DRIFT ESTIMATION ON NON COMPACT SUPPORT FOR DIFFUSION MODELS

F. COMTE AND V. GENON-CATALOT

Abstract. We study non parametric drift estimation for an ergodic diffusion process from discrete observations. The drift is estimated on a set A using an approximate regression equation by a least squares contrast, minimized over finite dimensional subspaces of $\mathbb{L}^2(A, dx)$. The novelty is that the set A is non compact and the diffusion coefficient unbounded. Risk bounds of a $\mathbb{L}^2$-risk are provided where new variance terms are exhibited. A data-driven selection procedure is proposed where the dimension of the projection space is chosen within a random set contrary to usual selection procedures.
July 30, 2020

1. Introduction

Non parametric drift estimation for ergodic diffusion processes whether from continuous or from discrete observations is a widely investigated subject and to some extent thoroughly known. Nevertheless, gaps in this field remain and the aim of this paper is to fill some of these. Two approaches are classically investigated. First, one can estimate the drift function by kernel. This is done, for instance, by Kutoyants (2004), Dalalyan (2005) or more recently by Aeckerle-Willems and Strauch (2018), Strauch (2015, 2016), Nickl and Ray (2020). With this method, the estimation of the drift is not direct: one has to estimate the product of the drift by the invariant density of the process and then divide the resulting estimator by an estimator of the invariant density. There are no support constraints with kernels but authors rather study pointwise risks than global $\mathbb{L}^2$-risks thus getting around some difficulties. The second approach is based on least-squares and sieves. An estimation set A is fixed and a collection of finite-dimensional subspaces of $\mathbb{L}^2(A, dx)$ is chosen. This leads by minimization of a least-squares contrast on each subspace, to a collection of estimators of the drift restricted to the estimation set, indexed by the dimension of the projection space. The estimators of the drift are defined directly but here, there is a support constraint as the drift is not estimated outside the estimation set. This method was initiated by Birgé and Massart (1998), Barron *et al.* (1999), Baraud (2002), for regression models, by Hoffmann (1999) and Comte *et al.* (2007) for diffusions. The estimation set A is assumed to be compact and the drift function square integrable on this set. This is a drawback of this approach. Moreover, on a compact set, the diffusion coefficient which is a continuous function, is obviously bounded.
In this paper, we investigate the sieves method approach relaxing the compactness constraint on the estimation set. The estimation set may be equal to $\mathbb{R}^+$ or the whole real line.

Key words and phrases. Discrete time observation. Mean square estimator. Model selection. Nonparametric drift estimation. Stochastic differential equations.

Université de Paris, Laboratoire MAP5, email: fabienne.comte@parisdescartes.fr, valentine.genon-catalot@parisdescartes.fr.

More precisely, we consider discrete observations with sampling interval Δ, $(X_{i\Delta})_{1\le i\le n+1}$, of the one-dimensional diffusion process $(X_t)_{t\ge 0}$, solution of

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = \eta, \tag{1}$$

where (W_t) is a Wiener process and η is a random variable independent of (W_t). Assume that model (1) is in stationary regime with marginal distribution $\pi(x)dx$ and exponentially β-mixing. The asymptotic framework is that, as n tends to infinity, the sampling interval $\Delta = \Delta_n$ tends to 0 while the total length time of observations $n\Delta_n$ tends to infinity (high frequency data). The functions b, σ are unknown and we are concerned here with nonparametric estimation of b. We only assume moment assumptions for $b(X_0), \sigma(X_0)$ and in particular, these two functions need not be bounded.

Drift estimation for diffusion models in high frequency framework share some features with heteroscedastic regression function estimation. Setting

$$Y_{i\Delta} = \frac{X_{(i+1)\Delta} - X_{i\Delta}}{\Delta}, \qquad Z_{i\Delta} = \frac{1}{\Delta}\sigma(X_{i\Delta})\left(W_{(i+1)\Delta} - W_{i\Delta}\right) \tag{2}$$

yields $Y_{i\Delta} = b(X_{i\Delta}) + Z_{i\Delta} +$ residual term, where the expression of the residual term is obtained using equation (1). This approximate regression equation is used to build a least-squares contrast leading by minimization to a collection of estimators on an unrestricted estimation set. Because the estimation set is unrestricted, the method quite differs from the one in Comte *et al.* (2007). Estimators are expressed using non compactly supported bases and the dimensions of the projection spaces are subject to a constraint which does not exist in the compact case. A similar constraint was introduced by Cohen *et al.* (2013) in the homoscedastic regression framework to obtain accuracy of regression function estimators. We study $\mathbb{L}^2$-risk bounds for our drift estimators on a fixed projection space and then propose a data-driven choice of the dimension by means of a random penalty. It is noteworthy that decomposition (2) is not the same as in Comte *et al.* (2007) (see Remark 3.1). This allows us to obtain novelties especially in the variance terms of the $\mathbb{L}^2$-risk bounds and in the penalty terms of the model selection procedure. Indeed, the penalty involves no longer upper or lower bounds of unknown functions, but only terms which can be more easily estimated.

Section 2 gives the framework and assumptions on the model. Section 3 concerns the estimation of the drift function and is divided in several subsections. First, the approximate regression model is precised. Then, the projection estimators of $b\mathbf{1}_A$ are built using a collection of finite-dimensional subspaces of $\mathbb{L}^2(A, dx)$ where the estimation set A is a general Borel subset of $\mathbb{R}$. Risk bounds are obtained based on the expectation of an empirical norm and of the $\mathbb{L}^2(A, \pi(x)dx)$- norm. The variance term is completely new and differs from the one obtained in Comte *et al.* (2007). Rates of convergence are discussed showing that, in the case where σ is a bounded function, for an appropriate choice of the projection dimension, our estimator is optimal. Comparison is done with respect to existing results in the litterature. A data-driven procedure to select among the collection of estimators is proposed. In the selection procedure, the dimension is selected via a penalization criterion within a random set which is non standard in these methods and induces difficulties in proofs. Section 4 is devoted to a simulation study. The examples of non compactly supported bases that we propose are the Laguerre bases for $A = \mathbb{R}^+$ and the Hermite bases for $A = \mathbb{R}$ which have been used in various contexts for nonparametric estimation (see Comte and Genon-Catalot, 2015, Belomestny *et al.*, 2016, Comte and Genon-Catalot, 2020a-b, Comte *et al.*, 2017, Mabon, 2017). Some concluding remarks are given in Section 5. Proofs are given in Section 6.

2. Assumptions on the diffusion model

Consider discrete observations with sampling interval Δ, $(X_{i\Delta})_{1\leq i\leq n+1}$, of the diffusion process $(X_t)_{t\geq 0}$, solution of (1) where (W_t) is a Wiener process and η is a random variable independent of (W_t). The drift function $b(.)$ is unknown and our aim is to propose non-parametric estimators for it, relying on the sample $(X_{i\Delta})_{1\leq i\leq n+1}$. The diffusion coefficient $\sigma^2(.)$ is also unknown and our estimation procedure will not depend on it. The asymptotic setting is: $\Delta = \Delta_n$ tends to 0 and $n\Delta_n$ tends to infinity as n tends to infinity. Without loss of generality, we assume $\log(n\Delta_n) \geq 1$. To simplify notations, we only write Δ without the subscript n. However, when speaking of constants, we mean quantities that depend neither on n nor on Δ. We consider the following assumptions.

(A1) $b, \sigma \in C^1(\mathbb{R})$ and there exists $L \geq 0$, such that, for all $x \in \mathbb{R}, |b'(x)| + |\sigma'(x)| \leq L$.

(A2) The scale density

$$s(x) = \exp\left\{-2\int_0^x \frac{b(u)}{\sigma^2(u)}du\right\}$$

satisfies $\int_{-\infty} s(x)dx = +\infty = \int^{+\infty} s(x)dx$ and the speed density $m(x) = 1/(\sigma^2(x)s(x))$ satisfies $\int_{-\infty}^{+\infty} m(x)dx = M < +\infty$.

(A3) $X_0 = \eta$ has distribution $\pi(x)dx$ given by $\pi(x) = M^{-1}m(x)$.

Under **(A1)**, Equation (1) has a unique strong solution adapted to the filtration $(\mathcal{F}_t = \sigma(\eta, W_s, s \leq t), t \geq 0)$. The functions b, σ have linear growth:

$$\exists K, \forall x \in \mathbb{R}, |b(x)| + |\sigma(x)| \leq K(1 + |x|). \tag{3}$$

Under the additional assumption **(A2)**, Model (1) admits a unique invariant probability $\pi(x)dx$. And under **(A3)**, (X_t) is strictly stationary and ergodic.

(A4) (X_t) is geometrically β-mixing: there exist constants $K > 0$ and $\theta > 0$ such that:

$$\beta_X(t) \leq Ke^{-\theta t}, \tag{4}$$

where $\beta_X(t) = \int_{-\infty}^{+\infty} \pi(x)dx\|P_t(x, dx') - \pi(x')dx'\|_{TV}$ denotes the β-mixing coefficient of (X_t). The norm $\|.\|_{TV}$ is the total variation norm and P_t denotes the transition probability.

(A5) $\|\pi\|_\infty < +\infty$ ($\|\pi\|_\infty = \sup_{x\in A} \pi(x)$ denotes the sup norm on A).

In Veretennikov (1988) or Pardoux and Veretennikov (2001), sufficient conditions ensuring **(A4)** may be found. Assumption **(A5)** is only used in Section 3.6. The following result is used below (see Proposition A, in Gloter (2000)):

Proposition 2.1. *Assume* **(A1)**-**(A3)** *and let* $f : \mathbb{R} \to \mathbb{R}$ *be* C^1 *and such that there exists a constant* $\gamma \geq 0$ *such that, for all* $x \in \mathbb{R}$, $|f'(x)| \leq c(1 + |x|^\gamma)$. *If, for* k *an interger,* $\mathbb{E}|\eta|^{k(1+\gamma)} < +\infty$,

$$\mathbb{E}(\sup_{s\in[t,t+h]} |f(X_s) - f(X_t)|^k) \leq ch^{k/2}(1 + \mathbb{E}|\eta|^{k(1+\gamma)}). \tag{5}$$

In particular, Proposition 2.1 applies for b and σ with $\gamma = 0$, under **(A1)**.

3. Drift estimation

In what follows, for a function h, we denote $h_A := h\mathbf{1}_A$, $\|.\|$ denotes the $\mathbb{L}^2$-norm, $\|.\|_\pi$ the norm in $\mathbb{L}^2(\pi(x)dx)$ and $\|.\|_{2,m}$ the Euclidian norm in $\mathbb{R}^m$.

3.1. **Approximate regression model for the drift.** Consider $Y_{i\Delta}, Z_{i\Delta}$ defined in (2) and set $R_{i\Delta} := R_i^{(1)} + R_i^{(2)}$,

$$R_i^{(1)} = \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}(\sigma(X_s) - \sigma(X_{i\Delta}))dW_s, \quad R_i^{(2)} = \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}(b(X_s) - b(X_{i\Delta}))ds.$$

The process (1) satisfies the following relation:

$$Y_{i\Delta} = b(X_{i\Delta}) + Z_{i\Delta} + R_{i\Delta}. \tag{6}$$

Equation (6) is close to an heteroscedastic regression equation where $Z_{i\Delta}$ plays the role of the noise and $R_{i\Delta}$ is an additional residual term to take into account. This leads us to apply part of the tools proposed for regression function estimation on non compact support in Comte and Genon-Catalot (2020a-b) to the present diffusion context.

Remark 3.1. Decomposition (6) is slightly different from the one used in Comte *et al.* (2007) which was $Y_{i\Delta} = b(X_{i\Delta}) + \tilde{Z}_{i\Delta} + R_{i\Delta,2}$, where $\tilde{Z}_{i\Delta} = \Delta^{-1}\int_{i\Delta}^{(i+1)\Delta}\sigma(X_s)dW_s$. The "new" noise $Z_{i\Delta}$ has simpler structure than $\tilde{Z}_{i\Delta}$. This is important for the technical tools used in the proofs (Talagrand deviation inequalities with coupling method instead of direct martingale deviations).

3.2. **Definition of the projection estimator of the drift.** In this section, several definitions and notations are common or close to the ones used in Comte and Genon-Catalot (2020a-b) inducing some unavoidable repetitions for the text to be self-contained. Consider model (1) with observations $(X_{i\Delta})_{1\le i\le n+1}$ decomposed as in (6). Let $A \subset \mathbb{R}$ and let $(\varphi_j, j = 0, \dots, m-1)$ be an orthonormal system of A-supported functions belonging to $\mathbb{L}^2(A, dx)$. Define $S_m = \mathrm{span}(\varphi_0, \dots, \varphi_{m-1})$, the linear space spanned by $(\varphi_0, \dots, \varphi_{m-1})$. The φ_j's may depend on m but this is omitted in notations for simplicity.
We will use further on the following assumption:

- **(A6)** For all j, $\int \varphi_j^2(x)\pi(x)dx < +\infty$, thus $S_m \subset \mathbb{L}^2(A, \pi(x)dx)$, and the collection of models S_m is nested, i.e. $m \le m' \Rightarrow S_m \subset S_{m'}$.

Define an estimator of the drift function b on A, element of S_m, by:

$$\hat{b}_m = \arg\min_{t\in S_m}\gamma_n(t)$$

where $\gamma_n(t)$ is a least-squares contrast given by

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^n \left[t^2(X_{i\Delta}) - 2Y_{i\Delta}t(X_{i\Delta})\right]. \tag{7}$$

For functions s, t, we set

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^n t^2(X_{i\Delta}),\ \langle s,t\rangle_n := \frac{1}{n}\sum_{i=1}^n s(X_{i\Delta})t(X_{i\Delta}) \text{ and } \langle \vec{u}, t\rangle_n = \frac{1}{n}\sum_{i=1}^n u_i t(X_{i\Delta})$$

when $\vec{u}$ is the vector $(u_1, \dots, u_n)'$, $\vec{u}\,'$ the transpose of $\vec{u}$, and t a function. Let

$$\widehat{\Phi}_m = (\varphi_j(X_{i\Delta}))_{1\le i\le n, 0\le j\le m-1},$$

and

$$(8)\quad \widehat{\Psi}_m = (\langle \varphi_j, \varphi_k\rangle_n)_{0\le j,k\le m-1} = \frac{1}{n}\widehat{\Phi}'_m\widehat{\Phi}_m,\ \Psi_m = \left(\int \varphi_j(x)\varphi_k(x)\pi(x)dx\right)_{0\le j,k\le m-1} = \mathbb{E}(\widehat{\Psi}_m).$$

Set $\overrightarrow{Y} = (Y_\Delta, \dots, Y_{n\Delta})'$. Assuming that $\widehat{\Psi}_m$ is invertible almost surely (a.s.) yields

$$(9)\qquad \hat{b}_m = \sum_{j=0}^{m-1} \hat{a}_j^{(m)}\varphi_j, \quad \text{with} \quad \vec{\hat{a}}^{(m)} = (\widehat{\Phi}'_m\widehat{\Phi}_m)^{-1}\widehat{\Phi}'_m\overrightarrow{Y} = \frac{1}{n}\widehat{\Psi}_m^{-1}\widehat{\Phi}'_m\overrightarrow{Y},$$

where $\vec{\hat{a}}^{(m)} = (\hat{a}_0^{(m)}, \dots, \hat{a}_{m-1}^{(m)})'$.

In what follows, the matrices $\widehat{\Psi}_m$ and Ψ_m play a central role for the comparability of the norms $\|.\|_\pi$ and $\|.\|_n$ uniformly over a space S_m. Key tools are deviation inequalities proved in Cohen *al.* (2013, 2019) and Comte and Genon-Catalot (2020a) for independent sequences of random variables. We extend these to a discretely observed diffusion process.

3.3. **Risk bounds for the drift estimator.**

Notations. For M a matrix, we denote by $\|M\|_{\text{op}}$ the operator norm defined as the square root of the largest eigenvalue of MM'. If M is symmetric, it coincides with $\sup\{|\lambda_i|\}$ where λ_i are the eigenvalues of M.

Decomposition (6) allows to handle a not necessarily bounded volatility function. It involves the empirical processes:

$$\nu_n(t) = \frac{1}{n\Delta}\sum_{i=1}^n t(X_{i\Delta})\sigma(X_{i\Delta})(W_{(i+1)\Delta} - W_{i\Delta}), \quad R_{n,k}(t) = \frac{1}{n}\sum_{i=1}^n t(X_{i\Delta})R_i^{(k)},\ k = 1, 2$$

($R_{n,2}$ is the same as in Comte *et al.* (2007)). The following assumption is required:

$$(10)\qquad L(m) := \sup_{x\in A}\sum_{j=0}^{m-1}\varphi_j^2(x) < +\infty.$$

One easily checks that $L(m)$ does not depend on the choice of the $\mathbb{L}^2(dx)$-orthonormal basis of S_m by taking two orthonormal bases and using the orthogonal matrix linking them to obtain the same $L(m)$. Note also that $L(m) = \sup_{t\in S_m \|t\|=1}\sup_{x\in A} t^2(x)$. Under **(A6)**, the spaces S_m are nested, and this implies that the map $m \mapsto L(m)$ is increasing. Assuming $\mathbb{E}\eta^2 < +\infty$ and using (3) and (10), we define

$$(11)\qquad \Psi_{m,\sigma^2} := \left(\int \varphi_j(x)\varphi_k(x)\sigma^2(x)\pi(x)dx\right)_{0\le j,k\le m-1}.$$

To ensure the stability of the least-squares estimator, we must consider a truncated version of $\hat{b}_m$ given by

$$(12)\qquad \widetilde{b}_m = \hat{b}_m\mathbf{1}_{\{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\text{op}}\vee 1)\le \mathfrak{c}n\Delta/\log^2(n\Delta)\}}, \quad \mathfrak{c} = \frac{\theta(3\log(3/2)-1)}{C_0}$$

where C_0 is a numerical constant, $C_0 \ge 72$. Actually, on the set $\{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \vee 1) \le \mathfrak{c}n\Delta/\log^2(n\Delta)\}$, the eigenvalues $(\lambda_i)_{1\le i\le m}$ of $\widehat{\Psi}_m$ are such that $\inf_{1\le i\le m}\lambda_i \ge$

$L(m)\log^2(n\Delta)/(\mathfrak{c}n\Delta)$; thus, the matrix $\widehat{\Psi}_m$ is invertible.
The choice of $\mathfrak{c}$ is done in Lemma 6.1, using Proposition 6.1 (i). Due to θ, the constant $\mathfrak{c}$ is unknown. To avoid this problem, for n large enough, we can change $\mathfrak{c}n\Delta/\log^2(n\Delta)$ into $\mathfrak{C}n\Delta/\log^{2+\epsilon}(n\Delta)$ with $\epsilon > 0$ and a known constant $\mathfrak{C}$.

