Should we estimate a product of density functions by a product of estimators ?

Fabienne Comte, Céline Duval

Should we estimate a product of density functions by a product of estimators ?

F. Comte and C. Duval

Université Paris Cité, CNRS, MAP5 UMR 8145,
F-75006 Paris, France. e-mail: fabienne.comte@u-paris.fr

Université de Lille, CNRS, UMR 8524,
Laboratoire Paul Painlevé, F-59000 Lille, France e-mail: celine.duval@univ-lille.fr

Abstract: In this paper, we consider the inverse problem of estimating the product fg of two densities, given a d-dimensional n-sample of i.i.d. observations drawn from each distribution. We propose a general method of estimation encompassing both projection estimators with model selection device and kernel estimators with bandwidth selection strategies. The procedures do not consist in making the product of each density estimator, but in plugging an overfitted estimator of one of the two densities, in an estimator based on the second sample. Our findings are a first step toward a better understanding of the good performances of overfitting in regression Nadaraya-Watson estimator.

MSC2020 subject classifications: Primary 62G05, 62G07.
Keywords and phrases: Bandwidth selection, Density estimation, Kernel estimator, Model selection, Nonparametric estimation, Penalized Comparison to Overfitting, Projection estimator.

1. Introduction

In this work, we consider that we have n observations $X_i, i = 1, \dots, n$ in $\mathbb{R}^d$ independent and identically distributed (i.i.d.) with density f and independent from n additional observations $Y_i, i = 1, \dots, n$ in $\mathbb{R}^d$, i.i.d. with density g. We study the question of estimating the product function fg from these observations. Note that the resulting function is not − in general − a density, and none of the observations are directly related to this product. In that sense, we face an inverse problem. Our framework contains the case where $f = g$ and the goal is to estimate f^2 by splitting a $2n$-sample. These quantities may be of interest in some testing problems or as a first step for estimating the $\mathbb{L}^2$-norm of f, see Laurent and Massart (2000); other product problems are considered in Butucea *et al.* (2018).

However, we must explain that we considered this problem as a simplified setting (a toy-problem, in some sense) for a more complicated question. Let us explain it. Consider a regression model in dimension $d = 1$ with $Y_i = b(X_i) + \varepsilon_i$ with i.i.d. and independent sequences $(X_i)_{1\leq i\leq n}$ and $(\varepsilon_i)_{1\leq i\leq n}$. The question is to estimate the regression function $b(\cdot)$ from observations $(X_i, Y_i)_{1\leq i\leq n}$. A

popular proposal is the Nadaraya-Watson estimator (see Györfi *et al.* (2002))

$$\widehat{b}_h(x) = \frac{\frac{1}{nh}\sum_{i=1}^n Y_i K\left(\frac{X_i-x}{h}\right)}{\frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)} = \sum_{i=1}^n w_{n,i,h} Y_i, \quad w_{n,i,h} = \frac{K\left(\frac{X_i-x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)},$$

where K is a kernel and h a bandwidth parameter. This estimator can be seen as a weighted combination of the Y_i's (second equality) or as a ratio of an estimator of bf, where f is still the density of the X_i's, divided by an estimator of f (first equality). In this last case, it is not clear that the same bandwidth h must be chosen for the numerator and the denominator. Surprisingly, Comte and Marie (2021) proposed sophisticated strategies for these two terms, but noticed in the simulation experiments that, if the numerical results obtained for both functions separately were excellent, the performance of the ratio was almost systematically defeated by the single bandwidth method selected from a least squares criterion relying on the weighted view of the question. The unique bandwidth selected in this case is small, but the ratio of these two bad and overfitted estimators is undoubtedly very good, at least for not too high noise level (see a related discussion in Section 4.1 of Bartlett *et al.* (2019)). This is why we wondered if the product of two functional estimators was a good estimator of the product of two functions; we took these functions as densities for simplicity. Thus our motivation is mainly theoretical, but we believe that the question is of general interest.

Now, let us see why making a product of density estimators can be seen as an inadequate (sub-optimal) strategy. Assume that we set $\widehat{fg} := \widehat{f} \times \widehat{g}$, where $\widehat{f}$ and $\widehat{g}$ are minimax optimal estimators of f and g respectively. To get an upper bound result, there is no other way than to separate the role of each estimate : both individual risks of $\widehat{f}$ and $\widehat{g}$ would emerge. Then, the resulting rate is the slowest between the rates of estimation of f and g: it is the rate induced by the less regular density between f and g, say g without loss or generality for the remaining of this discussion. Clearly, this is not optimal if the product fg is more regular than g. For instance, if $d = 1$ and for f a $\beta(p, p)$ density with $p \geq 2$, p integer and g a uniform density, i.e. a $\beta(1, 1)$. Then on $\mathbb{R}$, f has regularity $p - 1$ and g regularity 0, but $fg = f$ has regularity $p - 1$. Therefore, one can wonder if in these cases it is possible to build an estimator directly adapted to the regularity of the product fg.

A related disadvantage of an upper bound separating the roles of f and g is that it does not treat this problem as an inverse problem : both individual regularities of f and g intervene whereas one expects that the sole regularity of fg should matter. Especially since, depending on the regularity classes which are considered, there is often no universal rule relating the regularities of f and g to the one of the product.

To complete this discussion, notice that it is easy to derive a lower bound result, inspired by the former example on beta distributions. Denote by $\Sigma(\mathbf{s}, L)$, where $\mathbf{s} = (s_1, \ldots, s_d)$ with positive $s_i, i = 1, \ldots, d$ and L, a ball of radius L in space of functions with regularity $\mathbf{s}$. Then it holds, for any measurable function

T of $(X_i, Y_i)_{1\leq i\leq n}$,

$$\sup_{fg\in\Sigma(\mathbf{s},L)} \|T-fg\|^2 \geq \sup_{\substack{fg\in\Sigma(\mathbf{s},L)\\ \mathrm{Supp} f\subset[0,1]^d\\ g=\mathbf{1}_{[0,1]^d}}} \|T-fg\|^2 = \sup_{\substack{f\in\Sigma(\mathbf{s},L)\\ \mathrm{Supp} f\subset[0,1]^d}} \|T-f\|^2.$$

It follows that

$$\inf_T \sup_{fg\in\Sigma(\mathbf{s},L)} \|T-fg\|^2 \geq \inf_T \sup_{\substack{f\in\Sigma(\mathbf{s},L)\\ \mathrm{Supp} f\subset[0,1]^d}} \|T-f\|^2, \tag{1.1}$$

we recover on the right side the lower bound of the direct density estimation problem. To summarize, if the regularity set $\Sigma(\mathbf{s}, L)$ contains a $[0,1]^d$ supported density f_0, a lower bound for the product is given by a lower bound for the direct estimation of f_0. This is enough to state that the upper bound results presented below are optimal. For instance in dimension $d = 1$, we recover rates in $n^{-\frac{2s}{2s+1}}$ if $(fg) \in \Sigma(s, L)$, a Sobolev class of regularity s, that are minimax.

The plan of the paper is the following. We propose in section 2.1 a general estimation strategy: we define a function estimator of the product fg and prove a non-asymptotic risk bound under general assumptions. This general strategy encompasses projection and kernel methods, for which we check the assumptions and present spectific results in Section 2.2. Then we study in a particular projection case the resulting rate, related to the regularity of fg: it can be reached for a well-chosen dimension of the projection space (see section 2.4). As this choice depends on unknown parameters, we then propose in Section 3.1 a general parameter selection strategy of Goldenshluger and Lepski (2011) type, and prove that the resulting estimator automatically reaches the squared-bias/variance compromise. This general method applies to projection and kernel methods. However in both cases and in dimension 1 alternative methods are used in the simulations, which outperform Goldenshluger and Lepski procedures numerically; they are simpler to calibrate with faster execution times. Note that an invariable difficulty in the theoretical study of these procedures is that even if we have i.i.d. variables, our estimates are built from sums of dependent variables (see e.g. (2.6) or (2.7)). Numerical comparisons of the different methods and associated strategies for product estimators are conducted in section 4, and concur to our theoretical findings. Several additional questions are presented in the concluding remarks of section 5. Lastly, proofs are gathered in section 6 for the results of section 2 and in section 7 for the adaptive results stated in sections 3 and 4.

2. General estimates of the product and examples

2.1. Functional estimator and first risk bound

We consider a general family $\mathcal{K}$ of functions $\mathbb{K} : I^2 \mapsto \mathbb{R}$, where $I \subset \mathbb{R}^d$, that are symmetric (i.e. $\forall x, y \in I, \mathbb{K}(x, y) = \mathbb{K}(y, x)$).

In the sequel, we denote by $\psi_{\mathbb{K}}$ the quantity $\psi_{\mathbb{K}}(x) := \int \mathbb{K}(x,y)\psi(y)dy$ for any function ψ. The collection $\mathcal{K}$ must be chosen such that $\psi_{\mathbb{K}}$ can be a good approximation of ψ. Examples of possible collection $\mathcal{K}$ are provided in Section 2.2.

For such a function $\mathbb{K}^* \in \mathcal{K}$, we define an estimator of f by

$$\widehat{f}_{\mathbb{K}^*}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{K}^*(X_i, x), \quad x \in I. \tag{2.1}$$

This estimator is considered in Lerasle *et al.* (2016) and for specific choices of $\mathbb{K}^*$, it covers for instance projection estimators, kernel estimators or weighted projection estimators. The first two examples are detailed in Section 2.2. Remark that $\mathbb{E}(\widehat{f}_{\mathbb{K}^*}) = f_{\mathbb{K}^*}$ is expected to be a good approximation of f.

Following an analogous way, we propose as an estimator of fg, for $\mathbb{K} \in \mathcal{K}$,

$$\widehat{(fg)}_{\mathbb{K},\mathbb{K}^*}(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{\mathbb{K}^*}(Y_i)\mathbb{K}(Y_i, x), \quad x \in I. \tag{2.2}$$

Indeed, if $\widehat{f}_{\mathbb{K}^*}$ is near of f, and as $\mathbb{E}(f(Y_1)\mathbb{K}(Y_1,x)) = (fg)_{\mathbb{K}}$, we get a relevant estimator of fg if, again, for $\mathbb{K} \in \mathcal{K}, \psi_{\mathbb{K}}$ is a good approximation of fg. Note that in definition (2.2), X_1 and Y_1 must have the same dimension.

We set the following assumptions on $\mathbb{K}$:

($\mathcal{A}1$) $\exists C_1 > 0, \forall \psi \in \mathbb{L}^2(I),\ \|\psi_{\mathbb{K}}\|^2 = \int \left(\int \psi(y)\mathbb{K}(y,x)dy\right)^2 dx \le C_1\|\psi\|^2$
($\mathcal{A}2$) $L(\mathbb{K}) := \sup_{y\in I}\int \mathbb{K}^2(y,x)dx < \infty$.

Moreover, we require the following assumptions on $\widehat{f}_{\mathbb{K}^*}$:

($\mathcal{A}3$) $\exists C_3 > 0, \int \left(\mathbb{E}[\widehat{f}_{\mathbb{K}^*}(y)]\right)^2 dy = \|f_{\mathbb{K}^*}\|^2 \le C_3$.

($\mathcal{A}4$) $\exists C_4 > 0, \int \mathrm{Var}\left(\widehat{f}_{\mathbb{K}^*}(y)\right) dy \le C_4$.

($\mathcal{A}5$) $\exists C_5 > 0,\ \forall \psi \in \mathbb{L}^2(I),\ \mathbb{E}\left[\left\langle \psi, \widehat{f}_{\mathbb{K}^*} - \mathbb{E}[\widehat{f}_{\mathbb{K}^*}]\right\rangle^2\right] \le C_5\frac{\|\psi\|^2}{n}$.

These assumptions are related to $\mathbb{K}^*$ and therefore to the density f only. Now, we can establish the following upper bound on the mean integrated risk of $\widehat{(fg)}_{\mathbb{K},\mathbb{K}^*}$.

Proposition 2.1. *Let $\mathbb{K}, \mathbb{K}^*$ belong to $\mathcal{K}$. Assume that assumptions ($\mathcal{A}1$)-($\mathcal{A}5$) hold and that g is bounded on I with bound denoted by $\|g\|_\infty$. Let $\widehat{(fg)}_{\mathbb{K},\mathbb{K}^*}$ be the estimator defined by (2.2). Then, we have*

$$\mathbb{E}[\|\widehat{(fg)}_{\mathbb{K},\mathbb{K}^*} - fg\|^2] \le 2\|(fg)_{\mathbb{K}} - fg\|^2 + 2C_1\|g\|_\infty^2\|f - f_{\mathbb{K}^*}\|^2 + \mathfrak{C}(f,g)\frac{L(\mathbb{K})}{n}, \tag{2.3}$$

where $\mathfrak{C}(f,g) := \|g\|_\infty(C_3 + C_4 + C_5)$.

Strategy suggested by (2.3). The risk bound (2.3) contains two standard terms, the squared bias $\|(fg)_{\mathbb{K}} - fg\|^2$ and the variance $\mathfrak{C}(f,g)L(\mathbb{K})/n$, requiring

a standard selection for $\mathbb{K}$ (see Lerasle *et al.* (2016)). It also involves the bias term $\|f - f_{\mathbb{K}^*}\|^2$ which has no counterpart: thus $\mathbb{K}^*$ can and should be chosen in order to make it negligible and $\widehat{f}_{\mathbb{K}^*}$ can thus be taken overfitted.

If in the initial problem, f and g have symmetric roles, this is no longer true in the definition (2.2) of the estimator, where one of the two densities is estimated first. As a matter of fact, Proposition 2.1 suggests to plug in the product estimator (2.2) an over-fitted estimator, eliminating a selection issue for $\mathbb{K}^*$. When possible we select for this over-fitted estimator the one corresponding to the smoother density. Indeed, this should make the additional bias term decrease faster. However, the information about which is smoother between f and g, is not available. From theoretical viewpoint, both $\|f - f_{\mathbb{K}^*}\|^2$ and $\|g - g_{\mathbb{K}^*}\|^2$ are negligible by assuming a minimal regularity for f and g and choosing $\mathbb{K}^*$ such that Assumptions ($\mathcal{A}1$)-($\mathcal{A}5$) hold, makes the bias term the smallest as possible. From a practical point of view and in dimension 1, we propose to consider that the smoother density is the one for which a selection method for the direct density estimation of f and g leads to the smallest complexity $L(\mathbb{K})$ (i.e. the smallest selected dimension in projection or the largest bandwidth for kernels, see section 2.2 hereafter).

In the sequel, we may write $\widehat{(fg)}_{\mathbb{K}}$ instead of $\widehat{(fg)}_{\mathbb{K},\mathbb{K}^*}$ when there is no ambiguity, as $\mathbb{K}^*$ can be fixed.

2.2. Assumptions ($\mathcal{A}1$)-($\mathcal{A}5$) for projection and kernel estimators

We consider the following two examples:

[P] Projection function: for a multi-index $\mathbf{m} = (m_1, \dots, m_d) \in \mathbb{N}^d$, let

$$\mathbb{K}_{\mathbf{m}}(x, y) = \sum_{\mathbf{0} \le \mathbf{j} \le \mathbf{m-1}} \varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(y)$$

where $(\varphi_{\mathbf{j}})_{\mathbf{0} \le \mathbf{j} \le \mathbf{m-1}}$ is an $\mathbb{L}^2(I)$-orthonormal basis, $I = I_1 \times \ldots \times I_d$, with

$$\varphi_{\mathbf{j}}(x) = (\varphi_{j_1} \otimes \ldots \otimes \varphi_{j_d})(x) = \varphi_{j_1}(x_1) \times \ldots \times \varphi_{j_d}(x_d)$$

and

$$\mathbf{L}(\mathbf{m}) := \left\| \sum_{\mathbf{0} \le \mathbf{j} \le \mathbf{m-1}} \varphi_{\mathbf{j}}^2 \right\|_\infty = \prod_{j=1}^{d} \left(\sup_{x_j \in I_j} \sum_{k_j=0}^{m_j-1} \varphi_{k_j}(x_j)^2 \right) =: \prod_{j=1}^{d} L(m_j) < +\infty.$$

We denote by $S_{\mathbf{m}}$ the $D_{\mathbf{m}}$-dimensional linear subspace of $\mathbb{L}^2(I)$ spanned by $\varphi_{\mathbf{0}}, \dots, \varphi_{\mathbf{m-1}}$, where $D_{\mathbf{m}} = \prod_{j=1}^{d} m_j$.

[Ker] Kernel function:

$$\mathbb{K}_{\mathbf{h}}(x, u) = \mathbf{K}\left(\frac{x - u}{\mathbf{h}}\right) = \prod_{j=1}^{d} \frac{1}{h_j} K\left(\frac{x_j - u_j}{h_j}\right)$$

for K an integrable and square-integrable symmetric function ($K(-z) = K(z)$) defined on $\mathbb{R}$, such that $\int_I \mathbf{K} = 1$ and $\mathbf{h} \in [0,1]^d$. We denote by $\mathbf{K_h}(x) := \prod_{j=1}^d h_j^{-1} K(x_j/h_j)$ and set $\mathbf{L(h)} = \prod_{j=1}^d L(h_j)$ where $L(h) = \|K\|^2/h$.

First we state that the above functions fulfill assumptions (A1)-(A2):

Proposition 2.2. *The functions $\mathbb{K}_\mathbf{m}$ defined in* [P] *and $\mathbb{K}_\mathbf{h}$ defined in* [Ker] *satisfy assumptions (A1)-(A2) and belong to $\mathcal{K}$.*

The order of $L(\mathbb{K}_\mathbf{m}) := \mathbf{L(m)}$ depends on the choice of the basis. In dimension 1, for the trigonometric basis for m odd and $I = [0,1]$, it holds $L(m) = m$. For the Hermite basis where $I = \mathbb{R}$, we have $L(m) \leq C_H\sqrt{m}$ (see Lemma 1 in Comte and Lacour (2021) and section 2.4). For the Legendre polynomial basis where $I = [-1,1]$ it holds that $L(m) = m^2$, see Cohen *et al.* (2013, p.831). In any case, we consider that $L(m) \geq 1$, which holds at least for $m \geq m_0$. Extension to dimension d is straightforward by tensorization of the bases, and $\mathbf{L(m)} = \prod_{i=1}^d L(m_j)$. In the kernel case, we simply have $L(\mathbb{K}_\mathbf{h}) := \mathbf{L(h)}$.

Next we prove that Assumptions (A3)-(A5) are verified for the projection estimator of f

$$\widehat{f}_{\mathbb{K}_{\mathbf{m}^*}} := \widehat{f}_{\mathbf{m}^*} = \sum_{\mathbf{0}\leq\mathbf{j}\leq\mathbf{m}^*-\mathbf{1}} \widehat{a}_\mathbf{j}\varphi_\mathbf{j}, \quad \widehat{a}_\mathbf{j} = \frac{1}{n}\sum_{i=1}^n \varphi_\mathbf{j}(X_i), \tag{2.4}$$

where $\sum_{\mathbf{0}\leq\mathbf{j}\leq\mathbf{m}^*-\mathbf{1}}$ stands for $\sum_{j_1=0}^{m_1^*-1}\dots\sum_{j_d=0}^{m_d^*-1}$, and for the kernel estimator

$$\widehat{f}_{\mathbb{K}_{\mathbf{h}^*}} := \widehat{f}_{\mathbf{h}^*}(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{K}_{\mathbf{h}^*}(X_i - x). \tag{2.5}$$

Proposition 2.3. *Assume that f is bounded. Then the estimators of f:*

- $\widehat{f}_{\mathbf{m}^*}$ *in case [P] under* $\mathbf{L}(\mathbf{m}^*) \leq n$,
- $\widehat{f}_{\mathbf{h}^*}$ *in case [Ker] under* $\mathbf{L}(\mathbf{h}^*) \leq n$,

satisfy assumptions (A3)-(A5).

The conditions $\mathbf{L}(\mathbf{m}^*) \leq n$ and $\mathbf{L}(\mathbf{h}^*) \leq n$ represent a stronger version of Assumption (A2). Moreover $f_{\mathbb{K}_\mathbf{m}} = f_\mathbf{m} = \sum_{\mathbf{0}\leq\mathbf{j}\leq\mathbf{m}-\mathbf{1}} a_\mathbf{j}\varphi_\mathbf{j}$, with $a_\mathbf{j} = \langle f, \varphi_\mathbf{j}\rangle$ is the projection of f on $S_\mathbf{m}$: it is a good approximation of f for $\mathbf{m}$ large (in the sense that $\min_{1\leq j\leq d} m_j$ is large). Analogously, $f_{\mathbb{K}_\mathbf{h}} = f \star \mathbf{K_h}$ with $u \star v$ denoting the convolution product of two $\mathbb{L}^2(I)$ functions, gets close to f for small $\mathbf{h}$ (in the sense that $\max_{1\leq j\leq d} h_j$ is small) under mild conditions, see e.g. Tsybakov (2009) or Goldenshluger and Lepski (2014).

