq 1$:

$$\inf_{\varepsilon\in\mathscr{E}} \Big(K\varepsilon + \frac{\lambda_\varepsilon}{\sqrt{n}} \Big) \leq 2\Big(\frac{K^u}{n}\Big)^{1/(2+u)} \Big(\frac{2}{u} + C\Big)^2.$$

Hence we have the following result :

Corollary 4.4. *Assume that there exist* $u > 0$ *and* $C \geq 1$ *such that* $\ln N(\varepsilon) \leq C\varepsilon^{-u}$ *for every* $\varepsilon > 0$. *Then, under the conditions of Theorem 4.3, we have:*

$$\mathbb{E}L(\hat{g}) - L(g^\star) \leq 2\Big(\frac{K^u}{n}\Big)^{1/(2+u)} \Big(\frac{2}{u} + C\Big)^2 + \sqrt{\frac{\ln\Delta}{n}} + \frac{1}{\Delta\sqrt{n\ln\Delta}}.$$

Example 1. *Homogeneous diffusions* – As an illustration, consider the case of homogeneous diffusions, i.e. the coefficients of model (M) satisfy $f_0(t,x) = f_0(x)$, $b_0(t,x) = b_0(x)$ and $\sigma_0(t,x) = \sigma_0(x)$ for all $(t,x) \in [0,1]\times\mathbb{R}$. Also assume that the sets $\mathscr{F}_a$ and $\mathscr{F}_b$ only contain functions of class $\mathscr{C}^k$ such that there exists $A > 0$ satisying for all $j \in \mathbb{Z}$ and $i = 0, \cdots, k$:

$$\sup_{[j,j+1[} \Big| \frac{\mathrm{d}^i f}{\mathrm{d}x^i} \Big| \leq A(|j|+1)^r, \text{ if } f \in \mathscr{F}_a \cup \mathscr{F}_b.$$

For any $j \in \mathbb{Z}$, consider the ε-nets $\mathscr{F}_{a,j}(\varepsilon)$ and $\mathscr{F}_{b,j}(\varepsilon)$ of the restrictions of $\mathscr{F}_a$ and $\mathscr{F}_b$ to $[0,1]\times[j,j+1[$, as constructed in the proof of Theorem 2.7.1 in the

book by van der Vaart and Wellner (1996). We deduce from Proposition 4.1 (letting $\ell \geq r+2$) that for every $\varepsilon > 0$, $\ln N(\varepsilon) \leq C\varepsilon^{-1/k}$ where $C > 0$ only depends on k, a, b, M and L. Corollary 4.4 then gives:

$$\mathbb{E}L(\hat{g}) - L(g^\star) \leq \frac{C'}{n^{k/(2k+1)}},$$

for some $C' > 0$. Note that many usual models (Black-Scholes model, Langevin equation...) often have $\mathscr{C}^\infty$ coefficients, so that any value for k is suitable.

Example 2. Now assume that the coefficients of model (M) satisfy $f_0(t,x) = f_0(t)$, $b_0(t,x) = b_0(t)$ and $\sigma_0(t,x) = \sigma_0(t)$ for all $(t,x) \in [0,1] \times \mathbb{R}$. Also assume that $h_0 = (f_0/\sigma_0, f_0 b_0/\sigma_0 + f_0^2/2)$ belongs to some set $\mathscr{F}$ of functions of class $\mathscr{C}^k$, with uniformly bounded derivatives. Similar to the previous example, we deduce from Theorem 2.7.1 in the book by van der Vaart and Wellner (1996) that for every $\varepsilon > 0$, $\ln N(\varepsilon) \leq C\varepsilon^{-1/k}$ where $C > 0$ only depends on k, M and L. Therefore, by Corollary 4.4:

$$\mathbb{E}L(\hat{g}) - L(g^\star) \leq \frac{C''}{n^{k/(2k+1)}},$$

for some $C'' > 0$. Notice that in this case, $\alpha = \beta = a = b = 0$, hence $s = 0$ is suitable.

