A ridge estimator of the drift from discrete repeated observations of the solutions of a stochastic differential equation

Christophe Denis, Charlotte Dion, Miguel Martinez

A ridge estimator of the drift from discrete repeated observations of the solution of a stochastic differential equation.

Christophe Denis[(1,3)], Charlotte Dion[(2)], Miguel Martinez[(1)]
[(1)] LAMA, Université Gustave Eiffel [(2)] LPSM, Sorbonne Université
[(3)] MIA-PARIS, AgroParisTech, INRAE, Université Paris-Saclay

December 11, 2020

Abstract

This work focuses on the nonparametric estimation of a drift function from N discrete repeated independent observations of a diffusion process over a *fixed* time interval $[0, T]$. We study a ridge estimator obtained by the minimization of a constrained least squares contrast. The resulting projection estimator is based on the B-spline basis. Under mild assumptions, this estimator is universally consistent with respect to an integrate norm. We establish that, up to a logarithmic factor and when the estimation is performed on a compact interval, our estimation procedure reaches the best possible rate of convergence. Furthermore, we build an adaptive estimator that achieves this rate. Finally, we illustrate our procedure through an intensive simulation study which highlights the good performance of the proposed estimator in various models.
AMS Subject Classification: 62G05, 62M10, 62J07

1 Introduction

In this paper, we tackle the statistical problem of the estimation of the drift function of a one-dimensional time homogeneous diffusion process $X = (X_t)_{t\in[0,T]}$ where $T > 0$ is a fixed horizon time. The diffusion process X is given as the solution of the following equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x_0 \tag{1.1}$$

where $x_0 \in \mathbb{R}$ is known, and $(W_t)_{t\geq 0}$ denotes a standard Brownian motion. From $(X_t)_{t\in[0,T]}$ the solution of (1.1), we may construct an observation which is the high frequency discretized sample path $\overline{X} := (X_{k\Delta})_{k=0,\dots,n}$ with $T = n\Delta$, $n \in \mathbb{N}$, $0 < \Delta < 1$. We assume that $N \in \mathbb{N}^*$ independent discrete observations $\left(\overline{X}^{(1)}, \dots, \overline{X}^{(N)}\right)$ coming from independent solutions $\left(X^{(1)}, \dots, X^{(N)}\right)$ of (1.1) are available. We refer to the vector of observations $\left(\overline{X}^{(1)}, \dots, \overline{X}^{(N)}\right)$ as the learning sample. Based on this learning sample, the goal is to estimate the drift function $b : \mathbb{R} \to \mathbb{R}$ which can be interpreted as the instantaneous mean of the process. In this paper, we aim at studying a nonparametric ridge estimator of b.

1.1 State of the art

The estimation of the drift function of a diffusion process from a single path is a well known problem. For a review of parametric and nonparametric methods for diffusion processes we refer to Kutoyants (2004). More precisely, one can cite, for the case of continuous ergodic diffusions Yoshida (1992); Gobet (2002), Bibby & Sørensen (1995); Kessler *et al.* (1999) for martingale estimation functions, Gobet *et al.* (2004) in the low frequency context. In the nonparametric context, on can cite Hoffmann (1999); Dalalyan *et al.* (2005); Comte *et al.* (2007); Schmisser (2013). In the Bayesian

literature, the asymptotic properties of minimum contrast estimators are studied for example in Van der Meulen *et al.* (2013); Gugushvili *et al.* (2014); Koskela *et al.* (2019). Recently, Abraham, K. (2019) has studied a minimum contrast estimator over a class of functions on which a constraint on the supremum norm is imposed. Nevertheless, most of these works focus on the framework where the available data consists of only a single observation (continuous or discrete) of a diffusion path in the long run. The consistency of the methods is proved when the horizon time T is not fixed but tends to infinity. Within this context, it is often assumed that the process is ergodic.

On the contrary, in our setting we assume that the horizon time T is fixed and we do not require the ergodicity of the diffusion process. The key ingredient is that the available data consists of N discretized paths observations of the process. Hence, our natural asymptotic framework is given by $N \to +\infty$. Consequently the problem treated in this paper falls within the scope of general functional data analysis (see *e.g.* Ramsay & Silverman, 2007; Wang *et al.*, 2015) which covers a broad class of applications.

However, it seems that very few works investigate the estimation of the drift function from a sample of i.i.d. observations (diffusion paths) when the horizon time T is assumed to be fixed. Up to our knowledge, we only found the references Denis *et al.* (2019); Comte & Genon-Catalot (2019) in the literature. The first work deals with the parametric case and the second one deals with the nonparametric estimation from continuous observations. In Comte & Genon-Catalot (2019), the authors focus on a procedure which relies on a cutoff of a projection estimator where the m dimensional spaces of approximation are generated by Laguerre functions or Hermite functions. One of the main contributions of this work is that it provides theoretical guarantees for the estimation of $b\mathbb{1}_A$, where A is a non compact subset of $\mathbb{R}$.

Yet, in order to ensure the existence and the stability of the resulting estimator $\widehat{b}_m$, the authors insert a cutoff function. This leads the estimator $\widehat{b}_m$ defined as as follows $\widehat{b}_m = \widetilde{b}_m \mathbb{1}_{\{f(m) \leq \log(NT)/NT\}}$ where $\widetilde{b}_m$ is the minimizer of a suitable contrast function and f is an increasing function of the dimension m of the approximation space. We believe that one of the main drawbacks of this estimation procedure is that in regard to the size of the available learning sample, the dimension of the linear subspaces for which the estimator is not the zero function might appear be too small : in the case where the estimation of b necessitates to consider a space of approximation with a large dimension m, the estimator $\widehat{b}_m$ may perform badly.

To avoid this issue, a natural alternative is to seek for regularized methods such as ridge type procedures. These kinds of procedures, that include kernel methods, have been intensively studied in the regression setting (see Hastie *et al.*, 2001, and references therein), but, up to our knowledge, there is no prior study of a regularized procedure for the estimation of the drift function in the context of i.i.d. repeated observations.

1.2 Main contributions

In this paper, we present a new estimator based on the minimization of a least square contrast under an ℓ^2 constraint. Namely, the considered estimator of the drift function relies on projections on some finite dimensional subspaces and we impose an ℓ^2 constraint on the coefficients of the projection that ensures the existence, uniqueness, and stability of the resulting estimator. Notably, the resulting estimator is a ridge type estimator. We focus on the spaces of approximation generated by the popular B-spline basis (De Boor *et al.*, 1978; Györfi *et al.*, 2006) which is often used in practice.

In a first part we show that our procedure is universally consistent for the estimation of b with respect to the time average distribution L^2 risk of the process. Importantly, we emphasize that b is estimated over the whole real line and that this consistency result is obtained without any assumption on the existence of a density transition of the diffusion model. Our results extend those provided in Comte & Genon-Catalot (2019) to the context of discretely observed paths.

A second part of this paper is dedicated to the study of the rate of convergence for the estimation of $b\mathbb{1}_A$, where A is a compact subset of $\mathbb{R}$. To this end, we strengthen our assumptions and assume the uniform ellipticity of the diffusion model in order to ensure the existence of a transition density. We establish that, up to a logarithmic factor, the constructed estimator achieves the optimal rate of convergence over the Hölder balls in the minimax sense. Note that the previous estimator depends on the knowledge of the regularity of the drift b on the set A. As is customary, we deal with

this issue by proposing an adaptive estimator which is shown to reach the minimax rate (up to a logarithmic factor).

Finally, we show numerical experiments that support our study and give illustrations that our estimation performs well on several diffusion models. In various cases our numerical experiments highlight the benefit of considering a regularized estimator rather than using a cutoff procedure.

1.3 Outline of the paper

Section 2 provides the main notations and assumptions that are considered throughout the paper. A presentation of the B-spline basis is also given in this section. The general estimation procedure is described in Section 3 as well as our first consistency result. Section 4 focuses on the study of the rate of convergence when the estimation is restricted to a compact set. We propose an adaptive estimator that reaches the optimal rate of convergence up to a logarithmic factor. Finally the numerical performance of our procedures is investigated in Section 5. Proofs are relegated to the Appendix.

2 General framework

This section is devoted to the presentation of the general framework of our study. We detail the assumptions on the model defined by Equation (1.1) in Section 2.1 while important notation are provided in Section 2.2. The measure of performance from which we evaluate an estimator of the drift function b is introduced in Section 2.3. Finally, in Section 2.4 we present the space of approximation used to build our estimation procedures. This space is chosen to be that of B-spline functions and we recall some of their important properties.

2.1 Assumptions

Our study focuses on the possible solutions of the Stochastic Differential Equation (1.1). Let us introduce the following notations

$$\mathcal{Z}(\sigma) := \{x \in \mathbb{R} \ : \ \sigma(x) = 0\}, \quad \mathcal{Z}(b) = \{x \in \mathbb{R} \ : \ b(x) = 0\}$$

and

$$I(\sigma) := \left\{ x \in \mathbb{R} \ : \ \int_{x-\varepsilon}^{x+\varepsilon} \frac{1}{\sigma^2(y)} dy = +\infty, \quad \forall \varepsilon > 0 \right\}.$$

Throughout the paper, we make the following assumptions.

Assumption 2.1. *(Existence)*

(i) The function b/σ^2 is locally integrable on the complement of $I(\sigma)$ (with the convention that $0 \times \infty = 0$).

(ii) $I(\sigma) \subseteq \mathcal{Z}(\sigma) \cap \mathcal{Z}(b)$.

Assumption 2.2. *(Uniqueness)*

(i) $x \mapsto b(x)$ is globally Lipschitz: there exists $L_b > 0$ such that for all x, y, $|b(x)-b(y)| \leq L_b|x-y|$.

(ii) $x \mapsto \sigma(x)$ is Hölder with exponent $\alpha \in [1/2, 1]$: for all x, y, $|\sigma(x) - \sigma(y)| \leq L_\sigma |x-y|^\alpha$.

Assumption 2.1 ensures that there exists a (weak)-solution of (1.1). Also, it ensures that strong uniqueness holds for the solutions of (1.1), so that appealing to the celebrated result of Yamada and Watanabe, we ensure that there exists a unique strong solution to (1.1) (see *e.g.* Revuz & Yor, 1999).

Assumption (2.2) ensures the linear growth of the coefficients b and σ. Thus there exists a constant $C > 0$ such that

$$\forall x \in \mathbb{R}, \ |b(x)| + |\sigma(x)| \leq C(1 + |x|).$$

Using standard arguments (see for *e.g.* Graham & Talay, 2013), we can show that for any integer $q \geq 1$ there exists a constant $C < +\infty$ depending only on q, x_0, T (not α), such that for any $0 \leq s \leq t \leq T$,

$$\mathbb{E}|X_t - X_s|^{2q} \leq C\,(t-s)^q. \tag{2.1}$$

Note that for now, we do not impose a uniform ellipticity condition and the existence of a transition density function for the Markov process (X_t) is not ensured.

2.2 Notation for continuous and discrete norms

Let $\overline{X} = (X_{k\Delta})_{k=0,\dots,n}$ be a sample path independent of the discrete observations $\left(\overline{X}^{(1)}, \dots, \overline{X}^{(N)}\right)$ with $T = n\Delta$ a fixed time horizon ($n \in \mathbb{N}^*$, $0 < \Delta < 1$). In the following, $N \in \mathbb{N}^*$ independent discrete observations $\left(\overline{X}^{(1)}, \dots, \overline{X}^{(N)}\right)$ of the sample path, coming from $\left(X^{(1)}, \dots, X^{(N)}\right)$, are available. Our asymptotic framework is such that n and N go to infinity.

For a real valued function h defined on $\mathbb{R}$, we denote $\|h\|_{n,b}$ and $\|h\|_b$ the integrated norms defined as:

$$\|h\|_{n,b}^2 := \mathbb{E}_X\left[\frac{1}{n}\sum_{k=0}^{n-1} h^2(X_{k\Delta})\right], \quad \|h\|_b^2 := \mathbb{E}_X\left[\frac{1}{T}\int_0^T h^2(X_s)ds\right]$$

on $L^2\left(\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{P}\circ X_{k\Delta}^{-1}\right)$ and $L^2\left(\frac{1}{T}\int_0^T \mathbb{P}\circ X_s^{-1}\,ds\right)$ respectively, where $\mathbb{E}_X$ is the expectation with respect to the law $\mathbb{P}_X$ of the discrete path $\overline{X}$ defined by (1.1). Its standard L^2-norm is denoted by $\|h\|$. Let us also introduce the following empirical norms

$$\|h\|_n^2 := \frac{1}{n}\sum_{k=0}^{n-1} h^2(X_{k\Delta}), \;\; \|h\|_{N,n}^2 := \frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h^2\left(X_{k\Delta}^{(j)}\right).$$

Finally, $\|.\|_2$ stands for the standard Euclidean norm and $\|.\|_\infty$ for the supremum norm.

2.3 Measure of performance

We introduce the following risks

$$\mathcal{R}_n\left(h,b\right) := \|h-b\|_{n,b}^2, \quad \mathcal{R}\left(h,b\right) := \|h-b\|_b^2. \tag{2.2}$$

The performance of an estimator $\widehat{b}$ of b on a discrete path $\overline{X}$ is naturally assessed through the measure of performance $\mathcal{R}_n$. However, in view of considering the asymptotic $\Delta \to 0$, the major drawback of the risk $\mathcal{R}_n$ is that it depends on the discretization step Δ. Roughly speaking, the convergence of $\mathbb{E}\left[\mathcal{R}_n\left(\widehat{b},b\right)\right]$ to 0 (when n, N goes to infinity) does not ensure, without further investigations, that $\widehat{b}$ is "close" to b. In order to circumvent this issue, we consider more likely the risk measure $\mathcal{R}$ which does not depend on the time step Δ. Nevertheless, the following result shows that the norm $\|.\|_{n,b}$ and $\|.\|_b$ are equivalent up to a remainder term of order $\sqrt{\Delta}$.

Proposition 2.3. *Let h be an L_h-Lipschitz real function. Under Assumption 2.1 and 2.2, the following holds*

$$\left|\|h\|_b^2 - \|h\|_{n,b}^2\right| \leq C(h(0) \vee L_h)L_h\sqrt{\Delta}.$$

If h is bounded

$$\left|\|h\|_b^2 - \|h\|_{n,b}^2\right| \leq C\|h\|_\infty L_h\sqrt{\Delta}$$

where $C > 0$ is a constant which depends on x_0, b.

2.4 Spaces of approximation

Let $K_N \in \mathbb{N}^*$, $A_N, B_N \in \mathbb{R}$, $A_N < B_N$, and $M \in \mathbb{N}^*$. Let us introduce the sequence of knots $\mathbf{u} = (u_{-M}, \ldots, u_{K_N+M})$ such that for $i = 0, \ldots, K_N$

$$u_i = A_N + i\frac{(B_N - A_N)}{K_N},$$

$u_{-M} = \ldots = u_{-1} = u_0 = A_N$, and $u_{K_N} = u_{K_N+1} = \ldots = u_{K_N+M} = B_N$. We consider the B-splines functions $(B_{i,M,\mathbf{u}})_{i=-M,\ldots,K_N-1}$ of degree M associated to the knot vector $\mathbf{u}$. The B-splines functions are defined as follows (see for instance Györfi *et al.*, 2006, and references therein).

Definition 2.4. *the B-spline function of degree ℓ with knots vector $\mathbf{u}$ is recursively defined for all $x \in \mathbb{R}$ by,*

$$B_{i,\ell,\mathbf{u}}(x) = \mathbb{1}_{[u_i, u_{i+1})}(x),$$

for $\ell = 0$, and $i = -M, \ldots, K_N + M - 1$, and

$$B_{i,\ell+1,\mathbf{u}}(x) = \frac{x - u_i}{u_{i+\ell+1} - u_i} B_{i,\ell,\mathbf{u}}(x) + \frac{u_{i+\ell+2} - x}{u_{i+\ell+2} - u_{i+1}} B_{i+1,\ell,\mathbf{u}}(x),$$

for $\ell = 0, \ldots, M-1$, and $i = -M, \ldots, K_N + M - l - 2$. We use the convention $0/0 = 0$.

According to the choice of the knot vector $\mathbf{u}$, the B-spline functions are zero outside $[A_N, B_N]$. Besides, these functions are linearly independent even though their supports are not disjoint. The main advantage of these piecewise polynomial functions is that they satisfy some global smoothness conditions. This kind of attractive property is particularly interesting when we want to build smooth estimates. Finally, the B-spline space $\mathcal{S}_{K_N,M,\mathbf{u}}$ is defined as

$$\mathcal{S}_{K_N,M,\mathbf{u}} = \text{span}\{(B_{i,M,\mathbf{u}}) \ : \ i = -M, \ldots, K_N - 1\}.$$

Hence, the linear space $\mathcal{S}_{K_N,M,\mathbf{u}}$ has dimension $\dim(\mathcal{S}_{K_N,M,\mathbf{u}}) = K_N + M$. Let us recall some useful properties of the B-splines functions (see for instance Györfi *et al.*, 2006, Chapter 14).

Properties 2.5. *For all $i \in \{-M, \ldots, K_N - 1\}$, we have*

(i) For all $x \in \mathbb{R}$, $B_{i,M,\mathbf{u}}(x) \geq 0$.

(ii) For all $x \notin [u_i, u_{i+M+1})$, $B_{i,M,\mathbf{u}}(x) = 0$.

(iii) For all $x \in [A_N, B_N)$,

$$\sum_{i=-M}^{K_N-1} B_{i,M,\mathbf{u}}(x) = 1.$$

(iv) $\int_{A_N}^{B_N} B_{i,M,\mathbf{u}}(x)dx = \frac{u_{i+M+1} - u_i}{M+1} \leq \frac{B_N - A_N}{K_N}$.

