Adaptive greedy algorithm for moderately large dimensions in kernel conditional density estimation

Minh-Lien Jeanne Nguyen, Claire Lacour, Vincent Rivoirard

Adaptive greedy algorithm for moderately large dimensions in kernel conditional density estimation

Minh-Lien Jeanne Nguyen,
Mathematical Institute,
University of Leiden
Niels Bohrweg 1, 2333 CA Leiden, Netherlands
m.j.nguyen@math.leidenuniv.nl

Claire Lacour,
LAMA, CNRS
Univ Gustave Eiffel, Univ Paris Est Creteil
F-77447 Marne-la-Vallée, France
claire.lacour@univ-eiffel.fr

Vincent Rivoirard
CEREMADE, CNRS, UMR 7534
Université Paris-Dauphine, PSL University
75016 Paris, France
Vincent.Rivoirard@dauphine.fr

June 28, 2021

Abstract

This paper studies the estimation of the conditional density $f(x,\cdot)$ of Y_i given $X_i = x$, from the observation of an i.i.d. sample $(X_i, Y_i) \in \mathbb{R}^d$, $i \in \{1,\dots,n\}$. We assume that f depends only on r unknown components with typically $r \ll d$. We provide an adaptive fully-nonparametric strategy based on kernel rules to estimate f. To select the bandwidth of our kernel rule, we propose a new fast iterative algorithm inspired by the Rodeo algorithm (Wasserman and Lafferty, 2006) to detect the sparsity structure of f. More precisely, in the minimax setting, our pointwise estimator, which is adaptive to both the regularity and the sparsity, achieves the quasi-optimal rate of convergence. Our results also hold for density estimation. The computational complexity of our method is only $O(dn\log n)$. A deep numerical study shows nice performances of our approach.

Keywords: *Conditional density, Sparsity, Minimax rates, Kernel density estimators, Greedy algorithm.*

1 Introduction

1.1 Motivations

Consider $W = (W_1, \dots, W_n)$ a sample of a couple (X, Y) of multivariate random vectors: for $i \in \{1, \dots, n\}$,

$$W_i = (X_i, Y_i),$$

with X_i valued in $\mathbb{R}^{d_1}$ and Y_i in $\mathbb{R}^{d_2}$. We denote $d := d_1 + d_2$ the joint dimension. We assume that the marginal distribution of X and the conditional distribution of Y given X are absolutely continuous with respect to the Lebesgue measure, and we denote by f_X the marginal density of X (and more generally by f_Z the density of any random vector Z). Let us define $f : \mathbb{R}^d \to \mathbb{R}_+$ such that for any $x \in \mathbb{R}^{d_1}$, $f(x, \cdot)$ is the conditional density of Y conditionally on $X = x$:

$$f(x, y)dy = d\mathbb{P}_{Y|X=x}(y).$$

In this paper, we aim at estimating the conditional density f at a set point $w = (x, y)$ in $\mathbb{R}^d$.

Estimating a conditional density may be done in any regression framework, i.e. as soon as we observe a (possibly multidimensional) response Y associated with a (possibly multidimensional) covariate X. The regression function $\mathbb{E}[Y|X = x]$ is often studied, but this mean is in fact a summary of the entire distribution and may lose information (think in particular to the case of an asymmetric or multimodal distribution). Thus the problem of estimating the conditional distribution is considered in various application fields: meteorology, insurance, medical studies, geology, astronomy. See Nguyen (2018) and references therein. Moreover, the ABC methods (Approximate Bayesian Computation) are actually dedicated to find a conditional distribution (of the parameter given observations) in the case where the likelihood is not computable but simulable: see Izbicki et al. (2018) (and references therein) where the link between conditional density estimation and ABC is studied.

Several nonparametric methods have been proposed for estimating a conditional density: Hyndman et al. (1996) and Fan et al. (1996) have improved the seminal Nadaraya-Watson-type estimator of Rosenblatt (1969) and Lincheng and Zhijun (1985), as well as De Gooijer and Zerom (2003) who introduced another weighted kernel estimator. For these kernel estimators, different methods have been advocated to tackle the bandwidth selection issue: bootstrap approach (Bashtannyk and Hyndman, 2001) or cross-validation variants, see Fan and Yim (2004); Holmes et al. (2010), Ichimura and Fukuda (2010). Later, adaptive-in-smoothness estimators have been introduced: Brunel et al. (2007) with piecewise polynomial representation, Chagny (2013) with wraped base method, Le Pennec and Cohen (2013) with penalized maximum likelihood estimator, Bertin et al. (2016) with Lepski-type methods and Sart (2017) with tests-based histograms.

All above references do not really deal with the curse of dimensionality. From a theoretical point of view, the minimax rate of convergence for such nonparametric statistical problems is known to be $n^{-s/(2s+d)}$ (possibly up to a logarithmic term), where s is the smoothness of the target function. This illustrates that estimation gets increasingly hard when d is large. Moreover the computational complexity of above methods is often intractable as soon as d is larger than 3 or 4. A first answer to overcome this limitation is to consider single-index models as Bouaziz and Lopez (2010) or semi-parametric models as Fan et al. (2009), but this implies a strong structural assumption. A more general advance has been made by Hall et al. (2004) who assume that some components of X can be irrelevant, i.e. that they do not contain

any information about Y and should be dropped before conducting inference. Their cross-validation approach allows them to obtain a minimax rate for a r_1-dimensional C^2 function, where r_1 is the number of relevant X-components. Efromovich (2010) has improved these non-adaptive results by using thresholding and Fourier series and achieves the minimax rate $n^{-s/(2s+r_1)}$ without any knowledge of r_1 nor s. Note that above rates were established for the $\mathbb{L}^2$-loss whereas we shall consider the pointwise loss. Moreover these combinatorial approaches make their computation cost prohibitive when both n and d are large. In the same framework, Shiga et al. (2015) assume that the dependence of Y on the relevant components is additive. Another way is paved by Otneim and Tjøstheim (2018) who estimate the dependence structure in a Gaussian parametric way while estimating marginal distributions nonparametrically. More recently, Izbicki and Lee (2016, 2017) have proposed two attractive methodologies using orthogonal series estimators in the context of an eventual smaller unknown intrinsic dimension of the support of the conditional density. In particular, the Flexcode method originally proposes to transfer successful procedures for high dimensional regression to the conditional density estimation setting by interpreting the coefficients of the orthogonal series estimator as regression functions, which allows to adapt to data with different features (mixed data, smaller intrinsic dimension, relevant variables) in function of the regression method. However, the optimal tuning parameters depend in fact on the unknown intrinsic dimension. Furthermore, optimal minimax rates are not achieved, revealing the specific nature of the problem of conditional density estimation, more intricate, in full generality, than regression.

1.2 Objectives, methodology and contributions

In this paper, we wish to estimate the conditional density f by assuming that only $r \in \{0, \dots, d\}$ components are *relevant*, i.e. that there exists a subset $\mathcal{R} \subset \{1, \dots, d\}$ with cardinal r, such that for any fixed $\{z_j\}_{j\in\mathcal{R}}$, the function $\{z_k\}_{k\in\mathcal{R}^c} \mapsto f(z_1, \dots, z_d)$ is constant on the neighborhood of w, with $\mathcal{R}^c = \{1, \dots, d\} \setminus \mathcal{R}$. We denote $f_{\mathcal{R}}$ the restriction of f to the relevant directions. Assuming that f is s-Hölderian, our goal is to provide an estimation procedure such that it achieves the best adaptive rate. The meaning of *adaptation* is *twofold* in this paper; the first meaning corresponds to adaptation with respect to the smoothness, which is the classical meaning of adaptation. The second one corresponds to adaptation with respect to the sparsity. So, our goal is to propose an optimal procedure in this context, meaning that it does not depend on the knowledge of s and r, and even $\mathcal{R}$. Furthermore, for practical purposes in moderate large dimensions, it should be implemented with low computational time.

For this purpose, we consider a particular kernel estimator depending on a bandwidth $h \in \mathbb{R}_+^d$ to be selected. To circumvent the curse of dimensionality, we consider an iterative algorithm on a special path of bandwidths inspired by the RODEO procedures proposed by Wasserman and Lafferty (2006) and Lafferty and Wasserman (2008) for nonparametric regression, Liu et al. (2007) for density estimation and Nguyen (2018) for conditional density estimation. More precisely, our new procedure, called RevDir CDRODEO, is a variation of the CDRODEO proposed by Nguyen (2018) (and called Direct CDRODEO in the sequel). Each iteration step of this new algorithm is based on comparisons between partial derivatives of our kernel rule, denoted Z_{hj}, and specific thresholds λ_{hj}, respectively defined in (2.7) and (2.10). Let us mention that for variable selection in the regression model with very high ambient dimension, Comminges and Dalalyan (2012) used similar ideas to select the relevant variables by comparing some quadratic functionals of empirical Fourier coefficients to prescribed sig-

nificance levels. Consistency of this (non-greedy) procedure is established by Comminges and Dalalyan (2012).

We establish that, up to a logarithmic term whose exponent is positive but as close to 0 as desired, RevDir CDRodeo achieves the rate $((\log n)/n)^{s/(2s+r)}$, which is the optimal adaptive minimax rate on Hölder balls $\mathcal{H}_d(s, L)$, when the conditional density depends on r components. When r is much smaller than d, this rate is much faster than the usual rate $((\log n)/n)^{s/(2s+d)}$ achieved by classical kernel rules. Furthermore, unlike previous Rodeo-type procedures, our procedure is adaptive with respect to both the smoothness and the sparsity. To the best of our knowledge, our RevDir CDRodeo procedure is the first algorithm achieving quasi-minimax rates for conditional density estimation in this setting where both sparsity and smoothness are unknown. We lead a deep numerical study of parameters tuning of the algorithm. Then the numerical performances are presented for several examples of conditional densities. In particular RevDir CDRodeo is able to tackle the issue of sparsity detection. Moreover, for each relevant component, reconstructions are satisfying. Finally, we show that the total worst-case complexity of the RevDir CDRodeo algorithm is only $O(dn \log n)$. This last result is very important for modern statistics where many problems deal with very large datasets.

1.3 Plan of the paper and notation

The plan of the paper is the following. First we describe in Section 2 the estimation procedure. We give heuristic ideas based on the minimax approach and explain why some modifications of the Direct CDRodeo procedure are necessary. Then a detailed presentation of our algorithm is provided in Section 2.2.3. Next, the main result is stated in Section 3. The complexity of the algorithm is computed in Section 3.4. After tuning the method, the latter is illustrated via simulations and examples in Section 4. The proofs are gathered in Section 5.

In the sequel, we adopt the following notation. Given two functions $\phi, \psi : \mathbb{R}^d \to \mathbb{R}$, two integers j, k, two vectors h and h', two real numbers a and b, we denote

- $\|\phi\|_q = \left(\int |\phi(u)|^q du\right)^{1/q}$ the $\mathbb{L}_q$ norm of ϕ for any $q \geq 1$;
- $\phi \star \psi$ the convolution product $u \mapsto \int_{\mathbb{R}^d} \phi(u - v)\psi(v)dv$;
- $\partial_j \phi$ the partial derivative of ϕ with respect to the direction j (or $\frac{\partial}{\partial u_j}\phi$ when there is ambiguity on the variable);
- $j : k$ the set of integers from j to k;
- $|A|$ the cardinal of a set A,
- $h \preceq h'$ the partial order on vectors defined by: $h_k \leq h'_k$, for $k \in 1 : d$.
- $a \lesssim b$ (respectively $a \approx b$) means that the inequality (respectively the equality) is satisfied up to a constant.

2 Estimation procedure

As mentioned in Introduction, the goal of this paper is to provide an estimator of the conditional density achieving pointwise adaptive minimax rates, where the meaning of adaptation is twofold as explained in Section 1.2.

Our estimation procedure follows the kernel methodology. We use a specific family of kernel estimators (Bertin et al., 2016), called hereafter the BLR estimators and detailed in Section 2.1. The selection of the bandwidth is introduced with heuristic considerations and detailed in Section 2.2 in the spirit of RODEO (Lafferty and Wasserman, 2008; Nguyen, 2018). After presenting advantages and limitations of the latter, we propose a new algorithm called RevDir CDRODEO.

2.1 Kernel rule

We use the BLR family of kernel estimators as it presents some significant advantages explained below. The BLR family is defined as follows. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, namely K satisfies $\int_{\mathbb{R}} K(t)dt = 1$. Then, given a bandwidth $h = (h_1, \dots, h_d) \in (0, 1]^d$, the estimator of $f(w)$ associated with K and h is defined by

$$\hat{f}_h(w) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\tilde{\mathrm{f}}_X(X_i)} \mathrm{K}_h(w - W_i), \tag{2.1}$$

where for any $v \in \mathbb{R}^d$,

$$\mathrm{K}_h(v) = \prod_{j=1}^{d} h_j^{-1} K(v_j/h_j)$$

and $\tilde{\mathrm{f}}_X$ is an estimator of f_X, built from a sample $\tilde{X}$ not necessarily independent of W.

Remark 1. *Note that (non conditional) density estimation is a special case of this problem studied, as it corresponds to the setting where $d_1 = 0$ and $\mathrm{f}_X \equiv 1$ ($\equiv \tilde{\mathrm{f}}_X$). In this case, $\hat{f}_h(w)$ is the usual kernel density estimator.*

Several arguments justify the choice of the BLR family, rather than the intensively studied family expressed as a ratio of two density estimates of f_W and f_X, following:

$$f(x, y) = \frac{\mathrm{f}_W(x, y)}{\mathrm{f}_X(x)}.$$

Indeed, this last decomposition takes into account the characteristics (smoothness, sparsity) of f_W and f_X instead of those of our target f. More precisely, an irrelevant component of the conditional density may be relevant for both the joint density f_W and the marginal density f_X and it occurs in particular when a component of X is independent of Y. Similarly, the smoothness of f can be different from those of the functions f_W and f_X, which potentially would deteriorate the rates of convergence.

Conversely, the BLR estimators estimate f more directly: in particular, their expectations can be written as the usual kernel regularization of f: under some mild assumptions on K and f and with $\tilde{\mathrm{f}}_X = \mathrm{f}_X$,

$$\mathbb{E}[\hat{f}_h(w)] = \iint \frac{1}{\mathrm{f}_X(u)} \mathrm{K}_h(w - (u, v)) \mathrm{f}_W(u, v) du dv = \int \mathrm{K}_h(w - z) f(z) dz = (\mathrm{K}_h \star f)(w). \tag{2.2}$$

2.2 Selection of the bandwidth

The principal issue in kernel rules is the choice of the bandwidth. In particular, we consider a d-dimensional bandwidth, instead of a scalar one which would be easier and faster to select but would also deteriorate the performances of the estimator.

2.2.1 Heuristic minimax arguments

We consider $\mathcal{H}_{d|r}(s, L)$ the set of functions of $\mathcal{H}_d(s, L)$ with at most r relevant components, and its associated (squared) pointwise minimax risk

$$\inf_{\hat{T}_n} \sup_{f \in \mathcal{H}_{d|r}(s,L)} \mathbb{E}[(\hat{T}_n(w) - f(w))^2],$$

where the infimum is taken over all estimators of f built from the sample W.

In the case of kernel rules, let us denote h^* the *minimax bandwidth* minimizing this risk. We can decompose the squared risk in bias and variance terms:

$$R(h^*) := \mathbb{E}[(\hat{f}_{h^*}(w) - f(w))^2] = B^2(h^*) + \mathrm{Var}(\hat{f}_{h^*}(w)). \tag{2.3}$$

For any bandwidth $h \in (0, 1]^d$, the usual respective upper bounds for the bias and variance are typically

$$B^2(h) := \Big(\mathbb{E}[\hat{f}_h(w)] - f(w)\Big)^2 \lesssim \sum_{j \in \mathcal{R}} h_j^{2s} \tag{2.4}$$

and

$$\mathrm{V}_h := \mathrm{Var}(\hat{f}_h(w)) \lesssim \frac{1}{n \prod_{j=1}^d h_j}. \tag{2.5}$$

The minimizer h^* on $(0, 1]^d$ of the minimax risk is then of the form:

$$h_j^* = \begin{cases} n^{-1/(2s+r)} \text{ for } j \in \mathcal{R}, \\ 1 \text{ for } j \notin \mathcal{R}. \end{cases} \tag{2.6}$$

Given this bandwidth, which depends on s, r and $\mathcal{R}$, and given a sharp estimator $\tilde{\mathrm{f}}_X$, the Blr estimator achieves the minimax rates $n^{-\frac{s}{2s+r}}$. In the literature of conditional density estimation, to the best of our knowledge, no method provides theoretical results achieving the *twofold* adaptive rates. Besides, the smoothness-adaptive procedures of bandwidth selection are based on optimization over d-dimensional grids of bandwidths, thus require intensive computation, even in moderately high dimension as the grid grows exponentially fast with the dimension.