5 Proofs

5.1 Proof of Theorem 3.1

Let $\rho > 0$. By (3.3), we have for all $\varepsilon \in \mathscr{E}$:

$$\begin{aligned}
\mathbb{P}\Big(L(\hat{g}) - \hat{L}(\hat{g}_\varepsilon) > \frac{\lambda_\varepsilon}{\sqrt{n}} + \rho\Big) &\leq \mathbb{P}\Big(L(\hat{g}) - \hat{L}(\hat{g}) > \frac{\lambda_{\hat{\varepsilon}}}{\sqrt{n}} + \rho\Big) \\
&\leq \sum_{\varepsilon \in \mathscr{E}} \mathbb{P}\Big(L(\hat{g}_\varepsilon) - \hat{L}(\hat{g}_\varepsilon) > \frac{\lambda_\varepsilon}{\sqrt{n}} + \rho\Big) \\
&\leq \sum_{\varepsilon \in \mathscr{E}} \frac{N(\varepsilon)}{\varepsilon} \exp\Big(-2n\Big(\frac{\lambda_\varepsilon}{\sqrt{n}} + \rho\Big)^2\Big) \\
&\leq \Delta e^{-2n\rho^2},
\end{aligned}$$

according to the Hoeffding Inequality (Theorem 8.1 in the book by Devroye et al, 1996), and because $|\mathscr{G}(\varepsilon)| = |\mathscr{F}(\varepsilon) \times I(\varepsilon)| \leq N(\varepsilon)/\varepsilon$. Since for all $\varepsilon \in \mathscr{E}$,

$$\mathbb{E}L(\hat{g}) \leq \mathbb{E}\hat{L}(\hat{g}_\varepsilon) + \frac{\lambda_\varepsilon}{\sqrt{n}} + \int_0^\infty \mathbb{P}\Big(L(\hat{g}) - \hat{L}(\hat{g}_\varepsilon) > \frac{\lambda_\varepsilon}{\sqrt{n}} + \rho\Big) \mathrm{d}\rho,$$

we obtain, for all $u > 0$,

$$\mathbb{E}L(\hat{g}) \leq \mathbb{E}\hat{L}(\hat{g}_\varepsilon) + \frac{\lambda_\varepsilon}{\sqrt{n}} + u + \int_u^\infty e^{-2n\rho^2} \mathrm{d}\rho.$$

Now use the bound

$$\int_u^\infty e^{-2n\rho^2} \mathrm{d}\rho \leq \frac{1}{4nu} e^{-2nu^2},$$

to get, with the choice $u = \sqrt{(\ln \Delta)/(2n)}$:

$$\mathbb{E}L(\hat{g}) \leq \mathbb{E}\hat{L}(\hat{g}_\varepsilon) + \frac{\lambda_\varepsilon}{\sqrt{n}} + \sqrt{\frac{\ln \Delta}{n}} + \frac{1}{\Delta\sqrt{n \ln \Delta}},$$

for every $\varepsilon \in \mathscr{E}$. Since, by (3.1),

$$\mathbb{E}\hat{L}(\hat{g}_\varepsilon) \leq \mathbb{E}\hat{L}\big(g(h,p)\big) = L\big(g(h,p)\big),$$

for all $(h,p) \in \mathscr{G}(\varepsilon)$, we therefore obtain:

$$\mathbb{E}L(\hat{g}) - L(g^\star) \leq \inf_{\varepsilon \in \mathscr{E}} \inf_{(h,p) \in \mathscr{G}(\varepsilon)} \Big\{ L\big(g(h,p)\big) - L(g^\star) + \frac{\lambda_\varepsilon}{\sqrt{n}} \Big\} + \sqrt{\frac{\ln \Delta}{n}} + \frac{1}{\Delta\sqrt{n \ln \Delta}},$$

which is the desired result. □

5.2 Proof of Corollary 3.3

We begin the subsection with a lemma. From now on, $\|.\|$ is the Euclidean norm on $\mathbb{R}^2$.

Lemma 5.1. *For all $\varepsilon \in \mathscr{E}$, let $p_\varepsilon \in I(\varepsilon)$ and $h_\varepsilon \in \mathscr{F}(\varepsilon)$ be such that*

$$|p_\varepsilon - p_0| \leq \varepsilon \text{ and } \sup_{(t,x) \in [0,1] \times [-1/\varepsilon, 1/\varepsilon]} \|h_\varepsilon(t,x) - h_0(t,x)\| \leq \varepsilon.$$

Then, $F(h_\varepsilon, X)$ converges to $F(h_0, X)$ in probability when ε tends to 0, with $\varepsilon \in \mathscr{E}$.