From the above properties, we can deduce that if $h \in \mathcal{S}_{K_N,M,\mathbf{u}}$, then h is $M-1$ continuously differentiable on $[A_N, B_N)$ and zero outside of $[A_N, B_N)$. The following Lemma states a useful property which highlights the connections between the norms $\|h\|$ and $\|h\|_\infty$ of an element h of the linear space $\mathcal{S}_{K_N,M,\mathbf{u}}$ and the coefficients of its decomposition in the B-spline basis.

Lemma 2.6. *Let $h = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}} \in \mathcal{S}_{K_N,M,\mathbf{u}}$, then there exists constants $C_1 > 0$, $C_2 > 0$ which depend only on M such that*

$$C_1 K_N^{-1} \|\mathbf{a}\|_2^2 \leq \|h\|^2 \leq C_2 K_N^{-1} \|\mathbf{a}\|_2^2 \ \text{ and } \ \|h\|_\infty \leq \|\mathbf{a}\|_2.$$

3 Constrained estimation based on the B-spline basis

In this section, we describe our estimation procedure which relies on a projection estimator based on the B-spline basis. The estimation procedure is presented in Section 3.1 and the universal consistency of the proposed estimator is established in Section 3.2.

3.1 Estimation strategy

Let us consider the case where $B_N > 0$, $A_N = -B_N$ and $M \geq 1$. Throughout the paper, M is a fixed constant. For $L_N > 0$, we define the constrained subspace

$$\mathcal{S}_{K_N,L_N,M} := \left\{ h = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}} \in \mathcal{S}_{K_N,M,\mathbf{u}} \ : \ \|\mathbf{a}\|_2^2 \leq (K_N+M)L_N \right\}. \tag{3.1}$$

The subspace $\mathcal{S}_{K_N,L_N,M}$ is composed of functions $h = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}}$ for which we ensure uniform boundedness on the coefficients a_i. Then, in view of Lemma 2.6, functions of $\mathcal{S}_{K_N,L_N,M}$ are bounded w.r.t. $\|\cdot\|_\infty$ and $\|\cdot\|$. We consider the estimator $\widehat{b}_{N,n}$ defined as the minimizer of a least square contrast

$$\widehat{b}_{N,n} \in \underset{h \in \mathcal{S}_{K_N,L_N,M}}{\operatorname{argmin}} \ \gamma_{N,n}(h), \tag{3.2}$$

where for $h \in \mathcal{S}_{K_N,L_N,M}$,

$$\gamma_{N,n}(h) := \frac{1}{N\,n} \sum_{j=1}^{N} \sum_{k=0}^{n-1} \left(Z_{k\Delta}^{(j)} - h(X_{k\Delta}^{(j)}) \right)^2, \quad Z_{k\Delta}^{(j)} := \frac{X_{(k+1)\Delta}^{(j)} - X_{k\Delta}^{(j)}}{\Delta}. \tag{3.3}$$

Note that the resulting estimator is then defined as $\widehat{b}_{N,n}(.) = \sum_{i=-M}^{K_N-1} \widehat{a}_i B_{i,M,\mathbf{u}}(.)$ where the vector $\widehat{\mathbf{a}} = {}^t(\hat{a}_{-M}, \ldots, \widehat{a}_{K_N-1}) \in \mathbb{R}^{K_N+M}$ is the ridge estimator (see Hastie *et al.*, 2001):

$$\widehat{\mathbf{a}} = \underset{\|\mathbf{a}\|_2^2 \leq (K_N+M)L_N}{\operatorname{argmin}} \|\mathbf{Z} - \mathbf{B}\mathbf{a}\|_2^2,$$

where the vector $\mathbf{Z} = {}^t(Z_\Delta^{(j)}, \ldots, Z_{n\Delta}^{(j)},\ j = 1, \ldots n)$ belongs to $\mathbb{R}^{Nn}$ and the matrix $\mathbf{B} = (B_{i,M,\mathbf{u}}(\mathbf{X}_j))_{j,i} \in \mathbb{R}^{(Nn)\times(K_N+M)}$, with $\mathbf{X}_j = {}^t(X_\Delta^{(j)}, \ldots, X_{n\Delta}^{(j)})$. Thus, as for the ridge regression procedure, we minimize $\gamma_{N,n}$ given in Equation (3.3) over the constrained subspace $\mathcal{S}_{K_N,L_N,M}$ or equivalently over $\mathbf{a} \in \mathbb{R}^{K_N+M}$, under the constraint: $\|\mathbf{a}\|_2^2 \leq (K_N+M)L_N$. This problem has a unique solution which ensures that the resulting estimator $\widehat{b}_{N,n}$ is always well defined. Moreover, this procedure offers attractive numerical properties. The following result sums up this comment.

Proposition 3.1. *The estimator $\widehat{b}_{N,n}$ is defined as*

$$\widehat{b}_{N,n}(x) = \sum_{i=-M}^{K_N-1} \widehat{a}_i B_{i,M,\mathbf{u}}(x),$$

where $\widehat{\mathbf{a}} = {}^t(\widehat{a}_{-M}, \ldots, \widehat{a}_{K_N-1})$ is defined either by

$$\widehat{\mathbf{a}} = \left({}^t\mathbf{B}\mathbf{B}\right)^{-1} {}^t\mathbf{B}\mathbf{Z}$$

if the matrix $({}^t\mathbf{B}\mathbf{B})$ is invertible and $\|\mathbf{a}\|_2^2 \leq (K_N+M)L_N$, or by

$$\widehat{\mathbf{a}}_{\widehat{\lambda}} = \left({}^t\mathbf{B}\mathbf{B} + \widehat{\lambda}\mathbf{I}_{K_N+M}\right)^{-1} {}^t\mathbf{B}\mathbf{Z},$$

where $\widehat{\lambda}$ is the unique solution of $\|\widehat{\mathbf{a}}_\lambda\|_2^2 = (K_N+M)L_N$, where

$$\widehat{\mathbf{a}}_\lambda = ({}^t\mathbf{B}\mathbf{B} + \lambda\mathbf{I}_{K_N+M})^{-1} {}^t\mathbf{B}\mathbf{Z}.$$

In the recent work of Comte & Genon-Catalot (2019), the authors focus on a least squares contrast estimator (based on continuous observations). In order to ensure the stability of the estimator, the authors propose to insert a cutoff function. More precisely, the estimator is set to the zero function according to some threshold which depends on the dimension of the considered space of approximation. This procedure may reduce the dimension of the spaces of approximation on which the resulting estimator is non trivial and can lead to some limitations in practice (see Section 5.2.3 for more details). The estimator proposed in Proposition 3.1 may be viewed as an alternative.

3.2 Consistency of the procedure

In this section, we show that the proposed estimator $\widehat{b}_{N,n}$ is universally consistent with respect to the risk $\mathcal{R}$ given in Equation (2.2). The consistency of the estimator relies on the following result.

Proposition 3.2. *Under Assumptions 2.1 and 2.2, the estimator $\widehat{b}_{N,n}$ of b given in Equation* (3.2)*, satisfies*

$$\mathbb{E}\left[\|\widehat{b}_{N,n} - b\|_{N,n}^2\right] \leq C_1 \inf_{h \in \mathcal{S}_{K_N, L_N, M}} \mathcal{R}_n(h, b) + C_2 \left(\sqrt{\frac{(K_N + M)L_N}{N}} + \Delta \right), \tag{3.4}$$

where $C_1 > 1$ a numerical constant and $C_2 > 0$ is a constant depending on σ and T.

This Proposition gives a bound for the error of estimator $\widehat{b}_{N,n}$ which is expressed in terms of the empirical norm $\|.\|_{N,n}$. The first term in the r.h.s of Equation (3.4) is interpreted as the bias term while the second one is a bound of the variance term which is of order $((K_N + M)L_N/N)^{1/2}$.

Note that this rate is slower than could be expected due to the relative weak assumptions on the model and that there is no compactness assumption on the estimation interval for b. Indeed, we improve this rate in Section 4 by assuming ellipticity and mild regularity on σ and by estimating b over a compact interval. The last term in the r.h.s is of order Δ and highlights the error due to the discretization. Nevertheless, the result of Proposition 3.2 gives only an error bound on the observations $\overline{X}^1, \ldots, \overline{X}^N$. Hence, it is not sufficient to derive the consistency of $\widehat{b}_{N,n}$ w.r.t the empirical norm. The aim of the next result is to derive a consistency result w.r.t. to the risk $\mathcal{R}$. It relies on Proposition 3.2 and concentration arguments.

Theorem 3.3. *Under Assumptions 2.1 and 2.2, with $A_N = -B_N$, assume that $L_N \to +\infty$ and $\Delta = O\left(\frac{B_N^2}{K_N^2 N^2}\right)$. Furthermore, if $K_N, B_N \to +\infty$ such that*

$$\frac{(K_N + M)^2 L_N \log(N)}{N} \to 0, \quad \text{and} \quad \frac{B_N}{K_N} \to 0, \quad B_N > L_N,$$

the following holds

$$\mathbb{E}\left[\mathcal{R}(\widehat{b}_{N,n}, b)\right] \underset{N,n \to \infty}{\to} 0.$$

This result provides the consistency, under mild conditions, of our estimation procedure provided that the time step Δ is small enough w.r.t. N and $\frac{K_N}{B_N}$. Note that, the conditions required in Theorem 3.3 are not too difficult to fulfill. Indeed, Theorem 3.3 can be applied with $L_N = \dfrac{B_N}{2} = \log(N)$, and $K_N = N^{1/4}$. However, this result does not provide rates of convergence and this is the goal of the next section.

4 Optimal rates of convergence

This section is dedicated to the study of the rate of convergence of our method. More specifically, the aim of this section is to establish that our procedure is optimal in the minimax sense. To this end, we investigate the case where the estimation is performed a over compact interval. In Section 4.1, we introduce some additional assumptions on the model. The upper bound result is provided in Section 4.2 while the lower bound is derived in Section 4.3. Finally, in Section 4.4, we build an adaptive estimator which achieves the minimax rate (up to a logarithmic factor).

4.1 Assumptions

We consider the case where b is estimated on a compact set that for simplicity is assumed to be $[0, 1]$, thus we fix $A_N = 0$, and $B_N = 1$. We also assume that $x_0 \in (0, 1)$. Besides, the rate of convergence of our estimation procedure is studied over the class of Hölder functions. We make the following assumptions.

Assumption 4.1. *The diffusion coefficient σ belongs to $\mathcal{C}_b^2(\mathbb{R})$ and there exists some constants, $0 < \sigma_0 \leq \sigma_1$ such that*

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

Assumption 4.2. *For $\beta \in [1, M+1]$, and $R > 0$, the restriction $\tilde{b} := b_{|[0,1]}$ of b to $[0,1]$ belongs to the Hölder ball $\Sigma(\beta, R)$: the function $\tilde{b}$ is $l = \lfloor \beta \rfloor$ times differentiable on $(0,1)$ and its derivative $\tilde{b}^{(l)}$ satisfies*

$$\forall x, y \in (0,1), \quad \left| \tilde{b}^{(l)}(x) - \tilde{b}^{(l)}(y) \right| \leq R \left| x - y \right|^{\beta - l}.$$

Due to Assumption 4.1 the Markov process (X_t) admits a transition density $(t, y) \mapsto p(t, x_0, y)$ (see Fournier *et al.*, 2010). Following Gobet (2002) Proposition 1.2, from uniform ellipticity, we get that the transition densities of the diffusion process are bounded on a compact interval. In particular, one can derive the following result which connects the norm $\|.\|_{n,b}$ and the L^2-norm $\|.\|$ and then the risk $\mathcal{R}$ to the usual L^2 risk.

Lemma 4.3. *Under assumptions 2.1, 2.2 and 4.1, there exists $\pi_1 > \pi_0 > 0$, such that for all $y \in [0,1]$, we have*

(i) $\pi_0 \leq \frac{1}{n} \sum_{k=1}^{n-1} p(k\Delta, x_0, y) \leq \pi_1$, *(for any $n \geq 4$)*

(ii) $\pi_0 \leq \frac{1}{T} \int_0^T p(s, x_0, y) ds \leq \pi_1$.

In particular for a function h such that $\mathrm{supp}(h) \subseteq [0,1]$, *we have*

$$\|h\|^2 \leq \frac{1}{\pi_0} \|h\|_{n,b}^2.$$

4.2 Upper bound

One of the main ingredients to derive rates of convergence is the spline approximation result (see Chapter 14 in Györfi *et al.* (2006)) which allows to control the bias term in Equation (3.4). In order to derive optimal rate of convergence, we consider a slightly modified version of the estimator defined in Equation (3.2). The truncated estimator is defined as follow

$$\widehat{b}_{N,n}^{L_N}(x) := \begin{cases} \widehat{b}_{N,n}(x) & \text{if } |\widehat{b}_{N,n}(x)| \leq \sqrt{L_N}, \\ \mathrm{sgn}(\widehat{b}_{N,n}(x)) \sqrt{L_N} & \text{if } |\widehat{b}_{N,n}(x)| > \sqrt{L_N}. \end{cases} \tag{4.1}$$

We remind the reader that L_N is the multiplicative factor that controls the bound on the Euclidean norms of the parameter coefficients $\mathbf{a}$ for all functions belonging to $\mathcal{S}_{K_N, L_N, M}$ (see the definition of $\mathcal{S}_{K_N, L_N, M}$ (3.1)). First, for N large enough, since $\widetilde{b}$ is bounded, $\|\widetilde{b}\|_\infty \leq \sqrt{L_N}$ which implies $\|\widetilde{b} - \widehat{b}_{N,n}^{L_N}\|_b \leq \|\widetilde{b} - \widehat{b}_{N,n}\|_b$. Therefore, the consistency of $\widehat{b}_{N,n}$ implies the consistency of $\widehat{b}_{N,n}^{L_N}$. Moreover, let us notice that $\|\widehat{b}_{N,n}\|_\infty < \sqrt{(K_N + M) L_N}$ while the truncated estimator $\widehat{b}_{N,n}^{L_N}$ satisfies $\|\widehat{b}_{N,n}^{L_N}\|_\infty < \sqrt{L_N}$. This property is particularly important in Theorem 3.3 to reduce the order of the variance term with respect to K_N. Before announcing the main result of this section, we first establish a similar result to Proposition 3.2.

Proposition 4.4. *Grant Assumptions 2.1, 2.2, 4.1, and 4.2. The estimator $\widehat{b}_{N,n}$ of $\widetilde{b}$ given in Equation* (3.2)*, satisfies for N large enough,*

$$\mathbb{E}\left[\left\| \widehat{b}_{N,n}^{L_N} - \widetilde{b} \right\|_{N,n}^2 \right] \leq C \left(\left(\frac{M+1}{K_N} \right)^{2\beta} + \frac{K_N + L_N}{N} + \Delta \right),$$

where $C > 0$ is a constant depending only on σ_1, T, M, and R.

As for Proposition 3.2, the obtained bound is composed of three terms. The first one, which relies on the spline approximation properties, gives the order of the square bias under the assumption that the function $\tilde{b}$ is Hölder. The last two terms are similar to the ones obtained in Proposition 3.2. However, note that the variance term is of order $(K_N + L_N)/N$ which is faster than the one obtained

in Proposition 3.2, which is of order $((K_N+L_N)/N)^{1/2}$. Indeed in Proposition 4.4, we assume both the ellipticity of the diffusion coefficient σ and that the drift function b is estimated over the compact interval $[0,1]$. The combination of these assumptions implies that the empirical norm and the L_2-norm are equivalent over $[0,1]$ (see Lemma 4.3), which is the key point to get a variance rate of order of $(K_N+M)L_N/N$. Finally, the truncated version of the estimator given in Equation (4.1) allows to improve the variance rate to $(L_N+K_N)/N$ in the bound of Proposition 4.4.

Combining Proposition 4.4 with concentration arguments, we obtain the following result. Let us introduce $\mathcal{K}_N = \{1,\dots,K_N^*\}$ with $K_N^* = \sqrt{N/\log^2(N)}$.

Theorem 4.5. *Grant Assumptions 2.1, 2.2, 4.1, and 4.2. Let $K_N \in \mathcal{K}_N$. Assume that $L_N = \log(N)$ and $\Delta = O(K_N^{-1}N^{-2})$, then for N large enough the following holds*

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{b}_{N,n}^{L_N},\widetilde{b}\right)\right] \leq C\left(\left(\frac{M+1}{K_N}\right)^{2\beta} + \frac{\log^2(N)K_N}{N}\right), \tag{4.2}$$

where $C>0$ is a constant depending only on σ_1, T, M and R.

From this result, one can see that the rate of convergence is, up to a logarithmic factor, the optimal nonparametric rate in the regression setting (Tsybakov, 2009). Indeed, since $\beta \geq 1$, for $K_N = \left\lfloor \left(N/\log^2(N)\right)^{1/(2\beta+1)} \right\rfloor \in \mathcal{K}_N$, we obtain

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{b}_{N,n}^{L_N},\widetilde{b}\right)\right] \leq O\left(\left(\frac{\log^2(N)}{N}\right)^{\frac{2\beta}{2\beta+1}}\right).$$

Furthermore, as a consequence of the above inequality and using Lemma 4.3, we also deduce that

$$\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|^2\right] \leq O\left(\left(\frac{\log^2(N)}{N}\right)^{\frac{2\beta}{2\beta+1}}\right).$$

This inequality shows that, regarding to the L^2 risk, the problem of estimating the drift function on a compact set based on repeated observations is equivalent to the estimation of a function in the regression setting (provided that the time step Δ is small enough). Let us comment the logarithm factors. The first $\log(N)$ is due to the fact that there is no prior knowledge on the bound of $\|\widetilde{b}\|_\infty$. The second one is due to the control of the supremum of an empirical process over a subset of $\mathcal{S}_{K_N,L_N,M}$.

Finally let us conclude this paragraph noticing that, in the case where b is bounded, it is possible to derive the same result under a weaker condition on Δ. It relies on the following result which is similar to Proposition 2.3.

