Optimal adaptive estimation on R or R+ of the derivatives of a density

Fabienne Comte, Céline Duval, Ousmane Sacko

OPTIMAL ADAPTIVE ESTIMATION ON $\mathbb{R}$ OR $\mathbb{R}^+$ OF THE DERIVATIVES OF A DENSITY

FABIENNE COMTE[1], CÉLINE DUVAL[1], AND OUSMANE SACKO[1]

Abstract. In this paper, we consider the problem of estimating the d-order derivative $f^{(d)}$ of a density f, relying on a sample of n i.i.d. observations $X_1, \ldots, X_n$ with density f supported on $\mathbb{R}$ or $\mathbb{R}^+$. We propose projection estimators defined in the orthonormal Hermite or Laguerre bases and study their integrated $\mathbb{L}^2$-risk. For the density f belonging to regularity spaces and for a projection space chosen with adequate dimension, we obtain rates of convergence for our estimators, which are proved to be optimal in the minimax sense. The optimal choice of the projection space depends on unknown parameters, so a general data-driven procedure is proposed to reach the bias-variance compromise automatically. We discuss the assumptions and the estimator is compared to the one obtained by simply differentiating the density estimator. Simulations are finally performed and illustrate the good performances of the procedure and provide numerical comparison of projection and kernel estimators. September 20, 2019

Key words: Estimation of derivatives of a density; Hermite basis; Laguerre basis; Model selection; Projection estimator;

AMS Classification: 62G05 - 62G07

1. Introduction

1.1. **Motivations and content.** Let $X_1, \ldots, X_n$ be n i.i.d. random variables with common density f with respect to the Lebesgue measure. The problem of estimating f in this simple model has been widely studied. In some contexts, it is also of interest to estimate the d-th order derivative $f^{(d)}$ of f, for different values of the integer d. Several examples are developed in Singh (1977): regression curves $r(x) = \mathbb{E}(Y|X = x)$ for specific families of conditional distributions of Y given X, where $r(x) = f^{(1)}(x)/f(x)$ (see also Park and Kang (2008)); estimation and testing in one parameter scale of exponential families (see Genovese et al. (2016))... Derivative estimation can also be used as a mean of reaching information, such as mode seeking in mixture models and in data analysis, see *e.g.* Cheng (1995), Chacón and Duong (2013). Moreover, density derivatives also provide information about the slope of the curves, local extrema, saddle points...

Most proposals for estimating the derivative of a density are built as derivatives of kernel density estimators, see Bhattacharya (1967), Schuster (1969), Silverman (1978), Rao (1996), Chacón et al. (2011), Chacón and Duong (2013) or Giné and Nickl (2016), either in independent or in α-mixing settings, in univariate or multivariate contexts. A slightly different proposal still based on kernels can be found in Singh (1979). The question of bandwidth selection is not considered in the oldest of these papers, but is studied in more recent ones. For instance, Chacón and Duong (2013) propose a general cross-validation method in the multivariate case for a matrix bandwidth, see also the references therein. The case of estimation on $\mathbb{R}^+$ with gamma kernel estimator (and mixing data) is studied in Markovich (2016), and a risk bound is proved, but specifically for a first order derivative and a density with regularity of order 2. Projection estimators have also been considered for density and derivatives estimation. More precisely, using trigonometric basis, Efromovich (1998) proposes a complete study of optimality and sharpness of such estimators, on Sobolev periodic spaces. More recently, Giné and Nickl (2016) propose a projection

[1] Université Paris Descartes, MAP5, UMR CNRS 8145.

estimator and provide an upper bound for its $\mathbb{L}^p$-risk, $p \in [1, \infty]$. In a dependent context, Schmisser (2013) studies projection estimators in a compactly supported basis constrained on the borders or a non compact multi-resolution basis: she considers dependent β-mixing variables and a model selection method is proposed and proved to reach optimal rates on Besov spaces. In both contexts, the rate obtained for estimating $f^{(d)}$ the d-th order derivative belonging to a regularity space associated to a regularity α, is of order $n^{-2\alpha/(2\alpha+2d+1)}$.

In this work, we also consider projection estimators, defined as in Giné and Nickl (2016), but on specific projection spaces generated by Hermite or Laguerre basis. The integrated $\mathbb{L}^2$-risk of such estimators is classically decomposed into a squared bias and a variance term. The specificity of our context lies in the following facts.

(1) The bias term is studied on specific regularity spaces, namely Sobolev Hermite and Sobolev Laguerre spaces, as defined in Bongioanni and Torrea (2009), enabling to consider non compact estimation support $\mathbb{R}$ or $\mathbb{R}^+$.
(2) The order of the variance term depends on moment assumptions. This explains why, to perform a data driven selection of the projection space, we propose a random empirical estimator of the variance term, which has automatically the right order.
(3) In standard settings, the dimension of the projection space is the relevant parameter that needs to be selected to achieve the bias-variance compromise. In our context, this role is played by the square root of the dimension.

We also mention that our procedure provides very parsimonious estimators, as they require very few coefficients to reconstruct functions accurately. Moreover, our regularity assumptions are naturally set on f and not on its derivatives, contrary to what is done in several papers. We emphasize that we provide a complete panorama of the problem of estimating the derivatives of a density, providing a comparison of our estimators with those defined as derivatives of projection density estimators; a strategy usually applied with kernel methods. Finally, we also propose a numerical comparison between our projection procedure and a sophisticated kernel method inspired by Lacour et al. (2017).

The paper is organized as follows. In the remaining of this section, we define the Hermite and Laguerre bases and associated projection spaces. In Section 2, we define the estimators and establish general risk bounds, from which rates of convergence are obtained, and lower bounds in the minimax sense are proved. A model selection procedure is proposed, relying on a general variance estimate; it leads to a data-driven bias-variance compromise. Further questions are studied in Section 3: the comparison the derivatives of the density estimator leads in our setting to different developments depending on the considered basis: interestingly Hermite and Laguerre cases happen to behave differently from this point of view. Lastly, a simulation study is conducted in Section 4, in which kernel and projection strategies are compared.

1.2. **Notations and definition of the basis.** The following notations are used in the remaining of this paper. For a, b two real numbers, denote $a \vee b = \max(a, b)$ and $a_+ = \max(0, a)$. For u and v two functions in $\mathbb{L}^2(\mathbb{R})$, denote $\langle u, v\rangle = \int_{-\infty}^{+\infty} u(x)v(x)dx$ the scalar product on $\mathbb{L}^2(\mathbb{R})$ and $\|u\| = \left(\int_{-\infty}^{+\infty} u(x)^2 dx\right)^{1/2}$ the norm on $\mathbb{L}^2(\mathbb{R})$. Note that these definitions remain consistent if u and v are in $\mathbb{L}^2(\mathbb{R}^+)$.

1.2.1. *The Laguerre basis.* Define the Laguerre basis by:

$$\ell_j(x) = \sqrt{2}L_j(2x)e^{-x}, \quad L_j(x) = \sum_{k=0}^{j} \binom{j}{k}(-1)^k \frac{x^k}{k!}, \quad x \geqslant 0, \quad j \geqslant 0, \tag{1}$$

where L_j is the Laguerre polynomial of degree j. It satisfies: $\int_0^{+\infty} L_k(x)L_j(x)e^{-x}dx = \delta_{k,j}$ (see Abramowitz and Stegun (1964), 22.2.13), where $\delta_{k,j}$ is the Kronecher symbol. The family $(\ell_j)_{j\geqslant 0}$ is an orthonormal

basis on $\mathbb{L}^2(\mathbb{R}^+)$ such that $\|\ell_j\|_\infty = \sup_{x\in\mathbb{R}^+}|\ell_j(x)| \leqslant \sqrt{2}$. The derivative of ℓ_j satisfies a recursive formula (see Lemma 8.1 in Comte and Genon-Catalot (2018)) that plays an important role in the sequel:

$$\ell_0' = -\ell_0, \qquad \ell_j' = -\ell_j - 2\sum_{k=0}^{j-1}\ell_k, \quad \forall j \geqslant 1. \tag{2}$$

1.2.2. *The Hermite basis.* Define the Hermite basis $(h_j)_{j\geqslant 0}$ from Hermite polynomials $(H_j)_{j\geqslant 0}$:

$$h_j(x) = c_j H_j(x) e^{-x^2/2}, \quad H_j(x) = (-1)^j e^{x^2}\frac{d^j}{dx^j}(e^{-x^2}), \quad c_j = (2^j j!\sqrt{\pi})^{-1/2}, \quad x\in\mathbb{R},\, j\geqslant 0. \tag{3}$$

The family $(H_j)_{j\geqslant 0}$ is orthogonal with respect to the weight function e^{-x^2}: $\int_{\mathbb{R}} H_j(x)H_k(x)e^{-x^2}dx = 2^j j!\sqrt{\pi}\delta_{j,k}$ (see Abramowitz and Stegun (1964), 22.2.14). It follows that $(h_j)_{j\geqslant 0}$ is an orthonormal basis on $\mathbb{R}$. Moreover, h_j is bounded by

$$\|h_j\|_\infty = \sup_{x\in\mathbb{R}}|h_j(x)| \leqslant \phi_0, \text{ with } \phi_0 = \pi^{-1/4}, \tag{4}$$

(see Abramowitz and Stegun (1964), chap.22.14.17 and Indritz (1961)). The derivatives of h_j also satisfy a recursive formula (see Comte and Genon-Catalot (2018), Equation (52) in Section 8.2),

$$h_0' = -h_1/\sqrt{2}, \qquad h_j' = (\sqrt{j}\; h_{j-1} - \sqrt{j+1}h_{j+1})/\sqrt{2}, \quad \forall j\geqslant 1. \tag{5}$$

In the sequel, we denote by φ_j either for h_j in the Hermite case or for ℓ_j in the Laguerre case. Let $g\in\mathbb{L}^2(\mathbb{R})$ or $g\in\mathbb{L}^2(\mathbb{R}^+)$, g develops either in the Hermite basis or the Laguerre basis:

$$g = \sum_{j\geqslant 0} a_j(g)\varphi_j, \quad a_j(g) = \langle g, \varphi_j\rangle.$$

Define, for an integer $m\geqslant 1$, the space

$$S_m = \mathrm{Span}\{\varphi_0, \ldots, \varphi_{m-1}\}.$$

The orthogonal projection of g on S_m is given by: $g_m = \sum_{j=0}^{m-1} a_j(g)\varphi_j$.

2. Estimation of the first derivative without boundary issue

2.1. **Assumptions and projection estimator of $f^{(d)}$.** Let $X_1, \ldots, X_n$ be n i.i.d. random variables with common density f with respect to the Lebesgue measure and consider the following assumptions. Let d be an integer, $d\geqslant 1$.

- (**A**1) The density f is d-times differentiable and $f^{(d)}$ belongs to $\mathbb{L}^2(\mathbb{R}^+)$ in the Laguerre case or $\mathbb{L}^2(\mathbb{R})$ in the Hermite case.
- (**A**2) For all integer r, $0\leqslant r\leqslant d-1$, we have $\|f^{(r)}\|_\infty < +\infty$.
- (**A**3) For all integer r, $0\leqslant r\leqslant d-1$, it holds $\lim_{x\to 0} f^{(r)}(x) = 0$.

Assumption (**A**3) is specific to the Laguerre case and avoids boundary issue. In particular, it permits to establish Lemma 2.1 below that is central to define our estimator. This assumption can be removed at the expense of additional technicalities, see Section 3.
Under (**A**1), we develop $f^{(d)}$ in the Laguerre or Hermite basis, its orthogonal projection on S_m, $m\geqslant 1$, is

$$f_m^{(d)} = \sum_{j=0}^{m-1} a_j(f^{(d)})\varphi_j, \text{ where, } \quad a_j(f^{(d)}) = \langle f^{(d)}, \varphi_j\rangle. \tag{6}$$

The estimator is built by using the following result, proved in Appendix A.

Lemma 2.1. *Suppose that* (**A**1) *and* (**A**2) *hold in the Hermite case and that* (**A**1), (**A**2) *and* (**A**3) *hold in the Laguerre case. Then* $a_j(f^{(d)}) = (-1)^d \mathbb{E}[\varphi_j^{(d)}(X_1)]$, $\forall j \geqslant 0$.

Remark 1. *If the support of the density f is a strict compact subset $[a,b]$ of the estimation support (here $\mathbb{R}$ and $a<b$ or $\mathbb{R}^+$ and $0<a<b$), then the regularity condition* (**A**1) *implies that f must be null in a,b, as well as its derivatives up to order $d-1$. On the contrary, Assumption* (**A**3) *in the Laguerre case can be dropped out (see Section 3) and this shows that a specific problem occurs when the density support coincides with the estimation interval. This point presents a real difficulty and is either not discussed in the literature, or hidden by periodicity conditions.*

We derive the following estimator of $f^{(d)}$ (see also Giné and Nickl (2016) p.402): let $m \geqslant 1$,

$$\widehat{f}_{m,(d)} = \sum_{j=0}^{m-1} \widehat{a}_j^{(d)} \varphi_j, \quad \text{with} \quad \widehat{a}_j^{(d)} = \frac{(-1)^d}{n} \sum_{i=1}^{n} \varphi_j^{(d)}(X_i). \tag{7}$$

For $d=0$, we recover an estimator of the density f.

2.2. **Risk bound and rate of convergence.** We consider the $\mathbb{L}^2$-risk of $\widehat{f}_{m,(d)}$, defined in (7),

$$\mathbb{E}\big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\big] = \|f_m^{(d)} - f^{(d)}\|^2 + \mathbb{E}\big[\|\widehat{f}_{m,(d)} - f_m^{(d)}\|^2\big], \tag{8}$$

where $f_m^{(d)} := \sum_{k=0}^{m-1} a_j(f^{(d)})\varphi_j$. The study of the second right-hand-side term of the equality (variance term) leads to the following result.

Theorem 2.1. *Suppose that* (**A**1) *and* (**A**2) *hold in the Hermite case and that* (**A**1), (**A**2) *and* (**A**3) *hold in the Laguerre case. Assume that*

$$\mathbb{E}[X_1^{-d-1/2}] < +\infty \text{ in the Laguerre case and } \mathbb{E}[|X_1|^{2/3}] < +\infty \text{ in the Hermite case.} \tag{9}$$

Then, for sufficiently large $m \geqslant d$, it holds that

$$\mathbb{E}\big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\big] \leqslant \|f_m^{(d)} - f^{(d)}\|^2 + C\frac{m^{d+\frac{1}{2}}}{n} - \frac{\|f_m^{(d)}\|^2}{n}, \tag{10}$$

for a positive constant C depending on the moments in condition (9) (but not on m nor n).

Remark 2. *In the Laguerre case, condition* (9) *is a consequence of* (**A**3) *and $f^{(d)}(0) < +\infty$. Indeed,* (**A**3) *imposes that $f(x) \underset{x\to 0}{\sim} x^d f^{(d)}(x)$ which, under $f^{(d)}(0) < +\infty$, ensures integrability of $x^{-d-1/2}f(x)$ at 0; integrability at ∞ is a consequence of $f \in \mathbb{L}^1([0,\infty))$.*

The bound obtained for $\widehat{f}_{m,(d)}$ in Theorem 2.1 is sharp. Indeed, we can establish the following lower bound.

Proposition 2.1. *Under the Assumptions of Theorem 2.1, it holds, for some constant $c>0$, that*

$$\mathbb{E}\Big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\Big] \geqslant \|f_m^{(d)} - f^{(d)}\|^2 + c\frac{m^{d+\frac{1}{2}}}{n} - \frac{\|f_m^{(d)}\|^2}{n}.$$

2.3. **Definition of regularity classes and rate of convergence.** The first two terms in the right hand side of (10) have an antagonistic behavior with respect to m. Thus, the optimal choice of m requires a bias-variance compromise which allows to derive the rate of convergence of $\widehat{f}_{m,(d)}$. To evaluate the order of the bias term, we introduce Sobolev-Hermite and Sobolev-Laguerre regularity classes for f (see Bongioanni and Torrea (2009)).

2.3.1. *Sobolev-Hermite classes.* Let $s > 0$ and $D > 0$, define the Sobolev-Hermite ball

$$W_H^s(D) = \{\theta \in \mathbb{L}^2(\mathbb{R}), \sum_{k\geqslant 0} k^s a_k^2(\theta) \leqslant D\}, \tag{11}$$

where $a_k^2(\theta) = \langle \theta, h_k\rangle$. The following Lemma relates the regularity of $f^{(d)}$ and the one of f.