Proof. In the sequel, $\varepsilon \in \mathscr{E}$. Letting $h_\varepsilon = (r_{1,\varepsilon}, r_{2,\varepsilon})$, $h_0 = (r_{1,0}, r_{2,0})$ and omitting the dependency of (t, X_t) in the above expressions, observe that when $Y = i$:

$$F(h_\varepsilon, X) - F(h_0, X) = \int_0^1 (r_{1,\varepsilon} - r_{1,0})\big(\sigma_0 \mathrm{d}B_t + \gamma_i \mathrm{d}t\big) - \int_0^1 (r_{2,\varepsilon} - r_{2,0})\mathrm{d}t, \quad (5.1)$$

where $\gamma_0 = b_0$ and $\gamma_1 = b_0 + f_0\sigma_0$. For all $\rho > 0$:

$$\begin{aligned}
\mathbb{P}\Big(\Big|\int_0^1 \big(r_{2,\varepsilon} - r_{2,0}\big)\mathrm{d}t\Big| \geq \rho\Big) &\leq \mathbb{P}\Big(\int_0^1 \big|r_{2,\varepsilon} - r_{2,0}\big|\mathrm{d}t \geq \rho, \sup_{t\in[0,1]} |X_t| \leq 1/\varepsilon\Big) \\
&\quad + \mathbb{P}\Big(\sup_{t\in[0,1]} |X_t| \geq 1/\varepsilon\Big) \\
&\leq \mathbb{P}\Big(\sup_{t\in[0,1]} |X_t| \geq 1/\varepsilon\Big), \qquad (5.2)
\end{aligned}$$

as soon as $\varepsilon < \rho$, since $\int_0^1 |r_{2,\varepsilon} - r_{2,0}|\mathrm{d}t \leq \varepsilon$ when $\sup_{t\in[0,1]} |X_t| \leq 1/\varepsilon$. Similarly,

$$\begin{aligned}
\mathbb{P}\Big(\Big|\int_0^1 \big(r_{1,\varepsilon} - r_{1,0}\big)\gamma_i \mathrm{d}t\Big| \geq \rho\Big) &\leq \mathbb{P}\Big(\varepsilon \int_0^1 |\gamma_i|\mathrm{d}t \geq \rho, \sup_{t\in[0,1]} |X_t| \leq 1/\varepsilon\Big) \\
&\quad + \mathbb{P}\Big(\sup_{t\in[0,1]} |X_t| \geq 1/\varepsilon\Big) \\
&\leq \frac{\varepsilon}{\rho}\int_0^1 \mathbb{E}|\gamma_i|\mathrm{d}t + \mathbb{P}\Big(\sup_{t\in[0,1]} |X_t| \geq 1/\varepsilon\Big), \quad (5.3)
\end{aligned}$$

according to the Markov Inequality. Finally, since for all bounded stopping time T:

$$\mathbb{E}\Big(\int_0^T \big(r_{1,\varepsilon} - r_{1,0}\big)\sigma_0 \mathrm{d}B_t\Big)^2 = \mathbb{E}\int_0^T \big(r_{1,\varepsilon} - r_{1,0}\big)^2 \sigma_0^2 \mathrm{d}t,$$

we get with Lemma 4.6 in the book by Revuz and Yor (1999):

$$\begin{aligned}
&\mathbb{P}\Big(\Big(\int_0^1 \big(r_{1,\varepsilon} - r_{1,0}\big)\sigma_0 \mathrm{d}B_t\Big)^2 \geq \rho, \sup_{t\in[0,1]} |X_t| \leq 1/\varepsilon, \int_0^1 \sigma_0^2 \mathrm{d}t \leq 1/\varepsilon\Big) \\
&\leq \mathbb{P}\Big(\Big(\int_0^1 \big(r_{1,\varepsilon} - r_{1,0}\big)\sigma_0 \mathrm{d}B_t\Big)^2 \geq \rho, \int_0^1 \big(r_{1,\varepsilon} - r_{1,0}\big)^2 \sigma_0^2 \mathrm{d}t \leq \varepsilon\Big) \\
&\leq \frac{\varepsilon}{\rho}.
\end{aligned}$$

Therefore, using the fact that $\sup_{t\in[0,1]}|X_t|$ and $\int_0^1 \sigma_0^2 \mathrm{d}t$ are almost surely finite random variables, we get

$$\lim_{\varepsilon\to 0}\int_0^1 \big(r_{1,\varepsilon}-r_{1,0}\big)\sigma_0 \mathrm{d}B_t = 0,$$

in probability. We deduce from (5.2)-(5.3) that $F(h_\varepsilon,X)$ converges to $F(h_0,X)$ in probability, when ε tends to 0. □