We stress that, if Ψ_m (resp. $\widehat{\Psi}_m$) is invertible, then $\|\Psi_m^{-1}\|_{\rm op} = \sup_{t\in S_m, \|t\|_\pi=1}\|t\|^2$ (resp. $\|\widehat{\Psi}_m^{-1}\|_{\rm op} = \sup_{t\in S_m,\|t\|_n=1}\|t\|^2$). Thus under **(A6)**, $m \mapsto \|\Psi_m^{-1}\|_{\rm op}$ (resp. $m\mapsto \|\widehat{\Psi}_m^{-1}\|_{\rm op}$) is non-decreasing (see Proposition 2, Section 2.3 in Comte and Genon-Catalot (2020a)). We prove:

Proposition 3.1. *Let $(X_{i\Delta})_{1\le i\le n}$ be observations drawn from model (6) under assumptions* **(A1)-(A4)** *and (10), with $\Delta \le 1$, $\Delta = \Delta_n \to 0$ and $n\Delta \to +\infty$ when $n \to +\infty$. Assume that $\mathbb{E}(\eta^4) < +\infty$. Consider the estimator $\tilde{b}_m$ of b_A. Then for m such that*

$$L(m)(\|\Psi_m^{-1}\|_{\rm op} \vee 1) \le \frac{\mathfrak{c}n\Delta}{2\log^2(n\Delta)} \quad \textit{and} \quad m \le n\Delta \tag{13}$$

with $\mathfrak{c}$ given in (12), we have

$$\mathbb{E}[\|\widetilde{b}_m - b_A\|_n^2] \le 7\inf_{t\in S_m}\|b_A - t\|_\pi^2 + \frac{64}{n\Delta}{\rm Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] + c_1\Delta + \frac{c_2}{n\Delta},$$

$$\mathbb{E}[\|\widetilde{b}_m - b_A\|_\pi^2] \le c_3\{\inf_{t\in S_m}\|b_A - t\|_\pi^2 + \frac{1}{n\Delta}{\rm Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] + \Delta + \frac{1}{n\Delta}\},$$

where c_1, c_2, c_3 are positive constants.

Condition (13) is a stability condition analogous to the one proposed by Cohen *et al.* (2013, 2019). If m is too close to $n\Delta$, the least squares approximation becomes inaccurate. Note that, as for $\widehat{\Psi}_m$, for m satisfying condition (13), Ψ_m is invertible and its eigenvalues are lower bounded by $2L(m)\log^2(n\Delta)/\mathfrak{c}n\Delta$. The condition $m \le n\Delta$ is actually included in the first part of (13): indeed, if $(\theta_j, j = 0,\dots,m-1)$ is an orthonormal basis of S_m with respect to $\mathbb{L}^2(A,\pi(x)dx)$, and $K(m) = \sup_{x\in A}\sum_{j=0}^{m-1}\theta_j^2(x)$, then $K(m) \ge m$ and one can prove that $K(m) \le L(m)\|\Psi_m^{-1}\|_{\rm op}$ (see Lemma 4, section 6.3, in Comte and Genon-Catalot (2020a)).
Under **(A6)**, the bias term, $\inf_{t\in S_m}\|b_A - t\|_\pi^2$ decreases when m increases. The terms $c_1\Delta + c_2/(n\Delta)$ are residual terms tending to zero under our asymptotic framework. The novelty is the variance term ${\rm Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}]$. It is non-decreasing and can be upper bounded in several manners (see Proposition below). Note that if $\sigma(x) \equiv \sigma$, then $\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2} = \sigma^2{\rm Id}_m$ and ${\rm Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] = \sigma^2 m$.

Proposition 3.2. *Let m be an integer. Assume that Ψ_m is invertible and $\mathbb{E}\eta^2 < +\infty$.*

1. *Under* **(A6)**, *$m \mapsto {\rm Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}]$ is non-decreasing.*
2. *If σ is bounded on A, ${\rm Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] \le \|\sigma_A\|_\infty^2 m$.*
3. *Assume that (10) holds, then ${\rm Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] \le \mathbb{E}[\sigma_A^2(X_0)]L(m)\|\Psi_m^{-1}\|_{\rm op}$.*

From (2), if σ is bounded on A (compact or not), we recover the result of Comte *et al.* (2007), see the penalty term in Theorem 1 therein (which has order $\|\sigma_A\|_\infty m/(n\Delta)$).

3.4. **Rates in the compact case.** In this paragraph, we assume that A is compact. Then we can work under

$$0 < \pi_0 \le \pi(x) \le \pi_1, \forall x \in A, \tag{14}$$

where π_0 and π_1 are two unknown fixed constants.

Proposition 7 in Comte and Genon-Catalot (2020a, Section 3.3) implies that

$$\|\Psi_m^{-1}\|_{\text{op}} \le 1/\pi_0.$$

In such a case, condition (13) contains all indices up to order $n\Delta/\log^2(n\Delta)$.

When A is compact, we can also assume that b_A is square-integrable and that $\|\sigma_A\|_\infty < \infty$. Therefore

$$\inf_{t\in S_m} \|b_A - t\|_\pi^2 \le \pi_1 \inf_{t\in S_m} \|b_A - t\|^2 \quad \text{and} \quad \text{Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] \le \|\sigma_A\|_\infty^2 m. \tag{15}$$

The result of Proposition 3.1 thus encompasses the result given in Proposition 1 of Comte *et al.* (2007). Note that, in the latter paper, only the empirical risk is studied.

A benchmark for comparison with our framework is the result of Hoffmann (1999) and we also refer to it in our paper Comte *et al.* (2007). Hoffmann considers a compact estimation set A and a class of drift functions belonging to a ball of the Besov space $B_{\alpha,2,\infty}(A)$. He proves that the optimal rate obtained for the usual $\mathbb{L}^2$ risk is $(n\Delta)^{-2\alpha/(2\alpha+1)}$. The square bias has order $D_m^{-2\alpha}$ with D_m the dimension of the projection space S_m. In our case, if b_A belongs to a ball of the Besov space $B_{\alpha,2,\infty}(A)$, it follows from the last bound in Proposition 3.1 and from (15), that, if $\Delta = o(1/(n\Delta))$, and if $m_{opt} = (n\Delta)^{1/(2\alpha+1)}$, then indeed, we obtain the optimal rate $(n\Delta)^{-2\alpha/(2\alpha+1)}$ for the $\mathbb{L}^2(A)$-risk, thanks to condition (14); see also Inequality (16) in Comte *et al.* (2007). In this sense, we have optimality.

Note that other results exist; it is not always easy to compare our result with other rates obtained as the class of functions involved to assess the bias rate is not the same as ours and not comparable with ours. Moreover, the definition of the risk is not the same as ours. For instance, for drift estimation in ergodic diffusion models continuously observed throughout a time interval $[0, T]$, Dalalyan (2005) uses a weighted $\mathbb{L}^2$ risk, where the weight function is π^2 and not π and a class of weighted (by π^2) Sobolev balls with regularity s. In this framework, the square bias has order T^{-2s} yielding an optimal rate of order $T^{-2s/(2s+1)}$ (the correspondence here is $T = n\Delta$).

3.5. **Rates in the non compact case.** If A is not compact, we have to take into account condition (13), study the risk bound obtained in Proposition 3.1 with the more precise variance term $\text{Tr}[\Psi_m^{-1}\Psi_{m,\sigma^2}]/n\Delta$ and define an appropriate class of functions to assess the square bias rate.

Following Comte and Genon-Catalot (2020a), we introduce the regularity set:

$$W_\pi^s(A, R) = \left\{h \in \mathbb{L}^2(A, \pi(x)dx), \forall \ell \ge 1, \|h - h_\ell^\pi\|_\pi^2 \le R\ell^{-s}\right\},$$

where h_ℓ^π is the $\mathbb{L}^2(A, \pi(x)dx)$-orthogonal projection of h on S_ℓ. If b_A has a given (unknown) regularity s in the previous meaning, that is, if b_A belongs to $W_\pi^s(A, R)$, the square bias satisfies

$$\inf_{t\in S_m} \|b_A - t\|_\pi^2 = \|b_m^\pi - b_A\|_\pi^2 \le Rm^{-s}.$$

Let us justify the definition of the regularity space above by making an analogy. If we consider an orthonormal basis $(\varphi_j, j \geq 0)$ of $\mathbb{L}^2(A, dx)$, it is rather standard to define regularity spaces:

$$W^s(A,R) = \{h \in \mathbb{L}^2(A,dx), \quad \sum_{j\geq 0} j^s \langle h, \varphi_j\rangle^2 \leq R\}, \tag{16}$$

which describe the rate of decay of the coefficients of the function on the basis. If $h \in W^s(A,R)$, then $\forall \ell \geq 1$, $\|h - h_\ell\|^2 \leq R\ell^{-s}$ where h_ℓ is the $\mathbb{L}^2(A,dx)$-orthogonal projection of h on S_ℓ.
In Comte and Genon-Catalot (2020a), in the homoscedastic regression model with n independent observations, $Y_i = b(X_i) + \varepsilon_i$, projection estimators of the regression function are studied. The rate obtained for the $\mathbb{L}^2(f(x)dx)$-risk (f is the commmon density of the X_i's) is $n^{-s/s+1}$ if the regression function belongs to $W^s_f(A,R)$. This rate is optimal on this class of functions as a lower bound is proved. This rate and its optimality is extended to the heteroscedastic model in Comte and Genon-Catalot (2020b, Section 2.3), under the previous conditions and for σ bounded from above and below.

In the simulation study below, the following non compactly supported bases are used for implementation: if $A = \mathbb{R}$, the Hermite basis and if $A = \mathbb{R}^+$, the Laguerre basis. Definitions and elementary properties are recalled in this section and references are given in the introduction. If $A = \mathbb{R}$ and $(\varphi_j, j \geq 0)$ is the Hermite basis (resp. if $A = \mathbb{R}^+$ and $(\varphi_j, j \geq 0)$ is the Laguerre basis), then $W^s(A,R)$ is a Sobolev-Hermite (resp. Sobolev-Laguerre) ball. The index s (and not $2s$) is directly linked with regularity properties of functions (see Section 7 of Comte and Genon-Catalot (2015) and Appendix A.2 of Belomestny *et al.* (2016)).

Here, we do not assume that b_A belongs to $\mathbb{L}^2(A,dx)$ as this would be too restrictive for the drift function of model (1). Thus, the definition of $W^s_\pi(A,R)$. Note that, if $h \in W^s(A,R)$ and moreover π is bounded, then $h \in W^s_\pi(A, R\|\pi\|_\infty)$ as

$$\|h - h^\pi_\ell\|^2_\pi \leq \|\pi\|_\infty \|h - h_\ell\|^2.$$

Now assume that b_A belongs to $W^s_\pi(A,R)$, then, the bound given in Proposition 3.1 becomes,

$$\mathbb{E}\big[\|\tilde{b}_m - b_A\|^2_\pi\big] \lesssim Rm^{-s} + \frac{\mathrm{Tr}[\Psi_m^{-1}\Psi_{m,\sigma^2}]}{n\Delta} + \frac{1}{n\Delta}.$$

If σ is bounded on A (see Proposition 3.2) and if $m^\star = (n\Delta)^{1/(s+1)}$ satisfies (13), we find the rate

$$\mathbb{E}[\|\tilde{b}_{m^\star} - b_A\|^2_\pi] \lesssim (n\Delta)^{-s/(s+1)}.$$

This is coherent with the optimal rate obtained in the usual regression model. The condition that $m^\star = (n\Delta)^{1/(s+1)}$ satisfies (13) is actually mainly a constraint on π, see the discussion in Comte and Genon-Catalot (2020a), Section 3.2, 3.3, 3.4. Notes that these rates are specific to Laguerre and Hermite Sobolev spaces.

In the general case, the best compromise between square bias and variance terms is obtained defining $m^\star$ by the implicit relation $(m^\star)^{-s} = \mathrm{Tr}[\Psi^{-1}_{m^\star}\Psi_{m^\star,\sigma^2}]/n\Delta$ and yields a rate of implicit order $(m^\star)^{-s}$. The order of the quantity $\mathrm{Tr}[\Psi_m^{-1}\Psi_{m,\sigma^2}]$ is empirically illustrated in Figure 4 and seems to be close to cm, for c a constant, in rather general context.
In any case, the choice of $m^\star$ is not possible in practice, as s and R are unknown.

Note that, it is proved in Comte and Genon-Catalot (2020a), Lemma 3 and Proposition 8 (Section 3.4), that, with the non compactly supported Laguerre and Hermite bases,

a) for all $\Delta \leq 1$, for all $m \leq n\Delta$, $\widehat{\Psi}_m$ is a.s. invertible;
b) for all m, Ψ_m is invertible and there exists a constant $c^\star$ such that, $\|\Psi_m^{-1}\|_{\text{op}}^2 \geq c^\star m$.

Property b) enlightens that $\|\Psi_m^{-1}\|_{\text{op}}$ may have a real weight and increase the variance.

3.6. **Model selection.** We precise condition (10) as follows:

(B1) The collection of spaces S_m is such that, for each m, the basis $(\varphi_0, \dots, \varphi_{m-1})$ of S_m satisfies

$$L(m) = \|\sum_{j=0}^{m-1} \varphi_j^2\|_\infty \leq c_\varphi^2 m \quad \text{for} \quad c_\varphi^2 > 0 \quad \text{a constant.} \tag{17}$$

This assumption is shared by most classical bases on a compact support (histograms, trigonometric polynomials). For non compact support, we have in mind concrete examples of orthonormal bases. First, for $A = \mathbb{R}^+$, the basis of $\mathbb{L}^2(\mathbb{R}^+, dx)$ composed of Laguerre functions $(\ell_j, j \geq 0)$, and $S_m = \text{span}(\ell_0, \dots, \ell_{m-1})$; second, for $A = \mathbb{R}$, the basis of $\mathbb{L}^2(\mathbb{R}, dx)$ composed of Hermite functions $(h_j, j \geq 0)$, and $S_m = \text{span}(h_0, \dots, h_{m-1})$ (see Section 4.2 in Comte and Genon-Catalot (2018) and the simulation study below). Laguerre and Hermite functions being uniformly bounded functions, condition (17) holds.
We define, for π upper-bounded on A by $\|\pi\|_\infty$ (that is under **(A5)**), the collection:

$$\mathcal{M}_{n\Delta} = \left\{ m \in \mathbb{N}, c_\varphi^2 m \left(\|\Psi_m^{-1}\|_{\text{op}}^2 \vee 1\right) \leq \frac{\mathfrak{d}}{4} \frac{n\Delta}{\log^2(n\Delta)} \right\}, \quad \mathfrak{d} = \frac{\theta}{8\, C_0\, (\|\pi\|_\infty \vee 1 + \frac{1}{3})}, \tag{18}$$

where θ is defined in (4) and $C_0 \geq 72$ is the same as in $\mathfrak{c}$. The choice of $\mathfrak{d}$ comes from Lemma 6.4 and uses Proposition 6.1, `(ii)`. Due to $\|\pi\|_\infty$ and θ, the constant $\mathfrak{d}$ is unknown. As previously, for n large enough, we can take $\mathfrak{C} n\Delta/\log^{2+\epsilon}(n\Delta)$ instead of $\mathfrak{c} n\Delta/\log^2(n\Delta)$, with $\epsilon > 0$ and a known constant $\mathfrak{C}$. Note that the constraint on m in $\mathcal{M}_{n\Delta}$ is stronger than the one in (13) as $m\,(\|\Psi_m^{-1}\|_{\text{op}} \vee 1) \leq m\,(\|\Psi_m^{-1}\|_{\text{op}}^2 \vee 1)$.
Introducing the random collection of models $\widehat{\mathcal{M}}_{n\Delta}$ given by

$$\widehat{\mathcal{M}}_{n\Delta} = \left\{ m \in \mathbb{N},\, c_\varphi^2 m (\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \vee 1) \leq \mathfrak{d} \frac{n\Delta}{\log^2(n\Delta)} \right\}, \tag{19}$$

with $\mathfrak{d}$ defined in (18), we define the data-driven selection of m by

$$\hat{m} = \arg \min_{m \in \widehat{\mathcal{M}}_{n\Delta}} \left\{ -\|\hat{b}_m\|_n^2 + \kappa c_\varphi^2 s^2 \frac{m \|\widehat{\Psi}_m^{-1}\|_{\text{op}}}{n\Delta} \right\}, \quad s^2 = \mathbb{E}[\sigma^2(X_0)], \tag{20}$$

where κ is a numerical constant.

As usual, in (20), $\hat{m}$ is selected in order to realize automatically a bias-variance tradeoff. Indeed, $-\|\hat{b}_m\|_n^2$ is, up to a constant, an approximation of the squared bias. The second part estimates an upper bound of the variance given in Proposition 3.2-(3).

The set $\widehat{\mathcal{M}}_{n\Delta}$ is the empirical counterpart of $\mathcal{M}_{n\Delta}$ defined by (18), with constant multiplied by 4. This is different from the usual selection procedures where the set of possible values for choosing the dimension m is nonrandom. Note that for $m \in \widehat{\mathcal{M}}_{n\Delta}$, $\hat{b}_m = \widetilde{b}_m$, and these are the only m which are considered.

Theorem 3.1. *Let $(X_{i\Delta})_{1\le i\le n}$ be observations from model (6). Assume that* **(A1)**-**(A5)** *and* **(B1)** *hold and that $\mathbb{E}\eta^6 < +\infty$. Assume that $\Delta = \Delta_n \to 0$ and $n\Delta \to +\infty$ when $n \to +\infty$. Then, there exists a numerical constant κ_0 such that for $\kappa \ge \kappa_0$, we have*

$$(21)\quad \mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_n^2) \le C \inf_{m\in\mathcal{M}_{n\Delta}} [\inf_{t\in S_m} \|b_A - t\|_\pi^2 + \kappa c_\varphi^2 s^2 \frac{m\|\Psi_m^{-1}\|_{\text{op}}}{n\Delta}] + c_1\Delta + \frac{c_2\log^2(n\Delta)}{n\Delta},$$

and

$$(22)\quad \mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_\pi^2) \le C_1 \inf_{m\in\mathcal{M}_{n\Delta}} [\inf_{t\in S_m} \|b_A - t\|_\pi^2 + \kappa c_\varphi^2 s^2 \frac{m\|\Psi_m^{-1}\|_{\text{op}}}{n\Delta}] + c_1'\Delta + \frac{c_2'\log^2(n\Delta)}{n\Delta}$$

where C, C_1 are numerical constants and c_1, c_2, c_1', c_2' are constants depending on π, b, σ.

Inequalities (21) and (22) show that the estimator $\hat{b}_{\hat{m}}$ automatically realizes the compromise between the squared bias and the variance bound. The results are a substantial generalization of Theorem 1 in Comte *et al.* (2007). In Section 4, we explain how to estimate s^2 and how to fix κ.

The constant κ is a specific feature of this selection method. Theorem 3.1 states that, for any drift function b satisfying the assumptions of the theorem, there exists a numerical (universal) constant κ_0 such that inequalities (21)-(22) hold for all $\kappa \ge \kappa_0$. The proof provides a numerical value κ_0 which is not optimal and actually much too large. Finding the best value κ_0 for a given statistical problem is not easy. For instance, this topic is the subject of Birgé and Massart (2007) paper in the Gaussian white noise model where the authors prove that $\kappa > 1$ is required in this case. Thus, for practical implementation of the adaptive estimator, it is standard and commonly done that one starts by preliminary simulations to obtain a value of κ as close as possible to the true one. Afterwards, this value is fixed once and for all.

4. Simulation study

Samples $(X_{i\Delta})_{1\le i\le n}$ were generated for $(n,\Delta) = (100000, 0.02)$ ($n\Delta = 2000$), for $(n,\Delta) = (50000, 0.01)$ ($n\Delta = 500$) and for $(n,\Delta) = (5000, 0.05)$ ($n\Delta = 250$). The following models are considered.