The principle of Rodeo, and of its derived versions (Wasserman and Lafferty, 2006; Nguyen, 2018), is to progressively build a monotonous path of bandwidths through the bandwidths grid. The construction of this path is based on tests at each iteration to decide if a bandwidth component has a convenient level or still has to be multiplied by an iterative step factor. The tests rely on the partial derivatives of the estimator with respect to the components of the current bandwidth: for $h \in (0, 1]^d$ and $j \in 1 : d$,

$$Z_{hj} := \frac{\partial}{\partial h_j} \hat{f}_h(w). \tag{2.7}$$

The main idea is to use Z_{hj} as a proxy of $\frac{\partial}{\partial w_j} f$, relying on the natural intuition that the more f is varying, the smaller the bandwidth is needed to fit the curve. It is consistent with the minimax bandwidth level $h_j^* = 1$ for irrelevant j and the flatness of the curve in such a direction. Using the Blr family of conditional density estimators, the Z_{hj}'s are well defined as

soon as the kernel K is C^1. They are straightforwardly expressed, thus easily implementable, by using the following equation:

$$Z_{hj} = -\frac{1}{nh_j^2} \sum_{i=1}^{n} \frac{1}{\tilde{\mathrm{f}}_X(X_i)} J\left(\tfrac{w_j - W_{ij}}{h_j}\right) \prod_{k \neq j}^{d} h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right), \tag{2.8}$$

where J denotes the function $t \mapsto K(t) + tK'(t)$. Note that, under the condition $\tilde{\mathrm{f}}_X = \mathrm{f}_X$, if j is an irrelevant component,

$$\mathbb{E}[Z_{hj}] = 0, \tag{2.9}$$

which is expected in view of (2.7) (see Lemma 2 in Appendix or Lemma 6 of Nguyen (2018) for a rigorous proof). The tests involved in the RODEO procedure consist in comparing $|Z_{hj}|$ to a threshold λ_{hj}. The threshold is chosen as follows:

$$\lambda_{hj} := \mathrm{C}_\lambda \sqrt{\frac{(\log n)^a}{nh_j^2 \prod_{k=1}^{d} h_k}}, \tag{2.10}$$

with $\mathrm{C}_\lambda = 4\|J\|_2 \|K\|_2^{d-1}$ and an hyperparameter $a > 1$. It is determined by Bernstein's concentration inequalities to ensure that with high probability Z_{hj} is close to its expectation: $|Z_{hj} - \mathbb{E}[Z_{hj}]| \leq \frac{1}{2}\lambda_{hj}$. The hyperparameter a quantifies the degree of high probability. This definition is justified by following heuristic arguments. With $B(h) = \mathbb{E}[\hat{f}_h(w)] - f(w)$,

$$\frac{\partial}{\partial h_j} B(h) = \frac{\partial}{\partial h_j} \mathbb{E}[\hat{f}_h(w)] = \mathbb{E}\left[\frac{\partial}{\partial h_j} \hat{f}_h(w)\right] = \mathbb{E}[Z_{hj}].$$

If the upper bound of (2.4) is tight and since, with large probability, $Z_{hj} \approx \mathbb{E}[Z_{hj}]$, we obtain, for $j \in \mathcal{R}$

$$|Z_{hj}| \approx h_j^{s-1}.$$

We stop the algorithm when $|Z_{hj}| \approx \lambda_{hj}$ since for this bandwidth h, we expect

$$h_j{}^{s-1} \approx \lambda_{hj} \approx \frac{1}{h_j \sqrt{n \prod_{k=1}^{d} h_k}} \text{ (up to the logarithmic term)},$$

which corresponds to the minimax bandwidth h^* which satisfies the minimax trade-off:

$$h_j^{*2s} \approx \frac{1}{n \prod_{j=1}^{d} h_j^*},$$

for $j \in \mathcal{R}$.

2.2.2 Initialization of the algorithm and variants of CDRODEO

The previous paragraph explains quantities involved in the algorithm, its main ideas and the stopping criterion. We now study the initialization of the algorithm. We describe several alternatives.

Algorithm 0 Direct CDRodeo algorithm

Given a starting bandwidth $h^{(0)} = (h_0, \dots, h_0)$ with $h_0 > 0$, the decreasing iterative step factor $\beta \in (0,1)$, a hyperparameter $a > 1$, the activation of all components.

While there are still active components,
for all active component j, we test if $|Z_{hj}|$ is large (with respect to a threshold λ_{hj} defined in (2.10)):

- If $|Z_{hj}| > \lambda_{hj}$, then $h_j \leftarrow \beta h_j$, and j remains active.
- Else, j is deactivated and h_j remains unchanged for the next steps of the path.

Output The loop stops when either all components are deactivated or the bandwidth is too small $\left(\prod_{j=1}^d h_j < \frac{\log n}{n}\right)$, then the final bandwidth is selected and denoted $\hat{h}$.

Direct CDRodeo algorithm. The natural idea consists in initializing the bandwidth at a large enough level and then decreasing the components of the bandwidth until $|Z_{hj}| \leq \lambda_{hj}$. The detailed procedure is stated in Algorithm 0.

This procedure, called Direct CDRodeo, has been deeply studied by Nguyen (2018). Two cases can be distinguished for a component h_j. Either h_j is selected at the first iteration, or when $|Z_{hj}| \approx \lambda_{hj}$.

In the first case, remark that testing $|Z_{h^{(0)}j}| \leq \lambda_{h^{(0)}j}$ corresponds to testing the hypothesis $|\mathbb{E}[Z_{h^{(0)}j}]| \leq \frac{1}{2}\lambda_{h^{(0)}j}$, which is satisfied for any irrelevant component j: for any h, $\mathbb{E}[Z_{hj}] = 0$. So, with high probability the irrelevant bandwidth components are selected at the initialization level h_0, *i.e.* as large as allowed by the procedure, in line with the minimax approach.

In the second case, the component j is selected after a few iterations, and $|Z_{hj}| \approx \lambda_{hj}$ (where the approximation is due to the discretization in $\{\beta^k h_0, k \in \mathbb{N}\}^d$). Thus with high probability: $\frac{1}{2}\lambda_{hj} \lesssim |\mathbb{E}[Z_{hj}]| \lesssim \frac{3}{4}\lambda_{hj}$. For a relevant component j, for s an integer larger than 1, Nguyen (2018) proved that

$$|\mathbb{E}[Z_{hj}]| \approx h_j^{s-1},$$

if the derivative satisfies $|\partial_j^s f| > 0$ on the neighborhood of the evaluation point w. The assumption is quite restrictive. In particular, it excludes any density that is locally a polynomial of order smaller than s. Moreover, s has to be an integer.
When this assumption is not satisfied, Direct CDRodeo may stop with a too large bandwidth. Indeed, remember it begins with a large initial bandwidth in order to select large irrelevant bandwidth components, but the relevant components have to be selected much smaller. Between these two bandwidth levels, $\mathbb{E}[Z_{hj}]$ may have a change of sign, thus vanishes briefly before becoming larger (in absolute value) than λ_{hj} again. We have illustrated this problem in Figure 1 where we show that the initialization h_0 is not convenient.

In view of this issue, we consider in the following some variations to the *Direct* CDRodeo procedure.

A Reverse CDRodeo algorithm. The first variation which could be considered is the *Reverse* CDRodeo procedure in the same spirit as Liu et al. (2007) (see Section 4.2 therein).

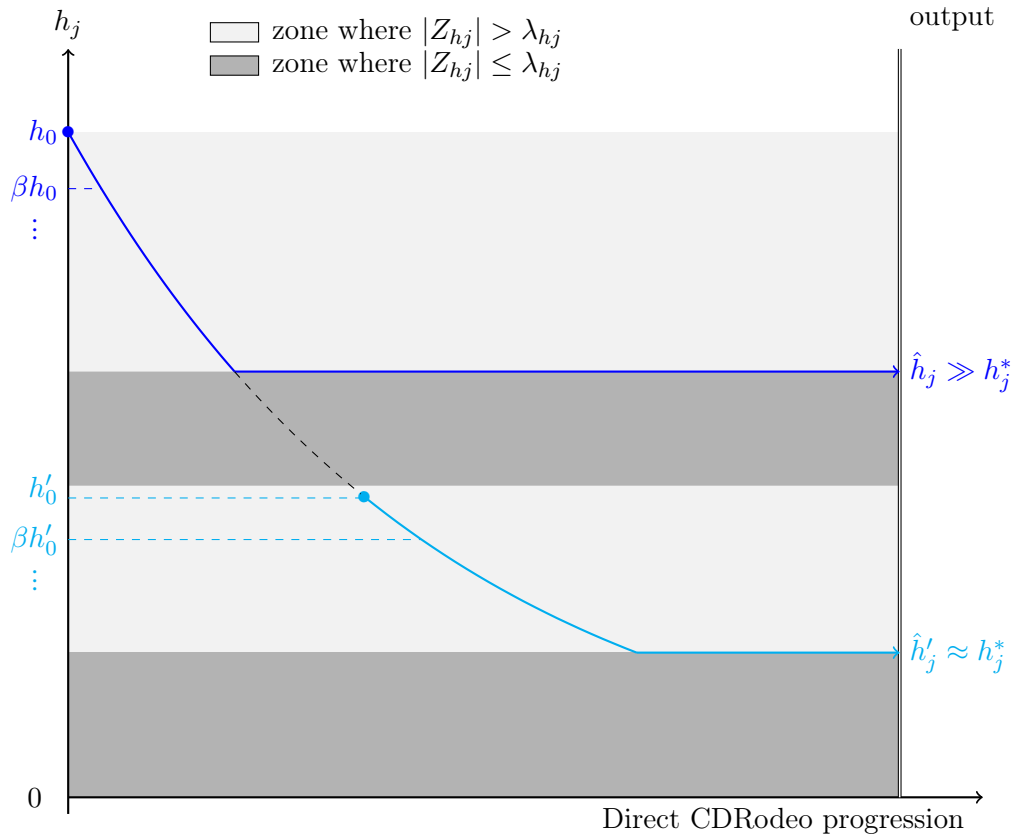

Figure 1: Two bandwidth paths for Direct CDRodeo with two different initializations, when $h_j \mapsto |Z_{hj}|/\lambda_{hj}$ is not monotonuous (larger than 1 in the lightgray zone and smaller than 1 in the darkgray zone). Starting with a large h_0, the algorithm stops when $|Z_{hj}|$ becomes smaller than λ_{hj} and provides a too large output bandwidth. Starting with h_0', the algorithm can provide the optimal bandwidth h_j^*. Observe that the area where $\mathbb{E}[Z_{hj}] \geq \lambda_{hj}$ is unknown, so h_0' is intractable.

We start with a small bandwidth and use a sequence of non-decreasing bandwidths to select the optimal value, still by comparing the Z_{hj}'s with the λ_{hj}'s. More precisely, instead of decreasing the bandwidth components by multiplied them by the factor β when $|Z_{hj}| \geq \lambda_{hj}$, the reverse algorithm increases them by *dividing* them by β when $|Z_{hj}| < \lambda_{hj}$. Note that with this second test, it does not matter if Z_{hj} vanishes. As illustrated by Liu et al. (2007), this approach is very useful for image data. However, the choice of the initial bandwidth is very sensitive. In particular, assume that f has a very low regularity and has only one relevant component, say the first one for instance. In this case, if h^* is the ideal bandwidth, h_1^* has to be as small as possible, i.e. $h_1^* = 1/n$ (up to a logarithmic term). Therefore, since $\mathcal{R}$ is unknown, the initialization of the bandwidth must be not larger than $h_{0,\text{rev}} = (1/n, \ldots, 1/n)$. However, such a small bandwidth leads to instability problems. In particular, the variance of $\hat{f}_{h_{0,\text{rev}}}(w)$ is of order n^{d-1} (see Equation (2.5)).

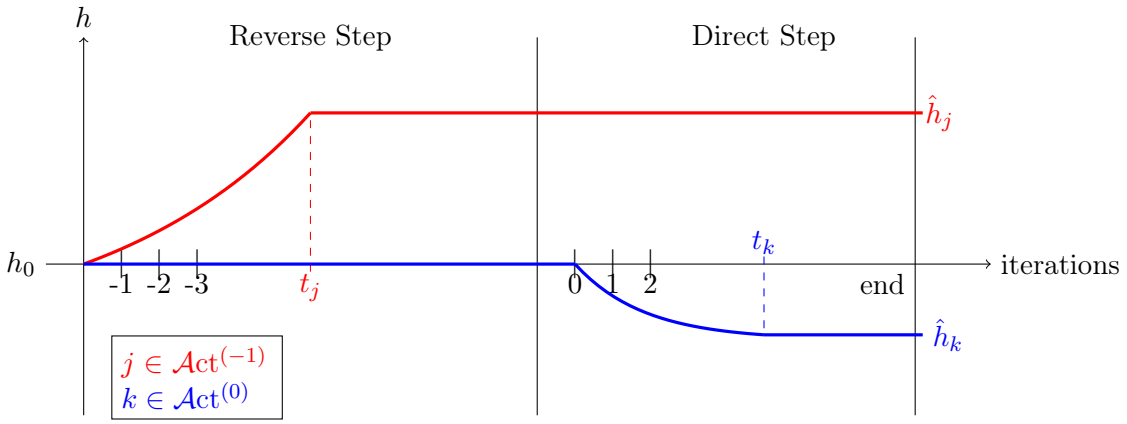

Figure 2: The two patterns of bandwidth path: the components $j \in \mathcal{A}\mathrm{ct}^{(-1)}$ with a deactivation time $t_j \leq 0$ in red, and in blue the components $k \in \mathcal{A}\mathrm{ct}^{(0)}$ with a deactivation time $t_k \geq 0$.

2.2.3 Our method: the RevDir CDRodeo procedure

In view of the analysis led in Section 2.2.2, we propose to give the option for each bandwidth component to either increase or decrease. The procedure is precisely described by Algorithm 1. The initial bandwidth can then be chosen at an intermediate level (and we show later that h_0 has to be chosen larger than the relevant components of the minimax bandwidth), then our procedure comprises the two following steps:

1. The first step consists in the execution of a Reverse CDRodeo procedure to increase the bandwidth components that need to be increased (including the irrelevant ones).

2. The second step executes a Direct CDRodeo procedure on the other bandwidth components.

The output bandwidth of the algorithm is denoted by $\hat{h}$, and the estimator of f by $\hat{f} := \hat{f}_{\hat{h}}$. Figure 2 illustrates the two kinds of path for the bandwidth components. If the component belongs to $\mathcal{A}\mathrm{ct}^{(-1)}$ (resp. $\mathcal{A}\mathrm{ct}^{(0)}$), it is deactivated during the Direct Step (resp. the Reverse Step) and has to be chosen larger (resp. smaller) than the initial bandwidth value h_0. Note that the RevDir procedure generalizes both the Direct and Reverse procedures in function of the choice of h_0. Indeed, if we set $h_0 = 1$, the RevDir procedure behaves as a Direct procedure with the same initialization. Conversely, setting $h_0 = 1/n$ brings us back on the Reverse procedure. Nonetheless, the purpose of our approach is to provide a better tuning of h_0, as discussed in the next section, to solve the initialization issue of the Direct and Reverse procedures.

3 Theoretical results

3.1 Sparsity and smoothness classes of functions

This section is devoted to the theoretical results satisfied by the RevDir CDRodeo procedure. We consider a kernel function $K : \mathbb{R} \to \mathbb{R}$ of class C^1, with compact support denoted $\mathrm{supp}(K)$.

Algorithm 1 RevDir CDRodeo algorithm

1. *Input:* the estimation point w, the observations W, the bandwidth decreasing factor $\beta \in (0,1)$, the bandwidth initialization value $h_0 > 0$, a tuning parameter $a > 1$.

2. *Initialization:*

   ▷ Initialize the trial bandwidth: *for* $k \in 1:d$, $H_k^{(0)} \leftarrow h_0$.

   ▷ Determine which variables are active for the Reverse Step or for the Direct Step:

   $$\mathcal{A}\mathrm{ct}^{(-1)} \leftarrow \{k \in 1:d,\ |Z_{H^{(0)}k}| \leq \lambda_{H^{(0)}k}\}$$
   $$\mathcal{A}\mathrm{ct}^{(0)} \leftarrow \{1:d\} \setminus \mathcal{A}\mathrm{ct}^{(-1)}$$

3. *Reverse Step:*

   ▷ Initialize the counter: $t \leftarrow -1$

   ▷ Initialize the current bandwidth: $\hat{h}^{(-1)} \leftarrow H^{(0)}$

   ▷ While $(\mathcal{A}\mathrm{ct}^{(t)} \neq \emptyset)$ & $(\max \hat{h}_k^{(t)} \leq \beta)$:

   ▶ Set the current trial bandwidth: $H_k^{(t)} = \begin{cases} \beta^{-1}\hat{h}_k^{(t)} & \text{if } k \in \mathcal{A}\mathrm{ct}^{(t)} \\ \hat{h}_k^{(t)} & \text{else.} \end{cases}$

   ▶ Set the next active set: $\mathcal{A}\mathrm{ct}^{(t-1)} \leftarrow \{k \in \mathcal{A}\mathrm{ct}^{(t)},\ |Z_{H^{(t)}k}| \leq \lambda_{H^{(t)}k}\}$

   ▶ Update the current bandwidth: $\hat{h}_k^{(t)} \leftarrow \begin{cases} H_k^{(t)} & \text{if } k \in \mathcal{A}\mathrm{ct}^{(t-1)} \\ \hat{h}_k^{(t)} & \text{else.} \end{cases}$

   ▶ Initialize the next bandwidth: $\hat{h}^{(t-1)} \leftarrow \hat{h}^{(t)}$

   ▶ Decrement the counter: $t \leftarrow t-1$

4. *Direct Step:*

   ▷ Initialize the current bandwidth: $\hat{h}^{(0)} \leftarrow \hat{h}^{(t)}$

   ▷ Reinitialize the counter: $t \leftarrow 0$

   ▷ While $\left(\mathcal{A}\mathrm{ct}^{(t)} \neq \emptyset\right)$ & $\left(\prod_{k=1}^{d} \hat{h}_k^{(t)} \geq \frac{(\log n)^{1+a}}{n}\right)$:

   ▶ Increment the counter: $t \leftarrow t+1$

   ▶ Set the current active set: $\mathcal{A}\mathrm{ct}^{(t)} \leftarrow \{k \in \mathcal{A}\mathrm{ct}^{(t-1)},\ |Z_{\hat{h}^{(t-1)}k}| > \lambda_{\hat{h}^{(t-1)}k}\}$

   ▶ Set the current bandwidth: $\hat{h}_k^{(t)} \leftarrow \begin{cases} \beta.\hat{h}_k^{(t-1)} & \text{if } k \in \mathcal{A}\mathrm{ct}^{(t)} \\ \hat{h}_k^{(t-1)} & \text{else.} \end{cases}$

5. *Output:* $\hat{h} \leftarrow \hat{h}^{(t)}$ (and compute $\hat{f}_{\hat{h}}(w)$).