Proposition 4.6. *Let h a measurable function such that $\|h\|_\infty < +\infty$. Under Assumptions 2.1, 2.2 and 4.1, the following holds*

$$\left|\|h\|_b^2 - \|h\|_{n,b}^2\right| \leq C\|h\|_\infty^2 \Delta \log\left(\frac{1}{\Delta}\right),$$

where $C \geq 1$ is a constant depending on $\|b\|_\infty$.

This upper bound for the difference of the norms is slightly better than the one obtain in Proposition 2.3. This is due to the compactness assumption and to the ellipticity assumption. In Theorem 4.5, this allows to alleviate the assumptions on Δ: the result holds for $\Delta = O(1/N)$.

4.3 Lower bound

As suggested by Equation (4.2), we then establish a lower bound on the risk $\mathcal{R}$ for the Hölder class of functions $\Sigma(\beta,R)$ with regularity parameter β, defined in Assumption 4.2. This lower bound shows that our proposed estimator is optimal, up to a logarithmic factor, in the minimax sense. More precisely, we obtain the following result

Theorem 4.7. *Grant Assumptions 2.1, 2.2, 4.1, and 4.2. There exists two constants $c_1, c_0 > 0$ such that for N large enough and $\widehat{b}$ constructed from $(X^1, \dots, X^N)$,*

$$\sup_{b\,:\,\widetilde{b}\in\Sigma(\beta,R)} \mathbb{E}\left[\left\|\widehat{b}-\widetilde{b}\right\|^2\right] \geq c_1 N^{-2\beta/(2\beta+1)},$$

$$\sup_{b\,:\,\widetilde{b}\in\Sigma(\beta,R)} \mathbb{E}\left[\mathcal{R}\left(\widehat{b},\widetilde{b}\right)\right] \geq c_0 N^{-2\beta/(2\beta+1)}.$$

The proof of the Theorem follows the same lines of Theorem 2.8 in Tsybakov (2009) except for the control of the Kullback-Leibler divergence. In Theorem 4.7, this control relies on the Girsanov formula. The result is then obtained for the risk $\mathcal{R}$ using Lemma 4.3.

Combining Theorem 4.7 and Inequality (4.2), we have shown, that for N large enough, there exists $C_1, c_1 > 0$ such that

$$c_1 N^{-2\beta/(2\beta+1)} \leq \inf_{\widehat{b}} \sup_{b\,:\,\widetilde{b}\in\Sigma(\beta,R)} \mathbb{E}\left[\left\|\widehat{b}-\widetilde{b}\right\|^2\right] \leq C_1 \left(\frac{\log^2(N)}{N}\right)^{\frac{2\beta}{2\beta+1}},$$

where the infimum is taken over all possible estimators $\widehat{b}$ and the supremum is taken over the set off all possible drift functions b such that $\widetilde{b} \in \Sigma(\beta, R)$ and for which Equation (1.1) satisfies Assumptions 2.1, 2.2, and 4.1. Hence, this inequality shows that the optimal rate is of order $N^{-2\beta/(2\beta+1)}$ (up to a logarithmic factor). From Inequality (4.2), we see that this rate is reached by the estimator $\widehat{b}^{L_N}_{N,n}$ for $K_N = \left\lfloor \left(N/\log^2(N)\right)^{1/(2\beta+1)} \right\rfloor$. However, this particular choice of K_N depends on the regularity β of the function $\widetilde{b}$ which is unknown in practice. To avoid this issue, it is usual to build an adaptive estimator to the regularity β.

4.4 Adaptive estimator

To alleviate the notations, the parametes K_N is denoted by K in the following last section. Besides, in order to highlight the dependency on K, the estimator $\widehat{b}^{L_N}_{N,n}$ is denoted $\widehat{b}_K$ (and we choose $L_N = \log(N)$). Our adaptive procedure relies on the dyadic B-splines. That is to say, we assume that K belongs to $\mathcal{K} = \{2^p,\ p = 0, \dots, p_{\max}\}$ with $p_{\max} \leq \sqrt{N/\log^2(N)}$. Hence, this particular choice ensures that the spaces $S_{K,M,\mathbf{u}}$ are nested (for $K < K'$, $S_{K,M,\mathbf{u}} \subset S_{K',M,\mathbf{u}}$) which is an important property in light of the proof of Theorem 4.8. We define the following estimator

$$\widehat{K} = \underset{K\in\mathcal{K}}{\operatorname{argmin}} \left\{\gamma_{N,n}(\widehat{b}_K) + \operatorname{pen}(K)\right\}, \tag{4.3}$$

and then consider the estimator $\widehat{b}_{\widehat{K}}$ defined as the minimizer of a penalized contrast. To penalize the complexity of $S_{K,L,M}$, we choose a penalty term $\operatorname{pen}(K) \geq 44\frac{\log^2(N)(K+M)}{N}$ for N large enough. Now, we state the following result

Theorem 4.8. *Under Assumptions 2.1, 2.2 and 4.1 and assume that $L_N = \log(N)$ and $\Delta = O(1/N^2)$, then the estimator $\widehat{b}_{\widehat{K}}$ of $\widetilde{b}$ satisfies*

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{b}_{\widehat{K}},\widetilde{b}\right)\right] \leq 2 \inf_{K\in\mathcal{K}} \left\{\inf_{h\in S_{K,L,M}} \mathcal{R}\left(h,\widetilde{b}\right) + \operatorname{pen}(K)\right\} + \frac{C}{N}, \tag{4.4}$$

where C is a positive constant depending on σ_1, T, M and R.

This result shows that the estimator $\widehat{b}_{\widehat{K}}$ achieves the bias-variance compromise over the model collection $(S_{K,L,M})_{K\in\mathcal{K}}$. In particular, whenever $\widetilde{b} \in \Sigma(\beta, R)$, with $\beta \leq M + 1$, the estimator $\widehat{b}_{\widehat{K}}$ reaches the optimal rate up to a logarithmic factor. Note that the penalty term can be chosen equal to $44\frac{\log^2(N)(K+M)}{N}$. Nevertheless, in practice it is better to consider $\operatorname{pen}(K) = c\frac{\log(N)^2(K+M)}{N}$ where the constant c is calibrated through intensive numerical experiments (see Section 5.1).

5 Numerical experiments

In this section, we investigate the performance of the adaptive estimator presented in Section 4.4. Section 5.1 is dedicated to the calibration of the penalty term whereas in Section 5.2, we assess the quality of the procedure in the case where the drift function is estimated either on a fixed compact interval, or on an interval which relies on the random data range.

Let us briefly talk about the simulation of the diffusion paths. We use the package `sde` presented in Iacus (2008) to implement an effective numerical method for the considered stochastic differential equations. Different discretization scheme can be used. We choose for example the exact simulation ("EA") for the Ornstein-Uhlenbeck process, "Euler" or "Milstein" for the others.

Throughout this section, the time step Δ between two subsequent observations is fixed but small, and n is large in accordance with our asymptotic context. We choose: $n\Delta = T = 1$ ($\Delta = 1/n$) $n \in \{100, 500\}$. The sample size N is chosen in $\{100, 1000\}$. Our estimators are based on the cubic ($M = 3$) B-spline basis. For the implementation of the B-spline, we use the package `fda`. We restrict our investigation to $\mathcal{K} = \{2^p, p = 0, 1, 2, 3, 4, 5\}$ (thus $\dim(\mathcal{S}_{K,M,\mathbf{u}}) = 2^p + 3,\ p = 0, 1, 2, 3, 4, 5$). Finally, according to our theoretical results, the constant coefficient L_N is chosen equal to $\log(N)$.

5.1 Calibration of the penalty

According to Theorem 4.8 the penalty function can be chosen, for $K \in \mathcal{K}$, as $\mathrm{pen}_c(K) = c\frac{\log(N)^2(K+M)}{N}$ with $c \geq 44$. Nevertheless, it is well known in practice that this constant is too large and has to be chosen through an intensive numerical study. To this purpose, we perform some preliminary simulations in the case where b is estimated over $[-1, 1]$.

We have investigated different models for the calibration step. More precisely, we fix $\sigma = 1$, $x_0 = 0$ and consider the three following drift functions

- model 1 $b(x) = -\frac{2x}{\sqrt{1+x^2}}$,
- model 2 $b(x) = \frac{3}{\sqrt{0.8\pi}}(\exp(-(4x-2)^2/0.8) + \exp(-(4x+2)^2/0.8))$,
- model 3 $b(x) = 0.6(\exp(-x^2) + \cos(10x) + \sin(5x))$.

Note that the three models satisfy the assumptions of Section 4. The first one is the simplest. On the contrary, the two other models are designed by multimodal functions where larger dimensions are required for the estimation.

We have considered these examples to ensure that the chosen constant for the dimension selection is optimal.

Now, we consider a grid $\mathcal{C} = \{0.01, 0.025, 0.05, 0.1, 0.5, 1, 2\}$ of possible values for the constant c. For each model and each $c \in \mathcal{C}$, we repeat 1000 times the following steps:

(i) simulate two independent datasets $\mathcal{D}_N$ and $\mathcal{D}_{N'}$ with $N' = 1000$;

(ii) based on $\mathcal{D}_N$, compute the estimator $\widehat{b}_{\widehat{K}}$ defined by Equation (4.3) with $\mathrm{pen}(K) = \mathrm{pen}_c(K)$;

(iii) based on $\mathcal{D}_{N'}$, evaluate the empirical risk $\|\widehat{b}_{\widehat{K}} - \tilde{b}\|^2_{n,b}$.

Finally, we compute the average $\mathrm{Err}\left(\widehat{b}_{\widehat{K}}(c)\right)$ of the empirical risk by using the Monte-Carlo method over the 1000 repetitions. For each of the three models, the functions $c \mapsto \mathrm{Err}\left(\widehat{b}_{\widehat{K}}(c)\right)$ are displayed in Figure 1. In most cases the constant $c_0 = 0.1$ is the best choice and we fix c with this value in the following.

5.2 Estimation results

We consider the four following models to illustrate the accuracy of the estimator. To prevent from over-fitting, they differ from models $1, 2, 3$ presented in the above section.

- model 4 Ornstein-Uhlenbeck $b(x) = 1 - x$, $\sigma(x) = 1$
- model 5 Cox-Ingersoll-Ross $b(x) = 1 - x$, $\sigma(x) = \sqrt{x}$

- model 6 $b(x) = (1-x^2)(-2\text{atanh}(x) - x)$, $\sigma(x) = 1 - x^2$
- model 7 $b(x) = 0.1(-\sin(2\pi x) + \cos(2\pi x) + 16\sin(3\pi x) - 5\cos(3\pi x))$, $\sigma(x) = 1$

The two first models are widely used diffusion models. Note that model 5 and model 6 possess a non constant diffusion coefficient and do not satisfy the ellipticity assumption 4.1. The model 7 has a multimodal drift function. It requires to explore more possible values of K (larger dimension).

This collection of models has been chosen to evaluate the performance of our procedure, but also to show its robustness to the assumptions. The benchmark of this study is the *oracle*-type estimator defined as the estimator of the collection which minimizes the risk. Note that this *estimator* is only available when the drift function is perfectly known.

5.2.1 Estimation on a fixed interval

In this section, we focus on the estimation of $\widetilde{b} = b\mathbb{1}_{[-1,1]}$. The theoretical guarantees of Section 4 are then in force. Note that for model 6 we have $\widetilde{b} = b$. As an illustration, Figure 2 displays ten realizations of the estimators $\widehat{b}_{\widehat{K}}$ on models 4,5,7. We can see that these estimates perform quite well. Then, we perform 1000 Monte-Carlo simulations of the following steps:

(i) simulate two independent dataset $\mathcal{D}_N$ and $\mathcal{D}_{N'}$ with $N' = 1000$;

(ii) based on $\mathcal{D}_N$, compute the estimator $\widehat{b}_K$ with $\mathcal{K} = \{2^p, p = 0, 1, 2, 3, 4, 5\}$, and $\widehat{K}$;

(iii) based on $\mathcal{D}_{N'}$, evaluate the empirical risks $\|\widehat{b}_{\widehat{K}} - \tilde{b}\|^2_{n,b}$ and $\|\widehat{b}_{K^*} - \tilde{b}\|^2_{n,b}$, with $K^* = \underset{K\in\mathcal{K}}{\text{argmin}} \|\widehat{b}_K - \tilde{b}\|^2_{n,b}$. In the following $\widehat{b}_{K^*}$ is referred as the *oracle* estimator.

Finally, we compute the mean and standard deviation of the empirical risk over the 1000 Monte-Carlo repetitions. The results are presented in Table 1. Let us make a few comments about these results. First, in terms of risk, the estimation procedure exhibits good performances. In particular, one can note that our estimator performs as well as the oracle estimator. Second, regarding the chosen dimension, for models 4,5,6 the value $\widehat{K} = 1$ is mostly chosen while $\widehat{K} = 8$ is mostly selected for model 7. This is not surprising since the drift functions of model 4,5,6 are quite simple whereas the multimodal aspect of the drift function of model 7 requires to select larger $\widehat{K}$. Hence, for model 7 the estimation of $\tilde{b}$ is more challenging. Finally, the influence of the parameter N is clearly illustrated : when N increases from 100 to 1000 the estimated values of risk are divided by 10. On the contrary, the performance of the procedure is not affected by the value of n for the chosen sample sizes. For this reason, in the sequel we set $n = 100$.

5.2.2 Estimation based on the random data range

Now, we investigate the estimation of b without restriction on the estimation interval. In this case, it is natural to build our procedure on the random interval defined as $\left[\min\left((\bar{X}^1, \ldots, \bar{X}^N\right), \max\left((\bar{X}^1, \ldots, \bar{X}^N\right)\right]$. We evaluate the performance of our procedure according to the procedure described in Section 5.2.1. The results are provided in Table 2. Similar comments to those given in Section 5.2.1 apply here. Nevertheless, let us notice that the estimated risks are a bit larger than the ones given in Table 1. This seems reasonable and in line with the theoretical results. Furthermore, due to the estimation on a larger interval, for model 7 the value $\widehat{K} = 32$ is the one that is mainly chosen. This last point shows that the dimension of the space of approximation should be large enough to ensure a good performance of the estimator.

5.2.3 Discussion and comparison

In this section we discuss two more points concerning the numerical study of model 7, the model for which we believe that the estimation of b is the most difficult. We focus on the case where the estimation procedure is based on the random data range.

N, n	$N = 100$ $n = 100$		$N = 100$ $n = 500$		$N = 1000$ $n = 100$		$N = 1000$ $n = 500$	
Estimator	$\widehat{b}_{\widehat{K}}$	$\widehat{b}_{K*}$	$\widehat{b}_{\widehat{K}}$	$\widehat{b}_{K*}$	$\widehat{b}_{\widehat{K}}$	$\widehat{b}_{K*}$	$\widehat{b}_{\widehat{K}}$	$\widehat{b}_{K*}$
Model 4	0.04 (0.04)	0.03 (0.03)	0.04 (0.03)	0.03 (0.02)	.004 (.003)	.004 (.003)	.004 (.003)	.004 (.003)
Model 5	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	.002 (.001)	.002 (.001)	.002 (.001)	.002 (.001)
Model 6	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)	0.02 (0.02)	.002 (.002)	.002 (.002)	.002 (.002)	.002 (.002)
Model 7	0.15 (0.05)	0.12 (0.04)	0.15 (0.05)	0.12 (0.04)	0.03 (.005)	0.02 (.006)	.026 (.005)	.020 (.006)

Table 1: Estimation on $[-1, 1]$. Average and standard deviation of the estimated risks $\|\widehat{b}_{\widehat{K}} - \tilde{b}\|^2_{n,b}$ and $\|\widehat{b}_{K^*} - \tilde{b}\|^2_{n,b}$ computed over 1000 repetitions.

N, n	$N = 100$ $n = 100$		$N = 1000$ $n = 100$	
Estimator	$\widehat{b}_{\widehat{K}}$	$\widehat{b}_{K^*}$	$\widehat{b}_{\widehat{K}}$	$\widehat{b}_{K^*}$
Model 4	0.04(0.03)	0.04 (0.02)	.005 (.003)	.005 (.003)
Model 5	0.03 (0.02)	0.02 (0.02)	.004 (.003)	.003 (.003)
Model 6	0.02 (0.02)	0.02 (0.02)	.002 (.002)	.002 (.002)
Model 7	0.22 (0.07)	0.21 (0.06)	.031 (.008)	.031 (.008)

Table 2: Estimation on the random data range. Average and standard deviation of the estimated risks $\|\widehat{b}_{\widehat{K}} - \tilde{b}\|^2_{n,b}$ and $\|\widehat{b}_{K^*} - \tilde{b}\|^2_{n,b}$ computed over 1000 repetitions.

About the choice of the parameter L_N. This tuning parameter is defined in Section 3 and is used to calibrate the ℓ^2-constraint in our procedure. The aim of this discussion is to highlight the influence of this parameter on the quality of the estimation. In the previous results, as suggested by our theoretical study (see Theorem 4.5), we have chosen $L_N = \log(N)$. Since the bound on the empirical error of the estimator (see Proposition 3.2) increases w.r.t. L_N, it can not be chosen too large. On the contrary, in view of the proof of Theorem 4.5, $\sqrt{L_N}$ should be larger than $\|b\|_\infty$. In our numerical study, this condition is satisfied for $L_N = \log(N)$. Nevertheless, if $\|b\|_\infty > \sqrt{\log(N)}$, we have to consider another choice for the value of parameter L_N.

Hereafter, we explore the cases where $L_N \in \{30, 300\}$ (thus $L_N > \log(N)$, for $N \in \{100, 1000\}$). According to the scheme described in Section 5.2.1, for $N = 100, n = 100$, we obtain an estimated risk of 0.35 (0.12) for $L_N = 30$, and 0.50 (0.23) for $L_N = 300$. For $N = 1000$, we obtain 0.04 (0.02) for $L_N = 30$ or 300. Hence, for moderate values of N ($N = 100$), choosing a too large value of L_N deteriorates the performance of the procedure. On the contrary, for large values of N ($N = 1000$), the performance of the procedure seems stable w.r.t the choice of L_N.