Example 1. Hyperbolic diffusion.

$$b(x) = -\theta x, \sigma(x) = \gamma\sqrt{1+x^2}, \theta = 2, \gamma = 1/\sqrt{2}.$$

Example 2.

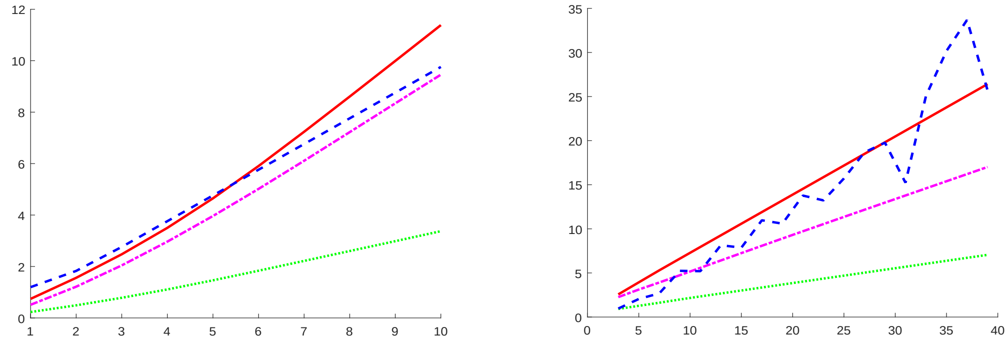
$$b(x) = (1-x^2)\left(-\frac{k}{2}\text{atanh}(x) - \frac{\gamma^2}{4}x\right),\ \sigma(x) = \frac{\gamma}{2}(1-x^2),\ \text{with } k = 2 \text{ and } \gamma = 4.$$

Example 3.

$$b(x) = x\left(-\frac{k}{2}\log(x) + \frac{\gamma^2}{8}\right),\ \sigma(x) = \frac{\gamma}{2}x,\ \text{with } k = 1 \text{ and } \gamma = 1/2.$$

Example 4. Square-root process.

$$b(x) = \frac{d\gamma^2}{4} - kx, \sigma(x) = \sigma\sqrt{x+}, d = 3, k = 2, \gamma = 1.$$



	Ex.1	Ex.2	Ex.3	Ex.4
Hermite basis	0.95	0.98	0.29	0.78
Trigonometric basis	0.71	0.60	0.20	0.49

FIGURE 1. Plots: mean of $m \mapsto \mathrm{Tr}[\widehat{\Psi}_{m,\sigma^2}\widehat{\Psi}_m^{-1}]$ over 200 paths in the four examples (Ex.1 red-full, Ex.2 blue-dashed, Ex.3 green-dotted, Ex.4 magenta-dash-dotted) for the Hermite basis (left) and $m = 1, \dots, 10$; for the trigonometric basis (right) with odd dimension, $m = 3, 5, \dots 39$. In all cases, $n\Delta = 2000$. Table: mean value over m of $\mathrm{Tr}[\widehat{\Psi}_{m,\sigma^2}\widehat{\Psi}_m^{-1}]/m$ for the curves above.

The model of Example 1 is simulated by an Euler scheme with step δ and started for simplicity with $X_0 = 0$. We keep one out of 10 observations i.e. $\Delta = 10\delta$. Assumptions **(A1)-(A2)** hold for $\theta > -\gamma^2/2$. The invariant density is proportional to $1/(1+x^2)^{1+(\theta/\gamma^2)}$ and $\int x^4\pi(x)dx < +\infty$ for $\theta > 3\gamma^2/2$. Setting $Y_t = \mathrm{arsinh}(X_t)$ (where arsinh denotes the inverse hyperbolic sine function), we see that the process (Y_t) satisfies the conditions of Pardoux and Veretennikov (2001) ensuring the exponential β-mixing property. Therefore, (X_t) satisfies **(A4)**.

The other examples are obtained from a d-dimensional Ornstein-Uhlenbeck process in stationary regime, $(U_t)_{t\geq 0}$, with dynamics given by

$$dU_t = -\frac{k}{2}U_t dt + \frac{\gamma}{2}dW_t, \quad U_0 \sim \mathcal{N}(0, \frac{\gamma^2}{4k}I_d). \tag{23}$$

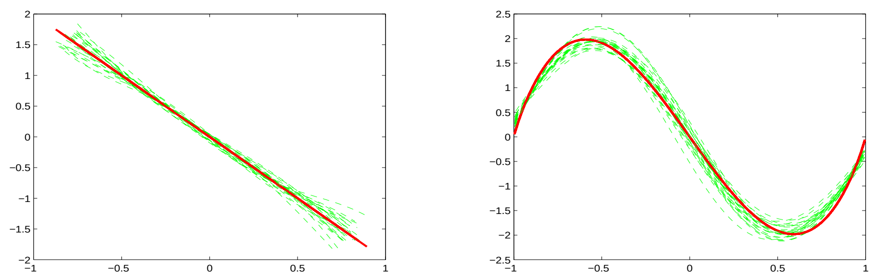
Exact simulation is generated with step Δ by computing

$$U_{(p+1)\Delta} = e^{-\frac{k\Delta}{2}}U_{p\Delta} + \varepsilon_{(p+1)\Delta}, \quad \varepsilon_{k\Delta} \sim_{\mathrm{iid}} \mathcal{N}(0, \frac{\gamma^2(1-e^{-k\Delta})}{4k}I_d).$$

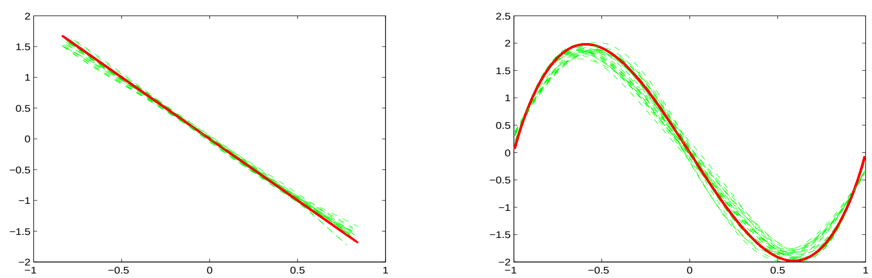
Example 2 corresponds to $X_t = \tanh(U_t)$ where U_t is defined by (23) with $d = 1$. Example 3 is $X_t = \exp(U_t)$ where U_t is defined by (23) with $d = 1$. The process of example 4 is $X_t = \|U_t\|_{2,d}^2$ where $\|.\|_{2,d}$ denotes the Euclidean norm in $\mathbb{R}^d$ and U_t is defined by (23) with $d = 3$.

In Examples 2,3,4, the models are strictly stationary, ergodic and β-mixing but the functions b, σ do not satisfy **(A1)**, and this would require a specific study (to get inequalities

Example 1 Example 2



$\bar{\hat{m}} = 5.2\ (0.9)$, $\bar{\hat{m}}_{\max} = 6.2\ (0.5)$ $\bar{\hat{m}} = 4.2\ (0.4)$, $\bar{\hat{m}}_{\max} = 5\ (0)$



$\bar{\hat{m}} = 6.0\ (0.6)$, $\bar{\hat{m}}_{\max} = 9.9\ (1.0)$ $\bar{\hat{m}} = 5.2\ (1.0)$, $\bar{\hat{m}}_{\max} = 7\ (0)$

	Example 1		Example 2	
	$\widehat{s^2}$	$\widetilde{s^2}$	$\widehat{s^2}$	$\widetilde{s^2}$
$n\Delta = 500$	$0.56_{(0.005)}$	$0.57_{(0.004)}$	$1.38_{(0.046)}$	$1.41_{(0.048)}$
$n\Delta = 2000$	$0.55_{(0.003)}$	$0.57_{(0.002)}$	$1.33_{(0.021)}$	$1.40_{(0.024)}$

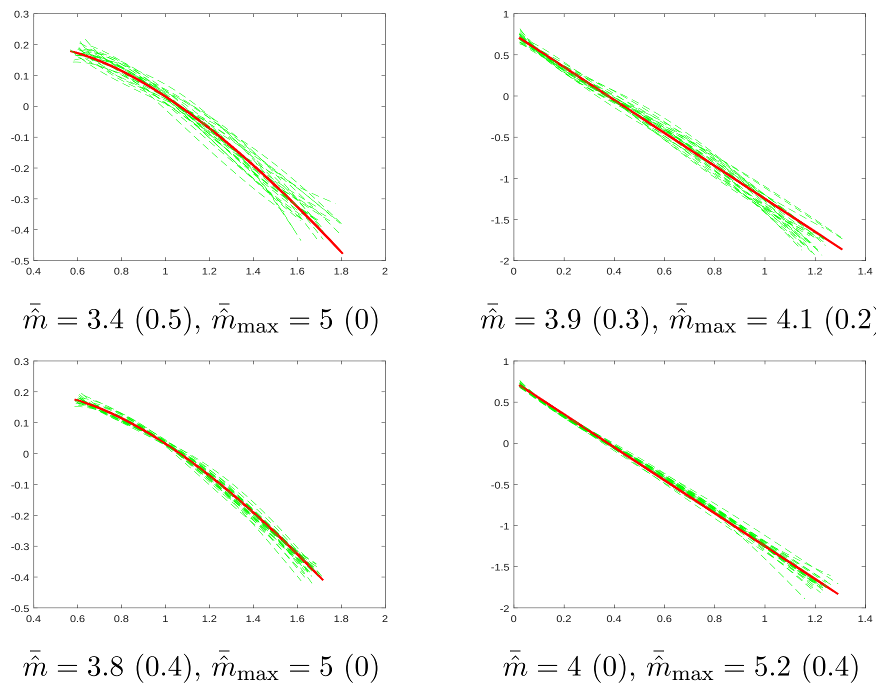
FIGURE 2. Plots: 25 estimated curves in Hermite basis (dotted-green), the true in bold (red), $n\Delta = 500$, top and $n\Delta = 2000$, bottom. Table: Estimation of $s^2 = \mathbb{E}[\sigma^2(X)]$ associated with the paths in the plots (with standard deviation in parenthesis).

of type (5), see Proposition 2.1). Nevertheless, we implement the estimation method. Examples 3,4 provide nonnegative processes and allow to use Laguerre basis.

Implementation is done with the compactly supported trigonometric basis, and the Hermite basis ($A = \mathbb{R}$). For nonnegative processes, we also use the Laguerre basis ($A = \mathbb{R}^+$). All these bases are easy to handle in practice, and we are more specifically interested in the last two bases, which have non-compact support.

• The trigonometric basis on $[a, b]$ is taken as $f_0(x) = 1/\sqrt{b-a}\,\mathbf{1}_{[a,b]}(x)$, $f_{2j-1}(x) = \sqrt{2/(b-a)}\cos(\pi(x-a)/(b-a))\,\mathbf{1}_{[a,b]}(x)$, $f_{2j}(x) = \sqrt{2/(b-a)}\sin(\pi(x-a)/(b-a))\,\mathbf{1}_{[a,b]}(x)$, $j = 1, \ldots, m/2$ for even m. Note that we should have 2π instead of π in the bases functions, but this implies a periodicity which is not true in general for the functions under estimation; we correct it by this "half-period" strategy, which implies that the basis

Example 3 Example 4



	Example 3		Example 4	
	$\widehat{s^2}$	$\widetilde{s^2}$	$\widehat{s^2}$	$\widetilde{s^2}$
$n\Delta = 500$	$0.07_{(0.004)}$	$0.07_{(0.004)}$	$0.37_{(0.01)}$	$0.37_{(0.01)}$
$n\Delta = 2000$	$0.07_{(0.002)}$	$0.07_{(0.002)}$	$0.36_{(0.007)}$	$0.37_{(0.007)}$

FIGURE 3. Plots: 25 estimated curves in Laguerre basis (dotted-green), the true in bold (red), $n\Delta = 500$, top and $n\Delta = 2000$, bottom. Table: Estimation of $s^2 = \mathbb{E}[\sigma^2(X)]$ associated with the paths in the plots (with standard deviation in parenthesis).

is no longer orthogonal (but almost). The collection of models ($S_m = \text{span}\{f_0, \dots, f_{m-1}\}$) is nested and it is easy to see that (17) holds with $c_\varphi^2 = 1/(b-a)$. We take in practice a the 2%-quantile of the $X_{i\Delta}$'s, and b the 98%-quantile.

• Laguerre basis, $A = \mathbb{R}^+$. The Laguerre polynomials (L_j) and the Laguerre functions (ℓ_j) are given by

$$L_j(x) = \sum_{k=0}^{j} (-1)^k \binom{j}{k} \frac{x^k}{k!}, \qquad \ell_j(x) = \sqrt{2} L_j(2x) e^{-x} \mathbf{1}_{x \geq 0}, \quad j \geq 0. \tag{24}$$

The collection $(\ell_j)_{j \geq 0}$ constitutes a complete orthonormal system on $\mathbb{L}^2(\mathbb{R}^+)$ satisfying (see Abramowitz and Stegun (1964)): $\forall j \geq 0,\ \ \forall x \in \mathbb{R}^+,\ \ |\ell_j(x)| \leq \sqrt{2}$. The collection of models ($S_m = \text{span}\{\ell_0, \dots, \ell_{m-1}\}$) is nested and obviously (17) holds with $c_\varphi^2 = 2$.

• Hermite basis, $A = \mathbb{R}$. The Hermite polynomial and the Hermite function of order j are given, for $j \geq 0$, by:

$$H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j}(e^{-x^2}), \quad h_j(x) = c_j H_j(x) e^{-x^2/2}, \quad c_j = \left(2^j j! \sqrt{\pi}\right)^{-1/2} \tag{25}$$

The sequence $(h_j, j \geq 0)$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$. Moreover (see Indritz (1961), Szegö (1959) p.242), $\|h_j\|_\infty \leq \Phi_0, \Phi_0 \simeq 1/\pi^{1/4} \simeq 0.8160$, so that (17) holds with $c_\varphi^2 = \Phi_0^2$. The collection of models $(S_m = \text{span}\{h_0, \dots, h_{m-1}\})$ is obviously nested. Laguerre polynomials were computed using formula (24) and Hermite polynomials with $H_0(x) \equiv 1$, $H_1(x) = x$ and the recursion $H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$.

When computing the set $\widehat{\mathcal{M}}_n$, we observed that the value $c_\varphi^2 m \|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2$ grows very fast. Therefore, to apply our theory, we took

$$\widehat{\mathcal{M}}_{n\Delta} = \left\{ m \in \mathbb{N},\ c_\varphi^2 m (\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^2 \vee 1) \leq \mathfrak{C} n\Delta / \log^{2+\epsilon}(n\Delta) \right\}, \tag{26}$$

with $\epsilon = 0.05$ and very large value for $\mathfrak{C}$: $\mathfrak{C} = 10^{12}$ for all bases. Numerically, we observe that the resulting set (26) is $\widehat{\mathcal{M}}_{n\Delta} = \{1, \dots, \hat{m}_{\max}\}$, with $\hat{m}_{\max}$ between 5 and 10 with the Laguerre or Hermite bases, and $\hat{m}_{\max}$ of order the maximal dimension $(n\Delta)/\log^{2+\epsilon}(n\Delta)$ in the trigonometric basis case (around 40 for $n\Delta = 2000$).

Then we apply the selection procedure described in (20) and $\hat{m}$ is selected as the minimizer of $-\|\hat{b}_m\|_n^2 + \widehat{\text{pen}}(m)$ where the penalty is equal to $\widehat{\text{pen}}(m) = \kappa c_\varphi^2 \widehat{s^2} m \|\widehat{\Psi}_m^{-1}\|_{\text{op}}/(n\Delta)$. Note that the term $s^2 := \mathbb{E}[\sigma^2(X_0)]$ is replaced by an estimator, corresponding to residual least squares associated with the estimator of b on the maximal dimension:

$$\widehat{s^2} = \frac{\Delta}{n} \sum_{i=1}^{n} [Y_{i\Delta} - \hat{b}_{\hat{m}_{\max}}(X_{i\Delta})]^2.$$

In the Tables of Figures 2-3, we compare $\widehat{s^2}$ with $\widetilde{s^2} = (1/n)\sum_{i=1}^n \sigma^2(X_{i\Delta})$, a pseudo-estimator using the (unavailable) knowledge of the function σ^2. The comparison is done for the 25 paths generated for Figures 2-3: we can see that the values of $\widehat{s^2}$ are $\widetilde{s^2}$ are nearly identical.

The constant κ is standardly calibrated by preliminary simulations and taken equal to $\kappa = 5.10^{-2}$ for the trigonometric basis and $\kappa = 2.10^{-4}$ for the Laguerre and Hermite bases. It is not surprising that κ must be chosen very small in these last cases: this is due to the fact that $m\|\widehat{\Psi}_m^{-1}\|_{\text{op}}$ is large, as noted above. In relation with the variance term obtained in the risk bounds of Proposition 3.1, we computed the mean of values of $\text{Tr}[\widehat{\Psi}_{m,\sigma^2}\widehat{\Psi}_m^{-1}]$ over 200 paths, for values of m going from 1 to 10 in the Hermite case, and odd dimensions between 3 and 39 for the trigonometric case. The results, for the four examples, are plotted in Figure 1: we obtain almost linear increase (which also holds path by path for these values of m): this means that in all cases the traces are of order m with trends estimated in the table associated to the plots, all between 0.2 and 1. In other words, the product of the two matrices is rather stable and with reasonable orders, compared to the ones obtained for $\|\widehat{\Psi}_m^{-1}\|_{\text{op}}$.The only exception is the Example 2 in the half-trigonometric basis, and this is probaly due to the fact that the function looks also like trigonometric function on a complete period.

	$n\Delta =$	MISE 250	MISE 500	MISE 2000	Selected dimension 250	Selected dimension 500	Selected dimension 2000
Ex.1	Hermite	$3.24_{(3.44)}$	$1.57_{(1.27)}$	$0.49_{(0.40)}$	$4.86_{(0.89)}$	$5.22_{(0.91)}$	$5.95_{(0.53)}$
	Trigo	$3.55_{(2.95)}$	$1.62_{(1.25)}$	$0.52_{(0.38)}$	$4.68_{(1.63)}$	$5.66_{(0.71)}$	$6.05_{(0.50)}$
Ex.2	Hermite	$15.4_{(9.41)}$	$6.07_{(3.88)}$	$4.38_{(1.81)}$	$4.28_{(0.45)}$	$4.25_{(0.44)}$	$5.38_{(0.89)}$
	Trigo	$72.4_{(50.9)}$	$7.84_{(4.39)}$	$4.11_{(2.20)}$	$3.35_{(1.48)}$	$5.33_{(0.47)}$	$6.00_{(0.05)}$
Ex.3	Laguerre	$0.16_{(.17)}$	$0.10_{(0.12)}$	$0.02_{(0.02)}$	$3.24_{(0.43)}$	$3.47_{(0.50)}$	$3.95_{(0.21)}$
	Hermite	$0.20_{(0.23)}$	$0.13_{(0.16)}$	$0.06_{(0.02)}$	$2.99_{(0.09)}$	$3.00_{(0.00)}$	$3.00_{(0.00)}$
	Trigo	$0.24_{(0.22)}$	$0.19_{(0.26)}$	$0.05_{(0.05)}$	$3.18_{(1.04)}$	$4.09_{(1.26)}$	$5.31_{(0.82)}$
Ex.4	Laguerre	$1.97_{(2.82)}$	$0.83_{(0.82)}$	$0.29_{(0.47)}$	$3.86_{(0.36)}$	$3.95_{(0.21)}$	$4.00_{(0.10)}$
	Hermite	$2.17_{(2.54)}$	$1.13_{(0.92)}$	$0.41_{0.25)}$	$3.76_{(0.43)}$	$3.97_{(0.17)}$	$4.00_{(0.00)}$
	Trigo	$4.69_{(4.99)}$	$2.72_{(2.32)}$	$0.90_{(0.72)}$	$5.15_{(1.13)}$	$6.08_{(1.41)}$	$8.23_{(1.59)}$

Table 1. MISE on the intervals of observations multiplied by 100 (with standard deviation multiplied by 100 in parenthesis), for the four examples with different bases. Mean of selected dimensions (with std in parenthesis) computed over 400 repetitions.