We shall also assume that K is of order p, *i.e.*: for $\ell \in 1:p-1$, $\int_{\mathbb{R}} t^\ell K(t)dt = 0$. Taking a kernel of order p is usual for the control of the bias of the estimator. Then, we define the neighborhood $\mathcal{U}$ of the point $w \in \mathbb{R}^d$ as follows:

$$\mathcal{U} := \left\{u \in \mathbb{R}^d : w - u \in (\mathrm{supp}(K))^d\right\}.$$

In the sequel, we denote

$$\|f\|_{\infty,\,\mathcal{U}} := \sup_{x\in\,\mathcal{U}} |f(x)|.$$

Remark 2. *The size of $\mathcal{U}$ is fixed. But $\mathcal{U}$ could be chosen so that its size goes to 0. In this case, we have to modify the stopping rule of the Reverse Step, namely $\max \hat{h}_k^{(t)} \leq \beta$, to force $\max \hat{h}_k^{(t)} \overset{n\to\infty}{\longrightarrow} 0$. For instance, if we impose $\max \hat{h}_k^{(t)} \leq \frac{1}{\log n}$, the rates of convergence of our estimate would typically be deteriorated by a logarithmic term.*

The notion of relevant components has already been introduced in Section 1.2 but subsequent results only need the function f to be locally sparse, so we shall consider the following definition depending on $\mathcal{U}$.

Definition 1. *We denote $\mathcal{R}$ the subset of $\{0,\dots,d\}$ with cardinal r such that for any fixed $\{z_j\}_{j\in\mathcal{R}}$, the function $\{z_k\}_{k\in\mathcal{R}^c} \mapsto f(z_1,\dots,z_d)$ is constant on $\mathcal{U}$. We call relevant any component in $\mathcal{R}$.*

The previous definition means that on $\mathcal{U}$, f depends only on r of its d variables.

Remark 3. *In fact,* CDRodeo *detects more complex sparsity structures. In particular, $\mathcal{R}^c$ could be enlarged to the components which are polynomial of degree smaller than the order p of the kernel, namely it suffices to consider $f(z) = z_j^l g(z_{-j})$ with $l \in 0 : p-1$, $z_{-j} = (z_1,\dots,z_{j-1},z_{j+1},\dots,z_d)$ and where g is an arbitrary function. Then j is considered as an irrelevant component by both our algorithm and in the bias-variance trade-off. Indeed, assume that $\tilde{\mathrm{f}}_X = \mathrm{f}_X$ for the sake of simplicity. Then for the algorithm, easy computations leads to $\mathbb{E}[Z_{hj}] = 0$ as an irrelevant bandwidth component. Then our algorithm behaves exactly as if j were irrelevant and select a large $\hat{h}_j$ (with high probability). For the bias-variance trade-off, the bias for f is proportional to the bias for g (multiplied by a term that does not depend on h):*

$$(\mathrm{K}_h \star f - f)(w) = w_j^l\,(\mathrm{K}_h^{(-j)} \star g - g)(w_{-j}), \qquad \mathrm{K}_h^{(-j)}(z_{-j}) := \prod_{k\neq j} K_h(z_k),$$

exactly as if j were irrelevant and $f(z) = cg(z_{-j})$, for c a constant. Then since only the variance depends on h_j, the bias-variance trade-off chooses a large value for h_j.
In particular there is no need for preliminary linear variable selection as suggested in Section 6.1 of (Lafferty and Wasserman, 2008).

In the sequel we derive rates on Hölder balls defined as follows.

Definition 2. *Let $L > 0$ and $s > 0$. We say that the conditional density f belongs to the Hölder ball of smoothness s and radius L, denoted $\mathcal{H}_d(s,L)$, if f is of class C^q and if it satisfies for all $z \in \mathcal{U}$ and for all $t \in \mathbb{R}$ such that $z + te_k \in \mathcal{U}$*

$$\left|\partial_k^q f(z+te_k) - \partial_k^q f(z)\right| \leq L|t|^{s-q},$$

where $q = \lceil s-1\rceil = \max\{l \in \mathbb{N} : l < s\}$ and e_k is the vector where all coordinates are null except the kth one which is equal to 1.

We investigate adaptive results in terms of sparsity and smoothness properties on Hölder balls $\mathcal{H}_d(s,L)$, with $s > 1$. Adaptation means that our procedure will not depend on the knowledge of $\mathcal{R}$ and (s,L). The condition on s means that f has to be at least C^1. This technical assumption is related to our methodology based on derivatives of $\hat{f}_h(w)$ as proxies of derivatives of f to detect relevant components.

3.2 Range of the algorithm inputs and assumptions

The RevDir CDRodeo procedure depends on three tuning parameters, namely h_0, β and a. In the sequel, we take $\beta \in (0,1)$. Since β is an exponential decay factor, its value has no influence on rates of convergence (up to the constant factor). The parameter a will be assumed to be larger than 1. Its value does not affect the main polynomial factor $n^{-\frac{s}{2s+r}}$ of the rate of convergence but only the logarithmic factor: the smaller a, the smaller the exponent of the logarithmic factor. See Section 4.2 for a detailed analysis of the practical choices for a and β. Finally, to initialize the procedure, we take h_0 such that

$$\mathrm{C}_\lambda^{2/d} \left(\frac{(\log n)^a}{n} \right)^{\frac{1}{d(2p+1)}} \leq h_0 \leq 1, \tag{3.1}$$

where C_λ, only depending on the kernel K, is defined in Section 2.2.1. Note in particular that the lower bound does not depend on any unknown value, and thus can be implemented as the bandwidth initialization. Besides, observe that each component of the minimax bandwidth for estimating f on $\mathcal{H}_d(s,L)$ is of order $n^{-1/(2s+r)}$ for relevant components and are constant for irrelevant ones. So, if $s \leq p$, as assumed in Theorem 1, then h_0 is larger than all relevant components of the optimal bandwidth, as required by the RevDir CDRodeo procedure.

To derive rates of convergence for $\hat{f}(w)$, we need three assumptions. The first two ones are related to f_X, the density of the X_i's. We recall that the evaluation point is $w = (x,y)$.

Assumption $\boldsymbol{\mathcal{L}_X}$ *[Lower bound on f_X]*
The density f_X is bounded away from 0 in the neighborhood of x:

$$\delta := \inf_{u \in \mathcal{U}_1} \mathrm{f}_X(u) > 0,$$

where $\mathcal{U}_1 := \left\{ u \in \mathbb{R}^{d_1} : x - u \in (\mathrm{supp}(K))^{d_1} \right\}$.

Remark 4. *Similarly to Remark 2, the size of $\mathcal{U}_1$ is fixed but it could decrease to 0 if we modify the stopping rule of the Reverse Step.*

This assumption is classical in the regression setting or for conditional density estimation. Indeed, if f_X is equal or close to 0 in the neighborhood of x, we have no or very few observations to estimate the distribution of Y given $X = x$. Thus, this assumption is required in all of the aforementioned works about conditional density estimation.

The next assumption specifies that we can estimate f_X very precisely.

Assumption $\boldsymbol{\mathcal{E}\mathrm{f}_X}$ *[Estimation of f_X]*
The estimator of f_X in (2.1) satisfies the following two conditions:

Condition (i) a positive lower bound: $\tilde{\delta}_X := \inf_{u \in \mathcal{U}_1} \tilde{\mathrm{f}}_X(u) > n^{-1/2}$,

Condition (ii) a concentration inequality in local sup norm:

$$\mathbb{P}\left(\sup_{u \in \mathcal{U}_1} \left| \mathrm{f}_X(u) - \tilde{\mathrm{f}}_X(u) \right| > M_X \frac{(\log n)^{\frac{a}{2}}}{\sqrt{n}} \right) \leq \exp(-(\log n)^{1+\frac{a-1}{2}}),$$

with $M_X := \frac{\delta \|J\|_2 \|K\|_2^{d-1}}{4\|f\|_{\infty,\mathcal{U}} \|J\|_1 \|K\|_1^{d-1}}$.

Remark 5. *For the simpler problem of density estimation, since* $\mathrm{f}_X \equiv 1 \equiv \tilde{\mathrm{f}}_X$*, Assumption* $\boldsymbol{\mathcal{E}\mathrm{f}_X}$ *is obviously satisfied.*

This $\sqrt{n}$-rate can be achieved either by restricting f_X to a parametric class, or by assuming we have at hand a larger sample of X. In particular, the following proposition provides precise conditions to satisfy Assumption $\boldsymbol{\mathcal{E}\mathrm{f}_X}$ using a well-tuned kernel density estimator $\tilde{\mathrm{f}}_X$. Furthermore, $\tilde{\mathrm{f}}_X$, the estimator provided by the proof of Proposition 1, is easily implementable.

Proposition 1. *Given a sample* $\tilde{X}$ *with same distribution as* X *and of size* $n_X = n^c$ *with* $c > 1$*, if* f_X *is of class* $C^{p'}$ *with* $p' \geq \frac{d_1}{2(c-1)}$*, there exists an estimator* $\tilde{\mathrm{f}}_X$ *which satisfies Assumption* $\boldsymbol{\mathcal{E}\mathrm{f}_X}$*.*

To prove Proposition 1, we build $\tilde{\mathrm{f}}_X$ as a truncated kernel estimator with a fixed bandwidth, but other methods can be used in practice, as, for instance, a Rodeo algorithm for density estimation. Actually any reasonable nonparametric estimator would have a rate of convergence in sup norm of the form $n_X^{-\beta}$ (typically $\beta = p'/(2p' + d_1)$) up to a logarithmic term. Then Condition (ii) of Assumption $\boldsymbol{\mathcal{E}\mathrm{f}_X}$ is verified as soon as $n_X^{-\beta} \leq n^{-1/2}$ and we need $c \geq 1 + d_1/(2p')$. Then, observe that if f_X is of class C^∞, then we just need $c = 1$ and we can take $\tilde{X} = X$. If we know that f_X is at least of class C^1 but its precise smoothness is unknown, taking $c \geq 1 + d_1/2$ is sufficient to satisfy assumptions of Proposition 1.

The next assumption is necessary to control the bias.

Assumption $\boldsymbol{\mathcal{C}}$
For all $j \in \mathcal{R}$*, for all* h *and* $h' \in [\frac{1}{n}, 1]^d$ *such that* $h \preceq h'$*,* $|\mathbb{E}[\bar{Z}_{h,j}]| \leq |\mathbb{E}[\bar{Z}_{h',j}]|$*, where* $\bar{Z}_{h,j}$ *is defined as* $Z_{h,j}$ *in (2.7) but with true* f_X *replacing* $\tilde{\mathrm{f}}_X$*.*

Let us comment Assumption $\boldsymbol{\mathcal{C}}$. First observe that it is verified by the sharp bound of $\mathbb{E}[\bar{Z}_{hj}]$ over the class $\mathcal{H}_d(s, L)$ ($s > 1$): denoting M_j the pseudo-kernel defined by $M_j(z) = J(z_j)\prod_{k\neq j} K(z_k)$,

$$|\mathbb{E}[\bar{Z}_{hj}]| = |\tfrac{\partial}{\partial h_j}(\mathrm{K}_h \star f - f)(w)| = |\tfrac{1}{h_j}\int M_j(z)[f(w - h.z) - f(w)]dz| \lesssim h_j^{-1}\sum_{k=1}^d h_k^s, \quad (3.2)$$

the last inequality coming from Taylor expansion and the s-Hölder smoothness of f. Then for all $h, h' \in [\frac{1}{n}, 1]^d$ such that $h \preceq h'$, $h_j^{-1}\sum_{k=1}^d h_k^s \leq {h'_j}^{-1}\sum_{k=1}^d {h'_k}^s$.

Assumption $\boldsymbol{\mathcal{C}}$ is named after convexity or concavity, as it requires monotony of $\mathbb{E}[\bar{Z}_{hj}]$, which is the derivative of the bias (after removing the potential perturbations of the pre-estimator $\tilde{\mathrm{f}}_X$ by replacing $\tilde{\mathrm{f}}_X$ by f_X). The absolute value in the assumption is simply a way to cover both cases (convexity and concavity), since in fact $\mathbb{E}[\bar{Z}_{hj}] \to 0$ as $h \to 0$ (at least in the scope of our results: when the smoothness s is larger than 1). In the context of the algorithm, this assumption prevents $\mathbb{E}[Z_{hj}]$ from vanishing temporarily and thus the algorithm from stopping prematurely. Ensuring that, the bias-variance trade-off is achieved.

Note that otherwise, the non convexity of the squared bias would reverberate on the squared risk, making its minimization much harder, especially when we target greedy algorithms to avoid a computationally intensive optimization over all bandwidths.

Remark 6. *If* f *is smooth enough so that* $\frac{\partial^p}{\partial h_j^p} f(h) \neq 0$ *with* p *such that* $\int u^p K(u)du \neq 0$*, then Assumption* $\boldsymbol{\mathcal{C}}$ *is not required. Nevertheless, the procedure cannot be adaptive in this case. See Nguyen (2018).*

3.3 Main result

We now derive the main result of our paper proved in Section 5 in which we show that $\hat{h}$ is closed to the ideal bandwidth h^* defined in Section 2.2.1. Thus our algorithm is able to both detect the irrelevant components and select the minimax bandwidth for relevant and irrelevant components.

Theorem 1. *For any $r \in 0:d$, $1 < s \leq p$ and $L > 0$, if f has only r relevant components and belongs to $\mathcal{H}_d(s, L)$, then under Assumptions $\boldsymbol{\mathcal{L}_X}$, $\boldsymbol{\mathcal{E}\mathbf{f}_X}$ and $\boldsymbol{\mathcal{C}}$, the pointwise risk of the RevDir* CDRodeo *estimator $\hat{f}_{\hat{h}}(w)$ is bounded as follows: for any $l \geq 1$, for n large enough,*

$$\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right]^{1/l} \leq C\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}} \tag{3.3}$$

where C only depends on $d, r, K, \beta, \delta, L, s, \|f\|_{\infty,\,\mathcal{U}}$.

We can compare the obtained rate with the classical pointwise adaptive minimax rate for estimating a s-regular r-dimensional density, which is $((\log n)/n)^{s/(2s+r)}$ (see Rebelles (2015)). Our procedure achieves this rate up to the term $(\log n)^{s(a-1)/(2s+r)}$. In Section 3.2, we specify that any value $a > 1$ is suitable. So, our procedure is nearly optimal. Actually, we need $a > 1$ to ensure that for n large enough,

$$(\log n)^{a-1} \geq \frac{\|f\|_{\infty,\,\mathcal{U}}}{\delta} \tag{3.4}$$

but if an upper bound (or a pre-estimator) of $\frac{\|f\|_{\infty,\,\mathcal{U}}}{\delta}$ were known, we could obtain the similar result with $a = 1$, and our procedure would be rate-optimal without any additional logarithmic term. Remember that the term $(\log n)^{s/(2s+r)}$ is the price to pay for adaptation with respect to the smoothness (see Tsybakov (1998)). Theorem 1 shows that, in our setting, there is no additive price for not knowing the sparsity, i.e. the value of r. This result is new for conditional density estimation.

Remark 7. *Assumption $\boldsymbol{\mathcal{C}}$ allows for a sharp control of the bias of our estimate and is only used in Section 5.5.1. Refining the decomposition of the term $\bar{B}_h$ in (5.19) shows that we can relax Assumption $\boldsymbol{\mathcal{C}}$. This is done in the supplementary file (Nguyen et al., 2021) where Assumption $\boldsymbol{\mathcal{C}}$ is replaced by Assumption $\boldsymbol{\mathcal{M}}$. The price to pay is an extra logarithmic term $(\log n)^{\frac{2s}{2s+r}}$ in the upper bound (3.3).*

3.4 Algorithm complexity

We now discuss the complexity of CDRodeo. Regarding the computation cost of $\tilde{\mathrm{f}}_X$, the estimator built for the proof of Proposition 1 has complexity $O(d_1 n^c)$ but in practice we use a Rodeo estimator with the same sample size n, which has a complexity $O(d_1 n \log n)$ for each computation of $\tilde{\mathrm{f}}_X(X_i)$ which causes an additional cost in $O(d_1 n^2 \log n)$ (applying following Proposition 2).