Lemma 2.2. *Let $s \geqslant d$ and $D > 0$, assume that f belongs to $W_H^s(D)$ and (**A**1), then there exist a constant $D_d > D$ such that $f^{(d)}$ is in $W_H^{s-d}(D_d)$.*

2.3.2. *Sobolev-Laguerre classes.* Similarly, consider the Sobolev-Laguerre ball

$$W_L^s(D) = \{\theta \in \mathbb{L}^2(\mathbb{R}^+), |\theta|_s^2 = \sum_{k\geqslant 0} k^s a_k^2(\theta) \leqslant D\}, \quad D > 0, \tag{12}$$

where $a_k(\theta) = \langle \theta, \ell_k\rangle$. If $s \geqslant 1$ an integer, there is an equivalent norm of $|\theta|_s^2$ (see Section 7.2 of Belomestny et al. (2016)) defined by

$$|||\theta|||_s^2 = \sum_{j=0}^{s} \|\theta\|_j^2, \quad \|\theta\|_j^2 = \|x^{j/2} \sum_{k=0}^{j} \binom{j}{k} \theta^{(k)}\|^2. \tag{13}$$

This inspires the definition, for $s \in \mathbb{N}$ and $D > 0$, of the subset $\widetilde{W}_L^s(D)$ as

$$\widetilde{W}_L^s(D) = \{\theta \in \mathbb{L}^2(\mathbb{R}^+),\, \theta^{(j)} \in C([0,\infty)),\, x \mapsto x^{k/2}\theta^{(j)}(x) \in \mathbb{L}^2(\mathbb{R}^+),\, 0 \leqslant j \leqslant k \leqslant s, |\theta|_s^2 \leqslant D\}. \tag{14}$$

It is straightforward to see that $\widetilde{W}_L^s(D) \subset W_L^s(D)$. Moreover, we can relate the regularity of $f^{(d)}$ and the one of f.

Lemma 2.3. *Let $s \in \mathbb{N}$, $s \geqslant d \geqslant 1$, $D > 0$ and $\theta \in \widetilde{W}_L^s(D)$, then, $\theta^{(d)} \in \widetilde{W}_L^{s-d}(D_d)$ where $D \leqslant D_d < \infty$.*

2.3.3. *Rate of convergence of $\widehat{f}_{m,(d)}$.* Assume that $f \in W_H^s(D)$ or $f \in \widetilde{W}_L^s(D)$, then Lemmas 2.2 and 2.3 enable a control of the bias term in (10)

$$\|f_m^{(d)} - f^{(d)}\|^2 = \sum_{j\geqslant m} (a_j(f^{(d)}))^2 = \sum_{j\geqslant m} j^{s-d}(a_j(f^{(d)}))^2 j^{-(s-d)} \leqslant D_d m^{-(s-d)}.$$

Injecting this in (10) yields

$$\mathbb{E}\big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\big] \leqslant D' m^{-(s-d)} + c\frac{m^{d+\frac{1}{2}}}{n}.$$

Consequently, selecting $m_{opt} = [n^{2/(2s+1)}]$ gives the rate of convergence

$$\mathbb{E}\big[\|\widehat{f}_{m_{opt},(d)} - f^{(d)}\|^2\big] \leqslant C(s,d,D) n^{-\frac{2(s-d)}{2s+1}}, \tag{15}$$

where $C(s,d,D)$ depends only on s, d and D, not on m. This rate coincides with the one obtained by Schmisser (2013) in the dependent case and by Giné and Nickl (2016). We can however mention that the squared-bias and variance terms do not have the same orders: the role of dimension in Schmisser (2013) is played in our setting by $\sqrt{m}$. This rate is better than the one obtained by Rao (1996) in the i.i.d. case, if we set a similar regularity condition. Note that, for $d = 0$ in (15), we recover the optimal rate for estimation of the density f.

Remark 3. *If f is a mixture of Gaussian densities in the Hermite case or a mixture of Gamma densities in the Laguerre case, it is known from Section 3.2 in Comte and Genon-Catalot (2018) that the bias decreases with exponential rate. The computations therein can be extended to the present setting and imply in both Hermite and Laguerre cases that m_{opt} is then proportional to $\log(n)$. Therefore the risk has*

order $[\log(n)]^{d+\frac{1}{2}}/n$*: for these collections of densities, the estimator converges much faster than in the general setting.*

2.4. **Lower bound.** Contrary to the lower bound given in Proposition 2.1, which ensures that the upper bound derived in Theorem 2.1 for the specific estimator $\widehat{f}_{m,(d)}$ is sharp, we provide a general lower bound that guarantees that the rate of the estimator $\widehat{f}_{m,(d)}$ is minimax optimal. The following Theorem states that the rate obtained in (15) is the optimal rate.

Theorem 2.2. *Let* $s \geqslant d$ *be an integer and* $\widetilde{f}_{n,d}$ *be any estimator of* $f^{(d)}$*. Then for* n *large enough, we have*

$$\inf_{\widetilde{f}_{n,d}} \sup_{f \in W^s(D)} \mathbb{E}[\|\widetilde{f}_{n,d} - f^{(d)}\|^2] \geqslant c n^{-\frac{2(s-d)}{2s+1}},$$

where the infimum is taken over all estimator of $f^{(d)}$*,* c *a positive constant depending on* s *and* d*, and* $W^s(D)$ *stands either for* $W^s_L(D)$ *or for* $W^s_H(D)$*.*

2.5. **Adaptive estimator of** $f^{(d)}$**.** The choice of $m_{opt} = [n^{2/(2s+1)}]$ leading to the optimal rate of convergence is not feasible in practice. In this section we provide an automatic choice of the dimension m, from the observations $(X_1, \ldots, X_n)$, that realizes the bias-variance compromise in (10). Assume that m belongs to a finite model collection $\mathcal{M}_{n,d}$, we look for m that minimizes the bias-variance decomposition (8) rewritten as

$$\mathbb{E}\big[\|\widehat{f}_{m,(d)} - f^{(d)}\|^2\big] = \|f^{(d)}_m - f^{(d)}\|^2 + \frac{1}{n}\sum_{j=0}^{m-1} \mathrm{Var}\left[\varphi^{(d)}_j(X_1)\right].$$

Note that the bias is such that $\|f^{(d)}_m - f^{(d)}\|^2 = \|f^{(d)}\|^2 - \|f^{(d)}_m\|^2$ where $\|f^{(d)}\|^2$ is independent of m and can be dropped out. The remaining quantity $-\|f^{(d)}_m\|^2$ is estimated by $-\|\widehat{f}_{m,(d)}\|^2$. The variance term is replaced by an estimator of a sharp upper bound, given by

$$\widehat{V}_{m,d} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{m-1}(\varphi^{(d)}_j(X_i))^2. \tag{16}$$

Finally, we set

$$\widehat{m}_n := \underset{m \in \mathcal{M}_{n,d}}{\mathrm{argmin}}\{-\|\widehat{f}_{m,(d)}\|^2 + \widehat{\mathrm{pen}}_d(m)\}, \quad \text{where} \quad \widehat{\mathrm{pen}}_d(m) = \kappa\frac{\widehat{V}_{m,d}}{n}, \tag{17}$$

where κ is a positive numerical constant. If we set $V_{m,d} := \sum_{j=0}^{m-1}\mathbb{E}[(\varphi^{(d)}_j(X_1)^2]$, it holds $\mathbb{E}[\widehat{\mathrm{pen}}_d(m)] = \kappa V_{m,d}/n$. In the sequel, we write $\mathrm{pen}_d(m) := \kappa V_{m,d}/n$. To implement the procedure a value for κ has to be set. Theorem 2.3 below provides a theoretical lower bound for κ, which is however generally too large. In practice this constant is calibrated by intensive preliminary experiments, see Section 4. General calibration methods can be found in Baudry et al. (2012) for theoretical explanations and heuristics, and in the associated package, for practical implementation.

Remark 4. *Note that in the definition of the penalty, instead of* (17)*, we can plug the deterministic upper bound on the variance and take* $c\,m^{d+\frac{1}{2}}/n$ *as a penalty (see Theorem 2.1) as Proposition 2.1 ensures its sharpness. However, this upper bound relies on additional assumptions given in* (9) *and depends on non explicit constants (see Askey and Wainger (1965)). This is why we choose to estimate directly the variance by* $\widehat{V}_{m,n}$ *and use* $\widehat{V}_{m,n}/n$ *as the penalty term.*

Theorem 2.3. *Let $\mathcal{M}_{n,d} := \{d, \dots, m_n(d)\}$, where $m_n(d) \geqslant d$. Assume that (A1) and (A2) hold, and that (A3) holds in the Laguerre case, and that $\|f\|_\infty < +\infty$.*

AL. Set $m_n(d) = \lfloor (n/\log^3(n))^{\frac{2}{2d+1}} \rfloor$, assume that $\sup_{x\in\mathbb{R}^+} \frac{f(x)}{x^d} < +\infty$ in the Laguerre case,

AH. Set $m_n(d) = \lfloor n^{\frac{2}{2d+1}} \rfloor$ in the Hermite case.

Then, for any $\kappa \geqslant \kappa_0 := 32$ it holds that

$$\mathbb{E}\Big[\|\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}\|^2\Big] \leqslant C \inf_{m\in\mathcal{M}_{n,d}} \Big(\|f_m^{(d)} - f^{(d)}\|^2 + \mathrm{pen}_d(m)\Big) + \frac{C'}{n}, \tag{18}$$

where C is a universal constant ($C = 3$ suits) and C' is a constant depending on $\sup_{x\in\mathbb{R}^+} \frac{f(x)}{x^d} < +\infty$ and $\mathbb{E}[X_1^{-d-1/2}] < +\infty$ (Laguerre case) or $\|f\|_\infty$ (Hermite case).

The constraint on the the largest element $m_n(d)$ of the collection $\mathcal{M}_{n,d}$ ensures that the variance term, which is upper bounded by $m^{d+\frac{1}{2}}/n$ vanishes asymptotically. The additional log term does not influence the rate of the optimal estimator: the optimal (and unknown) dimension $m_{opt} \asymp n^{\frac{2}{2s+1}}$, with s the regularity index of f, is such that $m_{opt} \ll n^{\frac{2}{2d+1}}$ as soon as $s > d$. For $s = d$, a log-loss in the rate would occur in the Laguerre case, but not in the Hermite case.

Note that, in the Laguerre case, condition $\sup_{x\in\mathbb{R}^+} \frac{f(x)}{x^d} < +\infty$ implies $\mathbb{E}(X_1^{-d-1/2}) < +\infty$ (see condition 9)) and is clearly related to (A3). Inequality (18) is a key result and expresses that $\widehat{f}_{\widehat{m}_n,(d)}$ realizes automatically a bias-variance compromise and is performing as well as the best model in the collection, up to the multiplicative constant C, since clearly, the last term C'/n is negligible. Thus, for f in $\widetilde{W}_L^s(D)$ or $W_H^s(D)$ and under the assumptions of Theorem 2.3, we have $\mathbb{E}\big[\|\widehat{f}_{\widehat{m},(d)} - f^{(d)}\|^2\big] = \mathcal{O}(n^{-2(s-d)/(2s+1)})$, which implies that the estimator is adaptive.

3. Further questions

We investigate here additional questions, and set for simplicity $d = 1$. Mainly, we compare our estimator to the derivative of a density estimator, and discuss condition (A3) in the Laguerre case.

3.1. Derivatives of the density estimator.

When using kernel strategies, it is classical to build an estimator of the derivative of f by differentiating the kernel density estimator, as already mentioned in the Introduction. For projection estimators, we find more relevant to proceed differently. Indeed, our aim is to obtain an estimator expressed in an orthonormal basis; unfortunately, the derivative of an orthonormal basis is a collection of functions but not an orthonormal basis. So, our proposal (7) is easier to handle. Moreover, our estimator can be seen as a contrast minimizer, which makes model selection possible to settle up.

However, Laguerre and Hermite cases are somehow different and can be more precisely compared. Let us recall that the projetion estimator of f on S_m is defined by (see Comte and Genon-Catalot (2018), or (7) for $d = 0$):

$$\widehat{f}_m := \sum_{k=0}^{m-1} \widehat{a}_k^{(0)} \varphi_k, \quad \text{where} \quad \widehat{a}_k^{(0)} := \frac{1}{n} \sum_{j=0}^{n} \varphi_k(X_j).$$

As the functions $(\varphi_j)_j$ are infinitely differentiable, both in Hermite and Laguerre settings, this leads to the natural estimator of $f^{(d)}$, $d \geqslant 1$,

$$(\widehat{f}_m)^{(d)} = \sum_{k=0}^{m-1} \widehat{a}_k^{(0)} \varphi_k^{(d)}. \tag{19}$$

For $d=1$, we write $(\widehat{f}_m)^{(1)} = (\widehat{f}_m)'$. We want to compare $(\widehat{f}_m)'$ to $\widehat{f}_{m,(1)}$. In both Hermite and Laguerre cases, this estimator is consistent, under adequate regularity assumptions and for adequate choice of m as a function of n.

3.2. **Comparison of $\widehat{f}_{m,(1)}$ with $(\widehat{f}_m)'$ in the Hermite case.** Using the recursive formula (5), in (19) and (7) respectively, straightforward computations give

$$(\widehat{f}_m)' = \frac{1}{\sqrt{2}}\widehat{a}_1^{(0)}h_0 + \sum_{j=1}^{m-1}\left(\sqrt{\frac{j+1}{2}}\widehat{a}_{j+1}^{(0)} - \sqrt{\frac{j}{2}}\widehat{a}_{j-1}^{(0)}\right)h_j - \sqrt{\frac{m}{2}}\left(\widehat{a}_m^{(0)}h_{m-1} + \widehat{a}_{m-1}^{(0)}h_m\right),$$

whereas
$$\widehat{f}_{m,(1)} = \frac{1}{\sqrt{2}}\widehat{a}_1^{(0)}h_0 + \sum_{j=1}^{m-1}\left(\sqrt{\frac{j+1}{2}}\widehat{a}_{j+1}^{(0)} - \sqrt{\frac{j}{2}}\widehat{a}_{j-1}^{(0)}\right)h_j.$$

Therefore, it holds that $\mathbb{E}[\|(\widehat{f}_m)' - \widehat{f}_{m,(1)}\|^2] = m/2\{\mathbb{E}\left[(\widehat{a}_m^{(0)})^2\right] + \mathbb{E}\left[(\widehat{a}_{m-1}^{(0)})^2\right]\}$ and

$$\mathbb{E}[\|(\widehat{f}_m)' - \widehat{f}_{m,(1)}\|^2] \leqslant \frac{m}{2}(a_{m-1}^2(f) + a_m^2(f)) + \frac{m}{2n}\left(\int h_m^2(x)f(x)dx + \int h_{m-1}^2(x)f(x)dx\right).$$

Using Lemma 8.5 in Comte and Genon-Catalot (2018) under $\mathbb{E}[|X_1|^{2/3}] < +\infty$ and for f in $W_H^s(D)$, $s>1$, it follows for some positive constant C that,

$$\mathbb{E}[\|(\widehat{f}_m)' - \widehat{f}_{m,(1)}\|^2] \leqslant \frac{D}{2}m^{-s+1} + C\frac{\sqrt{m}}{n}.$$

Under the same assumptions, (10) for $d=1$ implies

$$\mathbb{E}[\|(\widehat{f}_m)' - f'\|^2] \leqslant D'm^{-s+1} + c\frac{m^{3/2}}{n}.$$

Therefore, by triangle inequality, this implies that $(\widehat{f}_m)'$ reaches the same (optimal) rate as $\widehat{f}_{m,(1)}$, under the same assumptions.

3.3. **Comparison of $\widehat{f}_{m,(1)}$ with $(\widehat{f}_m)'$ in the Laguerre case.** In the Laguerre case, assumption (**A**3) is required for the estimator $\widehat{f}_{m,(1)}$ to be consistent, while it is not for the estimator $(\widehat{f}_m)'$.
Proceeding as previously and taking advantage of the recursive formula (2) in (19) and (7) respectively, straightforward computations give, for $m \geqslant 1$,

$$(\widehat{f}_m)' = \sum_{j=0}^{m-1}\left(\widehat{a}_j^{(0)} - 2\sum_{k=j}^{m-1}\widehat{a}_k^{(0)}\right)\ell_j, \quad \text{whereas} \quad \widehat{f}_{m,(1)} = \sum_{j=0}^{m-1}\left(\widehat{a}_j^{(0)} + 2\sum_{k=0}^{j-1}\widehat{a}_k^{(0)}\right)\ell_j. \tag{20}$$

Therefore, in the Laguerre case, the coefficients of $\widehat{f}_{m,(1)}$ in the basis $(\ell_j)_j$ do not depend on m while those of $(\widehat{f}_m)'$ do. Moreover, computing the difference between the estimators leads to $\widehat{f}_{m,(1)} - (\widehat{f}_m)' = 2\sum_{j=0}^{m-1}(\sum_{k=0}^{m-1}\widehat{a}_k^{(0)})\ell_j$ and

$$\|\widehat{f}_{m,(1)} - (\widehat{f}_m)'\|^2 = 4m\left(\sum_{k=0}^{m-1}\widehat{a}_k^{(0)}\right)^2.$$

Heuristically, if $f(0)=0$, as $f(0) = \sqrt{2}\sum_{j\geqslant 0}a_j(f) = 0$, it follows that $\sum_{j=0}^{m-1}a_j(f)$ should be small for m large enough. Consequently, its consistent estimator $\sum_{k=0}^{m-1}\widehat{a}_k^{(0)}$ should also be small. This would imply that, when $f(0)=0$, the distance $\|\widehat{f}_{m,(1)} - (\widehat{f}_m)'\|^2$ can be small; on the contrary, the distance should

tend to infinity with m if $f(0) \neq 0$. This is due to the fact that $\widehat{f}_{m,(1)}$ is not consistent, while $(\widehat{f}_m)'$ is. Indeed, in the general case ($f(0) \neq 0$), the risk bound we obtain for $(\widehat{f}_m)'$ is the following.