Comparison with the estimator proposed in Comte & Genon-Catalot (2019). We first recall the definition of the estimator studied in Comte & Genon-Catalot (2019). Let $(S_K)_{K=1,\dots,K_0}$ a family of linear subspace . For $K \in \{1, \dots, K_0\}$, the authors consider an estimator, denoted by $\widehat{b}_K$, defined as a truncated version of the minimizer of the continuous contrast

$$\gamma_N(h) = \frac{1}{NT} \sum_{i=1}^{N} \left(\int_0^T h^2(X_u^{(i)}) du - 2 \int_0^T h(X_u^{(i)}) dX_u^{(i)} \right), \tag{5.1}$$

over $h \in S_K$. In order to evaluate this procedure on the B-spline basis, we choose $S_K = S_{K,M,u}$ for $K \in \mathcal{K}$. Note that for practical implementation we have to consider the discretized version of the contrast γ_N given in Equation (5.1) which yields the contrast defined in Equation (3.3). In order to ensure the stability of the estimation procedure, we implement the cutoff proposed by the authors in the Simulation section. According to the scheme described in Section 5.2.1, we compute once again the estimated risk of the oracle estimator $\widehat{b}_{K^*}$ over the collection $(\widehat{b}_K)_{K \in \mathcal{K}}$. For $N = 100$, $n = 100$, we obtain 0.52 (0.32) and 0.12 (0.27) for $N = 1000, n = 100$. These results are not as good as the one obtained by our procedure in Table 1. The main reason is that the truncation (called cutoff in the paper) proposed by Comte & Genon-Catalot (2019) has the effect of reducing the dimension of the selected models. Our simulations give an illustration of the computational limitations of the cutoff procedures.

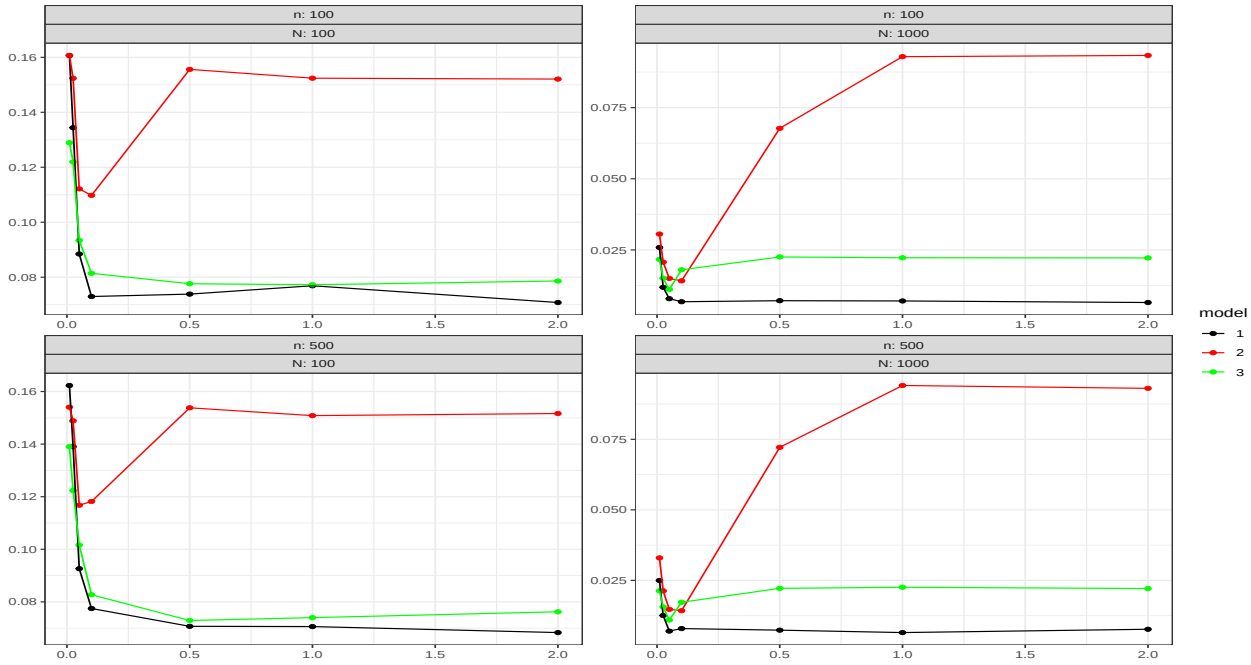

Figure 1: For models $1, 2, 3$ estimates of the empirical risk $\|\widehat{b}_{\widehat{K}} - \tilde{b}\|^2_{n,b}$ for the chosen estimator $\widehat{b}_{\widehat{K}}(c) = \widehat{b}_{\widehat{K}}$ as a function of $c \in \mathcal{C} = \{0.01, 0.025, 0.05, 0.1, 0.5, 1, 2\}$

6 Conclusion

In this paper we study a new nonparametric procedure for the drift function of general homogeneous stochastic differential equations in the framework where the data consist of N i.i.d. discrete observations of the sample path of the solution on a fixed time interval. The estimator is the minimizer of a least square contrast subject to a ridge constraint over the linear subspace spanned by the B-spline basis. We establish the consistency of our procedure. Furthermore, under mild assumptions and when the estimation is performed over a compact interval, we build an adaptive estimator which achieves, up to a logarithmic factor, the minimax rate of convergence.

As a possible guideline for further research, it is interesting to question the link between the framework investigated in this paper and the usual setting considered for the estimation of the drift b over $[0, T]$ (*e.g.* under assumptions ensuring ergodic properties), where only a single trajectory is observed but $T \to +\infty$.

Starting from our i.i.d. framework with fixed time horizon, a natural idea would be to try to build out of the N i.i.d. sample an original single sample path coming from a different S.D.E. with characteristics σ and some new drift function $\tilde{b}$. The drift $\tilde{b}$ would have to coincide with b on our estimation set (say $[0, 1]$) ensuring that the resulting diffusion satisfies nice ergodic properties. Our objective would be then fulfilled by using classical results for the estimation of $\tilde{b}$ in the setting of drift estimation for ergodic diffusions. However, although intuitive and elegant, this idea seems to raise some rather strong technical difficulties. One of the main issues is to paste the several pieces of i.i.d trajectories into a single one to obtain a sample path of the ergodic S.D.E with drift $\tilde{b}$. Furthermore, since we would paste only the observations which belong to $[0, 1]$, we may need to enlarge the probability space to determine how the ergodic diffusion might be driven outside the estimation set and links the different pieces.

On the other hand, let us assume that we start from only one observation consisting of a single trajectory of some (ergodic) diffusion process over a time interval $[0, T]$ with a time horizon T that is now allowed to tend to infinity. In this case, we may be tempted to construct N i.i.d. realizations

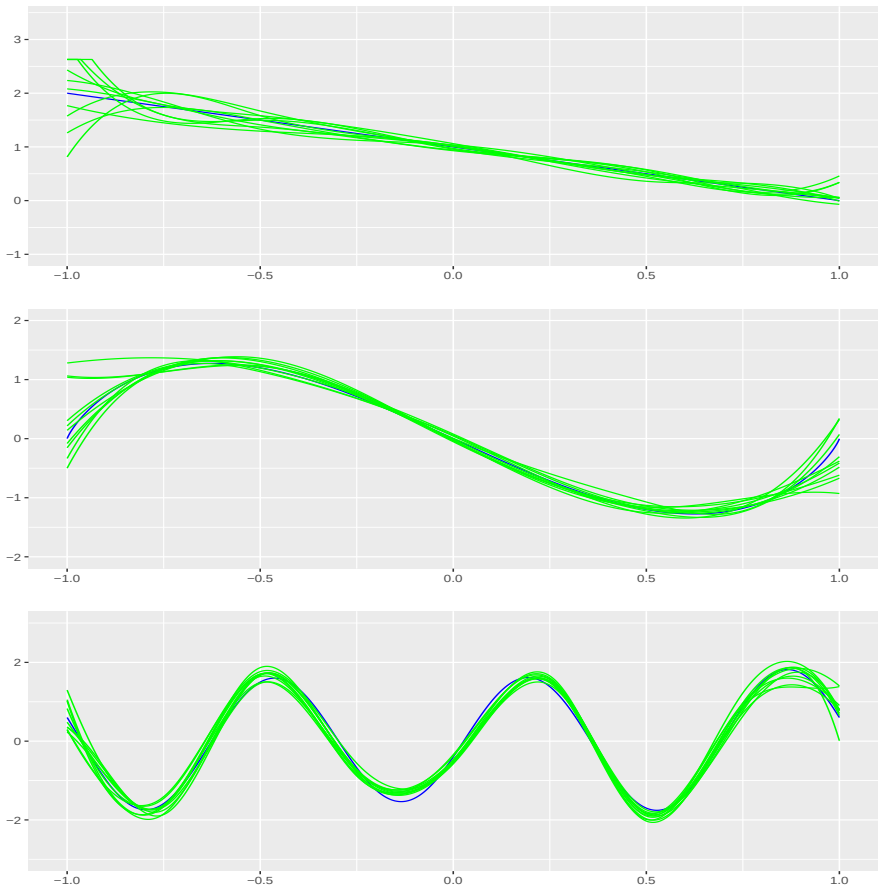

Figure 2: The three graphs show the three models 4-6-7 (top to bottom) and on each of them the true drift function in blue (dark) and 10 estimates $\widehat{b}_{\widehat{K}}$ in green (light grey), on the compact interval $[-1, 1]$

of sample paths with common fixed and finite time horizon by cutting the trajectory into several pieces. The idea would then be to apply similar results as those presented in this paper to estimate the drift coefficient b over a compact interval and finite time horizon. Here again, we believe that some technical issues have to be considered. For example, in our setting the sample paths start from a common starting point x_0, so we would have to make sure that our slicing permits to recover an N i.i.d. sample of observation paths that all start from x_0. The number N of trajectories would then become a random number and the methodology of proofs developed in the present paper could not be applied directly.

In future works, we also plan to extend our results to more general models such as inhomogeneous diffusion processes or stochastic differential equations driven by Levy processes.

References

Abraham, K. (2019) Nonparametric Bayesian posterior contraction rates for scalar diffusions with high-frequency data. *Bernoulli*, **25**(4A), 2696–2728.

Baraud, Y., Comte, F. & Viennet, G. (2001). Model selection for (auto-) regression with dependent data. *ESAIM: Probability and Statistics* **5**, 33–49.

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

Comte, F. & Genon-Catalot, V. (2019). Non parametric drift estimation for iid paths of stochastic differential equations. *The Annals of Statistics* To appear.

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

Dalalyan, A. *et al.* (2005). Sharp adaptive estimation of the drift function for ergodic diffusions. *The Annals of Statistics* **33**, 2507–2528.

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

Denis, C., Dion, C. & Martinez, M. (2019). Consistent procedures for multiclass classification of discrete diffusion paths. *Scandinavian Journal of Statistics* To appear.

Fournier, N., Printems, J. *et al.* (2010). Absolute continuity for some one-dimensional processes. *Bernoulli* **16**, 343–360.

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

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

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

Gugushvili, Shota, and Peter Spreij. (2014) *Consistent non-parametric Bayesian estimation for a time-inhomogeneous Brownian motion.* (2014) *ESAIM: Probability and Statistics* **18** 332–341.

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

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc.

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

Iacus, S. (2008). *Simulation and Inference for stochastic differential equation.* Springer Series in Statistics.

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

Konakov, V. & Menozzi, S. (2017). Weak error for the Euler scheme approximation of diffusions with non-smooth coefficients. *Electronic Journal of Probability* **22**.

Koskela, Jere, Dario Spano, and Paul A. Jenkins. (2019) *Consistency of Bayesian nonparametric inference for discretely observed jump diffusions. Bernoulli* **25.3** 2183–2205.

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

Lorentz, G. G., von Golitschek, M. & Makovoz, Y. (1996). *Constructive approximation: advanced problems*, vol. 304. Springer Berlin.

van der Meulen, Frank, and Harry Van Zanten. (2013). *Consistent nonparametric Bayesian inference for discretely observed scalar diffusions.*, *Bernoulli* **19.1** 44–63.

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

Revuz, D. & Yor, M. (1999). *Continuous martingales and Brownian motion*, vol. 293. Springer-Verlag, Berlin.

Schmisser, E. (2013). Penalized nonparametric drift estimation for a multidimensional diffusion process. *Statistics* **47**, 61–84.

Tsybakov, A. (2009). *Introduction to nonparametric estimation.* Springer Series in Satistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

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

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

Appendix

This section gathers the proofs of our results.

A Technical results

Throughout the paper, we will use the following results

Lemma A.1. *Let $X_1, \dots, X_N$ be independent copies of a random variable X such that $0 \leq X_1 \leq L$ for $L > 0$. Then for all $t > 0$,*

$$\mathbb{P}\left(\mathbb{E}[X] - \frac{2}{N}\sum_{i=1}^{N} X_i > t\right) \leq \exp\left(-\frac{Nt}{11L}\right).$$

Proof. The proof is a direct application of the Bernstein inequality. Indeed, we have for $t > 0$

$$\mathbb{P}\left(\mathbb{E}[X] - \frac{2}{N}\sum_{i=1}^{N} X_i > t\right) \leq \mathbb{P}\left(\mathbb{E}[X] - \frac{1}{N}\sum_{i=1}^{N} X_i > (t + \mathbb{E}[X])/2\right).$$

Hence, from the Bernstein inequality, we deduce

$$\mathbb{P}\left(\mathbb{E}[X] - \frac{2}{N}\sum_{i=1}^{N} X_i > t\right) \leq \exp\left(-\frac{N(t+\mathbb{E}[X])^2}{8\mathrm{Var}(X_1) + \frac{8}{3}Lt}\right). \tag{A.1}$$

Since $0 \leq \mathbb{E}[X]$ and $\mathrm{Var}(X_1) \leq \mathbb{E}\left[X_1^2\right] \leq L\mathbb{E}[X_1]$, we have $8\mathrm{Var}(X_1) + \frac{8}{3}Lt \leq 8L(t + \mathbb{E}[X]) + \frac{8}{3}L(\mathbb{E}[X]) \leq 11L(t + \mathbb{E}[X])$. Therefore, from Equation (A.1), we deduce

$$\mathbb{P}\left(\mathbb{E}[X] - \frac{2}{N}\sum_{i=1}^{N} X_i > t\right) \leq \exp\left(-\frac{N(t+\mathbb{E}[X])}{11L}\right) \leq \exp\left(-\frac{Nt}{11L}\right).$$

□

Lemma A.2. *Let $X_1, \dots, X_N$ be independent copies of a random variable $X \in \mathcal{X}$. Let $\mathcal{G}$ a class of real-valued functions on $\mathcal{X}$. For each $g \in \mathcal{G}$, and $x \in \mathcal{X}$, we assume that $0 \leq g(x) \leq L$, with $L > 0$. We consider $\mathcal{G}_\varepsilon$ an ε-net of $\mathcal{G}$ w.r.t $\|.\|_\infty$ and we denote by $\mathcal{N}_\infty(\varepsilon, \mathcal{G})$ its cardinality. Then, the following holds*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\mathbb{E}[g(X)] - \frac{2}{N}\sum_{i=1}^{N} g(X_i)\right)\right] \leq 3\varepsilon + \frac{11L\log(\mathcal{N}_\infty(\varepsilon,\mathcal{G}))}{N}.$$

Proof. Let $g \in \mathcal{G}$. We consider g_ε such that $\|g - g_\varepsilon\|_\infty \leq \varepsilon$. We obtain,

$$\mathbb{E}[g(X)] - \frac{2}{N}\sum_{i=1}^{N} g(X_i) \leq \mathbb{E}[g(X) - g_\varepsilon(X)] - \frac{2}{N}\sum_{i=1}^{N}(g(X_i) - g_\varepsilon(X_i)) + \mathbb{E}[g_\varepsilon(X)] - \frac{2}{N}\sum_{i=1}^{N} g_\varepsilon(X_i).$$

Therefore, we deduce from the above inequality that

$$\sup_{g\in\mathcal{G}} \mathbb{E}[g(X)] - \frac{2}{N}\sum_{i=1}^{N} g(X_i) \leq 3\varepsilon + \sup_{g\in\mathcal{G}_\varepsilon}\left\{\mathbb{E}[g(X)] - \frac{2}{N}\sum_{i=1}^{N} g(X_i)\right\}. \tag{A.2}$$

Now, we bound the last term in the r.h.s. We have that for $u \geq 0$,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_\varepsilon}\left\{\mathbb{E}\left[g(X)\right]-\frac{2}{N}\sum_{i=1}^{N}g(X_i)\right\}\right]\leq \\ u+\int_{t\geq u}\mathbb{P}\left(\sup_{g\in\mathcal{G}_\varepsilon}\left\{\mathbb{E}\left[g(X)\right]-\frac{2}{N}\sum_{i=1}^{N}g(X_i)\right\}\geq t\right)\mathrm{d}t.$$

Then, Lemma A.1 lead, for $t>0$, to

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_\varepsilon}\left\{\mathbb{E}\left[g(X)\right]-\frac{2}{N}\sum_{i=1}^{N}g(X_i)\right\}\right]\leq u+\mathcal{N}_\infty(\varepsilon,\mathcal{G})\int_{t\geq u}\exp\left(-\frac{Nt}{11L}\right)\mathrm{d}t.$$

Finally, setting $u=\dfrac{11L\log\left(\mathcal{N}_\infty(\varepsilon,\mathcal{G})\right)}{N}$ we obtain

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_\varepsilon}\left(\mathbb{E}\left[g(X)\right]-\frac{2}{N}\sum_{i=1}^{N}g(X_i)\right)\right]\leq\frac{11L\log\left(\mathcal{N}_\infty(\varepsilon,\mathcal{G})\right)}{N}.$$