We present in Figures 2-3 beams of 25 estimators $\hat{b}_{\hat{m}}$ corresponding to 25 simulated trajectories of each model using the Hermite basis for Examples 1 and 2 and Laguerre basis for Examples 3 and 4. The intervals of representation are $[a, b]$ with a the 2%-quantile of the $X_{i\Delta}$'s and b the 98% quantile, the same which were chosen as support of the trigonometric basis. We stress that the value of $\hat{m}$ is rather small: under each graph, we give the mean $\bar{\hat{m}}$ computed over the 25 estimators and the mean of the maximal value $\hat{m}_{\max}$, both with standard deviation in parenthesis. It is noteworthy that the function is very well reconstructed using a small number of coefficients. This is confirmed by Table 1, which also gives the mean selected dimensions, but now over 400 iterations. Note that they have comparable orders for all three bases, and slightly increase when $n\Delta$ increases, as expected (asymptotically, the optimal dimension increases). The MISE obtained for 400 iterations and the three sample sizes $(n, \Delta) = (100000, 0.02)$ ($n\Delta = 2000$), $(n, \Delta) = (50000, 0.01)$ ($n\Delta = 500$) and $(n, \Delta) = (5000, 0.05)$ ($n\Delta = 250$) are also presented; they are computed for each path on the same interquantile 2%-98% (random) interval and for the same data, and then averaged. As expected, the MISEs decrease when $n\Delta$ increases, and the results have the same orders for all bases. We note that the Laguerre basis gives better results than the two others basis. Example 2 for small $n\Delta$ seems to be overpenalized and would be improved by plugging a much smaller penalty constant. On the whole, Laguerre and Hermite bases are preferable to trigonometric basis. Their graphical representation and the range for computation of MISEs have to rely on the range of data, but their support is not random and the estimation procedure more intrinsic, which is not the case for compactly supported bases, whose support is in practice random.

5. Concluding remarks

In this paper, we revisit the problem of nonparametric drift estimation for an ergodic diffusion from discrete observations of the sample path. As in Comte *et al.* (2007), an estimation set A is chosen. The drift function is estimated on A using an approximate regression equation by a least squares contrast which is minimized over a finite dimensional subspace S_m of $\mathbb{L}^2(A, dx)$. This yields a collection of estimators indexed by the dimension of the projection space. A data-driven procedure is proposed to select the best dimension using a penalization criterion. While in Comte *et al.* (2007), the set A must be compact and the diffusion coefficient must be uniformly bounded, the novelty of the present paper is to get rid of these two assumptions. This leads to considerable modifications in the method of estimation and complicates the proofs a lot. First, we rely on a slightly different approximate regression equation. This allows to modify the way of dealing with the adaptive procedure. Second, the possible dimensions to define the projection estimators are restricted to a set involving the inversion of a matrix which does not even appear when the set A is compact. In the risk bounds, the variance term is different from the case where A is compact and $\sigma(.)$ bounded; it is very relevant in those cases also, as it provides a natural estimator of unknown quantities (infinite norm of σ or lower bound of the stationary density). Moreover, to define the adaptive procedure, the adequate dimension is to be selected within a random set. This induces difficulties and a non standard treatment of the classical method of penalization. The estimator obtained is nevertheless adaptive in the sense that its $\mathbb{L}^2$-risk achieves the best compromise between the squared bias and the new variance term.
An important question may be to look at rates of convergence and optimality in this new setting. This is treated in the case of the simple regression model with independent data on non compact support in Comte and Genon-Catalot (2020a) and is worth of interest in the diffusion context.
Estimation of σ^2 could be investigated too under the same set of assumptions but leads to rather lengthy developments.

6. Proofs

We denote by $x \lesssim y$, $x \leq cy$ for some constant c which does not depend on n, Δ, m.

6.1. **Preliminary properties.** Consider the set where the empirical and $\mathbb{L}^2(A, \pi(x)dx)$-norms on S_m are equivalent:

$$\Omega_m(u) = \left\{ \sup_{t \in S_m,\, t \neq 0} \left| \frac{\|t\|_n^2}{\|t\|_\pi^2} - 1 \right| \leq u \right\}. \tag{27}$$

We generalize to the diffusion context Proposition 3 of Comte and Genon-Catalot (2020a)(see also Theorem 1 of Cohen *et al.* (2013) for part `(i)`).

Proposition 6.1. *Let $(X_{i\Delta})_i$ be a discrete sampling of the process (X_t) given by (1) and assume* **(A1)-(A4)** *(thus $(X_{i\Delta})_i$ is strictly stationary and geometrically β-mixing with β-mixing coefficients satisfying $\beta(i) = \beta_X(i\Delta) \leq Ke^{-\theta i\Delta}$ for some constants $K > 0, \theta > 0$). Consider a basis satisfying (10) and let Id_m denote the $m \times m$ identity matrix.*

(i) *Assume that Ψ_m is invertible. For $\widehat{\Psi}_m$ defined by Equation (8), for all $u \in [0,1]$*

$$\begin{aligned}\mathbb{P}(\Omega_m(u)^c) &= \mathbb{P}\left[\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > u\right] \\ &\leq 4m\exp\left(-\frac{n\Delta\theta\, c(u)}{12\log(n\Delta)\ \ L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}} \vee 1)}\right) + \frac{\theta}{6(n\Delta)^5},\end{aligned}$$

where $c(u) = u + (1-u)\log(1-u)$.
(ii) *If in addition* **(A5)** *holds, then, for all $u > 0$,*

$$\mathbb{P}\left[\|\Psi_m - \widehat{\Psi}_m\|_{\mathrm{op}} \geq u\right] \leq 4m\exp\left(-\frac{n\Delta\theta u^2/2}{12L(m)\log(n\Delta)\ (\|\pi\|_\infty \vee 1 + 2u/3)}\right) + \frac{\theta}{6(n\Delta)^5}.$$

6.2. **Proof of Proposition 6.1.** For $t = \sum_{j=0}^{m-1} x_j\varphi_j$ in S_m, $\|t\|_\pi^2 = \vec{x}'\Psi_m\vec{x}$, $\|t\|_n^2 = \vec{x}'\widehat{\Psi}_m\vec{x}$. Thus,

$$\sup_{t\in S_m, \|t\|_\pi=1}\left|\frac{1}{n}\sum_{i=1}^n[t^2(X_i) - \mathbb{E}t^2(X_i)]\right| = \sup_{\vec{x}\in\mathbb{R}^m, \|\Psi_m^{1/2}\vec{x}\|_{2,m}=1}\left|\vec{x}'\widehat{\Psi}_m\vec{x} - \vec{x}'\Psi_m\vec{x}\right|$$

$$= \sup_{\vec{u}\in\mathbb{R}^m, \|\vec{u}\|_{2,m}=1}\left|\vec{u}'\Psi_m^{-1/2}(\widehat{\Psi}_m - \Psi_m)\Psi_m^{-1/2}\vec{u}\right| = \|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}}.$$

Hence,

$$\Omega_m(u)^c = \left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m\|_{\mathrm{op}} > u\right\}. \tag{28}$$

Now, we consider the coupling method and the associated variables $(X^*_{i\Delta})$ with Berbee's Lemma, see Berbee (1979), with the method described in Viennet (1997, Prop.5.1 and its proof p.484). Assume for simplicity that $n = 2p_nq_n$ for integers p_n, q_n. Then there exist random variables $X^*_{i\Delta}$, $i = 1, ..., n$ satisfying the following properties:

- For $\ell = 0, ..., p_n - 1$, the random vectors

  $$\vec{X}_{\ell,1} = \left(X_{(2\ell q_n+1)\Delta}, ..., X_{(2\ell+1)q_n\Delta}\right)' \text{ and } \vec{X}^*_{\ell,1} = \left(X^*_{(2\ell q_n+1)\Delta}, ..., X^*_{(2\ell+1)q_n\Delta}\right)'$$

  have the same distribution, and so have the random vectors

  $$\vec{X}_{\ell,2} = \left(X_{[(2\ell+1)q_n+1]\Delta}, ..., X_{(2\ell+2)q_n\Delta}\right)' \text{ and } \vec{X}^*_{\ell,2} = \left(X^*_{[(2\ell+1)q_n+1]\Delta}, ..., X^*_{(2\ell+2)q_n\Delta}\right)'.$$

- For $\ell = 0, ..., p_n - 1$,

$$\mathbb{P}\left[\vec{X}_{\ell,1} \neq \vec{X}^*_{\ell,1}\right] \leq \beta(q_n) = \beta_X(q_n\Delta) \text{ and } \mathbb{P}\left[\vec{X}_{\ell,2} \neq \vec{X}^*_{\ell,2}\right] \leq \beta_X(q_n\Delta). \tag{29}$$

- For each $\delta \in \{1,2\}$, the random vectors $\vec{X}^*_{0,\delta}, ..., \vec{X}^*_{p_n-1,\delta}$ are independent.

Then let $\Omega^* = \{X_{i\Delta} = X^*_{i\Delta}, i = 1, \dots, n\}$ and write that

$$\begin{aligned}\mathbb{P}\left[\|\mathrm{Id}_m - \Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}\|_{\mathrm{op}} > u\right] &\leq \mathbb{P}\left[\{\|\mathrm{Id}_m - \Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}\|_{\mathrm{op}} > u\} \cap \Omega^*\right] \\ &\quad + \mathbb{P}[(\Omega^*)^c].\end{aligned} \tag{30}$$

Using the definition of the variables $X^*_{i\Delta}$, we get $\mathbb{P}[(\Omega^*)^c] \leq 2p_n\beta_X(q_n\Delta) \leq 2p_ne^{-\theta q_n\Delta}$.

Choosing $q_n\Delta = 6\log(n\Delta)/\theta$, thus $2p_n = n/q_n = n\Delta\theta/6\log(n\Delta)$, yields

$$\mathbb{P}[(\Omega^*)^c] \leq \frac{\theta n\Delta}{6\log(n\Delta)(n\Delta)^6} \leq \frac{\theta}{6(n\Delta)^5}. \tag{31}$$

Now, we write $\mathbf{S}_m = (1/2)(\mathbf{S}_{m,1} + \mathbf{S}_{m,2})$ where $\mathbf{S}_m$ is given by

$$\mathbf{S}_m = \Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \mathrm{Id}_m = \frac{1}{n}\sum_{i=1}^n\{\mathbf{K}_m(X_{i\Delta}) - \mathbb{E}\left[\mathbf{K}_m(X_{i\Delta})\right]\},$$

$$\mathbf{K}_m(X_{i\Delta}) = \Psi_m^{-1/2}\widetilde{\mathbf{K}}_m(X_{i\Delta})\Psi_m^{-1/2}, \quad \widetilde{\mathbf{K}}_m(X_{i\Delta}) = (\varphi_j(X_{i\Delta})\varphi_k(X_{i\Delta}))_{0\leq j,k\leq m-1}, \tag{32}$$

with obviously $\mathbb{E}\left[\mathbf{K}_m(X_{i\Delta})\right] = \mathrm{Id}_m$. Here, $\mathbf{S}_{m,1}$ is built with the $\vec{X}_{\ell,1}$:

$$\mathbf{S}_{m,1} = \frac{1}{p_n}\sum_{\ell=0}^{p_n-1}\frac{1}{q_n}\sum_{r=1}^{q_n}\{\mathbf{K}_m(X_{(2\ell q_n+r)\Delta}) - \mathbb{E}(\mathbf{K}_m(X_{(2\ell q_n+r)\Delta}))\}$$

and $\mathbf{S}_{m,2}$ is analogously defined with the $\vec{X}_{\ell,2}$. We have

$$\begin{aligned}\mathbb{P}\left[\{\|\mathrm{Id}_m - \Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}\|_{\mathrm{op}} > u\}\cap\Omega^*\right] &= \mathbb{P}\left[\{\|\mathbf{S}_{m,1} + \mathbf{S}_{m,2}\|_{\mathrm{op}} > 2u\}\cap\Omega^*\right] \\ &\leq \mathbb{P}\left[\|\mathbf{S}^*_{m,1}\|_{\mathrm{op}} > u\right] + \mathbb{P}\left[\|\mathbf{S}^*_{m,2}\|_{\mathrm{op}} > u\right], \end{aligned} \tag{33}$$

where $\mathbf{S}^*_{m,\delta}$, $\delta = 1, 2$ are built on the $\vec{X}^*_{\ell,\delta}$. The two terms are similar so we only treat one. We write

$$S^*_{m,1} = \sum_{\ell=0}^{p_n-1}\mathbf{G}^*_\ell - \mathrm{Id}_m \quad \text{where} \quad \mathbf{G}^*_\ell = \frac{1}{p_n}\frac{1}{q_n}\sum_{r=1}^{q_n}\mathbf{K}_m(X^*_{(2\ell q_n+r)\Delta}).$$

By (32), we have $\|\mathbf{G}^*_\ell\|_{\mathrm{op}} \leq L(m)\|\Psi_m^{-1}\|_{\mathrm{op}}/p_n$. Then, as in Theorem 1 in Cohen *et al.* (2013), we use the Chernoff bound of Tropp (2015) and we get

$$\mathbb{P}(\|\mathbf{S}^*_{m,1}\|_{\mathrm{op}} > u) \leq 2m\exp\left(-\frac{c(u)p_n}{L(m)\|\Psi_m^{-1}\|_{\mathrm{op}}}\right).$$

Using the definition of p_n, gives the result `(i)` of Proposition 6.1.

To prove `(ii)`, we proceed similarly and bound $\mathbb{P}\left[\{\|\Psi_m - \widehat{\Psi}_m\|_{\mathrm{op}} \geq u\}\cap\Omega^*\right]$. We write $\widetilde{\mathbf{S}}_m = (1/2)(\widetilde{\mathbf{S}}_{m,1} + \widetilde{\mathbf{S}}_{m,2})$ where ($\widetilde{\mathbf{K}}_m$ is defined by (32)):

$$\begin{aligned}\widetilde{\mathbf{S}}_m &= \frac{1}{n}\sum_{i=1}^n\{\widetilde{\mathbf{K}}_m(X_{i\Delta}) - \mathbb{E}\left[\widetilde{\mathbf{K}}_m(X_{i\Delta})\right]\} = \widehat{\Psi}_m - \Psi_m, \\ \widetilde{\mathbf{S}}_{m,1} &= \frac{1}{p_n}\sum_{\ell=0}^{p_n-1}\frac{1}{q_n}\sum_{r=1}^{q_n}\{\widetilde{\mathbf{K}}_m(X_{(2\ell q_n+r)\Delta}) - \mathbb{E}(\widetilde{\mathbf{K}}_m(X_{(2\ell q_n+r)\Delta}))\}\end{aligned}$$

is built with the $\vec{X}_{\ell,1}$ and $\widetilde{\mathbf{S}}_{m,2}$ is analogously defined with the $\vec{X}_{\ell,2}$. As above,

$$\mathbb{P}\left[\{\|\Psi_m - \widehat{\Psi}_m\|_{\mathrm{op}} \geq u\}\cap\Omega^*\right] \leq \mathbb{P}\left[\|\widetilde{\mathbf{S}}^*_{m,1}\|_{\mathrm{op}} \geq u\right] + \mathbb{P}\left[\|\widetilde{\mathbf{S}}^*_{m,2}\|_{\mathrm{op}} \geq u\right], \tag{34}$$

where $\widetilde{\mathbf{S}}^*_{m,\delta}$, for $\delta = 1, 2$ are built on the $\vec{X}^*_{\ell,\delta}$. We treat only the first term applying Tropp's result to $\widetilde{\mathbf{S}}^*_{m,1}$ which is a sum of p_n independent matrices.

It is clear that for all ℓ, r, $\|\widetilde{\mathbf{K}}_m(X^*_{(2\ell q_n+r)\Delta})\|_{\rm op} \le L(m)$ a.s. and thus

$$\frac{1}{p_n q_n}\|\sum_{r=1}^{q_n}\widetilde{\mathbf{K}}_m(X^*_{(2\ell q_n+r)\Delta}) - \mathbb{E}(\widetilde{\mathbf{K}}_m(X^*_{(2\ell q_n+r)\Delta}))\|_{\rm op} \le 2\frac{L(m)}{p_n} = \frac{24}{\theta}\frac{L(m)\log(n\Delta)}{n\Delta} := \mathbf{L}.$$

Now, we must bound the variance of $\mathbf{S}^*_{m,1}$. We have

$$\nu(\widetilde{\mathbf{S}}^*_{m,1}) = \frac{1}{p_n}\sup_{\|\vec{x}\|_{2,m}=1}\mathbb{E}\left(\frac{1}{q_n^2}\left\|\left[\sum_{r=1}^{q_n}(\widetilde{\mathbf{K}}_m(X^*_{r\Delta}) - \mathbb{E}(\widetilde{\mathbf{K}}_m(X^*_{r\Delta})))\right]\vec{x}\right\|^2_{2,m}\right)$$

Next,

$$\begin{aligned}\widetilde{\mathbb{E}}_1 &= \mathbb{E}\left(\frac{1}{q_n^2}\left\|\left[\sum_{r=1}^{q_n}(\widetilde{\mathbf{K}}_m(X^*_{r\Delta}) - \mathbb{E}(\widetilde{\mathbf{K}}_m(X^*_{r\Delta})))\right]\vec{x}\right\|^2_{2,m}\right)\\ &= \frac{1}{q_n^2}\sum_{j=0}^{m-1}\operatorname{Var}\left[\sum_{r=1}^{q_n}\sum_{k=0}^{m-1}\varphi_j(X_{r\Delta})\varphi_k(X_{r\Delta})x_k\right]\\ &= \frac{1}{q_n}\sum_{j=0}^{m-1}\sum_{r=1}^{q_n}\operatorname{Var}\left[\sum_{k=0}^{m-1}\varphi_j(X_{r\Delta})\varphi_k(X_{r\Delta})x_k\right] = \sum_{j=0}^{m-1}\operatorname{Var}\left[\sum_{k=0}^{m-1}\varphi_j(X_{r\Delta})\varphi_k(X_{r\Delta})x_k\right]\end{aligned}$$

Therefore, for $\|x\|_{2,m}=1$,

$$\begin{aligned}\widetilde{\mathbb{E}}_1 &\le \sum_{j=0}^{m-1}\mathbb{E}\left[\left(\sum_{k=0}^{m-1}\varphi_j(X_{r\Delta})\varphi_k(X_{r\Delta})x_k\right)^2\right] \le L(m)\mathbb{E}\left[\left(\sum_{k=0}^{m-1}\varphi_k(X_{r\Delta})x_k\right)^2\right]\\ &= L(m)\int\left(\sum_{k=0}^{m-1}\varphi_k(u)x_k\right)^2\pi(u)du \le L(m)\|\pi\|_\infty.\end{aligned}$$

Thus, $\nu(\widetilde{\mathbf{S}}^*_{m,1}) \le L(m)\,\|\pi\|_\infty/p_n = (12/\theta)(L(m)\log(n\Delta)/n\Delta)$ using the definition of q_n. Finally, applying Theorem 6.1 given in appendix and joining the analogous of (30), (31), (34), the value of $\mathbf{L}$ and the bound on $\nu(\widetilde{\mathbf{S}}^*_{m,1})$ gives `(ii)`. □

6.3. **Proof of Proposition 3.1.** We define the sets

$$\Lambda_m = \left\{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\rm op}\vee 1) \le \mathfrak{c}\frac{n\Delta}{\log^2(n\Delta)}\right\} \text{ and } \Omega_m := \Omega_m(1/2) = \left\{\left|\frac{\|t\|_n^2}{\|t\|_\pi^2} - 1\right| \le \frac{1}{2}, \forall t\in S_m\right\}.$$

Below, we prove the following lemma

Lemma 6.1. *Under the assumptions of Proposition 3.1, for m satisfying (13), we have*

$$\mathbb{P}(\Lambda_m^c) \lesssim 1/(n\Delta)^5, \quad \mathbb{P}(\Omega_m^c) \lesssim 1/(n\Delta)^5,$$

for any $\mathfrak{c} \le \frac{\theta(3\log(3/2)-1)}{6\times 12}$.