Proposition 3.1. *Assume that* (**A**1) *and* (**A**2) *hold for $d = 1$ and that f belongs to $W_L^s(D)$. Then, it holds*

$$\mathbb{E}\|(\widehat{f}_m)' - f'\|^2 \leqslant Cm^{-s+2} + \frac{3}{n}\|f\|_\infty m^2. \tag{21}$$

Obviously, for suitably chosen m the estimator is consistent and by selecting $m_{\text{opt}} \asymp n^{1/s}$, it reaches the rate: $\mathbb{E}[\|(\widehat{f}_{m_{\text{opt}}})' - f'\|^2] \leqslant C(s,D)n^{-(s-2)/s}$. This rate is worse than the one obtained for $\widehat{f}_{m,(1)}$ but it is valid without (**A**3), and thus $\widehat{f}_{m,(1)}$ is consistent to estimate an exponential density, or any mixture involving exponential densities. Note that both the order of the bias and the variance in (21) are deteriorated compared to (10), and we believe these orders are sharp.
In the following section, we investigate if the rate can be improved, if (**A**3) is not satisfied, by correcting our estimator (6).

3.4. **Estimation of f' on $\mathbb{R}^+$ with $f(0) > 0$.** Assumption (**A**3) excludes some classical distribution such as the exponential distribution or Beta distributions $\beta(a,b)$ with $a = 1$. If $f(0) > 0$, Lemma 2.1 no longer holds, and one has $a_j(f') = -f(0)\ell_j(0) - \mathbb{E}[\ell_j'(X_1)]$ instead. Therefore, $f(0)$ has to be estimated and we consider

$$\widehat{a}_{j,K}^{(1)} = -\ell_j(0)\widehat{f}_K(0) - \frac{1}{n}\sum_{i=1}^n \ell_j'(X_i), \text{ with } \widehat{f}_K = \sum_{j=0}^{K-1} \widehat{a}_j^{(0)}\ell_j,\ \widehat{a}_j^{(0)} = \frac{1}{n}\sum_{i=1}^n \ell_j(X_i). \tag{22}$$

We estimate f' as follows

$$\widetilde{f}_{m,K}' = \sum_{j=0}^{m-1} \widehat{a}_{j,K}^{(1)}\ell_j, \text{ with } \widehat{a}_{j,K}^{(1)} = -\frac{1}{n}\sum_{i=1}^n \ell_j'(X_i) - \widehat{f}_K(0)\ell_j(0). \tag{23}$$

Obviously, $\widehat{a}_{j,K}^{(1)}$ is a biased estimator of $a_j(f')$, implying that $\widetilde{f}_{m,K}'$ is a biased estimator of f_m'. Now there are two dimensions m and K to be optimized. We can establish the following upper bound.

Proposition 3.2. *Suppose* (**A**1) *is satisfied for $d = 1$, then it holds that*

$$\mathbb{E}\big[\|\widetilde{f}_{m,K}' - f'\|^2\big] \leqslant \|f' - f_m'\|^2 + \frac{2}{n}\sum_{j=0}^{m-1} \mathbb{E}\left[\left(\ell_j'(X_1)\right)^2\right] + 4m(\text{Var}(\widehat{f}_K(0)) + (f(0) - f_K(0))^2), \tag{24}$$

where f_K is the orthogonal projection of f on S_K defined by: $f_K = \sum_{j=0}^{K-1} a_j(f)\ell_j$.

The first two terms of the upper bound seem similar to the ones obtained under (**A**3), but as we no longer assume $f(0) = 0$, Assumption (9) for $d = 1$ cannot hold and the tools used to bound the variance term $V_{m,1}$ by $m^{3/2}$ no longer apply and we only get an order m^2 for this term, under $\|f\|_\infty < +\infty$.
The last two terms of (24) correspond to m times the pointwise risk of $\widehat{f}_K(0)$. Then, using $\|\ell_j\|_\infty \leqslant \sqrt{2}$, we obtain $\text{Var}(\widehat{f}_K(x)) \leqslant 4K^2/n$. If $\|f\|_\infty < \infty$, this can be improved in $\text{Var}(\widehat{f}_K(x)) \leqslant \|f\|_\infty K/n$, using the orthonormality of $(\ell_j)_j$.
To sum up, if $f \in \widetilde{W}_L^s(D)$, and $\|f\|_\infty < \infty$, then

$$\mathbb{E}\big[\|\widetilde{f}_{m,K}' - f'\|^2\big] \leqslant C(s,D,\|f\|_\infty)\left\{m^{-s+2} + \frac{m^2}{n} + m\left(K^{-s+1} + \frac{K}{n}\right)\right\}.$$

Choosing $K_{\rm opt} = cn^{1/s}$ and $m_{\rm opt} = cn^{1/s}$ gives the rate $\mathbb{E}\big[\|\widetilde{f}'_{m_{\rm opt},K_{\rm opt}} - f'\|^2\big] \leqslant Cn^{-(s-2)/s}$, that is the same rate as the one obtained for $(\widehat{f}_{m_{\rm opt}})'$. Then, renouncing to Assumption (**A**3) has a cost, it renders the procedure burdensome and leads to slower rates.
We propose a model selection procedure adapted to this new estimator. Let

$$\widehat{f}'_{m,K} = \arg\min_{t\in S_m} \gamma_n(t) \tag{25}$$

where $\gamma_n(t) = \|t\|^2 + \frac{2}{n}\sum_{i=1}^n t'(X_i) + 2t(0)\widehat{f}_K(0)$. Here, we consider that $K = K_n$ is chosen so that $\widehat{f}_{K_n}$ satisfies

$$\Big[\mathbb{E}(\widehat{f}_{K_n}(0)) - f(0)\Big]^2 \leqslant \frac{K_n \log(n)}{n}. \tag{26}$$

This assumption is likely to be fulfilled for a K selected in order to provide a squared-bias/variance compromise, see the pointwise adaptive procedure for density estimation in Plancade (2009); however therein, the choice of K is random while we set K_n as fixed, here. Then, we select m as follows:

$$\widehat{m}_K = \arg\min_{m\in\mathcal{M}_n} \Big\{\gamma_n(\widehat{f}'_{m,K}) + {\rm pen}_K(m)\Big\},\ \mathcal{M}_n = \{1,\dots,[\sqrt{n}]\} \tag{27}$$

with

$${\rm pen}_K(m) = c_1\|f\|_\infty \frac{m^2\,\log(n)}{n} + c_2(\|f\|_\infty \vee 1)\frac{m\,K\,\log(n)}{n} := {\rm pen}_1(m) + {\rm pen}_{2,K}(m). \tag{28}$$

It is easy to ckeck that $\gamma_n(\widehat{f}'_{m,K}) = -\|\widehat{f}'_{m,K}\|^2$. We prove the following result

Theorem 3.1. *Let $\widehat{f}'_{m,K_n}$ be defined by (25) with $m = \widehat{m}_{K_n}$ selected by (27)-(28) and K_n such that (26) holds. Then for c_1 and c_2 larger than fixed constants $c_{0,1}, c_{0,2}$, we have*

$$\mathbb{E}\left(\|f' - \widehat{f}'_{\widehat{m},K_n}\|^2\right) \leqslant C\left(\|f' - f'_m\|^2 + m^2\frac{\log(n)}{n} + m\frac{K_n\log(n)}{n}\right) + \frac{C'}{n},$$

where C is a numerical constant and C' depends on f.

Theorem 3.1 implies that the adaptive estimator $\widehat{f}'_{m,K_n}$ provides the adequate compromise, up to log terms.

4. Numerical study

4.1. **Simulation setting and implementation.** We illustrate the performances of the adaptive estimator $\widehat{f}_{\widehat{m}_n,(d)}$ defined in (7), with $\widehat{m}$ selected by (16)-(17), for different distributions and values of d ($d = 1, 2$). In the *Hermite case* we consider the following distributions which are estimated on the interval I, which we fix to ensure reproducibility of our experiments:

(i) Gaussian standard $\mathcal{N}(0,1)$, $I = [-4,4]$,
(ii) Mixed Gaussian $0.4\mathcal{N}(-1,1/2) + 0.6\mathcal{N}(1,1/2)$, $I = [-2.5,2.5]$,
(iii) Cauchy standard, density: $f(x) = (\pi(1+x^2))^{-1}$, $I = [-6,6]$,
(iv) Gamma $\Gamma(5,5)/10$, $I = [0,7]$,
(v) Beta $5\beta(4,5)$, $I = [0,5]$.

In the **Laguerre case** we consider densities (iv), (v) and the two following additional distributions

(vi) Weibull $W(4,1)$, I $= [0,1.5]$,
(vii) Maxwell with density $\sqrt{2}x^2e^{-x^2/(2\sigma^2)}/(\sigma^3\sqrt{\pi})$, with $\sigma = 2$ and I $= [0,8]$.

All these distributions satisfy Assumptions (**A**1), (**A**2) and densities (iv)-(vii) satisfy (**A**3). The moment conditions given in (9) are fulfilled for $d = 1, 2$, even by the Cauchy distribution (iii) which has finite moments of order $2/3 < 1$. For the adaptive procedure, the model collection considered is $\mathcal{M}_{n,d} = \{d, \ldots, m_n(d)\}$, where the maximal dimension is $m_n(d) = 50$ in the Laguerre case and $m_n(d) = 40$ in the Hermite case, for all values of n and d (smaller values may be sufficient and spare computation time). In practice, the adaptive procedure follows the steps:

- For m in $\mathcal{M}_{n,d}$, compute $-\sum_{j=0}^{m-1}(\widehat{a}_j^{(d)})^2 + \widehat{\text{pen}}_d(m)$, with $\widehat{a}_j^{(d)}$ given in (7) and $\widehat{\text{pen}}_d(m)$ in (17),
- Choose $\widehat{m}_n$ via $\widehat{m}_n = \underset{m\in\mathcal{M}_{n,d}}{\text{argmin}}\, \{-\sum_{j=0}^{m-1}(\widehat{a}_j^{(d)})^2 + \widehat{\text{pen}}_d(m)\}$,
- Compute $\widehat{f}_{\widehat{m}_n,(d)} = \sum_{j=0}^{\widehat{m}-1} \widehat{a}_j^{(d)}\varphi_j$.

Then, we compute the empirical *Mean Integrated Squared Errors (MISE)* of $\widehat{f}_{\widehat{m}_n,(d)}$. For that, we first compute the ISE by Riemann discretization in 100 points: for the j-th path, and the j-th estimate $\widehat{g}_{\widehat{m}}^{(j)}$ of g, where g stands either for the density f or for its derivative f', we set

$$\|g - \widehat{g}_{\widehat{m}}^{(j)}\|^2 \approx \frac{\text{length}(I)}{K}\sum_{k=1}^{K}(\widehat{g}_{\widehat{m}}^{(j)}(x_k)) - g(x_k))^2, \quad x_k = \min(I) + k\frac{\text{length}(I)}{K}, \quad k = 1, \ldots, K,$$

for $j = 1, \ldots R$. To get the MISE, we average over j of these R values of ISEs.
The *constant* κ in the penalty is calibrated by preliminary experiments. A comparison of the MISEs for different values of κ and different distributions (distinct from the previous ones to avoid overfitting) allows to choose a relevant value. We take $\kappa = 3.5$ in the Laguerre case or $\kappa = 4$ in the Hermite case.

Comparison with kernel estimators. We compare the performances of our method with those of kernel estimators, and start by density estimation ($d = 0$). The density kernel estimator is defined as follows

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \quad x \in \mathbb{R}$$

where $h > 0$ is the bandwidth and K a kernel such that $\int K(x)dx = 1$. These two quantities (h and K) are user-chosen. For density estimation, we use the function implemented in the statistical software `R` called `density`, where the kernel is chosen Gaussian and the bandwidth selected by cross-validation (`R`-function `bw.SJ`), see Tables 2 and 4.

f		Hermite case		Laguerre case	
Density		(ii)		(vi)	
n		500	2000	500	2000
Mean of m_{opt}	$d = 0$	7.95	9.45	5.95	7.65
	$d = 1$	8.50	9.50	6.30	7.05
	$d = 2$	8.70	9.80	5.80	6.80

TABLE 1. Mean of selected dimensions $\widehat{m}_n$ presented in Figures 1 and 2.

For the estimation of the derivative, the kernel estimator we compare with (see Tables 3 and 5) is defined by:

$$\widehat{f}'_h(x) = -\frac{1}{nh^2}\sum_{i=1}^{n} K'\left(\frac{X_i - x}{h}\right).$$

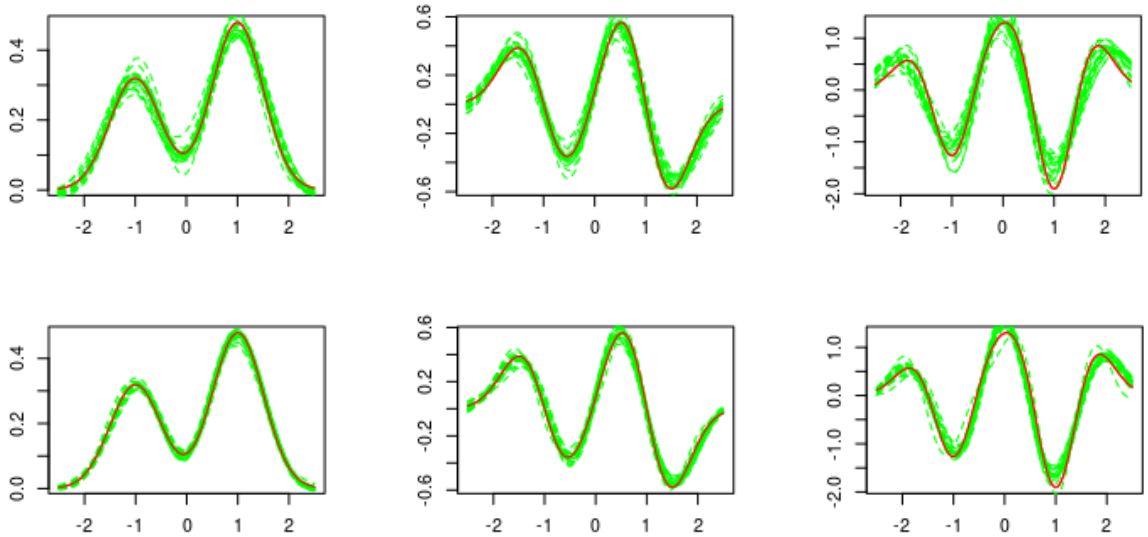

FIGURE 1. 20 estimates $\widehat{f}_{\widehat{m}_n,(d)}$ in the Hermite basis of a Mixed Gaussian distribution (ii), with $n = 500$ (first line) and $n = 2000$ (second line). The true quantity is in bold red and the estimate in dotted lines (left $d = 0$, middle $d = 1$ and right $d = 2$).

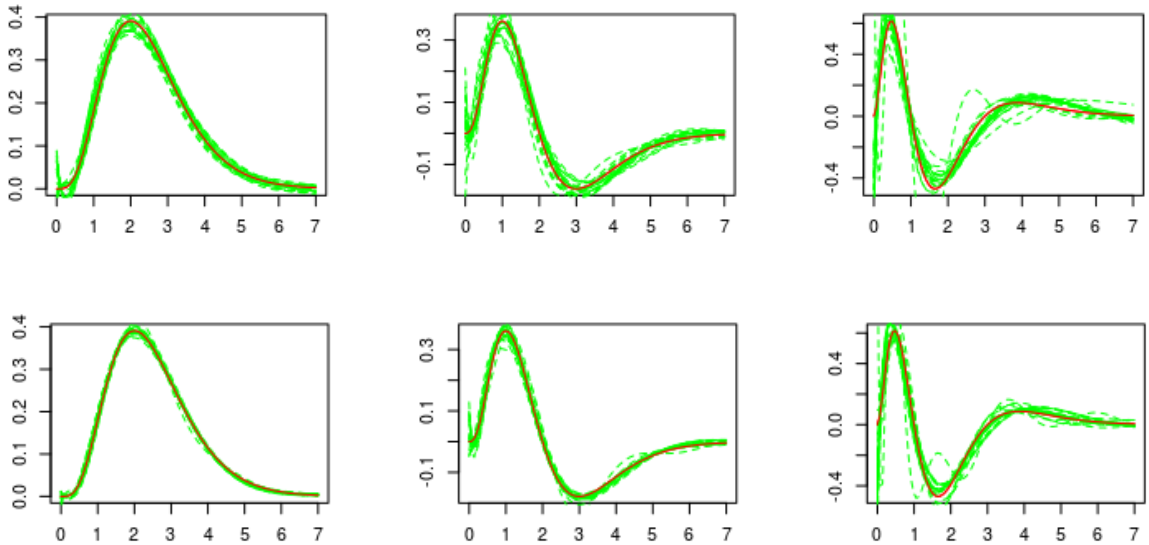

FIGURE 2. 20 estimates $\widehat{f}_{\widehat{m}_n,(d)}$ in the Laguerre basis of a Gamma distribution (iv), with $n = 500$ (first line), and $n = 2000$ (second line). The true quantity is in bold red and the estimate in dotted lines (left $d = 0$, middle $d = 1$ and right $d = 2$).