This allows to derive from Proposition 2.1 in the following section upper-bound results for projection and kernel estimators that we introduce below.

2.3. A Corollary in the projection or kernel case

In a purely projection approach where we take $\mathbb{K}$ and $\mathbb{K}^*$ in the same collection of the form [P], we obtain the following estimator of fg:

$$\widehat{(fg)}_{\mathbb{K}_m,\mathbb{K}_{m^*}} := \widehat{(fg)}_{\mathbf{m},\mathbf{m}^*} = \sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}-\mathbf{1}} \widehat{a}_{\mathbf{j}}^{(\mathbf{m}^*)}\varphi_{\mathbf{j}}, \quad \widehat{a}_{\mathbf{j}}^{(\mathbf{m}^*)} = \frac{1}{n}\sum_{i=1}^{n}\varphi_{\mathbf{j}}(Y_i)\widehat{f}_{\mathbf{m}^*}(Y_i). \tag{2.6}$$

Clearly, $\mathbb{E}(\widehat{a}_{\mathbf{j}}^{(\mathbf{m}^*)}) = \langle\varphi_{\mathbf{j}}, f_{\mathbf{m}^*}g\rangle$, which shows that our estimator is indeed close to $f_{\mathbf{m}^*}g$, which in turn should be near of fg for large $\mathbf{m}^*$. Choosing $\mathbf{m}^*$ large is possible since the variance of $\widehat{f}_{\mathbf{m}^*}$ does not appear in the risk bound. It is worth noting that the sum in the right hand side of (2.6) is composed of identically distributed but dependent variables: indeed, dependency appears through $\widehat{f}_{\mathbf{m}^*}$, which is based on the X-sample.

In a purely kernel approach where we take $\mathbb{K}$ and $\mathbb{K}^*$ in the same collection of the form [Ker], we consider the estimator of fg:

$$\widehat{(fg)}_{\mathbf{h},\mathbf{h}^*}(x) = \frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{\mathbf{h}^*}(Y_i)\mathbf{K}_{\mathbf{h}}(Y_i - x). \tag{2.7}$$

A straightforward consequence of Proposition 2.1 is the corollary:

Corollary 2.1. *Assume that f and g are bounded.*
Let $\widehat{(fg)}_{\mathbf{m},\mathbf{m}^*}$ be defined by (2.6). For any $\mathbf{m}^*$ such that $\mathbf{L}(\mathbf{m}^*) \le n$, we have

$$\mathbb{E}\left(\|\widehat{(fg)}_{\mathbf{m},\mathbf{m}^*} - fg\|^2\right) \le 2\|(fg)_{\mathbf{m}} - fg\|^2 + 2\|g\|_\infty^2\|f - f_{\mathbf{m}^*}\|^2 + \mathfrak{C}_1\frac{\mathbf{L}(\mathbf{m})}{n}, \tag{2.8}$$

where $fg = fg\mathbf{1}_I$, $\mathfrak{C}_1 := \|g\|_\infty(1 + 2\|f\|_\infty)$ and $(fg)_m$ is the orthogonal projection of fg on $S_{\mathbf{m}}$, $f_{\mathbf{m}^*}$ the orthogonal projection of f on $S_{\mathbf{m}^*}$.
Let $\widehat{(fg)}_{\mathbf{h},\mathbf{h}^*}$ be defined by (2.7). For any $\mathbf{h}^*$ such that $\mathbf{L}(\mathbf{h}^*) \le n$, we have

$$\mathbb{E}\left(\|\widetilde{(fg)}_{\mathbf{h},\mathbf{h}^*} - fg\|^2\right) \le 2\|(fg)_{\mathbf{h}} - fg\|^2 + 2\|g\|_\infty^2\|\mathbf{K}\|_1^2\|f_{\mathbf{h}^*} - f\|^2 + \frac{\mathfrak{C}_2\mathbf{L}(\mathbf{h})}{n}, \tag{2.9}$$

where $\mathfrak{C}_2 = \|g\|_\infty(1 + 2\|\mathbf{K}\|_1^2\|f\|_\infty)$.

It is also possible to consider a mixed strategy where $\mathbb{K}^*$ is selected in collection [P] and $\mathbb{K}$ in [Ker], or vice-versa, and a similar result is obtained. This strategy is investigated in the numerical section. Then, the two bias terms $\|(fg)_{\mathbb{K}} - fg\|^2$ and $\|f - f_{\mathbb{K}^*}\|^2$ refer to different regularity spaces, see the discussion below and the next section.

The bound (2.9) suggests to choose $\mathbf{h}^*$ the smallest as possible, in order to make this term negligible. For instance if f belongs to a Nikols'ki ball with regularity parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ (see Goldenshluger and Lepski (2014), Definition 1), $\|f_{\mathbf{h}^*} - f\|^2$ has order $\sum_{j=1}^{d}(h_j^*)^{2\alpha_j}$ if the kernel $\mathbf{K}$ has order at least $\max_j\lfloor\alpha_j\rfloor$ (where $\lfloor\alpha\rfloor$ is the largest integer such that $\lfloor\alpha\rfloor < \alpha$ and the order of

the kernel is understood as in section 3.2 of Goldenshluger and Lepski (2014)). It follows that if $\min_j \alpha_j > 1/2$ and $h_j^* = 1/n$, $\forall j$, this term has order less than $1/n$ and is negligible. Then, only the bandwidth $\mathbf{h}$ requires to be selected.

If in addition fg belongs to a Nikols'ki ball with regularity parameter $\boldsymbol{\beta}$ and $\mathbf{K}$ has order at least $\max_j \lfloor \beta_j \rfloor$, then the estimator reaches the minimax rate

$$n^{-2\overline{\boldsymbol{\beta}}/(2\overline{\boldsymbol{\beta}}+d)} \quad \text{for} \quad \frac{1}{\overline{\boldsymbol{\beta}}} = \frac{1}{d}\sum_{j=1}^{d}\frac{1}{\beta_j},$$

(see Goldenshluger and Lepski (2014)), for h_j chosen of order $n^{-\overline{\boldsymbol{\beta}}/\left(\beta_j(2\overline{\boldsymbol{\beta}}+1)\right)}$. Such a choice of $\mathbf{h}$ is not feasible since $\boldsymbol{\beta}$ is unknown, a data driven procedure for selecting $\mathbf{h}$ must be proposed.

In the following section we provide a more precise evaluation of the rates of estimation in the projection case in dimension 1.

2.4. Projection strategy: Rates on Sobolev Hermite spaces

In this section we focus on the case $d = 1$ and give an example of rate induced by the bound (2.8), in the case of the Hermite basis and associated Sobolev spaces (see Comte and Lacour (2021) Definition 1 for the general case $d \geq 1$). The Hermite functions $(\varphi_j)_{j\geq 0}$ are defined from Hermite polynomials $(H_j)_{j\geq 0}$ by: for $x \in \mathbb{R}$

$$\varphi_j(x) = c_j H_j(x) e^{-x^2/2}, \quad H_j(x) = (-1)^j e^{x^2}\frac{d^j}{dx^j}(e^{-x^2}), \quad c_j = (2^j j!\sqrt{\pi})^{-1/2}. \tag{2.10}$$

The Hermite polynomials $(H_j)_{j\geq 0}$ are orthogonal with respect to the weight function e^{-x^2}, that is: $\int_{\mathbb{R}} H_j(x)H_k(x)e^{-x^2}dx = 2^j j!\sqrt{\pi}\delta_{j,k}$ (see Abramowitz and Stegun (1964), chap 22.2.14). Therefore, the Hermite basis $(\varphi_j)_{j\geq 0}$ is an orthonormal basis on $\mathbb{R}$. We note also that φ_j is bounded:

$$||\varphi_j||_\infty = \sup_{x\in\mathbb{R}}|\varphi_j(x)| \leq \Phi_0, \text{ with } \Phi_0 \simeq 1,086435/\pi^{1/4} \simeq 0,8160 \tag{2.11}$$

(see Abramowitz and Stegun (1964), chap.22.14.17). Moreover, as recalled above $\sup_{x\in\mathbb{R}}\sum_{j=0}^{m-1}\varphi_j^2(x) \leq C_H\sqrt{m}$ for a finite constant $C_H > 0$ (see Lemma 1 of Comte and Lacour (2021)).

For $s > 0$, the Sobolev-Hermite ball (see Bongioanni and Torrea (2006)) is defined by :

$$W_H^s(D) = \Big\{\theta \in \mathbb{L}^2(\mathbb{R}), \sum_{k\geq 0} k^s a_k^2(\theta) \leq D\Big\}, \quad D > 0, \tag{2.12}$$

where $a_k(\theta) = \langle \theta, \varphi_k\rangle$. It is proved in Belomestny *et al.* (2019) that, for s an integer, $s \geq 1$, $f \in W_H^s = \{\theta \in \mathbb{L}^2(\mathbb{R}), \sum_{k\geq 0} k^s a_k^2(\theta) < +\infty\}$ is equivalent to: f admits derivatives up to order s which satisfy: f, f', ..., $f^{(s)}$, $x^{s-\ell}f^{(\ell)}$ for

$\ell = 0, \dots, s-1$ belong to $\mathbb{L}^2(\mathbb{R})$. Moreover, for any function $f \in W_H^s(D)$, we have $\|f - f_m\|^2 \leq Dm^{-s}$. It is also easy to see that if, in addition, $s > 1$, then f is bounded. Indeed

$$|\sum_{j\geq 0} a_j\varphi_j| \leq \Phi_0\left(|a_0| + \sum_{j\geq 1}(|a_j|j^{s/2})j^{-s/2}\right) \leq \Phi_0\left(\|f\| + \sqrt{\sum_{j\geq 1} j^s a_j^2 \sum_{j\geq 1} j^{-s}}\right).$$

As f and g are assumed to be bounded, it holds

$$|\langle fg, \varphi_j\rangle| \leq \min(\|f\|_\infty |\langle g, \varphi_j\rangle|, \|g\|_\infty |\langle f, \varphi_j\rangle|).$$

Thus if $fg \in W_H^s(D)$, $f \in W_H^{s'}(D')$ and $g \in W_H^{s''}(D'')$, then $s \geq \max(s', s'')$.

Consequently, we obtain as a straightforward consequence of bound (2.8), the following result.

Proposition 2.4. *Let $s \geq s' \geq 1/2$ and assume that $fg \in W_H^s(D)$, $f \in W_H^{s'}(D')$ with f and g bounded and $g \in \mathbb{L}^2(\mathbb{R})$. Then choosing $m_{\text{opt}} = [n^{1/(s+1/2)}]$ and $m_n^* = n^2/C_H^2$, we have*

$$\mathbb{E}\left(\|\widehat{(fg)}_{m_{\text{opt}}, m_n^*} - fg\|^2\right) \leq C(D, D', \|f\|_\infty, \|g\|_\infty) n^{-\frac{2s}{2s+1}}.$$

We can conclude that the resulting rate is of order $n^{-2s/(2s+1)}$, and is optimal, see (1.1).

3. Adaptive procedure

3.1. General adaption result

We propose a Goldenschluger and Lepski (2011) method. Define $\mathcal{K}_n = \{\mathbb{K}_\tau\}_{\tau\in\mathcal{T}_n}$ a family of symmetric functions indexed by a parameter τ, satisfying ($\mathcal{A}1$)-($\mathcal{A}2$). For simplicity, we write ψ_τ instead of $\psi_{\mathbb{K}_\tau}$ and $L(\tau)$ instead of $L(\mathbb{K}_\tau)$. For example in the previous examples the parameter τ is a d dimensional vector of integers $\mathbf{m}$ in the projection case and a bandwidth $\mathbf{h} \in [0,1]^d$ in the kernel context. We add the following assumptions

($\mathcal{A}6$) $\forall x, y \in I, \forall \tau, \tau' \in \mathcal{T}_n$, $\int \mathbb{K}_\tau(x,z)\mathbb{K}_{\tau'}(z,y)dz = \int \mathbb{K}_\tau(y,z)\mathbb{K}_{\tau'}(z,x)dz$.

($\mathcal{A}7$) f and g are bounded and $\forall \tau \in \mathcal{T}_n$, $\|f_\tau\|_\infty \leq C_7$.

($\mathcal{A}8$) The collection of models is such that $\text{Card}(\mathcal{T}_n) \leq n^d$, and $\forall c > 0$,

$$\sum_{\tau\in\mathcal{T}_n} e^{-c\sqrt{L(\tau)}} \leq \Sigma < +\infty$$

where $\Sigma = \Sigma(c)$ is a constant depending on c but not on n.

($\mathcal{A}9$) $\forall \tau \in \mathcal{T}_n$, $\sup_{x,y} |\mathbb{K}_\tau(x,y)|^2 \leq C_9 L^2(\tau)$.

($\mathcal{A}10$) The parameter τ^* is such that $L(\tau^*) \leq \dfrac{\|f\|_\infty}{2(d+3)(1+\sqrt{C_9})} \dfrac{n}{\log(n)}$.

Define

$$\widehat{(fg)}_{\tau,\tau'}(x) = \int \mathbb{K}_{\tau'}(y,x)\widehat{(fg)}_{\tau}(y)dy.$$

Under (A6) note that $\widehat{(fg)}_{\tau,\tau'} = \widehat{(fg)}_{\tau',\tau}$. Now, set

$$\begin{cases} A(\tau) = \sup_{\tau' \in \mathcal{T}_n} \left[\|\widehat{(fg)}_{\tau,\tau'} - \widehat{(fg)}_{\tau'}\|^2 - \kappa V(\tau') \right]_+, \\ V(\tau) = \kappa(C_1 \vee C_1^2)(\|f\|_\infty^2 + \|g\|_\infty^2)\frac{L(\tau)}{n}. \end{cases} \tag{3.1}$$

The selection of τ is done by the rule

$$\widehat{\tau} = \arg\min_{\tau \in \mathcal{K}_n} \{A(\tau) + \kappa' V(\tau)\}$$

for some positive constants κ and κ' to be selected. The estimator $\widehat{f}_{\mathbb{K}_{\tau^*}}$ of f defined in (2.1) relies now on the symmetric function $\mathbb{K}_{\tau^*} \in \mathcal{K}_n$ and is rewritten $\widehat{f}_{\tau^*}$ with $\mathbb{K}^* = \mathbb{K}_{\tau^*}$ and we note $f_{\mathbb{K}_{\tau^*}}$ by f_{τ^*}.

Theorem 3.1. *Assume that Assumptions (A1)-(A10) are fulfilled. Then we have, for $\kappa' \geq \kappa$,*

$$\begin{aligned} \mathbb{E}(\|\widehat{(fg)}_{\tau,\widehat{\tau}} - fg\|^2) \quad \leq \quad & C \inf_{\tau \in \mathcal{T}_n} \left\{ (C_1 \vee 1)\|(fg)_\tau - fg\|^2 + \kappa' V(\tau) \right\} \\ & + 42(C_1 \vee C_1^2)\|g\|_\infty^2 \|f - f_{\tau^*}\|^2 + \frac{C'}{n}, \end{aligned}$$

where C is a numerical and C' depends on $\|f\|_\infty$, $\|g\|_\infty$ and on a constant specific to the collection $\mathcal{K}$.

Theorem 3.1 shows that the adaptive procedure automatically realizes the squared bias-variance tradeoff up to negligible terms, as soon as τ^* is such that $\|f - f_{\tau^*}\|^2$ has order less than $O(1/n)$. This result is general and allows to consider any collection $\mathcal{K}$ such that (A1)-(A10) are satisfied. For example, a mixed strategy where a kernel estimator of f is plugged in a projection estimator of the product is possible. It is worth stressing that the proof of Theorem 3.1 relies on the Talagrand inequality and does not involve the study of a U-statistics (see Lerasle *et al.* (2016) and our Theorem 4.2 below).

3.2. Assumptions (A6)-(A10) for projection and kernel estimators

In this section, we present assumptions ensuring (A1)-(A10) for the previous procedures. First, in the projection case (2.6), we define the collection of proposals for $\mathbf{m}$ as follows

$$\mathcal{M}_n = \{\mathbf{m} \in \{1, \ldots, n\}^d, \quad \mathbf{L}(\mathbf{m}) \leq n\},$$

and set the following set of assumptions:

[P1] f and g are bounded on I.

[P2] The basis functions are bounded: $\forall \mathbf{j} \in \mathbb{N}^d, \forall x \in I$, $|\varphi_{\mathbf{j}}(x)| \leq C_\varphi$.

[P3] The model $\mathbf{m}^*$ is such that $\mathbf{L}(\mathbf{m}^*) \leq \dfrac{\|f\|_\infty}{4(d+3)} \dfrac{n}{\log(n)}$.

[P4] The collection of models is such that $\mathrm{Card}(\mathcal{M}_n) \leq n^d$, and $\forall c > 0$, $\sum_{\mathbf{m} \in \mathcal{M}_n} e^{-c\sqrt{\mathbf{L}(\mathbf{m})}} \leq \Sigma < +\infty$ where $\Sigma = \Sigma(c)$ is a constant depending on c but not on n.

[P5] There exist $\mathbf{b} \in (1, \infty)^d$ and $C_{\mathbf{b}} > 0$, which need not to be known, such that $\sum_{\mathbf{j} \geq 1} \mathbf{j}^{\mathbf{b}} a_{\mathbf{j}}^2(f) \leq C_{\mathbf{b}} < +\infty$, where $\mathbf{j}^{\mathbf{b}} = j_1^{b_1} \dots j_d^{b_d}$.

Corollary 3.1. *In the projection case [P], assumptions* [P1]-[P5] *imply (A1)-(A10).*

In Assumption [P3], the maximal value of $\mathbf{m}^*$ depends on $\|f\|_\infty$. This constraint can be replaced by $\mathbf{L}(\mathbf{m}^*) \leq n/\log^{3/2}(n)$ and the result follows for n large enough. Assumptions [P2] and [P4] are classical, for instance they are fulfilled by trigonometric and Hermite bases. Condition [P5] is a minimal regularity constraint. For instance for $d = 1$, it requires that the function f has a minimal regularity of $1/2$ on Sobolev-Fourier spaces for $I = [0, 1]$ and 1 on Hermite Sobolev spaces.

Second, for the kernel estimator (2.7), denote by $\mathcal{T}_n$ a discrete collection $\mathcal{H}_n$ of bandwidths in $(1/n, 1)$ with cardinality less than n. We consider the following set of assumptions:

[K1] f and g are bounded on I.

[K2] The kernel K is even, bounded and integrable.

[K3] The bandwidth $\mathbf{h}^*$ is such that $\mathbf{L}(\mathbf{h}^*) \leq \frac{\|f\|_\infty}{2(d+3)(1+\frac{\|\mathbf{K}\|_\infty}{\|\mathbf{K}\|^2})} \frac{n}{\log(n)}$.

[K4] The discrete collection $\mathcal{H}_n$ of bandwidths in $(1/n, 1)^d$ has cardinality less than n^d and for any $c_1 > 0$, $\sum_{\mathbf{h} \in \mathcal{H}_n} \exp(-c_1\sqrt{\mathbf{L}(\mathbf{h})}) \leq \Sigma = \Sigma(c_1) < +\infty$.

Note that as $\int \mathbf{K} = 1$, $\|\mathbf{K}\|_1 \geq 1$. Similarly to [P3], We can replace in [K3] the bound by $n \log(n)^{-3/2} \geq \mathbf{L}(\mathbf{h}^*)$, for large enough n, to get rid of the unknown constant $\|f\|_\infty$ in the bound defining $\mathbf{h}^*$. Assumption [K4] is fulfilled for

$$\mathcal{H}_n = \left\{ \mathbf{h} \in \left\{ \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1 \right\}^d \right\}.$$

Contrary to the projection, the kernel method does not require any regularity constraint of type [P5].

Corollary 3.2. *In the kernel case [K], assumptions* [K1]-[K4] *imply (A1)-(A10).*

4. Numerical study

4.1. Numerically efficient adaptive procedures in dimension 1

For the numerical study we focus on dimension $d = 1$ and do not implement the above adaptive procedure. Indeed, the Goldenschluger and Lepski method is often difficult to calibrate from an implementation viewpoint (see Comte and Rebafka (2012)) and suffers from important computational costs. Indeed, it involves the calibration of two constants, κ' and κ. This preliminary calibration step is difficult, probably because these constants act simultaneously on the bias and variance terms. In the kernel case, the "double" convolution which appears when computing $\widehat{fg}_{h,h'}$ is numerically time consuming. Instead we propose two numerically efficient procedures model selection and PCO, for which the squared bias-variance tradeoff is also attained under [P1]-[P5] and [K1]-[K4] respectively.