The last inequality and Equation (A.2) yield the expected result. □

B Proof of Section 2

Proof of Proposition 2.3. Let h denote some measurable function. We have,

$$\begin{aligned}\left|\|h\|_b^2-\|h\|_{n,b}^2\right| &= \left|\mathbb{E}\frac{1}{T}\int_0^T\left(h^2\left(X_s\right)-h^2(X_{\eta(s)})\right)ds\right| \\ &= \left|\mathbb{E}\frac{1}{T}\int_0^T\left(h\left(X_s\right)-h(X_{\eta(s)})\right)\left(h\left(X_s\right)+h(X_{\eta(s)})\right)ds\right|\end{aligned}$$

with $\eta(s)=k\Delta,\ k\Delta\leq s<(k+1)\Delta$. Then, since h is L_h-Lipschitz, there exists C such that for each $s\in[0,T]$

$$\begin{aligned}\left|h\left(X_s\right)+h(X_{\eta(s)})\right| &\leq \left|h\left(X_s\right)-h(0)\right|+\left|h(X_{\eta(s)})-h(0)\right|+2|h(0)| \\ &\leq 2L_h\sup_{u\in[0,T]}|X_u|+2|h(0)| \\ &\leq 2\left(L_h\vee|h(0)|\right)\left(1+\sup_{u\in[0,T]}|X_u|\right).\end{aligned}$$

Since $\mathbb{E}\left[1+\sup_{u\in[0,T]}|X_u|\right]<\infty$, using Cauchy-Schwarz inequality and Equation (2.1) we obtain

$$\left|\|h\|_b^2-\|h\|_{n,b}^2\right|\leq \\ \frac{2L_h\left(L_h\vee|h(0)|\right)}{T}\int_0^T\mathbb{E}\left[(X_s-X_{\eta(s)})^2\right]^{1/2}ds\,\mathbb{E}\left[\left(1+\sup_{u\in[0,T]}|X_u|\right)^2\right]^{1/2},$$

which yields

$$\left|\|h\|_b^2-\|h\|_{n,b}^2\right|\leq CL_h\left(L_h\vee|h(0)|\right)\Delta^{1/2}.$$

Then, besides if h is bounded, we naturally obtain

$$\left|\|h\|_b^2-\|h\|_{n,b}^2\right|\leq CL_h\|h\|_\infty\Delta^{1/2}.$$

□

Proof of Lemma 2.6. Let us denote $h = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}}$. Then,

$$\sum_{i=-M}^{K_N-1} |a_i B_{i,M,\mathbf{u}}| \leq \left(\sum_{i=-M}^{K_N-1} a_i^2 B_{i,M,\mathbf{u}} \right)^{1/2} \left(\sum_{i=-M}^{K_N-1} B_{i,M,\mathbf{u}} \right)^{1/2}$$

thus as for all $x \in [A_N, B_N)$, $\sum_{i=-M}^{K_N-1} B_{i,M,\mathbf{u}}(x) = 1$,

$$\left(\sum_{i=-M}^{K_N-1} |a_i B_{i,M,\mathbf{u}}(x)| \right)^2 \leq \sum_{i=-M}^{K_N-1} a_i^2 B_{i,M,\mathbf{u}}(x)$$

and

$$\|h\|^2 \leq \sum_{i=-M}^{K_N-1} a_i^2 \int_{A_N}^{B_N} B_{i,M,\mathbf{u}}(x) dx.$$

But we have that $\int_{A_N}^{B_N} B_{i,M,\mathbf{u}}(x) dx = \frac{u_{i+M+1} - u_i}{M+1} \leq \frac{B_N - A_N}{K_N}$ from point (iv) of Proposition 2.5 (see *e.g.* De Boor *et al.*, 1978). Thus

$$\|h\|^2_{[A_N,B_N]} \leq \frac{B_N - A_N}{K_N} \|a\|_2^2.$$

For the second inequality, we first observe that for a polynomial P of degree M or less,

$$\sup_{x \in [0,1]} |P(x)| \leq C \int_0^1 |P(u)| du$$

where $C > 0$ does only depend on M (this result is the consequence of the norm equivalence in finite dimensional vector space). Therefore with a change of variable the result holds for an arbitrary interval $[a, b]$:

$$\sup_{x \in [a,b]} |P(x)| \leq \frac{C}{b-a} \int_a^b |P(u)| du.$$

Since h is a polynomial of degree M or less on each interval $[u_i, u_{i+1})$ for all i,

$$\max_{u_i \leq x \leq u_{i+1}} |h(x)| \leq \frac{C K_N}{B_N - A_N} \int_{u_i}^{u_{i+1}} |h(y)| dy. \tag{B.1}$$

Then, as in De Boor *et al.* (1978) Equation (5) Chapter XI, we have

$$|a_i| \quad \leq \quad C \max_{u_i \leq x \leq u_{i+M+1}} |h(x)| \leq \sum_{j=1}^{M+1} \max_{u_{i+j-1} \leq x \leq u_{i+j}} |h(x)|.$$

From the above Equation applying Chasles relation and with Equation (B.1), we obtain

$$|a_i| \quad \leq \quad \frac{C K_N}{B_N - A_N} \int_{u_i}^{u_{i+M+1}} |h(y)| dy.$$

Then,

$$\sum_{i=-M}^{K_n-1} a_i^2 \leq \frac{C K_N}{B_N - A_N} \left(\sum_{i=-M}^{K_n-1} \int_{u_i}^{u_{i+M+1}} h^2(y) dy \right)$$

$$\begin{aligned}
\sum_{i=-M}^{K_n-1} \int_{u_i}^{u_{i+M+1}} h^2(y) dy &= \sum_{i=-M}^{K_n-1} \sum_{j=i}^{M-2} \int_{u_j}^{u_{j+1}} h^2(y) dy \\
&= \sum_{j=-M}^{M-2} \sum_{i=-M}^{j} \int_{u_j}^{u_{j+1}} h^2(y) dy \\
&\leq \sum_{j=-M}^{M-2} \|h\|^2.
\end{aligned}$$

Finally, applying Cauchy-Schwarz inequality and summing over $i = -M, \dots, K_N - 1$ we get

$$\|\mathbf{a}\|_2^2 \leq C\|h\|^2 \frac{K_N}{(B_N - A_N)}.$$

Let us now prove the second equation of the Lemma. For all $x \in [A_N, B_N]$, from Cauchy-Schwarz inequality,

$$|h(x)| = \left| \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}} \right| \leq \|\mathbf{a}\|_2 \left(\sum_{i=-M}^{K_N-1} B_{i,M,\mathbf{u}}^2 \right)^{1/2}.$$

Since for all $x \in [A_N, B_N)$, $0 \leq B_{i,M-1,\mathbf{u}}(x) \leq 1$ and $\sum_{i=-M}^{K_N-1} B_{i,M,\mathbf{u}}(x) = 1$, the above inequality yields the result.

□

C Proof of Section 3

Proof of Proposition 3.2. Let us denote:

$$\begin{aligned} Z_{k\Delta}^{(j)} &:= \frac{X_{(k+1)\Delta}^{(j)} - X_{k\Delta}^{(j)}}{\Delta} \\ &= b(X_{k\Delta}^{(j)}) + \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} \sigma(X_s^{(j)}) dW_s^{(j)} + \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} (b(X_s^{(j)}) - b(X_{k\Delta}^{(j)}))ds \\ &:= b(X_{k\Delta}^{(j)}) + \Sigma_{k,j} + R_{k,j}. \end{aligned}$$

Let $h = \sum_{i=-M}^{K_N-1} h_i B_{i,M,\mathbf{u}} \in \mathcal{S}_{K,L,M}$, we first introduce the following notation

$$\nu_{N,n}(h) := \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} h(X_{k\Delta}^{(j)}) \Sigma_{k,j}. \tag{C.1}$$

From Equation (3.3), we get

$$\gamma_{N,n}(h) - \gamma_{N,n}(b) = \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} (h-b)^2(X_{k\Delta}^{(j)}) + 2\nu_{N,n}(b-h) + \frac{2}{N} \sum_{j=1}^{N} \frac{1}{n} \sum_{k=0}^{n-1} (b-h)(X_{k\Delta}^{(j)}) R_{k,j}$$

Besides, $\gamma_{N,n}(\widehat{b}_{N,n}) - \gamma_{N,n}(b) \leq \gamma_{N,n}(h) - \gamma_{N,n}(b)$, therefore we get

$$\|\widehat{b}_{N,n} - b\|_{N,n}^2 \leq \|h - b\|_{N,n}^2 + 2\nu_{N,n}(\widehat{b}_{N,n} - h) + \frac{2}{N} \sum_{j=1}^{N} \frac{1}{n} \sum_{k=0}^{n-1} (\widehat{b}_{N,n} - h)(X_{k\Delta}^{(j)}) R_{k,j}.$$

Using for $a > 0$ the relation $2xy \leq \frac{1}{a}x^2 + ay^2$, we have

$$\begin{aligned} \frac{2}{N} \sum_{j=1}^{N} \frac{1}{n} \sum_{k=0}^{n-1} (\widehat{b}_{N,n} - h)(X_{k\Delta}^{(j)}) R_{k,j} &\leq \frac{1}{a}\|\widehat{b}_{N,n} - h\|_{N,n}^2 + \frac{a}{N}\frac{1}{n} \sum_{j=1}^{N} \sum_{k=0}^{n-1} R_{k,j}^2 \\ &\leq \frac{2}{a}\left(\|h - b\|_{N,n}^2 + \|\widehat{b}_{N,n} - b\|_{N,n}^2 \right) + \frac{a}{N}\frac{1}{n} \sum_{j=1}^{N} \sum_{k=0}^{n-1} R_{k,j}^2. \end{aligned}$$

Hence, as $\mathbb{E}\left[R_{k,j}^2\right] \leq C\Delta$ (see Lemma 7.3 Denis *et al.*, 2019), we obtain

$$\left(1 - \frac{2}{a}\right) \mathbb{E}\left[\|\widehat{b}_{N,n} - b\|_{N,n}^2 \right] \leq \left(1 + \frac{2}{a}\right) \|h - b\|_{n,b} + 2\mathbb{E}\left[\nu_{N,n}(\widehat{b}_{N,n} - h) \right] + aC\Delta. \tag{C.2}$$

Since the functions h and $\widehat{b}_{N,n}$ are in $\mathcal{S}_{K_N,L_N,M}$, we observe that thanks to the Cauchy-Schwarz inequality

$$\begin{aligned}\nu_{N,n}(\widehat{b}_{N,n}-h) &= \sum_{i=-M}^{K_N-1}(\widehat{a}_i-h_i)\nu_{N,n}(B_{i,M,\mathbf{u}}) \\ &\leq 2\sqrt{(K_N+M)L_N}\sqrt{\sum_{i=-M}^{K_N-1}\nu_{N,n}^2(B_{i,M,\mathbf{u}})}.\end{aligned}$$

Therefore, applying again the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\mathbb{E}\left[\nu_{N,n}(\widehat{b}_{N,n}-h)\right] &\leq \sqrt{\mathbb{E}\left[\nu_{N,n}^2(\widehat{b}_{N,n}-h)\right]} \\ &\leq 2\sqrt{(K_N+M)L_N}\sqrt{\mathbb{E}\left[\sum_{i=-M}^{K_N-1}\nu_{N,n}^2(B_{i,M,\mathbf{u}})\right]}.\end{aligned}$$

Since $\sum_{i=-M}^{K_N-1}B_{i,M,\mathbf{u}}^2(x)\leq 1$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=-M}^{K_N-1}\nu_{N,n}^2(B_{i,M,\mathbf{u}})\right] &= \frac{1}{n^2\Delta^2N^2}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\mathbb{E}\left[\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_s^{(j)})ds\sum_{i=-M}^{K_N-1}B_{i,M,\mathbf{u}}^2(X_{k\Delta}^{(j)})\right] \\ &\leq \frac{1}{n^2\Delta^2N^2}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\mathbb{E}\left[\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_s^{(j)})ds\right] \\ &= \frac{1}{T^2N^2}\sum_{j=1}^{N}\mathbb{E}\left[\int_0^T\sigma^2(X_s^{(j)})ds\right]\leq C\frac{1}{NT} \qquad \text{(C.3)}\end{aligned}$$

as $(1/T)\mathbb{E}\left[\int_0^T\sigma^2(X_s^{(j)})ds\right]<\infty$ which concludes the proof (for example with $a=3$).

□

Proof of Theorem 3.3. Let us define

$$\widetilde{b}(x):=\begin{cases}b(x) & \text{if } |b(x)|\leq\sqrt{L_N},\\ \operatorname{sgn}(b(x))\sqrt{L_N} & \text{if } |b(x)|>\sqrt{L_N}.\end{cases}$$

Let us start with the following lemma. We remind the reader that L_b is the notation for the Lipschitz constant of b.

Lemma C.1. *Under assumption of Theorem 3.3, the following holds : for any $\alpha>1$,*

$$\inf_{h\in\mathcal{S}_{K_N,L_N,M}}\|h-b\|_{n,b}^2\leq\left(\frac{2L_b(M+1)B_N}{K_N}\right)^2+\frac{C_\alpha}{L_N^{\alpha-1}}+C_b\sqrt{\Delta}+2\|b-\widetilde{b}\|_b^2.$$

Proof. For each $h\in\mathcal{S}_{K_N,L_N,M}$ since b and $\widetilde{b}$ are Lipschitz with the same constant, from Proposition 2.3, we deduce

$$\begin{aligned}\|h-b\|_{n,b}^2 &\leq 2\|h-\widetilde{b}\|_{n,b}^2+2\|b-\widetilde{b}\|_{n,b}^2 \\ &\leq 2\|h-\widetilde{b}\|_{n,b}^2+2\|b-\widetilde{b}\|_b^2+C_b\sqrt{\Delta}. \qquad \text{(C.4)}\end{aligned}$$

Now, we study the first term in the r.h.s. We define

$$\tilde{h}=\sum_{i=-M}^{K_N-1}\widetilde{b}(u_i)B_{i,M,\mathbf{u}}.$$

Since by definition we have $\sum_{i=-M}^{K_N-1}\widetilde{b}^2(u_i) \leq (K_N+M)L_N$, we have that $\tilde{h} \in \mathcal{S}_{K_N,L_N,M}$. Let $x \in [u_{i_0}, u_{i_0+1})$, for $0 \leq i_0 \leq K_N - 1$. Then, as the $B_{i,M,u}$ are nonnegative functions we get

$$\begin{aligned}
|\tilde{h}(x) - \widetilde{b}(x)| &= \left| \sum_{i=-M}^{K_N-1} \left(\widetilde{b}(x) - \widetilde{b}(u_i)\right) B_{i,M,\mathbf{u}}(x) \right| \\
&= \left| \sum_{i=i_0-M}^{i_0} \left(\widetilde{b}(x) - \widetilde{b}(u_i)\right) B_{i,M,\mathbf{u}}(x) \right| \\
&\leq \max_{i=i_0-M,\ldots,i_0} \left|\widetilde{b}(x) - \widetilde{b}(u_i)\right| \sum_{i=i_0-M}^{i_0} B_{i,M,\mathbf{u}}(x) \\
&\leq L_b |u_{i_0+1} - u_{i_0-M}| \\
&\leq L_b \frac{2(M+1)B_N}{K_N}.
\end{aligned}$$

Hence, the last inequality implies that for all $x \in (-B_N, B_N)$ (remember $A_N = -B_N$),

$$|\tilde{h}(x) - \widetilde{b}(x)| \leq \frac{2L_b(M+1)B_N}{K_N}.$$

Now, since

$$\inf_{h \in \mathcal{S}_{K_N,L_N,M}} \|h - \widetilde{b}\|_{n,b}^2 \leq \|\tilde{h} - \widetilde{b}\|_{n,b}^2$$

and $B_N > L_N$ by assumption, we have

$$\begin{aligned}
\|\tilde{h} - \widetilde{b}\|_{n,b}^2 &\leq \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widetilde{h}(X_{k\Delta}) - \widetilde{b}(X_{k\Delta})\right)^2 \left(\mathbb{1}_{X_{k\Delta} \notin [-L_N, L_N]} + \mathbb{1}_{X_{k\Delta} \in [-L_N, L_N]}\right)\right] \\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widetilde{h}(X_{k\Delta}) - \widetilde{b}(X_{k\Delta})\right)^2 \mathbb{1}_{X_{k\Delta} \notin [-L_N, L_N]}\right] + \frac{2L_b^2(M+1)^2 B_N^2}{K_N^2}. \qquad \text{(C.5)}
\end{aligned}$$

Let us deal with the first term in the r.h.s. Since $\|\tilde{h}\|_\infty \leq \|\widetilde{b}\|_\infty$, we have for $\alpha > 1$

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widetilde{h}(X_{k\Delta}) - \widetilde{b}(X_{k\Delta})\right)^2 \mathbb{1}_{X_{k\Delta} \notin [-L_N, L_N]}\right] &\leq 4\|\widetilde{b}\|_\infty^2 \sup_{s\in[0,T]} \mathbb{P}\left(|X_s| > L_N\right) \\
&= 4\, L_N \sup_{s\in[0,T]} \mathbb{P}\left(|X_s|^\alpha > L_N^\alpha\right).
\end{aligned}$$

Using Markov's inequality, we obtain

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widetilde{h}(X_{k\Delta}) - \widetilde{b}(X_{k\Delta})\right)^2 \mathbb{1}_{X_{k\Delta} \notin [-L_N, L_N]}\right] \leq \frac{C}{L_N^{\alpha-1}}. \qquad \text{(C.6)}$$

Combining Equations (C.4), (C.5), and (C.6) we get the desired result. □

Now, we go back to the proof of the theorem. We have that:

$$\|\widehat{b}_{N,n} - b\|_b^2 \leq 2\|\widehat{b}_{N,n} - \widetilde{b}\|_b^2 + 2\|\widetilde{b} - b\|_b^2. \qquad \text{(C.7)}$$