In that latter case there is no ready-to-use procedure implemented in `R`; therefore, we generalize the adaptive procedure of Lacour et al. (2017) from density to derivative estimation. To that aim, we consider

a kernel of order 7 (*i.e.* $\int x^j K(x)dx = 0$, *for* $j = 1, \ldots, 7$) built as a Gaussian mixture defined by:

$$K(x) = 4n_1(x) - 6n_2(x) + 4n_3(x) - n_4(x), \tag{29}$$

where $n_j(x)$ is the density of a centered Gaussian with a variance equal to j: the higher the order, the better the results, in theory (see Tsybakov (2009)) and in practice (see Comte and Marie (2019)). By analogy with the proposal of Lacour et al. (2017) for density estimation, we select h by:

$$\widehat{h} = \underset{h\in\mathcal{H}}{\text{argmin}}\{\|\widehat{f}'_h - \widehat{f}'_{h_{min}}\|^2 + \text{pen}(h)\}, \text{ with } \text{pen}(h) = \frac{4}{n}\langle K'_h, K_{h'_{min}}\rangle,$$

where $h_{min} = \min \mathcal{H}$, for $\mathcal{H}$ the collection of bandwidths chosen in $[c/n, 1]$ and $K_h(x) = \frac{1}{h}K(\frac{x}{h})$. Note that

$$\text{pen}(h) = \frac{4}{n}\langle K'_h, K_{h'_{min}}\rangle = \frac{4}{nh^2h^2_{min}}\int K'(\frac{u}{h})K'(\frac{u}{h_{min}})du$$

and this term can be explicitely computed with the definition of K in (29).

4.2. **Results and discussion.** Figures 1 and 2 show 20 estimated f, f', f'' in case (ii), for two values of n, 500 and 2000. These plots can be read as variability bands illustrating the performance and the stability of the estimator. We observe that increasing n improves the estimation and, on the contrary, that increasing the order of the derivative makes the problem more difficult. The means of the dimensions selected by the adaptive procedure are given in Table 1. Unsurprisingly, this dimension increases with the sample size n. In average, these dimensions are comparable for $d \in \{0, 1, 2\}$, this is in accordance with the theory: the optimal value m_{opt} does not depend on d.
Tables 2 and 4 for $d = 0$ and Tables 3 and 5 for $d = 1$ allow to compare the MISEs obtained with our method and the kernel method for different sample sizes and densities.The error decreases when the sample size increases for both methods. For density estimation ($d = 0$), the results obtained with our Hermite projection method in Table 2 are better in most cases than the kernel competitor, except for smallest sample size $n = 100$ and Gamma (iv) and Beta (v) distributions. Table 3 gives the risks obtained for derivative estimation in the Hermite basis: our method is better for densities (i), (ii), (iii) (except for $n = 100$ for Gaussian distribution (i)), but the kernel method is often better for densities (iv) and (v); they correspond to Gamma and beta densities which are in fact with support included in $\mathbb{R}^+$.
In Table 4, we compare the errors obtained for densities (iv)-(vii) with support in $\mathbb{R}^+$. Our method is always better than the R-kernel estimate. For the derivatives, in Table 5, our method and the kernel estimator seem equivalent.

	Our method				Kernel method			
f \ n	100	500	1000	2000	100	500	1000	2000
Gaussian (i)	0.12	0.03	0.02	4.10^{-3}	0.74	0.23	0.13	0.07
Mixed Gaussian (ii)	1.01	0.26	0.13	0.07	1.46	0.44	0.22	0.14
Cauchy (iii)	0.63	0.38	0.19	0.10	4.26	3.42	1.75	0.89
Gamma (iv)	1.46	0.36	0.18	0.09	0.99	0.26	0.14	0.08
Beta (v)	1.09	0.18	0.10	0.05	0.96	0.26	0.151	0.09

TABLE 2. Empirical MISE $100 \times \mathbb{E}\|\widehat{f}_{\widehat{m}} - f\|^2$ (left) and $100 \times \mathbb{E}\|\widehat{f}_{\widehat{h}} - f\|^2$ (right, Kernel Estimator) for $R = 100$ in the Hermite case.

	Our method				Kernel method			
f \ n	100	500	1000	2000	100	500	1000	2000
Gaussian (i)	1.21	0.30	0.15	0.10	1.16	0.81	0.53	0.25
Mixed Gaussian (ii)	10.08	2.39	1.89	1.07	14.13	3.56	2.00	1.2
Cauchy (iii)	2.91	1.28	0.87	0.56	4.14	1.58	1.19	0.88
Gamma (iv)	5.88	1.89	1.43	0.60	2.45	1.25	0.75	0.63
Beta (v)	5.84	1.76	0.91	0.87	5.62	3.19	0.59	0.33

TABLE 3. Empirical MISE $100 \times \mathbb{E}\|\widehat{f}_{\widehat{m},(1)} - f'\|^2$ (left) and $100 \times \mathbb{E}\|\widehat{f}'_{\widehat{h}} - f'\|^2$ (right) for $R = 100$ in the Hermite case.

	Our method				Kernel method			
f \ n	100	500	1000	2000	100	500	1000	2000
Gamma (iv)	0.54	0.16	0.08	0.04	0.99	0.26	0.14	0.08
Beta (v)	0.86	0.20	0.10	0.06	0.96	0.26	0.15	0.09
Weibull (vi)	2.61	0.60	0.33	0.17	3.55	0.80	0.46	0.29
Maxwell (vii)	0.64	0.11	0.06	0.04	0.59	0.16	0.10	0.06

TABLE 4. Empirical MISE $(100 \times \mathbb{E}\|\widehat{f}_{\widehat{m},(0)} - f\|^2$ (left) and $100 \times \mathbb{E}\|\widehat{f}_{\widehat{h}} - f\|^2$ (right) for $R = 100$ in the Laguerre case.

	Our method				Kernel method			
f \ n	100	500	1000	2000	100	500	1000	2000
Gamma (iv)	5.21	0.95	0.48	0.17	2.45	1.25	0.75	0.63
Beta (v)	4.55	1.55	0.95	0.45	5.62	3.19	0.59	0.33
Weibull (vi)	126.95	34.54	22.31	14.10	127.38	38.60	35.47	11.36
Maxwell (vii)	1.46	0.60	0.24	0.13	0.87	0.21	0.18	0.10

TABLE 5. Empirical MISE: $100 \times \mathbb{E}\|\widehat{f}_{\widehat{m},(1)} - f'\|^2$ (left) and $100 \times \mathbb{E}\|\widehat{f}'_{\widehat{h}} - f'\|^2$ (right) for $R = 100$ in the Laguerre case.

5. PROOFS

In the sequel C denotes a generic constant whose value may change from line to line and whose dependency is sometimes given in indexes.

5.1. **Proof of Theorem 2.1.** Following (8) we study the variance term, notice that $\mathbb{E}\big[\|\widehat{f}_{m,(d)} - f_m^{(d)}\|^2\big] = \sum_{j=0}^{m-1} \mathrm{Var}(\widehat{a}_j^{(d)})$. By definition of $\widehat{a}_j^{(d)}$ given in (7), we have

$$\mathrm{Var}(\widehat{a}_j^{(d)}) = \mathrm{Var}\left(\frac{(-1)^d}{n}\sum_{i=1}^n \varphi_j^{(d)}(X_i)\right) = \frac{1}{n}\mathrm{Var}(\varphi_j^{(d)}(X_1)) = \frac{1}{n}\mathbb{E}[(\varphi_j^{(d)}(X_1))^2] - \frac{a_j^2(f^{(d)})}{n}. \tag{30}$$

Clearly, $\sum_{j=0}^{m-1} a_j^2(f^{(d)}) = \|f_m^{(d)}\|^2$. In the sequel we denote by $V_{m,d}$ the quantity

$$V_{m,d} = \sum_{j=0}^{m-1} \mathbb{E}[(\varphi_j^{(d)}(X_1))^2]. \tag{31}$$

The remaining of the proof consists in showing that under (9) we have $V_{m,d} \leqslant cm^{d+1/2}$. For that, write

$$V_{m,d} = \sum_{j=0}^{m-1} \int (\varphi_j^{(d)}(x))^2 f(x)dx = \left(\sum_{j=0}^{d-1} \int (\varphi_j^{(d)}(x))^2 f(x)dx + \sum_{j=d}^{m-1} \int (\varphi_j^{(d)}(x))^2 f(x)dx \right), \tag{32}$$

where

$$\sum_{j=0}^{d-1} \int (\varphi_j^{(d)}(x))^2 f(x)dx \leqslant \sum_{j=0}^{d-1} \|\varphi_j^{(d)}\|_\infty^2 := c(d). \tag{33}$$

To bound the second term in (32), we consider separately Hermite and Laguerre cases.

5.1.1. *The Laguerre case.* We derive from (1) that

$$\ell_j^{(d)}(x) = \sqrt{2} \sum_{k=0}^{d} (-1)^{d-k} \binom{d}{k} L_j^{(k)}(2x) e^{-x}.$$

Using Koekoek (1990), Equation 2.10, we derive

$$L_j^{(k)}(x) = \frac{d^k}{dx^k} L_j(x) = (-1)^k L_{j-k,(k)}(x), \quad \text{where} \quad L_{p,(\delta)}(x) = \frac{1}{p!} e^x x^{-\delta} \frac{d^p}{dx^p} \left(x^{\delta+p} e^{-x} \right) \mathbf{1}_{\delta \leqslant p}.$$

Moreover, introduce the orthonormal basis on $\mathbb{L}^2(\mathbb{R}^+)$ $(\ell_{k,(\delta)})_{0 \leqslant k < \infty}$ by

$$\ell_{k,(\delta)}(x) = 2^{\frac{\delta+1}{2}} \left(\frac{k!}{\Gamma(k+\delta+1)} \right)^{1/2} L_{k,(\delta)}(2x) x^{\frac{\delta}{2}} e^{-x}. \tag{34}$$

Therefore, $(L_j(2x))^{(k)} = 2^k L_{j-k,(k)}(2x) \mathbf{1}_{j \geqslant k}$, so that

$$\ell_j^{(d)}(x) = (-1)^d \sum_{k=0}^{d} \binom{d}{k} 2^{\frac{k}{2}} x^{-k/2} \left(\frac{j!}{(j-k)!} \right)^{\frac{1}{2}} \ell_{j-k,(k)}(x), \tag{35}$$

where $\ell_{j,(\delta)}$ is defined in (34). Using the Cauchy Schwarz inequality in (35), we derive that

$$\begin{aligned} \sum_{j=d}^{m-1} \int_0^\infty [\ell_j^{(d)}(x)]^2 f(x)dx &\leqslant 3^d \sum_{j=d}^{m-1} \sum_{k=0}^{d} \binom{d}{k} \frac{j!}{(j-k)!} \int_0^{+\infty} x^{-k} [\ell_{j-k,(k)}(x)]^2 f(x)dx \\ &\leqslant C_d \sum_{j=d}^{m-1} \sum_{k=0}^{d} j^d \int_0^{+\infty} x^{-k} (\ell_{j-k,(k)}(x/2))^2 f(x/2)dx. \end{aligned}$$

Now we rely on the following Lemma, proved in Appendix A.

Lemma 5.1. *Let $j \geqslant k \geqslant 0$ and suppose that $\mathbb{E}[X^{-k-1/2}] < +\infty$, it holds, for a positive constant C depending only on k, that*

$$\int_0^{+\infty} x^{-k} \left[\ell_{j-k,(k)}(x/2) \right]^2 f(x/2)dx \leqslant \frac{C}{\sqrt{j}}.$$

From Lemma 5.1, we obtain

$$\sum_{j=d}^{m-1}\int(\ell_j^{(d)}(x))^2f(x)dx\leqslant C\sum_{j=d}^{m-1}\sum_{k=0}^{d}j^{d-1/2}\leqslant Cm^{d+1/2}.$$

Plugging this and (33) in (32), gives the result (10) and Theorem 2.1 in the Laguerre case.

5.1.2. *The Hermite case.* We first introduce a useful technical result, its proof is given in Appendix A.

Lemma 5.2. *Let h_j given in (3), the d-th derivative of h_j is such that*

$$h_j^{(d)}=\sum_{k=-d}^{d}b_{k,j}^{(d)}h_{j+k},\quad \textit{where}\quad b_{k,j}^{(d)}=\mathcal{O}(j^{d/2}),\quad j\geqslant d\geqslant|k|. \tag{36}$$

Using successively Lemma 5.2, the Cauchy Schwarz inequality and Lemma 8.5 in Comte and Genon-Catalot (2018) (using that $\mathbb{E}[|X_1|^{2/3}]<\infty$), we obtain, for $k+j$ large enough,

$$\begin{aligned}\sum_{j=d}^{m-1}\int(h_j^{(d)}(x))^2f(x)dx\leqslant&(2d+1)\sum_{j=d}^{m-1}\sum_{k=-d}^{d}(b_{k,j}^{(d)})^2\int h_{j+k}(x)^2f(x)dx\leqslant d(2d+1)^2\sum_{k=-d}^{d}\sum_{j=d}^{m-1}cj^{d-\frac12}\\ \leqslant&c'(d)m^{d+\frac12}.\end{aligned} \tag{37}$$

Plugging (37) and (33) in (32) leads to inequality (10) and Theorem 2.1 in the Hermite case.

5.2. **Proof of Proposition 2.1.** We build a lower bound for (8). Recalling (30) and notation $V_{m,d}=\sum_{j=0}^{m-1}\mathbb{E}[(\varphi_j^{(d)}(X_1))^2]$, to establish Proposition 2.1, we have to build a minorant for $V_{m,d}$. We consider separately the Laguerre and Hermite cases.

5.2.1. *The Laguerre case.* Using (35), we have

$$\begin{aligned}\ell_j^{(d)}(x)=&(-1)^d2^{d/2}x^{-d/2}\Big(\frac{j!}{(j-d)!}\Big)^{1/2}\ell_{j-d,(d)}(x)+(-1)^d\sum_{k=0}^{d-1}\binom{d}{k}2^{\frac{k}{2}}x^{-k/2}\left(\frac{j!}{(j-k)!}\right)^{\frac12}\ell_{j-k,(k)}(x)\\ :=&T_1(x)+T_2(x).\end{aligned}$$

It follows that

$$\int_0^{+\infty}(\ell_j^{(d)})^2(x)f(x)dx\geqslant\int_0^{+\infty}T_1(x)^2f(x)dx+2\int_0^{+\infty}T_1(x)T_2(x)f(x)dx:=E_1+E_2.$$

For the first term, as (**A**1) ensures that f is a continuous density, there exist $0\leqslant a<b$ and $c>0$, such that $\inf_{a\leqslant x\leqslant b}f(x)\geqslant c>0$. We derive

$$E_1\geqslant 2^d\frac{j!}{(j-d)!}\int_0^{+\infty}x^{-d}\ell_{j-d,(d)}^2(x)f(x)dx\geqslant c2^d(j-d)^db^{-d}\int_a^b\ell_{j-d,(d)}^2(x)dx.$$

By Theorem 8.22.5 in Szegö (1959), for $\delta>-1$ an integer, and for $\underline{b}/j\leqslant x\leqslant\bar{b}$, where $\underline{b}$, $\bar{b}$ are arbitrary positive constants, it holds

$$\ell_{j,(\delta)}(x)=\mathfrak{d}(jx)^{-\frac14}\left(\cos(2\sqrt{2}\sqrt{jx}-\frac{\delta\pi}{2}-\frac{\pi}{4})+(jx)^{-\frac12}\mathcal{O}(1)\right), \tag{38}$$

where $\mathcal{O}(1)$ is uniform on $[\underline{b}/j,\bar{b}]$ and $\mathfrak{d}=2^{1/4}/\sqrt{\pi}$. It follows that,

$$\ell_{j,(\delta)}^2(x)=\frac{\mathfrak{d}^2}{2}(jx)^{-\frac12}\left[1+\cos(4\sqrt{2}\sqrt{jx}-\delta\pi-\frac{\pi}{2})\right]+(jx)^{-1}\mathcal{O}(1).$$

We derive that $\int_a^b \ell^2_{j-d,(d)}(x)dx \geqslant C(j-d)^{-1/2}$, after a change of variable $y=\sqrt{x}$, for some positive constant C depending on a,b and d. Consequently, it holds

$$E_1 \geqslant C(j-d)^{d-\frac{1}{2}} \geqslant C' j^{d-\frac{1}{2}}, \quad \forall j \geqslant 2d, \tag{39}$$

where C' depends on a,b,c and d. For the second term, we have

$$\begin{aligned}|E_2| &\leqslant 2\int_0^{+\infty} |T_1(x)T_2(x)|f(x)dx \\ &\leqslant 2j^{\frac{d}{2}} j^{\frac{d-1}{2}} \sum_{k=0}^{d-1} \binom{d}{k} 2^{\frac{k+d}{2}} \left[\int_0^{+\infty} x^{-d}\ell^2_{j-d,(d)}(x)f(x)dx + \int_0^{+\infty} x^{-k}\ell^2_{j-k,(k)}(x)f(x)dx\right].\end{aligned}$$

By Lemma 5.1, it follows that

$$|E_2| \leqslant Cj^{\frac{d}{2}} j^{\frac{d-1}{2}} j^{-\frac{1}{2}} \sum_{k=0}^{d-1} \binom{d}{k} 2^{\frac{k+d}{2}} \leqslant Cj^{d-1}.$$

This together with (39), lead to $\int_0^{+\infty} (\ell_j^{(d)})^2(x)f(x)dx \geqslant C' j^{d-\frac{1}{2}}, \quad j \geqslant 2d$ where C depends on a,b,c and d. We derive

$$V_{m,d} \geqslant Cm^{d+\frac{1}{2}}, \tag{40}$$

which ends the proof in the Laguerre case.