4.1.1. Model selection for projection estimators

We present, under the above assumptions [P1]-[P5], a result for a more standard and simpler model selection procedure. More precisely, define

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{m^*}(Y_i)t(Y_i). \tag{4.1}$$

Then select $\widetilde{m}$ with the criterion

$$\widetilde{m} := \arg\min_{m\in\mathcal{M}_n}\left\{\min_{t\in S_m}\gamma_n(t) + \mathrm{pen}(m)\right\}, \quad \mathrm{pen}(m) = \kappa(\|f\|_\infty^2 + \|g\|_\infty^2)\frac{L(m)}{n}$$

where κ is a numerical constant. Note that

$$\min_{t\in S_m}\gamma_n(t) = \gamma_n(\widehat{(fg)}_{m,m^*}) = -\|\widehat{(fg)}_{m,m^*}\|^2.$$

With the same tools as those used in the proof of Theorem 3.1, we can prove the following result.

Theorem 4.1. *In the projection case [P], if Assumptions* [P1]-[P5] *hold, then, there exists κ_0 such that, for any $\kappa \geq \kappa_0$, we have*

$$\mathbb{E}(\|\widehat{(fg)}_{\widetilde{m},m^*} - fg\|^2) \leq \inf_{m\in\mathcal{M}_n}\left\{3\|fg - (fg)_m\|^2 + 4\kappa(\|f\|_\infty^2 + \|g\|_\infty^2)\frac{L(m)}{n}\right\}$$
$$+16\|g\|_\infty^2\|f - f_{m^*}\|^2 + \frac{C}{n}, \tag{4.2}$$

where C is a constant depending on $\|f\|_\infty$, C_a.

The proof is omitted but details can be found in the preprint version Comte and Duval (2022), version 1, which also indicates that $\kappa_0 = 8 \times 12 = 96$ would

suit. In practice, this theoretical value is always too large, and has to be calibrated on preliminary simulation experiments. Note that, the estimate $\widehat{(fg)}_{\widetilde{m},m^*}$ is replaced by its positive part, for which the same risk bound holds. Moreover, it can be easily checked that this proof also hold in a mixed strategy where a kernel estimator for f is plugged in a projection estimator of the product, i.e. $\mathbb{K}^*$ in [Ker] and $\mathbb{K}$ in [P]. This justifies why for numerical results we experiment this mixed strategy with a penalised criterion for adaptation.

The values $\|f\|_\infty$, $\|g\|_\infty$ in the penalty term are unknown and must be replaced by estimates. The bound $\|f\|_\infty$ can be estimated by the maximal value of a projection estimate of f on a middle-sized space, for instance $\sup_{x\in I}|\widehat{f}_{[\sqrt{n}]}(x)|$ and an analogous approach can be adopted for $\|g\|_\infty$. Let us denote these estimators by $\widehat{\|f\|}_\infty$ and $\widehat{\|g\|}_\infty$. This strategy is theoretically studied in Theorem 12 p.594 (Appendix A: Random penalty) in Lacour (2007).
We consider in the numerical Section the estimator $\widehat{f}_{\widetilde{m},m^*}$ and adopt the following strategy. The penalty is obtained from the theory as the sum of the bounds on two terms, a bound on $\frac{1}{n}\sum_{j=1}^{m-1}\mathbb{E}\left(\varphi_j^2(Y_1)[\widehat{f}_{m^*}(Y_1)]^2\right)$ and a bound on an additional term given by $\|f\|_\infty\|g\|_\infty L(m)/n$. Following ideas in Massart (2007) (see also Theorem 7.6 p.216, in the density case), we replace the first term by

$$\widehat{\text{pen}}_1(m)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=0}^{m-1}\left[\widehat{f}_{m^*}(Y_i)\varphi_j(Y_i)\right]^2$$

and the second term by $\widehat{\text{pen}}_2(m)=\widehat{\|f\|}_\infty\widehat{\|g\|}_\infty L(m)/n$. So, in the Hermite basis where $L(m)=C_H\sqrt{m}$ (with unknown C_H), our global penalty is

$$\widehat{\text{pen}}_1(m)+\kappa\widehat{\|f\|}_\infty\widehat{\|g\|}_\infty\frac{\sqrt{m}}{n}. \tag{4.3}$$

The constant κ is calibrated by preliminary simulations, see section 4.

4.1.2. Bandwidth selection with PCO

We describe here a method called "PCO" ("Penalized Comparison to Overfitting") introduced for kernel density estimation by Lacour *et al.* (2017). The PCO method is more complicated from theoretical point of view, because it involves the study of several U-statistics of order 2. But, it is much easier to calibrate and implement in practice.

Let us describe the PCO method. Denote by $\widehat{(fg)}_h$ (omitting h^* for simplicity), we select

$$\widetilde{h}=\arg\min_{h\in\mathcal{H}_n}\left\{\|\widehat{(fg)}_h-\widehat{(fg)}_{h_{\min}}\|^2+2\text{pen}(h)\right\}$$

with $h_{\min}=\min\{h,h\in\mathcal{H}_n\}$,

$$\text{pen}(h)=\text{pen}_1(h)+\text{pen}_2(h)\text{ where }\text{pen}_1(h)=\frac{1}{n^2}\langle K_h,K_{h_{\min}}\rangle\sum_{i=1}^{n}\widehat{f}_{h^*}^2(Y_i), \tag{4.4}$$

$$\mathrm{pen}_2(h) = \kappa \frac{c_0(f,g,K)}{nh}, \quad c_0(f,g,K) = 4\|K\|_1^3 \|K\|_\infty (\|g\|_\infty^2 + \|f\|_\infty^2). \quad (4.5)$$

Note that pen_1 and pen_2 are both of order $1/(nh)$. This is obvious for pen_2; for pen_1, observe that $|\langle K_h, K_{h_{\min}}\rangle| \leq \|K\|_\infty \|K\|_1 / h$ and that, under [K3], $(1/n)\sum_{i=1}^n \widehat{f}_{h^*}^2(Y_i)$ is bounded with large probability, see (7.6).

Theorem 4.2. *Consider the kernel case [K]. Assume that Assumptions* [K1]-[K4] *hold and that* $1/(nh_{\min}) \leq 1$. *Then, for any* $\kappa \geq 1/4$, *we have*

$$\mathbb{E}\left(\|\widehat{(fg)}_{\widetilde{h}} - fg\|^2\right) \leq c_1 \inf_{h\in\mathcal{H}_n} \left\{ \|(fg)_h - fg\|^2 + (1 + \frac{\kappa}{1+\theta}) \frac{c_0(f,g,K)}{nh} \right\}$$
$$+ c_2 \|(fg)_{h_{\min}} - fg\|^2 + c_3 \|f_{h^*} - f\|^2 + C\frac{\log(n)}{n},$$

where c_3 *and* C *are positive constants depending on* f, g, K *and* $(c_1, c_2) = (3.6, 10.4)$ *would suit.*

The risk bound of Theorem 4.2 involves four terms. The first term in the first line is the minimal risk among the collection of estimators, up to multiplicative a constants. The two following terms, $\|(fg)_{h_{\min}} - fg\|^2$ and $\|f_{h^*} - f\|^2$ are bias terms corresponding to small bandwidths, they are negligible if $h_{\min}$ is of order $1/n$ and h^* of order $\log(n)/n$ (as required by assumption [K3]), and if the functions f and fg have regularity larger that $1/2$. The last term $\log(n)/n$ has negligible order compared to the first one. Therefore, the adaptive estimator achieves the intended squared-bias variance compromise given at the end of section 2.3.

4.2. Description of the implemented procedures

We illustrate the performances of the projection estimator with Hermite basis (see section 2.4) and kernel estimator with kernel built as a Gaussian mixture defined by:

$$K(x) = 2n_1(x) - n_2(x), \quad (4.6)$$

where $n_j(x)$ is the density of a centered Gaussian with a variance equal to j. This kernel is of order 3 (*i.e.* $\int x^j K(x)dx = 0$, for $j = 1, 2, 3$). We consider four examples:

[BU] $X \sim f = \mathcal{B}(7,5)$ and $Y \sim g = \mathcal{U}(0,1)$,
[GE] $X \sim f = \Gamma(4, 1/4)$ and $Y \sim g = \mathcal{E}(1/4)$,
[NL] $X \sim f = \mathcal{N}(0,3)$ and $Y \sim g$ Laplace,
[NC] $X \sim f = \mathcal{N}(0,3)$ and $Y \sim g$ Cauchy.

We compute normalized $\mathbb{L}^2$-risks to allow the numerical comparaison of the different examples for which $\int (fg)^2$ varies a lot. Namely, we evaluate

$$\frac{\mathbb{E}[\|\widehat{(fg)} - fg\|^2]}{\|fg\|^2}$$

and the associated deviations, from $N = 100$ independent datasets with different values of sample size $n = 200,\ 1000$ and 2000. All adaptive methods require the calibration of constants κ's in penalties. This is done by preliminary simulation experiments. For calibration strategies (dimension jump and slope heuristics), the reader is referred to Baudry *et al.* (2012), and to Lerasle (2012) for theoretical justifications. Here, we test a grid of values of κ's, the tests are conducted on a set of densities which are different from the one considered hereafter, to avoid overfitting. The different estimators are computed on the same datasets and compared.

• **Product** : This estimator is obtained as the product of $\widehat{f}\widehat{g}$ where each estimator is an adaptive optimal estimator. In the projection case, the product is $\widehat{f}_{\widehat{m}_1}\widehat{g}_{\widehat{m}_2}$, where $\widehat{f}_m$ is defined by (2.4) with

$$\widehat{m}_1 = \arg \min_{m\in\{1,\dots,D_{\max}\}} \left\{ -\|\widehat{f}_m\|^2 + \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j=0}^{m-1}\varphi_j^2(X_i) \right\},$$

and $\widehat{g}_{\widehat{m}_2}$ is defined analogously. In the kernel case, $\widetilde{f}_{\widetilde{h}_1}\widetilde{g}_{\widetilde{h}_2}$, where $\widetilde{f}_{\widetilde{h}_1}$ is defined by (2.5) with

$$\widetilde{h}_1 = \arg \min_{h\in\{1/k,k=1,\dots,n\}} \left\{ \|\widetilde{f}_h - \widetilde{f}_{\frac{1}{n}}\|^2 + \frac{4}{n}\langle K_h, K_{\frac{1}{n}}\rangle \right\},$$

and $\widetilde{g}_{\widetilde{h}_2}$ is defined analogously.

• **First X**: In all our examples f is smoother than g. The theoretical results suggest that one should consider for the preliminary estimate the dataset X which has density f. In the projection setting our estimate is $\widehat{(fg)}_{\widetilde{m},m^*}$ of Theorem 4.1 and penalty given by (4.3), where $\widehat{\|f\|^2_\infty}$ is estimated by $\sup_{x\in I}|\widehat{f}^2_{10}(x)|$, $\widehat{\|g\|^2_\infty}$ is estimated similarly, and with $\kappa = 0.15$ after calibration. In the kernel case we consider the estimator $\widetilde{(fg)}_{\widetilde{h}}$ of Theorem 4.2 with penalty given by (4.4) where pen_2 is replaced by

$$\widehat{\mathrm{pen}}_2 = 0.32\frac{\widehat{\|f\|^2_\infty} + \widehat{\|g\|^2_\infty}}{nh},$$

where $\widehat{\|f\|^2_\infty}$ is estimated by $\sup_{x\in I}|\widetilde{f}^2_{\log n/\sqrt{n}}(x)|$, $\widehat{\|g\|^2_\infty}$ is estimated similarly. Note that $\|K\|_1 \simeq 1.133$ and $\|K\|_\infty \simeq 0.516$, $\|K\|_1^3\|K\|_\infty \simeq 0.75$.

• **Optimal first** : As the information about compared smoothness of f and g is unavailable in practice, we have proposed an adaptive method for choosing which estimate is plugged in: we perform a classical penalized (resp. PCO) procedure (see step **Product**) to the datasets X and Y and we take as preliminary projection (resp. kernel) estimate the one for which $\widehat{m}$ (resp. $\widetilde{h}$) is the smallest (resp. largest). Indeed, the optimal dimension (resp. bandwidth) is asymptotically a decreasing (resp. increasing) function of the regularity. For instance, if $\widehat{m}_1 < \widehat{m}_2$ we proceed as in **First X** step, otherwise the roles of X

and Y are switched. We count the number of times where Y is selected first; thus, when this count is zero, **First X** and **Optimal first** are the same and give the same result.

• **Oracle (optimal first)** : Our benchmark is computed as follows. We consider for all dimensions or bandwidths the estimators of the step **Optimal first** and select the oracle that minimizes $m \mapsto \mathbb{E}[\|\widehat{(fg)}_{m,m^*} - fg\|^2]$ or $h \mapsto \mathbb{E}[\|\widetilde{(fg)}_h - fg\|^2]$. This quantity provides a numerical lower bound for the $\mathbb{L}^2$-risk of our procedure.

4.3. Numerical results

n	Product		First X		Optimal first		Oracle	
	Hermite	Kernel	Hermite	Kernel	Hermite	Kernel	Hermite	Kernel
200	18.2 (1.85)	3.21 (2.97)	12.3 (3.32)	21.0 (3.96)	7.79 (2.32,**100**)	5.80 (8.60,**79**)	2.08 (1.64)	4.69 (8.00)
1000	4.41 (0.57)	1.22 (1.05)	3.79 (1.09)	1.05 (0.68)	2.19 (0.66,**100**)	0.93 (0.6,**26**)	0.54 (0.40)	0.74 (0.54)
2000	2.26 (0.38)	0.43 (0.29)	1.92 (0.46)	0.45 (0.32)	1.18 (0.37,**100**)	0.40 (0.28,**47**)	0.27 (0.21)	0.31 (0.23)

TABLE 1
Case [BU]*: $\mathbb{L}^2$-risks with std in parenthesis (both multiplied by 10^2), $D_{\max} = 130$. For "Optimal first", the bold sub-script is the number of times where Y is selected first.*

n	Product		First X		Optimal first		Oracle	
	Hermite	Kernel	Hermite	Kernel	Hermite	Kernel	Hermite	Kernel
200	27.4 (8.13)	16.6 (9.11)	10.5 (5.07)	5.21 (3.36)	10.5 (5.07,**0**)	5.21 (3.36,**0**)	4.27 (2.87)	3.79 (2.18)
1000	38.1 (6.63)	10.1 (7.29)	3.86 (1.43)	2.77 (2.12)	3.86 (1.43,**0**)	2.80 (2.13,**2**)	$1.52_{(1.02)}$ (1.02)	1.25 (0.77)
2000	36.2 (4.98)	11.0 (10.8)	2.62 (0.80)	1.75 (1.57)	2.62 (0.80,**0**)	1.75 (1.57,**0**)	0.90 0.64)	0.80 (0.55)

TABLE 2
Case [GE]*: $\mathbb{L}^2$-risks with std in parenthesis (both multiplied by 10^2), $D_{\max} = 100$. For the Optimal first the bold sub-script is the number of times where Y is selected first.*

Let us comment the results of Tables 1-4. First, we compare separately projection and kernel procedures. Let us start with the two fully data driven methods **Product** and **Optimal first**. We observe that the results of the corresponding columns nicely confirm the theory: the risks of our procedure is almost systematically and significantly smaller (see Table 2 in particular). Besides, the risk of **Optimal first** is always comparable and even sometimes better than the risk of the **First X** method which uses the unavailable knowledge of the smoothest density. The risk of **Optimal first** has the same order as the **Oracle** even if a multiplicative factor larger than 2 is observed. Lastly, as the risks are normalized we can compare the risks of the different Tables; we see that the first two

n	Product		First X		Optimal first		Oracle	
	Hermite	Kernel	Hermite	Kernel	Hermite	Kernel	Hermite	Kernel
200	5.45	6.81	5.68	6.50	5.65	6.66	3.43	3.18
	(2.61)	(7.42)	(1.47)	(3.49)	(1.44,**51**)	(3.59,**29**)	(1.58)	(2.04)
1000	4.62	4.36	2.25	2.94	2.53	2.80	1.16	0.98
	(0.93)	(6.96)	(0.93)	(2.39)	(0.95,**100**)	(2.32,**42**)	(0.60)	(0.57)
2000	2.27	2.46	1.39	2.34	1.46	2.62	0.74	0.62
	(1.57)	(2.37)	(0.33)	(2.44)	(0.41,**55**)	(2.68,**26**)	(0.34)	(0.35)

TABLE 3
Case [NL]: $\mathbb{L}^2$-*risks with std in parenthesis (both multiplied by* 10^2*),* $D_{\max} = 100$. *For the Optimal first the bold sub-script is the number of times where* Y *is selected first.*

n	Product		First X		Optimal first		Oracle	
	Hermite	Kernel	Hermite	Kernel	Hermite	Kernel	Hermite	Kernel
200	4.60	49.8	3.24	5.19	3.24	5.18	2.43	2.65
	(2.59)	(71.8)	(2.14)	(2.48)	(2.14,**48**)	(2.48,**23**)	(1.67)	(1.86)
1000	2.31	72.3	1.84	2.51	1.87	2.46	0.84	0.78
	(0.74)	(74.6)	(0.54)	(1.87)	(0.50,**100**)	(1.87,**13**)	(0.52)	(0.68)
2000	0.92	64.1	0.73	1.88	0.83	1.77	0.43	0.39
	(0.74)	85.4)	(0.49)	(1.73)	(0.53,**95**)	(1.66,**10**)	(0.30)	(0.31)

TABLE 4
Case [NC]: $\mathbb{L}^2$-*risks with std in parenthesis (both multiplied by* 10^2*),* $D_{\max} = 50$. *For the Optimal first the bold sub-script is the number of times where* Y *is selected first.*

examples (Tables 1 and 2) are slightly more difficult which was expected: these densities are less regular as functions on $\mathbb{R}$.

Second, we can compare projection and kernel methods. The kernel method is much more time consuming that the projection method (by a factor more than 10). We can see that for the operational **Optimal first** method the kernel strategy seems better for the first two examples while the projection method wins in the two other cases. Nevertheless, the gap between the risks is never very large.

Mixed Kernel and projection strategy. Finally, we implement a mixed strategy where a kernel estimator for f, based on the kernel (4.6), is plugged in a Hermite projection estimator of the product fg. The selection procedure is performed via model selection as described above. Following the same strategies as previously, we display the estimated normalized $\mathbb{L}^2$- risks for the four examples. The adaptive **Optimal first** strategy is conducted by adaptively selecting the smoothest density with a model selection procedure, even if the plugged-in estimator is a kernel. Comparing the results displayed in Table 5 with above results, we observe an analog behavior for the mixed strategy as for the projection or kernel ones.

	Product		First X		Optimal first	
n	1000	2000	1000	2000	1000	2000
[BU]	0.71 (0.45)	0.69 (0.29)	2.52 (0.67)	1.14 (0.32)	4.43 (0.94,**100**)	2.10 (0.37,**100**)
[GE]	21.14 (6.82)	22.16 (9.40)	4.54 (1.66)	3.22 (0.86)	4.54 (1.66,**0**)	3.22 (0.86,**0**)
[NL]	3.65 (2.49)	2.52 (0.61)	3.96 (1.21)	2.49 (0.71)	3.44 (1.20,**98**)	2.40 (0.60,**100**)
[NC]	2.03 (2.44)	1.01 (0.38)	2.07 (0.43)	1.50 (0.49)	2.05 (0.45,**99**)	1.47 (0.50,**98**)

TABLE 5

Mixed Kernel-Projection : $\mathbb{L}^2$-risks with std in parenthesis (both multiplied by 10^2). For the Optimal first the bold sub-script is the number of times where Y is selected first.

5. Concluding remarks

In this paper, we have shown that an optimal strategy for estimating a product of densities was not to make a product of estimators but to plug an overfitted estimator of one of the densities in the estimator of the product. This can be done both with projection and kernel estimators and adequate model or bandwidth selection methods are proved to deliver adaptive estimators. We have implemented these methods and proved their good numerical performances in dimension 1.