Let us show that $\widehat{b}_{N,n}$ is Lipschitz with a Lipschitz constant bounded by $C\sqrt{(K_N+M)L_N}\frac{K_N}{B_N}$. Indeed from Györfi *et al.* (2006) Lemma 14.6 we get that for $x \in [A_N, B_N)$

$$\left|\sum_{i=-M}^{K_N-1} \widehat{a}_i B'_{i,M,\mathbf{u}}(x)\right| \leq \sum_{i=-(M-1)}^{K_N-1} \frac{M}{u_{i+M} - u_i} |\widehat{a}_i - \widehat{a}_{i-1}| |B_{i,M-1,\mathbf{u}}(x)|$$

thus

$$\sum_{i=-M}^{K_N-1} \widehat{a}_i B'_{i,M,\mathbf{u}}(x) \leq \left(\sum_{i=-(M-1)}^{K_N-1} \left(\frac{M}{u_{i+M}-u_i} B_{i,M-1,\mathbf{u}}(x) \right)^2 \right)^{1/2} \left(\sum_{i=-(M-1)}^{K_N-1} (\widehat{a}_i - \widehat{a}_{i-1})^2 \right)^{1/2}$$

then,

$$\left(\sum_{i=-(M-1)}^{K_N-1} (\widehat{a}_i - \widehat{a}_{i-1})^2 \right)^{1/2} \leq \sqrt{2} \|\widehat{\mathbf{a}}\|_2 \leq \sqrt{2} \sqrt{(K_N+M)L_N}$$

and

$$\left(\sum_{i=-(M-1)}^{K_N-1} \left(\frac{M}{u_{i+M}-u_i} B_{i,M-1,\mathbf{u}}(x) \right)^2 \right)^{1/2} \leq \frac{K_N}{B_N - A_N} \left(\sum_{i=-(M-1)}^{K_N-1} B^2_{i,M-1,\mathbf{u}}(x) \right)^{1/2}.$$

Since, for all $x \in [A_N, B_N)$, $\sum_{i=-(M-1)}^{K_N-1} B^2_{i,M-1,\mathbf{u}}(x) \leq 1$, we deduce from the above inequality with $A_N = -B_N$ that

$$\left(\sum_{i=-(M-1)}^{K_N-1} \left(\frac{M}{u_{i+M}-u_i} B_{i,M-1,\mathbf{u}}(x) \right)^2 \right)^{1/2} \leq \frac{K_N}{2B_N}.$$

Besides, we have that $\widetilde{b}$ is Lipschitz and $\|\widetilde{b}\|_\infty \leq \sqrt{L_N}$. Furthermore, according to Lemma 2.6 we have $\|\widehat{b}_{N,n}\|_\infty \leq \|\mathbf{a}\|_2 \leq \sqrt{(K_N+M)L_N}$. Hence, applying Proposition 2.3 with $h = (\widehat{b}_{N,n} - \widetilde{b})$, we obtain for N large enough ($L_h < 2\sqrt{(K_N+M)L_N} K_N/B_N$),

$$\|\widehat{b}_{N,n} - \widetilde{b}\|_b^2 \leq \|\widehat{b}_{N,n} - \widetilde{b}\|_{n,b}^2 + C((K_N+M)L_N) \frac{K_N}{B_N} \Delta^{1/2} \tag{C.8}$$

Now, adding and subtracting an artificial term, we have

$$\|\widehat{b}_{N,n} - \widetilde{b}\|_{n,b}^2 = \|\widehat{b}_{N,n} - \widetilde{b}\|_{n,b}^2 - 2\|\widehat{b}_{N,n} - \widetilde{b}\|_{N,n}^2 + 2\|\widehat{b}_{N,n} - \widetilde{b}\|_{N,n}^2.$$

Since by definition of $\widetilde{b}$ we have $\|\widehat{b}_{N,n} - \widetilde{b}\|_{N,n}^2 \leq \|\widehat{b}_{N,n} - b\|_{N,n}^2$, we deduce with Proposition 3.2, Equation (C.7), and Equation (C.8)

$$\begin{aligned} \mathbb{E}\left[\left\| \widehat{b}_{N,n} - b \right\|_b^2 \right] \quad \leq \quad & 2\mathbb{E}\left[\|\widehat{b}_{N,n} - \widetilde{b}\|_{n,b}^2 - 2\|\widehat{b}_{N,n} - \widetilde{b}\|_{N,n}^2 \right] \\ & + 2 \inf_{h \in \mathcal{S}_{K_N, L_N, M}} \|h - b\|_{n,b}^2 + C\left(\sqrt{\frac{(K_N+M)L_N}{N}} + \Delta \right) \\ & + C((K_N+M)L_N) \frac{K_N}{B_N} \Delta^{1/2} + 2\|\widetilde{b} - b\|_b^2. \end{aligned} \tag{C.9}$$

Now, we deal with the first term in the right hand side. We observe that

$$\mathbb{E}\left[\|\widehat{b}_{N,n} - \widetilde{b}\|_{n,b}^2 - \frac{2}{N} \sum_{j=1}^{N} \|\widehat{b}_{N,n} - \widetilde{b}\|_n^{2,(j)} \right] \leq \mathbb{E}\left[\sup_{h \in \mathcal{S}_{K_N, L_N, M}} \left(\|h - \widetilde{b}\|_{n,b}^2 - 2\|h - \widetilde{b}\|_{N,n}^2 \right) \right].$$

For $h \in \mathcal{S}_{K_N, L_N, M}$, we define the function g_h as

$$g_h(x_1, \ldots, x_n) = \frac{1}{n} \sum_{k=1}^{n} (h(x_k) - \tilde{b}(x_k))^2, \ \ \forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{C.10}$$

and set $\mathcal{G} = \{g_h, \ \ h \in \mathcal{S}_{K_N,L_N,M}\}$. Then, as by definition $\|h\|^2_{N,n} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n}\sum_{k=0}^{n-1} h^2(X^{(j)}_{k\Delta}))$, we have

$$\mathbb{E}\left[\|\widehat{b}_{N,n} - \widetilde{b}\|^2_{n,b} - 2\|\widehat{b}_{N,n} - \widetilde{b}\|^2_{N,n}\right] \leq \mathbb{E}\left[\sup_{g_h\in\mathcal{G}}\left(\mathbb{E}\left[g_h(\overline{X})\right] - \frac{2}{N}\sum_{j=1}^{N} h\left(\overline{X}^j\right)\right)\right].$$

Since, for each $h \in \mathcal{S}_{K_N,L_N,M}$, $\|h\|_\infty \leq \sqrt{(K_N+M)L_N}$, and $\|\widetilde{b}\|_\infty \leq \sqrt{L_N}$, we deduce that for each $g_h \in \mathcal{G}$

$$0 \leq g_h(x) \leq 2\left(\|h\|^2_\infty + \|\widetilde{b}\|^2_\infty\right) \leq 4(K_N+M)L_N.$$

Therefore, for $\varepsilon > 0$ a direct application of Lemma A.2 yields

$$\mathbb{E}\left[\|\widehat{b}_{N,n} - \widetilde{b}\|^2_{n,b} - 2\|\widehat{b}_{N,n} - b\|^2_{N,n}\right] \leq 3\varepsilon + \frac{44(K_N+M)L_N\log(\mathcal{N}_\infty(\varepsilon,\mathcal{G}))}{N}. \tag{C.11}$$

Hence it remains to control $\mathcal{N}_\infty(\varepsilon,\mathcal{G})$. It is known that for an euclidean ball of radius R denoted $\overline{B}_2(0,R) \subset \mathbb{R}^{K_N+M}$ the covering numbers are controlled as

$$\mathcal{N}\left(\varepsilon, \overline{B}_2(0,R), \|.\|_2\right) \leq \left(\frac{3R}{\varepsilon}\right)^{K_N+M}$$

see for example Lorentz *et al.* (1996) Chapter 15 Prop 1.3. Then, since for a function $h = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}}$ we have that $\|h\|_\infty \leq \|\mathbf{a}\|_2$, we deduce

$$\mathcal{N}_\infty\left(\varepsilon, \mathcal{S}_{K_N,L_N,M}\right) \leq \left(\frac{3\sqrt{(K_N+M)L_N}}{\varepsilon}\right)^{K_N+M}. \tag{C.12}$$

Finally, considering an ε-net of $\mathcal{S}_{K_N,L_N,M}$ w.r.t $\|.\|_\infty$, we observe that for $g_h \in \mathcal{G}$ and g_{h_ε} such that $\|h - h_\varepsilon\|_\infty \leq \varepsilon$,

$$\begin{aligned}
|g_h(x) - g_{h_\varepsilon}(x)| &= \left|\frac{1}{n}\sum_{k=0}^{n-1}(h(x_k) - \widetilde{b}(x_k))^2 - (h_\varepsilon(x_k) - \widetilde{b}(x_k))^2\right| \\
&= \left|\frac{1}{n}\sum_{k=0}^{n-1}(h(x_k) - (h_\varepsilon(x_k))(h(x_k) + h_\varepsilon(x_k) - 2\widetilde{b}(x_k))\right| \\
&\leq \frac{\varepsilon}{n}\sum_{k=0}^{n-1}\left|h(x_k) + h_\varepsilon(x_k) - 2\widetilde{b}(x_k)\right| \\
&\leq 4\varepsilon\sqrt{(K_N+M)L_N}.
\end{aligned}$$

Hence, from the last inequality, we deduce

$$\mathcal{N}_\infty\left(\varepsilon,\mathcal{G}\right) \leq \mathcal{N}_\infty\left(\frac{\varepsilon}{4\sqrt{(K_N+M)L_N}}, \mathcal{S}_{K_N,L_N,M}\right) \leq \left(\frac{12(K_N+M)L_N}{\varepsilon}\right)^{K_N+M}. \tag{C.13}$$

Therefore, setting $\varepsilon = \frac{12(K_N+M)L_N}{N}$ in Equation (C.11) yields

$$\mathbb{E}\left[\|\widehat{b}_{N,n} - \widetilde{b}\|^2_{n,b} - 2\|\widehat{b}_{N,n} - b\|^2_{N,n}\right] \leq \frac{36(K_N+M)L_N}{N} + \frac{44(K_N+M)^2L_N\log(N)}{N}.$$

Since $\|\widetilde{b} - b\|^2_b \to 0$ from Lebesgue's dominated convergence theorem and $\Delta = O\left(\left(\frac{B_N}{K_N N}\right)^2\right)$, we deduce from Equation (C.9) and Lemma C.1 the desired result provided that

$$\frac{(K_N+M)^2L_N\log(N)}{N} \to 0 \ \text{ and } \ \frac{B_N}{K_N} \to 0.$$

□

D Proof of Section 4

We start this section with an important result which provide a control of the set where the empirical norm $\|.\|_{N,n}$ and the norm $\|.\|_{n,b}$ are equivalent. We also consider $\mathcal{K}_N = \{1, \dots, K_N^*\}$, where $K_N^* \leq N$.

Lemma D.1. *Let us define the set for $K_N \in \mathcal{K}_N$,*

$$\Omega_{N,n,K_N} = \bigcap_{h \in \mathcal{S}_{K_N,M,\mathbf{u}} \setminus \{0\}} \left\{ \omega \in \Omega, \ \left| \frac{\|h\|_{N,n}^2}{\|h\|_{n,b}^2} - 1 \right| \leq \frac{1}{2} \right\}. \tag{D.1}$$

Then if $K_N^* = o\left(\sqrt{N/\log(N)}\right)$, for each $K_N \in \mathcal{K}_N$ and N large enough

$$\mathbb{P}\left(\Omega_{N,n,K_N}^c\right) \leq \frac{C_{\pi_0}}{N}.$$

Proof. First, observe that, for $h \in \mathcal{S}_{K,L_N,M}$,

$$\frac{\mathbb{E}\|h\|_{N,n}^2}{\|h\|_{n,b}^2} = 1 \ \text{ and } \ \Omega_{N,n,K}^c = \left\{ \omega \in \Omega, \ \exists h_0 \in \mathcal{S}_{K,M,\mathbf{u}} \setminus \{0\}, \left| \frac{\|h_0\|_{N,n}^2}{\|h_0\|_{n,b}^2} - 1 \right| > 1/2 \right\}.$$

Therefore

$$\sup_{h \in \mathcal{S}_{K_N,M,\mathbf{u}} \setminus \{0\}} \left| \frac{\|h\|_{N,n}^2}{\|h\|_{n,b}^2} - 1 \right| = \sup_{h \in \mathcal{S}_{K_N,M,\mathbf{u}}, \|h\|_{n,b}^2 = 1} \left| \|h\|_{N,n}^2 - \mathbb{E}[\|h\|_{N,n}^2] \right|$$

We denote by $\mathcal{H} = \{h \in \mathcal{S}_{K_N,M,\mathbf{u}}, \ \ \|h\|_{n,b}^2 = 1\}$ and consider $\mathcal{H}_\varepsilon$ an ε-net of $\mathcal{H}$ w.r.t $\|.\|_\infty$. Thus, for all $h \in \mathcal{H}$ there exists an element $\overline{h} \in \mathcal{H}_\varepsilon$ such that $\|h - \overline{h}\|_\infty \leq \varepsilon$. Then,

$$\left| \|h\|_{N,n}^2 - 1 \right| \leq \left| \|h\|_{N,n}^2 - \|\overline{h}\|_{N,n}^2 \right| + \left| \|\overline{h}\|_{N,n}^2 - 1 \right|.$$

We have for all $x \in [0,1]$ and $h = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}} \in \mathcal{H}$

$$h(x) \leq \left(\sum_{i=-M}^{K_N-1} a_i^2 \right)^{1/2} \left(\sum_{i=-M}^{K_N-1} B_{i,M,\mathbf{u}}^2(x) \right)^{1/2}.$$

Since for all $x \in [0,1)$, $0 \leq B_{i,M,\mathbf{u}}(x) \leq 1$, from the above inequality we get $\|h\|_\infty \leq \|\mathbf{a}\|_2$. Besides, Proposition 4.3, $\|h\|_{n,b} = 1$ implies $\|h\| \leq \dfrac{1}{\pi_0}$, then from Lemma 2.6, we deduce that for each $h = \sum_{i=-M}^{K-1} a_i B_{i,M,\mathbf{u}} \in \mathcal{H}$, with $\kappa = C_1^{-1}$,

$$\|h\|_\infty \leq \|\mathbf{a}\|_2 \leq \sqrt{\frac{\kappa K_N}{\pi_0^2}}. \tag{D.2}$$

Hence, we get

$$\left| \|h\|_{N,n}^2 - \|\overline{h}\|_{N,n}^2 \right| \leq 2\sqrt{\frac{\kappa K_N}{\pi_0^2}}\varepsilon,$$

Therefore

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| \|h\|_{N,n}^2 - 1 \right| \geq \delta \right) \leq \mathbb{P}\left(\sup_{\bar{h} \in \mathcal{H}_\varepsilon} \left| \|\bar{h}\|_{N,n}^2 - 1 \right| \geq \delta/2 \right) + \mathbb{1}_{4\sqrt{\frac{\kappa K_N}{\pi_0^2}}\varepsilon \geq \delta}.$$

Besides, Hoeffding inequality leads to

$$\mathbb{P}\left(\sup_{\bar{h} \in \mathcal{H}_\varepsilon} \left| \|\bar{h}\|_{N,n}^2 - 1 \right| \geq \delta/2 \right) \leq 2\mathcal{N}_\infty(\varepsilon, \mathcal{H}) \exp\left(-\frac{N\pi_0^2\delta^2}{2\kappa K_N} \right)$$

thus, for $K_N \in \mathcal{K}_N$, if $4\sqrt{\frac{\kappa K_N^*)}{\pi_0^2}}\varepsilon < \delta$ we obtain:

$$\mathbb{P}\left(\Omega^c_{N,n,K_N}\right) \leq 2\mathcal{N}_\infty(\varepsilon, \mathcal{H}) \exp\left(-\frac{N\delta^2}{2\kappa K_N^*}\right).$$

Now, it remains to control $\mathcal{N}_\infty(\varepsilon, \mathcal{H})$. For $K \in \mathcal{K}$, from Equation (D.2),

$$\mathcal{N}_\infty\left(\mathcal{H}, \varepsilon\right) \leq \left(\frac{3\sqrt{\kappa K_N}}{\pi_0 \varepsilon}\right)^{K_N+M} \leq \left(\frac{3\sqrt{\kappa K_N^*}}{\pi_0 \varepsilon}\right)^{K_N+M}.$$

We choose $\varepsilon = \dfrac{3\sqrt{\kappa K_N^*}}{\pi_0 N}$ which implies with $K_N^* \leq N$

$$\mathbb{P}\left(\Omega^c_{N,n,K_N}\right) \leq 2N^{(K_N^*+M)} \exp\left(-\frac{N\pi_0^2\delta^2}{2\kappa K_N^*}\right),$$

provided that $4\sqrt{\frac{\kappa K_N^*}{\pi_0^2}}\varepsilon < \delta$. Finally, we can choose δ such that

$$\delta^2 \geq \frac{2\kappa K_N^*}{\pi_0^2 N} \log\left(N^{(K_N^*+M)}\right).$$

Since $K_N^* = o\left(\sqrt{N/\log(N)}\right)$, we can take $\delta = 1/2$ which gives the result. □

Proof of Lemma 4.3. The upper bounds may be derived easily by using the Gaussian bounds for the transition density (see Gobet (2002)).