5.2.2. *The Hermite Case.* The proof is similar to the Laguerre case. Consider the following expression of h_j (see Szegö (1959), p.248):

$$h_j(x) = \lambda_j \cos\left((2j+1)^{\frac{1}{2}}x - \frac{j\pi}{2}\right) + \frac{1}{(2j+1)^{\frac{1}{2}}}\xi_j(x), \quad \forall x \in \mathbb{R}, \tag{41}$$

where $\lambda_j = |h_j(0)|$ for j even or $\lambda_j = |h_j'(0)|/(2j+1)^{1/2}$ for j odd and

$$\xi_j(x) = \int_0^x \sin\left((2j+1)^{\frac{1}{2}}(x-t)\right) t^2 h_j(t)dt.$$

By Stirling Formula, it holds

$$\lambda_{2j} = \frac{(2j)!^{\frac{1}{2}}}{2^j j! \pi^{1/4}} \sim \pi^{-1/2} j^{-1/4} \text{ and } \lambda_{2j+1} = \lambda_{2j}\frac{\sqrt{2j+1}}{\sqrt{2j+3/2}} \sim \pi^{-1/2} j^{-1/4}. \tag{42}$$

Differentiating (41), we get

$$h_j^{(d)}(x) = \lambda_j (2j+1)^{\frac{d}{2}} \cos\left((2j+1)^{\frac{1}{2}}x - \frac{j\pi}{2} + \frac{d\pi}{2}\right) + \frac{1}{\sqrt{2j+1}}\xi_j^{(d)}(x).$$

Note that if $d=2$ it holds

$$\xi_j^{(2)}(x) = \sqrt{2j+1}x^2 h_j(x) - (2j+1)\xi_j(x). \tag{43}$$

From (**A**1), there exists $a<b$ and $c>0$ such that $\inf_{a\leqslant x\leqslant b} f(x) \geqslant c > 0$. It follows

$$\begin{aligned}\int_{\mathbb{R}} h_j^{(d)}(x)^2 f(x)dx \geqslant & c(2j+1)^d \lambda_j^2 \int_a^b \cos^2\left((2j+1)^{\frac{1}{2}}x - (j+d)\frac{\pi}{2}\right)dx \\ &+ 2c\lambda_j (2j+1)^{\frac{d-1}{2}} \int_a^b \cos\left((2j+1)^{\frac{1}{2}}x - (j+d)\frac{\pi}{2}\right)\xi_j^{(d)}(x)dx := E_1 + E_2.\end{aligned}$$

For the first term, using $\cos^2(x) = (1+\cos(2x))/2$ and (42), we get

$$E_1 = c(2j+1)^d \lambda_j^2 \left(\frac{b-a}{2} + \mathcal{O}(\frac{1}{\sqrt{j}}) \right) \geqslant c' j^{d-\frac{1}{2}} \left(\frac{b-a}{2} + \mathcal{O}(\frac{1}{\sqrt{j}}) \right).$$

For the second term we first show that

$$\forall x \in [a,b],\ \forall j \geqslant 0,\ \forall d \geqslant 0,\ \xi_j^{(d)}(x) = \mathcal{O}(j^{d/2}). \tag{44}$$

To establish (44) we first note, using (43), that for $d \geqslant 2$, $\forall x \in \mathbb{R}$,

$$\xi_j^{(d)}(x) + (2j+1)\xi_j^{(d-2)}(x) = (\xi_j^{(2)}(x) + (2j+1)\xi_j(x))^{(d-2)} = \sqrt{2j+1}(x^2 h_j(x))^{(d-2)} =: \Psi_{j,d}(x).$$

Together with Lemma 5.2, one easily obtains by induction that $\forall x \in [a,b]$, $\forall j \geqslant 0$, $\Psi_{j,d}(x) = \mathcal{O}(j^{\frac{d-1}{2}})$. The latter result gives $\xi_j^{(d)} = -j\xi_j^{(d-2)} + \Psi_{j,d}$ and an immediate induction on d leads to (44). Injecting this in E_2 gives, together with (42), $|E_2| \leqslant C j^{d-\frac{3}{4}}$, for a positive constant C depending on a, b, c and d. Gathering the bound on E_1 and E_2 lead to

$$\int_{\mathbb{R}} h_j^{(d)}(x)^2 f(x)dx \geqslant c' j^{d-\frac{1}{2}} \left(\frac{b-a}{2} + \mathcal{O}(\frac{1}{\sqrt{j}}) \right) - \mathcal{O}(j^{d-\frac{3}{4}}) \geqslant C'_d j^{d-\frac{1}{2}},$$

and

$$V_{m,d} \geqslant c_d m^{d+\frac{1}{2}}, \tag{45}$$

which ends the proof of the Hermite case.

5.3. **Proof of Theorem 2.2.** We apply Theorem 2.7 in Tsybakov (2009). We start by the construction of a family of hypotheses $(f_\theta)_\theta$. The construction is inspired by Belomestny et al. (2017). Define f_0 by

$$f_0(x) = P(x)\mathbf{1}_{]0,1[}(x) + \frac{1}{2}x\mathbf{1}_{[1,2]}(x) + Q(x)\mathbf{1}_{]2,3]}(x), \tag{46}$$

where P and Q are positive polynomials, for $0 \leqslant k \leqslant s$, $P^{(k)}(0) = Q^{(k)}(3) = 0$, $P^{(k)}(1) = \lim_{x\downarrow 1}(x/2)^{(k)}$, $Q^{(k)}(2) = \lim_{x\uparrow 2}(x/2)^{(k)}$ and finally $\int_0^1 P(x)dx = \int_2^3 Q(x)dx = \frac{1}{8}$. Consider f_θ defined as a perturbation of f_0

$$f_\theta(x) = f_0(x) + \delta K^{-(\gamma+d)} \sum_{k=0}^{K-1} \theta_{k+1}\psi\big((x-1)(K+1)-k\big), \text{ with } K \in \mathbb{N}, \tag{47}$$

for some $\delta > 0$, $\theta = (\theta_1, \dots, \theta_K) \in \{0,1\}^K$, $\gamma > 0$ and ψ which is supported on $[1,2]$, admits bounded derivatives up to order s and is such that $\int_1^2 \psi(x)dx = 0$. Theorem 2.2 is a consequence of the following Lemma.

Lemma 5.3. *(i). Let $s \geqslant d$, $\forall\ \theta \in \{0,1\}^K$, there exist δ small enough and $\gamma > 0$ such that f_θ is density. There exists $D > 0$ such that f_θ belongs to $W_H^s(D)$. If in addition $\gamma \geqslant s-d$, f_θ belongs to $W_L^s(D)$.*

(ii). Let M an integer, for all $j < l \leqslant M$, $\forall \theta^{(j)}$, $\theta^{(l)}$ in $\{0,1\}^K$, it holds $\|f_{\theta^{(j)}}^{(d)} - f_{\theta^{(l)}}^{(d)}\|^2 \geqslant C\delta^2 K^{-2\gamma}$.

(iii). For δ small enough, $K = n^{1/(2\gamma+2d+1)}$ and for all $(\theta^{(j)})_{1\leqslant j\leqslant M} \in (\{0,1\}^K)^M$, it holds

$$\frac{1}{M}\sum_{j=1}^{M} \chi^2\left({f_{\theta^{(j)}}}^{\otimes n}, {f_0}^{\otimes n}\right) \leqslant \alpha M,$$

where $0 < \alpha < 1/8$ and $\chi^2(g,h)$ denotes the χ^2 divergence between the distributions g and h.

Choosing $\gamma = s - d$, $K = n^{1/(2\gamma+2d+1)}$ and δ small enough, we derive from Lemma 5.3 that,

$$\|f^{(d)}_{\theta^{(j)}} - f^{(d)}_{\theta^{(l)}}\|^2 \geqslant C\delta^2 n^{-2\frac{(s-d)}{2s+1}}, \quad \forall \theta^{(j)}, \theta^{(l)} \in \{0,1\}^K.$$

The announced result is then a consequence of Theorem 2.7 in Tsybakov (2009).

5.4. **Proof of Theorem 2.3.** Consider the contrast function defined as follows:

$$\gamma_{n,d}(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^n (-1)^d t^{(d)}(X_i), \quad t \in \mathbb{L}^2(\mathbb{R}),$$

for which $\widehat{f}_{m,(d)} = \underset{t\in S_m}{\operatorname{argmin}} \gamma_{n,d}(t)$ (see (7)) and $\gamma_n(\widehat{f}_{m,(d)}) = -\|\widehat{f}_{m,(d)}\|^2$. For two functions $t, s \in \mathbb{L}^2(\mathbb{R})$, consider the decomposition:

$$\gamma_{n,d}(t) - \gamma_{n,d}(s) = \|t - f^{(d)}\|^2 - \|s - f^{(d)}\|^2 - 2\nu_{n,d}(t-s), \tag{48}$$

where

$$\nu_{n,d}(t) = \frac{1}{n}\sum_{i=1}^n \left((-1)^d t^{(d)}(X_i) - \langle t, f^{(d)}\rangle\right).$$

By (17), it holds for all $m \in \mathcal{M}_{n,d}$, that $\gamma_{n,d}(\widehat{f}_{\widehat{m}_n,(d)}) + \widehat{\operatorname{pen}}_d(\widehat{m}_n) \leqslant \gamma_{n,d}(f^{(d)}_m) + \widehat{\operatorname{pen}}_d(m)$. Plugging this in (48) yields, for all $m \in \mathcal{M}_{n,d}$,

$$\|\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}\|^2 \leqslant \|f^{(d)}_m - f^{(d)}\|^2 + \widehat{\operatorname{pen}}_d(m) + 2\nu_{n,d}\left(\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}_m\right) - \widehat{\operatorname{pen}}_d(\widehat{m}_n). \tag{49}$$

Note that for $t \in \mathbb{L}^2(\mathbb{R})$, $\nu_{n,d}(t) = \|t\|\nu_{n,d}\big(t/\|t\|\big) \leqslant \|t\| \sup_{s\in S_m + S_{\widehat{m}}, \|s\|=1} |\nu_{n,d}(s)|$. Consequently, using $2xy \leqslant x^2/4 + 4y^2$, we obtain

$$2\nu_{n,d}\left(\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}_m\right) \leqslant \frac{1}{2}\|\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}\|^2 + \frac{1}{2}\|f^{(d)}_m - f^{(d)}\|^2 + 4\sup_{t\in S_m + S_{\widehat{m}}, ||t||=1} |\nu_{n,d}(t)|^2. \tag{50}$$

It follows from (49) and (50) that:

$$\frac{1}{2}\|\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}\|^2 \leqslant \frac{3}{2}\|f^{(d)}_m - f^{(d)}\|^2 + \widehat{\operatorname{pen}}_d(m) + 4\sup_{t\in S_m + S_{\widehat{m}}, ||t||=1} |\nu_{n,d}(t)|^2 - \widehat{\operatorname{pen}}_d(\widehat{m}_n).$$

Introduce the function $p(m,m') = 4\frac{V_{m\vee m',d}}{n}$, we get, after taking the expectation,

$$\begin{aligned}\frac{1}{2}\mathbb{E}\left[\|\widehat{f}_{\widehat{m}_n,(d)} - f^{(d)}\|^2\right] \leqslant \frac{3}{2}\|f^{(d)}_m - f^{(d)}\|^2 + \operatorname{pen}_d(m) + 4\mathbb{E}\left[\left(\sup_{t\in S_m + S_{\widehat{m}}, ||t||=1} |\nu_{n,d}(t)|^2 - p(m,\widehat{m}_n)\right)_+\right] \\ + \mathbb{E}[4p(m,\widehat{m}_n) - \operatorname{pen}_d(\widehat{m}_n)] + \mathbb{E}\left[(\operatorname{pen}_d(\widehat{m}_n) - \widehat{\operatorname{pen}}_d(\widehat{m}_n))_+\right].\end{aligned}$$

The remaining of the proof is a consequence of the following Lemma.

Lemma 5.4. *Under the assumptions of Theorem 2.3, the following hold.*

(i) There exists a constant Σ_1 such that:

$$\mathbb{E}\left[\left(\sup_{t\in S_m + S_{\widehat{m}}, ||t||=1} |\nu_{n,d}(t)|^2 - p(m,\widehat{m}_n)\right)_+\right] \leqslant \frac{\Sigma_1}{n}.$$

(ii) There exists a constant Σ_2 such that:

$$\mathbb{E}\left[(\operatorname{pen}_d(\widehat{m}_n) - \widehat{\operatorname{pen}}_d(\widehat{m}_n))_+\right] \leqslant \frac{1}{2}\mathbb{E}[\operatorname{pen}_d(\widehat{m}_n)] + \frac{\Sigma_2}{n}.$$

Lemma 5.4 yields

$$\frac{1}{2}\mathbb{E}\left[\|\widehat{f}_{\widehat{m}_n,(d)}-f^{(d)}\|^2\right]\leqslant\frac{3}{2}\|f_m^{(d)}-f^{(d)}\|^2+\mathrm{pen}_d(m)+4\frac{\Sigma_1}{n}+\mathbb{E}[4p(m,\widehat{m}_n)-\frac{1}{2}\mathrm{pen}_d(\widehat{m}_n)]+\frac{\Sigma_2}{n}.$$

Next, for $\kappa\geqslant 32=:\kappa_0$, we have, $4p(m,\widehat{m}_n)\leqslant \mathrm{pen}_d(\widehat{m}_n)/2+\mathrm{pen}_d(m)/2$. Therefore, we derive

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}_n,(d)}-f^{(d)}\|^2\right]\leqslant 3\|f_m^{(d)}-f^{(d)}\|^2+3\mathrm{pen}_d(m)+2\frac{4\Sigma_1+\Sigma_2}{n},\quad \forall m\in\mathcal{M}_{n,d}.$$

Taking the infimum on $\mathcal{M}_{n,d}$, $C=3$ and $C'=2(4\Sigma_1+\Sigma_2)/n$ completes the proof.

5.5. **Proof of Proposition 3.1.** First, it holds that

$$\begin{aligned}\mathbb{E}\Big[\|(\widehat{f}_m)'-f'\|^2\Big]&\leqslant 2\Big[\|(f_m)'-f'\|^2+\mathbb{E}[\|(\widehat{f}_m)'-(f_m)'\|^2]\Big]\\&=2\int_0^{+\infty}(\sum_{j\geqslant m}a_j(f)\ell_j'(x))^2dx+2\mathbb{E}\left[\|\sum_{j=0}^{m-1}(\widehat{a}_j^{(0)}-a_j(f))\ell_j'\|^2\right].\end{aligned}$$

For the first bias term, we derive from (2) that $\langle\ell_j',\ell_k'\rangle=2+4j\wedge k$ for $j\neq k$ and $\langle\ell_j',\ell_j'\rangle=1+4j$, and we derive that

$$\int_0^{+\infty}(\sum_{j\geqslant m}a_j(f)\ell_j'(x))^2dx=\sum_{j\geqslant m}a_j(f)^2(1+4j)+2\sum_{m\leqslant j<k}a_j(f)a_k(f)(2+4j).$$

First, for f in $W_L^s(D)$, we have

$$\sum_{j\geqslant m}a_j(f)^2(1+4j)\leqslant m^{-s}\sum_{j\geqslant m}j^sa_j(f)^2+4m^{-s+1}\sum_{j\geqslant m}j^sa_j(f)^2\leqslant 5Dm^{-s+1},$$

and by the Cauchy-Schwarz inequality, it holds for a positive constant C,

$$\sum_{m\leqslant j<k}a_j(f)a_k(f)\leqslant\left(\sum_{m\leqslant j<k}j^sa_j(f)^2k^sa_k(f)^2\right)^{\frac12}\left(\sum_{m\leqslant j<k}j^{-s}k^{-s}\right)^{\frac12}\leqslant\sum_{j\geqslant m}j^sa_j(f)^2\sum_{j\geqslant m}j^{-s}\leqslant DCm^{-s+1}$$

$$\sum_{m\leqslant j<k}j|a_j(f)a_k(f)|\leqslant\sum_{j\geqslant m}j|a_j(f)|\left(\sum_{k\geqslant j}k^sa_k(f)^2\sum_{k\geqslant j}k^{-s}\right)^{\frac12}\leqslant\sqrt{DC}\sum_{j\geqslant m}j^{\frac s2-s+\frac32}|a_j(f)|\leqslant DCm^{-s+2}.$$

Thus, it comes

$$2\|(f_m)'-f'\|^2\leqslant Cm^{-(s-2)},\tag{51}$$

where $C>0$ depends on D. Second, for the variance term, straightforward computations lead to

$$\mathbb{E}\Big[\|\sum_{j=0}^{m-1}(\widehat{a}_j^{(0)}-a_j(f))\ell_j'\|^2\Big]=\frac{1}{n}\int_0^{+\infty}\mathrm{Var}(\sum_{j=0}^{m-1}\ell_j(X_1)\ell_j'(x))dx\leqslant\frac{1}{n}\int_0^{+\infty}\mathbb{E}\left[(\sum_{j=0}^{m-1}\ell_j(X_1)\ell_j'(x))^2\right]dx.$$

By the orthonormality of $(\ell_j)_j$ and (**A2**), we obtain

$$\int_0^{+\infty}\mathbb{E}\left[(\sum_{j=0}^{m-1}\ell_j(X_1)\ell_j'(x))^2\right]dx\leqslant\|f\|_\infty\sum_{j,k=0}^{m-1}\int_0^{+\infty}\int_0^{+\infty}\ell_j(u)\ell_j'(x)\ell_k(u)\ell_k'(x)dudx=\|f\|_\infty\sum_{j=0}^{m-1}(1+4j)\leqslant 3\|f\|_\infty m^2.$$

From this and (51), the result follows.