Now, we write

$$\begin{aligned}\|\widetilde{b}_m - b_A\|_n^2 &= \|\hat{b}_m - b_A\|_n^2\mathbf{1}_{\Lambda_m} + \|b_A\|_n^2\mathbf{1}_{\Lambda_m^c}\\ &= \|\hat{b}_m - b_A\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m} + \|\hat{b}_m - b_A\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c} + \|b_A\|_n^2\mathbf{1}_{\Lambda_m^c}.\end{aligned}\tag{35}$$

We bound successively the expectation of the three terms.

• Study of $\mathbb{E}(\|\hat{b}_m - b_A\|_n^2 \mathbf{1}_{\Lambda_m\cap\Omega_m})$.
The following equality holds, for all functions s, t in S_m:

$$\gamma_n(t) - \gamma_n(s) = \|t-b_A\|_n^2 - \|s-b_A\|_n^2 - 2\nu_n(t-s) - 2R_{n,1}(t-s) - 2R_{n,2}(t-s). \tag{36}$$

Thus $\gamma_n(\hat{b}_m) \le \gamma_n(b_m)$, for any function b_m in S_m, implies

$$\|\hat{b}_m - b_A\|_n^2 \le \|b_m - b_A\|_n^2 + 2\nu_n(\hat{b}_m - b_m) + 2R_{n,1}(\hat{b}_m - b_m) + 2R_{n,2}(\hat{b}_m - b_m). \tag{37}$$

We first study the last term. We have

$$2R_{n,2}(\hat{b}_m - b_m) \le \frac{1}{8}\|\hat{b}_m - b_m\|_n^2 + \frac{8}{n\Delta^2}\sum_{k=1}^n \left(\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\right)^2$$

Now, using (5) for $f=b$, we get

$$\begin{aligned}\mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta}(b(X_s)-b(X_{k\Delta}))ds\right)^2 &\le \Delta\int_{k\Delta}^{(k+1)\Delta}\mathbb{E}[(b(X_s)-b(X_{k\Delta}))^2]ds \\ &\le c\Delta^3(1+\mathbb{E}\eta^2).\end{aligned}$$

Thus, we have:

$$\begin{aligned}2\mathbb{E}\left[R_{n,2}(\hat{b}_m - b_m)\mathbf{1}_{\Lambda_m\cap\Omega_m}\right] &\le \frac{1}{8}\mathbb{E}[\|\hat{b}_m - b_m\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m}] + 8c'\Delta \\ &\le \frac{1}{4}\mathbb{E}[\|\hat{b}_m - b_A\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m}] + \frac{1}{4}\mathbb{E}[\|b_A - b_m\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m}] + 8c'\Delta. \qquad (38)\end{aligned}$$

For the two other terms to study in (37), we write

$$\begin{aligned}&2\mathbb{E}\left[(\nu_n(\hat{b}_m - b_m) + R_{n,1}(\hat{b}_m - b_m))\mathbf{1}_{\Lambda_m\cap\Omega_m}\right] \\ \le\ &\mathbb{E}\left[\frac{1}{8}\|\hat{b}_m - b_m\|_\pi^2\mathbf{1}_{\Lambda_m\cap\Omega_m} + 8\sup_{t\in S_m,\|t\|_\pi=1}[\nu_n(t)+R_{n,1}(t)]^2\right]\end{aligned}$$

Then, by the definition of Ω_m, we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{8}\|\hat{b}_m - b_m\|_\pi^2\mathbf{1}_{\Lambda_m\cap\Omega_m}\right] &\le \mathbb{E}\left[\frac{1}{4}\|\hat{b}_m - b_m\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m}\right] \\ &\le \mathbb{E}\left[\frac{1}{2}\|\hat{b}_m - b_A\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m}\right] + \mathbb{E}\left[\frac{1}{2}\|b_A - b_m\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m}\right]. \qquad (39)\end{aligned}$$

Moreover

$$8\mathbb{E}(\sup_{t\in S_m,\|t\|_\pi=1}[\nu_n(t)+R_{n,1}(t)]^2]) \le 16\mathbb{E}(\sup_{t\in S_m,\|t\|_\pi=1}\nu_n^2(t) + \sup_{t\in S_m,\|t\|_\pi=1}R_{n,1}^2(t)).$$

Now the following result holds.

Lemma 6.2. *Under the assumptions of Proposition 3.1, we have*

$$\mathbb{E}\left(\sup_{t\in S_m,\|t\|_\pi=1}\nu_n^2(t)\right) = \frac{1}{n\Delta}\mathrm{Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}], \tag{40}$$

$$\mathbb{E}(\sup_{t\in S_m,\|t\|_\pi=1} R^2_{n,1}(t)) \le C\Delta. \tag{41}$$

Therefore, gathering (37), (38), (39), (40) and (41) we get, for m satisfying (13),

$$\frac{1}{4}\mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2 \mathbf{1}_{\Lambda_m\cap\Omega_m}\big] \le \frac{7}{4}\mathbb{E}\left[\|b_m - b_A\|_n^2 \mathbf{1}_{\Lambda_m\cap\Omega_m}\right] + \frac{16}{n\Delta}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] + c_1\Delta.$$

• Study of $\mathbb{E}(\|b_A\|_n^2\mathbf{1}_{\Lambda_m^c})$.
We use Lemma 6.1. By the Cauchy-Schwarz inequality, $\mathbb{E}(\|b_A\|_n^2\mathbf{1}_{\Lambda_m^c}) \le c\mathbb{E}^{1/2}[b^4(X_0)]/(n\Delta)^{5/2}$ with the bound on $\mathbb{P}(\Lambda_m^c)$ given in Lemma 6.1.

• Study of $\mathbb{E}(\|\hat{b}_m - b_A\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c})$.
We introduce the operator Π_m of orthogonal projection, for the scalar product of $\mathbb{R}^n$, onto the subspace $\{(t(X_\Delta),\dots,t(X(_{n\Delta})), t\in S_m\}$ of $\mathbb{R}^n$ and denote by $\Pi_m b_A$ the projection of $\vec{b}_A = (b_A(X_\Delta),\dots,b_A(X_{n\Delta}))'$. We can write:

$$\|\hat{b}_m - b_A\|_n^2 = \|\hat{b}_m - \Pi_m b_A\|_n^2 + \|\Pi_m b_A - b_A\|_n^2 \le \|\hat{b}_m - \Pi_m b_A\|_n^2 + \|b_A\|_n^2. \tag{42}$$

Recall that (see (6)) $Y_{k\Delta} = b(X_{k\Delta}) + E_{k\Delta}$ with $E_{k\Delta} = Z_{k\Delta} + R_{k\Delta}$. Elementary computations yield:

$$\Pi_m b_A = (\sum_{j=0}^{m-1}\tilde{a}_j^{(m)}\varphi_j(X_{k\Delta}), k=1,\dots,n)' \quad \text{with} \quad \vec{\tilde{a}}^{(m)} = \frac{1}{n}\widehat{\Psi}_m^{-1}\widehat{\Phi}_m'\vec{b}_A,$$

while $\hat{b}_m = (\sum_{j=0}^{m-1}\hat{a}_j^{(m)}\varphi_j(X_{k\Delta}), k=1,\dots,n)'$ with $\vec{\hat{a}}^{(m)} = \frac{1}{n}\widehat{\Psi}_m^{-1}\widehat{\Phi}_m'\vec{Y}$. Therefore, setting $\vec{E} = (R_\Delta + Z_\Delta,\dots,R_{n\Delta}+Z_{n\Delta})'$, we have

$$\begin{aligned}\|\hat{b}_m - \Pi_m b_A\|_n^2 &= \frac{1}{n}\sum_{i=1}^n\left(\sum_{j=0}^{m-1}(\hat{a}_j^{(m)} - \tilde{a}_j^{(m)})\varphi_j(X_{i\Delta})\right)^2 = (\vec{\hat{a}}^{(m)} - \vec{\tilde{a}}^{(m)})'\widehat{\Psi}_m(\vec{\hat{a}}^{(m)} - \vec{\tilde{a}}^{(m)}) \\ &= \frac{1}{n^2}\vec{E}'\widehat{\Phi}_m\widehat{\Psi}_m^{-1}(\widehat{\Phi}_m)'\vec{E} \le \|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}\|(\widehat{\Phi}_m)'\vec{E}/n\|_{2,m}^2,\end{aligned}$$

where $\|\vec{x}\|_{2,k}^2 = x_1^2 + \cdots + x_k^2$ denotes the Eucidean norm of the vector $\vec{x} = (x_1,\dots,x_k)'$ in $\mathbb{R}^k$.

On Λ_m, $\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \le \mathfrak{c}n\Delta/(L(m)\log^2(n\Delta))$. Consequently,

$$\mathbb{E}\big[\|\hat{b}_m - \Pi_m b_A\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c}\big] \le \mathfrak{c}\frac{n\Delta}{L(m)\log^2(n\Delta)}\mathbb{E}^{1/2}(\|(\widehat{\Phi}_m)'\vec{E}/n\|_{2,m}^4)\mathbb{P}^{1/2}(\Omega_m^c)$$

We can prove the following Lemma:

Lemma 6.3. *Under the assumptions of Proposition 3.1, we have for $C>0$ a constant,*

$$\mathbb{E}(\|(\widehat{\Phi}_m)'\vec{E}/n\|_{2,m}^4) \le CmL^2(m)(\frac{\Delta^2}{n^3} + \frac{1}{(n\Delta)^2}).$$

Using Lemma 6.3, the bound on $\mathbb{P}(\Omega_m^c)$ given in Lemma 6.1 and the fact that $m \le n\Delta$, we get $\mathbb{E}(\|\hat{b}_m - b_A\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c}) \lesssim 1/(n\Delta)^2$, and $\mathbb{E}(\|b_A\|_n^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c}) \lesssim \mathbb{E}^{1/2}[b^4(X_0)]/(n\Delta)^{5/2}$. Gathering the three bounds and plugging them in (35) implies the first result of Proposition 3.1.

To get the result in $\mathbb{L}^2(\pi)$-norm, we write analogously:

$$\|\widetilde{b}_m - b_A\|_\pi^2 = \|\hat{b}_m - b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m} + \|\hat{b}_m - b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m^c} + \|b_A\|_\pi^2 \mathbf{1}_{\Lambda_m^c}. \tag{43}$$

For any $t\in S_m$, we have using $(x+y)^2 \le (1+1/\theta)x^2 + (1+\theta)y^2$ with $\theta = 4$,

$$\begin{aligned}\|\hat{b}_m - b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m} &\le \frac{5}{4}\|\hat{b}_m - t\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m} + 5\|t-b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m} \\ &\le \frac{5}{2}\|\hat{b}_m - t\|_n^2 \mathbf{1}_{\Lambda_m\cap\Omega_m} + 5\|t-b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m},\end{aligned}$$

by using the definition of Ω_m. We insert b_A again and get:

$$\|\hat{b}_m - b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m} \le 5\|\hat{b}_m - b_A\|_n^2 \mathbf{1}_{\Lambda_m\cap\Omega_m} + 5\|b_A - t\|_n^2 \mathbf{1}_{\Lambda_m\cap\Omega_m} + 5\|t-b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m}.$$

Therefore taking expectation and applying the first result of Proposition 3.1 yield

$$\mathbb{E}\left(\|\hat{b}_m - b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m}\right) \le 45 \inf_{t\in S_m}(\|t-b_A\|_\pi^2) + 5\times 64 \frac{\mathrm{Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}]}{n\Delta} + c\Delta,$$

for c a constant. Next, we study $\mathbb{E}\left(\|\hat{b}_m - b_A\|_\pi^2 \mathbf{1}_{\Lambda_m\cap\Omega_m^c}\right) + \mathbb{E}(\|b_A\|_\pi^2 \mathbf{1}_{\Lambda_m^c})$. For the first term, we write that $\|\hat{b}_m - b_A\|_\pi^2 \le 2(\|\hat{b}_m\|_\pi^2 + \|b_A\|_\pi^2)$ and

$$\|\hat{b}_m\|_\pi^2 = \int\left(\sum_{j=0}^{m-1}\hat{a}_j\varphi_j(x)\right)^2 \pi(x)dx = (\vec{\hat{a}}^{(m)})'\Psi_m\,\vec{\hat{a}}^{(m)} \le \|\Psi_m\|_{\mathrm{op}}\|\vec{\hat{a}}^{(m)}\|_{2,m}^2.$$

First, under $\|\sum_{j=0}^{m-1}\varphi_j^2\|_\infty \le L(m)$, we get

$$\begin{aligned}\|\Psi_m\|_{\mathrm{op}} &= \sup_{\|\vec{x}\|_{2,m}=1}\vec{x}'\Psi_m\vec{x} = \sup_{\|\vec{x}\|_{2,m}=1}\int\left(\sum_{j=0}^{m-1}x_j\varphi_j(u)\right)^2\pi(u)du \\ &\le \sup_{\|\vec{x}\|_{2,m}=1}\int\left(\sum_{j=0}^{m-1}x_j^2\sum_{j=0}^{m-1}\varphi_j^2(u)\right)\pi(u)du \le L(m).\end{aligned}$$

Next, $\|\vec{\hat{a}}^{(m)}\|_{2,m}^2 = (1/n^2)\|\widehat{\Psi}_m^{-1}\widehat{\Phi}'_m\vec{Y}\|_{2,m}^2 \le (1/n^2)\|\widehat{\Psi}_m^{-1}\widehat{\Phi}'_m\|_{\mathrm{op}}^2\|\vec{Y}\|_{2,n}^2$ and

$$\|\widehat{\Psi}_m^{-1}\widehat{\Phi}'_m\|_{\mathrm{op}}^2 = \lambda_{\max}\left(\widehat{\Psi}_m^{-1}\widehat{\Phi}'_m\widehat{\Phi}_m\widehat{\Psi}_m^{-1}\right) = n\lambda_{\max}(\widehat{\Psi}_m^{-1}) = n\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}$$

Therefore, on Λ_m we have $L(m)(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}\vee 1) \le \mathfrak{c}n\Delta$, and thus

$$\|\hat{b}_m\|_\pi^2 \le \frac{L(m)\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}}{n}\left(\sum_{i=1}^n Y_{i\Delta}^2\right) \lesssim \Delta\left(\sum_{i=1}^n Y_{i\Delta}^2\right).$$

Then as $\mathbb{E}[(\sum_{i=1}^n Y_{i\Delta}^2)^2] \le n^2\mathbb{E}(Y_\Delta^4)$, we get

$$\mathbb{E}(\|\hat{b}_m\|_\pi^2\mathbf{1}_{\Lambda_m\cap\Omega_m^c}) \le \sqrt{\mathbb{E}(\|\hat{b}_m\|_\pi^4)\mathbb{P}(\Omega_m^c)} \lesssim \mathbb{E}^{1/2}(Y_\Delta^4)n\Delta\mathbb{P}^{1/2}(\Omega_m^c) \lesssim 1/(n\Delta)$$

as $\mathbb{E}[Y_\Delta^4] \lesssim \Delta^2$. On the other hand $\mathbb{E}(\|b_A\|_\pi^2\mathbf{1}_{\Omega_m^c}) \le \|b_A\|_\pi^2\mathbb{P}(\Omega_m^c) \lesssim 1/(n\Delta)^5$. Thus $\mathbb{E}\left(\|\hat{b}_m - b_A\|_\pi^2\mathbf{1}_{\Omega_m^c}\right) \lesssim 1/(n\Delta)$. Joining the bounds for the three terms of (43) ends the proof of the second Inequality of Proposition 3.1. □

6.4. **Proof of Lemma 6.1.** We use Proposition 6.1, (i) for $u = 1/2$ and condition (10) to get that

$$\mathbb{P}(\Omega_m^c) = \mathbb{P}(\Omega_m(1/2)^c) \le 4n\Delta \exp\left[-\frac{\theta(3\log(3/2)-1)\log n\Delta}{24\mathfrak{c}/2}\right] + \frac{\theta}{6(n\Delta)^5} \lesssim \frac{1}{(n\Delta)^5}$$

for $\mathfrak{c} \le (\theta(3\log(3/2)-1))/(6\times 12)$, see also Cohen *et al.* (2019) for the change of the constant $1-\log(2)$ into $3\log(3/2)-1$. Besides, in Comte and Genon-Catalot (2020a), Lemma 5, it is proved that $\mathbb{P}(\Lambda_m^c) \le \mathbb{P}(\Omega_m^c)$. The proof of Lemma 6.1 is thus complete. $\square$

6.5. **Proof of Lemma 6.2.** • Proof of equality (40).
Write that

$$\sup_{t\in S_m, \|t\|_\pi=1} \nu_n^2(t) = \sup_{t\in S_m, \|\Psi_m^{1/2}\vec{a}\|_{2,m}=1, t=\sum_{j=0}^{m-1} a_j\varphi_j} \langle\overrightarrow{\sigma\varepsilon}, t\rangle_n^2$$

where $\overrightarrow{\sigma\varepsilon}$ is the n-dimensional vector with coordinates $\sigma(X_{i\Delta})\varepsilon_i$, $i = 1, \dots n$ and $\varepsilon_i = (W_{(i+1)\Delta} - W_{i\Delta})/\Delta$. Let $t = \sum_{j=0}^{m-1} a_j\varphi_j$ where $\vec{a} = \Psi_m^{-1/2}\vec{u}$, that is $a_j = \sum_{k=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}u_k$ and $\|\vec{u}\|_{2,m} = 1$. Then $t = \sum_{k=0}^{m-1} u_k(\sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j)$. By the Cauchy-Schwarz inequality, we have $\langle\overrightarrow{\sigma\varepsilon}, t\rangle_n^2 \le \sum_{k=0}^{m-1}\langle\overrightarrow{\sigma\varepsilon}, \sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j\rangle_n^2$, and more precisely,

$$\sup_{t\in S_m, \|\Psi_m^{1/2}\vec{a}\|_{2,m}=1} \langle\overrightarrow{\sigma\varepsilon}, t\rangle_n^2 = \sum_{k=0}^{m-1}\langle\overrightarrow{\sigma\varepsilon}, \sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j\rangle_n^2.$$