Regarding the main part of the algorithm, during the Reverse Step, $|\mathcal{A}\mathrm{ct}^{(-1)}|$ components are updated, and, for fixed h, the computation of all Z_{hj}'s and the comparisons to the thresholds λ_{hj} need $O(|\mathcal{A}\mathrm{ct}^{(-1)}|n)$ operations. In the same way, during the Direct Step, $|\mathcal{A}\mathrm{ct}^{(0)}|$ components are updated and each update needs $O(|\mathcal{A}\mathrm{ct}^{(0)}|n)$ operations. Since the

number of updates is at worse of order $\log(n)$ (because of the stopping conditions), and $|\mathcal{A}\text{ct}^{(-1)}| + |\mathcal{A}\text{ct}^{(0)}| \leq d$, we obtain the following proposition. More details can be found in the proof (see Section 5.6).

Proposition 2. *Apart from the computation of* $\tilde{\mathrm{f}}_X$*, the total worst-case complexity of RevDir* CDRodeo *algorithm is*

$$O(dn \log n).$$

Notice that for classical methods with optimization on a bandwidths grid, the complexity is of ordrer $dn|H|^d$, where $|H|$ denotes the size of the grid for each component. In practice, the grid has to include at least $\log n$ points, which leads to a computational cost $O(dn(\log n)^d)$. For illustration, $d = 5$ and $n = 10^5$, the ratio of complexities $\frac{dn(\log n)^d}{dn \log n}$ is already larger than 1.7×10^4.

4 Simulations

This section is devoted to the numerical analysis of our algorithms. In Section 4.1, we first describe the three examples on which we test CDRodeo. Then we calibrate its parameters in Section 4.2. We finally look at its numerical performances in Section 4.3: we first analyse the behavior of CDRodeo for different examples then assess the sparsity detection by adding an increasing number of irrelevant components. In particular, our analysis relies on the fact that the behavior of CDRodeo is easily explainable from the bandwidth it selects.

4.1 Examples

We describe 3 examples. For this purpose, we denote $\mathcal{N}(a, b)$ the Gaussian distribution with mean a and variance b, $\mathcal{U}_{[a,b]}$ the uniform distribution on the compact set $[a, b]$ and $\mathcal{IG}(a, b)$ the inverse-gamma distribution with parameters (a, b).

- Example (a): We consider $d_2 = 2$ response variables and $d_1 \in 1 : 4$ auxiliary variables with the following hierarchical structure:

$$Y_{i2} \sim \mathcal{IG}(4, 3), \quad Y_{i1}|Y_{i2} \sim \mathcal{N}(0, Y_{i2}), \quad X_{ij}|Y_i \overset{iid}{\sim} \mathcal{N}(Y_{i1}, Y_{i2}),$$

  which leads to the following conditional density (derived in (Nguyen, 2019, Chapter IV, Section 5.a)):

$$f : (x, y) \mapsto \mathbb{1}_{\{y_2>0\}} \frac{\sqrt{d_1+1}}{\sqrt{2\pi}\Gamma(4+\frac{d_1}{2})} (\beta_1(x))^{4+\frac{d_1}{2}} y_2^{-(5+\frac{d_1+1}{2})} e^{-\frac{\beta_1(x)}{y_2} - \frac{\left(y_1 - \frac{\sum_{j=1}^{d_1} x_j}{d_1+1}\right)^2}{\left(\frac{2y_2}{d_1+1}\right)}}$$

  with $\beta_1(x) := \frac{1}{2}\left(6 + \sum_{j=1}^{d_1} x_j^2 - \frac{(\sum_{j=1}^{d_1} x_j)^2}{d_1+1}\right)$.

  This example is an usual Bayesian model (see for example (Raynal et al., 2018)) where one of the tasks is to retrieve the posterior distribution f of the mean Y_{i1}'s and the variance Y_{i2}'s given the normal observations X_i's, which is exactly what our method performs in this paper.

- Example (b): We consider $d_2 = 1$ response variable and $d_1 \in 1:12$ auxiliary variables with the following hierarchical structure:

$$X_{ij} \overset{iid}{\sim} \mathcal{N}(0,1), \quad Y_{i1}|X_i \sim \mathcal{N}(3X_{i1}^3, 0.5^2).$$

  In this case, the conditional density is then

$$f : (x,y) \mapsto \sqrt{\frac{2}{\pi}} e^{-2(y-3x_1^3)^2}.$$

- Example (c): We consider $d_2 = 1$ response variable and $d_1 \in 1:12$ auxiliary variables with the following hierarchical structure:

$$X_{ij} \overset{iid}{\sim} \mathcal{U}_{[-1,1]}, \quad Y_{i1}|X_i \sim \mathcal{N}(3X_{i1}^3, 0.5^2).$$

  In this case, the conditional density is then

$$f : (x,y) \mapsto \sqrt{\frac{2}{\pi}} e^{-2(y-3x_1^3)^2} \mathbb{1}_{\{x \in [-1,1]^{d_1}\}}.$$

Example (a), in which $r = d$, will be used as reference for estimation without sparsity structure and will illustrate the estimation difficulty when we have to face with the curse of dimensionality. Examples (b) and (c) circumvent the curse of dimensionality given their sparsity structure: $r = 2$ (Y_i is scalar and depends only on X_{i1}). Note that Example (c) is discontinuous, whereas our method rather targets C^1-functions.

4.2 Calibration...

In this section, we focus on the calibration of the threshold λ_{hj} and the decay factor β, whereas some other parameters are fixed: in particular, we are using the Gaussian kernel, and the initialization value h_0 is chosen as the lower bound provided in (3.1):

$$h_0 = \mathrm{C}_\lambda^{2/d} \left(\frac{(\log n)^a}{n} \right)^{\frac{1}{d(2p+1)}} \tag{4.1}$$

with $\mathrm{C}_\lambda = 4\|J\|_2 \|K\|_2^{d-1}$. The choice of the threshold is quite sensitive since it influences the bias-variance trade-off, and intensive simulations have been performed to determine the convenient tuning, while the decay factor (which only quantifies the size of the step) rather impacts the running times of the procedure.

We determine each parameter separately, since their respective impact is rather independent. Moreover, to avoid the influence of a chosen $\tilde{\mathrm{f}}_X$ (and its peculiar specificities), the calibration is run with known f_X (which is plugged as input of the algorithm).

4.2.1 ... of the threshold

Since the calibration of β is not done yet, we fix for this section $\beta = 0.9$.

Given Definition (2.10), two parameters influence the threshold: a and C_λ, but they are clearly redundant. Therefore only the calibration of a will be performed while we take the theoretical value of C_λ.

We compare on a grid of values of a the absolute error of our estimator, i.e. $\left|\hat{f}_{\hat{h}}(w) - f(w)\right|$. We abbreviate it AE in the following. Several settings are considered, each corresponding to a separate graph. In particular, we consider for each example a variety of sample sizes ($n \in \{10\ 000; 50\ 000; 100\ 000; 200\ 000\}$) and X of different dimensions ($d_1 \in 1 : d_{\max}$, with $d_{\max} = 6$ in Example (a) and $d_{\max} = 9$ in Examples (b) and (c)). Moreover, in each graph, we consider 3 samples (in the graphs with different line types) and several evaluation points $\{w^k\}_{k=1:16}$ randomly drawn according to the joint distribution f_W (the 16 pastel curves in the graphs). Note that, to refine our selection of a, we add a logarithmic grid to the standard grid of integers, after observing that the AE minimizers increase sublinearly with d_1. We simply

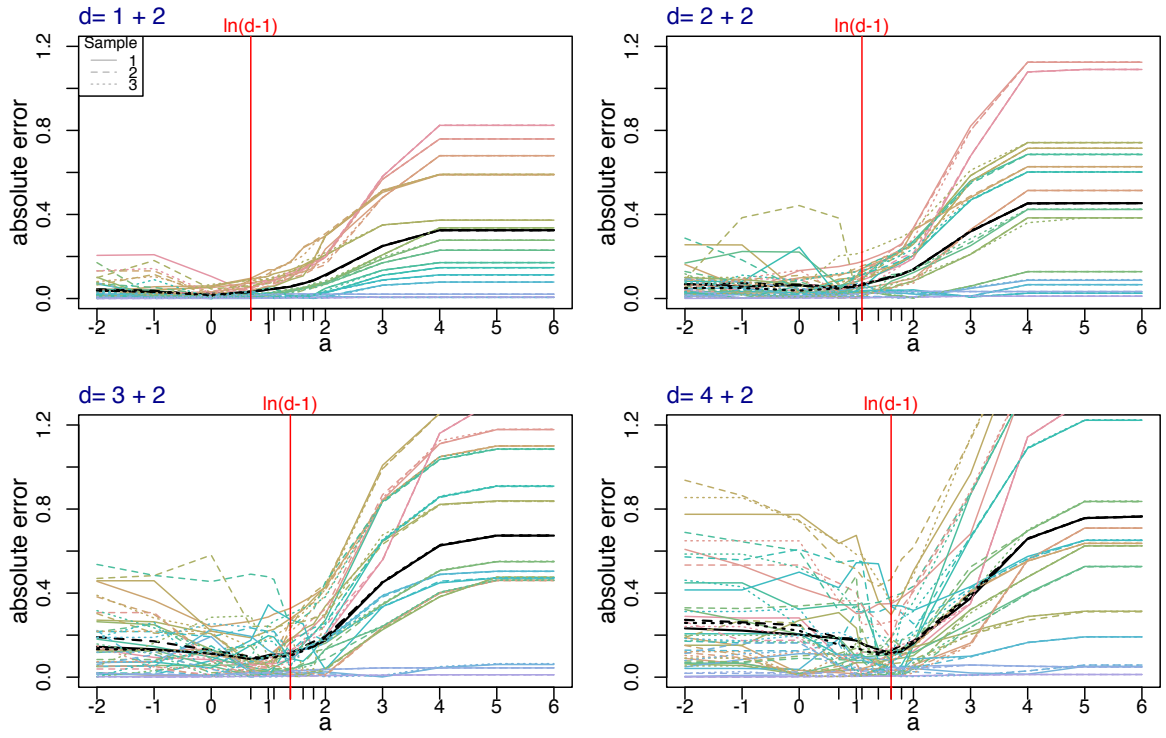

Figure 3: **Illustration of the tuning of a for Example (a) with $n = 200\ 000$ for growing dimensions.** In each subgraph: AE curves in function of a for 16 evaluation points w^k (the warmer the pastel color, the larger $f(w^k)$) given $B = 3$ samples (differentiated by line type) at fixed dimension (specified top left in the form $d = d_1 + d_2$). In black lines: the average per sample of the 16 pastel curves. The vertical straight red line: our final choice.

provide here one case (Example (a) with $n = 200\ 000$ in Figure 3), but the whole set of figures can be found in supplementary material (see Nguyen et al. (2021)).

For ease of interpretation, the average per sample over the different evaluation points has been added in thicker black line. Then, our goal is to determine this minimizer as a function of the varying parameters mentioned above. A good point is that the minimizers do not seem to depend on the sample size (cf the whole set of figures). However the effect of the dimension is more sensitive. First note that the larger a, the larger the thresholds λ_{hj}, thus the larger $\hat{h}$. We observe the chaotic behavior of CDRodeo for small values of a (especially for large dimension and small sample size) and, for large values of a, the superposition of the curves

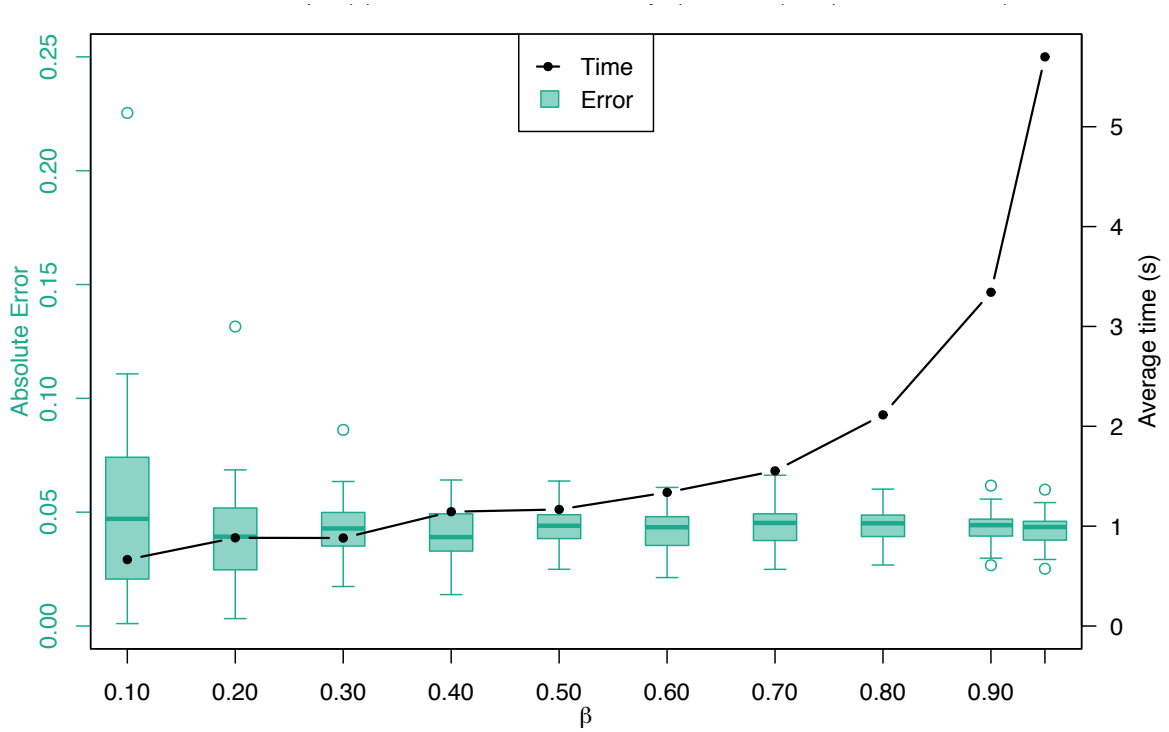

Figure 4: **Illustration of the calibration of β.** For Example (a) with $d_1 = 3$, given $B = 50$ samples of size $n = 100\ 000$, boxplots of the AEs and their average running times (in black) in function of β.

built from different samples, meaning low variance but large bias of the estimators. This corresponds to the usual phenomenon of under- and over-smoothing.

Finally, a good trade-off is achieved by the tuning

$$a = \log(d - 1),$$

and all the following simulations will be implemented with this choice.

4.2.2 ... of the step size

Let us now tune the step parameter, namely β the multiplicative decay factor of the bandwidth. As one can expect, the calibration of β is a compromise between running times and estimation sharpness: the smaller the parameter β, the bigger the step size leading to a faster procedure but a larger approximation error.

In Figure 4 (corresponding to Example (a) with $d_1 = 3$ and $n = 100\ 000$), we put in perspective the boxplots (built given 50 samples) of the AEs with their mean running times. As one could expect with a multiplicative factor, the computational time increases exponentially fast with β: in particular, the running time explodes when $\beta \geq 0.9$. Conversely the smaller β, the larger standard deviation of the boxplots, therefore β should not be taken too small.

To sum up, the range of values satisfying a good compromise is quite large. To fix the parameter, we take

$$\beta = 0.8,$$

and all the following simulations will be implemented with this choice.

4.3 Numerical performances

In this section, we assess the performances of our procedure according to two directions: we first visualize how our procedure reconstructs functions, then we focus on the sparsity detection, the key property of our algorithm to circumvent the curse of dimensionality.

4.3.1 Reconstructions: direction-by-direction visualization and estimation of f_X

We first focus on a global visualization of the estimation of the function f: in particular, we are interested in the performances of our estimator evaluated on a grid. Two kinds of estimates are considered: one in which the true f_X is plugged, the other in which $\tilde{\mathrm{f}}_X$ is estimated by our procedure with the following methodology.

Density estimation: a RevDir CDRodeo procedure for the input $\{\tilde{\mathrm{f}}_X(X_i)\}_{i=1:n}$. First, for the sake of practicality, we use the same sample to compute $\tilde{\mathrm{f}}_X$ and $\hat{f}_{\hat{h}}$. Note that there is no requirement of independence in the theoretical results.