We assume that the two samples have the same sizes but the case where the X-sample has size n_X and the Y sample size n_Y is worth being studied, for instance if $n_X = \gamma n$, $n_Y = n$, $\gamma \in (0, \infty)$. Following the steps of the proof in the projection case suggests that the procedure can be adapted and leads to similar results with rate induced by the smallest sample size $(1 \wedge \gamma)n$.

We considered a product of two densities but generalizations to product of other functions or product of more than two densities may be worth studying. Lastly, it is likely that our methods would extend to dependent variables, provided that the two sequences remain independent, but this should be further investigated.

If we come back to the Nadaraya-Watson problem that initiated our question, we justified in our context that plugging an overfitted estimator is an optimal strategy. Other contexts where overfitting has been recognized as judicious exists (see Chinot and Lerasle (2020)). The next step, as the original problem is a ratio, is to address the question of estimating $1/f$ when f is a density.

6. Proofs of section 2

6.1. Proof of Proposition 2.1

Consider the classical bias variance decomposition

$$\mathbb{E}[\|\widehat{(fg)}_{\mathbb{K}} - fg\|^2] = \mathbb{E}[\|\widehat{(fg)}_{\mathbb{K}} - \mathbb{E}[\widehat{(fg)}_{\mathbb{K}}]\|^2] + \|\mathbb{E}[\widehat{(fg)}_{\mathbb{K}}] - fg\|^2 := \mathbb{V} + \mathbb{B}.$$

We first study the bias term $\mathbb{B}$, note that $\mathbb{E}[\widehat{(fg)}_{\mathbb{K}}] = (\mathbb{E}[\widehat{f}_{\mathbb{K}^*}]g)_{\mathbb{K}} = \left(f_{\mathbb{K}^*}g\right)_{\mathbb{K}}$, it follows that

$$\mathbb{B} \leq 2\left(\|\left(f_{\mathbb{K}^*}g\right)_{\mathbb{K}} - (fg)_{\mathbb{K}}\|^2 + \|(fg)_{\mathbb{K}} - fg\|^2\right) := 2(\mathbb{B}_1 + \mathbb{B}_2),$$

where $\mathbb{B}_2 = \|(fg)_{\mathbb{K}} - fg\|^2$ is the standard integrated squared bias of fg. Using $(\mathcal{A}1)$ we write

$$\begin{aligned}\mathbb{B}_1 &= \int \left(\int (f_{\mathbb{K}^*}(y) - f(y))g(y)\mathbb{K}(y,x)dy\right)^2 dx \leq C_1\|(f_{\mathbb{K}^*} - f)g\|^2 \\ &\leq C_1\|g\|_\infty^2\|f_{\mathbb{K}^*} - f\|^2.\end{aligned}$$

Next we split $\mathbb{V}$ to involve the conditional variance given $(X_1, \ldots, X_n) = \mathbf{X}$

$$\begin{aligned}\mathbb{V} = \mathbb{E}&\left[\left\|\frac{1}{n}\sum_{i=1}^n \widehat{f}_{\mathbb{K}^*}(Y_i)\mathbb{K}(Y_i,.) - \mathbb{E}\left[\widehat{f}_{\mathbb{K}^*}(Y_1)\mathbb{K}(Y_1,.)\middle|\mathbf{X}\right]\right\|^2\right] \\ &+ \mathbb{E}\left[\|\left((\widehat{f}_{\mathbb{K}^*} - f_{\mathbb{K}^*})g\right)_{\mathbb{K}}\|^2\right] := \mathbb{V}_1 + \mathbb{V}_2,\end{aligned}$$

as $\mathbb{E}\left[\widehat{f}_{\mathbb{K}^*}(Y_1)\mathbb{K}(Y_1,.)\middle|\mathbf{X}\right] = (\widehat{f}_{\mathbb{K}^*}g)_{\mathbb{K}}$. For the first term, as the sum is composed of centered terms we get

$$\begin{aligned}\mathbb{V}_1 &= \mathbb{E}\left[\int \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \widehat{f}_{\mathbb{K}^*}(Y_i)\mathbb{K}(Y_i,x)\middle|\mathbf{X}\right)dx\right] \\ &\leq \frac{1}{n}\int \mathbb{E}\left[\widehat{f}_{\mathbb{K}^*}(Y_1)^2\mathbb{K}(Y_1,x)^2\right]dx \\ &= \frac{1}{n}\left\{\int \mathbb{E}\left[\left(\widehat{f}_{\mathbb{K}^*}(Y_1) - f_{\mathbb{K}^*}(Y_1)\right)^2\mathbb{K}(Y_1,x)^2\right]dx + \mathbb{E}\left[\int f_{\mathbb{K}^*}(Y_1)^2\mathbb{K}(Y_1,x)^2dx\right]\right\} \\ &\leq \frac{1}{n}\left[\iint \mathbb{E}\left[(\widehat{f}_{\mathbb{K}^*}(y) - f_{\mathbb{K}^*}(y))^2\right]\mathbb{K}(y,x)^2g(y)dydx + \iint f_{\mathbb{K}^*}(y)^2\mathbb{K}(y,x)^2g(y)dydx\right] \\ &= \frac{1}{n}\left[\int \mathrm{Var}(\widehat{f}_{\mathbb{K}^*}(y))g(y)\int \mathbb{K}(y,x)^2dxdy + \int f_{\mathbb{K}^*}(y)^2g(y)\int \mathbb{K}(y,x)^2dxdy\right] \\ &\leq \|g\|_\infty\frac{L(\mathbb{K})}{n}\left(C_4 + C_3\right),\end{aligned}$$

where we used successively $(\mathcal{A}2)$, $(\mathcal{A}4)$ and $(\mathcal{A}3)$. Finally, we write using $(\mathcal{A}5)$

$$\begin{aligned}\mathbb{V}_2 &= \mathbb{E}\left[\int \left\langle \widehat{f}_{\mathbb{K}^*} - f_{\mathbb{K}^*}, g\mathbb{K}(.,x)\right\rangle^2 dx\right] \leq \frac{C_5}{n}\iint \mathbb{K}(y,x)^2g(y)^2dydx \\ &\leq \frac{C_5}{n}L(\mathbb{K})\|g\|^2,\end{aligned}$$

where we used $(\mathcal{A}2)$ to obtain the last line. Using $\|g\|^2 \leq \|g\|_\infty$ and gathering all bounds completes the proof.

6.2. Proof of Proposition 2.2

In case [P], we have, denoting by $\psi_{\mathbf{m}}$ the orthogonal projection of ψ on the linear space $S_{\mathbf{m}} :=\text{span}(\varphi_{\mathbf{0}},\dots,\varphi_{\mathbf{m-1}})$,

$$\int\left(\int\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m-1}}\varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(y)\psi(y)dy\right)^2dx=\int\left[\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m-1}}\left(\int\varphi_{\mathbf{j}}(y)\psi(y)dy\right)\varphi_{\mathbf{j}}(x)\right]^2dx$$
$$=\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m-1}}\left(\int\varphi_{\mathbf{j}}(y)\psi(y)dy\right)^2=\|\psi_{\mathbf{m}}\|^2\le\|\psi\|^2.$$

So Assumption $(\mathcal{A}1)$ holds with $C_1=1$. On the other hand, for $(\mathcal{A}2)$, we have

$$\sup_{y\in I}\int\left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m-1}}\varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(y)\right)^2dx=\sup_{y\in I}\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m-1}}\varphi_{\mathbf{j}}^2(y)=\mathbf{L}(\mathbf{m})<+\infty.$$

In case [Ker], we denote by $u\star v(x)=\int u(y)v(x-y)dy$ the convolution product. Then by the Young inequality (8.1), we get

$$\int\left(\int\psi(y)\mathbf{K_h}(x-y)dy\right)^2dx=\|\psi\star\mathbf{K_h}\|^2\le\|\mathbf{K_h}\|_1^2\|\psi\|^2=\|K\|_1^{2d}\|\psi\|^2.$$

Thus $(\mathcal{A}1)$ holds with $C_1=\|K\|_1^{2d}$. As $\forall j$, $\int K_h^2(y_j-x_j)dx_j=\int K^2(u)du/h_j<+\infty$, Assumption $(\mathcal{A}2)$ is satisfied and $\mathbf{L}(\mathbf{K_h})=\|K\|^{2d}\prod_{j=1}^d h_j^{-1}$.

6.3. Proof of Proposition 2.3

In case [P], we have $\mathbb{E}[\widehat{f}_{\mathbf{m}^*}(y)]=f_{\mathbf{m}^*}(y)$ where $f_{\mathbf{m}^*}$ is the projection of f on $S_{\mathbf{m}^*}=\text{span}(\varphi_{\mathbf{0}},\dots,\varphi_{\mathbf{m}^*-1})$. Therefore $\int\left(\mathbb{E}[\widehat{f}_{\mathbf{m}^*}(y)]\right)^2dy=\|f_{\mathbf{m}^*}\|^2\le\|f\|^2\le\|f\|_\infty$, and $(\mathcal{A}3)$ holds with $C_3=\|f\|^2$ or $C_3=\|f\|_\infty$. Next we write

$$\int\text{Var}\left(\widehat{f}_{\mathbf{m}^*}(y)\right)dy=\int\text{Var}\left[\frac{1}{n}\sum_{i=1}^n\left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}}\varphi_{\mathbf{j}}(X_i)\varphi_{\mathbf{j}}(y)\right)\right]dy$$
$$=\frac{1}{n}\int\text{Var}\left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}}\varphi_{\mathbf{j}}(X_1)\varphi_{\mathbf{j}}(y)\right)dy$$
$$\le\frac{1}{n}\int\mathbb{E}\left[\left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}}\varphi_{\mathbf{j}}(X_1)\varphi_{\mathbf{j}}(y)\right)^2\right]dy$$
$$=\frac{1}{n}\mathbb{E}\left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}}\varphi_{\mathbf{j}}^2(X_1)\right)\le\frac{\mathbf{L}(\mathbf{m}^*)}{n}.$$

Therefore ($\mathcal{A}4$) holds with $C_4 = 1$. Lastly, recalling that $\widehat{a}_{\mathbf{j}} = n^{-1}\sum_{i=1}^n \varphi_{\mathbf{j}}(X_i)$, and noting that $a_{\mathbf{j}} = \langle \varphi_{\mathbf{j}}, f\rangle = \mathbb{E}(\widehat{a}_{\mathbf{j}})$, we have

$$\begin{aligned}
&\mathbb{E}\left(|\langle \psi, \widehat{f}_{\mathbf{m}^*} - \mathbb{E}\widehat{f}_{\mathbf{m}^*}\rangle|^2\right)\\
=\;& \iint \psi(u)\psi(v)\mathbb{E}\left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}} (\widehat{a}_{\mathbf{j}} - a_{\mathbf{j}})\varphi_{\mathbf{j}}(u) \sum_{\mathbf{0}\le\mathbf{k}\le\mathbf{m}^*-\mathbf{1}} (\widehat{a}_{\mathbf{k}} - a_{\mathbf{k}})\varphi_{\mathbf{k}}(v)\right) dudv\\
=\;& \frac{1}{n}\iint \psi(u)\psi(v) \sum_{\mathbf{0}\le\mathbf{j},\mathbf{k}\le\mathbf{m}^*-\mathbf{1}} \mathrm{cov}(\varphi_{\mathbf{j}}(X_1)\varphi_{\mathbf{j}}(u), \varphi_{\mathbf{k}}(X_1)\varphi_{\mathbf{k}}(v))dudv\\
\le\;& \frac{1}{n}\iint \psi(u)\psi(v) \sum_{\mathbf{0}\le\mathbf{j},\mathbf{k}\le\mathbf{m}^*-\mathbf{1}} \mathbb{E}(\varphi_{\mathbf{j}}(X_1)\varphi_{\mathbf{j}}(u)\varphi_{\mathbf{k}}(X_1)\varphi_{\mathbf{k}}(v))dudv
\end{aligned}$$

as the omitted term is the opposite of a nonnegative quantity (it can be written as the opposite of a square). It follows that

$$\begin{aligned}
&\mathbb{E}\left(|\langle \psi, \widehat{f}_{\mathbf{m}^*} - \mathbb{E}\widehat{f}_{\mathbf{m}^*}\rangle|^2\right)\\
\le\;& \frac{1}{n}\sum_{\mathbf{0}\le\mathbf{j},\mathbf{k}\le\mathbf{m}^*-\mathbf{1}} \iiint \psi(u)\psi(v)\varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(u)\varphi_{\mathbf{k}}(x)\varphi_{\mathbf{k}}(v)f(x)dudvdx\\
=\;& \frac{1}{n}\int \left[\int \left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}} \varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(u)\right)\psi(u)du\right]^2 f(x)dx\\
\le\;& \frac{\|f\|_\infty}{n}\int \left[\int \left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}} \varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(u)\right)\psi(u)du\right]^2 dx\\
\le\;& \frac{\|f\|_\infty}{n}\sum_{\mathbf{0}\le\mathbf{j},\mathbf{k}\le\mathbf{m}^*-\mathbf{1}} \iint \psi(u)\psi(v)\left(\int \varphi_{\mathbf{j}}(x)\varphi_{\mathbf{k}}(x)dx\right)\varphi_{\mathbf{j}}(u)\varphi_{\mathbf{k}}(v)dudv\\
=\;& \frac{\|f\|_\infty}{n}\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}} \left(\int \psi(u)\varphi_{\mathbf{j}}(u)du\right)^2 = \frac{\|f\|_\infty}{n}\|\psi_{\mathbf{m}^*}\|^2 \le \frac{\|f\|_\infty}{n}\|\psi\|^2.
\end{aligned}$$

We obtain that ($\mathcal{A}5$) holds with $C_5 = \|f\|_\infty$.

Now we consider case [Ker] and note that $\mathbb{E}(\widehat{f}_{\mathbf{h}^*}(x)) = f \star \mathbf{K}_{\mathbf{h}^*}(x)$. As by Young Inequality (8.1), $\|f \star \mathbf{K}_{\mathbf{h}^*}\|^2 \le \|\mathbf{K}_{\mathbf{h}^*}\|_1^2\|f\|^2 = \|\mathbf{K}\|_1^2\|f\|^2$, Assumption ($\mathcal{A}3$) is satisfied with $C_3 = \|K\|_1^{2d}\|f\|^2$ (or $C_3 = \|K\|_1^{2d}\|f\|_\infty$). A standard bound (see Tsybakov (2009), proposition 1.4) yields $\int \mathrm{Var}\left(\widehat{f}_{\mathbf{h}^*}(x)\right)dx \le$

$\|K\|^{2d}/(n\prod_{j=1}^d h_j^*) \leq 1$, so that ($\mathcal{A}4$) holds with $C_4 = 1$. Lastly

$$\begin{aligned}
&\mathbb{E}\left(\langle\psi, \frac{1}{n}\sum_{i=1}^n \mathbf{K}_{\mathbf{h}^*}(X_i - \cdot) - f \star \mathbf{K}_{\mathbf{h}^*}\rangle^2\right) = \frac{1}{n}\mathrm{Var}\left(\langle\psi, \mathbf{K}_{\mathbf{h}^*}(X_1 - .)\rangle\right)\\
\leq\ & \frac{1}{n}\mathbb{E}\left(\langle\psi, \mathbf{K}_{\mathbf{h}^*}(X_1 - .)\rangle^2\right) \leq \frac{1}{n}\int\left(\int \psi(u)\mathbf{K}_{\mathbf{h}^*}(x-u)du\right)^2 f(x)dx\\
\leq\ & \frac{\|f\|_\infty}{n}\|\psi \star \mathbf{K}_{\mathbf{h}^*}\|^2 \leq \|f\|_\infty\|\mathbf{K}\|_1^2\frac{\|\psi\|^2}{n} = \|f\|_\infty\|K\|_1^{2d}\frac{\|\psi\|^2}{n},
\end{aligned}$$

where we used Young's inequality. Therefore ($\mathcal{A}5$) is holds with $C_5 = \|f\|_\infty\|K\|_1^{2d}$.

7. Proofs of adaptive results

7.1. Proof of Theorem 3.1

The proof starts by decompositions which are standard when studying Goldenschluger and Lepski (2011) methods and bounds. For $\kappa' \geq \kappa$, we get (see Comte (2017), sec 4.2)

$$\mathbb{E}\left(\|\widehat{(fg)}_{\widehat{\tau}} - fg\|^2\right) \leq 3\mathbb{E}(\|\widehat{(fg)}_\tau - fg\|^2) + 6\kappa' V(\tau) + 6\mathbb{E}(A(\tau)). \tag{7.1}$$

The first right-hand-side term has expectation controlled by applying Proposition 2.1 and the term $V(\tau)$ can be associated with it. Only $A(\tau)$ must be studied. We write

$$\|\widehat{(fg)}_{\tau'} - \widehat{(fg)}_{\tau,\tau'}\|^2 \leq 3(\|\widehat{(fg)}_{\tau'} - (fg)_{\tau'}\|^2 + \|(fg)_{\tau'} - (fg)_{\tau,\tau'}\|^2 + \|\widehat{(fg)}_{\tau,\tau'} - (fg)_{\tau,\tau'}\|^2).$$

The bound on the middle term is obtained with ($\mathcal{A}1$)

$$\|(fg)_{\tau'} - (fg)_{\tau,\tau'}\|^2 \leq C_1\|(fg)_\tau - (fg)\|^2$$

which refers to an adequate bias term. For the last term, we use ($\mathcal{A}1$) and get

$$\|\widehat{(fg)}_{\tau,\tau'} - (fg)_{\tau,\tau'}\|^2 \leq C_1\|\widehat{(fg)}_\tau - (fg)_\tau\|^2.$$

The expectation of this term is also studied in Proposition 2.1, with $\mathbb{E}(\|\widehat{(fg)}_\tau - (fg)_\tau\|^2) = \mathbb{V} + \mathbb{B}_1$ with the notation of this proof. Next

$$\|\widehat{(fg)}_{\tau'} - (fg)_{\tau'}\|^2 \leq 2(\|\widehat{(fg)}_{\tau'} - (f_{\tau^*}g)_{\tau'}\|^2 + \|(f_{\tau^*}g)_{\tau'} - (fg)_{\tau'}\|^2)$$

where we have the bound $\|(f_{\tau^*}g)_{\tau'} - (fg)_{\tau'}\|^2 \leq C_1\|g\|_\infty^2\|f_{\tau^*} - f\|^2$ from ($\mathcal{A}1$). Now we notice that $\|\widehat{(fg)}_\tau - (f_{\tau^*}g)_\tau\|^2 = \sup_{t\in\mathcal{B}(0,1)}\nu_n^2(t)$ where $\mathcal{B}(0,1)$ is a countable set of square integrable functions with $\|t\| = 1$ and the empirical process is defined by

$$\nu_n(t) = \langle\widehat{(fg)}_\tau - (f_{\tau^*}g)_\tau, t\rangle.$$

Therefore we have

$$\|\widehat{(fg)}_\tau - (fg)_\tau\|^2 \le 2\left(\sup_{t\in\mathcal{B}(0,1)} \nu_n^2(t) + C_1\|g\|_\infty^2\|f_{\tau^*}-f\|^2\right).$$

Reminding the definition of $A(\tau)$ given by (3.1), we have

$$\mathbb{E}(A(\tau)) \le 6\mathbb{E}\left(\sup_{\tau'\in\mathcal{K}_n}\sup_{t\in\mathcal{B}(0,1)} \nu_n^2(t) - \frac{\kappa}{6}V(\tau')\right) + 3C_1\mathbb{E}(\|\widehat{(fg)}_\tau - (fg)_\tau\|^2)$$
$$+ 6C_1^2\|g\|_\infty^2\|f_{\tau^*}-f\|^2 + 6C_1\|(fg)_\tau - (fg)\|^2. \tag{7.2}$$

Thus, the result of Theorem 3.1 holds if we prove that, for two constants $\mathfrak{c}_1, \mathfrak{c}_2$, we have

$$\sum_{\tau'\in\mathcal{K}_n} \mathbb{E}\left(\sup_{t\in\mathcal{B}(0,1)} \nu_n^2(t) - \mathfrak{c}_1 V(\tau')\right) \le \frac{C}{n}. \tag{7.3}$$

Indeed, plugging (7.3) in (7.2) and the result in (7.1) is the result of Theorem 3.1.