Therefore,

$$\mathbb{E}\left(\sup_{t\in S_m, \|\Psi_m^{1/2}\vec{a}\|_{2,m}=1} \langle\overrightarrow{\sigma\varepsilon}, t\rangle_n^2\right) = \sum_{k=0}^{m-1}\mathbb{E}\left(\langle\overrightarrow{\sigma\varepsilon}, \sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j\rangle_n^2\right).$$

Then using that, for any bounded function ψ, $\mathbb{E}[\varepsilon_i\varepsilon_k\psi(X_{i\Delta})\psi(X_{k\Delta})]$ is equal to 0 if $i \ne k$ and equal to $\mathbb{E}(\psi^2(X_0))/\Delta$ if $i = k$, we get

$$\begin{aligned}\mathbb{E}\left(\langle\overrightarrow{\sigma\varepsilon}, \sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j\rangle_n^2\right) &= \frac{1}{n\Delta}\mathbb{E}\left[\sigma^2(X_0)\left(\sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j(X_0)\right)^2\right] \\ &= \frac{1}{n\Delta}\sum_{0\le j,\ell\le m-1}[\Psi_m^{-1/2}]_{j,k}[\Psi_m^{-1/2}]_{\ell,k}[\Psi_{m,\sigma^2}]_{j,\ell}.\end{aligned}$$

We thus obtain equality (40) as

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in S_m, \|\Psi_m^{1/2}\vec{a}\|_{2,m}=1} \langle\overrightarrow{\sigma\varepsilon}, t\rangle_n^2\right) &= \frac{1}{n\Delta}\sum_{0\le j,k,\ell\le m-1}[\Psi_m^{-1/2}]_{j,k}[\Psi_m^{-1/2}]_{\ell,k}[\Psi_{m,\sigma^2}]_{j,\ell} \\ &= \frac{1}{n\Delta}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right].\end{aligned}$$

• Proof of bound (41). Now write

$$\sup_{t\in S_m, \|t\|_\pi=1} R_{n,1}^2(t) = \sup_{t\in S_m, \|\Psi_m^{1/2}\vec{a}\|_{2,m}=1, t=\sum_{j=0}^{m-1} a_j\varphi_j} R_{n,1}^2(t).$$

Let $t = \sum_{j=0}^{m-1} a_j \varphi_j$ where $\vec{a} = \Psi_m^{-1/2}\vec{u}$, $\|\vec{u}\|_{2,m} = 1$. Then (as above):

$$t = \sum_{k=0}^{m-1} u_k \left(\sum_{j=0}^{m-1} [\Psi_m^{-1/2}]_{j,k} \varphi_j \right) \quad \text{and} \quad R_{n,1}^2(t) \le \sum_{k=0}^{m-1} R_{n,1}^2 \left(\sum_{j=0}^{m-1} [\Psi_m^{-1/2}]_{j,k} \varphi_j \right)$$

Therefore,

$$\mathbb{E}\left(\sup_{t \in S_m, \|\Psi_m^{1/2}\vec{a}\|_{2,m}=1} R_{n,1}^2(t) \right) \le \sum_{k=0}^{m-1} \mathbb{E}\left[R_{n,1}^2 \left(\sum_{j=0}^{m-1} [\Psi_m^{-1/2}]_{j,k} \varphi_j \right) \right].$$

Then using that the $\int_{i\Delta}^{(i+1)\Delta} (\sigma(X_s) - \sigma(X_{i\Delta}))dW_s$ and $X_{j\Delta}, j \le i$ are uncorrelated and the terms are centered, and that the process (X_t) is stationary, we get

$$\mathbb{E}\left[R_{n,1}^2 \left(\sum_{j=0}^{m-1} [\Psi_m^{-1/2}]_{j,k} \varphi_j \right) \right] \;=\; \frac{1}{n}\mathbb{E}\left[R_0^{(1)} \left(\sum_{j=0}^{m-1} [\Psi_m^{-1/2}]_{j,k} \varphi_j(X_0) \right)^2 \right].$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left(\sup_{t \in S_m, \|t\|_\pi=1} R_{n,1}^2(t) \right) \le \frac{1}{n} \sum_{k=0}^{m-1} \sum_{0 \le j,\ell \le m-1} [\Psi_m^{-1/2}]_{j,k} [\Psi_m^{-1/2}]_{\ell,k} \mathbb{E}\left(\varphi_j(X_0)\varphi_\ell(X_0)(R_0^{(1)})^2 \right) \\
\le\ & \frac{1}{n} \sum_{0 \le j,\ell \le m-1} \left(\sum_{k=0}^{m-1} [\Psi_m^{-1/2}]_{j,k} [\Psi_m^{-1/2}]_{\ell,k} \right) \mathbb{E}\left(\varphi_j(X_0)\varphi_\ell(X_0)(R_0^{(1)})^2 \right) \\
=\ & \frac{1}{n} \sum_{0 \le j,\ell \le m-1} [\Psi_m^{-1}]_{j,\ell} \mathbb{E}\left(\varphi_j(X_0)\varphi_\ell(X_0)(R_0^{(1)})^2 \right) \\
\le\ & \frac{L(m)\|\Psi_m^{-1}\|_{\text{op}}}{n\Delta^2} \mathbb{E}\left[\left(\int_0^\Delta (\sigma(X_s) - \sigma(X_0))dW_s \right)^2 \right] \lesssim \frac{L(m)\|\Psi_m^{-1}\|_{\text{op}}}{n}
\end{aligned}$$

since by (5) for $f = \sigma$, $\mathbb{E}\left[\int_0^\Delta (\sigma(X_s) - \sigma(X_0))^2 ds \right] \lesssim \Delta^2(1 + \mathbb{E}\eta^2)$. Now, for m satisfying (13), $L(m)\|\Psi_m^{-1}\|_{\text{op}} \le \mathfrak{c}n\Delta/(2\log^2(n\Delta)) \le \mathfrak{c}n\Delta$.

Thus,

$$\mathbb{E}\left(\sup_{t \in S_m, \|t\|_\pi=1} R_{n,1}^2(t) \right) \lesssim \Delta.$$

This ends the proof of (41) and of Lemma 6.2. □

6.6. **Proof of Lemma 6.3.** Set

$$N^2 = \|n^{-1}\widehat{\Phi}'_m \vec{E}\|_{2,m}^2 = n^{-2} \sum_{j=0}^{m-1} (\sum_{i=1}^n \varphi_j(X_{i\Delta})E_{i\Delta})^2$$

where $E_{i\Delta}=\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}(b(X_s)-b(X_{i\Delta}))ds+\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta}\sigma(X_s)dW_s$. We can write $\sum_{i=1}^n \varphi_j(X_{i\Delta})E_{i\Delta}=\int_0^{(n+1)\Delta}H_s^{(j)}ds+\int_0^{(n+1)\Delta}K_s^{(j)}dW_s$ with

$$H_s^{(j)}=\frac{1}{\Delta}\sum_{i=1}^n \mathbf{1}_{]i\Delta,(i+1)\Delta]}(s)\varphi_j(X_{i\Delta})(b(X_s)-b(X_{i\Delta})),\ K_s^{(j)}=\frac{1}{\Delta}\sum_{i=1}^n \mathbf{1}_{]i\Delta,(i+1)\Delta]}(s)\varphi_j(X_{i\Delta})\sigma(X_s).$$

Therefore,

$$\begin{aligned}N^4 &= \frac{1}{(n\Delta)^4}\left[\sum_{j=0}^{m-1}\left(\int_0^{(n+1)\Delta}H_s^{(j)}ds+\int_0^{(n+1)\Delta}K_s^{(j)}dW_s\right)^2\right]^2\\
&\leq 8\frac{m}{(n\Delta)^4}\sum_{j=0}^{m-1}\left[\left(\int_0^{(n+1)\Delta}H_s^{(j)}ds\right)^4+\left(\int_0^{(n+1)\Delta}K_s^{(j)}dW_s\right)^4\right]\\
&\leq 8\frac{m}{n^4}\sum_{j=0}^{m-1}\left[\frac{1}{\Delta}\int_0^{(n+1)\Delta}(H_s^{(j)})^4ds+\frac{1}{\Delta^4}\left(\int_0^{(n+1)\Delta}K_s^{(j)}dW_s\right)^4\right].\end{aligned}$$

We bound successively the expectation of the two terms. By (10), $\sum_{j=0}^{m-1}\varphi_j^4\leq L^2(m)$, so:

$$\begin{aligned}\frac{1}{\Delta}\mathbb{E}\sum_{j=0}^{m-1}\int_0^{(n+1)\Delta}(H_s^{(j)})^4ds &\leq L^2(m)\frac{1}{\Delta}\mathbb{E}\sum_{i=1}^{n}\int_{i\Delta}^{(i+1)\Delta}(b(X_s)-b(X_{i\Delta}))^4ds\\
&\lesssim L^2(m)n\Delta^2(1+\mathbb{E}(\eta^4).\end{aligned}$$

Next, using the Burkholder-Davis-Gundy and the Cauchy-Schwarz inequalities yields

$$\frac{1}{\Delta^4}\mathbb{E}\sum_{j=0}^{m-1}\left(\int_0^{(n+1)\Delta}K_s^{(j)}dW_s\right)^4\lesssim\frac{1}{\Delta^4}\mathbb{E}\sum_{j=0}^{m-1}\left(\int_0^{(n+1)\Delta}(K_s^{(j)})^2ds\right)^2$$

$$\lesssim \frac{1}{\Delta^4}\sum_{j=0}^{m-1}\mathbb{E}\sum_{j=0}^{m-1}n\Delta\int_{\Delta}^{(n+1)\Delta}(K_s^{(j)})^4ds\lesssim\frac{1}{\Delta^4}L^2(m)n\Delta\mathbb{E}\sum_{i=1}^{n}\int_{i\Delta}^{(i+1)\Delta}\sigma^4(X_s)ds$$

$$\lesssim \frac{1}{\Delta^4}L^2(m)(n\Delta)^2\mathbb{E}\sigma^4(\eta)\lesssim\frac{1}{\Delta^4}L^2(m)(n\Delta)^2\mathbb{E}(1+\eta^4)$$

Finally, $\mathbb{E}N^4\lesssim mL^2(m)(\Delta^2n^{-3}+(n\Delta)^{-2})$, which is the result of Lemma 6.3. □

6.7. **Proof of Proposition 3.2.** (1) The result follows from equality (40) and the fact that the spaces are nested.

(2) For the second point, we use: $\mathrm{Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}]\leq m\|\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\|_{\mathrm{op}}$. Then,

$$\|\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\|_{\mathrm{op}}=\sup_{\|x\|_{2,m}=1}x'\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}x=\sup_{y,\|\Psi_m^{1/2}y\|_{2,m}=1}y'\Psi_{m,\sigma^2}y.$$

Now, if σ is bounded on A, $y'\Psi_{m,\sigma^2}y$ is equal to

$$\int(\sum_{j=0}^{m-1}y_j\varphi_j(x))^2\sigma^2(x)\pi(x)dx\leq\|\sigma_A^2\|_\infty\int(\sum_{j=0}^{m-1}y_j\varphi_j(x))^2\pi(x)dx=\|\sigma_A^2\|_\infty\|\Psi_m^{1/2}y\|_{2,m}.$$

Thus, $\mathrm{Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] \leq m\|\sigma_A^2\|_\infty$.
(3) Let us prove the other variance bound. We recall that $\|A\|_{\mathbf{F}}^2 = \mathrm{Tr}(AA') = \mathrm{Tr}(A'A)$. Writing that $\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2} = (\Psi_m^{-1/2}\Psi_{m,\sigma^2}^{1/2})(\Psi_m^{-1/2}\Psi_{m,\sigma^2}^{1/2})'$, we have

$$\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] = \|\Psi_m^{-1/2}\Psi_{m,\sigma^2}^{1/2}\|_F^2$$

Using $\|AB\|_{\mathbf{F}}^2 \leq \|A\|_{\mathbf{F}}^2\|B\|_{\mathrm{op}}^2$,

$$\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] \leq \|\Psi_m^{-1/2}\|_{\mathrm{op}}^2\|\Psi_{m,\sigma^2}^{1/2}\|_F^2 = \|\Psi_m^{-1}\|_{\mathrm{op}}\mathrm{Tr}(\Psi_{m,\sigma^2}). \tag{44}$$

Lastly $\mathrm{Tr}(\Psi_{m,\sigma^2}) = \sum_{j=0}^{m-1}\int \varphi_j^2(x)\sigma^2(x)\pi(x)dx \leq L(m)\mathbb{E}[\sigma_A^2(X_0)]$ gives bound (3). □

6.8. **Proof of Theorem 3.1.** We follow the scheme of Theorem 2 in Comte and Genon-Catalot (2020a). But here, the variables are not independent, the function $\sigma(.)$ is unbounded and there are two other main differences:

- the penalty

$$\widehat{\mathrm{pen}}(m) = \kappa c_\varphi^2 s^2 \frac{m\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}}{n} \tag{45}$$

  is random and has to be compared to its deterministic counterpart, $\mathrm{pen}(m) = \kappa' c_\varphi^2 s^2 m\|\Psi_m^{-1}\|_{\mathrm{op}}/n$,
- there are the two additional terms, $R_{n,1}$ and $R_{n,2}$.

We denote by $\widehat{M}_{n\Delta}$ the maximal element of $\widehat{\mathcal{M}}_{n\Delta}$ defined by (19), by $M_{n\Delta}$ the maximal element of $\mathcal{M}_{n\Delta}$ defined by (18) and by $M_{n\Delta}^+$ the maximal element of the set defined by

$$\mathcal{M}_{n\Delta}^+ = \left\{m \in \mathbb{N}, \quad c_\varphi^2 m(\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \vee 1) \leq 4\mathfrak{d}\frac{n\Delta}{\log^2(n\Delta)}\right\}, \text{ with } \mathfrak{d} \text{ is given in (18).} \tag{46}$$

The value $\widehat{M}_{n\Delta}$ is random but thanks to the constants associated with the sets, with large probability, we prove $M_{n\Delta} \leq \widehat{M}_{n\Delta} \leq M_{n\Delta}^+$ or equivalently $\mathcal{M}_{n\Delta} \subset \widehat{\mathcal{M}}_{n\Delta} \subset \mathcal{M}_{n\Delta}^+$.

Set

$$\Xi_{n\Delta} := \left\{\mathcal{M}_{n\Delta} \subset \widehat{\mathcal{M}}_{n\Delta} \subset \mathcal{M}_{n\Delta}^+\right\}, \quad \Omega_{n\Delta} = \cap_{m\in\mathcal{M}_{n\Delta}^+}\Omega_m. \tag{47}$$

Lemma 6.4. *Under the assumptions of Theorem 3.1,* $\mathbb{P}(\Omega_{n\Delta}^c) \leq c/(n\Delta)^4$ *and* $\mathbb{P}(\Xi_{n\Delta}^c) \leq c'/(n\Delta)^4$*, where* c*,* c' *are positive constants.*

We do not give a detailed proof of this Lemma. As $4\mathfrak{d} \leq \mathfrak{c}/2$, the first bound of Lemma 6.4 is a simple consequence of Lemma 6.1. The proof of the second bound is not immediate but quite similar to the one of Lemma 7 in Comte and Genon-Catalot (2020a). The order obtained is different due to larger constants $\mathfrak{c}$, $\mathfrak{d}$ in the present problem. Lemma 6.4 relies on Inequality `(ii)` of Proposition 6.1 and this is the only place where this inequality is applied.

Now we write the decomposition:

$$\begin{aligned}\hat{b}_{\hat{m}} - b_A &= (\hat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_{n\Delta}} + (\hat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_{n\Delta}^c} \\ &= (\hat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}} + (\hat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}^c} + (\hat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_{n\Delta}^c}\end{aligned} \tag{48}$$

Lemma 6.5. *Under the assumptions of Theorem 3.1,* $\mathbb{E}\left[\|\hat{b}_{\hat{m}} - b_A\|_n^4\right] \leq c(n\Delta)^2$.

Applying Lemma 6.4, we get

$$\mathbb{E}\left[\|\hat{b}_{\hat{m}} - b_A\|_n^2(\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}^c} + \mathbf{1}_{\Xi_{n\Delta}^c})\right] \leq \frac{c}{n\Delta}.$$

Therefore it remains to study $\mathbb{E}(\|\hat{b}_{\hat{m}} - b_A\|_n^2\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}})$. We have

$$\hat{m} = \arg\min_{m\in\widehat{\mathcal{M}}_{n\Delta}} \{\gamma_n(\hat{b}_m) + \widehat{\text{pen}}(m)\},$$

with $\widehat{\text{pen}}(m)$ defined by (45). Thus, using the definition of the contrast, we have, for any $m \in \widehat{\mathcal{M}}_{n\Delta}$, and any $b_m \in S_m$,

$$\gamma_n(\hat{b}_{\hat{m}}) + \widehat{\text{pen}}(\hat{m}) \leq \gamma_n(b_m) + \widehat{\text{pen}}(m). \tag{49}$$

Now, on the set $\Xi_{n\Delta} = \left\{\mathcal{M}_{n\Delta} \subset \widehat{\mathcal{M}}_{n\Delta} \subset \mathcal{M}_{n\Delta}^+\right\}$, we have in all cases that $\hat{m} \leq \widehat{M}_{n\Delta} \leq M_{n\Delta}^+$ and either $M_{n\Delta} \leq \hat{m} \leq \widehat{M}_{n\Delta} \leq M_{n\Delta}^+$ or $\hat{m} < M_{n\Delta} \leq \widehat{M}_{n\Delta} \leq M_{n\Delta}^+$. In the first case, $\hat{m}$ is upper and lower bounded by deterministic bounds, and in the second,

$$\hat{m} = \arg\min_{m\in\mathcal{M}_{n\Delta}} \{\gamma_n(\hat{b}_m) + \widehat{\text{pen}}(m)\}.$$

Thus, on $\Xi_{n\Delta}$, Inequality (49) holds for any $m \in \mathcal{M}_{n\Delta}$ and any $b_m \in S_m$. With decomposition (36), it yields, for any $m \in \mathcal{M}_{n\Delta}$ and any $b_m \in S_m$, on $\Xi_{n\Delta} \cap \Omega_{n\Delta}$,

$$\begin{aligned}
\|\hat{b}_{\hat{m}} - b\|_n^2 &\leq \|b_m - b\|_n^2 + \frac{1}{8}\|\hat{b}_{\hat{m}} - b_m\|_\pi^2 + 16\sup_{t\in B_{\hat{m},m}^\pi(0,1)} \nu_n^2(t) + \widehat{\text{pen}}(m) - \widehat{\text{pen}}(\hat{m}) \\
&\quad +16\sup_{t\in B_{\hat{m},m}^\pi(0,1)} R_{n,1}^2(t)\mathbf{1}_{\Xi_{n\Delta}} + 2R_{n,2}(\hat{b}_{\hat{m}} - b) \\
&\leq \left(1 + \frac{1}{2}\right)\|b_m - b\|_n^2 + \frac{1}{2}\|\hat{b}_{\hat{m}} - b\|_n^2 + 16\Big(\sup_{t\in B_{\hat{m},m}^\pi(0,1)} \nu_n^2(t) - p(m,\hat{m})\Big)_+ \\
&\quad +16\sup_{t\in B_{\hat{m},m}^\pi(0,1)} R_{n,1}^2(t)\mathbf{1}_{\Xi_{n\Delta}} + 2R_{n,2}(\hat{b}_{\hat{m}} - b_m) + \widehat{\text{pen}}(m) + 16p(m,\hat{m}) - \widehat{\text{pen}}(\hat{m}).
\end{aligned} \tag{50}$$

where $B_{m,m'}^\pi(0,1) = \{t \in S_m + S_{m'}, \|t\|_\pi = 1\}$ and $p(m,m')$ is defined by

$$p(m,m') = \kappa_0 s^2 c_\varphi^2 \frac{(m \vee m')\|\Psi_{m\vee m'}^{-1}\|_{\text{op}}}{n\Delta}, \quad s^2 = \mathbb{E}[\sigma^2(X_0)]. \tag{51}$$

Then, for $\hat{m}$ a random index in $\widehat{\mathcal{M}}_{n\Delta}$, using (41),

$$\mathbb{E}(\sup_{t\in S_m+S_{\hat{m}},\|t\|_\pi=1} R_{n,1}^2(t)\mathbf{1}_{\Xi_{n\Delta}}) \leq \mathbb{E}(\sup_{t\in S_{M_n^+},\|t\|_\pi=1} R_{n,1}^2(t)) \leq C\,\Delta.$$

The bound on $R_{n,2}$ is straightforward (see the proof of Proposition 3.1, non adaptive case) and we get

$$\mathbb{E}\left[R_{n,2}(\hat{b}_{\hat{m}} - b_m)\mathbf{1}_{\Xi_{n\Delta}}\right] \leq \frac{1}{8}\mathbb{E}[\|\hat{b}_{\hat{m}} - b_m\|_n^2\mathbf{1}_{\Xi_{n\Delta}}] + 8c'\Delta. \tag{52}$$

The main point is the study of $\nu_n(t)$.