Proof of Corollary 3.3. In the sequel, $\varepsilon \in \mathscr{E}$. Let $p_\varepsilon \in I(\varepsilon)$ and $h_\varepsilon \in \mathscr{F}(\varepsilon)$ be such that

$$|p_\varepsilon - p_0| \le \varepsilon \text{ and } \sup_{[0,1]\times[-1/\varepsilon,1/\varepsilon]} \|h_\varepsilon - h_0\| \le \varepsilon.$$

Consider the rule $g_\varepsilon = \mathbf{1}\{\eta_\varepsilon > 1/2\}$, where

$$\eta_\varepsilon(X) = \frac{(1-p_\varepsilon)\exp F(h_\varepsilon,X)}{p_\varepsilon + (1-p_\varepsilon)\exp F(h_\varepsilon,X)}.$$

By Lemma 5.1, $F(h_\varepsilon,X)$ converges to $F(h_0,X)$ in probability as ε tends to 0, hence $\eta_\varepsilon(X)$ converges in probability to

$$\mathbb{E}(Y|X) = \frac{(1-p_0)\exp F(h_0,X)}{p_0 + (1-p_0)\exp F(h_0,X)},$$

according to (2.3). Since $\eta_\varepsilon(X)$ is bounded by 1, the convergence also holds in $\mathbb{L}^1(\mathbb{P})$. Recalling now that by Theorem 2.2 in the book by Devroye et al (1996), we have

$$L(g_\varepsilon) - L(g^\star) \le 2\mathbb{E}|\mathbb{E}(Y|X) - \eta_\varepsilon(X)|,$$

we deduce that

$$\lim_{\varepsilon\to 0}\inf_{(h,p)\in\mathscr{G}(\varepsilon)} L\big(g(h,p)\big) = L(g^\star).$$

Therefore,

$$\lim_{n\to\infty}\inf_{\varepsilon\in\mathscr{E}}\inf_{(h,p)\in\mathscr{G}(\varepsilon)}\Big\{L\big(g(h,p)\big) - L(g^\star) + \frac{\lambda_\varepsilon}{\sqrt{n}}\Big\} = 0,$$

hence the result, using Theorem 3.1. □

5.3 Proof of Proposition 4.1

For all $j \in \mathbb{Z}$, let

$$I_j = [0,1] \times [j, j+1[,\ \varepsilon_j = \frac{\varepsilon \max(|j|,1)^\ell}{\sqrt{6ML^4}},\ \bar{N}_j = N_{a,j}(\varepsilon_j)$$

and $\{f_{j,1}, \cdots, f_{j,\bar{N}_j}\}$ be the ε_j-net for $\mathscr{F}_{a,j}$ equipped with the supremum norm. For all $j \in \mathbb{Z}$ and $i = 1, \cdots, \bar{N}_j$, we let

$$\tilde{f}_{j,i} = \min(f_{j,i}, \psi_a)\mathbf{1}\{f_{j,i} \geq 0\} + \max(f_{j,i}, -\psi_a)\mathbf{1}\{f_{j,i} < 0\}.$$

Observe that $|\tilde{f}_{j,i}| \leq \psi_a$. Hence, we only need to prove that the set of functions

$$\mathscr{F}_a(\varepsilon) = \Big\{ \sum_{j \in \mathbb{Z}} \tilde{f}_{j,k_j} \mathbf{1}_{I_j},\, k_j \in \{1, \cdots, \bar{N}_j\} \ \text{ for all } j \in \mathbb{Z} \Big\},$$

forms an ε-net for $\mathscr{F}_a$ endowed with the $\mathbb{L}^2(Q)$-norm. Let $f \in \mathscr{F}_a$. For all $j \in \mathbb{Z}$, there exists $k_j = 1, \cdots, \bar{N}_j$ such that

$$\sup_{I_j} |f - f_{j,k_j}| \leq \varepsilon_j.$$

Consequently, by definition of the $\tilde{f}_{j,i}$'s,

$$\sup_{I_j} |f - \tilde{f}_{j,k_j}| \leq \varepsilon_j.$$

Since the I_j's form a partition of $[0,1] \times \mathbb{R}$, we have:

$$\begin{aligned}
\big\| f - \sum_{j \in \mathbb{Z}} \tilde{f}_{j,k_j} \mathbf{1}_{I_j} \big\|^2_{\mathbb{L}^2(Q)} &= \int \Big(\sum_{j \in \mathbb{Z}} (f - \tilde{f}_{j,k_j}) \mathbf{1}_{I_j} \Big)^2 \mathrm{d}Q \\
&= \sum_{j \in \mathbb{Z}} \int_{I_j} (f - \tilde{f}_{j,k_j})^2 \mathrm{d}Q \\
&\leq \frac{\varepsilon^2}{6ML^4} \Big(1 + \sum_{j \in \mathbb{Z}} |j|^{2\ell} Q(I_j) \Big).
\end{aligned}$$