We use the RevDir CDRodeo procedure, since it can perfectly be used for estimating standard densities (cf Remark 1). Since our method is pointwise, we need to compute $\mathrm{f}_X(X_i)$ for each $i \in 1:n$. Note that sparsity structures are rarer in standard densities, for which all variables are of interest, than in conditional densities. Therefore the straightforward estimation of f_X is limited by the dimension of X due to the curse. To circumvent this fact, we propose to add conditioning, artificially, by decomposing f_X as follows:

$$\mathrm{f}_X(x) = \mathrm{f}_{X_1}(x_1) \prod_{j=2}^{d_1} f_{X_j|X_{1:(j-1)}}(x_{1:j}).$$

Notice that the n estimates $\{\tilde{\mathrm{f}}_{X_1}(X_{i,1})\}_{i=1}^n$ are needed as input to compute the $\{\tilde{f}_{X_2|X_1}(X_{i,1:2})\}_{i=1}^n$, which are needed to compute the $\{\tilde{f}_{X_3|X_{1:2}}(x_{1:3})\}_{i=1}^n$, and so on.

Observe also that the previous calibration of a, namely $a = \log(d-1)$, does not extend for univariate densities. Based on preliminary numerical experiments, we set $a = -1$ for the univariate case.

Implemented in R, with a 3.1 GHz Intel Core i7 processor, the running times for $\tilde{\mathrm{f}}_X$ in Example (a) in dimension $d_1 = 2$ and in Examples (b) and (c) in dimension $d_1 = 3$ is summarized in the following table:

	Mean time per run (seconds)			Total time for 100 000 runs
	$\tilde{\mathrm{f}}_{X_1}$	$\tilde{f}_{X_2\|X_1}$	$\tilde{f}_{X_3\|X_{1:2}}$	$\tilde{\mathrm{f}}_X$
Model (a)	0.734	0.654	N.A.	138 780s (around 1d 15h)
Model (b)	1.31	1.61	1.72	463 559s (around 5d 9h)
Model (c)	0.675	1.17	1.05	289 695s (around 3d 8h)

The running times strongly depend on the distance between the initialization bandwidth and the selected one, which explains non increasing running times when the dimension grows for Models (a) and (c).

One may object that several days of computation for the preliminary estimator is quite long. But, note that it is done without parallelization. Given a powerful enough cluster, the running time can be divided by n using parallelization over the evaluation points.

Visualization. In Figures 5, 6, and 7, the two kinds of estimates are built from a sample of dimension $d = 4$ and size $n = 100\ 000$ for respectively Examples (a), (b) and (c). Limited to two-dimensional visualizations, we vary only one component at a time, the others being fixed to a set point: $w = (0, 0, 0, 0.4)$ for Example (a) and $w = (0, 0, 0, 0)$ for Examples (b) and (c).

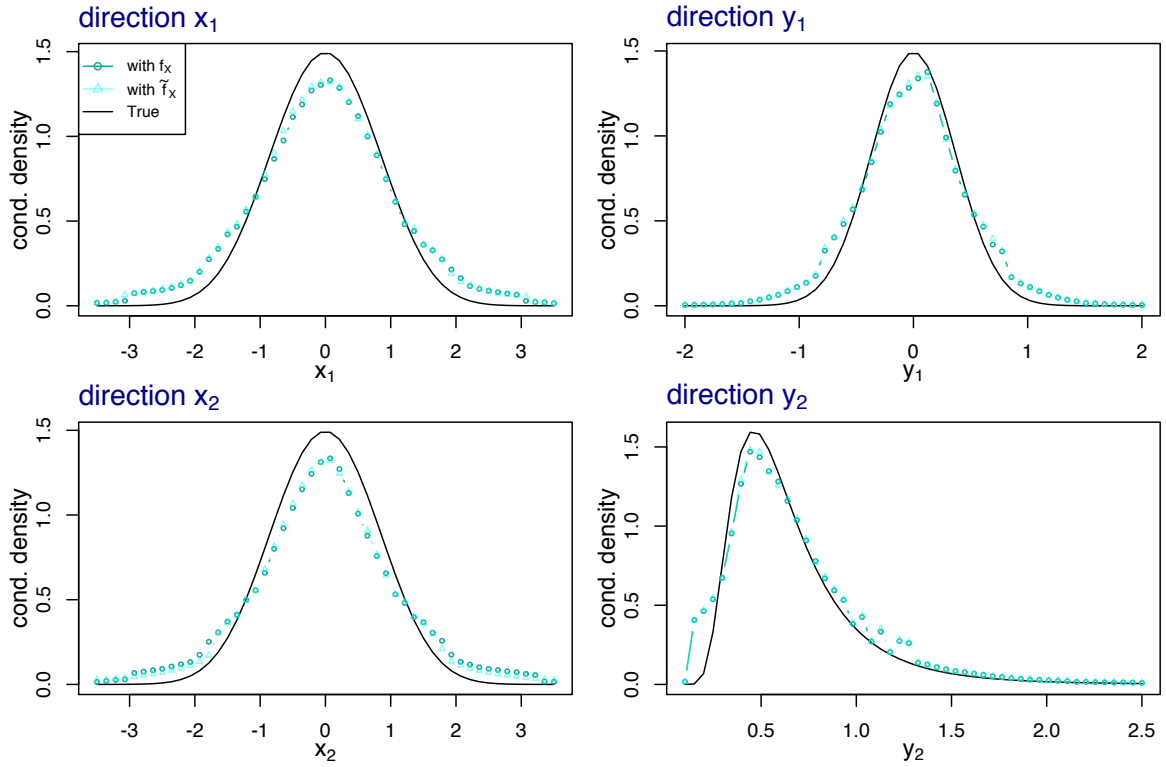

Figure 5: **Reconstruction for Example (a).** Compared to the true conditional density (in black line), our pointwise estimates with the true marginal density f_X (circles in darker shade) or with the estimate $\tilde{f}_X$ (triangles in lighter shade). Only one direction is varying specified at the top left of each graph, the others being fixed to a set point: $w = (0, 0, 0, 0.4)$. The sample size is $n = 100\ 000$.

The overall signal is nicely recovered. Comparing the different examples, Example (a) is the least accurately estimated: the estimates are oversmoothed near the modes. It was expected since it is the example without sparsity and even in dimension as small as 4, the curse deteriorates the convergence rate.

Thanks to the strong similarity between Example (b) and Example (c), the impact of the discontinuity can be properly visualized: (b) is clearly more accurately estimated than (c), even though the focus point $w = 0$ is not really close to the discontinuity points ± 1 (in the directions x_j). The loss of accuracy is once again due to the curse, as the directions x_j, $j \geq 2$, in Example (c) are not completely irrelevant. The CDRodeo procedure does not consider the relevance of a variable as a binary answer: in fact, when a variable is relevant, it can be more or less relevant. See the analysis of the selected bandwidths in the next section for more details.

Besides, in all examples, the estimation is less accurate at the specific points where Assumption $\mathcal{C}$ is not satisfied. Taking account that the chosen kernel is Gaussian, thus of order 2, it especially happens around the zeros of the second derivative.

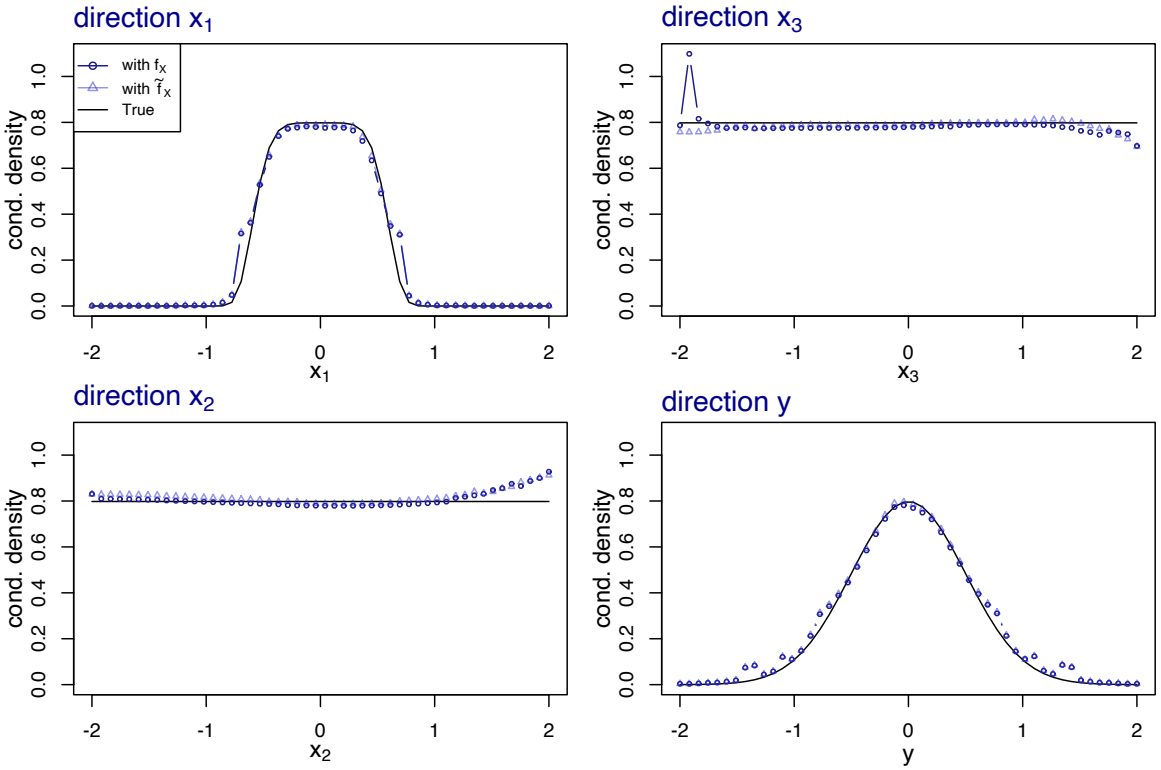

Figure 6: **Reconstruction for Example (b).** See the description in Figure 5, except: $w = (0, 0, 0, 0)$.

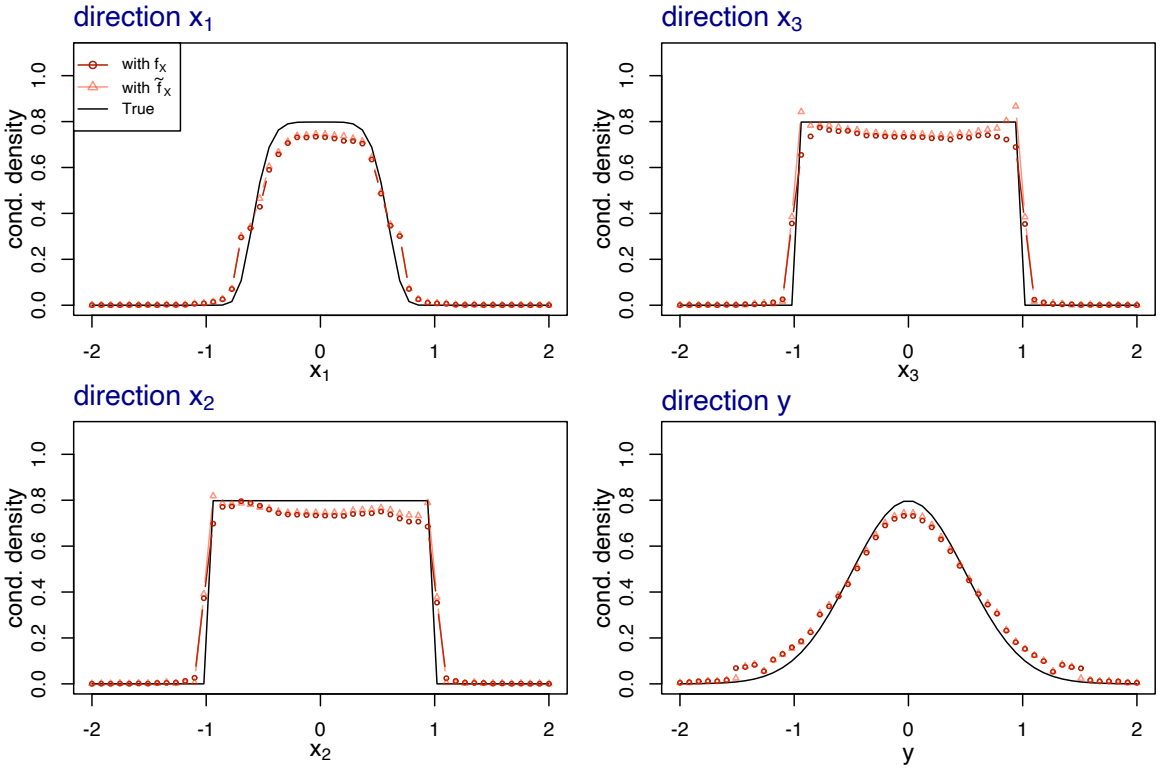

Figure 7: **Reconstruction for Example (c).** See the description in Figure 5, except: $w = (0, 0, 0, 0)$.

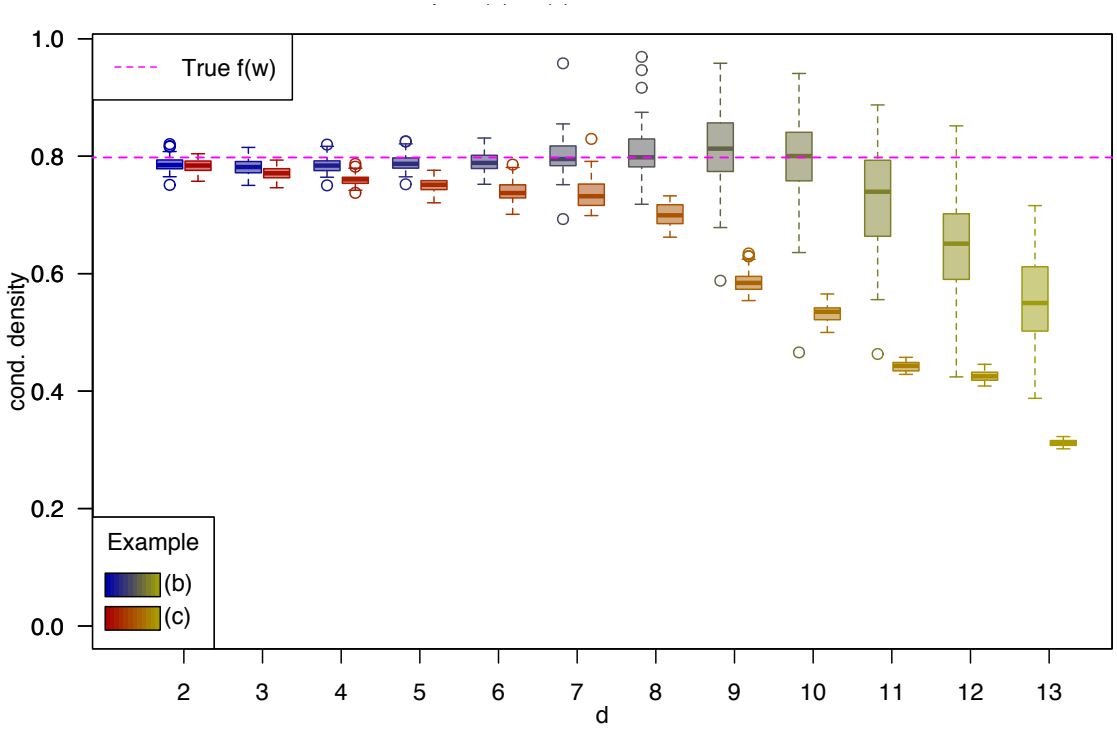

Figure 8: **Robustness to addition of irrelevant variables.** Boxplots of the estimates $\hat{f}_{\hat{h}}(w)$ in function of d, given 50 samples of size $n = 100\ 000$ for Example (b) (in red shades) and Example (c) (in green shades). The dashed horizontal line is the true value $f(w)$ (at the evaluation point $w = 0$).

Then, note that RevDir CDRodeo may stop during the direct Step (as $|Z_{hj}|$ may become smaller than λ_{hj}) but has no impact on the increasing step (Reverse Step). That is why the initialization h_0 is set as the lower bound of its range (see Equation (3.1)) to minimize the undesirable impacts. For illustration of the improvement made by the RevDir algorithm, see Figures IV.3, IV.4 and IV.5 in Ph.D. thesis (Nguyen, 2019) which compares the Direct and the RevDir procedures.

Note lastly that the estimates with either f_X or $\tilde{\mathrm{f}}_X$ are very close to each other. More precisely, the estimates with $\tilde{\mathrm{f}}_X$ is slightly better (in particular, near the modes and near the discontinuity in Example (c)): Delyon et al. (2016) actually prove that dividing by an estimator of the density produces better results than if the density itself was used. That is the reason why the reliability of our results is maintained in the following part even if the true f_X is used in order to save the running times of computing the $\tilde{\mathrm{f}}_X(X_i)$'s for several samples and dimensions.