Lemma 6.6. *Let $(X_{i\Delta}, i=1,\dots,n)$ be observations from model (1) under* **(A1)-(A4)**, *with basis satisfying* **(B1)**. *Assume that $\mathbb{E}\eta^6<+\infty$. Then there exists κ_0 such that $\nu_n(t)$ satisfies*

$$\mathbb{E}\left[\Big(\sup_{t\in B^{\pi}_{\hat m,m}(0,1)}\nu_n^2(t)-p(m,\hat m)\Big)_+\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}\right]\le C\frac{\log^2(n\Delta)}{n\Delta}$$

where $p(m,m')$ is defined by (51).

For $\kappa'\ge 16\kappa_0$, $16p(m,m')\le \mathrm{pen}(m)+\mathrm{pen}(m')$. Therefore, plugging the result of Lemma 6.6 and (52) in (50) and taking expectation yield that

$$\begin{aligned}\frac14\mathbb{E}(\|\hat b_{\hat m}-b\|_n^2\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}})\le&\frac74\|b_m-b\|_n^2+\mathrm{pen}(m)+C\frac{\log^2(n\Delta)}{n\Delta}+C'\Delta\\&+\mathbb{E}(\widehat{\mathrm{pen}}(m)\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}})+\mathbb{E}[(\mathrm{pen}(\hat m)-\widehat{\mathrm{pen}}(\hat m))_+\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}).\end{aligned}$$

Lemma 6.7. *Under the assumptions of Theorem 3.1, there exist constants $c_1,c_2>0$ such that for $m\in\mathcal{M}_{n\Delta}$ and $\hat m\in\widehat{\mathcal{M}}_{n\Delta}$,*

$$\mathbb{E}(\widehat{\mathrm{pen}}(m)\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}})\le c_1\mathrm{pen}(m)+\frac{c_2}{n\Delta}\tag{53}$$

$$\mathbb{E}[(\mathrm{pen}(\hat m)-\widehat{\mathrm{pen}}(\hat m))_+\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}})\le\frac{c_2}{n\Delta}.\tag{54}$$

Lemma 6.7 concludes the study of the expectation of the empirical risk on $\Xi_{n\Delta}\cap\Omega_{n\Delta}$. The proof of (21) is now complete. For the step from the empirical norm to the $\mathbb{L}^2(\pi)$-norm, we proceed as in the proof of Proposition 3.1 and get (22). □

6.9. **Proof of Lemma 6.5.** We start as in (42) with m replaced by $\hat m$. We have $\|\hat b_{\hat m}-\Pi_{\hat m}b_A\|_n^2=(1/n^2)\vec E'\widehat\Phi_{\hat m}\widehat\Psi_{\hat m}^{-1}\widehat\Phi'_{\hat m}\vec E\le\|\widehat\Psi_{\hat m}^{-1}\|_{\mathrm{op}}\|\widehat\Phi'_{\hat m}\vec E/n\|^2_{2,\hat m}$. Now as $\hat m\in\widehat{\mathcal{M}}_{n\Delta}$, $\|\widehat\Psi_{\hat m}^{-1}\|_{\mathrm{op}}\lesssim\sqrt{n\Delta}$. Moreover, $m\mapsto\|\widehat\Phi'_m\vec E/n\|^2_{2,m}$ is increasing, so

$$\|\hat b_{\hat m}-\Pi_{\hat m}b_A\|_n^2\le\sqrt{n\Delta}\|\widehat\Phi'_{n\Delta}\vec E/n\|^2_{2,n\Delta}.$$

Thus, using the bound proved in Lemma 6.3, we get $\mathbb{E}(\|\hat b_{\hat m}-\Pi_{\hat m}b_A\|_n^4)\lesssim(n\Delta)^2$. □

6.10. **Proof of Lemma 6.6.** To apply the Talagrand Inequality (see Theorem 6.2 in appendix), we make the following decompositions. Set $u_i=W_{(i+1)\Delta}-W_{i\Delta}$, and let k_n and ℓ_n be integers given below (see formula (56) and (57)). Then, define

$$u_i^{(1)}=u_i\mathbf{1}_{|u_i|\le k_n\sqrt\Delta}-\mathbb{E}\big[u_i\mathbf{1}_{|u_i|\le k_n\sqrt\Delta}\big],\quad u_i^{(2)}=u_i-u_i^{(1)},$$

and set $\tau(x)=\sigma_A(x)\mathbf{1}_{\sigma_A^2(x)\le\ell_n\sqrt\Delta}$ and $\theta(x)=\sigma_A(x)-\tau(x)$. We have $\nu_n(t)=\nu_{n,1}(t)+\nu_{n,2}(t)+\nu_{n,3}(t)$, where

$$\nu_{n,1}(t)=\frac{1}{n\Delta}\sum_{i=1}^n t(X_{i\Delta})\tau(X_{i\Delta})u_i^{(1)},\quad \nu_{n,2}(t)=\frac{1}{n\Delta}\sum_{i=1}^n t(X_{i\Delta})\theta(X_{i\Delta})u_i^{(1)},\text{ and}$$

$$\nu_{n,3}(t)=\frac{1}{n\Delta}\sum_{i=1}^n t(X_{i\Delta})\sigma(X_{i\Delta})u_i^{(2)}.$$

Then we write

$$\Big(\sup_{t\in B^{\pi}_{\hat m,m}(0,1)}\nu_n^2(t)-p(m,\hat m)\Big)_+ \leq \Big(\sup_{t\in B^{\pi}_{\hat m,m}(0,1)}3\nu_{n,1}^2(t)-p(m,\hat m)\Big)_+ \tag{55}$$
$$+3\sup_{t\in B^{\pi}_{\hat m,m}(0,1)}\nu_{n,2}^2(t)+3\sup_{t\in B^{\pi}_{\hat m,m}(0,1)}\nu_{n,3}^2(t),$$

and we bound the three terms.

• First, we study the second term in (55). Recall that $M^+_{n\Delta}\leq 4\mathfrak{d}n\Delta/\log^2(n\Delta)$ is the dimension of the largest space of the collection $\mathcal{M}^+_{n\Delta}$. We proceed as in the proof of Lemma 6.2, bound (41), to obtain:

$$\mathbb{E}\Big[\Big(\sup_{t\in B^{\pi}_{\hat m,m}(0,1)}\nu_{n,2}^2(t)\Big)_+\mathbf{1}_{\Xi_{n\Delta}}\Big]\leq \|\Psi^{-1}_{M^+_{n\Delta}}\|_{\mathrm{op}}\sum_{j=0}^{M^+_{n\Delta}-1}\mathbb{E}\big[\nu_{n,2}^2(\varphi_j)\big]$$
$$=\|\Psi^{-1}_{M^+_{n\Delta}}\|_{\mathrm{op}}\sum_{j=0}^{M^+_{n\Delta}-1}\mathrm{Var}\Big(\frac{1}{n\Delta}\sum_{i=1}^n u_i^{(1)}\theta(X_{i\Delta})\varphi_j(X_{i\Delta})\Big)\leq\frac{c_\varphi^2M^+_{n\Delta}\|\Psi^{-1}_{M^+_{n\Delta}}\|_{\mathrm{op}}}{n\Delta^2}\mathbb{E}[(u_1^{(1)})^2]\mathbb{E}[\theta_A^2(X_0)]$$

Now we use that $\mathbb{E}\big[(u_1^{(1)})^2\big]\leq\mathbb{E}[u_1^2]=\Delta$ and that $M^+_{n\Delta}$ is in $\mathcal{M}^+_{n\Delta}$, i.e.

$$c_\varphi^2M^+_{n\Delta}\|\Psi^{-1}_{M^+_{n\Delta}}\|_{\mathrm{op}}=c_\varphi^2\sqrt{M^+_{n\Delta}}\sqrt{M^+_{n\Delta}\||\Psi^{-1}_{M^+_{n\Delta}}\|^2_{\mathrm{op}}}\leq 4\mathfrak{d}\frac{n\Delta}{\log^2(n\Delta)}$$

and we get

$$\mathbb{E}\Big[\Big(\sup_{t\in B^{\pi}_{\hat m,m}(0,1)}\nu_{n,2}^2(t)\Big)_+\Big]\leq c_\varphi^2\frac{1}{\log^2(n\Delta)}\mathbb{E}\big[\sigma_A^2(X_0)\mathbf{1}_{\sigma_A^2(X_0)>\ell_n\sqrt{\Delta}}\big]\leq c_\varphi^2\frac{\mathbb{E}\big[|\sigma_A(X_0)|^{2+q}\big]}{\log^2(n\Delta)(\ell_n\sqrt{\Delta})^{q/2}}$$
$$\leq\frac{c_\varphi^2}{c_\star^{q/2}}\mathbb{E}\big[|\sigma_A(X_0)|^{2+q}\big]\frac{\log^{q-2}(n\Delta)}{(n\Delta)^{q/4}}=\frac{c_\varphi^2}{c_\star^2}\mathbb{E}\big[|\sigma_A(X_0)|^6\big]\frac{\log^2(n\Delta)}{n\Delta},$$

by taking $q=4$, and

$$\ell_n=\Big[c_\star\frac{\sqrt{n}}{\log^2(n\Delta)}\Big], \tag{56}$$

where $[x]$ denotes the integer part of x and $c_\star$ is to be chosen later.

• Let us now study the third term in (55). We have, relying on similar arguments,

$$\mathbb{E}\Big[\Big(\sup_{t\in B^{f}_{\hat m,m}(0,1)}\nu_{n,3}^2(t)\mathbf{1}_{\Xi_n}\Big)_+\Big]\leq\|\Psi^{-1}_{M^+_{n\Delta}}\|_{\mathrm{op}}\sum_{j=0}^{M^+_{n\Delta}-1}\mathbb{E}\big[\nu_{n,3}^2(\varphi_j)\big]$$
$$=\|\Psi^{-1}_{M^+_{n\Delta}}\|_{\mathrm{op}}\sum_{j=0}^{M^+_{n\Delta}-1}\mathrm{Var}\Big(\frac{1}{n\Delta}\sum_{i=1}^n u_i^{(2)}\sigma_A(X_{i\Delta})\varphi_j(X_{i\Delta})\Big)\leq\frac{c_\varphi^2M^+_{n\Delta}\|\Psi^{-1}_{M^+_{n\Delta}}\|_{\mathrm{op}}}{n\Delta^2}\mathbb{E}[\sigma_A^2(X_0)]\mathbb{E}\big[(u_1^{(2)})^2\big]$$
$$\leq\frac{c_\varphi^2\mathbb{E}[\sigma_A^2(X_0)]}{\Delta\log^2(n\Delta)}\mathbb{E}\big[u_1^2\mathbf{1}_{|u_1|>k_n\sqrt{\Delta}}\big]\leq\frac{c_\varphi^2\mathbb{E}[\sigma_A^2(X_0)]\mathbb{E}\big[u_1^6\big]}{\Delta\log^2(n\Delta)(k_n\sqrt{\Delta})^4}\lesssim\mu_6\frac{\log^2(n\Delta)}{n\Delta},$$

where the last line follows from the Markov inequality, $\mu_6 = 15 = \mathbb{E}[u_1^6/\Delta^3]$ (sixth moment of the standard gaussian) and the choices $p = 4$ and for $c'_\star$ to be chosen later,

$$k_n = \left[c'_\star \frac{(n\Delta)^{1/4}}{\log(n\Delta)} \right]. \tag{57}$$

• To bound the first term, we use the Talagrand inequality (see Theorem 6.2 in appendix) applied to the process $\nu_{n,1}$. As the variables are not independent we must split again this term into several parts.

We proceed by the coupling strategy used in the proof of Proposition 6.1, applied to $v_i = (u_i, X_{i\Delta})$ which is also a β-mixing sequence with mixing coefficient such that $\beta_k = \beta_X(k\Delta) \lesssim e^{-\theta k\Delta}$, as in Baraud *et al.* (2001). We denote by $\Omega^\star = \{v_i = v_i^\star,\ i = 1, \dots, n\}$. We still have $\mathbb{P}((\Omega^\star)^c) \le p_n \beta_X(q_n\Delta) \lesssim 1/(n\Delta)^4$ for $q_n\Delta = 5\log(n\Delta)/\theta$.

On $\Omega^\star$, we replace the v_i by the $v_i^\star$ and split the term between odd and even blocks. We have to bound, say

$$\mathbb{E}\Big(\sup_{t \in B^\pi_{\hat m, m}(0,1)} (\nu^{\star,1}_{n,1})^2(t) - \frac{1}{6} p(m, \hat m) \Big)_+$$

by using Talagrand inequality (see Theorem 6.2 in appendix) applied to mean of p_n independent random variables

$$\nu^{\star,1}_{n,1}(t) = \frac{1}{p_n} \sum_{\ell=0}^{p_n-1} \left(\frac{1}{2q_n\Delta} \sum_{r=1}^{q_n} u^{(1)\star}_{2\ell q_n + r} \tau(X^\star_{(2\ell q_n + r)\Delta}) t(X^\star_{(2\ell q_n + r)\Delta}) \right).$$

Note that the random variables inside the sum in large brackets are not independent but uncorrelated.

Set $Y_\ell = (u^{(1)\star}_{\ell,1}, X^*_{\ell,1}) \in \mathbb{R}^{q_n} \times \mathbb{R}^{q_n}$, where $u^{(1)\star}_{\ell,1} = (u^{(1)\star}_{2\ell q_n + r})_{1 \le r \le q_n}$ and $X^*_{\ell,1} = (X^\star_{(2\ell q_n + r)\Delta})_{1 \le r \le q_n}$. Then we have $\nu^{\star,1}_{n,1}(t) = \frac{1}{p_n} \sum_{\ell=0}^{p_n-1} [f^{(t)}(Y_\ell) - \mathbb{E} f^{(t)}(Y_\ell)]$, $\quad f^{(t)} : \mathbb{R}^{q_n} \times \mathbb{R}^{q_n} \to \mathbb{R}$ with

$$f^{(t)}(z, x) = \frac{1}{2q_n\Delta} \sum_{r=1}^{q_n} z_r \mathbf{1}_{|z_r| \le k_n\sqrt{\Delta}} t(x_r)\tau(x_r), \quad z = (z_1, \dots, z_{q_n}), x = (x_1, \dots, x_{q_n}),$$

and $\mathcal{F} = \{f^{(t)}, t \in B^\pi_{m',m}(0,1)\}$. Using analogous tools as above, we get

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{t \in B^\pi_{m',m}(0,1)} [\nu^{\star,1}_{n,1}(t)]^2 \Big] \le \|\Psi^{-1}_{m \vee m'}\|_{\rm op} \sum_{j=0}^{(m-1)\vee(m'-1)} \frac{1}{n\Delta^2} \mathrm{Var}\Big(u_1^{(1)} \tau(X_0)\varphi_j(X_0) \Big) \\
\le\ & \|\Psi^{-1}_{m \vee m'}\|_{\rm op} \sum_{j=0}^{(m-1)\vee(m'-1)} \frac{1}{n\Delta^2} \mathbb{E}\left[\Big(u_1^{(1)} \tau(X_0)\varphi_j(X_0) \Big)^2 \right] \\
\le\ & \frac{\mathbb{E}[u_1^2] \|\Psi^{-1}_{m \vee m'}\|_{\rm op}}{n\Delta^2} \sum_{j=0}^{(m-1)\vee(m'-1)} \mathbb{E}\left[\tau^2(X_0)\varphi_j^2(X_0) \right] \le \mathbb{E}[\sigma_A^2(X_0)] c_\varphi^2 \frac{(m \vee m')\|\Psi^{-1}_{m \vee m'}\|_{\rm op}}{n\Delta} := H^2.
\end{aligned}$$

Next, we have

$$
\begin{aligned}
& \sup_{t\in B^\pi_{m,m'}(0,1)} \mathrm{Var}\left(\frac{1}{q_n\Delta}\sum_{r=1}^{q_n} u_r^* t(X^*_{r\Delta})\sigma_A(X^*_{r\Delta})\right) \\
= & \sup_{t\in B^\pi_{m,m'}(0,1)} \mathrm{Var}\left(\frac{1}{q_n\Delta}\sum_{r=1}^{q_n} u_r t(X^*_{r\Delta})\sigma_A(X^*_{r\Delta})\right) = \frac{\mathbb{E}(u_1^2)}{q_n\Delta^2}\sup_{t\in B^\pi_{m',m}(0,1)} \mathbb{E}(t^2(X_0)\sigma^2(X_0)) \\
\leq & \frac{\mathbb{E}(\varepsilon_1^2)}{q_n\Delta^2}\sup_{t\in B^\pi_{m',m}(0,1)} \mathbb{E}^{1/2}\left[\sigma_A^4(X_0)\right]\mathbb{E}^{1/2}[t^2(X_0)]\|t\|_\infty \\
\leq & \frac{c_\varphi}{q_n\Delta}\mathbb{E}^{1/2}\left[\sigma_A^4(X_0)\right]\sqrt{(m\vee m')\|\Psi^{-1}_{m\vee m'}\|_{\mathrm{op}}} := v
\end{aligned}
$$

Lastly

$$
\begin{aligned}
& \sup_{t\in B^\pi_{m',m}(0,1)}\sup_{\vec z\in\mathbb{R}^{q_n},\vec x\in\mathbb{R}^{q_n}}\left(\frac{1}{q_n\Delta}\sum_{r=1}^{q_n}|z_r|\mathbf{1}_{|z_r|\leq k_n\sqrt{\Delta}}|\sigma(x_r)\mathbf{1}_{|\sigma^2(x_r)|\leq \ell_n\sqrt{\Delta}}|t(x_r)|\right) \\
\leq & \frac{k_n\sqrt{\ell_n}}{\Delta^{1/4}}\sup_{t\in B^\pi_{m',m}(0,1)}\sup_x |t(x)| \leq c_\varphi\frac{k_n\sqrt{\ell_n}}{\Delta^{1/4}}\sqrt{(m\vee m')\|\Psi^{-1}_{m\vee m'}\|_{\mathrm{op}}} := M.
\end{aligned}
$$