Let $j \in \mathbb{Z}$ be different from 0 and -1. By definition of Q and M, we get

$$\begin{aligned} Q(I_j) &= L^4 \int_0^1 \mathbb{E}\mathbf{1}\{X_t \in [j, j+1[\}(1+|X_t|)^q \mathrm{d}t \\ &\leq L^4 \int_0^1 \sqrt{\mathbb{P}(j \leq X_t < j+1)\mathbb{E}(1+|X_t|)^{2q}}\, \mathrm{d}t \\ &\leq \frac{L^4 M^{1/2}}{j^q} \int_0^1 \left(\mathbb{E}|X_t|^{2q}\right)^{1/2} \mathrm{d}t \\ &\leq \frac{L^4 M}{j^q}, \end{aligned}$$

using the Cauchy-Schwarz Inequality. Consequently,

$$\begin{aligned} \Big\| f - \sum_{j\in\mathbb{Z}} \tilde{f}_{j,k_j} \mathbf{1}_{I_j} \Big\|^2_{\mathbb{L}^2(Q)} &\leq \frac{\varepsilon^2}{6ML^4}\Big(2 + 2L^4 M \sum_{j\geq 1} j^{2\ell-q}\Big) \\ &\leq \frac{\varepsilon^2}{6ML^4}\Big(2 + 2L^4 M \sum_{j\geq 1} j^{-2}\Big) \\ &\leq \varepsilon^2, \end{aligned}$$

because $q \geq 2\ell+2$ and $L, M \geq 1$. Thus $\mathscr{F}_a(\varepsilon)$ is an ε-net for $\mathscr{F}_a$. □

5.4 Proof of Theorem 4.3

We begin the subsection with two lemmas.

Lemma 5.2. *Let $h_\varepsilon = (r_{1,\varepsilon}, r_{2,\varepsilon}) \in \mathscr{F}(\varepsilon)$ be such that*

$$\|f_0/\sigma_0 - r_{1,\varepsilon}\|_{\mathbb{L}^2(Q)} \leq \varepsilon \text{ and } \|f_0 b_0/\sigma_0 + f_0^2/2 - r_{2,\varepsilon}\|_{\mathbb{L}^2(Q)} \leq \varepsilon.$$

Then, if $2\max(\alpha,\beta) \leq q$:

$$\mathbb{E}\big(F(h_\varepsilon, X) - F(h_0, X)\big)^2 \leq 9\varepsilon^2.$$

Proof. For notational simplicity, we let $r_1 = f_0/\sigma_0$, $r_2 = f_0 b_0/\sigma_0 + f_0^2/2$, and we omit the dependency on (t, X_t) in the expressions of the functions r_1, r_2 as well as for $r_{1,\varepsilon}$ and $r_{2,\varepsilon}$. Observe that when $Y = i$:

$$F(h_\varepsilon, X) - F(h_0, X) = \int_0^1 (r_{1,\varepsilon} - r_1)\big(\sigma_0 \mathrm{d}B_t + \gamma_i \mathrm{d}t\big) - \int_0^1 (r_{2,\varepsilon} - r_2)\mathrm{d}t, \quad (5.4)$$

where $\gamma_0 = b_0$ and $\gamma_1 = b_0 + f_0\sigma_0$. First consider the first term on the right-hand side of (5.4). Since $(\int_0^s (r_{1,\varepsilon} - r_1)\sigma_0 \mathrm{d}B_t)_{s\in[0,1]}$ is a martingale with quadratic variation $(\int_0^s (r_{1,\varepsilon} - r_1)^2\sigma_0^2 \mathrm{d}t)_{s\in[0,1]}$, the definition of Q gives:

$$\begin{aligned}\mathbb{E}\Big(\int_0^1 (r_{1,\varepsilon} - r_1)\sigma_0 \mathrm{d}B_t\Big)^2 &= \mathbb{E}\int_0^1 (r_{1,\varepsilon} - r_1)^2\sigma_0^2 \mathrm{d}t \\ &\leq \|r_{1,\varepsilon} - r_1\|^2_{\mathbb{L}^2(Q)} \\ &\leq \varepsilon^2, \end{aligned} \tag{5.5}$$

because $|\sigma_0| \leq \psi_\beta$ with $2\beta \leq q$, and $L \geq 1$. We now turn to the second term on the right-hand side of (5.4). Observing that $|\gamma_i| \leq \psi_\alpha$, we deduce from the Jensen Inequality that, since $2\alpha \leq q$:

$$\begin{aligned}\mathbb{E}\Big(\int_0^1 (r_{1,\varepsilon} - r_1)\gamma_i \mathrm{d}t\Big)^2 &\leq \mathbb{E}\int_0^1 (r_{1,\varepsilon} - r_1)^2\gamma_i^2 \mathrm{d}t \\ &\leq \|r_{1,\varepsilon} - r_1\|^2_{\mathbb{L}^2(Q)} \\ &\leq \varepsilon^2. \end{aligned} \tag{5.6}$$

Similarly,

$$\mathbb{E}\Big(\int_0^1 (r_{2,\varepsilon} - r_2)\mathrm{d}t\Big)^2 \leq \varepsilon^2. \tag{5.7}$$

We conclude from (5.4)-(5.7) that

$$\mathbb{E}\big(F(h_\varepsilon, X) - F(h_0, X)\big)^2 \leq 9\varepsilon^2,$$

hence the lemma. □

Lemma 5.3. *Let $h = (r_1, r_2) : [0,1] \times \mathbb{R} \to \mathbb{R}^2$ satisfying $|r_1| \leq \psi_a$ and $|r_2| \leq \psi_b$, and let $s \geq 0$ be such that*

$$B = \mathbb{E}\exp\Big(16L^4 \int_0^1 (1 + |X_t|)^s \mathrm{d}t\Big) < \infty.$$

If $2(a+\beta)$, $a+\alpha$ and b do not exceed s, we have:

$$\mathbb{E}\exp\big(2F(h, X)\big) \leq B^{1/2}.$$

Proof of Lemma 5.3. In the sequel, we omit the dependency on (t,X_t) in the expressions of the functions r_1, r_2 and f_0, b_0 and σ_0, and we let $\gamma_0 = b_0$ and $\gamma_1 = b_0 + f_0\sigma_0$. Observe that when $Y = i$:

$$\begin{aligned}\exp\big(2F(h,X)\big) &= \exp\Big(2\int_0^1 r_1\sigma_0 \mathrm{d}B_t + 2\int_0^1 r_1\gamma_i \mathrm{d}t - 2\int_0^1 r_2 \mathrm{d}t\Big) \\ &= \mathscr{E}_1 \exp\Big(4\int_0^1 r_1^2\sigma_0^2 \mathrm{d}t + 2\int_0^1 r_1\gamma_i \mathrm{d}t - 2\int_0^1 r_2 \mathrm{d}t\Big),\end{aligned}$$

where $(\mathscr{E}_s)_{s\in[0,1]}$ is the process defined for all $s \in [0,1]$ by

$$\mathscr{E}_s = \exp\Big(2\int_0^s r_1\sigma_0 \mathrm{d}B_t - 4\int_0^s r_1^2\sigma_0^2 \mathrm{d}t\Big).$$

Observe that the stochastic exponential process $(\mathscr{E}_s^2)_{s\in[0,1]}$ is a martingale; indeed,

$$\mathbb{E}\exp\Big(4\int_0^1 r_1^2\sigma_0^2 \mathrm{d}t\Big) \leq \mathbb{E}\exp\Big(4L^4\int_0^1 (1+|X_t|)^{2(a+\beta)} \mathrm{d}t\Big) \leq B,$$

since $2(a+\beta) \leq s$, implying the Novikov Criterion (Proposition 1.15, Chapter 8 in the book by Revuz and Yor, 1999). Hence $\mathbb{E}\mathscr{E}_1^2 = 1$, so that by the Cauchy-Schwarz Inequality:

$$\begin{aligned}\big(\mathbb{E}\exp\big(2F(h,X)\big)\big)^2 &\leq \mathbb{E}\exp\Big(8\int_0^1 r_1^2\sigma_0^2 \mathrm{d}t + 4\int_0^1 r_1\gamma_i \mathrm{d}t - 4\int_0^1 r_2 \mathrm{d}t\Big) \\ &\leq \mathbb{E}\exp\Big(16L^4\int_0^1 (1+|X_t|)^s \mathrm{d}t\Big),\end{aligned}$$