We prove (7.3). We split ν_n in four terms: $\nu_n = \nu_{n,1} + \nu_{n,2} + \nu_{n,3} + \nu_{n,4}$ where for some positive constant c_0 to be defined in the sequel, we set

$$A(x) = \{|\widehat{f}_{\tau^*}(x) - f_{\tau^*}(x)| < c_0\},$$

and

$$\nu_{n,1}(t) = \frac{1}{n}\sum_{i=1}^n [(\widehat{f}(Y_i) - f_{\tau^*}(Y_i))\mathbf{1}_{A(Y_i)}\langle \mathbb{K}_\tau(Y_i,.),t\rangle - \langle((\widehat{f}-f_{\tau^*})\mathbf{1}_A g)_\tau, t\rangle],$$

$$\nu_{n,2}(t) = \frac{1}{n}\sum_{i=1}^n [(\widehat{f}(Y_i) - f_{\tau^*}(Y_i))\mathbf{1}_{(A(Y_i))^c}\langle \mathbb{K}_\tau(Y_i,.),t\rangle - \langle((\widehat{f}-f_{\tau^*})\mathbf{1}_{A^c} g)_\tau, t\rangle],$$

$$\nu_{n,3}(t) = \langle((\widehat{f}-f_{\tau^*})g)_\tau, t\rangle = \frac{1}{n}\sum_{i=1}^n \psi_t(X_i), \quad \psi_t(X) = \left\langle\left((\mathbb{K}_{\tau^*}(X,.)-f_{\tau^*})g\right)_\tau, t\right\rangle,$$

$$\nu_{n,4}(t) = \frac{1}{n}\sum_{i=1}^n [f_{\tau^*}(Y_i)\langle \mathbb{K}_\tau(Y_i,.),t\rangle - \langle (f_{\tau^*}g)_\tau, t\rangle].$$

Study of $\nu_{n,2}$ We start by the study of $\nu_{n,2}$ as it leads to fix c_0, and we first establish that $\mathbb{E}(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,2}(t)|) \le n^{-p}$ for some positive p. It holds that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\in\mathcal{B}(0,1)}[\nu_{n,2}^2(t)]\Big|\mathbf{X}\right] \\
\le\ & \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n [(\widehat{f}_{\tau^*}(Y_i) - f_{\tau^*}(Y_i))\mathbf{1}_{(A(Y_i))^c}\mathbb{K}_\tau(Y_i,.) - ((\widehat{f}_{\tau^*}-f_{\tau^*})\mathbf{1}_{A^c}g)_\tau]\right\|^2\Big|\mathbf{X}\right] \\
\le\ & \frac{1}{n}\int \mathbb{E}\left[(\widehat{f}_{\tau^*}(Y_1) - f_{\tau^*}(Y_1))^2\mathbf{1}_{(A(Y_1))^c}\mathbb{K}_\tau(Y_1,x)^2\Big|\mathbf{X}\right]dx \\
\le\ & 2\frac{L(\tau)}{n}\int \mathbb{E}\left[(\widehat{f}_{\tau^*}(Y_1)^2 + f_{\tau^*}(Y_1))^2\mathbf{1}_{(A(Y_1))^c}\Big|\mathbf{X}\right]dx.
\end{aligned}$$

Using ($\mathcal{A}9$) we obtain that $\widehat{f}_{\tau^*}(Y_1)^2 \leq C_9 L^2(\tau^*) \leq C_{10} n^2$ (where the last inequality follows from ($\mathcal{A}10$)), and with ($\mathcal{A}7$) that

$$\mathbb{E}\left[\sup_{t\in\mathcal{B}(0,1)}[\nu_{n,2}^2(t)]\Big|\mathbf{X}\right] \leq 2(C_{10}n^2 + C_7^2)\frac{L(\tau)}{n}\mathbb{P}(A^c(Y_1)) \leq 2(C_{10}n^2 + C_7^2)\mathbb{P}(A^c(Y_1)).$$

Therefore,

$$\mathbb{E}\left(\sup_{t\in\mathcal{B}(0,1)}[\nu_{n,2}^2(t)]\right) \leq 2(C_{10}n + C_7^2)\int \mathbb{P}(|\widehat{f}(y) - f_{\tau^*}(y)| > c_0)g(y)dy. \quad (7.4)$$

We complete by applying the Bernstein inequality to $Z_i = \mathbb{K}_{\tau^*}(X_i, y)$ yielding

$$\mathbb{P}\left(|\widehat{f}_{\tau^*}(y) - f_{\tau^*}(y)| > c_0\right) \leq 2\exp\left(-\frac{nc_0^2}{2(v_2^2 + b_2c_0)}\right)$$

with v_2^2 a bound on $\mathrm{Var}(Z_i)$ and b_2 an a.s. bound on Z_i. We find with ($\mathcal{A}9$) that $b_2 = \sqrt{C_9}L(\tau^*)$ suits and with ($\mathcal{A}2$) $\mathrm{Var}(Z_i) \leq \|f\|_\infty L(\tau^*) = v_2^2$. Therefore, choosing

$$c_0 = \|f\|_\infty, \quad (7.5)$$

and using ($\mathcal{A}10$) , it follows that

$$\mathbb{P}\left(|\widehat{f}_{\tau^*}(y) - f_{\tau^*}(y)| > c_0\right) \leq 2n^{-(d+3)}. \quad (7.6)$$

Then, gathering (7.4) and (7.6) leads to $\mathbb{E}(\sup_{t\in\mathcal{B}(0,1)}\nu_{n,2}^2(t)) \leq 4(C_{10}n^2 + C_7^2)n^{-(d+3)}$.
As a consequence under ($\mathcal{A}8$), we get

$$\sum_{\tau\in\mathcal{K}_n}\mathbb{E}(\sup_{t\in\mathcal{B}(0,1)}[\nu_{n,2}^2(t)]) \leq \frac{C}{n}. \quad (7.7)$$

Study of $\nu_{n,1}$. We apply the Talagrand inequality (see Lemma 8.1) to $\nu_{n,1}$ conditionally to $\mathbf{X}$. Using that $t \mapsto \nu_{n,1}(t)$ is linear and the Cauchy-Schwarz inequality and $\|t\|^2 = 1$, we get

$$\begin{aligned}\left(\mathbb{E}\left(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,1}(t)|\,|\mathbf{X}\right)\right)^2 \leq \mathbb{E}\left(\sup_{t\in\mathcal{B}(0,1)}\nu_{n,1}^2(t)|\mathbf{X}\right)&\\ \leq \frac{1}{n}\int \mathrm{Var}[(\widehat{f}_{\tau^*}(Y_1) - f_{\tau^*}(Y_1))\mathbf{1}_{A(Y_1)}\mathbb{K}_\tau(Y_1, y)|\mathbf{X}]dy&\\ \leq \frac{1}{n}\int \mathbb{E}[(\widehat{f}_{\tau^*}(Y_1) - f_{\tau^*}(Y_1))^2\mathbf{1}_{A(Y_1)}\mathbb{K}_\tau(Y_1, y)^2|\mathbf{X}]dy&\\ \leq \frac{c_0^2 L(\tau)}{n} := H_1^2.&\end{aligned}$$

Next, note that

$$\sup_{x,t}|(\widehat{f}_{\tau^*}(x)-f_{\tau^*}(x))\mathbf{1}_{A(x)}\langle\mathbb{K}_\tau(x,.),t\rangle| \le c_0\sup_{t,x}|\langle\mathbb{K}_\tau(x,.),t\rangle| \le c_0\|\mathbb{K}_\tau(x,.)\| = c_0\sqrt{L(\tau)} := b_1 \quad (7.8)$$

and

$$\begin{aligned}\sup_t \mathrm{Var}((\widehat{f}_{\tau^*}(Y_1)-f_{\tau^*}(Y_1))\mathbf{1}_{A(Y_1)}\langle\mathbb{K}_\tau(Y_1,.),t\rangle|\mathbf{X}) &\le \sup_t \mathbb{E}[(\widehat{f}_{\tau^*}(Y_1)^2\mathbf{1}_{A(Y_1)}\langle\mathbb{K}_\tau(Y_1,.),t\rangle|\mathbf{X})\\ &= \sup_t\int \widehat{f}_{\tau^*}(u)^2\mathbf{1}_{A(y)}\langle\mathbb{K}_\tau(y,.),t\rangle|\mathbf{X})g(y)dy\\ &\le C_1C_7^2c_0^2\|g\|_\infty := v_1^2, \end{aligned} \quad (7.9)$$

where we used $(\mathcal{A}1)$ and $(\mathcal{A}7)$. Applying Lemma 8.1 with $\delta=\frac{1}{2}$, it follows that

$$\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,1}(t)|^2-4\frac{c_0^2L(\tau)}{n}\right)_+\Big|\mathbf{X}\right] \le \frac{4c_0^2}{nK_1}\left(C_1C_7^2\|g\|_\infty\exp\left(-K_1\frac{L(\tau)}{2\|g\|_\infty}\right)+\frac{49}{K_1C^2(1/2)}\exp\left(-\frac{K_1C(1/2)}{7}\sqrt{n}\right)\right),$$

using $L(\tau)\le n$. Since the latter bound does not depend on $\mathbf{X}$, the inequality holds unconditionally in expectation. Therefore under $(\mathcal{A}8)$ and as $L(\tau)\ge 1$, we get, for C a positive constant, and using (7.5),

$$\sum_{\tau\in\mathcal{K}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,1}(t)|^2-4\frac{\|f\|_\infty^2L(\tau)}{n}\right)_+\right]\le\frac{C}{n}. \quad (7.10)$$

Study of $\nu_{n,3}$ Similarly to $\nu_{n,1}$ we apply the Talagrand inequality

$$\begin{aligned}&\left(\mathbb{E}\left(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,3}(t)|\right)\right)^2 \le \mathbb{E}\left[\left\|\left((\widehat{f}_{\tau^*}-f_{\tau^*})g\right)_\tau\right\|^2\right]\\ &= \frac{1}{n}\int \mathrm{Var}\left(\left(\mathbb{K}_{\tau^*}(X_1,.)g\right)_\tau(y)\right)dy\\ &\le \frac{1}{n}\int\mathbb{E}\left(\left(\mathbb{K}_{\tau^*}(X_1,.)g\right)_\tau(y)^2\right)dy = \frac{1}{n}\iint\left(\int\mathbb{K}_\tau(x,y)\mathbb{K}_{\tau^*}(z,x)g(x)dx\right)^2f(z)dzdy\\ &\le C_1\frac{\|f\|_\infty}{n}\iint g(x)^2\mathbb{K}_\tau(x,y)^2dxdy \le C_1\frac{\|f\|_\infty\|g\|^2}{n}L(\tau) := H_3^2,\end{aligned}$$

using $(\mathcal{A}1)$ and $(\mathcal{A}2)$. Next, note that using Cauchy-Schwarz inequality and $(\mathcal{A}1)$

$$\sup_{x,t}\left|\left\langle\left((\mathbb{K}_{\tau^*}(x,.)-f_{\tau^*})g\right)_\tau,t\right\rangle\right| \le C_1\sup_x\left\|\left(\mathbb{K}_{\tau^*}(x,.)-f_{\tau^*}\right)g\right\| \le C_1\|g\|_\infty\left(\sqrt{L(\tau^*)}+\sqrt{C_7}\right)\le 2C_1\sqrt{n}\|g\|_\infty := b_3,$$

using ($\mathcal{A}7$) with $\|f_{\tau^*}\| \leq \sqrt{\|f_{\tau^*}\|_\infty}$ and ($\mathcal{A}10$). Finally, ($\mathcal{A}1$) enables to write

$$\begin{aligned}\sup_t \mathrm{Var}\left(\langle((\mathbb{K}_{\tau^*}(X,.)-f_{\tau^*})g)_\tau,t\rangle\right) &\leq \sup_t \mathbb{E}\left[\langle\left(\mathbb{K}_{\tau^*}(X,.)\right)g)_\tau,t\rangle^2\right]\\ &\leq C_1\|f\|_\infty \sup_t \int\left(\int \mathbb{K}_\tau(u,v)t(u)du\right)^2 g^2(v)dv\\ &\leq C_1^2\|f\|_\infty\|g\|_\infty^2 := v_3^2.\end{aligned}$$

Applying Lemma 8.1 with $\delta=\frac{1}{2}$, it follows that

$$\begin{aligned}&\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,3}(t)|^2-4C_1\|f\|_\infty\|g\|_\infty\frac{L(\tau)}{n}\right)_+\right]\\ &\leq \frac{4C_1^2\|g\|_\infty^2}{K_1\,n}\left(\|f\|_\infty e^{-K_1\frac{L(\tau)}{2C_1\|g\|_\infty}}+\frac{49\times 4C_1^2}{K_1C^2(1/2)}e^{-\frac{K_1C(1/2)\sqrt{\|f\|_\infty}}{14\sqrt{C_1\|g\|_\infty}}\sqrt{L(\tau)}}\right).\end{aligned}$$

Therefore under ($\mathcal{A}8$), we get, for C a positive constant,

$$\sum_{\tau\in\mathcal{K}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,3}(t)|^2-4C_1\|f\|_\infty\|g\|_\infty\frac{L(\tau)}{n}\right)_+\right]\leq\frac{C}{n}. \tag{7.11}$$

Study of $\nu_{n,4}$ Again we apply the Talagrand inequality, similar computations enable to derive from ($\mathcal{A}1$) and ($\mathcal{A}7$) $H_4^2=C_1\|f\|_\infty\|g\|_\infty L(\tau)/n$, $v_4^2=C_1C_7\|g\|_\infty$ and $b_4=C_7\sqrt{n}$. It follows, by applying Lemma 8.1 with $\delta=\frac{1}{2}$, that

$$\begin{aligned}&\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,4}(t)|^2-4C_1\|f\|_\infty\|g\|_\infty\frac{L(\tau)}{n}\right)_+\right]\\ &\leq\frac{4}{K_1}\left(\frac{C_1C_7\|g\|_\infty}{n}e^{-K_1\frac{L(\tau)\|f\|_\infty}{2C_7}}+\frac{49\times C_7^2}{K_1nC^2(1/2)}e^{-\frac{K_1C(1/2)}{14}\frac{\sqrt{C_1\|f\|_\infty\|g\|_\infty}}{C_7}\sqrt{n}}\right).\end{aligned}$$

Therefore under ($\mathcal{A}8$) we get, for C a positive constant,

$$\sum_{\tau\in\mathcal{K}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}(0,1)}|\nu_{n,4}(t)|^2-4C_1\frac{\|f\|_\infty\|g\|_\infty L(\tau)}{n}\right)_+\right]\leq\frac{C}{n}. \tag{7.12}$$

As a consequence, gathering (7.7)-(7.10)-(7.11) and (7.12) gives the result (7.3) for C a positive finite constant, depending on a, $\|f\|_\infty$, $\|g\|_\infty$, C_1 and C_7.

7.2. *Proof of Corollary 3.1*

We already checked that [P] satisfies ($\mathcal{A}1$) and ($\mathcal{A}2$), and that for f bounded, Assumptions ($\mathcal{A}3$)-($\mathcal{A}5$) hold under $\mathbf{L}(\mathbf{m}^*)\leq n$, which is ensured by [P3]. For

($\mathcal{A}6$), we write

$$\begin{aligned}\int \sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}-\mathbf{1}} \varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(z) \sum_{\mathbf{0}\le\mathbf{k}\le\mathbf{m}'-\mathbf{1}} \varphi_{\mathbf{k}}(z)\varphi_{\mathbf{k}}(y)dz \\ = \sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}-\mathbf{1}} \sum_{\mathbf{0}\le\mathbf{k}\le\mathbf{m}'-\mathbf{1}} \varphi_{\mathbf{j}}(x)\varphi_{\mathbf{k}}(y) \underbrace{\int \varphi_{\mathbf{j}}(z)\varphi_{\mathbf{k}}(z)dz}_{=\delta_{j,k}} \\ = \sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}\wedge\mathbf{m}'-\mathbf{1}} \varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(y),\end{aligned}$$

which is clearly symmetric in $\mathbf{m}$, $\mathbf{m}'$. Therefore ($\mathcal{A}6$) is fulfilled.

To check ($\mathcal{A}7$) we have to bound $\|f_{\mathbf{m}^*}\|_\infty$. Under [P2], we have

$$|f_{\mathbf{m}^*}(x)| = \left|a_{\mathbf{0}}(f)\varphi_{\mathbf{0}}(x) + \sum_{\mathbf{1}\le\mathbf{j}\le\mathbf{m}^*-\mathbf{1}} a_{\mathbf{j}}(f)\varphi_{\mathbf{j}}(x)\right| \le C_\varphi \left(|a_{\mathbf{0}}| + \sum_{\mathbf{j}\ge\mathbf{1}} |a_{\mathbf{j}}(f)|\right).$$

Then using [P5], we have

$$\begin{aligned}|f_{\mathbf{m}^*}(x)| &\le C_\varphi \left(C_\varphi + \sqrt{\sum_{\mathbf{j}\ge\mathbf{1}} \mathbf{j}^{\mathbf{b}} a_{\mathbf{j}}^2(f) \sum_{\mathbf{j}\ge\mathbf{1}} \mathbf{j}^{-\mathbf{b}}}\right) \\ &\le C_\varphi \left(C_\varphi + \sqrt{C_{\mathbf{b}} \sum_{\mathbf{j}\ge\mathbf{1}} \mathbf{j}^{-\mathbf{b}}}\right) := C(\mathbf{b},\varphi) < +\infty\end{aligned}$$

since $\min_{1\le j\le d} b_j > 1$. Thus, Assumptions [P5] and [P2] imply that ($\mathcal{A}7$) holds with $C_7 = C(\mathbf{b},\varphi)$.

Assumption [P4] is analogous to ($\mathcal{A}8$). Lastly

$$\sup_{x,y} \left|\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}-\mathbf{1}} \varphi_{\mathbf{j}}(x)\varphi_{\mathbf{j}}(y)\right|^2 \le \sup_x \left(\sum_{\mathbf{0}\le\mathbf{j}\le\mathbf{m}-\mathbf{1}} \varphi_{\mathbf{j}}^2(x)\right)^2 = \mathbf{L}^2(\mathbf{m}),$$

so that ($\mathcal{A}9$) holds with $C_9 = 1$, and therefore [P3] is ($\mathcal{A}10$).

7.3. *Proof of Corollary 3.2*

We already checked that [Ker] satisfies ($\mathcal{A}1$) and ($\mathcal{A}2$), and that for f bounded, Assumptions ($\mathcal{A}3$)-($\mathcal{A}5$) hold under $\mathbf{L}(\mathbf{h}^*) \le n$, which is ensured by [K3] (for n large enough). Next we compute

$$\begin{aligned}\int \mathbf{K}_{\mathbf{h}}(x-z)\mathbf{K}_{\mathbf{h}'}(z-y)dz &= \int \mathbf{K}_{\mathbf{h}}(u)\mathbf{K}_{\mathbf{h}'}(x-u-y)du \\ &= \int \mathbf{K}_{\mathbf{h}}(v-y)\mathbf{K}_{\mathbf{h}'}(x-v)dz,\end{aligned}$$

where we made successively the affine changes of variables $u = x - z$ and $v = u + y$. Using that K is even it follows that Assumption (A6) is fulfilled. Using [K1], we easily bound $\|f_{\mathbf{h}^*}\|_\infty \leq \|f\|_\infty \|\mathbf{K}\|_1$, implying that (A7) holds with $C_7 = \|f\|_\infty \|\mathbf{K}\|_1$. Assumption [K4] implies (A8). Lastly, $\sup_{x,y} |\mathbf{K}_\mathbf{h}(x,y)|^2 \leq \mathbf{L}^2(\mathbf{h})\|\mathbf{K}\|_\infty^2/\|\mathbf{K}\|^4$, so that (A9) holds with $C_9 = \|\mathbf{K}\|_\infty^2/\|\mathbf{K}\|^4$ and [K3] implies (A10).