Let us now give the proof of the lower bound for the integral of the transition density. In order to simplify the proof and without loss of generality, we restrict ourselves to the particular case where $T = 1$. From Gobet (2002), we know that there exist constants $c > 1$ and $K > 1$ such that for all $(x, y) \in \mathbb{R}$ and for all $t \in (0, 1]$ the following lower bound holds :

$$p(t, x_0, y) \geq \frac{1}{Kt^{1/2}} \exp\left(-\frac{c|x_0 - y|^2}{t}\right) \exp\left(-c|x_0|^2 t\right).$$

Hence we have

$$K\mathrm{e}^{c|x_0|^2} p(t, x_0, y) \geq \frac{1}{t^{1/2}} \exp\left(-\frac{c|x_0 - y|^2}{t}\right). \tag{D.3}$$

Let us fix $y \in [0, 1]$. An easy computation (change of variable and integration by parts) shows that

$$\int_0^1 \frac{1}{\sqrt{s}} \exp\left(-\frac{c|x_0 - y|^2}{s}\right) ds = 2\mathrm{e}^{-c|x_0-y|^2} - 2\sqrt{c}|x_0 - y|\sqrt{\pi}\,\mathrm{erfc}\left(\sqrt{c}|x_0 - y|\right)$$

where $\mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-u^2} du$ stands for the complementary error function. Set $\ell : d \mapsto 2\mathrm{e}^{-d^2} - 2d\sqrt{\pi}\,\mathrm{erfc}\,(d)$ defined on $\mathbb{R}^{+*}$. The function ℓ is positive and strictly decreasing with $\lim_{d \searrow 0} \ell(d) = 2$. Since $|x_0 - y| \leq 1$, this shows that $\ell(\sqrt{c}|x_0 - y|) \geq \ell(\sqrt{c}) > 0$. Thus, we get

$$\int_0^1 \frac{1}{\sqrt{s}} \exp\left(-\frac{c|x_0 - y|^2}{s}\right) ds \geq \ell(\sqrt{c}) > 0. \tag{D.4}$$

Setting $\tilde{\pi}_0 = \frac{\mathrm{e}^{-c|x_0|^2}}{K} \ell\left(\sqrt{c}\right)$, we deduce from (D.3) that

$$\int_0^1 p(t, x_0, y) dt \geq \tilde{\pi}_0.$$

Now, we prove the first point of the lemma. From (D.3) it is enough to bound $\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp(-c|x_0-y|^2/k\Delta)$ from below. First for $y=x_0$, we observe that

$$\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\geq 1.$$

Now, let us fix $y\in[0,1]\setminus\{x_0\}$. Set $a=\sqrt{c}|x_0-y|$. Since $a>0$, we are allowed to make use of the convention $\mathrm{e}^{-a^2/0}=0$.

An elementary study of the function $p\ :\ s\mapsto\dfrac{\exp\left(-a^2/s\right)}{\sqrt{s}}$ on $(0,1]$ gives us insurance that p is strictly increasing on $(0,2a^2)$ and strictly decreasing on $(2a^2,1]$.

A lower bound for $a\geq\dfrac{1}{2}$. We start with the following decomposition.

$$\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\mathrm{e}^{-\frac{a^2}{k\Delta}}=\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\int_{(k-1)\Delta}^{k\Delta}\left(\mathrm{e}^{-a^2/s}+\left(\mathrm{e}^{-\frac{a^2}{k\Delta}}-\mathrm{e}^{-\frac{a^2}{s}}\right)\right)ds.$$

Hence, since $\frac{k-1}{k}\geq\frac{1}{2}$ for $k\geq 2$, a change of variable gives

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\mathrm{e}^{-\frac{a^2}{k\Delta}} &\geq \frac{1}{n\Delta}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\int_{a^2/k\Delta}^{a^2/(k-1)\Delta}\frac{a^2}{u^2}\mathrm{e}^{-u}du\\ &\geq \frac{1}{2n\Delta}\sum_{k=1}^{n-1}\sqrt{k\Delta}\left(\mathrm{e}^{-\frac{a^2}{k\Delta}}-\mathrm{e}^{-\frac{a^2}{(k-1)\Delta}}\right)\end{aligned}$$

Remember that we have set $n\Delta=1$, so

$$\begin{aligned}\frac{1}{2n\Delta}\sum_{k=1}^{n-1}\sqrt{k\Delta}\left(\mathrm{e}^{-\frac{a^2}{k\Delta}}-\mathrm{e}^{-\frac{a^2}{(k-1)\Delta}}\right) &\geq \frac{1}{2n\Delta}\sum_{k=1}^{n-1}\sqrt{k\Delta}\int_{a^2/k\Delta}^{a^2/(k-1)\Delta}\mathrm{e}^{-\theta}d\theta\\ &= \frac{1}{2n\Delta}\sum_{k=1}^{n-1}\sqrt{k\Delta}\int_{(k-1)\Delta}^{k\Delta}\frac{a^2}{u^2}\mathrm{e}^{-a^2/u}du\\ &\geq \frac{1}{2n\Delta}\sum_{k=1}^{n-1}\int_{(k-1)\Delta}^{k\Delta}\frac{a^2}{u^{3/2}}\mathrm{e}^{-a^2/u}du,\end{aligned}$$

then, since $\sqrt{k\Delta}\geq\sqrt{u}$ if $u\in[(k-1)\Delta,k\Delta]$ finishing the computation of the integral leads to

$$\frac{1}{2n\Delta}\sum_{k=1}^{n-1}\sqrt{k\Delta}\left(\mathrm{e}^{-\frac{a^2}{k\Delta}}-\mathrm{e}^{-\frac{a^2}{(k-1)\Delta}}\right)=\frac{\sqrt{\pi}}{2}a\mathrm{erfc}(a).$$

Finally as we assume that $\frac{1}{2}\leq a$ and $|x_0-y|\leq 1$, then $\frac{1}{2}\leq a\leq\sqrt{c}$ and $\frac{\sqrt{\pi}}{2}a\mathrm{erfc}(a)\geq\frac{\sqrt{\pi}}{4}\mathrm{erfc}(\sqrt{c})>0$ which ends this part of the proof.

A lower bound for $0<a\leq\dfrac{1}{2}$.

Notice that

$$\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp(-a^2/k\Delta)\geq\frac{1}{n}\sum_{k=\lceil n/2\rceil}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp(-a^2/k\Delta).$$

Since $a\leq\frac{1}{2}$ we have $2a^2\leq\frac{1}{2}$ and $k\Delta\in(2a^2,1]$ for all $k\in\{\lceil n/2\rceil,\dots,n-1\}$. Besides, p is strictly decreasing on $(2a^2,1]$, using the fact that $n\geq 4$, and $n\Delta=1$, we get

$$\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp(-a^2/k\Delta)\geq\frac{n-1-\lceil n/2\rceil}{n\sqrt{(n-1)\Delta}}\exp\left(-a^2/(n-1)\Delta\right)\geq\frac{\exp\left(\left(-2a^2\right)\right)}{4}\geq\frac{\mathrm{e}^{-1/2}}{2}.$$

To conclude, we have obtained that for all $n \geq 4$

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n-1} p(k\Delta, x_0, y) &\geq K\mathrm{e}^{-c|x_0|^2}\frac{1}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp(-c|x_0-y|^2/k\Delta)\\ &\geq K\mathrm{e}^{-c|x_0|^2}\left(\frac{\sqrt{\pi}}{4}\mathrm{erfc}(\sqrt{c})\wedge\frac{\mathrm{e}^{-1/2}}{2}\wedge 1\right)\end{aligned}$$

which is non negative.

Set $\tilde{\tilde{\pi}}_0 := K\mathrm{e}^{-c|x_0|^2}\left(\frac{\sqrt{\pi}}{4}\mathrm{erfc}(\sqrt{c})\wedge\frac{\mathrm{e}^{-1/2}}{2}\wedge 1\right)$, the common lower bound follows by setting $\pi_0 := \tilde{\tilde{\pi}}_0 \wedge \tilde{\pi}_0 > 0$. Moreover, for a square integrable h such that $\mathrm{supp}(h) \subseteq [0,1]$, we get

$$\pi_0\|h\|^2 \leq \int_0^1 h^2(y)\left(\frac{1}{n}\sum_{k=1}^{n-1} p\left(k\Delta, x_0, y\right)\right)dy \leq \frac{1}{n}\sum_{k=1}^{n-1}\mathbb{E}h^2(X_{k\Delta}) \leq \|h\|_{n,b}^2$$

which proves our lemma. □

Proof of Proposition 4.4. The following lemma is a result on spline approximation.

Lemma D.2. *For $\tilde{b} \in \Sigma(\beta, R)$, for N large enough, it comes*

$$\inf_{h\in\mathcal{S}_{K_N,L_N,M}}\|h-\widetilde{b}\|_{n,b}^2 \leq C\left(\frac{M+1}{K_N}\right)^{2\beta},$$

where C is a constant which depends only on M and R.

Proof. Let $l = \lfloor\beta\rfloor$ denote the greatest integer strictly less than β. A direct application of Theorem 14.3 and 14.4 and Problem 14.3 in Györfi *et al.* (2006) shows that there exists $\tilde{h} = \sum_{i=-M}^{K_N-1} a_i B_{i,M,\mathbf{u}}$ such that

$$|a_i| \leq C\|\tilde{b}\|_\infty \text{ and } |\tilde{h}(x)-\tilde{b}(x)| \leq C\frac{L_b}{l!}\left(\frac{M+1}{K_N}\right)^{\beta},$$

where $C > 0$ depends only on M. Since for N large enough we have $\|\tilde{b}\|_\infty \leq \frac{\sqrt{L_N}}{C}$, we note that $\tilde{h} \in \mathcal{S}_{K_N,L_N,M}$. Let us denote the density f_n

$$f_n(y) := \frac{1}{n-1}\sum_{k=1}^{n-1} p(k\Delta, x_0, y).$$

Therefore, we deduce that

$$\begin{aligned}\inf_{h\in\mathcal{S}_{K_N,L_N,M}}\|h-\widetilde{b}\|_{n,b}^2 &\leq \frac{1}{n}\left(\tilde{h}(x_0)-\tilde{b}(x_0)\right)^2 + \frac{n-1}{n}\int_0^1\left(\tilde{h}(x)-\tilde{b}(x)\right)^2 f_n(x)\mathrm{d}x\\ &\leq \left(C\frac{L}{l!}\right)^2\left(\frac{M+1}{K_N}\right)^{2\beta}.\end{aligned}$$

□

Now, we go back to the proof of Proposition 4.4. From Equation (C.2), we have that for each $h \in \mathcal{S}_{K_N,L_N,M}$

$$\left(1-\frac{2}{a}\right)\mathbb{E}\left[\|\widehat{b}_{N,n}-\widetilde{b}\|_{N,n}^2\right] \leq \left(1+\frac{2}{a}\right)\|h-\widetilde{b}\|_{n,b} + 2\mathbb{E}\left[\nu_{N,n}(\widehat{b}_{N,n}-h)\right] + aC\Delta,$$

for $a > 0$ and where for any $g \in \mathcal{S}_{K_N,L_N,M}$, we used the notation

$$\nu_{N,n}(g) = \frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} g(X_{k\Delta}^{(j)})\Sigma_{k,j}$$

(see (C.1)). Moreover, by linearity it comes for $d > 0$,

$$\begin{aligned} 2\nu_{N,n}(\widehat{b}_{N,n} - h) &= 2\|\widehat{b}_{N,n} - h\|_{n,b}\nu_{N,n}\left(\frac{\widehat{b}_{N,n} - h}{\|\widehat{b}_{N,n} - h\|_{n,b}}\right) \\ &\leq \frac{1}{d}\|\widehat{b}_{N,n} - h\|_{n,b}^2 + d \sup_{\{g \in \mathcal{S}_{K_N,M,\mathbf{u}};\ \|g\|_{n,b}=1\}} \nu_{N,n}^2(g). \end{aligned}$$

We have $\|\widehat{b}_{N,n} - h\|_{N,n}^2 \leq 2\left(\|\widehat{b}_{N,n} - b\|_{N,n}^2 + \|h - b\|_{N,n}^2\right)$ and on $\Omega_{N,n}$, we have $\|\widehat{b}_{N,n} - h\|_{n,b}^2 \leq 2\|\widehat{b}_{N,n} - h\|_{N,n}^2$.

Therefore, we get

$$\begin{aligned} \left(1 - \frac{2}{a} - \frac{4}{d}\right)\|\widehat{b}_{N,n} - \widetilde{b}\|_{N,n}^2 &\leq \left(1 + \frac{2}{a} + \frac{4}{d}\right)\|h - \widetilde{b}\|_{N,n}^2 \\ &\quad + d \sup_{\{g \in \mathcal{S}_{K,M,\mathbf{u}};\ \|g\|_{n,b}=1\}} \nu_{N,n}^2(g) + aC\Delta. \end{aligned} \tag{D.5}$$

Using the fact that $\|g\|_{n,b} = 1$ implies $\|g\| \leq \frac{1}{\sqrt{\pi_0}}$ according to Lemma 4.3, we obtain

$$\mathbb{E}\left[\sup_{\{g \in \mathcal{S}_{K,M,\mathbf{u}};\ \|g\|_{n,b}=1\}} \nu_{N,n}^2(h)\right] \leq \frac{1}{\pi_0}\mathbb{E}\left[\sup_{\{g \in \mathcal{S}_{K,M,\mathbf{u}};\ \|g\|=1\}} \nu_{N,n}^2(g)\right]$$

Applying Lemma 2.6, we have that $\|\mathbf{a}\|_2^2 \leq C_1^{-1}K_N\|h\|^2$ (where $\mathbf{a}$ denotes the coefficient corresponding to $h \in \mathcal{S}_{K,M,\mathbf{u}}$).

Hence, we deduce from the Cauchy-Schwarz Inequality that, for $h \in \mathcal{S}_{K,M,\mathbf{u}}$ such that $\|h\| = 1$,

$$\nu_{N,n}^2(h) \leq CK_N \sum_{i=-M}^{K_N-1} \nu_{N,n}^2(B_{i,M,\mathbf{u}}).$$

Finally, following Equation (C.3), it comes

$$\mathbb{E}\left[\sup_{\{h \in \mathcal{S}_{K,M,\mathbf{u}};\ \|h\|=1\}} \nu_{N,n}^2(h)\right] \leq C\frac{K_N}{N}.$$

Therefore, Equation (D.5) with $a = d = 8$ yields

$$\mathbb{E}\left[\|\widehat{b}_{N,n} - \tilde{b}\|_{N,n}^2 \mathbb{1}_{\Omega_{N,n}}\right] \leq 7 \inf_{h \in \mathcal{S}_{K_N,L_N,M}} \|h - \widetilde{b}\|_{n,b}^2 + C\left(\frac{K_N}{N} + \Delta\right). \tag{D.6}$$

Then, since $\|\widehat{b}_{N,n}^{L_N}\|_\infty \leq \sqrt{L_N}$ and for N large enough $\|\widetilde{b}\|_\infty \leq \sqrt{L_N}$, we have (for N large enough)

$$\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|_{N,n}^2 \leq \|\widehat{b}_{N,n} - \widetilde{b}\|_{N,n}^2 \text{ and } \|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|_{N,n}^2 \leq 4L_N.$$

Therefore Equation (D.6) yields

$$\begin{aligned} \mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_{N,n}^2\right] &\leq \mathbb{E}\left[\|\widehat{b}_{N,n} - \tilde{b}\|_{N,n}^2 \mathbb{1}_{\Omega_{N,n}}\right] + \mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_{N,n}^2 \mathbb{1}_{\Omega_{N,n}^c}\right] \\ &\leq 7 \inf_{h \in \mathcal{S}_{K_N,L_N,M}} \|h - b\|_{n,b}^2 + C\left(\frac{K_N}{N} + \Delta\right) + 4L_N\mathbb{P}\left(\Omega_{N,n,K}^c\right). \end{aligned}$$

The proof is then completed by applying Lemma D.2 and Lemma D.1. □

Proof of Proposition 4.6. Let h denote some measurable function s.t. $\|h\|_\infty < \infty$. We have

$$\begin{aligned} \|h\|_b^2 - \|h\|_{n,b}^2 &= \mathbb{E}\left[\frac{1}{T}\int_0^T \left(h^2(X_s) - h^2(X_{\eta(s)})\right) ds\right] \\ &= \mathbb{E}\left[\frac{1}{T}\int_\Delta^T \left(h^2(X_s) - h^2(X_{\eta(s)})\right) ds\right] + \mathbb{E}\left[\frac{1}{T}\int_0^\Delta h^2(X_s) - h^2(x_0) ds\right]. \end{aligned}$$

From the above equality, we get

$$\|h\|_b^2 - \|h\|_{n,b}^2 = \frac{1}{T}\int_\Delta^T ds \int_{\mathbb{R}} h^2(y)\left(p(s,x_0,y) - p\left(\eta(s),x_0,y\right)\right)dy \\ + \frac{1}{T}\int_0^\Delta ds \int_{\mathbb{R}} \left(h^2(y) - h^2(x_0)\right) p(s,x_0,y)dy.$$

Using the estimates of the transition probability density stated in Konakov & Menozzi (2017) (Proposition 3.1 p.16), we find that there exists a constant $C \geq 1$ and $c \in (0,1]$ such that the second term in the RHS of the previous equality bounds as

$$\int_0^\Delta ds \int_{\mathbb{R}} \left(h^2(y) - h^2(x_0)\right) p(s,x_0,y)dy \\ \leq \int_0^\Delta ds \int_{\mathbb{R}} \left(h^2(y) - h^2(x_0)\right) \frac{C}{(2\pi\, cs)^{1/2}} \mathrm{e}^{-|x_0-y|^2/2cs} dy \leq C\|h\|_\infty^2 \Delta.$$

Let us now turn to the first term in the RHS. We have

$$I := \left| \int_\Delta^T ds \int_{\mathbb{R}} h^2(y)\left(p(s,x_0,y) - p\left(\eta(s),x_0,y\right)\right)dy \right| \leq \\ \int_\Delta^T ds \int_{\mathbb{R}} dy\, h^2(y) \int_{\eta(s)}^s du \left| \frac{\partial}{\partial u} p\left(u,x_0,y\right) \right|.$$

Using the Fubini-Tonelli theorem, the Kolmogorov-Backward equation and the Gaussian estimates given in Konakov & Menozzi (2017) (Proposition 3.1 p.16), we find that there exists a constant $C \geq 1$ and $c \in (0,1]$ with g_c the Gaussian density function ($g_c(u,x_0,y) := e^{-c(x_0-y)^2/(2u)}/\sqrt{2\pi u/c}$), such that

$$\begin{aligned} I &\leq C\|h\|_\infty^2 \int_\Delta^T ds \int_{\eta(s)}^s \frac{du}{u} \left(\int_{\mathbb{R}} g_c(u,x_0,y)\, dy \right) \\ &\leq C\|h\|_\infty^2 \int_\Delta^T \log(s/\eta(s)) ds \\ &\leq C\|h\|_\infty^2 \sum_{k=1}^{n-1} \int_{k\Delta}^{(k+1)\Delta} \log(s/k\Delta)\, ds. \end{aligned}$$