5.6. **Proof of Proposition 3.2.** By the Pythagoras Theorem, we have the bias-variance decomposition $\mathbb{E}\big[\|\widetilde{f}'_{m,K}-f'\|^2\big]=\|f'-f'_m\|^2+\mathbb{E}\big[\|\widetilde{f}'_{m,K}-f'_m\|^2\big]$. As $\ell_j(0)=\sqrt{2}$, it follows that

$$\widetilde{f}'_{m,K}-f'_m=\sum_{j=0}^{m-1}\left[-\sqrt{2}(\widehat{f}_K(0)-f(0))-\frac{1}{n}\sum_{i=1}^n(\ell'_j(X_i)-\mathbb{E}[\ell'_j(X_i)])\right]\ell_j.$$

From the orthonormality of $(\ell_j)_j$, it follows

$$\begin{aligned}\mathbb{E}\big[\|\widetilde{f}'_{m,K}-f'_m\|^2\big]&=\sum_{j=0}^{m-1}\mathbb{E}\left[-\sqrt{2}(\widehat{f}_K(0)-f(0))-\frac{1}{n}\sum_{i=1}^n(\ell'_j(X_i)-\mathbb{E}[\ell'_j(X_i)])\right]^2\\&\leqslant 4m\mathbb{E}\left[(\widehat{f}_K(0)-f(0))^2\right]+2\sum_{j=0}^{m-1}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n(\ell'_j(X_i)-\mathbb{E}[\ell'_j(X_i)])\right)^2\right].\end{aligned}$$

Finally, using that the $(X_i)_i$ are i.i.d. lead to the result in the second variance term.

5.7. **Proof of Theorem 3.1.** We have the decomposition:

$$\gamma_n(t)-\gamma_n(s)=\|t-f'\|^2-\|s-f'\|^2-2\langle s-t,f'\rangle-\frac{2}{n}\sum_{i=1}^n(s'-t')(X_i)-2(s(0)-t(0))\widehat{f}_K(0)$$

and as $\langle t,f'\rangle=-t(0)f(0)-\int t'f$, we get

$$\gamma_n(t)-\gamma_n(s)=\|t-f'\|^2-\|s-f'\|^2-2\nu_n(s-t)-2(s(0)-t(0))(\widehat{f}_K(0)-f(0)),\tag{52}$$

$$\text{with}\qquad \nu_n(t)=\frac{1}{n}\sum_{i=1}^n(t'(X_i)-\langle t',f\rangle.$$

First note that for

$$f'_{m,K}=\sum_{j=0}^{m-1}a^{(1)}_{j,K}\ell_j,\quad a^{(1)}_{j,K}=\mathbb{E}[\widehat{a}^{(1)}_{j,K}]=\langle f',\ell_j\rangle+\ell_j(0)(f(0)-\mathbb{E}[\widehat{f}_K(0)],$$

it holds that

$$\begin{aligned}\|f'-f'_{m,K}\|^2&=\left\|\sum_{j=0}^{\infty}\langle f',\ell_j\rangle\ell_j-\sum_{j=0}^{m-1}\langle f',\ell_j\rangle\ell_j-\sum_{j=0}^{m-1}\ell_j(0)\big(f(0)-\mathbb{E}[\widehat{f}_K(0)]\big)\ell_j\right\|^2\\&=\sum_{j\geqslant m}\langle f',\ell_j\rangle^2+2\sum_{j=0}^{m-1}\big(f(0)-\mathbb{E}[\widehat{f}_K(0)]\big)^2=\|f'-f'_m\|^2+2m\big(f(0)-\mathbb{E}[\widehat{f}_K(0)]\big)^2.\end{aligned}$$

Let us start by writing that, by definition of $\widehat{m}_K$, it holds, $\forall m\in\mathcal{M}_n$,

$$\gamma_n(\widehat{f}'_{\widehat{m}_K,K})+\mathrm{pen}_K(\widehat{m}_K)\leqslant\gamma_n(f'_{m,K})+\mathrm{pen}_K(m),$$

which yields, with (52) and notations introduced in (28),

$$\begin{aligned}\|\widehat{f}'_{\widehat{m}_K,K}-f'\|^2&\leqslant\|f'_{m,K}-f'\|^2+\mathrm{pen}_K(m)+2\nu_n(f'_{m,K}-\widehat{f}'_{\widehat{m}_K,K})-\mathrm{pen}_1(\widehat{m}_K)\\&\quad+2(f'_{m,K}(0)-\widehat{f}'_{\widehat{m}_K,K}(0))(\widehat{f}_K(0)-f(0))-\mathrm{pen}_{2,K}(\widehat{m}_K)\\&\leqslant\|f'_{m,K}-f'\|^2+\mathrm{pen}_K(m)+\frac{1}{4}\|f'_{m,K}-\widehat{f}'_{\widehat{m}_K,K}\|^2+8\sup_{t\in S_{m\vee\widehat{m}_K}}\nu_n^2(t)-\mathrm{pen}_1(\widehat{m}_K)\\&\quad+16(m\vee\widehat{m}_K)[\widehat{f}_K(0)-f(0)]^2-\mathrm{pen}_{2,K}(\widehat{m}_K).\end{aligned}$$

To get the last line, we write that, for any $t \in S_m$,

$$|t(0)| = \sqrt{2}\left|\sum_{j=0}^{m-1} a_j(t)\right| \leqslant \sqrt{2m\sum_{j=0}^{m} a_j^2(t)} \leqslant \sqrt{2m}\|t\|,$$

and we use that $2xy \leqslant x^2/8 + 8y^2$ for all real x, y. We obtain

$$\begin{aligned}\frac{1}{2}\|\widehat{f}'_{\widehat{m}_K,K} - f'\|^2 \quad \leqslant \quad & \frac{3}{2}\|f'_{m,K} - f'\|^2 + \mathrm{pen}_K(m) + 16m(\widehat{f}_K(0) - f(0))^2 \\ & +8\left(\sup_{t\in S_{m\vee\widehat{m}_K},\|t\|=1} \nu_n^2(t) - p_1(m\vee\widehat{m}_K)\right)_+ + 8p_1(m\vee\widehat{m}_K) - \mathrm{pen}_1(\widehat{m}_K) \\ & +16\widehat{m}_K\left[(\widehat{f}_K(0) - f(0))^2 - c_2(\|f\|_\infty \vee 1)K\frac{\log(n)}{n}\right], \end{aligned} \tag{53}$$

where

$$p_1(m) = \mathtt{b}(1 + 2\log(n))\|f\|_\infty \frac{m^2}{n}, \quad \mathtt{b} > 0.$$

The following Lemma can be proved using the Talagrand Inequality (see Section B.2).

Lemma 5.5. *Under the assumptions of Theorem 3.1, and* $\mathtt{b} \geqslant 6$,

$$\sum_{m\in\mathcal{M}_n} \mathbb{E}\left[\sup_{t\in S_m,\|t\|=1} \nu_n^2(t) - p_1(m)\right]_+ \leqslant \frac{c}{n}.$$

It follows that

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in S_{m\vee\widehat{m}_K},\|t\|=1} \nu_n^2(t) - p_1(m\vee\widehat{m}_K)\right)_+ \quad &\leqslant \quad \sum_{m'\in\mathcal{M}_n} \mathbb{E}\left(\sup_{t\in S_{m'\vee m},\|t\|=1} \nu_n^2(t) - p_1(m\vee m')\right)_+ \\ &\leqslant \quad \frac{c}{n}. \end{aligned} \tag{54}$$

This implies that $8p_1(m\vee\widehat{m}_K) \leqslant \mathrm{pen}_1(m) + \mathrm{pen}_1(\widehat{m}_K)$ for c_1 –defined in (28)– large enough. Moreover, let $\mathtt{a} > 0$ and

$$\Omega_K := \left\{\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i^K - \mathbb{E}(Z_i^K))\right| \leqslant \sqrt{\mathtt{a}(\|f\|_\infty \vee 1)\frac{K\log(n)}{n}}\right\},$$

where $Z_i^K := \sum_{j=0}^{K-1} \ell_j(X_i)$. To apply the Bernstein Inequality (see Section B.3), we compute $s^2 = \|f\|_\infty K$ and $b = \sqrt{2}K$ and note that $K\log(n)/n \leqslant 1$. Thus, we get that there exist constants c_0, c such that

$$\text{For } \mathtt{a} > c_0, \quad \mathbb{P}(\Omega_K^c) \leqslant \frac{c}{n^4}. \tag{55}$$

On Ω_K, it holds that

$$(\widehat{f}_K(0) - f_K(0))^2 = \left(\frac{1}{n}\sum_{i=1}^{n}(Z_i^K - \mathbb{E}(Z_i^K))\right)^2 \leqslant 2\mathtt{a}(\|f\|_\infty \vee 1)K\frac{\log(n)}{n}. \tag{56}$$

For any $K_n \leqslant [n/\log(n)]$ satisfying condition (26), we have

$$\begin{aligned} & \mathbb{E}\left\{\widehat{m}_{K_n}\left[(\widehat{f}_{K_n}(0)-f(0))^2 - c_2(\|f\|_\infty \vee 1)K_n\frac{\log(n)}{n}\right]\right\} \\ \leqslant\ & \mathbb{E}\left\{\widehat{m}_{K_n}\left[(\widehat{f}_{K_n}(0)-f_{K_n}(0))^2 - (c_2-2)(\|f\|_\infty \vee 1)K_n\frac{\log(n)}{n}\right]\right\} \end{aligned}$$

Now we note that $|\widehat{f}_K(x)| \leqslant 2K$ for all $x \in \mathbb{R}^+$ and any integer K and by using the definition of (56), provided that $c_2 > 2\mathtt{a}+2$, we obtain

$$\begin{aligned} & \mathbb{E}\left\{\widehat{m}_{K_n}\left[(\widehat{f}_{K_n}(0)-f_{K_n}(0))^2 - (c_2-2)(\|f\|_\infty \vee 1)K_n\frac{\log(n)}{n}\right]\right\} \\ \leqslant\ & \mathbb{E}\left\{\widehat{m}_{K_n}\left[(\widehat{f}_{K_n}(0)-f_{K_n}(0))^2 - (c_2-2)(\|f\|_\infty \vee 1)K_n\frac{\log(n)}{n}\right]\mathbf{1}_{\Omega_{K_n}}\right\} \\ & +\mathbb{E}\left\{\widehat{m}_{K_n}\left[(\widehat{f}_{K_n}(0)-f_{K_n}(0))^2 - (c_2-2)(\|f\|_\infty \vee 1)K_n\frac{\log(n)}{n}\right]\mathbf{1}_{\Omega^c_{K_n}}\right\} \\ \lesssim\ & Cn^{5/2}\mathbb{P}(\Omega^c_{K_n}) \lesssim \frac{1}{n}, \end{aligned}$$

the term on Ω_{K_n} being less than or equal to 0. Plugging this and (54) into (53), we get

$$\mathbb{E}\left(\|\widehat{f}'_{\widehat{m}_K,K} - f'\|^2\right) \leqslant 3\|f'_{m,K} - f'\|^2 + 4\mathrm{pen}_K(m) + 32m(\widehat{f}_K(0)-f(0))^2 + \frac{c}{n}$$

which gives the result of Theorem 3.1. □

Appendix A. Proofs of auxiliary results

A.1. **Proof of Lemma 2.1.** In the Hermite case $\varphi_j = h_j$ and $f : \mathbb{R} \mapsto [0,\infty)$, allowing d successive integration by parts, it holds that

$$a_j(f^{(d)}) = \int_{\mathbb{R}} f^{(d)}(x)h_j(x)dx = \left[\sum_{k=0}^{d-1}(-1)^k f^{(d-1-k)}(x)h_j^{(k)}(x)\right]_{-\infty}^{+\infty} + (-1)^d\int_{\mathbb{R}} h_j^{(d)}(x)f(x)dx. \tag{57}$$

By definition for all $j \geqslant 0$, $h_j(x) = c_jH_j(x)e^{-\frac{x^2}{2}}$ where H_j is a polynomial. Then, its k-th derivative, $0 \leqslant k \leqslant d-1$, is a polynomial multiplied by $e^{-x^2/2}$ and $\lim_{|x|\to+\infty} h_j^{(k)}(x) = 0$. This together with (A2), gives that the bracket in (57) is null and the result follows.
Similarly in the Laguerre case, (57) holds integrating on $[0,\infty)$ instead of $\mathbb{R}$ and replacing h_j by ℓ_j. The term in the bracket is null at 0 from (A3). It is also null at infinity using (A2) together with the fact that ℓ_j are polynomials multiplied by e^{-x} leading similarly to $\lim_{x\to\infty} f^{(d-1-k)}(x)\ell_j^{(k)}(x) = 0$, $0 \leqslant k \leqslant d-1$, $j \geqslant 0$. The result follows.

A.2. **Proof of Lemma 2.2.** We control the quantity

$$\sum_{j\geqslant 0} j^{s-d}\langle f^{(d)}, h_j\rangle^2 = \sum_{j=0}^{d-1} j^{s-d}\langle f^{(d)}, h_j\rangle^2 + \sum_{j\geqslant d} j^{s-d}\langle f^{(d)}, h_j\rangle^2. \tag{58}$$

The first term is a constant which depending on d. For the second term using Lemma 5.2, we obtain

$$\begin{aligned}\sum_{j\geqslant d} j^{s-d}\langle f^{(d)},h_j\rangle^2 &= \sum_{j\geqslant d} j^{s-d}\left(\sum_{k=-d}^{d} b_{k,j}^{(d)}\int h_{j+k}(x)f(x)dx\right)^2\\ &\leqslant C_d\sum_{j\geqslant d} j^s\sum_{k=-d}^{d}\left(\int h_{j+k}(x)f(x)dx\right)^2 = C_d\sum_{k=-d}^{d}\sum_{j\geqslant d} j^s\langle h_{j+k},f\rangle^2\\ &= C_d\sum_{k=-d}^{d}\left(\sum_{j\geqslant d+k}|j-k|^s\langle h_j,f\rangle^2\right)\leqslant C_d\sum_{k=-d}^{d}\left(\sum_{j\geqslant 0}2^s j^s\langle h_j,f\rangle^2\right) = (2d+1)2^s DC_d.\end{aligned}$$

Inserting this in (58), we obtain the announced result.

A.3. **Proof of Lemma 2.3.** We establish the result for $d=1$, the general case is an immediate consequence. It follows from the definition of $\widetilde{W}_L^s(D)$ that $(\theta')^{(j)}$, $0\leqslant j\leqslant s-1$ are in $C([0,\infty))$. Moreover, it holds that $x\mapsto x^{k/2}(\theta')^{(j)}(x)\in\mathbb{L}^2(\mathbb{R}^+)$ for all $0\leqslant j<k\leqslant s-1$. The case $k=j$ is obtained using that $\theta^{(j)}$ is continuous on $C([0,\infty))$ and that $x\mapsto x^{(j+1)/2}(\theta')^{(j)}(x)\in\mathbb{L}^2(\mathbb{R}^+)$. It follows that

$$\begin{aligned}|||\theta'|||_s^2 &= \sum_{j=0}^{s-1}\left\|x^{j/2}\sum_{k=0}^{j}\binom{j}{k}(\theta')^{(k)}\right\|^2\leqslant 2\sum_{j=0}^{s-1}\left\|x^{j/2}\sum_{k=0}^{j-1}\binom{j}{k}(\theta')^{(k)}\right\|^2+2\sum_{j=0}^{s-1}\left\|x^{j/2}(\theta')^{(j)}\right\|^2\\ &\leqslant C+2\sum_{j=0}^{s-1}\|x^{(j+1)/2}(\theta')^{(j)}(x)\|^2<\infty,\end{aligned}$$

where C depends on D. Finally, using the equivalence of the norms $|.|_s$ and $|||.|||_s$, the value of D' follows from the latter inequality.