provided $2(a+\beta)$, $a+\alpha$ and b are smaller than s, hence the lemma. □

Proof of Theorem 4.3. By Theorem 3.1, we only need to bound the term

$$m = \inf_{\varepsilon\in\mathscr{E}} \inf_{(h,p)\in\mathscr{G}(\varepsilon)} \Big\{L\big(g(h,p)\big) - L(g^\star) + \frac{\lambda_\varepsilon}{\sqrt{n}}\Big\}.$$

For all $\varepsilon \in \mathscr{E}$, let $h_\varepsilon = (r_{1,\varepsilon}, r_{2,\varepsilon}) \in \mathscr{F}(\varepsilon)$ be such that

$$\|f_0/\sigma_0 - r_{1,\varepsilon}\|_{\mathbb{L}^2(Q)} \leq \varepsilon \text{ and } \|f_0 b_0/\sigma_0 + f_0^2/2 - r_{2,\varepsilon}\|_{\mathbb{L}^2(Q)} \leq \varepsilon,$$

and $p_\varepsilon \in I(\varepsilon)$ be such that $|p_\varepsilon - p_0| \leq \varepsilon$. Letting $g_\varepsilon = g(h_\varepsilon, p_\varepsilon)$, we have

$$m \leq L(g_\varepsilon) - L(g^\star) + \frac{\lambda_\varepsilon}{\sqrt{n}} \ \forall \varepsilon \in \mathscr{E},$$

hence one only needs to bound the term $L(g_\varepsilon) - L(g^\star)$. Similarly to (2.4), observe that $g_\varepsilon = \mathbf{1}\{\eta_\varepsilon > 1/2\}$, where

$$\eta_\varepsilon(X) = \frac{(1-p_\varepsilon)\exp F(h_\varepsilon, X)}{p_\varepsilon + (1-p_\varepsilon)\exp F(h_\varepsilon, X)}.$$

Therefore, by Theorem 2.2 in the book by Devroye et al (1996), we have

$$L(g_\varepsilon) - L(g^\star) \leq 2\mathbb{E}|\mathbb{E}(Y|X) - \eta_\varepsilon(X)|.$$

Moreover, easy calculations leads to the inequality :

$$\Big|\frac{(1-x)e^y}{x+(1-x)e^y} - \frac{(1-x')e^{y'}}{x'+(1-x')e^{y'}}\Big| \leq \Big(\frac{1}{2} + \max\big(e^y, e^{-y}\big)\Big)\big(|x-x'| + |y-y'|\big)$$

for all $x, x' \in [0,1]$ and $y, y' \in \mathbb{R}$. Thus, using the explicit representation of $\mathbb{E}(Y|X)$ given in (2.3), we deduce that

$$|\mathbb{E}(Y|X) - \eta_\varepsilon(X)| \quad \leq \quad \Big(\frac{1}{2} + H\Big)\big(|F(h_\varepsilon, X) - F(h_0, X)| + |p_\varepsilon - p_0|\big),$$

where

$$H = \max\big(\exp(F(h_0, X)), \exp(-F(h_0, X))\big).$$

According to the Cauchy-Schwarz Inequality and Lemma 5.2, we get

$$\begin{aligned} L(g_\varepsilon) - L(g^\star) &\leq \big(1 + 2\sqrt{\mathbb{E}H^2}\big)\Big(\sqrt{\mathbb{E}\big(F(h_\varepsilon, X) - F(h_0, X)\big)^2} + |p_\varepsilon - p_0|\Big) \\ &\leq 8\big(1 + \sqrt{\mathbb{E}H^2}\big)\varepsilon. \end{aligned}$$