Therefore, we deduce

$$\begin{aligned} I &\leq C\|h\|_\infty^2\, \Delta \sum_{k=1}^{n-1} \log\left(1+\frac{1}{k}\right) \\ &\leq C\|h\|_\infty^2 \frac{T}{n} \sum_{k=1}^{n-1} \frac{1}{k} \leq C\,T\,\|h\|_\infty^2 \left(\frac{\log(n)+1}{n}\right). \end{aligned}$$

Consequently, there exists a constant $C \geq 1$ such that for any $h \in L^\infty(\mathbb{R})$:

$$\left| \|h\|_b^2 - \|h\|_{n,b}^2 \right| \leq C\|h\|_\infty^2 (-\Delta\, \log \Delta).$$

□

Proof of Theorem 4.5. For N large enough $\|\widetilde{b}\|_\infty \leq \sqrt{L_N}$, and by definition $\|\widehat{b}_{N,n}^{L_N}\|_\infty \leq \sqrt{L_N}$ (see Equation (4.1)). Besides, $\hat{b}_{N,n}^{L_N}$ is Lipschitz with a lipschitz constant bounded by $\sqrt{(K_N+M)L_N}K_N$ (see Equation C.8), then according to Proposition 2.3 we have :

$$\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_b^2\right] \leq \mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_{n,b}^2\right] + CK_N^{3/2} L_N \sqrt{\Delta}$$

Therefore, it remains to control the first term in the r.h.s. We start with the following decomposition

$$\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_{n,b}^2 = \|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_{n,b}^2 - 2\|\widehat{b}_{N,n}^{L_N} - b\|_{N,n}^2 + 2\|\widehat{b}_{N,n}^{L_N} - b\|_{N,n}^2,$$

which yields, with Proposition 4.4, for N large enough

$$\begin{aligned}\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_{n,b}^2\right] &\leq \mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_{n,b}^2\right] - \mathbb{E}\left[2\|\widehat{b}_{N,n}^{L_N} - \tilde{b}\|_{N,n}^2\right] \\ &\quad + C\left(\left(\frac{M+1}{K_N}\right)^{2\beta} + \frac{(K_N + L_N)}{N} + \Delta\right). \end{aligned} \tag{D.7}$$

The end of the proof follows the same lines as the proof of Theorem 3.3. We observe that

$$\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|_{n,b}^2 - 2\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|_{N,n}^2\right] \leq \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\|h - \widetilde{b}\|_{n,b}^2 - 2\|h - \widetilde{b}\|_{N,n}^2\right)\right],$$

where $\mathcal{H} = \left\{h^{L_N},\ \ h \in \mathcal{S}_{K_N,L_N,M}\right\}$. For $h \in \mathcal{H}$, we define the function g_h as in Equation (C.10), and consider $\mathcal{G} = \{g_h,\ \ h \in \mathcal{H}\}$. Then, we have

$$\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|_{n,b}^2 - 2\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|_{N,n}^2\right] \leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[g(\overline{X})\right] - \frac{2}{N}\sum_{j=1}^{N} g\left(\overline{X}^j\right)\right)\right].$$

Since, for each $h \in \mathcal{H}$, $\|h\|_\infty \leq \sqrt{L_N}$, and $\|\widetilde{b}\|_\infty \leq \sqrt{L_N}$ for N large enough, we deduce that for each $g \in \mathcal{G}$ and N large enough

$$0 \leq g(x) \leq 4L_N.$$

Therefore, for $\varepsilon > 0$ a direct application of Lemma A.2 yields

$$\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|_{n,b}^2 - 2\|\widehat{b}_{N,n}^{L_N} - b\|_{N,n}^2\right] \leq 3\varepsilon + \frac{44L_N \log(\mathcal{N}_\infty(\varepsilon,\mathcal{G}))}{N}. \tag{D.8}$$

Since an ε-net of $\mathcal{S}_{K_N,L_N,M}$ w.r.t $\|.\|_\infty$ is also an ε-net of $\mathcal{H}$ w.r.t $\|.\|_\infty$, we have from Equation (C.13)

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

Finally, setting $\varepsilon = \frac{12(K_N+M)L_N}{N}$ in Equation (D.8) yields

$$\mathbb{E}\left[\|\widehat{b}_{N,n}^{L_N} - \widetilde{b}\|_{n,b}^2 - 2\|\widehat{b}_{N,n}^{L_N} - b\|_{N,n}^2\right] \leq \frac{36(K_N+M)L_N}{N} + \frac{44L_N\log(N)(K_N+M)}{N}. \tag{D.9}$$

The above inequality with Equation (D.7) lead to the result.

□

Proof of Theorem 4.7. The proof follows the scheme of the proof of Theorem 2.8 of Tsybakov (2009) Chapter 2. The main point is to apply Theorem 2.5 which is based on three condition. For sake of clarity we recall the main lines of the proof. For $c_0 >$ and $m \geq 1$,

$$\begin{aligned} m &= \lceil c_0 N^{\frac{1}{2\beta+1}}\rceil,\ h_N = 1/m,\ x_k = \frac{k-1/2}{m} \\ \varphi_k(x) &= Lh_N^\beta K\left(\frac{x - x_k}{h_N}\right),\ k = 1,\ldots,m, \end{aligned}$$

where $K : \mathbb{R} \to [0,+\infty[$ satisfies $K \in \Sigma(\beta,1/2)\bigcap C^\infty(\mathbb{R})$ and $K(u) > 0 \Leftrightarrow u \in (-1/2,1/2)$.

The hypotheses $b_{N,j}$ are taken in the following space:

$$\mathcal{E} = \left\{b_{N,i}(x) = \sum_{k=1}^{m} \omega_k^{(i)}\varphi_k(x),\ \omega \in \{\omega^{(0)},\ldots,\omega^{(D)}\}\right\},$$

where $\{\omega^{(0)}, \dots, \omega^{(D)}\}$ is a subset of $\{0,1\}^m$, such that $\omega^{(0)} = (0, \dots, 0)$ and the Hamming distance ρ satisfies the Varshamov-Gilbert bound:

$$\rho(\omega^{(i)}, \omega^{(k)}) \geq m/8, \ \forall 0 \leq i < k \leq D, \quad D \geq 2^{m/8}.$$

To apply Theorem 2.5, we have to check the following points for a fixed $\alpha \in (0, 1/2)$:

1. $b_{N,i} \in \Sigma(\beta, R), \ j = 0, \dots, M,$
2. $\|b_{N,i} - b_{N,k}\| \geq 2s > 0, \ 0 \leq i < k \leq D$, with $s = s_N \asymp N^{-\beta/(2\beta+1)}$,
3. $\dfrac{1}{D}\sum_{i=1}^{D} K(P_i^{\otimes N}, P_0^{\otimes N}) \leq \alpha \log(D), \ i = 1, \dots, D.$

The first two conditions are satisfied by construction. Let us deal with the third point: the control of the Kullback divergence between two hypothesis. We denote $\mathbb{P}_i$ the probability measure under which $(X_t)_{t\geq 0}$ is solution of $dX_t = b_{N,i}(X_t)dt + \sigma(X_t)d\widetilde{W}_t$ where $\widetilde{W}$ is a Brownian motion under $\mathbb{P}_j$. Finally we define $P_i := \mathbb{P}_{i|\mathcal{F}_T^X}$. Since b satisfies Assumptions 2.1 and 4.1, the Novikov condition is satisfied

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

Then Girsanov's Theorem (see Revuz & Yor (1999)) ensures that the probability measure $\mathbb{P}_0$ is absolutely continuous w.r.t. $\mathbb{P}_i$ and that

$$\frac{dP_i}{dP_0} = \exp\left(\int_0^T \frac{b_{N,i}}{\sigma^2}(X_s)dX_s - \frac{1}{2}\int_0^T \frac{b_{N,i}^2}{\sigma^2}(X_s)ds\right).$$

Hence,

$$\begin{aligned} K(P_i, P_0) &= \int \log\left(\frac{dP_i}{dP_0}\right) dP_i = \mathbb{E}\left[\int_0^T \frac{b_{N,i}}{\sigma}(X_s)dW_s + \frac{1}{2}\int_0^T \frac{b_{N,i}^2}{\sigma^2}(X_s)ds\right] \\ &= \frac{1}{2}\mathbb{E}\left[\int_0^T \frac{b_{N,i}^2}{\sigma^2}(X_s)ds\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} K(P_i^{\otimes N}, P_0^{\otimes N}) &= \frac{1}{2}\sum_{j=1}^{N}\mathbb{E}\left[\int_0^T \frac{(\sum_{k=1}^m \omega_k^{(i)}\varphi_k)^2}{\sigma^2}(X_s^{(j)})ds\right] \\ &\leq \frac{1}{2\sigma_0^2}\sum_{i=1}^{N}\mathbb{E}\left[\int_0^T \sum_{k=1}^m \omega_k^{(i)}\varphi_k^2(X_s^{(j)})\mathbb{1}_{\{X_s^{(j)} \in]\frac{(k-1)}{m}, \frac{k}{m}]\}}ds\right] \\ &\leq \frac{L^2 h_N^{2\beta}\|K\|_\infty^2}{2\sigma_0^2}\mathbb{E}\left[\int_0^T \sum_{i=1}^{N}\sum_{k=1}^{m} \mathbb{1}_{\{X_s^{(j)} \in]\frac{(k-1)}{m}, \frac{k}{m}]\}}ds\right] \\ &= \frac{L^2 h_N^{2\beta}\|K\|_\infty^2 TN}{2\sigma_0^2} \leq \frac{L^2\|K\|_\infty^2 T c_0^{-(2\beta+1)} m}{2\sigma_0^2}. \end{aligned}$$

Then, with $m \leq 8\log(D)/\log(2)$ choosing,

$$c_0 = \left(\frac{8TL^2\|K\|_\infty^2}{\alpha\log(2)}\right)^{\frac{1}{2\beta+1}}$$

it comes that $K(P_j^{\otimes N}; P_0^{\otimes N}) < \alpha\log(D)$, which concludes the proof.

□

Proof of Theorem 4.8. Let us remind the reader that in this section, we assume that $\mathcal{K} = \{2^p,\ p = 0, \dots, p_{\max}\}$ and $L_N = \log(N)$.

Lemma D.3. *For any positive numbers $\varepsilon,\ v$, we have*

$$\mathbb{P}\left(\sum_{j=1}^{N}\sum_{k=0}^{n-1} t(X_{k\Delta}^{(j)})\Sigma_{k\Delta}^{(j)} \geq Nn\varepsilon,\ \|t\|_{N,n}^2 \leq v^2\right) \leq \exp\left(-\frac{Nn\Delta\varepsilon^2}{2v^2\sigma_1^2}\right).$$

Proof of Lemma D.3. The proof is based on the martingale $M_s = \int_0^s \sum_{j=1}^N H_u^{(j)} dW_u^{(j)}$, with $H_u^{(j)} = \sum_{k=0}^{n-1} \mathbb{1}_{[k\Delta,(k+1)\Delta[}(u) t(X_{k\Delta}^{(j)})\sigma(X_u^{(j)})$ and can be easily adapted from Lemma 2 in Comte *et al.* (2007). □

We go back to the proof of Theorem 4.8. Since, for each $K \in \mathcal{K}_{N,n}$, we have

$$\gamma_{N,n}(\widehat{b}_{\widehat{K}}) + \text{pen}(\widehat{K}) \leq \gamma_{N,n}(\widehat{b}_K) + \text{pen}(K),$$

and as in inequality (D.5) with $a = d = 8$, we get for each $h \in \mathcal{S}_{K_N,L_N,M}$ on $\Omega_{N,n,K_{\max}}$

$$\begin{aligned}\left\|\widehat{b}_{\widehat{K}} - \widetilde{b}\right\|_{N,n}^2 &\leq 7\left\|h - \widetilde{b}\right\|_{N,n}^2 + 8 \sup_{\{h \in \mathcal{S}_{K,M,\mathbf{u}} + \mathcal{S}_{\widehat{K},M,\mathbf{u}},\ \|h\|_{n,b}=1\}} \nu_{N,n}^2(h) \\ &\quad + C\Delta + 4\left(\text{pen}(K) - \text{pen}(\widehat{K})\right).\end{aligned}$$

Let us denote:

$$G_K(K') := \sup_{\{h \in \mathcal{S}_{K,M,\mathbf{u}} + \mathcal{S}_{K',M,\mathbf{u}},\ \|h\|_{n,b}=1\}} |\nu_{N,n}(h)|.$$

Therefore, as for N we obtain,

$$\begin{aligned}\left\|\widehat{b}_{\widehat{K}} - \widetilde{b}\right\|_{N,n}^2 &\leq \left\{7\left\|h - \widetilde{b}\right\|_{N,n}^2 + 8G_K^2(\widehat{K}) + C\Delta + 4\left(\text{pen}(K) - \text{pen}(\widehat{K})\right)\right\}\mathbb{1}_{\Omega_{N,n,K_{\max}}} \\ &\quad + 2L_N\mathbb{1}_{\Omega_{N,n,K_{\max}}^c}.\end{aligned}$$

Now, for all $h \in \mathcal{S}_{K,L_N,M}$, let us define,

$$E_{N,n}(h) := 2\left\|h - \widetilde{b}\right\|_{N,n}^2 - \left\|h - \widetilde{b}\right\|_{n,b}^2.$$

For $\mathcal{H} = \left\{h^{L_N},\ \ h \in \mathcal{S}_{K,L_N,M}\right\}$, it comes,

$$\begin{aligned}\mathbb{E}\left[-E_{N,n}(\widehat{b}_{\widehat{K}}^{L_N}) - \text{pen}(\widehat{K})\right] &\leq \sum_{K\in\mathcal{K}} \mathbb{E}\left[-E_{N,n}(\widehat{b}_K^{L_N}) - \text{pen}(K)\right] \\ &\leq \sum_{K\in\mathcal{K}} \mathbb{E}\left[\sup_{\mathcal{H}}(-E_{N,n}(h))\right] - \text{pen}(K) \\ &\leq 0,\end{aligned}$$

where, in the last inequality we use that $\text{pen}(K) \geq 44(K+M)\log(N)/N$. Indeed, as in proof of Theorem 4.5, Equation (D.9),

$$\mathbb{E}\left[\sup_{\mathcal{H}}(-E_{N,n}(h))\right] \leq 44(K+M)\frac{\log(N)}{N}.$$

Finally,

$$\|\widehat{b}_{\widehat{K}}^{L_N} - \widetilde{b}\|_{n,b}^2 \leq \text{pen}(\widehat{K}) + 2\left\|\widehat{b}_{\widehat{K}} - \widetilde{b}\right\|_{N,n}.$$

Therefore, since for N large enough $\|\tilde{b}\|_\infty \leq \frac{\sqrt{L_N}}{C}$ and $|\widehat{b}_K| \leq \sqrt{L_N}$ by definition, we obtain

$$\mathbb{E}\left[\left\|\widehat{b}_{\widehat{K}} - \widetilde{b}\right\|_{n,b}^2\right] \leq \left(14\left\|h - \widetilde{b}\right\|_{n,b}^2 + \text{pen}(K)\right) + 16\mathbb{E}\left[G_K^2(\widehat{K})\mathbb{1}_{\Omega_{N,n,K_{\max}}}\right] + 7\mathbb{E}\left[\text{pen}(K) - \text{pen}(\widehat{K})\right] + \frac{CL_N}{N}.$$

Finally, choosing $16p(K,K') \leq 7(\text{pen}(K) + \text{pen}(K'))$ and since

$$G_K^2(\widehat{K})\mathbb{1}_{\Omega_{N,n,K_{\max}}} \leq \sum_{K' \in \mathcal{K}} \left(G_K^2(K') - p(K,K')\right)_+ \mathbb{1}_{\Omega_{N,n,K_{\max}}} + p(K,K'),$$

it comes

$$\mathbb{E}\left[\left\|\widehat{b}_{\widehat{K}} - \widetilde{b}\right\|_{n,b}^2\right] \leq 15\left(\left\|h - \widetilde{b}\right\|_{n,b}^2 + \text{pen}(K)\right) + 16\mathbb{E}\left[\sum_{K' \in \mathcal{K}} \left(G_K^2(K') - p(K,K')\right)_+ \mathbb{1}_{\Omega_{N,n,K_{\max}}}\right] + \frac{CL_N}{N}.$$

Therefore, Lemma D.3 used together with a chaining argument detailed in Baraud *et al.* (2001), gives

$$\mathbf{E}\left[\left(G_K^2(K') - p(K,K')\right)_+ \mathbb{1}_{\Omega_{N,n,K_{\max}}}\right] \leq c\sigma_1^2 \exp\left(-(K'+M)\right)/N$$

with $p(K,K') = \kappa_1\sigma_1^2\frac{K+K'+2M}{N}$, $\kappa_1 > 0$ a constant. This implies that,

$$\begin{aligned}\sum_{K' \in \mathcal{K}} \left(G_K^2(K') - p(K,K')\right)_+ \mathbb{1}_{\Omega_{N,n,K_{\max}}} &\leq \frac{c\sigma_1^2}{N}\sum_{K' \in \mathcal{K}} \exp((-(K'+M)) \\ &\leq \frac{C}{N}\sum_{k \geq 1}\exp\left(-k\right) \leq \frac{C}{N}.\end{aligned}$$

Finally, we must have:

$$\text{pen}(K) \geq \max\left(44\frac{\log(N)(K+M)}{N}; \kappa_1\sigma_1^2\frac{(K+M)}{N}\right)$$

and for N large enough, the first term $\text{pen}(K) \geq 44\frac{\log(N)(K+M)}{N}$.

□