4.3.2 Impact of the dimension and sparsity detection

Let us now consider how the RevDir CDRodeo procedure detects the sparsity structure. For examples with sparsity structure – namely Examples (b) and (c) –, we check the robustness to irrelevant explanatory variables: starting with the fully relevant example at dimension $d = 2$, we gradually add irrelevant variables until dimension $d = 13$. In Figure 8, the boxplots are built from 50 simulated samples of size $n = 100\ 000$ with varying dimension d_1 from

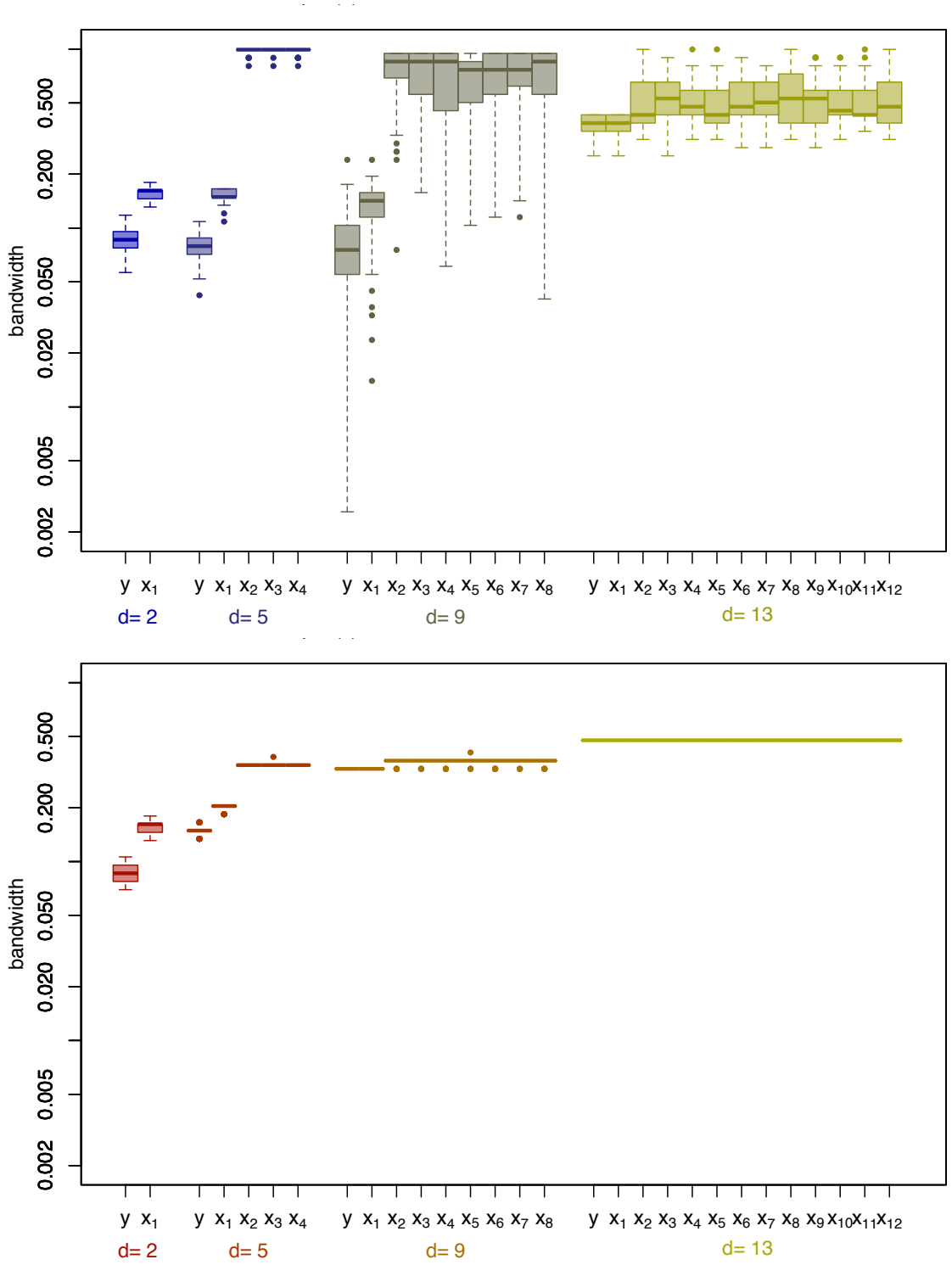

Figure 9: Bandwidths associated to the estimates of Figure 8 for the dimensions $d \in \{2, 5, 9, 13\}$. Top: Example (b). Bottom: Example (c).

1 to 12: in bluish shades, the estimates of Example (b) and in reddish shades, the ones of Example (c). We also provide in Figure 9 the boxplots of the selected bandwidths for the

dimensions $d \in \{2, 5, 9, 13\}$. Notice that our fully nonparametric procedure actually ends within reasonable times for dimensions as large as 13 (e.g. 40 minutes for the whole 600 estimates of Figure 8 on samples of size 100 000), while most nonparametric methods struggle to handle dataset of dimension higher than 4.

Usually, without sparsity, each added variable worsens the estimation: see for instance Example (a) with increasing (relevant) dimension in Figure 13 of the supplementary material Nguyen et al. (2021), in which our method struggles providing good estimates as soon as the dimension 5. For Example (b) (where the relevant dimension is $r = 2$), until the dimension 6, our method has the same behavior as for dimension 2. For larger dimensions, the estimation is progressively noised by the too many irrelevant variables, to finally lose the signal beyond the dimension 10. The bandwidths in Figure 9 give a good understanding of how the procedure handles the extra variables. Comparing the dimensions $d = 2$ and $d = 5$, the relevant bandwidth components (namely the directions y and x_1) are selected at very similar values (called hereafter their "expected values"), while the irrelevant components (in dimension $d = 5$) are taken as high as possible, around the value 1 (the upper limit of the bandwidth grid): thus, the bias-variance trade-off is unchanged, ensuring a quality of estimation as good as in dimension $d = 2$. In dimension $d = 9$, the larger dimension makes the detection of irrelevant variables more difficult, producing variance in the bandwidth selection. Nevertheless, the relevant components are still selected at their expected value (but with more variance), producing rather good estimates. In dimension $d = 13$, the sparsity is less accurately detected: the irrelevant bandwidths decrease to 0.5. Their product 0.5^{11} reaches numerically the emergency stop $\approx \frac{1}{n}$. Therefore, there is not enough room left for the relevant components to decrease until their expected value, which explains the loss of signal observed in Figure 8. Note that in this last setting $d > \log n$, and that is the reason why the emergency stop is reached. More generally, this framework seemed to be out of reach for RODEO-type procedures: in particular, in (Lafferty and Wasserman, 2008) where growing dimensions with n are considered, the framework is also restricted to dimensions $d \ll \log n$.

Let now consider Example (c). The same phenomenon occurs, but complicated by the discontinuity of f in the directions x_j: away from ± 1, the relevant dimension is $r = 2$, but in the neighborhood of ± 1, these components are highly relevant. In fact, these neighborhoods depend on the bandwidth: the larger the bandwidth, the larger the support of K_h until reaching the points ± 1, and once ± 1 belongs to the support of K_h, the components $x_j, j > 1$, are detected as relevant. This is the reason why these bandwidth components are much smaller in Figure 9 (bottom) (around the value 0.48 instead of 1 in Example (c)). These smaller components amplify the phenomenon described for Example (b): as soon as dimension $d = 5$, the relevant components can no longer decrease to their expected value; in dimension $d = 9$, there is almost no room left for the relevant components, and in dimension $d = 13$, the relevant components are completely lost.

All in all, the overall behavior of our procedure is very satisfying: the RevDir CDRODEO procedure nicely detects relevant variables and is robust to extra irrelevant in moderate dimensions ($d \leq \log n$). The difficulties described in the last paragraphs are inherent to the curse of dimensionality and is bound to occur with any nonparametric procedure.

5 Proofs

5.1 Notations

In order to prove the theorem, some intermediate lemmas are needed. See Appendix for their statements. First, we define some general notations: We denote

- $\partial_j g$ the partial derivative of a function g with respect to its j-th component;
- $v \cdot v'$ the multiplication term by term of two vectors v and v';
- $v_{\mathcal{I}}$ the vector v restricted to its components indexed in $\mathcal{I}$;
- $b \vee c = \max(b, c)$ the maximum value of two reals b and c.

Let us now introduce the key quantities of the proofs. For any bandwidth $h \in (\mathbb{R}_+^*)^d$ and any component $k \in 1:d$, we consider the estimator $\bar{f}_h(w)$ that we would have used if the density f_X were known:

$$\bar{f}_h(w) := \frac{1}{n}\sum_{i=1}^{n} \bar{f}_{hi}(w), \quad \bar{f}_{hi}(w) := \frac{\mathrm{K}_h(w - W_i)}{\mathrm{f}_X(X_i)}$$

and we denote Δ_h its difference with the real estimator:

$$\Delta_h := \hat{f}_h(w) - \bar{f}_h(w).$$

We denote $\bar{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$ the bias of $\bar{f}_h(w)$. We also consider its partial derivative $\bar{Z}_{hk}$:

$$\bar{Z}_{hk} := \frac{\partial}{\partial h_k}\bar{f}_h(w).$$

We can write

$$\bar{Z}_{hk} := \frac{1}{n}\sum_{i=1}^{n} \bar{Z}_{hik}, \quad \bar{Z}_{hik} := \frac{1}{\mathrm{f}_X(X_i)}\frac{\partial}{\partial h_k}\left(\prod_{k=1}^{d} h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right)\right).$$

We shall consider $\Delta_{Z,hk}$ the difference between Z_{hk} and $\bar{Z}_{hk}$:

$$\Delta_{Z,hk} := Z_{hk} - \bar{Z}_{hk}.$$

Note that the value of the final bandwidth of our procedure provides the value of the bandwidth at each iteration. More precisely, if a bandwidth h is the output of the RevDir procedure, we denote $(h^{(t)})_{t\in\mathbb{Z}}$, the different values of the bandwidth for all iterations t.
- On the one hand, if $h_k > h_0$, it means that at Initialization, the component k was in $\mathcal{A}\mathrm{ct}^{(-1)}$ and then the bandwidth path of this component has increased during the Reverse Step according to the following path $h_0\beta^{-1}, h_0\beta^{-2}, ...$ until $h_k := h_0\beta^{-|t_k|}$, and remains fixed during the whole Direct Step ($t \geq 0$).
- On the other hand, if $h_k < h_0$, the component k was in $\mathcal{A}\mathrm{ct}^{(0)}$ at Initialization. Thus the value of the bandwidth component was fixed and equals to h_0 during the Reverse Step (i.e for every $t < 0$). Then, it decreases during the Direct step: $h_0\beta, h_0\beta^2, ...$ until $h_k := h_0\beta^{t_k}$ is achieved (see Figure 2). This gives the following formula: for any $k \in 1:d$, during the Reverse Step (when $t < 0$),

$$h_k^{(t)} := \max(h_0, \min\left(h_k, \beta^t h_0\right)) = \begin{cases} h_0 & \text{if } k \text{ is active during the Direct Step,} \\ \beta^t h_0 & \text{if } k \text{ is active during the Reverse Step and not deactivated yet,} \\ h_k & \text{if } k \text{ has already been deactivated during the Reverse Step,} \end{cases}$$

and during Direct Step (when $t \geq 0$),

$$h_k^{(t)} := \max\left(h_k, \beta^t h_0\right) = \begin{cases} \beta^t h_0 & \text{if } k \text{ is active during the Direct Step and not deactivated yet,} \\ h_k & \text{if } k \text{ has already been deactivated (during the Reverse or the Direct Step).} \end{cases}$$

Now we can define the set of bandwidths $\mathcal{H}_{\text{hp}}$ which contains with high probability the bandwidth selected by the RevDir procedure:

$$\begin{aligned} \mathcal{H}_{\text{hp}} := \{h \in \left(\mathbb{R}_+^*\right)^d : \forall k \in 1:d, h_k = \beta^{t_k} h_0 \leq 1 \text{ with } t_k \in \mathbb{Z}, \\ \text{and } \prod_{k=1}^d h_k \geq \beta^r \frac{(\log n)^{a+1}}{n}, \\ \text{and } \forall k \in \mathcal{R}^c, h_k = h_{\text{irr}}\}, \end{aligned}$$

where h_{irr} is uniquely defined by $t_{\text{irr}} \in \mathbb{Z}$ such that $\beta < h_{\text{irr}} := \beta^{t_{\text{irr}}} h_0 \leq 1$. We also denote $\mathcal{H}_{\text{hp}}^{\text{Rev}}$ (respectively $\mathcal{H}_{\text{hp}}^{\text{Dir}}$) the set which contains the different states of the bandwidth during the Reverse Step (respectively the Direct Step) provided that the selected bandwidth is in $\mathcal{H}_{\text{hp}}$:

$$\mathcal{H}_{\text{hp}}^{\text{Rev}} := \{h^{(t)} : h \in \mathcal{H}_{\text{hp}}, t < 0\} \tag{5.1}$$

$$\mathcal{H}_{\text{hp}}^{\text{Dir}} := \{h^{(t)} : h \in \mathcal{H}_{\text{hp}}, t \geq 0\}. \tag{5.2}$$

Finally, we introduce the high probability event $\mathcal{E}_{\text{hp}}$ on which $\hat{h}$ systematically belongs to $\mathcal{H}_{\text{hp}}$:

$$\mathcal{E}_{\text{hp}} := \widetilde{\mathcal{A}}_n \cap \bigcap_{h \in \mathcal{H}_{\text{hp}}} \left(\mathcal{B}\text{ern}_{\bar{f}}(h) \cap \mathcal{B}\text{ern}_{|\bar{f}|}(h)\right) \cap \bigcap_{h \in (\mathcal{H}_{\text{hp}}^{\text{Rev}} \cup \mathcal{H}_{\text{hp}}^{\text{Dir}})} \bigcap_{k=1}^d \left(\mathcal{B}\text{ern}_{\bar{Z}}(h,k) \cap \mathcal{B}\text{ern}_{|\bar{Z}|}(h,k)\right), \tag{5.3}$$

where $\widetilde{\mathcal{A}}_n$ is the high probability event of Condition (ii) in Assumption $\boldsymbol{\mathcal{E}\mathbf{f}_X}$:

$$\widetilde{\mathcal{A}}_n = \left\{ \sup_{u \in\, \mathcal{U}_1} \left| \mathrm{f}_X(u) - \tilde{\mathrm{f}}_X(u) \right| \leq M_X \frac{(\log n)^{\frac{a}{2}}}{\sqrt{n}} \right\},$$

and $\mathcal{B}\text{ern}_\dagger(\ddagger)$ is the high probability event resulting of Bernstein's Inequality applied on the random variable $\dagger$ with parameter(s) $\ddagger$. More formally:

$$\mathcal{B}\text{ern}_{\bar{f}}(h) := \{|\bar{f}_h(w) - \mathbb{E}[\bar{f}_h(w)]| \leq \sigma_h\},$$

$$\mathcal{B}\text{ern}_{|\bar{f}|}(h) := \left\{ \left| \frac{1}{n} \sum_{i=1}^n |\bar{f}_{hi}(w)| - \mathbb{E}[|\bar{f}_h(w)|] \right| \leq \mathrm{C}_{\bar{\mathrm{E}}} \right\},$$

$$\mathcal{B}\text{ern}_{\bar{Z}}(h,k) := \left\{ |\bar{Z}_{hk} - \mathbb{E}\bar{Z}_{hk}| \leq \frac{1}{2}\lambda_{hk} \right\},$$

$$\mathcal{B}\text{ern}_{|\bar{Z}|}(h,k) := \left\{ \left| \frac{1}{n} \sum_{i=1}^n |\bar{Z}_{hik}| - \mathbb{E}|\bar{Z}_{h1k}| \right| \leq \mathrm{C}_{E|\bar{Z}|} h_k^{-1} \right\},$$

where

$$\sigma_h = \mathrm{C}_\sigma \sqrt{\frac{(\log n)^a}{n \prod_{k=1}^d h_k}}$$

with $\mathrm{C}_\sigma = \frac{2\|K\|_2^d \|f\|_{\infty,\,\mathcal{U}}^{\frac{1}{2}}}{\delta^{\frac{1}{2}}}$. See Lemmas 1 and 2 in Appendix for the details and definitions of constants $\mathrm{C}_{\bar{\mathrm{E}}}, \mathrm{C}_{E|\bar{Z}|}$.

5.2 Main steps of the proof

Proposition 3 describes the form of the bandwidth selected by the RevDir procedure with high probability. Given this selection, Proposition 4 gives upper bounds on the bias and the deviation of the estimator $\bar{f}_{\hat{h}}(w)$.

Proposition 3. *The selected bandwidth belongs to $\mathcal{H}_{hp}$ with high probability. More precisely:*

$$\mathcal{E}_{hp} \subset \{\hat{h} \in \mathcal{H}_{hp}\} \tag{5.4}$$

and for n large enough:

$$\mathbb{P}\left(\mathcal{E}_{hp}^c\right) \leq 2e^{-(\log n)^{1+\frac{a-1}{2}}}. \tag{5.5}$$

Note in particular that with high probability the irrelevant components of the selected bandwidth are equal to h_{irr}.

Recall that $\bar{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$ is the bias of $\bar{f}_h(w)$.