Since $|r_{1,\varepsilon}|$ and $|r_{2,\varepsilon}|$ are respectively bounded by ψ_a and ψ_b by construction of the ε-nets $\mathscr{F}_a(\varepsilon)$ and $\mathscr{F}_b(\varepsilon)$ (see Proposition 4.1), and similarly for $|f_0/\sigma_0|$, $|f_0 b_0/\sigma_0 + f_0^2/2|$ by assumption, we deduce from Lemma 5.3 that

$$\begin{aligned} \sqrt{\mathbb{E}H^2} &\leq \sqrt{\mathbb{E}\exp(2F(h_0, X))} + \sqrt{\mathbb{E}\exp(-2F(h_0, X))} \\ &\leq 2\Big[\mathbb{E}\exp\Big(16L^4 \int_0^1 (1 + |X_t|)^s \mathrm{d}t\Big)\Big]^{1/4}, \end{aligned}$$

hence the theorem. □

6 Appendix : proof of equality in Remark 2.2

Note that

$$L(g^\star) = p_0\mathbb{P}(g^\star(X) \neq Y|Y=0) + (1-p_0)\mathbb{P}(g^\star(X) \neq Y|Y=1).$$

Compute for instance

$$\begin{aligned}\mathbb{P}(g^\star(X) \neq Y|Y=1) &= \mathbb{P}\Big(F(h_0,X) \leq \ln\frac{p_0}{1-p_0}|Y=1\Big) \\ &= \mathbb{P}\Big(\int_0^1 f_0(t)\mathrm{d}B_t + \frac{1}{2}\|f_0\|^2 \leq \ln\frac{p_0}{1-p_0}|Y=1\Big).\end{aligned}$$

Since $\int_0^1 f_0(t)\mathrm{d}B_t \sim \mathcal{N}(0,\|f_0\|^2)$, we then have

$$\mathbb{P}(g^\star(X) \neq Y|Y=1) = \Phi\Big(-\frac{\|f_0\|}{2} + \frac{1}{\|f_0\|}\ln\frac{p_0}{1-p_0}\Big),$$

and similarly for $\mathbb{P}(g^\star(X) \neq Y|Y=0)$. □

References

Abraham, C., Biau, G. and Cadre, B. (2003). On the Kernel Rule for Function Classification, *Ann. Inst. Statist. Math.*, pp. 619-633.

Baìllo, A., Cuesta-Alberto, J. and Cuevas, A. (2011). Supervised Classification for a Family of Gaussian Functional Models, *Scand. J. Statist.*, pp. 480-498.

Biau, G., Bunea, F. and Wegkamp, M.H. (2005). Functional Classification in Hilbert Spaces, *IEEE Trans. Inf. Theory*, pp. 2163-2172.

Cérou, F. and Guyader, A. (2006). Nearest Neighbor Classification in Infinite Dimension, *ESAIM P&S*, pp. 340-355.

Chonavel, T. (2002). *Statistical Signal Processing*, Springer-Verlag, New-York.

Cover, T.M. and Hart, P.E. (1967). Nearest Neighbor Pattern Classification, *IEEE Trans. Inf. Theory*, pp. 13-21.

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

Ferraty, F. and Vieu, P. (2003). Curves Discrimination: a Nonparametric Functional Approach, *Comp. Statist. Data Anal.*, pp. 161-173.

Hall, P., Poskitt, D.S. and Presnell, B. (2001). A Functional Data-Analytic Approach to Signal Discrimination, *Technometrics*, pp. 1-9.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized Discriminant Analysis, *Ann. Statist.*, pp. 73-102.

Huang, W. (2009). Exponential Integrability of Itô's Processes, *J. Math. Anal. Appl.*, pp. 427-433.

Kloeden, P.E. and Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, New-York.

Kulkarni, S.R. and Posner, S.E. (1995). Rates of Convergence of Nearest Neighbor Estimation under Arbitrary Sampling, *IEEE Trans. Inf. Theory*, pp. 1028-1039.

Lamberton, D. and Lapeyre, B. (1996). *Introduction to Stochastic Calculus Applied to Finance*, Chapman and Hall, CRC Press, London.

Lande, R., Engen, S. and Sæther, B.E. (2003). *Stochastic Populations Dynamics in Ecology and Conservation*, Oxford University Press Inc., New-York.

Lipster, R.S. and Shiryaev, A.N. (2001). *Statistics of Random Processes. I. General Theory*, 2nd Edition, Springer-Verlag, New-York.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis*, Springer-Verlag, New-York.

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

Renshaw, E. (1991). *Modelling Biological Populations in Space and Time*, Cambridge University Press.

Revuz, D. and Yor, M. (1999). *Continuous Martingales and Brownian Motion*, Springer-Verlag, New-York.

van der Vaar, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*, Springer-Verlag, New-York.

van Kampen, N.G. (2007). *Stochastic Processes in Physics and Chemistry, 3rd Edition*, Elsevier, New-York.