7.4. Proof of Theorem 4.2

Following the first step in Lacour *et al.* (2017), we write

$$\|\widehat{(fg)}_{\widetilde{h}} - fg\|^2 \leq \|\widehat{(fg)}_h - fg\|^2 + (\mathrm{pen}(h) - \psi_n(h)) - (\mathrm{pen}(\widetilde{h}) - \psi_n(\widetilde{h})) \quad (7.13)$$

with

$$\psi_n(h, h_{\min}) = \langle \widehat{(fg)}_h - fg, \widehat{(fg)}_{h_{\min}} - fg \rangle.$$

As in Comte and Marie (2021), we decompose ψ_n in

$$\psi_n(h, h_{\min}) = \psi_{1,n}(h, h_{\min}) + \psi_{2,n}(h, h_{\min}) + \psi_{3,n}(h, h_{\min}).$$

First,

$$\psi_{1,n}(h, , h_{\min}) := \frac{\langle K_h, K_{h_{\min}} \rangle}{n^2} \sum_{i=1}^n \widehat{f}^2_{h^*}(Y_i) + \frac{U(h, h_{\min})}{n^2} = \mathrm{pen}_1(h) + \frac{U(h, h_{\min})}{n^2},$$

where

$$U_n(h, h') := \sum_{1 \leq i \neq j \leq n} \langle \widehat{f}_{h^*}(Y_i) K_h(Y_i - \cdot) - (fg)_h, \widehat{f}_{h^*}(Y_j) K_{h'}(Y_j - \cdot) - (fg)_{h'} \rangle. \quad (7.14)$$

Indeed,

$$\frac{\langle K_h, K_{h_{\min}} \rangle}{n^2} \sum_{i=1}^n \widehat{f}^2_{h^*}(Y_i) = \frac{1}{n^2} \sum_{i=1}^n \langle \widehat{f}_{h^*}(Y_i) K_h(Y_i - \cdot), \widehat{f}_{h^*}(Y_i) K_{h_{\min}}(Y_i - \cdot) \rangle.$$

Second,

$$\psi_{2,n}(h, h_{\min}) := +\frac{1}{n} \langle (fg)_h, (fg)_{h_{\min}} \rangle - \frac{1}{n^2} \left(\sum_{i=1}^n \langle \widehat{f}_{h^*}(Y_i) K_{h_{\min}}(Y_i - \cdot), (fg)_h \rangle + \sum_{i=1}^n \langle \widehat{f}_{h^*}(Y_i) K_h(Y_i - \cdot), (fg)_{h_{\min}} \rangle \right) \quad (7.15)$$

and lastly

$$\psi_{3,n}(h, h_{\min}) := V_n(h, h_{\min}) + V_n(h_{\min}, h) + \langle (fg)_h - fg, (fg)_{h_{\min}} - fg \rangle$$

with $V_n(h, h') := \langle \widehat{(fg)}_h - (fg)_h, (fg)_{h'} - fg \rangle$.

We state a series of Lemmas that permit to establish Theorem 4.2.

Lemma 7.1. *Under the assumptions of Theorem 4.2, it holds that*

$$\mathbb{E}\left(\sup_{h,h'\in\mathcal{H}_n}|\psi_{2,n}(h,h')|\right)\leq\frac{C}{n},$$

where $C=C(f,g,K)$ is a positive constant depending on f, g, K.

Lemma 7.2. *Under the assumptions of Theorem 4.2, for every $\vartheta\in(0,1)$, it holds*

$$\mathbb{E}\left(\sup_{h,h'}\left\{|V_n(h,h')|-\vartheta\|(fg)_{h'}-fg\|^2\right\}\right)\leq\frac{1}{2\vartheta}\|K\|_1^2\|g\|_\infty\|f_{h^*}-f\|^2+C\frac{\log(n)}{n}.$$

Lemma 7.3. *Under the assumptions of Theorem 4.2, for every $\vartheta\in[0,1]$, it holds*

$$\mathbb{E}\left(\sup_{h\in\mathcal{H}_n}\left\{\frac{|U_n(h,h_{\min})|}{n^2}-\vartheta\frac{c_0(f,g,K)}{nh}\right\}\right)\leq\vartheta\|f_{h^*}-f\|^2+\frac{C\log(n)}{n}.$$

We deduce from (7.13) that

$$\begin{aligned}\|\widehat{(fg)}_{\widetilde{h}}-fg\|^2 \quad\leq\quad & \|\widehat{(fg)}_h-fg\|^2+2\left(\mathrm{pen}_1(h)-\psi_n(h)\right)+2\mathrm{pen}_2(h)\\ & -2(\mathrm{pen}_1(\widetilde{h})-\psi_n(\widetilde{h}))-2\mathrm{pen}_2(\widetilde{h}).\end{aligned}\tag{7.16}$$

We have for all positive h

$$\begin{aligned}\psi_n(h)-\mathrm{pen}_1(h)=\frac{U_n(h,h_{\min})}{n^2}&+\psi_{2,n}(h,h_{\min})+V_n(h,h_{\min})+V_n(h_{\min},h)\\ &+\langle(fg)_h-fg,(fg)_{h_{\min}}-fg\rangle.\end{aligned}$$

We note that for all positive θ

$$|\langle(fg)_h-fg,(fg)_{h_{\min}}-fg\rangle|\leq\frac{\theta}{2}\|(fg)_h-fg\|^2+\frac{1}{2\theta}\|(fg)_{h_{\min}}-fg\|^2.$$

Applying these for $h=\widetilde{h}$ we get

$$\begin{aligned}&\mathbb{E}\left(\left|\psi_n(\widetilde{h})-\mathrm{pen}_1(\widetilde{h})\right|-\mathrm{pen}_2(\widetilde{h})\right)\\ &\leq\mathbb{E}\left(\left|\frac{U_n(\widetilde{h},h_{\min})}{n^2}\right|-\theta\frac{c_0(f,g,K)}{n\widetilde{h}}\right)\\ &+\mathbb{E}\left(|V_n(\widetilde{h},h_{\min})|-\frac{\theta}{2}\|(fg)_{h_{\min}}-fg\|^2\right)+\mathbb{E}\left(|V_n(h_{\min},\widetilde{h})|-\frac{\theta}{2}\|(fg)_{\widetilde{h}}-fg\|^2\right)\\ &+\mathbb{E}\left(\theta\|(fg)_{\widetilde{h}}-fg\|^2+(\theta-\kappa)\frac{c_0(f,g,K)}{n\widetilde{h}}\right)+\frac{1}{2}(\theta+\frac{1}{\theta})\|(fg)_{h_{\min}}-fg\|^2+\frac{C}{n}.\end{aligned}$$

where we used Lemma 7.1. Now, using Lemmas 7.2 and 7.3, we get

$$\begin{aligned}
&\mathbb{E}\left(\left|\psi_n(\widetilde{h}) - \mathrm{pen}_1(\widetilde{h})\right| - \mathrm{pen}_2(\widetilde{h})\right) \\
\leq\ & \frac{1}{2}(\theta + \frac{1}{\theta})\|(fg)_{h_{\min}} - fg\|^2 + c_1(f,g,K,\theta)\|f_{h^*} - f\|^2 \\
&+\theta\mathbb{E}\left(\|(fg)_{\widetilde{h}} - fg\|^2 + (1 - \frac{\kappa}{\theta})\frac{c_0(f,g,K)}{n\widetilde{h}}\right) + C\frac{\log(n)}{n}.
\end{aligned}$$

Observe that $\|\widehat{(fg)}_h - fg\|^2 = \|\widehat{(fg)}_h - (fg)_h\|^2 + \|(fg)_h - fg\|^2 + 2V_n(h,h)$. It follows that for all $\theta \in (0, \frac{1}{2})$,

$$\begin{aligned}
&(1-2\theta)\left(\|(fg)_h - fg\|^2 + (1 - \frac{\kappa}{\theta})\frac{c_0(f,g,K)}{nh}\right) - \|\widehat{(fg)}_h - fg\|^2 \\
&= -2\left(V_n(h,h) + \theta\|(fg)_h - fg\|^2\right) + (1-2\theta)(1 - \frac{\kappa}{\theta})\frac{c_0(f,g,K)}{nh} - \|\widehat{(fg)}_h - (fg)_h\|^2 \\
&\leq -2\left(V_n(h,h) + \theta\|(fg)_h - fg\|^2\right) \leq 2(|V_n(h,h)| - \theta\|(fg)_h - fg\|^2)
\end{aligned}$$

provided that $1 - \kappa/\theta \leq 0$. Therefore we choose $\kappa \geq \theta$ and apply Lemma 7.2 again. We obtain

$$\begin{aligned}
\mathbb{E}\left(\left|\psi_n(\widetilde{h}) - \mathrm{pen}_1(\widetilde{h})\right| - \mathrm{pen}_2(\widetilde{h})\right) \leq\ & \frac{1}{2}(\theta + \frac{1}{\theta})\|(fg)_{h_{\min}} - fg\|^2 + c_2(f,g,K)\|f_{h^*} - f\|^2 \\
&+ \frac{\theta}{1-2\theta}\mathbb{E}\left(\|\widehat{(fg)}_{\widetilde{h}} - fg\|^2\right) + C\frac{\log(n)}{n}.
\end{aligned}$$

Similarly, we get

$$\begin{aligned}
\mathbb{E}\left(|\psi_n(h) - \mathrm{pen}_1(h)| + \mathrm{pen}_2(h)\right) \leq\ & \frac{1}{2}(\theta + \frac{1}{\theta})\|(fg)_{h_{\min}} - fg\|^2 + c_2(f,g,K)\|f_{h^*} - f\|^2 \\
&+ \theta\left(\|(fg)_h - fg\|^2 + (1 + \frac{\kappa}{\theta})\frac{c_0(f,g,K)}{nh}\right) + C\frac{\log(n)}{n}.
\end{aligned}$$

Plugging the last two bounds in the expectation of (7.16) implies

$$\begin{aligned}
\left(1 - \frac{2\theta}{1-2\theta}\right)\mathbb{E}\left(\|\widehat{(fg)}_{\widetilde{h}} - fg\|^2\right) \leq\ & \mathbb{E}\left(\|\widehat{(fg)}_h - fg\|^2\right) \\
&+ 2\theta\left(\|(fg)_h - fg\|^2 + (1 + \frac{\kappa}{\theta})\frac{c_0(f,g,K)}{nh}\right) \\
&+ 2(\theta + \frac{1}{\theta})\|(fg)_{h_{\min}} - fg\|^2 + 4c_2(f,g,K)\|f_{h^*} - f\|^2 + C\frac{\log(n)}{n}.
\end{aligned}$$

Now applying Proposition 2.9 with a rougher bound on the variance (the constant is larger), we get for $\theta \in (0, 1/4)$, and $\kappa \geq \theta$,

$$\begin{aligned}
\left(1 - \frac{2\theta}{1-2\theta}\right)\mathbb{E}\left(\|\widehat{(fg)}_{\widetilde{h}} - fg\|^2\right) \leq\ & 2(1+\theta)\|(fg)_h - fg\|^2 + 2(1+\theta+\kappa)\frac{c_0(f,g,K)}{nh} \\
&+ 2(\theta + \frac{1}{\theta})\|(fg)_{h_{\min}} - fg\|^2 + 4c_2(f,g,K)\|f_{h^*} - f\|^2 + C\frac{\log(n)}{n}.
\end{aligned}$$

We conclude that for $\kappa \geq 1/4$ and all $\theta \in (0, 1/4)$,

$$\mathbb{E}\left(\|\widehat{(fg)}_{\widetilde{h}} - fg\|^2\right) \leq 2(1+c_1(\theta)) \inf_{h\in\mathcal{H}_n}\left\{\|(fg)_h - fg\|^2 + (1+\frac{\kappa}{1+\theta})\frac{c_0(f,g,K)}{nh}\right\}$$
$$+c_2(\theta)\|(fg)_{h_{\min}} - fg\|^2 + c_3\|f_{h^*} - f\|^2 + C\frac{\log(n)}{n},$$

where

$$c_1(\theta) = 2\theta(1-\theta)/(1-3\theta) > 0, \quad c_2(\theta) = 2\frac{(1+\theta^2)(1-2\theta)}{\theta(1-2\theta)} > 0,$$

and c_3 and C are positive constants depending on θ, f, g, K. Taking $\theta = 0.2$ we get the result given in Theorem 4.2.

7.4.1. Proof of Lemma 7.1

For the study of $\psi_{2,n}$, note that $|\langle (fg)_h, (fg)_{h'}\rangle| \leq \|fg \star K_h\|\|fg \star K_{h'}\| \leq \|K\|_1^2\|fg\|^2$ and observe that

$$\mathbb{E}(|\widehat{f}_{h^*}(Y_1)|) \leq \sqrt{\mathbb{E}(\widehat{f}^2_{h^*}(Y_1))} \leq \sqrt{\|g\|_\infty(\|K\|^2 + \|K\|_1^2\|f\|^2)}.$$

As a consequence, for all positive h, h',

$$\mathbb{E}\left(\sup_{h,h'\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^n \langle \widehat{f}_{h^*}(Y_i)K_h(Y_i - \cdot), (fg)_{h'}\rangle\right|\right)$$
$$\leq \mathbb{E}\left(|\widehat{f}_{h^*}(Y_1)|\sup_{h,h'}|\langle K_h(Y_1 - \cdot), (fg)_{h'}\rangle|\right)$$
$$\leq \mathbb{E}(|\widehat{f}_{h^*}(Y_1)|)\|K\|_1^2\|fg\|_\infty$$
$$\leq \sqrt{\|g\|_\infty(\|K\|^2 + \|K\|_1^2\|f\|^2)}\|K\|_1^2\|fg\|_\infty.$$

From the definition of $\psi_{2,n}$ given by (7.15), il follows that the result of Lemma 7.1 holds with $C = (2\sqrt{\|g\|_\infty(\|K\|^2 + \|K\|_1^2\|f\|^2)} + \|g\|_\infty)\|fg\|_\infty\|K\|_1^2$.

7.4.2. Proof of Lemma 7.2 and study of $V_n(h, h')$

We decompose $V_n(h, h') = V_{n,1}(h, h') + V_{n,2}(h, h') + V_{n,3}(h, h')$ with

$$\begin{aligned} V_{n,1}(h,h') &= \langle \widehat{(fg)}_h - (\widehat{f}_{h^*}g)_h, (fg)_{h'} - fg\rangle, \\ V_{n,2}(h,h') &= \langle (\widehat{f}_{h^*}g)_h - (f_{h^*}g)_h, (fg)_{h'} - fg\rangle, \\ V_{n,3}(h,h') &= \langle (f_{h^*}g)_h - (fg)_h, (fg)_{h'} - fg\rangle. \end{aligned}$$

We have for all positive θ

$$|V_{n,3}(h,h')| = |\langle (f_{h^*}g)_h - (fg)_h, (fg)_{h'} - fg\rangle| \leq \frac{\vartheta}{2}\|(fg)_{h'} - fg\|^2 + \frac{1}{2\vartheta}\|K\|_1^2\|g\|_\infty^2\|f_{h^*} - f\|^2,$$

implying that

$$\mathbb{E}\left(\sup_{h,h'}\left\{|V_{n,3}(h,h')|-\frac{\vartheta}{2}\|(fg)_{h'}-fg\|^2\right\}\right)\leq\frac{1}{2\vartheta}\|K\|_1^2\|g\|_\infty\|f_{h^*}-f\|^2. \tag{7.17}$$

Next, we write

$$V_{n,2}(h,h')=\frac{1}{n}\sum_{i=1}^n(Z_i-\mathbb{E}(Z_i)),$$

where $Z_i-\mathbb{E}(Z_i):=\langle[(K_{h^*}(X_i-\cdot)-f_{h^*}(\cdot)g(\cdot)]\star K_h,(fg)_{h'}-fg\rangle$ and apply Bernstein Inequality. Using that K is even, the variance bound is obtained by

$$\begin{aligned}
\mathrm{Var}Z_1 &\leq \mathbb{E}(Z_1^2)=\int\langle[(K_{h^*}(u-\cdot)\,g(\cdot)]\star K_h,(fg)_{h'}-fg\rangle^2 f(u)du\\
&= \int\left[K_{h^*}\star(g\,[K_h\star((fg)_{h'}-fg)])(u)\right]^2 f(u)du\\
&\leq \|f\|_\infty\|K_{h^*}\star(gK_h\star((fg)_{h'}-fg))\|^2\\
&\leq \|f\|_\infty\|K_{h^*}\|_1^2\|(gK_h\star((fg)_{h'}-fg))\|^2\\
&\leq \|f\|_\infty\|g\|_\infty^2\|K\|_1^4\|(fg)_{h'}-fg)\|^2:=v_{h,h'}^2.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
|Z_1| &\leq \sup_z\left|\int K_{h^*}(z-x)g(x)((fg)_{h'}-fg)\star K_h)(x)dx\right|\\
&\leq \|g\|_\infty\sup_x|((fg)_{h'}-fg)\star K_h)(x)|\int|K_{h^*}(z)|dz\\
&\leq \|g\|_\infty\|K\|_1\sup_x|((fg)_{h'}-fg)(x)|\int|K_h(u)|du\\
&\leq \|g\|_\infty\|K\|_1^2(1+\|K\|_1)\|fg\|_\infty:=b_{h,h'}.
\end{aligned}$$

Therefore, Bernstein Inequality (8.2) implies that with probability larger than $1-2e^{-\lambda}$

$$\begin{aligned}
|V_{n,2}(h,h')| &\leq \sqrt{\frac{2\lambda v_{h,h'}^2}{n}}+\frac{\lambda}{n}b_{h,h'}\\
&\leq \frac{\theta}{4}\|(fg)_{h'}-fg)\|^2+\frac{\lambda}{n}\|f\|_\infty\|g\|_\infty^2\|K\|_1^2(2\frac{\|K\|_1^2}{\theta}+1+\|K\|_1).
\end{aligned}$$

This together with (8.3) and $|\mathcal{H}_n|=n$ leads to

$$\mathbb{E}\left(\sup_{h,h'}\left\{|V_{n,2}(h,h')|-\frac{\vartheta}{4}\|(fg)_{h'}-fg\|^2\right\}\right)\leq C\frac{\log(n)}{n}. \tag{7.18}$$

For $V_{n,1}$, we write $V_{n,1}(h,h')=V_{n,1}^b(h,h')+V_{n,1}^c(h,h')$ with

$$V_{n,1}^b(h,h')=\frac{1}{n}\sum_{i=1}^n\int\left(\widehat{f}_{h^*}^b(Y_i)K_h(Y_i-x)-(\widehat{f}_{h^*}^bg)\star K_h(x)\right)((fg)_{h'}-fg)(x)dx$$

where

$$\widehat{f}^b_{h^*}(x) = \widehat{f}_{h^*}(x)\mathbf{1}_{|\widehat{f}_{h^*}(x)|\leq d_0} \quad d_0 = 2\|f\|_\infty\|K\|_1.$$

Here, we apply the Bernstein inequality conditionally to $\mathbf{X}$, with $b_{h,h'} = \|f\|_\infty\|fg\|_\infty\|K\|_1^2(1+\|K\|_1)$ and $v^2_{h,h'} = \|f\|^2_\infty\|g\|_\infty\|K\|_1^4\|(fg)_{h'} - fg\|^2$. The orders of $b_{h,h'}$ and $v_{h,h'}$ being the same as for $V_{n,2}$ and independent of $\mathbf{X}$, the result for $V^b_{n,1}$ is :

$$\mathbb{E}\left(\sup_{h,h'}\left\{|V^b_{n,1}(h,h')| - \frac{\vartheta}{4}\|(fg)_{h'} - fg\|^2\right\}\right) \leq C\frac{\log(n)}{n}. \tag{7.19}$$

Lastly, by noticing that $|\widehat{f}_{h^*}(x)| \leq |\widehat{f}_{h^*}(x) - f_{h^*}(x)| + \|f\|_\infty\|K\|_1$, we get

$$\mathbb{P}(|\widehat{f}_{h^*}(x)| > d_0) \leq \mathbb{P}(|\widehat{f}_{h^*}(x) - f_{h^*}(x)| > \|f\|_\infty\|K\|_1) \leq \mathbb{P}(|\widehat{f}_{h^*}(x) - f_{h^*}(x)| > c_0)$$

as $\|f\|_\infty\|K\|_1 \geq \|f\|_\infty = c_0$ by reminding of (7.5). Now using as in (7.6) Bernstein Inequality, we get, under [K3],

$$\mathbb{P}(|\widehat{f}_{h^*}(x)| > d_0) \leq 2n^{-4}. \tag{7.20}$$

Indeed condition [K3] is the translation of ($\mathcal{A}1$) with $d = 1$ and $C = \|K\|^2_\infty/\|K\|^4$. Then as $\|(fg)_h - fg\|_\infty \leq \|fg\|_\infty(\|K\|_1 + 1) \leq 2\|fg\|_\infty\|K\|_1$, we get that

$$\begin{aligned}\mathbb{E}\left(\sup_{h,h'\in\mathcal{H}_n}|V^c_{n,1}(h,h')|\right) &\leq \sum_{h,h'\in\mathcal{H}_n} 4\frac{\|K\|_\infty}{h^*}\|fg\|_\infty\|K\|_1\int\mathbb{P}(|\widehat{f}_{h^*}(x)| > d_0)g(x)dx \\ &\leq \frac{C(f,g,K)|\mathcal{H}_n|^2}{n^4h^*} \leq \frac{C}{n}.\end{aligned}$$

Note that the last bound is obtained using $|\mathcal{H}_n| \leq n$, $L(h^*) \lesssim n$ under assumption [K3]. Gathering this with (7.17), (7.18), (7.19) gives the result of Lemma 7.2.