Proposition 4. *The following upper bounds are satisfied for all $h \in \mathcal{H}_{hp}$, and any constants $A \in \mathbb{R}$ and $C_A > 0$:*

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} \left|\bar{B}_h\right| \leq r\, C_{\bar{B}} C_A{}^s \frac{(\log n)^{As}}{n^{\frac{s}{2s+r}}} + r \max\left(\frac{7C_\lambda}{4\beta^{\frac{d-r}{2}} C_A{}^{\frac{r}{2}}} \frac{(\log n)^{\frac{a-Ar}{2}}}{n^{\frac{s}{2s+r}}}, \frac{7}{4}\left(\frac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}}\right), \tag{5.6}$$

$$\begin{aligned}\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} \sigma_h \\ &\leq \max\left(\frac{C_\sigma}{\beta^{(d-r)/2} C_A{}^{r/2}} (\log n)^{(a-Ar)/2}, \frac{4C_A{}^s C_{E\bar{Z}} C_\sigma \beta^{-\frac{r}{2}-s}}{C_\lambda} (\log n)^{sA}\right) n^{-\frac{s}{2s+r}},\end{aligned} \tag{5.7}$$

where C_λ is the constant defined in (2.10) and $C_{\bar{B}}, C_\sigma, C_{E\bar{Z}}$ are constants defined in Lemmas 1 and 2 in Appendix.

5.3 Proof of Theorem 1

Let us fix $l > 1$. From Proposition 3: $\mathcal{E}_{\mathrm{hp}} \subset \{\hat{h} \in \mathcal{H}_{\mathrm{hp}}\}$, thus:

$$\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right] = \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\mathrm{hp}}^c}\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right] + \sum_{h\in\mathcal{H}_{\mathrm{hp}}} \mathbb{E}\left[\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathrm{hp}}}\left|\hat{f}_h(w) - f(w)\right|^l\right]. \tag{5.8}$$

We first control the terms $\mathbb{E}\left[\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathrm{hp}}}\left|\hat{f}_h(w) - f(w)\right|^l\right]$. We fix $h \in \mathcal{H}_{\mathrm{hp}}$. Then, we decompose the difference $\hat{f}_h(w) - f(w)$ as follows:

$$\hat{f}_h(w) - f(w) = \Delta_h + \left(\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right) + \bar{B}_h, \tag{5.9}$$

where we recall the notations $\Delta_h := \hat{f}_h(w) - \bar{f}_h(w)$ and $\bar{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$. Remark that $\prod_{k=1}^d h_k \leq 1$, since $h \in \mathcal{H}_{\rm hp}$. We apply 2. of Lemma 3 and 3. of Lemma 1. Since $\mathcal{E}_{\rm hp} \subset \left(\widetilde{\mathcal{A}}_n \cap \mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)\right) \cap \mathcal{B}\mathrm{ern}_{\bar{f}}(h)$:

$$\mathbb{1}_{\mathcal{E}_{\rm hp}} |\Delta_h| \leq \mathrm{C}_{\mathrm{M}\Delta}\sigma_h$$

and

$$\mathbb{1}_{\mathcal{E}_{\rm hp}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| \leq \sigma_h.$$

Therefore:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}} \left|\hat{f}_h(w) - f(w)\right| \leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}} \left((\mathrm{C}_{\mathrm{M}\Delta}+1)\sigma_h + \left|\bar{B}_h\right|\right). \tag{5.10}$$

From Proposition 4 which controls both σ_h and $\left|\bar{B}_h\right|$, we deduce:

$$\begin{aligned}
&\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}} \left|\hat{f}_h(w) - f(w)\right| \\
&\quad\leq (\mathrm{C}_{\mathrm{M}\Delta}+1)\max\left(\tfrac{\mathrm{C}_\sigma}{\beta^{(d-r)/2}{C_A}^{r/2}}(\log n)^{\frac{a-Ar}{2}}, \tfrac{4{C_A}^s \mathrm{C}_{E\bar{Z}}\mathrm{C}_\sigma \beta^{-\frac{r}{2}-s}}{\mathrm{C}_\lambda}(\log n)^{sA}\right) n^{-\frac{s}{2s+r}} \\
&\qquad + r\mathrm{C}_{\bar{B}}{C_A}^s (\log n)^{As} n^{-\frac{s}{2s+r}} + r\max\left(\tfrac{7\mathrm{C}_\lambda}{4\beta^{\frac{d-r}{2}}{C_A}^{\frac{r}{2}}}\tfrac{(\log n)^{\frac{a-Ar}{2}}}{n^{\frac{s}{2s+r}}}, \tfrac{7}{4}\left(\tfrac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}}\right).
\end{aligned}$$

We optimize in A and C_A: With $A = \frac{a}{2s+r}$, we obtain

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}} \left|\hat{f}_h(w) - f(w)\right| \leq \max\left(C_1\left(\tfrac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}}, \tfrac{7}{4}r\left(\tfrac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}}\right).$$

where C_1 depends on $\beta, d, r, s, \mathrm{C}_{\bar{B}}, \mathrm{C}_{E\bar{Z}}, \mathrm{C}_\sigma, \mathrm{C}_{\mathrm{M}\Delta}, \mathrm{C}_\lambda$. If $r = 0$, the last term in the right hand side vanishes, otherwise $p/(2p+1) \geq s/(2s+r)$ (since $p \geq s$). Therefore, for n large enough:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}} \left|\hat{f}_h(w) - f(w)\right| \leq \mathrm{C}'\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}}. \tag{5.11}$$

To prove the theorem, it then remains to control $\left|\hat{f}_h(w) - f(w)\right|$ on $\mathcal{E}_{\rm hp}^c$. Recall that:

$$\prod_{k=1}^d \hat{h}_k \geq \beta^r \frac{(\log n)^{1+a}}{n},$$

and Condition (i):

$$\tilde{\delta}_X := \inf_{u\in\,\mathcal{U}_1} \tilde{\mathrm{f}}_X(u) > n^{-1/2},$$

then we can roughly bound $\hat{f}_{\hat{h}}(w)$ by:

$$\left|\hat{f}_{\hat{h}}(w)\right| \leq \frac{\|K\|_\infty^d n}{\tilde{\delta}_X \beta^r (\log n)^{1+a}} = o(n^2).$$

So:

$$\left|\hat{f}_h(w) - f(w)\right|^l = o(n^{2l}) = o(e^{2l\log n}).$$

Besides, from Proposition 3:

$$\mathbb{P}\left(\mathcal{E}_{\text{hp}}^c\right) \leq 2e^{-(\log n)^{1+\frac{a-1}{2}}}.$$

Note that, since $a > 1$,

$$2l\log n + l\log(n^{\frac{1}{2}}) = o((\log n)^{1+\frac{a-1}{2}}), \tag{5.12}$$

therefore:

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\text{hp}}^c}\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right]^{1/l} \leq \left(\mathbb{P}\left(\mathcal{E}_{\text{hp}}^c\right)e^{2l\log n}\right)^{1/l} = o(n^{-\frac{1}{2}}).$$

To conclude, we combine Equation (5.8) with the above upper bound and Inequality (5.11):

$$\begin{aligned}\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right]^{1/l} &\leq o(n^{-\frac{1}{2}}) + \left\{\left(\mathrm{C}'\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}}\right)^l \sum_{h\in\mathcal{H}_{\text{hp}}}\mathbb{E}[\mathbb{1}_{\hat{h}=h}]\right\}^{1/l} \\ &\leq \mathrm{C}\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}},\end{aligned}$$

with C depending on $d, r, \|f\|_\infty, \mathcal{U}, \delta, L, s, K, \beta$.

5.4 Proof of Proposition 3

By definition of the procedure, any selected bandwidth $\hat{h}$ satisfies

$$\exists(t_1,\ldots,t_d) \in \mathbb{Z}^d, \forall k \in 1:d, \hat{h}_k = \beta^{t_k}h_0$$

The loop condition in the Reverse Step imposes for any active component k that at the beginning of an iteration $t \in \mathbb{Z}_-$:

$$\hat{h}_k^{(t)} \leq \beta.$$

At most, $\hat{h}_k^{(t)}$ is multiplied by β^{-1}. Then after the last update of the component $\hat{h}_k$:

$$\hat{h}_k \leq 1 = \beta^{-1}\beta.$$

Now let us prove that on $\mathcal{E}_{\text{hp}}$, the irrelevant components are deactivated at value h_{irr}. It suffices to show that during the initialization, the irrelevant components activate for Reverse Step, *i.e.*:

$$\mathcal{R}^c \subset \mathcal{A}\text{ct}^{(-1)},$$

and in the case where $h_0 \leq \beta$, it suffices to prove that they remain active at all iterations $t \in -1 : t_{\text{irr}}$. Remember that $t_{\text{irr}} \in \mathbb{Z}$ is defined such that: $h_{\text{irr}} = \beta^{t_{\text{irr}}}h_0$.

Note that if the irrelevant components remain active at all iteration $t \in -1 : t_{\text{irr}}$, then for $k \in \mathcal{R}^c$, $\hat{h}_k^{(t)} = H_k^{(t)} = \beta^t h_0$. It corresponds to the definition of $\mathcal{H}_{\text{hp}}$, since for all $h \in \mathcal{H}_{\text{hp}}$, $t \in -1 : t_{\text{irr}}$ and $k \in \mathcal{R}^c$,

$$h_k^{(t)} = \beta^t h_0.$$

Therefore, there exists $h \in \mathcal{H}_{\text{hp}}$ such that $\hat{h}^{(t)} = h^{(t)}$ for all iterations $t \in -1 : t_{\text{irr}}$.We will then prove that for any $h \in \mathcal{H}_{\text{hp}}$, $t \in -1 : t_{\text{irr}}$ and $k \in \mathcal{R}^c$,

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}}|Z_{h^{(t)}k}| \leq \lambda_{h^{(t)}k}.$$

Let us fix $h \in \mathcal{H}_{\text{hp}}$, $t \in -1 : t_{\text{irr}}$ and $k \in \mathcal{R}^c$. We decompose $Z_{h^{(t)}k}$ as follows:

$$Z_{h^{(t)}k} = \left(Z_{h^{(t)}k} - \bar{Z}_{h^{(t)}k}\right) + \left(\bar{Z}_{h^{(t)}k} - \mathbb{E}\bar{Z}_{h^{(t)}k}\right) + \mathbb{E}\bar{Z}_{h^{(t)}k}. \tag{5.13}$$

We use:

- 1. of Lemma 3: Recall the notation $\Delta_{Z,h^{(t)}k} := Z_{h^{(t)}k} - \bar{Z}_{h^{(t)}k}$, then remark that $\forall h' \in \mathcal{H}_{\text{hp}}^{\text{Rev}} \cup \mathcal{H}_{\text{hp}}^{\text{Dir}}$, $\prod_{k=1}^{d} h'_k \leq 1$, and $\mathcal{E}_{\text{hp}} \subset \mathcal{B}\text{ern}_{|\bar{Z}|}(h^{(t)}, k) \cap \widetilde{\mathcal{A}}_n$, therefore:

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}}\left(Z_{h^{(t)}k} - \bar{Z}_{h^{(t)}k}\right) \leq \frac{1}{4}\lambda_{h^{(t)}k},$$

- the definition of $\mathcal{B}\text{ern}_{\bar{Z}}(h^{(t)}, k)$: since $\mathcal{E}_{\text{hp}} \subset \mathcal{B}\text{ern}_{\bar{Z}}(h^{(t)}, k)$,

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}}\left|\bar{Z}_{h^{(t)}k} - \mathbb{E}\bar{Z}_{h^{(t)}k}\right| \leq \frac{1}{2}\lambda_{h^{(t)}k},$$

- 2. of Lemma 2: since $k \in \mathcal{R}^c$,

$$\mathbb{E}\bar{Z}_{h^{(t)}k} = 0.$$

Therefore:

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}}|Z_{h^{(t)}k}| \leq \frac{3}{4}\lambda_{h^{(t)}k} \leq \lambda_{h^{(t)}k},$$

and so, every irrelevant component is active during Reverse Step until Iteration t_{irr}. In particular, we have proved that:

$$\mathcal{E}_{\text{hp}} \subset \{\forall k \in \mathcal{R}^c : \hat{h}_k = h_{\text{irr}}\}.$$

Let us now prove that on $\mathcal{E}_{\text{hp}}$,

$$\prod_{k=1}^{d} \hat{h}_k \geq \beta^r \frac{(\log n)^{1+a}}{n}.$$

The loop condition in the Direct Step imposes that at the beginning of any iteration $t \geq 0$:

$$\prod_{k=1}^{d} \hat{h}_k^{(t)} \geq \frac{(\log n)^{1+a}}{n}.$$

For our algorithm, the bandwidth can only decrease during the Direct Step. Since on $\mathcal{E}_{\text{hp}}$, the irrelevant components are active the during Reverse Step, they are inactive during the Direct Step. This is the reason why during the last iteration, only relevant components could decrease and be multiplied by β. Therefore:

$$\prod_{k=1}^{d} \hat{h}_k \geq \beta^r \frac{(\log n)^{1+a}}{n},$$

which ends the proof of the inclusion (5.4) of Proposition 3.

Finally, we control $\mathbb{P}\left(\mathcal{E}_{\text{hp}}^c\right)$. We first control the cardinal of $\mathcal{H}_{\text{hp}}$ by enumerating the possible values for a component of a bandwidth in $\mathcal{H}_{\text{hp}}$. For $h \in \mathcal{H}_{\text{hp}}$ and $k \in \mathcal{R}$,

$$\beta(\log n)^{1+a} n^{-1} \leq h_k \leq 1,$$

thus:

$$|\{h_k : h \in \mathcal{H}_{\mathrm{hp}}\}| = \left|\left\{\beta^t h_0 \in [\beta(\log n)^{1+a} n^{-1}, 1], t \in \mathbb{Z}\right\}\right| \leq 1 + \log_{\frac{1}{\beta}} \left(\frac{1}{\beta(\log n)^{1+a} n^{-1}}\right) \leq \log_{\frac{1}{\beta}} n$$

(for n large enough). For $k \in \mathcal{R}^c$,

$$h_k = h_{\mathrm{irr}},$$

thus, we have

$$|\{h_k : h \in \mathcal{H}_{\mathrm{hp}}\}| = 1.$$

Therefore:

$$|\mathcal{H}_{\mathrm{hp}}| \leq \left(\log_{\frac{1}{\beta}} n\right)^r. \tag{5.14}$$

Let us also control the cardinal of $\mathcal{H}_{\mathrm{hp}}^{\mathrm{Rev}} \cup \mathcal{H}_{\mathrm{hp}}^{\mathrm{Dir}}$. The only supplementary bandwidths are the ones whose irrelevant components are smaller than h_{irr}. We consider the irrelevant components as the relevant ones, and we obtain the rough bound

$$\left|\mathcal{H}_{\mathrm{hp}}^{\mathrm{Rev}} \cup \mathcal{H}_{\mathrm{hp}}^{\mathrm{Dir}}\right| \leq \left(\log_{\frac{1}{\beta}} n\right)^d. \tag{5.15}$$

By Assumption $\boldsymbol{\mathcal{E}\mathbf{f}_{\boldsymbol{X}}}$, Condition (ii):

$$\mathbb{P}\left(\widetilde{\mathcal{A}}_n^c\right) \leq \exp(-(\log n)^{1+\frac{a-1}{2}}).$$

We bound the events $\mathcal{B}\mathrm{ern}_{\bar{f}}(h)^c$'s and $\mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)^c$'s using Lemma 1. Since for all $h \in \mathcal{H}_{\mathrm{hp}}$,

$$\prod_{k=1}^{d} h_k \geq \beta^r \frac{(\log n)^{a+1}}{n},$$

note that:

- $\mathrm{Cond}(h)$: $\prod_{k=1}^{d} h_k \geq \frac{4\|K\|_\infty^{2d}}{9\delta^2 \mathrm{C}_\sigma^2} \frac{(\log n)^a}{n}$ is satisfied for any $h \in \mathcal{H}_{\mathrm{hp}}$ for n large enough (when $\log n \geq \frac{4\|K\|_\infty^{2d}}{9\beta^r \delta^2 \mathrm{C}_\sigma^2}$). So, we have

$$\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{\bar{f}}(h)^c\right) \leq 2e^{-(\log n)^a}.$$

- Moreover,

$$\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)^c\right) \leq 2e^{-\mathrm{C}_{\gamma|f|} n \prod_{k=1}^{d} h_k} \leq 2e^{-\mathrm{C}_{\gamma|f|} \beta^r (\log n)^{a+1}}.$$

Similarly, we bound the probability of events $\mathcal{B}\mathrm{ern}_{\bar{Z}}(h)^c$'s and $\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h)^c$'s using Lemma 2. Note that for all $h \in \mathcal{H}_{\mathrm{hp}}^{\mathrm{Rev}} \cup \mathcal{H}_{\mathrm{hp}}^{\mathrm{Dir}}$:

- $\mathrm{Cond}_{\bar{Z}}(h)$: $\prod_{k=1}^{d} h_k \geq \mathrm{cond}_{\bar{Z}} \frac{(\log n)^a}{n}$ is satisfied for n large enough (when $\log n \geq \frac{\mathrm{cond}_{\bar{Z}}}{\beta^r}$). So, we have

$$\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{\bar{Z}}(h, j)^c\right) \leq 2e^{-\frac{\delta}{\|f\|_\infty, \mathcal{U}} (\log n)^a}.$$

- Moreover,

$$\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h,j)^c\right) \leq 2e^{-\mathrm{C}_{\gamma|\bar{Z}|}n\prod_{k=1}^d h_k} \leq 2e^{-\mathrm{C}_{\gamma|\bar{Z}|}\beta^r(\log n)^{a+1}}.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_{\mathrm{hp}}^c\right) &\leq \mathbb{P}\left(\widetilde{\mathcal{A}}_n^c\right) + \sum_{h\in\mathcal{H}_{\mathrm{hp}}}\left(\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{\bar{f}}(h)^c\right) + \mathbb{P}\left(\mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)^c\right)\right) \\
&\quad + \sum_{h\in(\mathcal{H}_{\mathrm{hp}}^{\mathrm{Rev}}\cup\mathcal{H}_{\mathrm{hp}}^{\mathrm{Dir}})}\sum_{k=1}^d\left(\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{\bar{Z}}(h,k)^c\right) + \mathbb{P}\left(\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h,k)^c\right)\right) \\
&\leq e^{-(\log n)^{1+\frac{a-1}{2}}} + \sum_{h\in\mathcal{H}_{\mathrm{hp}}}\left(2e^{-(\log n)^a} + 2e^{-\mathrm{C}_{\gamma|f|}\beta^r(\log n)^{a+1}}\right) \\
&\quad + \sum_{h\in(\mathcal{H}_{\mathrm{hp}}^{\mathrm{Rev}}\cup\mathcal{H}_{\mathrm{hp}}^{\mathrm{Dir}})}\sum_{k=1}^d\left(2e^{-\frac{\delta}{\|f\|_\infty,\,\mathcal{U}}(\log n)^a} + 2e^{-\mathrm{C}_{\gamma|\bar{Z}|}\beta^r(\log n)^{a+1}}\right) \\
&\leq e^{-(\log n)^{1+\frac{a-1}{2}}}\left(1 + 4\left(\log_{\frac{1}{\beta}} n\right)^r e^{-(\log n)^{\frac{a-1}{2}}} + 4d\left(\log_{\frac{1}{\beta}} n\right)^d e^{-\frac{\delta}{\|f\|_\infty,\,\mathcal{U}}(\log n)^{\frac{a-1}{2}}}\right) \\
&\leq 2e^{-(\log n)^{1+\frac{a-1}{2}}},
\end{aligned}$$

for n large enough.