Therefore, by applying Theorem 6.2 (Talagrand Inequality) recalled in section 6.12:

$$
\mathbb{E}\Big(\sup_{t\in B^\pi_{m,m'}(0,1)}(\nu^{\star,1}_{n,1})^2(t) - 2H^2\Big)_+ \leq C_1\left(\frac{v}{p_n}\exp(-C_2\frac{p_nH^2}{v}) + \frac{M^2}{p_n^2}\exp(-C_3\frac{p_nH}{M}\right)
$$

we obtain, recalling that $2p_nq_n = n$ and $q_n = (5/\theta)(\log(n\Delta)/\Delta)$, and $m^* = m\vee m'$,

$$
\begin{aligned}
& \mathbb{E}\Big(\sup_{t\in B^\pi_{m',m}(0,1)}(\nu^{\star,1}_{n,1})^2(t) - 2H^2\Big)_+ \\
\leq & C_1'\left(\frac{\sqrt{m^*\|\Psi^{-1}_{m^*}\|_{\mathrm{op}}}}{p_nq_n\Delta}\exp(-C_2'\sqrt{m^*\|\Psi^{-1}_{m^*}\|_{\mathrm{op}}}) + \frac{k_n^2\sqrt{\ell_n}m^*\|\Psi^{-1}_{m^*}\|_{\mathrm{op}}}{p_n^2\sqrt{\Delta}}\exp\left(-C_3'\frac{p_n}{\sqrt{n\Delta}k_n\sqrt{\ell_n}}\right)\right),
\end{aligned}
$$

Now we use that, for the first right-hand-side term, $\sqrt{x}e^{-C_2\sqrt{x}} \leq c'e^{-(C_2/2)\sqrt{x}}$ and $c_\varphi^2 m^*\|\Psi^{-1}_{m^*}\|_{\mathrm{op}} \leq 4\mathfrak{d}n\Delta/\log^2(n\Delta)$. For the second right-hand-side term, we use the definition (56) and (57) of ℓ_n and k_n, the value of q_n and $c_\varphi^2 m^*\|\Psi^{-1}_{m^*}\|_{\mathrm{op}} \leq 4\mathfrak{d}n\Delta/\log^2(n\Delta)$. This implies

$$
\mathbb{E}\Big(\sup_{t\in B^\pi_{m',m}(0,1)}[\nu^{\star,1}_{n,1}]^2(t) - \frac{1}{6}p(m,m')\Big)_+ \lesssim \frac{1}{n\Delta}\left(e^{-(C_2'/2)\sqrt{m^*\|\Psi^{-1}_{m^*}\|_{\mathrm{op}}}} + \frac{n\sqrt{\Delta}}{\log^4(n\Delta)}e^{-C_3''\log(n\Delta)}\right).
$$

where $p(m,m') = 12H^2$. Next note that $\|\Psi_m^{-1}\|_{\mathrm{op}} \geq 1/\|\Psi_m\|_{\mathrm{op}} \geq 1/\|\pi\|_\infty$, and choose $c_\star, c_\star'$ in the definition of k_n and ℓ_n so that $C_3'' = 2$. This yields

$$
\mathbb{E}\Big(\sup_{t\in B^\pi_{m',m}(0,1)}[\nu^{\star,1}_{n,1}]^2(t) - \frac{1}{6}p(m,m')\Big)_+ \lesssim \frac{1}{n\Delta}\left(\exp(-C_4\sqrt{m^*}) + \frac{1}{n\log^4(n\Delta)}\right).
$$

By summing up all terms over $m' \in \mathcal{M}_{n\Delta}^+$, we deduce

$$\mathbb{E}\Big(\sup_{t\in B_{\hat{m},m}^{\pi}(0,1)}[\nu_{n,1}^{\star,1}]^2(t)-\frac{1}{6}p(m,\hat{m})\Big)_+ \leq \sum_{m'}\mathbb{E}\Big(\sup_{t\in B_{m',m}^{\pi}(0,1)}[\nu_{n,1}^{\star,1}]^2(t)-\frac{1}{6}p(m,m')\Big)_+$$

$$\lesssim \frac{1}{n\Delta}. \tag{58}$$

It remains to bound $\mathbb{E}[(\sup_{t\in B_{\hat{m},m}^{\pi}(0,1)}(\nu_{n,1})^2(t)-p(m,\hat{m}))\mathbf{1}_{(\Omega^\star)^c}]_+$. We use the infinite norm computed to evaluate M and the bound on $\mathbb{P}[(\Omega^\star)^c]$. □

6.11. **Proof of Lemma 6.7.** First write that

$$\widehat{\mathrm{pen}}(m) \leq \kappa c_\varphi^2 s^2 \frac{m\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}}{n\Delta}+\frac{\kappa}{\kappa'}\mathrm{pen}(m)$$

Moreover, for $m\in\mathcal{M}_{n\Delta}$ and on $\Xi_{n\Delta}$ (thus $m\in\widehat{\mathcal{M}}_{n\Delta}$), $c_\varphi^2 m\|\Psi_m^{-1}\|_{\mathrm{op}} \leq (\mathfrak{d}/4)n\Delta/\log^2(n\Delta)$ and $c_\varphi^2 m\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \leq \mathfrak{d}n\Delta/\log^2(n\Delta)$. Thus,

$$\begin{aligned} c_\varphi^2 m\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}} &= c_\varphi^2 m\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}\mathbf{1}_{\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}\leq\|\Psi_m^{-1}\|_{\mathrm{op}}\}} \\ &\quad + c_\varphi^2 m\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}\mathbf{1}_{\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\}} \\ &\leq c_\varphi^2 m\|\Psi_m^{-1}\|_{\mathrm{op}}\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}+\frac{5}{4}\frac{\mathfrak{d}n\Delta}{\log^2(n\Delta)}\mathbf{1}_{\{\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\}} \end{aligned}$$

We obtain:

$$\mathbb{E}(\widehat{\mathrm{pen}}(m)\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}) \leq 2\frac{\kappa}{\kappa'}\mathrm{pen}(m)+\frac{5}{4}\frac{\mathfrak{d}n\Delta}{\log^2(n\Delta)}\mathbb{P}\left(\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\right).$$

Now by Proposition 2.4 in Comte and Genon-Catalot (2018),

$$\mathbb{P}\left(\|\widehat{\Psi}_m^{-1}-\Psi_m^{-1}\|_{\mathrm{op}}>\|\Psi_m^{-1}\|_{\mathrm{op}}\right)\leq\mathbb{P}\left(\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-\mathrm{Id}_m\|_{\mathrm{op}}>\frac{1}{2}\right)\lesssim 1/(n\Delta)^5$$

for $m\in\mathcal{M}_{n\Delta}$, by Lemma 6.1. This completes the proof of (53).
Now we turn to the proof of (54). Writing that $\|\widehat{\Psi}_{\hat{m}}^{-1}\|_{\mathrm{op}}\geq\|\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}-\|\widehat{\Psi}_{\hat{m}}^{-1}-\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}$, we get

$$\mathrm{pen}(\hat{m})-\widehat{\mathrm{pen}}(\hat{m})\leq\kappa c_\varphi^2 s^2\frac{\hat{m}\|\widehat{\Psi}_{\hat{m}}^{-1}-\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}}{n\Delta}+(\kappa'-\kappa)\frac{s^2c_\varphi^2\hat{m}\|\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}}{n}.$$

Next we decompose similarly to previously, with a change in the cutoff,

$$\begin{aligned} c_\varphi^2 \hat{m}\|\widehat{\Psi}_{\hat{m}}^{-1}-\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}} &= c_\varphi^2 \hat{m}\|\widehat{\Psi}_{\hat{m}}^{-1}-\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}\mathbf{1}_{\{\|\widehat{\Psi}_{\hat{m}}^{-1}-\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}\leq\frac{1}{2}\|\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}\}} \\ &\quad + c_\varphi^2 \hat{m}\|\widehat{\Psi}_{\hat{m}}^{-1}-\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}\mathbf{1}_{\{\|\widehat{\Psi}_{\hat{m}}^{-1}-\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}>\frac{1}{2}\|\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}\}} \\ &\leq \frac{1}{2}c_\varphi^2 \hat{m}\|\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}\mathbf{1}_{\Xi_{n\Delta}\cap\Omega_{n\Delta}}+5\frac{\mathfrak{d}n\Delta}{\log^2(n\Delta)}\mathbf{1}_{\{\|\widehat{\Psi}_{\hat{m}}^{-1}-\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}>\frac{1}{2}\|\Psi_{\hat{m}}^{-1}\|_{\mathrm{op}}\}}. \end{aligned}$$

Now, $\hat{m} \in \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+$, implies that $c_\varphi^2 \hat{m} \|\Psi_{\hat{m}}^{-1}\|_{\text{op}} \leq 4\mathfrak{d} n \log^2(n\Delta)$. We get

$$\begin{aligned}(\text{pen}(\hat{m}) - \widehat{\text{pen}}(\hat{m}))_+ \mathbf{1}_{\Xi_{n\Delta} \cap \Omega_{n\Delta}} \quad &\leq \quad 4\kappa c_\varphi^2 s^2 \frac{\mathfrak{d} n\Delta}{\log^2(n\Delta)} \mathbf{1}_{\{\|\widehat{\Psi}_{\hat{m}}^{-1} - \Psi_{\hat{m}}^{-1}\|_{\text{op}} > \frac{1}{2}\|\Psi_{\hat{m}}^{-1}\|_{\text{op}}\}} \mathbf{1}_{\Xi_{n\Delta} \cap \Omega_{n\Delta}} \\ &\quad + (\kappa' - \frac{\kappa}{2}) \frac{s^2 c_\varphi^2 \hat{m} \|\Psi_{\hat{m}}^{-1}\|_{\text{op}}}{n}.\end{aligned}$$

We choose $\kappa' - \frac{\kappa}{2} \leq 0$ that is $\kappa \geq 2\kappa'$, so that the last term vanishes and then obtain:.

$$\begin{aligned}&\mathbb{E}\left[(\text{pen}(\hat{m}) - \widehat{\text{pen}}(\hat{m}))_+ \mathbf{1}_{\Xi_{n\Delta} \cap \Omega_{n\Delta}}\right] \\ \leq \quad & 4\kappa c_\varphi^2 s^2 \frac{\mathfrak{d} n\Delta}{\log^2(n\Delta)} \mathbb{P}\left(\{\|\widehat{\Psi}_{\hat{m}}^{-1} - \Psi_{\hat{m}}^{-1}\|_{\text{op}} > \frac{1}{2}\|\Psi_{\hat{m}}^{-1}\|_{\text{op}}\} \cap \Xi_{n\Delta} \cap \Omega_{n\Delta}\right).\end{aligned}$$

Now

$$\mathbb{P}\left(\{\|\widehat{\Psi}_{\hat{m}}^{-1} - \Psi_{\hat{m}}^{-1}\|_{\text{op}} > \frac{1}{2}\|\Psi_{\hat{m}}^{-1}\|_{\text{op}}\} \cap \Xi_{n\Delta} \cap \Omega_{n\Delta}\right) \leq \sum_{m \in \mathcal{M}_{n\Delta}^+} \mathbb{P}\left(\|\widehat{\Psi}_m^{-1} - \Psi_m^{-1}\|_{\text{op}} > \frac{1}{2}\|\Psi_m^{-1}\|_{\text{op}}\right)$$

Then we use Proposition 2.4 `(ii)` in Comte and Genon-Catalot (2018), to get

$$\mathbb{P}\left(\{\|\widehat{\Psi}_{\hat{m}}^{-1} - \Psi_{\hat{m}}^{-1}\|_{\text{op}} > \frac{1}{2}\|\Psi_{\hat{m}}^{-1}\|_{\text{op}}\} \cap \Xi_{n\Delta} \cap \Omega_{n\Delta}\right) \leq \mathbb{P}\left(\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2} - \text{Id}_m\|_{\text{op}} > \frac{1}{4}\right).$$

The choice of $\mathfrak{d}$ implies that this probability is less than $K/(n\Delta)^5$. This leads to

$$\mathbb{E}\left[(\text{pen}(\hat{m}) - \widehat{\text{pen}}(\hat{m}))_+ \mathbf{1}_{\Xi_{n\Delta} \cap \Omega_{n\Delta}}\right] \lesssim 1/(n\Delta).$$

This ends the proof of (54) and we can set $\kappa' = \kappa/2$ and take $\kappa \geq 2 \times 12 \times 16$. □

6.12. **Appendix.** We recall the following result of Tropp (2015) and the Talagrand concentration inequality given in Klein and Rio (2005).

Theorem 6.1 (Bernstein Matrix inequality)**.** *Consider a finite sequence $\{\mathbf{S}_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that $\mathbb{E}\mathbf{S}_k = 0$ and $\|\mathbf{S}_k\|_{\text{op}} \leq L$ for all k. Introduce the random matrix $\mathbf{Z} = \sum_k \mathbf{S}_k$. Let $\nu(\mathbf{Z})$ be the variance statistic of the sum: $\nu(\mathbf{Z}) = \max\{\lambda_{\max}(\mathbb{E}[\mathbf{Z}'\mathbf{Z}]), \lambda_{\max}(\mathbb{E}[\mathbf{Z}\mathbf{Z}'])\}$. Then for all $t \geq 0$*

$$\mathbb{P}\left[\|\mathbf{Z}\|_{\text{op}} \geq t\right] \leq (d_1 + d_2) \exp\left(-\frac{t^2/2}{\nu(\mathbf{Z}) + Lt/3}\right).$$

Theorem 6.2. *Consider $n \in \mathbb{N}^*$, $\mathcal{F}$ a class at most countable of measurable functions, and $(X_i)_{i \in \{1,\dots,n\}}$ a family of real independent random variables. Define, for $f \in \mathcal{F}$, $\nu_n(f) = (1/n)\sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)])$, and assume that there are three positive constants M, H and v such that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq M$, $\mathbb{E}[\sup_{f \in \mathcal{F}} |\nu_n(f)|] \leq H$, and $\sup_{f \in \mathcal{F}} (1/n)\sum_{i=1}^n \text{Var}(f(X_i)) \leq v$. Then for all $\alpha > 0$, with $C(\alpha) = (\sqrt{1+\alpha} - 1) \wedge 1$, and $b = \frac{1}{6}$,*

$$\mathbb{E}\left[\left(\sup_{f \in \mathcal{F}} |\nu_n(f)|^2 - 2(1+2\alpha)H^2\right)_+\right] \quad \leq \quad \frac{4}{b}\left(\frac{v}{n} e^{-b\alpha\frac{nH^2}{v}} + \frac{49M^2}{bC^2(\alpha)n^2} e^{-\frac{\sqrt{2}bC(\alpha)\sqrt{\alpha}}{7}\frac{nH}{M}}\right).$$

By density arguments, this result can be extended to the case where $\mathcal{F}$ is a unit ball of a linear normed space.

References

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition.

[2] Aeckerle-Willems, C. and Strauch, C. (2018). Sup-norm adaptive simultaneous drift estimation for ergodic diffusions. Preprint arXiv:1808.10660

[3] Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.* **6**, 127-146.

[4] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301-413.

[5] Baraud, Y., Comte, F. and Viennet, Y. (2001) Adaptive estimation in autoregression or β-mixing regression via model selection. *The Annals of Statistics*, **29**, 839-875.

[6] Belomestny, D., Comte, F., and Genon-Catalot, V. (2016). Nonparametric Laguerre estimation in the multiplicative censoring model. *Electron. J. Statist.*, 10(2):3114–3152.

[7] Berbee, H.C.P. (1979). Random walks with stationary increments and renewal theory. *Math. Tracts. Mathematisch Centrum*, Amsterdam, **112**.

[8] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329–375.

[9] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138**, 33-73.

[10] Cohen, A., Davenport, M.A. and Leviatan, D. (2013). On the stability and acuracy of least squares approximations. *Found. Comput. math.* **13**, 819-834.

[11] Cohen, A., Davenport, M.A. and Leviatan, D. (2019). Correction to: On the stability and accuracy of least squares approximations. *Found. Comput. Math.* **19**, 239.

[12] Comte, F. and Genon-Catalot, V. (2015). Adaptive Laguerre density estimation for mixed Poisson models. *Electron. J. Stat.* **9**, 1113-1149.

[13] Comte, F. and Genon-Catalot, V. (2020a). Regression function estimation on non compact support as a partly inverse problem. *Ann Inst Stat Math* **72**, 1023-1054, doi.org/10.1007/s10463-019-00718-2.

[14] Comte, F. and Genon-Catalot, V. (2020b). Regression function estimation on non compact support in an heteroscedastic model. *Metrika* **83**, 93-128.

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

[16] Comte, F., Cuenod, C.-A., Pensky, M., and Rozenholc, Y. (2017). Laplace deconvolution and its application to dynamic contrast enhanced imaging. *J. R. Stat. Soc., Ser. B* **79**, 69-94.

[17] Dalalyan, A. (2005). Sharp adaptive estimation of the drift function for ergodic diffusions. *Ann. Statist.* **33**(6), 2507-2528.

[18] Gloter, A. (2000). Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient. *ESAIM Probab. Statist.* **4**, 205-227.

[19] Hoffmann, M. (1999). Adaptive estimation in diffusion processes. *Stochastic Process. Appl.* **79**, 135-163.

[20] Indritz, J. (1961). An inequality for Hermite polynomials. *Proc. Amer. Math. Soc.* **12**, 981-983.

[21] Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33**, no. 3, 1060-1077.

[22] Kutoyants, Y.A. (2004). Statistical inference for ergodic diffusion processes. Springer Series in Statistics. Springer-Verlag London, Ltd., London.

[23] Nickl, R. and Ray, K. (2020) Nonparametric statistical inference for drift vector fields of multidimensional diffusions. *The Annals of Statistics* **48**, 1383-1408.

[24] Mabon, G. (2017). Adaptive deconvolution on the nonnegative real line. *Scandinavian Journal of Statistics*, 44:707-740.

[25] Pardoux, E. and Veretennikov, A. Yu. (2001). On the Poisson equation and diffusion approximation. I. *Ann. Probab.* **29**, 3, 1061-1085.

[26] Strauch, C. (2015). Sharp adaptive drift estimation for ergodic diffusions: the multivariate case. *Stochastic Process. Appl.*, **125**, 2562-2602.

[27] Strauch, C. (2016). Exact adaptive pointwise drift estimation for multidimensional ergodic diffusions. *Probab. Theory Related Fields* **164**, no. 1-2, 361-400.

[28] Szegö, G. (1975). *Orthogonal polynomials.* Fourth edition. American Mathematical Society, Colloquium Publications, Vol. XXIII. American mathematical Society, Providence, R.I.

[29] Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230.

[30] Veretennikov, A. Y. (1988). Bounds for the mixing rate in the theory of stochastic differential equation, *Theory Probab. Appl.*, **32**, 273-281.

[31] Viennet, G. (1997). Inequalities for absolutely regular processes: application to density estimation. *Probab. Theory Relat. Fields* **107**, 467-492.