7.4.3. Proof of Lemma 7.3 and study of $U_n(h,h')$

Warning. For the study of this term, in order to avoid burdensome technicalities, we assume that $\widehat{f}_{h^*}$ is bounded by $2\|K\|_1\|f\|_\infty$. We proved in the study of V_n (see (7.20)) that that the probability of the complement is $1/n^4$ under [K3].

Recall that $U_n(h,h')$ is defined by (7.14). We write that

$$\widehat{f}_{h^*}(Y_i)K_h(Y_i - \cdot) - (fg)_h = \underbrace{\widehat{f}_{h^*}(Y_i)K_h(Y_i - \cdot) - (\widehat{f}_{h^*}g)\star K_h}_{(1)_h}$$
$$+\underbrace{(\widehat{f}_{h^*}g)\star K_h - (f_{h^*}g)\star K_h}_{(2)_h} + \underbrace{(f_{h^*}g)\star K_h - (fg)_h}_{(3)_h}$$

so that $U_n(h,h_{\min})$ can be splitted into 9 terms, denoted with obvious super-indices (k,ℓ) for $k,\ell\in\{1,2,3\}$. These 9 terms can be reduced to 6 by symmetry arguments, denoted by $U_n^{(i),(j)}(h,h_{\min})$ for $i\leq j\in\{1,2,3\}$.

• **Treatment of** $U_n^{(1),(1)}(h,h_{\min})$ we have, by analogy with Lemma 6.2 in Comte and Marie (2021), that, for every $\vartheta\in[0,1]$,

$$\mathbb{E}\left(\sup_{h\in\mathcal{H}_n}\left\{\frac{|U_n^{(1),(1)}|(h,h_{\min})}{n^2}-\frac{\vartheta\|K\|^2\|K\|_1^2\|f\|_\infty^2}{nh}\right\}|\mathbf{X}\right)\leq\frac{C\log(n)}{n} \quad (7.21)$$

and it is easy to see that all bounds do note depend on $\mathbf{X}$ so de-conditioning is straightforward.

• **Treatment of** $U_n^{(3),(3)}(h,h_{\min})$ it is easy to handle thanks to the equality

$$U_n^{(3),(3)}(h,h_{\min})=n(n-1)\langle(f_{h^*}g)\star K_h-(fg)_h,(f_{h^*}g)\star K_{h_{\min}}-(fg)_{h_{\min}}\rangle$$

leading to the bound

$$\begin{aligned}\frac{|U_n^{(3),(3)}(h,h_{\min})}{n^2}|&\leq\|[(f_{h^*}-f)g]\star K_h\|\|[(f_{h^*}-f)g]\star K_{h_{\min}}\|\\&\leq\|K\|_1^2\|g\|_\infty^2\|f_{h^*}-f\|^2. \quad (7.22)\end{aligned}$$

• **Treatment of** $U_n^{(2),(3)}(h,h_{\min})$ first note that $U_n^{(2),(3)}(h,h_{\min})/n^2=[(n-1)/n]\langle(2)_h,(3)_{h_{\min}}\rangle$ where $\langle(2)_h,(3)_{h_{\min}}\rangle$ is equal to

$$\begin{aligned}&\langle[(\widehat{f}_{h^*}g)-(f_{h^*}g)]\star K_h,((f_{h^*}-f)g)\star K_{h_{\min}}\rangle\\&\qquad=\langle[(\widehat{f}_{h^*}g)-(f_{h^*}g)],((f_{h^*}-f)g)\star K_{h_{\min}}\star K_h\rangle\\&\qquad=\langle[(\widehat{f}_{h^*}g)-(f_{h^*}g)],((f_{h^*}-f)g)\star K_h\star K_{h_{\min}}\rangle\\&\qquad=\langle[(\widehat{f}_{h^*}g)-(f_{h^*}g)]\star K_{h_{\min}},((f_{h^*}-f)g)\star K_h\rangle\end{aligned}$$

and thus $\langle(2)_h,(3)_{h_{\min}}\rangle=\langle(2)_{h_{\min}},(3)_h\rangle$, so that $U_n^{(2),(3)}(h,h_{\min})=U_n^{(3),(2)}(h,h_{\min})$. Now, the process can be written as

$$\frac{1}{n}\sum_{i=1}^n(Z_i^{2,3}-\mathbb{E}(Z_i^{2,3})),\quad Z_i^{2,3}:=\langle K_{h^*}(X_i-\cdot)g,[(f_{h^*}-f)g]\star K_h\star K_{h_{\min}}\rangle.$$

To apply Bernstein Inequality, we need to bound the variance and infinite norm of the $Z_i^{2,3}$'s. For the moment of order 2, we have

$$\begin{aligned}\mathbb{E}[(Z_1^{2,3})^2]&=\int f(x)\left(\int K_{h^*}(x-u)g(u)[(f-f_{h^*})g]\star K_h\star K_{h_{\min}}(u)du\right)^2dx\\&\leq\|f\|_\infty\|K_{h^*}\star[g[(f_{h^*}-f)g]\star K_h\star K_{h_{\min}}]\|^2\\&\leq\|f\|_\infty\|K_{h^*}\|_1^2\|[g[(f_{h^*}-f)g]\star K_h\star K_{h_{\min}}]\|^2\\&\leq\|f\|_\infty\|g\|_\infty^2\|K\|_1^6\|f_{h^*}-f\|^2:=\mathfrak{v}.\end{aligned}$$

For the upper bound, it holds:

$$\begin{aligned}
&\sup_x \left| \int K_{h^*}(x-u) g(u) [(f-f_{h^*})g] \star K_h \star K_{h_{\min}}(u) du \right| \\
\leq\ & \|g\|_\infty \sup_u |[(f-f_{h^*})g] \star K_h \star K_{h_{\min}}(u)| \sup_x \int |K_{h^*}(x-u)| du \\
\leq\ & \|g\|_\infty \|K\|_1 \sup_v ||[(f-f_{h^*})g] \star K_h(v)| \sup_u \int |K_{h_{\min}}(u)| du \\
\leq\ & \|g\|_\infty^2 \|K\|_1^3 (\|f\|_\infty + \sup_u |f \star K_{h^*}(u)|) \leq \|g\|_\infty^2 \|K\|_1^3 (1+\|K\|_1) \|f\|_\infty := \mathfrak{b}.
\end{aligned}$$

Then Bernstein Inequality implies that with probability larger than $1-2e^{-\lambda}$, $\lambda > 0$, for any $\vartheta \in (0,1)$,

$$|U_n^{(2),(3)}(h, h_{\min})|/n^2 \leq |\langle (2)_h, (3)_{h_{\min}} \rangle| \leq \sqrt{\frac{2\mathfrak{v}\lambda}{n}} + \frac{\lambda}{n}\mathfrak{b} \leq \vartheta \|f - f_{h^*}\|^2 + C(K,f,g)\frac{\lambda}{\vartheta n}.$$

As a consequence, we obtain

$$\mathbb{P}\left(\sup_{h,h'} \left(\left| \frac{U^{(2),(3)}(h,h')}{n^2} \right| - \vartheta \|f - f_{h^*}\|^2 \right) \geq C(K,f,g)\frac{\lambda}{\vartheta n} \right) \leq 2|\mathcal{H}_n|^2 e^{-\lambda}.$$

Then it follows from (8.3) and $|\mathcal{H}_n| = n$ that

$$\mathbb{E}\left(\sup_{h,h'} \left| \frac{U^{(2),(3)}(h,h')}{n^2} \right| - \vartheta \|f - f_{h^*}\|^2 \right)_+ \leq \frac{C' \log(n)}{n}. \tag{7.23}$$

• **Treatment of** $U_n^{(1),(3)}(h, h_{\min})$ write that $U_n^{(1),(3)}(h, h_{\min})/n^2 = [(n-1)/n]\langle (1)_h, (3)_{h_{\min}} \rangle$ and $\langle (1)_h, (3)_{h_{\min}} \rangle$ is

$$\frac{1}{n}\sum_{i=1}^n \langle \widehat{f}_{h^*}(Y_i) K_h(Y_i - \cdot) - (\widehat{f}_{h^*} g) \star K_h, [(f_{h^*} - f)g] \star K_{h_{\min}} \rangle.$$

We apply Bernstein Inequality conditionally to $\mathbf{X}$, recalling that we consider $\widehat{f}_{h^*}$ bounded by $2\|f\|_\infty \|K\|_1$.

$$\begin{aligned}
&\mathbb{E}\left(\langle \widehat{f}_{h^*}(Y_i) K_h(Y_i - \cdot), [(f_{h^*} - f)g] \star K_{h_{\min}} \rangle^2 \,|\mathbf{X} \right) \\
&= \int \left(\int \widehat{f}_{h^*}(y) K_h(y-u) [(f_{h^*} - f)g] \star K_{h_{\min}}(u) du \right)^2 g(y) dy \\
&\leq 4\|K\|_1^2 \|f\|_\infty^2 \|g\|_\infty \, \| |K_h| \star |(f_{h^*} - f)g| \star |K_{h_{\min}}| \|^2 \leq 4\|K\|_1^6 \|f\|_\infty^2 \|g\|_\infty^3 \|f_{h^*} - f\|^2,
\end{aligned}$$

by iterative application of Young Inequality. Next for the infinite norm

$$\begin{aligned}
&\sup_y |\langle \widehat{f}_{h^*}(y)K_h(y-\cdot), [(f_{h^*}-f)g]\star K_{h_{\min}}\rangle| \\
\leq\ & 2\|f\|_\infty\|K\|_1 \sup_y \int |K_h(y-u)||[(f_{h^*}-f)g]\star K_{h_{\min}}(u)|du \\
\leq\ & 2\|f\|_\infty\|K\|_1 \int |K_h(v)|dv \sup_u |[(f_{h^*}-f)g]\star K_{h_{\min}}(u)| \\
\leq\ & 2\|f\|_\infty\|K\|_1^3 \sup_z |(f_{h^*}-f)(z)g(z)| \leq 2\|f\|_\infty^2\|g\|_\infty\|K\|_1^3(1+\|K\|_1).
\end{aligned}$$

The bounds do not depend on $\mathbf{X}$, it holds:

$$\mathbb{E}\left(\sup_{h,h'}\left|\frac{U^{(1),(3)}(h,h')}{n^2}\right| - \vartheta\|f-f_{h^*}\|^2\right)_+ \leq C\frac{\log(n)}{n} \tag{7.24}$$

for a constant $C >$ depending on f, g, K. Moreover, the bounds do not depend on $h, h_{\min}$ so the same bound hold for $U_n^{(3),(1)}(h,h_{\min})/n^2$.

• **Treatment of** $U_n^{(1),(2)}(h,h_{\min})$ write $U_n^{(1),(2)}(h,h_{\min})/n^2 = [(n-1)/n]\langle(1)_h,(2)_{h_{\min}}\rangle$ where $\langle(1)_h,(2)_{h_{\min}}\rangle$ is

$$\frac{1}{n}\sum_{i=1}^n \langle \widehat{f}_{h^*}(Y_i)K_h(Y_i-\cdot) - (\widehat{f}_{h^*}g)\star K_h, [(\widehat{f}_{h^*}-f_{h^*})g]\star K_{h_{\min}}\rangle.$$

First, we apply a Bernstein Inequality conditionally to $\mathbf{X}$. The variance term is:

$$\mathbb{E}\left(\langle \widehat{f}_{h^*}(Y_i)K_h(Y_i-\cdot), [(\widehat{f}_{h^*}-f_{h^*})g]\star K_{h_{\min}}\rangle^2\right) \leq 4\|f\|_\infty\|g\|_\infty\|K\|_1^4\|K_h\star[(\widehat{f}_{h^*}-f_{h^*})g]\|^2 := \mathfrak{v}.$$

For the upper bound, we get

$$\sup_y |\langle \widehat{f}_{h^*}(y)K_h(y-\cdot), [(\widehat{f}_{h^*}-f_{h^*})g]\star K_{h_{\min}}\rangle| \leq 6\|K\|_1^4\|f\|_\infty^2\|g\|_\infty := \mathfrak{b}$$

with usual tricks. Now, we can notice that

$$\mathbb{E}\left(\|K_h\star[(\widehat{f}_{h^*}-f_{h^*})g]\|^2\right) \leq \frac{\|g\|_\infty\|K\|_1^2\|K\|^2}{nh}. \tag{7.25}$$

So we write

$$\begin{aligned}
&\mathbb{E}\left[\sup_{h\in\mathcal{H}_n}\left(\frac{U_n^{(1),(2)}(h,h_{\min})}{n^2} - 4\vartheta\frac{\|g\|_\infty\|K\|_1^2\|K\|^2}{nh}\right)\right] \\
\leq\ & \mathbb{E}\left\{\mathbb{E}\left[\sup_{h\in\mathcal{H}_n}\left(\frac{U_n^{(1),(2)}(h,h_{\min})}{n^2} - \vartheta\|K_h\star[(\widehat{f}_{h^*}-f_{h^*})g]\|^2\right)|\mathbf{X}\right]\right\} \\
&+\vartheta\mathbb{E}\left[\sup_{h\in\mathcal{H}_n}\left(\|K_h\star[(\widehat{f}_{h^*}-f_{h^*})g]\|^2 - 4\frac{\|g\|_\infty\|K\|_1^2\|K\|^2}{nh}\right)\right]
\end{aligned}$$

The first term is bounded by taking the expectation of the conditional Bernstein, where constants are independent of the X_i, which writes with the terms $\mathfrak{b}, \mathfrak{v}$:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}_n}\left(\frac{U_n^{(1),(2)}(h,h_{\min})}{n^2}-\vartheta\|K_h\star[(\widehat{f}_{h^*}-f_{h^*})g]\|^2\right)\right]\leq C\frac{\log(n)}{n}. \quad (7.26)$$

For the second, we use Talagrand Inequality, relying on the linear process

$$\nu_n(t)=\langle K_h\star[(\widehat{f}_{h^*}-f_{h^*})g],t\rangle$$

which fulfills $\sup_{t\in\mathcal{B}(0,1)}\nu_n^2(t)=\|K_h\star[(\widehat{f}_{h^*}-f_{h^*})g]\|^2$ where $\mathcal{B}(0,1)$ is a countable dense subset of $\{t\in\mathbb{L}^2(\mathbb{R}),\|t\|^2=1\}$. To apply Talagrand inequality, we compute H^2,v,b. We have from (7.25) that

$$H^2=\frac{\|g\|_\infty\|K\|_1^2\|K\|^2}{nh}.$$

Then we compute v^2.

$$\begin{aligned}
&\sup_{\|t\|=1}\mathrm{Var}\left(\iint K_h(x-u)K_{h^*}(X_1-u)t(x)dxdu\right)\\
&\quad\leq\sup_{\|t\|=1}\mathbb{E}\left[\left(\iint K_h(x-u)K_{h^*}(X_1-u)t(x)dxdu\right)^2\right]\\
&\quad=\sup_{\|t\|=1}\mathbb{E}\left[(K_{h^*}\star K_h\star t(X_1))^2\right]\\
&\quad\leq\|f\|_\infty\sup_{\|t\|=1}\|K_{h^*}\star K_h\star t\|^2\leq\|f\|_\infty\|K\|_1^4:=v^2.
\end{aligned}$$

Next, for b, we find

$$\begin{aligned}
\sup_{\|t\|=1}\sup_y|K_{h^*}\star K_h\star t(y)| &\leq \sup_{\|t\|=1}\sup_y\left[\int t^2(x)dx\int(K_h\star K_{h^*}(y-x))^2dx\right]^{1/2}\\
&= \|K_h\star K_{h^*}\|\leq\frac{\|K\|_1\|K\|}{\sqrt{h}}:=b.
\end{aligned}$$

The Talagrand Inequality gives

$$\begin{aligned}
&\mathbb{E}\left[\sup_{h\in\mathcal{H}_n}\left(\|K_h\star[(\widehat{f}_{h^*}-f_{h^*})g]\|^2-4\frac{\|g\|_\infty\|K\|_1^2\|K\|^2}{nh}\right)_+\right]\\
\leq\ &\frac{C}{n}\left(\sum_{h\in\mathcal{H}_n}\exp(-c_1/h)+\mathrm{card}(\mathcal{H}_n)\exp(-C_2\sqrt{n})\right).
\end{aligned}$$

As a consequence, as under [K4] and $h\leq 1$, $\sum_{h\in\mathcal{H}_n}\exp(-c_1/h)\leq\Sigma<+\infty$ and $\mathrm{card}(\mathcal{H}_n)\leq n$, we get

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}_n}\left(\frac{U_n^{(1),(2)}(h,h_{\min})}{n^2}-4\vartheta\frac{\|g\|_\infty\|K\|_1^2\|K\|^2}{nh}\right)\right]\leq C\frac{\log(n)}{n}. \quad (7.27)$$

• **Treatment of** $U_n^{(2),(2)}(h,h_{\min})$ write $U_n^{(2),(2)}(h,h_{\min})/n^2=[(n-1)/n]\langle(2)_h,(2)_{h_{\min}}\rangle$ where $\langle(2)_h,(2)_{h_{\min}}\rangle$ is

$$\langle[(\widehat{f}_{h^*}-f_{h^*})g]\star K_h,[(\widehat{f}_{h^*}-f_{h^*})g]\star K_{h_{\min}}\rangle.$$

The decomposition of this term involves first a U-statistics related to X:

$$\frac{U_n^{\mathbf{X}}(h,h_{\min})}{n^2}:=\frac{1}{n^2}\sum_{1\le i\neq j\le n}\int\quad\left(\int(K_{h^*}(X_i-u)-f\star K_{h^*}(u))g(u)K_h(x-u)du\right)$$
$$\times\left(\int(K_{h^*}(X_j-v)-f\star K_{h^*}(v))g(v)K_{h_{\min}}(x-v)dv\right)dx$$

and terms corresponding to $i=j$ that are studied separately:

$$\frac{1}{n^2}\sum_{i=1}^n\int\left(\int(K_{h^*}(X_i-u)-f\star K_{h^*}(u))g(u)K_h(x-u)du\right)$$
$$\times\left(\int(K_{h^*}(X_i-v)-f\star K_{h^*}(v))g(v)K_{h_{\min}}(x-v)dv\right)dx.$$

First, developing the latter product leads to the study of the four following terms. Two cross-terms that are are bounded by

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^n\left|\int f\star(K_{h^*}g)\star K_h\star K_{h_{\min}}(x)K_{h^*}(X_i-x)g(x)dx\right|\\
\le\ &\frac{\|g\|_\infty}{n}\sup_x|f\star(K_{h^*}g)\star K_h\star K_{h_{\min}}(x)|\int|K_{h^*}(z)|dz\\
\le\ &\frac{\|g\|_\infty\|K\|_1^2}{n}\sup_x|f\star(K_{h^*}g)\star K_h(x)|\le\frac{\|g\|_\infty^2\|K\|_1^3}{n}\sup_x|f\star K_{h^*}(x)|\\
\le\ &\frac{\|f\|_\infty\|g\|_\infty^2\|K\|_1^4}{n}.
\end{aligned}$$

The product of last terms can be written

$$\begin{aligned}
&\frac{1}{n}\left|\int(f\star K_{h^*}g)\star K_h(x)(f\star K_{h^*}g)\star K_{h_{\min}}(x)dx\right|\\
&\quad\le\frac{1}{n}\|(f\star K_{h^*}g)\star K_h\|\|(f\star K_{h^*}g)\star K_{h_{\min}}\|\\
&\quad\le\frac{1}{n}\|K_h\|_1\|K_{h_{\min}}\|_1\|g\|_\infty^2\|f\star K_{h^*}\|^2\\
&\quad\le\frac{\|K\|_1^4\|g\|_\infty^2\|f\|^2}{n}.
\end{aligned}$$

Finally, the product of the first terms is

$$\begin{aligned}
&\left|\frac{1}{n^2}\sum_{i=1}^n\iiint K_{h^*}(X_i-u)K_{h^*}(X_i-v)K_h(x-u)K_{h_{\min}}(x-v)g(u)g(v)dudvdx\right|\\
\leq\ & \frac{\|g\|_\infty^2}{n^2}\sum_{i=1}^n\int |K_{h^*}|\star|K_h|(X_i-x)|K_{h^*}|\star|K_{h_{\min}}|(X_i-x)dx\\
=\ & \frac{\|g\|_\infty^2}{n}\int |K_{h^*}|\star|K_h|(z)|K_{h^*}|\star|K_{h_{\min}}|(z)dz\\
\leq\ & \frac{\|g\|_\infty^2}{n}\sup_z|K_{h^*}|\star|K_h|(z)\int |K_{h^*}|\star|K_{h_{\min}}|(z)dz\\
\leq\ & \frac{\|g\|_\infty^2\|K\|_1^3\|K\|_\infty}{nh},
\end{aligned}$$

implying that, for $H_i(u,v,x):=K_{h^*}(X_i-u)K_{h^*}(X_i-v)K_h(x-u)K_{h_{\min}}(x-v)$

$$\sup_h\left(\left|\frac{1}{n^2}\sum_{i=1}^n\iiint H_i(u,v,x)g(u)g(v)dudvdx\right|-\frac{\|g\|_\infty^2\|K\|_1^3\|K\|_\infty}{nh}\right)\leq 0.$$

Let us deal with the U-statistics $U_n^{\mathbf{X}}(h,h_{\min})$. We follow the line of the proof of Lemma 6.2 in Comte and Marie (2021) and write $U_n^{\mathbf{X}}(h,h_{\min})=\sum_{1\leq i\neq j\leq n}G_{h,h_{\min}}(X_i,X_j)$ where

$$G_{h,h_{\min}}(X_i,X_j)=\langle[(K_{h^*}(X_i-\cdot)-f_{h^*})g]\star K_h,[(K_{h^*}(X_j-\cdot)-f_{h^*})g]\star K_{h_{\min}}\rangle.$$

Indeed, $G_{h,h_{\min}}(X_i,X_j)=G_{h_{\min},h}(X_i,X_j)$ as for all functions u,v it holds

$$\langle u\star K_h,v\star K_{h_{\min}}\rangle=\langle u,v\star K_{h_{\min}}\star K_h\rangle=\langle u,v\star K_{h_{\min}}\star K_h\rangle=\langle u\star K_{h_{\min}},v\star K_h\rangle.$$

We apply the deviation inequality for U-statistics of order 2, as in Lacour *et al.* (2017), see Theorem 3.4 in Houdré and Reynaud-Bouret (2003). Following the notations of the aforementionned papers, we have to compute four bounds $\mathfrak{a}_n,\mathfrak{b}_n,\mathfrak{c}_n,\mathfrak{d}_n$.