A.4. **Proof of Lemma 5.1.** Consider the decomposition

$$\int_0^{+\infty} x^{-k}(\ell_{j-k,(k)}(x/2))^2 f(x/2)dx = \sum_{i=1}^{6} I_i,$$

where for $\nu=4j-2k+2$, $j\geqslant k$, we used the decomposition $(0,\infty)=(0,\frac{1}{\nu}]\cup(\frac{1}{\nu},\frac{\nu}{2}]\cup(\frac{\nu}{2},\nu-\nu^{1/3}]\cup(\nu-\nu^{1/3},\nu+\nu^{/13}]\cup(\nu+\nu^{1/3},3\nu/2]\cup(3\nu/2,\infty)$. Using Askey and Wainger (1965) (see Appendix B.1) and straightforward inequalities give

$$\begin{aligned}I_1 &\lesssim \int_0^{\frac{1}{\nu}} x^{-k}(x\nu)^k f(x/2)dx\leqslant\int_0^{\frac{1}{\nu}} x^{-k}(x\nu)^{-1/2}f(x/2)dx\lesssim\nu^{-1/2}\mathbb{E}[X^{-k-1/2}],\\ I_2 &\lesssim \int_{1/\nu}^{\frac{\nu}{2}} x^{-k}((x\nu)^{-1/4})^2 f(x/2)dx=\nu^{-1/2}\int_{1/\nu}^{\frac{\nu}{2}} x^{-k-1/2}f(x/2)dx\leqslant\nu^{-1/2}\mathbb{E}[X^{-k-1/2}],\\ I_3 &\lesssim \int_{\frac{\nu}{2}}^{\nu-\nu^{1/3}} x^{-k}(\nu^{-1/4}(\nu-x)^{-1/4})^2 f(x/2)dx=\nu^{-1/2}\int_{\frac{\nu}{2}}^{\nu-\nu^{1/3}} x^{-k}(\nu-x)^{-1/2}f(x/2)dx\lesssim\nu^{-1/2},\\ I_4 &\lesssim \int_{\nu-\nu^{1/3}}^{\nu+\nu^{1/3}} x^{-k}(\nu^{-1/3})^2 f(x/2)dx\leqslant\nu^{-2/3}\int_{\frac{\nu}{2}}^{\nu+\nu^{1/3}} x^{-k}f(x/2)dx\lesssim\nu^{-1/2}\nu^{-k}\leqslant\nu^{-1/2},\end{aligned}$$

$$I_5 \lesssim \int_{\nu+\nu^{1/3}}^{3\nu/2} x^{-k}\nu^{-1/2}(x-\nu)^{-1/2}e^{-2\gamma_1\nu^{-1/2}(x-\nu)^{3/2}}f(x/2)dx \lesssim \nu^{-1/2}\nu^{-1/6}\nu^{-k}\int f(x/2)dx \lesssim \nu^{-1/2},$$
$$I_6 \lesssim \int_{3\nu/2}^{+\infty} x^{-k}e^{-2\gamma_2 x}f(x/2)dx \lesssim e^{-3\gamma_2\nu/2} = \mathcal{O}(\nu^{-1/2}).$$

Gathering these inequalities give the announced result.

A.5. **Proof of Lemma 5.2.** The result is obtained by induction on d. If $d=1$, h'_j is given by (5), with $b^{(1)}_{-1,j-1} = j^{1/2}/\sqrt{2}$, $b_{0,j}=0$ and $b^{(1)}_{1,j} = (j+1)^{1/2}/\sqrt{2}$, $\forall j \geqslant 1$. Thus, it holds $b^{(1)}_{k,j} = \mathcal{O}(j^{1/2})$ and (36) is satisfied for $d=1$. Let $\mathrm{P}(d)$ the proposition given by Equation (36) and assume $\mathrm{P}(d)$ holds and we establish $\mathrm{P}(d+1)$. It holds using successively $\mathrm{P}(d)$ and (5) that

$$\begin{aligned} h^{(d+1)}_j(x) &= \sum_{k=-d}^{d} b^{(d)}_{k,j}\left[\frac{\sqrt{j+k}}{\sqrt{2}}h_{j+k-1} - \frac{\sqrt{j+k+1}}{\sqrt{2}}h_{j+k+1}\right] \\ &= \sum_{k'=-d-1}^{d-1} b^{(d)}_{k'+1,j}\frac{\sqrt{j+k'+1}}{\sqrt{2}}h_{j+k'} - \sum_{k'=-d+1}^{d+1} b^{(d)}_{k'-1,j}\frac{\sqrt{j+k'}}{\sqrt{2}}h_{j+k'} := \sum_{k=-d-1}^{d+1} b^{(d+1)}_{k,j}h_{j+k'}, \end{aligned}$$

where $\quad b^{(d)}_{k,j} = \mathcal{O}(j^{d/2})$, $\forall j \geqslant d \geqslant |k|$ $\quad$ and $\quad b^{(d+1)}_{k,j} = b^{(d)}_{k+1,j}\frac{\sqrt{j+k+1}}{\sqrt{2}}\mathbf{1}_{|k|\leqslant d-1} - b^{(d)}_{k-1,j}\frac{\sqrt{j+k}}{\sqrt{2}}\mathbf{1}_{|k|\leqslant d+1}$.

It follows that $|b^{(d+1)}_{k,j}| \leqslant 2\sqrt{(j+d+1)/2}\,j^{\frac{d}{2}} \leqslant C_d j^{\frac{d+1}{2}}$, $|k| \leqslant d \leqslant j$, which completes the proof.

A.6. **Proof of Lemma 5.3.**

A.6.1. *Proof of part (i).* By construction, f_0 is positive and $\forall \theta \in \{0,1\}^K$, $\int f_\theta(x)dx = \int f_0(x)dx = 1$. It remains to show that f_θ is nonnegative. The supports of $\big(\psi((.-1)(K+1)-k)\big)_{0\leqslant k\leqslant K-1}$ are disjoint and are in $[1,2]$, then $f_\theta(x) \geqslant 0$ for all $x \in \mathbb{R}\backslash[1,2]$. Now, for all x in $[1,2]$, there exists k_0 such that

$$f_\theta(x) = \frac{x}{2} + \delta K^{-\gamma-d}\theta_{k_0+1}\psi\big((x-1)(K+1)-k_0\big) \geqslant \frac{1}{2} - \delta\|\psi\|_\infty K^{-\gamma-d},$$

which is nonnegative if $\delta \leqslant \|\psi\|_\infty^{-1}/2$. Now, let us show that f_0 and f_θ belong to $W^s(D)$.
The Laguerre case. We use the equivalent norm $|||.|||_s$ of $|.|_s$ (see (13)) and start with f_0. As f_0 is s-th differentiable, we have

$$|||f_0|||_s^2 = \sum_{j=0}^{s}\int_0^3\left(x^{j/2}\sum_{k=0}^{j}\binom{j}{k}f_0^{(k)}(x)\right)^2 dx \leqslant \sum_{j=0}^{s}2^j\sum_{k=0}^{j}\binom{j}{k}\int_0^3(x^{j/2}f_0^{(k)}(x))^2dx.$$

As $\int_0^3(x^{j/2}f_0^{(k)}(x))^2dx \leqslant c(s) < +\infty$, $0 \leqslant k \leqslant j \leqslant s$, it follows $|f|_s^2 \leqslant 3D/4$, D depends on s. For f_θ, we have

$$\begin{aligned} |||f_\theta - f_0|||_s^2 =& \delta^2K^{-2\gamma-2d}\sum_{j=0}^{s}\int_1^2\left(\sum_{l=0}^{j}\binom{j}{l}\sum_{k=0}^{K-1}x^{j/2}\theta_{k+1}(K+1)^l\psi^{(l)}\big((x-1)(K+1)-k\big)\right)^2dx \\ \leqslant& \delta^2K^{-2\gamma-2d}\sum_{j=0}^{s}\sum_{l=0}^{j}2^j\binom{j}{l}\int_1^2\left(x^{j/2}\sum_{k=0}^{K-1}\theta_{k+1}(K+1)^l\psi^{(l)}\big((x-1)(K+1)-k\big)\right)^2dx. \end{aligned}$$

Using that $\psi^{(l)}\big((x-1)(K+1)-k\big), \psi^{(l)}\big((x-1)(K+1)-k'\big)$ have disjoint supports for $k \neq k'$ and that $\psi^{(l)}$ are bounded by c, we get after the change of variable $y=(x-1)(K+1)-k$,

$$|||f_\theta - f_0|||_s^2 \leqslant \delta^2 2^{3s} c^2 K^{-2\gamma-2d} \sum_{j=0}^{s} \sum_{k=0}^{K-1} (K+1)^{2j-1} \leqslant C(s)\delta^2 K^{-2\gamma-2d+2s}.$$

For $\gamma \geqslant s-d$ and δ small enough, it holds $|f_\theta - f_0|_s \leqslant D/4$ and therefore $|f_\theta|_s \leqslant |f_\theta - f_0|_s + |f_0|_s \leqslant D$.
The Hermite case. The usual Sobolev space W^s, if s is integer, is defined by

$$W^s = \{f \in \mathbb{L}^2(\mathbb{R}), f \text{ admits derivatives up to order } s, \text{such that } |||f|||_{s,sob} = \sum_{j=0}^{s} \|f^{(j)}\|^2 < +\infty\}.$$

It is proved in Bongioanni and Torrea (2006) that: if $f \in W^s$ has compact support, then f belongs to W_H^s. By construction f_0 and f_θ have a compact support and as they admit derivatives up to order s, they belong to W^s. It follows that f_0 and f_θ belong W_H^s. This completes the proof of (i).

A.6.2. *Proof of part (ii).* As for $k \neq k'$, $\psi\big((.-1)(K+1)-k\big), \psi\big((.-1)(K+1)-k'\big)$ have disjoint supports, we have, $\forall \theta^{(j)}, \theta^{(l)} \in \{0,1\}^K$,

$$\begin{aligned}\|f^{(d)}_{\theta^{(j)}} - f^{(d)}_{\theta^{(l)}}\|^2 =& \delta^2 \sum_{k=0}^{K-1} (\theta^{(j)}_{k+1} - \theta^{(l)}_{k+1})^2 K^{-2\gamma-2d}(K+1)^{2d} \int_1^2 \psi^{(d)}\big((x-1)(K+1)-k\big)^2 dx \\ \geqslant& \delta^2 \|\psi^{(d)}\|^2 K^{-2\gamma-1} \rho(\theta^{(j)}, \theta^{(l)}),\end{aligned}$$

where $\rho(\theta^{(j)}, \theta^{(l)}) = \sum_{k=1}^K \mathbf{1}_{\theta^{(j)}_k \neq \theta^{(l)}_k}$ is the Hamming distance. By Lemma 2.7 in Tsybakov (2009), for $K \geqslant 8$, there exist $\{\theta^{(0)}, \dots, \theta^{(M)}\}$ in $\{0,1\}^K$ such that

$$\rho(\theta^{(j)}, \theta^{(l)}) \geqslant \frac{K}{8}, \ \forall \quad 0 \leqslant j < l \leqslant M \text{ and } M \geqslant 2^{\frac{K}{8}}.$$

Thus, it holds, $\forall \theta^{(j)}, \theta^{(l)} \in \{0,1\}^K$, $\|f^{(d)}_{\theta^{(j)}} - f^{(d)}_{\theta^{(l)}}\|^2 \geqslant \delta^2/8 \|\psi^{(d)}\|^2 K^{-2\gamma}$, which gives (ii) if we set $C = \|\psi^{(d)}\|^2/8$.

A.6.3. *Proof of part (iii).* For M integer and $(\theta^{(j)})_{1 \leqslant j \leqslant M}$ in $(\{0,1\}^K)^M$, we have

$$\sum_{j=1}^{M} \chi^2\left({f_{\theta^{(j)}}}^{\otimes n}, {f_0}^{\otimes n}\right) = \sum_{j=1}^{M} \left(\left(1+\chi^2(f_{\theta^{(j)}}, f_0)\right)^n - 1\right) = \sum_{j=1}^{M} \left(e^{n\log(1+\chi^2(f_{\theta^{(j)}}, f_0))} - 1\right). \tag{59}$$

Since $f_0 \geqslant c > 0$ on $[1,2]$, it holds for any $\theta \in \{0,1\}^K$,

$$\begin{aligned}\chi^2(f_\theta, f_0) = \int_1^2 \frac{(f_\theta(x) - f_0(x))^2}{f_0(x)} dx \leqslant \frac{\delta^2}{c} K^{-2\gamma-2d} \sum_{k=0}^{K-1} \int_1^2 \Big(\psi\big((x-1)(K+1)-k\big)\Big)^2 dx \\ \leqslant \frac{\delta^2}{c} K^{-2\gamma-2d} \|\psi\|^2 \leqslant \frac{8\delta^2}{c\log 2} \log(M) K^{-2\gamma-2d-1},\end{aligned}$$

where we used that $M \geqslant 2^{\frac{K}{8}}$. Consequently, using in (59) that $\log(1+x) \leqslant x$, for any $x \geqslant 0$, and the latter inequality, give

$$\frac{1}{M} \sum_{j=1}^{M} \chi^2\left({f_{\theta^{(j)}}}^{\otimes n}, {f_0}^{\otimes n}\right) \leqslant e^{n\frac{8\delta^2}{c\log 2}\log(M) K^{-2\gamma-2d-1}} - 1.$$

For δ well chosen and $K = n^{1/(2\gamma+2d+1)}$, comes the result.

A.7. **Proof of Lemma 5.4.**

A.7.1. *Proof of part (i).* First, it holds that

(60)
$$\mathbb{E}\left[\left(\sup_{t\in S_m+S_{\widehat{m}},||t||=1}|\nu_{n,d}(t)|^2-p(m,\widehat{m}_n)\right)_+\right]\leqslant\sum_{m'\in\mathcal{M}_{n,d}}\mathbb{E}\left[\left(\sup_{t\in S_m+S_{m'},||t||=1}|\nu_{n,d}(t)|^2-p(m,m')\right)_+\right],$$

which we bound applying a Talagrand Inequality (see Section B.2). Following notations of Section B.2, we have three terms H^2, v and M_1 to compute. Let us denote by $m^*=m\vee m'$, for $t\in S_m+S_{m'}$, $\|t\|=1$, it holds

$$\|t\|^2=\|\sum_{j=0}^{m^*-1}a_j\varphi_j\|^2=\sum_{j=0}^{m^*-1}a_j^2=1.$$

Computing H^2. By the linearity of $\nu_{n,d}$ and the Cauchy Schwarz inequality, we have

$$\nu_{n,d}(t)^2=\left(\sum_{j=0}^{m^*-1}a_j\nu_{n,d}(\varphi_j)\right)^2\leqslant\sum_{j=0}^{m^*-1}a_j^2\sum_{j=0}^{m^*-1}\nu_{n,d}^2(\varphi_j)=\sum_{j=0}^{m^*-1}\nu_{n,d}^2(\varphi_j).$$

One can check that the latter is an equality for $a_j=\nu_{n,d}(\varphi_j)$. Therefore, taking expectation, it follows

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in S_m^*,||t||=1}\nu_{n,d}^2(t)\right]&=\sum_{j=0}^{m^*-1}\mathrm{Var}(\nu_{n,d}(\varphi_j))=\frac{1}{n}\sum_{j=0}^{m^*-1}\mathrm{Var}(\varphi_j^{(d)}(X_1))\\&\leqslant\frac{1}{n}\sum_{j=0}^{m^*-1}\mathbb{E}\left[\varphi_j^{(d)}(X_1)^2\right]=\frac{V_{m^*,d}}{n}=:H^2.\end{aligned}$$

Computing v. It holds for $t\in S_m+S_{m'}$, $\|t\|=1$,

$$\begin{aligned}\mathrm{Var}\left((-1)^dt^{(d)}(X_1)\right)&\leqslant\int t^{(d)}(x)^2f(x)dx=\int\left(\sum_{j=0}^{m^*-1}a_j\varphi_j^{(d)}(x)\right)^2f(x)dx\\&\leqslant 2\int\left(\sum_{j=0}^{d-1}a_j\varphi_j^{(d)}(x)\right)^2f(x)dx+2\int\left(\sum_{j=d}^{m^*-1}a_j\varphi_j^{(d)}(x)\right)^2f(x)dx.\end{aligned}\tag{61}$$

The first term of the previous inequality is a constant depending only on d. For the second term, we consider separately the Laguerre and Hermite cases.
The Laguerre Case ($\varphi_j=\ell_j$). Using (35) and the Cauchy Schwarz inequality, it holds that

$$\begin{aligned}\int\left(\sum_{j=d}^{m^*-1}a_j\ell_j^{(d)}(x)\right)^2f(x)dx&\leqslant 3^d\sum_{k=0}^{d}\binom{d}{k}\int\left(\sum_{j=d}^{m^*-1}a_j\left(\frac{j!}{(j-k)!}\right)^{\frac{1}{2}}x^{-\frac{k}{2}}\ell_{j-k,(k)}(x)\right)^2f(x)dx\\&\leqslant 3^d\sum_{k=0}^{d}\binom{d}{k}\sup_{x\in\mathbb{R}^+}\frac{f(x)}{x^k}\sum_{j=d}^{m^*-1}a_j^2\frac{j!}{(j-k)!}\leqslant C(d)(m^*)^d,\end{aligned}\tag{62}$$

where we used the orthonormality of $(\ell_{j,(k)})_{j\geqslant 0}$ and where $C(d)$ is a constant depending only on d and $\sup_{x\in\mathbb{R}^+}\frac{f(x)}{x^k}$.

The Hermite case ($\varphi_j = h_j$). Similarly, using Lemma 5.2 and the orthonormality of h_j, it follows

$$\int \left(\sum_{j=d}^{m^*-1} a_j h_j^{(d)}(x)\right)^2 f(x)dx \leqslant (2d+1) \sum_{k=-d}^{d} \int \left(\sum_{j=d}^{m^*-1} a_j b_{k,j} h_{j+k}(x)\right)^2 f(x)dx$$
$$\leqslant C(d)\|f\|_\infty (m^*)^d. \tag{63}$$

Plugging (62) or (63) in (61), we set in the two cases $v := c_1(m^*)^d$ where c_1 depends on d and either on $\sup_{x\in\mathbb{R}^+} \frac{f(x)}{x^k}$ (Laguerre case) or $\|f\|_\infty$ (Hermite case).

Computing M_1**.** The Cauchy Schwarz Inequality and $\|t\| = 1$ give

$$\|(-1)^d t^{(d)}\|_\infty = \|\sum_{j=0}^{m^*-1} (-1)^d a_j \varphi_j^{(d)}\|_\infty \leqslant \sup_{x\in\mathbb{R}} \sqrt{\sum_{j=0}^{m^*-1} \varphi_j^{(d)}(x)^2}. \tag{64}$$

The Laguerre case. We use the following Lemma whose proof is a consequence of (2) and an induction on d.

Lemma A.1. *For ℓ_j given in (1), the d-th derivative of ℓ_j is such that $\|\ell_j^{(d)}\|_\infty \leqslant C_d(j+1)^d$, $\forall j \geqslant 0$ and where C_d is a positive constant depending on d.*

Using Lemma A.1, we obtain

$$\sum_{j=0}^{m^*-1} \ell_j^{(d)}(x)^2 \leqslant C_d^2 (m^*)^{2d+1}. \tag{65}$$

The Hermite case. The d first terms in the sum in (64) can be bounded by a constant depending only on d. For the remaining terms, Lemma 5.2 and $\|h_j\|_\infty \leqslant \phi_0$ (see (4)) give

$$\sum_{j=d}^{m^*-1} [h_j^{(d)}(x)]^2 \leqslant C_d^2 \phi_0^2 \sum_{k=-d}^{d} \sum_{j=d}^{m^*-1} j^d \leqslant C(m^*)^{d+1}, \tag{66}$$

where C is a positive constant depending on d and ϕ_0.
Injecting either (65) or (66) in (64), we set $M_1 = \mathcal{O}(m^{d+\frac{1}{2}})$ in the Laguerre case or $M_1 = \mathcal{O}(m^{\frac{d}{2}+\frac{1}{2}})$ in the Hermite case.

Now, we apply the Talagrand Inequality see Appendix B.2 with $\varepsilon = 1/2$, it follows

$$\mathbb{E}\left[\left(\sup_{t\in S_m+S_{m'},\|t\|=1} |\nu_{n,d}(t)|^2 - 4H^2\right)_+\right] \leqslant \frac{C_1}{n}\left(v\exp\left(-C_2\frac{nH^2}{v}\right) + C_3\frac{M_1^2}{n}\exp\left(-C_4\frac{nH}{M_1}\right)\right)$$
$$:= \frac{C_1}{n}\left(U_d(m^*) + V_d(m^*)\right).$$

The Laguerre Case. We have

$$U_d(m^*) = c_1(m^*)^d \exp\left(-C_2\frac{V_{m^*,d}}{c_1(m^*)^d}\right) \text{ and } V_d(m^*) = C_3c_2\frac{(m^*)^{2d+1}}{n}\exp\left(-C_4\sqrt{n}\frac{\sqrt{V_{m^*,d}}}{c_2(m^*)^{d+\frac{1}{2}}}\right).$$

From (40) and the value of $m_n(d)$, we obtain

$$U_d(m^*) \leqslant c_1(m^*)^d \exp(-C_2' m^{*\frac{1}{2}}) \quad \text{and} \quad V_d(m^*) \leqslant C_3c_2(m^*)^{d+\frac{1}{2}}\exp(-C_4'\sqrt{n}(m^*)^{-\frac{d}{2}-\frac{1}{4}}).$$

Using the value $m_n(d)$, it holds $(m^*)^{d+1/2} \leqslant n/\log^3(n)$, which implies (recall $m^* = m \vee m'$)

$$\sum_{m' \in \mathcal{M}_{n,d}} V_d(m^*) \leqslant C \sum_{m' \in \mathcal{M}_{n,d}} (m^*)^{d+\frac{1}{2}} \exp\left(-C_4 \log^2(n)\right) \leqslant \Sigma_{d,2},$$

where $\Sigma_{d,2}$ is a constant depending only on d. Next, it follows

$$\sum_{m'=1}^{n} U_d(m^*) = \sum_{m'=1}^{m} U_d(m^*) + \sum_{m'=m}^{n} U_d(m^*) = c_1 m^{d+1} \exp(-C_2' m^{\frac{1}{2}}) + \sum_{m'=m}^{n} c_1 (m')^d \exp(-C_2' m'^{\frac{1}{2}}).$$

The function $m \mapsto m^{d+1} \exp(-C_2' m^{\frac{1}{2}})$ is bounded and the sum is finite on m', it holds

$$C_1 \sum_{m'=1}^{n} U_d(m^*) \leqslant \Sigma_{d,1}, \text{ where } \Sigma_{d,1} \text{ depends only on } d.$$

The Hermite case. Only the second term $V_d(m^*)$ changes. Here, it is given by

$$\begin{aligned} V_d(m^*) &= C_3 c_2 \frac{(m^*)^{d+1}}{n} \exp\left(-C_3 \sqrt{n} \frac{\sqrt{V_{m^*,d}}}{c_2 (m^*)^{\frac{d}{2}+\frac{1}{2}}}\right) \leqslant C_3 c_2 (m^*)^{1/2} \exp(-C_4' \sqrt{n} (m^*)^{-\frac{1}{4}}) \\ &\leqslant C_3 c_2 (m^*)^{1/2} \exp(-C_4' (m^*)^{\frac{d}{2}}), \end{aligned}$$

where we used (45) and the value of $m_n(d)$. We derive that $\sum_{m' \in \mathcal{M}_{n,d}} V_d(m^*) \leqslant \Sigma_{d,2}$.

Gathering all terms, it follows

$$\mathbb{E}\left[\left(\sup_{t \in S_m + S_{m'}, ||t||=1} |\nu_{n,d}(t)|^2 - 4H^2\right)_+\right] \leqslant \frac{\Sigma}{n}, \text{ where } \Sigma = \Sigma_{d,1} + \Sigma_{d,2}$$

Plugging this in (60) gives the announced result.

A.7.2. *Proof of part (ii).* We use the Bernstein Inequality (see Appendix B.3) to prove the result. Define

$$Z_i^{(m)} = \sum_{j=0}^{m-1} (\varphi_j^{(d)}(X_i))^2, \quad \text{then,} \quad \widehat{V}_{m,d} = \frac{1}{n} \sum_{i=1}^{n} Z_i^{(m)}$$

We select s^2 and b such that $\mathrm{Var}(Z_i^{(m)}) \leqslant s^2$ and $|Z_i^{(m)}| \leqslant b$. By the computation of M_1 (see Proof of part (i)), we set $b := C^* m^\alpha$, with $\alpha = 2d+1$ (Laguerre case) or $\alpha = d+1$ (Hermite case), where C^* depends on d. For s^2, using that $\mathrm{Var}(Z_i^{(m)}) \leqslant \mathbb{E}[(Z_i^{(m)})^2] \leqslant b \sum_{j=0}^{m-1} \mathbb{E}\left[(\varphi_j^{(d)}(X_i))^2\right] = C^* m^\alpha V_{m,d} =: s^2$. Applying the Bernstein Inequality, we have for $S_n = n(\widehat{V}_{m,d} - V_{m,d})$

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| \geqslant \sqrt{\frac{2xC^* m^\alpha V_{m,d}}{n}} + \frac{C^* m^\alpha x}{3n}\right) \leqslant 2e^{-x}, \quad \forall x > 0. \tag{67}$$

Choose $x = 2\log(n)$ and define the set

$$\Omega := \left\{ m \in \mathcal{M}_{n,d},\ \frac{1}{n}|S_n| \leqslant 2\sqrt{\frac{C^* m^\alpha \log(n) V_{m,d}}{n}} + \frac{2C^* m^\alpha \log(n)}{3n} \right\}.$$

Consider the decomposition,

$$\mathbb{E}\left[(\mathrm{pen}_d(\widehat{m}_n) - \widehat{\mathrm{pen}}_d(\widehat{m}_n))_+\right] \leqslant \mathbb{E}\left[(\mathrm{pen}_d(\widehat{m}_n) - \widehat{\mathrm{pen}}_d(\widehat{m}_n))_+ \mathbf{1}_\Omega\right] + \mathbb{E}\left[(\mathrm{pen}_d(\widehat{m}_n) - \widehat{\mathrm{pen}}_d(\widehat{m}_n))_+ \mathbf{1}_{\Omega^c}\right].$$

Using $2xy \leqslant x^2 + y^2$, we have on Ω

$$|\widehat{V}_{\widehat{m},d} - V_{\widehat{m},d}| \leqslant \frac{V_{\widehat{m},d}}{2} + \frac{2C^*\widehat{m}^\alpha \log(n)}{n} + \frac{2C^*\widehat{m}^\alpha \log(n)}{3n} = \frac{V_{\widehat{m},d}}{2} + \frac{8}{3}\frac{C^*\widehat{m}^\alpha \log(n)}{n}.$$

The constraint on m_n gives $\widehat{m}^{d+1/2} \leqslant Cn/(\log(n))^2$ together with (40) giving $V_{\widehat{m},d} \geqslant c^*\widehat{m}^{d+1/2}$ give for $\alpha = 2d+1$ (Laguerre case) that $\frac{8C^*}{3}\frac{\widehat{m}^\alpha \log(n)}{n} \leqslant \frac{8CC^*}{3c^*}\frac{V_{\widehat{m},d}}{\log(n)} \leqslant \frac{V_{\widehat{m},d}}{4}$, for n large enough and

$$\mathbb{E}\left[(\mathrm{pen}_d(\widehat{m}_n) - \widehat{\mathrm{pen}}_d(\widehat{m}_n))_+ \mathbf{1}_\Omega\right] \leqslant \frac{3}{4}\mathbb{E}[\mathrm{pen}_d(\widehat{m}_n)]. \tag{68}$$

In the Hermite case ($\alpha = d+1$) computations are similar as $\widehat{m}^{d+1} \leqslant \widehat{m}^{2d+1}$. For the control on Ω^c, we write, using (67),

$$\mathbb{E}\left[(\mathrm{pen}_d(\widehat{m}_n) - \widehat{\mathrm{pen}}_d(\widehat{m}_n))_+ \mathbf{1}_{\Omega^c}\right] \leqslant 2\kappa\mathbb{P}(\Omega^c) \leqslant 2\kappa \sum_{m\in\mathcal{M}_{n,d}} 2e^{-2\log(n)} := \frac{\Sigma_2}{n}. \tag{69}$$

Gathering (68) and (69), we get the desired result.

Appendix B. Some inequalities

B.1. **Asymptotic Askey and Wainger formula.** From Askey and Wainger (1965), we have for $\nu = 4k + 2\delta + 2$, and k large enough

$$|\ell_{k,(\delta)}(x/2)| \leqslant C \begin{cases} a) & (x\nu)^{\delta/2} & \text{if } 0 \leqslant x \leqslant 1/\nu \\ b) & (x\nu)^{-1/4} & \text{if } 1/\nu \leqslant x \leqslant \nu/2 \\ c) & \nu^{-1/4}(\nu - x)^{-1/4} & \text{if } \nu/2 \leqslant x \leqslant \nu - \nu^{1/3} \\ d) & \nu^{-1/3} & \text{if } \nu - \nu^{1/3} \leqslant x \leqslant \nu + \nu^{1/3} \\ e) & \nu^{-1/4}(x-\nu)^{-1/4}e^{-\gamma_1\nu^{-1/2}(x-\nu)^{3/2}} & \text{if } \nu + \nu^{1/3} \leqslant x \leqslant 3\nu/2 \\ f) & e^{-\gamma_2 x} & \text{if } x \geqslant 3\nu/2 \end{cases}$$

where γ_1 and γ_2 are positive and fixed constants.

B.2. **A Talagrand Inequality.** The Talagrand inequalities have been proven in Talagrand (1996) and reworked by Ledoux (9597). This version is given in Klein and Rio (2005). Let $(X_i)_{1\leqslant i\leqslant n}$ be independent real random variables and

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^n (t(X_i) - \mathbb{E}[t(X_i)]),$$

for t in $\mathcal{F}$ a class of measurable functions. If there exist M_1, H and v such that:

$$\sup_{t\in\mathcal{F}} \|t\|_\infty \leqslant M_1, \quad \mathbb{E}[\sup_{t\in\mathcal{F}} \mid \nu_n(t) \mid] \leqslant H, \quad \sup_{t\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(t(X_i)) \leqslant v,$$

then, for $\varepsilon > 0$,

$$\mathbb{E}\left[\left(\sup_{t\in\mathcal{F}} |\nu_n^2(t)| - 2(1+2\varepsilon)H^2\right)_+\right] \leqslant \frac{4}{K_1}\left(\frac{v}{n}\exp\left(-K_1\varepsilon\frac{nH^2}{v}\right) + \frac{49M_1^2}{K_1C^2(\varepsilon)n^2}\exp\left(-K_1'C(\varepsilon)\sqrt{\varepsilon}\frac{nH}{M_1}\right)\right),$$

where $C(\varepsilon) = (\sqrt{1+\varepsilon} - 1) \wedge 1$, $K_1 = 1/6$ and K_1' a universal constant.

B.3. **Bernstein Inequality (Massart (2007)).** Let $X_1, \dots X_n$, n independent real random variables. Assume there exist two constants s^2 and b, such that $\mathrm{Var}(X_i) \leqslant s^2$ and $|X_i| \leqslant b$. Then, for all x positive, we have

$$\mathbb{P}\left(|S_n| \geqslant \sqrt{2ns^2x} + \frac{bx}{3}\right) \leqslant 2e^{-x}, \text{ with } S_n = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]).$$

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.

Askey, R. and Wainger, S. (1965). Mean convergence of expansions in Laguerre and Hermite series. *Amer. J. Math.*, 87:695–708.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470.

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

Belomestny, D., Comte, F., Genon-Catalot, V., et al. (2017). Correction to: Nonparametric laguerre estimation in the multiplicative censoring model. *Electronic Journal of Statistics*, 11(2):4845–4850.

Bhattacharya, P. (1967). Estimation of a probability density function and its derivatives. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 373–382.

Bongioanni, B. and Torrea, J. L. (2006). Sobolev spaces associated to the harmonic oscillator. *Proc. Indian Acad. Sci. Math. Sci.*, 116(3):337–360.

Bongioanni, B. and Torrea, J. L. (2009). What is a Sobolev space for the Laguerre function systems? *Studia Math.*, 192(2):147–172.

Chacón, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*, 7:499–532.

Chacón, J. E., Duong, T., and Wand, M. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, pages 807–840.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799.

Comte, F. and Genon-Catalot, V. (2018). Laguerre and Hermite bases for inverse problems. *J. Korean Statist. Soc.*, 47(3):273–296.

Comte, F. and Marie, N. (2019). Bandwidth Selection for the Wolverton-Wagner Estimator. working paper or preprint.

Efromovich, S. (1998). Simultaneous sharp estimation of functions and their derivatives. *Ann. Statist.*, 26(1):273–278.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2016). Non-parametric inference for density modes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(1):99–126.

Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press.

Indritz, J. (1961). An inequality for Hermite polynomials. *Proc. Amer. Math. Soc.*, 12:981–983.

Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077.

Koekoek, R. (1990). Generalizations of laguerre polynomials. *Journal of Mathematical Analysis and Applications*, 153(2):576–590.

Lacour, C., Massart, P., and Rivoirard, V. (2017). Estimator selection: a new method with applications to kernel density estimation. *Sankhya A*, 79(2):298–335.

Ledoux, M. (1995/97). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87.

Markovich, L. (2016). Gamma kernel estimation of the density derivative on the positive semi-axis by dependent data. *REVSTAT–Statistical Journal*, 14(3):327–348.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Park, C. and Kang, K.-H. (2008). Sizer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis*, 52(8):3954–3970.

Plancade, S. (2009). Estimation of the density of regression errors by pointwise model selection. *Math. Methods Statist.*, 18(4):341–374.

Rao, B. L. S. P. (1996). Nonparametric estimation of the derivatives of a density by the method of wavelets. *Bull. Inform. Cybernet.*, 28(1):91–100.

Schmisser, E. (2013). Nonparametric estimation of the derivatives of the stationary density for stationary processes. *ESAIM Probab. Stat.*, 17:33–69.

Schuster, E. F. (1969). Estimation of a probability density function and its derivatives. *The Annals of Mathematical Statistics*, 40(4):1187–1195.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, pages 177–184.

Singh, R. (1979). Mean squared errors of estimates of a density and its derivatives. *Biometrika*, 66(1):177–180.

Singh, R. S. (1977). Applications of estimators of a density and its derivatives to certain statistical problems. *J. Roy. Statist. Soc. Ser. B*, 39(3):357–363.

Szegö, G. (1959). *Orthogonal polynomials*. American Mathematical Society Colloquium Publications, Vol. 23. Revised ed. American Mathematical Society, Providence, R.I.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563.

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