◇ First $\mathfrak{a}_n$ is a bound on $\sup_{z,z'}|G_{h,h_{\min}}(z,z')|$.

$$\begin{aligned}
&\sup_{z,z'}|G_{h,h_{\min}}(z,z')|\\
&\leq\sup_{z,z'}\left(\sup_x|[(K_{h^*}(z-\cdot)-f_{h^*})g]\star K_{h_{\min}}(x)|\int|[(K_{h^*}(X_i-\cdot)-f_{h^*})g]\star K_h(x)|dx\right)\\
&\leq\sup_{z,z'}\left(\|K_{h_{\min}}\|_\infty\|(K_{h^*}(z-\cdot)-f_{h^*})g\|_1\int|[(K_{h^*}(X_i-\cdot)-f_{h^*})g](x)|dx\int|K_h(x)|dx\right)\\
&\leq 2\frac{\|K\|_\infty}{h_{\min}}\|g\|_\infty\|K\|_1\times 2\|g\|_\infty\|K\|_1^2=4\frac{\|g\|_\infty^2\|K\|_1^3\|K\|_\infty}{h_{\min}}:=\mathfrak{a}_n.
\end{aligned}$$

Thus

$$\frac{\mathfrak{a}_n\lambda^2}{n^2}\leq 4\lambda^2\frac{\|g\|_\infty^2\|K\|_1^3\|K\|_\infty}{n}.$$

$\diamond$ Next $\mathfrak{b}_n^2$ is a bound on $n\sup_z \mathbb{E}[G^2_{h,h_{\min}}(z,X_1)]$, we write

$$\begin{aligned}
&n\sup_z \mathbb{E}[G^2_{h,h_{\min}}(z,X_1)]\\
&\leq n\sup_z \|[(K_{h^*}(z-\cdot)-f_{h^*})g]\star K_h\|^2\mathbb{E}(\|[(K_{h^*}(X_1-\cdot)-f_{h^*})g]\star K_{h_{\min}}\|^2)\\
&\leq \|K_h\|^2\|K_{h_{\min}}\|^2\sup_z \|[(K_{h^*}(z-\cdot)-f_{h^*})g]\|_1^2\mathbb{E}(\|[(K_{h^*}(X_1-\cdot)-f_{h^*})g]\|_1^2)\\
&\leq 4n\frac{\|K\|^4\|K\|_1^4\|g\|_\infty^2}{hh_{\min}} := \mathfrak{b}_n^2.
\end{aligned}$$

We obtain

$$\frac{\mathfrak{b}_n\lambda^{3/2}}{n^2} \leq 2\lambda^{3/2}\frac{\|K\|^2\|K\|_1^2\|g\|_\infty}{\sqrt{hh_{\min}}n^{3/2}} \leq \theta\frac{\|K\|^2\|K\|_1^2\|g\|_\infty^2}{nh} + \frac{\lambda^3}{\theta}\frac{\|K\|^2\|K\|_1^2}{n^2h_{\min}}.$$

$\diamond$ We compute $\mathfrak{c}_n^2$ which is a bound on $n^2\mathbb{E}\left[G^2_{h,h_{\min}}(X_1,X_2)\right]$. Decompose

$$\mathbb{E}\left[G^2_{h,h_{\min}}(X_1,X_2)\right] = \mathbb{E}\left[\langle[(K_{h^*}(X_1\cdot)-f_{h^*})g]\star K_h, [(K_{h^*}(X_2-\cdot)-f_{h^*})g]\star K_{h_{\min}}\rangle^2\right]$$

into four squared terms. First,

$$\langle f_{h^*}g\star K_h, f_{h^*}g\star K_{h_{\min}}\rangle^2 \leq \|K\|_1^4\|f_{h^*}g\|^4 \leq \|g\|_\infty^4\|K\|_1^8\|f\|^4 \leq \|g\|_\infty^4\|K\|_1^8\|f\|_\infty^2.$$

Second

$$\begin{aligned}
&\langle f_{h^*}g\star K_h, (K_{h^*}(X_2-\cdot)g)\star K_{h_{\min}}\rangle^2\\
&\qquad\leq \left\{\sup_z|f_{h^*}g\star K_h(z)|\int|(K_{h^*}(X_2-\cdot)g)\star K_{h_{\min}}(z)|dz\right\}^2\\
&\qquad\leq \left\{\|K\|_1\sup_z|(f_{h^*}g)(z)|\|K\|_1\int|K_{h^*}(X_1-u)g(u)|du\right\}^2\\
&\qquad\leq \left\{\|K\|_1^2\|f\|_\infty\|g\|_\infty \times \|g\|_\infty\|K\|_1^2\right\}^2 = (\|f\|_\infty\|g\|_\infty^2\|K\|_1^4)^2.
\end{aligned}$$

The twin term in $h_{\min}, h$ has clearly the same bound. Lastly

$$\begin{aligned}
&\mathbb{E}\left[\langle(K_{h^*}(X_1-\cdot)g)\star K_h, (K_{h^*}(X_2-\cdot)g)\star K_{h_{\min}}\rangle^2\right]\\
=\;& \iint\left(\int K_{h^*}(u-\cdot)g\star K_h(x)K_{h^*}(v-\cdot)g\star K_{h\min}(x)dx\right)^2 f(u)f(v)dudv\\
\leq\;& \|g\|_\infty^4\iint\left(\int|K_{h^*}|\star|K_h|(u-x)|K_{h^*}|\star|K_{h\min}|(v-x)dx\right)^2 f(u)f(v)dudv\\
=\;& \|g\|_\infty^4\iint\left[|K_{h^*}|\star|K_h|\star|K_{h^*}|\star|K_{h\min}|(u-v)\right]^2 f(u)f(v)dudv\\
\leq\;& \|g\|_\infty^4\|f\|_\infty^2\||K_{h^*}|\star|K_h|\star|K_{h^*}|\star|K_{h\min}|\|^2\\
\leq\;& \|g\|_\infty^4\|f\|_\infty^2\|K\|_1^6\frac{\|K\|^2}{h}.
\end{aligned}$$

We get

$$\mathfrak{c}_n^2 = \frac{n^2}{h}\|g\|_\infty^4\|f\|_\infty^2\|K\|_1^6\left(\|K\|^2+3\|K\|_1^2\right).$$

Thus , for all positive θ, λ it holds

$$\frac{\mathfrak{c}_n\sqrt{\lambda}}{n^2} \leq \theta\frac{\|g\|_\infty^4\|f\|_\infty^2\|K\|_1^6}{nh} + \frac{\lambda(\|K\|^2+3\|K\|_1^2)}{4n\,\theta}.$$

◇ Lastly, the term $\mathfrak{d}_n$ is a bound on

$$\sup_{a,b}\sum_{1\leq i\neq j\leq n}\mathbb{E}\left[G_{h,h_{\min}}(X_i,X_j)a_i(X_i)b_j(X_j)\right],$$

where $a_k(\cdot)$, $b_k(\cdot)$ for $k=1,\dots,n$ is such that $\mathbb{E}(\sum_{k=1}^n a_k^2(X_k)) \leq 1$ and $\mathbb{E}(\sum_{k=1}^n b_k^2(X_k)) \leq 1$. Using the independence for $i\neq j$ between functions of X_i and functions of X_j, we get that the term inside the sup is less than

$$\langle\sum_{i=1}^n\mathbb{E}\left(H_i\star|K_h||a_i(X_i)|\right),\sum_{i=1}^n\mathbb{E}\left(H_j\star|K_{h_{\min}}||b_j(X_j)|\right)\rangle, \tag{7.28}$$

where $H_i = |K_{h^*}(X_i-\cdot)g - f_{h^*}g|$. First we have

$$\begin{aligned}
&\sum_{i=1}^n\mathbb{E}\left(|K_{h^*}(X_i-\cdot)g-f_{h^*}g|\star|K_h||a_i(X_i)|\right)\\
&\qquad\leq\sqrt{n}\left\{\sum_{i=1}^n\left[\mathbb{E}\left(|K_{h^*}(X_i-\cdot)g-f_{h^*}g|\star|K_h||a_i(X_i)|\right)\right]^2\right\}^{1/2}\\
&\qquad\leq\sqrt{n}\left\{\sum_{i=1}^n\mathbb{E}\left[\left(|K_{h^*}(X_i-\cdot)g-f_{h^*}g|\star|K_h|\right)^2\right]\mathbb{E}(a_i^2(X_i))\right\}^{1/2}
\end{aligned}$$

As

$$\mathbb{E}\left[\left(|K_{h^*}(X_i-\cdot)g-f_{h^*}g|\star|K_h|\right)^2\right] \leq \|f\|_\infty\|g\|_\infty^2[\||K_{h^*}|\star|K_h|\|^2+\|f_{h^*}\star|K_h|\|^2]$$

we get

$$\sum_{i=1}^n\mathbb{E}\left(|K_{h^*}(X_i-\cdot)g-f_{h^*}g|\star|K_h||a_i(X_i)|\right) \leq \sqrt{2n\|f\|_\infty}\|g\|_\infty\|K\|_1\|K_h\|.$$

Plugging this in formula (7.28), we get

$$\begin{aligned}
&\langle\sum_{i=1}^n\mathbb{E}\left(H_i\star|K_h||a_i(X_i)|\right),\sum_{i=1}^n\mathbb{E}\left(H_j\star|K_{h_{\min}}||b_j(X_j)|\right)\rangle\\
&\leq\sqrt{2n\|f\|_\infty}\|g\|_\infty\|K\|_1\|K_h\|\sum_{j=1}^n\mathbb{E}\left(\int|K_{h^*}(X_j-\cdot)g-f_{h^*}g|\star|K_{h_{\min}}(u)|du\,|b_j(X_j)|\right)\\
&\leq\sqrt{2n\|f\|_\infty}\|g\|_\infty\|K\|_1\|K_h\|\times 2\|K\|_1^2\|g\|_\infty\left(\sum_{j=1}^n\mathbb{E}(b_j(X_j)|)\right)\\
&\leq 2\sqrt{2n\|f\|_\infty}\|g\|_\infty^2\|K\|_1^2\|K_h\|\times\sqrt{n}.
\end{aligned}$$

Therefore,

$$\mathfrak{d}_n := 2\sqrt{2\|f\|_\infty}\|g\|_\infty^2\|K\|_1^2\|K\|\frac{n}{\sqrt{h}}.$$

It follows that

$$\frac{\mathfrak{d}_n\lambda}{n^2} \leq \theta\frac{\|f\|_\infty^2\|g\|_\infty^4\|K\|_1^4\|K\|^2}{nh} + 2\frac{\lambda^2}{n}.$$

Applying the deviation inequality for U-statistics of order 2 (see Lacour *et al.* (2017) and Theorem 3.4 in Houdré and Reynaud-Bouret (2003)) leads thus to

$$\mathbb{E}\left\{\sup_{h\in\mathcal{H}_n}\left(\frac{U_n^{\mathbf{X}}(h,h_{\min})}{n^2} - \theta\frac{\|f\|_\infty^2\|g\|_\infty^4\|K\|_1^4\|K\|^2}{nh}\right)\right\} \leq C\frac{\log(n)}{n}.$$

Therefore

$$\mathbb{E}\left\{\sup_{h\in\mathcal{H}_n}\left(\frac{U_n^{(2),(2)}(h,h_{\min})}{n^2} - \theta\frac{\|f\|_\infty^2\|g\|_\infty^4\|K\|_1^4\|K\|^2}{nh}\right)\right\} \leq C\frac{\log(n)}{n}. \quad (7.29)$$

The result of Lemma 7.3 follows by gathering the bounds (7.21), (7.26), (7.27), (7.29), (7.24), (7.23), (7.22).

8. Appendix

In the paper we make an extensive use of the following:

- The Young Inequality: for $u \in L^p(\mathbb{R}^d)$ and $v \in L^q(\mathbb{R}^d)$, $1 \leq p, q \leq r \leq \infty$,

$$\|u \star v\|_r \leq \|u\|_p\|v\|_q, \quad \frac{1}{r} + 1 = \frac{1}{p} + \frac{1}{q}, \quad (8.1)$$

- The Bernstein inequality: For i.i.d. random variables Z_i, set $S_n = \frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])$. If $\mathbb{E}[Z_1^2] \leq \mathfrak{v}$ and $|Z_1| \leq \mathfrak{b}$ a.s. then with probability larger than $1 - 2e^{-\lambda}$, for any $\lambda > 0$,

$$|S_n| \leq \sqrt{\frac{2\mathfrak{v}\lambda}{n}} + \frac{\lambda}{n}\mathfrak{b}. \quad (8.2)$$

- Deriving a bound in expectation from a bound on probability: If $\mathbb{P}(Z \geq \frac{\kappa\lambda}{n}) \leq n^2e^{-\lambda}$ for all $\lambda > 0$, then it holds

$$\mathbb{E}[Z_+] \leq c\frac{\log(n)}{n}. \quad (8.3)$$

  Indeed, for all positive A we have

$$\mathbb{E}[Z_+] = \int_0^\infty \mathbb{P}(Z \geq x)dx = \frac{\kappa}{n}\int_0^\infty \mathbb{P}(Z \geq \frac{\kappa\lambda}{n})d\lambda \leq \frac{\kappa}{n}\left(A + 2n^2e^{-A}\right),$$

  and $A = 2\log(n)$ gives the result.

The Talagrand inequality. The result below follows from the Talagrand concentration inequality given in Klein and Rio (2005) and arguments in Birgé and Massart (1998) (see the proof of their Corollary 2 page 354).

Lemma 8.1. *(Talagrand Inequality) Let $Y_1, \dots, Y_n$ be independent random variables and let $\mathcal{F}$ be a countable class of uniformly bounded measurable functions. Consider ν_n, the centered empirical process defined by*

$$\nu_n(f) = \frac{1}{n}\sum_{i=1}^{n}[f(Y_i) - \mathbb{E}(f(Y_i))]$$

for $f \in \mathcal{F}$. Assume there exists three positive constants M, H and v such that

$$\sup_{f\in\mathcal{F}} \|f\|_\infty \le b, \quad \mathbb{E}\Big[\sup_{f\in\mathcal{F}} |\nu_n(f)|\Big] \le H, \quad \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{k=1}^{n} \mathrm{Var}(f(Y_k)) \le v^2.$$

Then, for any $\delta > 0$ the following holds

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} |\nu_n(f)|^2 - 2(1+2\delta)H^2\Big]_+ \le \frac{4}{K_1}\left(\frac{v^2}{n}\exp\left(-K_1\delta\frac{nH^2}{v^2}\right) + \frac{49b^2}{K_1 n^2 C^2(\delta)}\exp\left(-\frac{K_1 C(\delta)\sqrt{2\delta}}{7}\frac{nH}{b}\right)\right),$$

with $C(\delta) = \sqrt{1+\delta} - 1$ and $K_1 = 1/6$.

By standard density arguments, this result can be extended to the case where $\mathcal{F}$ is a unit ball of a linear normed space, after checking that $f \mapsto \nu_n(f)$ is continuous and $\mathcal{F}$ contains a countable dense family.

References

[1] Abramowitz, M. and Stegun, I. A. *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., 1964.

[2] Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta numerica*, **30**, 87-201, 2021.

[3] Belomestny, D., Comte, F. and Genon-Catalot, V. Sobolev-Hermite versus Sobolev nonparametric density estimation on $\mathbb{R}$. *Ann. Inst. Statist. Math.* **71**, 1, 29-62, 2019.

[4] Bongioanni, B. and Torrea, J. L. Sobolev spaces associated to the harmonic oscillator. *Proc. Indian Acad. Sci. Math. Sci.*, 116 no. 3:337–360, 2006.

[5] Butucea, C., Delmas, J.-F., Dutfoy, A. and Fischer, R. Fast adaptive estimation of log-additive exponential models in Kullback-Leibler divergence. *Electron. J. Stat.* **12** , no. 1, 1256-1298, 2018.

[6] Chinot, G and Lerasle, M. On the robustness of the minimum $\ell 2$ interpolator. To appear in *Bernoulli* arXiv:2003.05838, 2020.

[7] Cohen, A., Davenport, M. A. and Leviatan, D. On the stability and accuracy of least squares approximations. *Found. Comput. Math.* 13 , no. 5, 819-834, 2013.

[8] Comte, F. *Nonparametric estimation.* Spartacus-Idh. 128 p., Paris, 2017.
[9] Comte, F. and Duval, C. Should we estimate a product of density functions by a product of estimators? *Preprint hal-03602694* version 1, 2022, https://hal.archives-ouvertes.fr/hal-03602694v1.
[10] Comte, F. and Lacour, C. (2021). Noncompact estimation of the conditional density from direct or noisy data, Preprint Hal-03276251, To appear in *Ann. Inst. Henri Poincaré* .
[11] Comte, F. and Marie, N. On a Nadaraya-Watson estimator with two bandwidths. *Electron. J. Stat.* **15** no. 1, 2566-2607, 2021.
[12] Comte, F. and Rebafka, T. Adaptive density estimation in the pile-up model involving measurement errors. *Electronic journal of Statistics* **6**, 2002-2037, 2012.
[13] Goldenshluger, A. and Lepski, O. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** no. 3, 1608-1632, 2011.
[14] Goldenshluger, A., and Lepski, O. On adaptive minimax density estimation on R^d. *Probability Theory and Related Fields*, **159** no. 3, 479-543, 2014.
[15] Györfi, L., Kohler, M. , Krzyzak, A. and Walk, H. . *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.
[16] Houdré, C and Reynaud-Bouret, P. Exponential Inequalities, with Constants, for U-Statistics of Order Two. *Stochastic Inequalities and Applications,* Progr. Probab. 56, Birkhäuser, Basel, 55-69, 2003.
[17] Lacour, C. Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincar Probab. Statist.* **43**, no. 5, 571-597, 2007.
[18] Lacour, C., Massart, P. and Rivoirard, V. Estimator selection: a new method with applications to kernel density estimation. *Sankhya* **A**, *The Indian Journal of Statistics*, **79** no. 2,298-335, 2017.
[19] Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28**, no. 5, 1302-1338, 2000.
[20] Lerasle, M. Optimal model selection in density estimation, *Ann. Inst. Henri Poincaré,* **48**, 3 884–908, 2012.
[21] Lerasle, M., Magalhães, N. M. and Reynaud-Bouret, P. Optimal kernel selection for density estimation. *High dimensional probabilities,* VII: The Cargese Volume:425460, 2016.
[22] Massart, P. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
[23] Tsybakov, A. B. Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. Springer Series in Statistics. Springer, New York, 2009.