5.5 Proof of Proposition 4

We fix $h \in \mathcal{H}_{\mathrm{hp}}$ and consider the event $\{\hat{h} = h\} \cap \mathcal{E}_{\mathrm{hp}}$. Let $(t_1, \ldots, t_d) \in \mathbb{Z}^d$ such that for all $k \in 1:d$,

$$h_k = \beta^{t_k} h_0.$$

Given positive constants A and C_A (to be opimized), we call $C_A\,(\log n)^A\, n^{-\frac{1}{2s+r}}$ the minimax bandwidth level and we define $t(A, C_A) \in \mathbb{R}$ such that

$$\beta^{t(A,C_A)} h_0 = C_A\,(\log n)^A\, n^{-\frac{1}{2s+r}}.$$

Using the definition (3.1) of h_0, observe that $t(A, C_A) > 0$ (for n large enough). To simplify the notation (permutation of the labels), we consider:

$$\mathcal{R} = 1:r$$

and

$$t_1 \geq t_2 \geq \cdots \geq t_r. \tag{5.16}$$

5.5.1 Proof of Inequality (5.6)

The bias of $\bar{f}_h(w)$ is denoted $\bar{B}_h$. Note that it does not depend on $\{h_k\}_{k\in\mathcal{R}^c}$. Indeed, we have

$$\begin{aligned}
\bar{B}_h :&= \mathbb{E}\left[\bar{f}_h(w)\right] - f(w) \\
&= \int_{u\in\mathbb{R}^d} \mathrm{K}_h(w-u)\frac{\mathrm{f}_W(u)}{\mathrm{f}_X(u_{1:d_1})}du - f(w) \\
&= \int_{u\in\mathbb{R}^d} \mathrm{K}_h(w-u)f(u)du - f(w) \\
&= \int_{z\in\mathbb{R}^d} \left(\prod_{k=1}^{d} K(z_k)\right)\left[f(w-h\cdot z) - f(w)\right]dz \qquad (5.17)\\
&= \int_{z'\in\mathbb{R}^r} \left(\prod_{k=1}^{r} K(z'_k)\right)\left[f_{\mathcal{R}}\left(w_{1:r} - h_{1:r}\cdot z'\right) - f_{\mathcal{R}}(w_{1:r})\right]dz'.
\end{aligned}$$

We consider the following disjunction of cases:

(Case A) without relevant component: $\mathcal{R} = \emptyset$

(Case B) with small relevant bandwidth components: $\min_{j\in\mathcal{R}} t_j \geq t(A, C_A)$

(Case C) with at least one large relevant bandwidth component: $\exists j \in \mathcal{R}, t_j < t(A, C_A)$.

Then we control the bias in each case.

(Case A) Assume $\mathcal{R} = \emptyset$. In particular, f is constant on the neighborhood $\mathcal{U}$. Note that for any $z \in \mathrm{supp}\,(K)^d$, $w - h\cdot z \in \mathcal{U}$. We then derive from Equation (5.17):

$$\bar{B}_h = 0.$$

(Case B) Assume $\min_{j\in\mathcal{R}} t_j \geq t(A, C_A)$. We apply 2. of Lemma 1

$$\begin{aligned}
\left|\bar{B}_h\right| &\leq \mathrm{C}_{\bar{B}} \sum_{j\in\mathcal{R}} h_j^s = \mathrm{C}_{\bar{B}} \sum_{j\in\mathcal{R}} \left(\beta^{t_j} h_0\right)^s \\
&\leq \mathrm{C}_{\bar{B}} \times r \left(\beta^{t(A,C_A)} h_0\right)^s = r\mathrm{C}_{\bar{B}} {C_A}^s \left(\log n\right)^{As} n^{-\frac{s}{2s+r}}
\end{aligned}$$

(Case C) Assume $\exists j \in \mathcal{R}, t_j < t(A, C_A)$. Then we consider

$$j_A = \min\left(j \in \mathcal{R} : t_j < t(A, C_A)\right).$$

In particular, for all $j \geq j_A$, the bandwidth components are larger than the minimax level:

$$h_j \geq C_A (\log n)^A n^{-\frac{1}{2s+r}}. \qquad (5.18)$$

For the previously fixed bandwidth h (and its relevant deactivation times $(t_1, \ldots, t_r)$), we define the following intermediate bandwidths $h^{(\mathrm{int},t)}$, $t \in \mathbb{R}$:

$$h_k^{(\mathrm{int},t)} = \begin{cases} \beta^{t\vee t_k} h_0 & \text{if } k \in \mathcal{R} \\ h_k & \text{else.} \end{cases}$$

Then we decompose the bias by splitting $f(w-h\cdot z)-f(w)$ (note that $h^{(\mathrm{int},t_r)}=h$):

$$\begin{aligned}\bar{B}_h = \int_{z\in\mathbb{R}^d}\left(\prod_{k=1}^d K(z_k)\right)[&f(w-h^{(\mathrm{int},t(A,C_A))}\cdot z)-f(w)\\ &+f(w-h^{(\mathrm{int},t_{j_A})}\cdot z)-f(w-h^{(\mathrm{int},t(A,C_A))}\cdot z)\\ &+\sum_{j_0=j_A+1}^{r} f(w-h^{(\mathrm{int},t_{j_0})}\cdot z)-f(w-h^{(\mathrm{int},t_{j_0-1})}\cdot z)]dz\\ = \bar{B}_{h^{(\mathrm{int},t(A,C_A))}} + (\bar{B}_{h^{(\mathrm{int},t_{j_A})}} - \bar{B}_{h^{(\mathrm{int},t(A,C_A))}}) &+ \sum_{j_0=j_A+1}^{r}\left(\bar{B}_{h^{(\mathrm{int},t_{j_0})}} - \bar{B}_{h^{(\mathrm{int},t_{j_0-1})}}\right).\end{aligned} \tag{5.19}$$

For the first term, note that $h^{(\mathrm{int},t(A,C_A))}$ satisfies the condition of (Case B), thus:

$$\left|\bar{B}_{h^{(\mathrm{int},t(A,C_A))}}\right| \le r\mathrm{C}_{\bar{B}}{C_A}^s\,(\log n)^{As}\,n^{-\frac{s}{2s+r}}. \tag{5.20}$$

Let us now control the other terms. The same arguments are used to control the terms in the sum $\bar{B}_{h^{(\mathrm{int},t_{j_0})}}-\bar{B}_{h^{(\mathrm{int},t_{j_0-1})}}$ (for $j_0\in(j_A+1):r$) and the second term $\bar{B}_{h^{(\mathrm{int},t_{j_A})}}-\bar{B}_{h^{(\mathrm{int},t(A,C_A))}}$. To shorten the proof, the followings lines are also applied to control the second term: for the added case $j_0=j_A$, one just has to replace $h_j^{(\mathrm{int},t_{j_0-1})}$ by $h^{(\mathrm{int},t(A,C_A))}$.

Let us now fix $j_0\in j_A:r$ and consider the path between $h_j^{(\mathrm{int},t_{j_0-1})}$ and $h_j^{(\mathrm{int},t_{j_0})}$. Namely for $u\in[0,1]$, we denote $h^{[j_0,u]}:=h^{(\mathrm{int},t_{j_0-1})}+u\left(h^{(\mathrm{int},t_{j_0})}-h^{(\mathrm{int},t_{j_0-1})}\right)$. Remark that, for any $j\in 1:d$,

$$\begin{aligned}h_j^{(\mathrm{int},t_{j_0})}-h_j^{(\mathrm{int},t_{j_0-1})}\neq 0 &\Rightarrow \left(j\in\mathcal{R}\text{ and }\beta^{t_{j_0}\vee t_j}\neq\beta^{(t_{j_0-1})\vee t_j}\right)\\ &\Rightarrow (j\in\mathcal{R}\text{ and }t_j<t_{j_0}\text{ or }t_j<t_{j_0-1})\\ &\Rightarrow (j\in\mathcal{R}\text{ and }t_j\le t_{j_0}).\end{aligned}$$

The last implication is due to the fact that a component could not be deactived between the consecutive deactivation times t_{j_0} and t_{j_0-1}.

Then, we introduce the function $g:u\in[0,1]\mapsto f(w-h^{[j_0,u]}\cdot z)$ (for a fixed $z\in\mathbb{R}^d$). In particular, using the above remark:

$$g'(u)=\sum_{\substack{j\in\mathcal{R}\\ t_j\le t_{j_0}}}\left(h_j^{(\mathrm{int},t_{j_0})}-h_j^{(\mathrm{int},t_{j_0-1})}\right)\times z_j\partial_j f(w-h^{[j_0,u]}\cdot z).$$

Then we write:

$$\begin{aligned}f(w-h^{(\mathrm{int},t_{j_0})}\cdot z)-&f(w-h^{(\mathrm{int},t_{j_0-1})}\cdot z)\\ &= g(1)-g(0)=\int_{u=0}^1 g'(u)du\\ &=\sum_{\substack{j\in\mathcal{R}\\ t_j\le t_{j_0}}}\int_{u=0}^1\left(h_j^{(\mathrm{int},t_{j_0})}-h_j^{(\mathrm{int},t_{j_0-1})}\right)\times z_j\partial_j f(w-h^{[j_0,u]}\cdot z)du.\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
\bar{B}_{h^{(\mathrm{int},t_{j_0})}} - \bar{B}_{h^{(\mathrm{int},t_{j_0-1})}} &= \int_{z\in\mathbb{R}^d} \left(\prod_{k=1}^{d} K(z_k)\right) [f(w - h^{(\mathrm{int},t_{j_0})}\cdot z) - f(w - h^{(\mathrm{int},t_{j_0-1})}\cdot z)]dz \\
&= \sum_{\substack{j\in\mathcal{R}\\ t_j\le t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \int_{z\in\mathbb{R}^d} \left(\prod_{k=1}^{d} K(z_k)\right) z_j \partial_j f(w - h^{[j_0,u]}\cdot z)dz\ du \\
&= \sum_{\substack{j\in\mathcal{R}\\ t_j\le t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right] du, \qquad (5.21)
\end{aligned}$$

using Equation (6.4):

$$\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right] = \int_{\mathbb{R}^d} \left(\prod_{k=1}^{d} K(z_k)\right) z_j \partial_j f(w - h^{[j_0,u]}\cdot z)dz.$$

Now the idea is to control $\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right|$ with the test at the iteration t_j on $|Z_{h^{(t_j)},j}|$. More precisely, we will first apply Assumption $\boldsymbol{\mathcal{C}}$ to move from $\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right|$ to $\left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$. Then, we will apply Bernstein's inequality to convert the control on $\left|Z_{h^{(t_j)},j}\right|$ to a control on $\left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$.
Let us fix $j \in \mathcal{R}$ such that $t_j \le t_{j_0}$. We distinguish the cases where the component j is deactivated during the Reverse Step or when it happens during the Direct Step.

Subcase (C.a) $t_j \ge 0$, *i.e.*: j is deactivated during the Direct Step.
Let us show $h^{[j_0,u]} \preccurlyeq h^{(t_j)}$:

- for $k \in \mathcal{R}^c$, since $h_k^{(\mathrm{int},t_{j_0-1})} = h_k = h_k^{(\mathrm{int},t_{j_0})}$,

$$h_k^{[j_0,u]} = h_k.$$

  Remember that the irrelevant components deactivate during the Reverse Step, therefore they already have their final value during the Direct Step. Formally, since $t_k < 0 \le t_j$, we have

$$h_k^{[j_0,u]} = h_k = \beta^{t_k} h_0 = \beta^{t_j \wedge t_k} h_0 = h_k^{(t_j)}.$$

- for $k \in \mathcal{R}$, notice $h^{(\mathrm{int},t_{j_0-1})} \preccurlyeq h^{(\mathrm{int},t_{j_0})}$. Therefore:

$$\begin{aligned}
h_k^{[j_0,u]} &\le h_k^{(\mathrm{int},t_{j_0})} = \beta^{t_{j_0}\vee t_k} h_0 \\
&\le \beta^{t_j\wedge t_k} h_0 = h_k^{(t_j)}.
\end{aligned}$$

Then, we have proved $h^{[j_0,u]} \preccurlyeq h^{(t_j)}$. Using Assumption $\boldsymbol{\mathcal{C}}$:

$$\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| \le \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|.$$

Subcase (C.b) $t_j < 0$, *i.e.*: j is deactivated during Reverse Step.
As well as $h' \mapsto \bar{B}_{h'}$, $h' \mapsto \mathbb{E}\left[\bar{Z}_{h',j}\right]$ is independent of the irrelevant components of the bandwidth (see for instance Equation (6.4)).
Then we modify the irrelevant components of $h^{[j_0,u]}$ and use the value of the irrelevant components of $h^{(t_j)}$. Formally, we introduce the notation $h^{\{j_0,u\}}$ such that

$$h_k^{\{j_0,u\}} = \begin{cases} h_k^{[j_0,u]} & \text{if } k \in \mathcal{R} \\ h_k^{(t_j)} & \text{else,} \end{cases}$$

so that:

$$\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right] = \mathbb{E}\left[\bar{Z}_{h^{\{j_0,u\}},j}\right].$$

Now we just have to verify $h^{\{j_0,u\}} \preccurlyeq h^{(t_j)}$:

- for $k \in \mathcal{R}^c$, by definition of $h^{\{j_0,u\}}$:

$$h_k^{\{j_0,u\}} = h_k^{(t_j)}$$

- for $k \in \mathcal{R}$,

$$\begin{aligned} h_k^{\{j_0,u\}} &= h_k^{[j_0,u]} \\ &\le h_k^{(\mathrm{int},t_{j_0})} = \beta^{t_{j_0} \vee t_k} h_0 \\ &\le \beta^{t_j \vee t_k} h_0, \text{ since } t_j \le t_{j_0}, \\ &\le \max\left(h_k, \beta^{t_j} h_0\right) =: h_k^{(t_j)}. \end{aligned}$$

Then we have proved $h^{\{j_0,u\}} \preccurlyeq h^{(t_j)}$. Using Assumption $\boldsymbol{\mathcal{C}}$:

$$\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| = \left|\mathbb{E}\left[\bar{Z}_{h^{\{j_0,u\}},j}\right]\right| \le \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|.$$

In each case (C.a and C.b), we have proved $\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| \le \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$, then we apply this inequality in Equation (5.21):

$$\left|\bar{B}_{h^{(\mathrm{int},t_{j_0})}} - \bar{B}_{h^{(\mathrm{int},t_{j_0-1})}}\right| \le \sum_{\substack{j \in \mathcal{R} \\ t_j \le t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| du \quad (5.22)$$

$$\le \sum_{\substack{j \in \mathcal{R} \\ t_j \le t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| du$$

$$\le \sum_{\substack{j \in \mathcal{R} \\ t_j \le t_{j_0}}} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|. \quad (5.23)$$

Then, the previous decomposition of the bias (5.19) leads to:

$$\begin{aligned}
\left|\bar{B}_h\right| &\leq \left|\bar{B}_{h^{(\mathrm{int},t(A,C_A))}}\right| + \sum_{j_0=j_A}^{r} \left|\bar{B}_{h^{(\mathrm{int},t_{j_0})}} - \bar{B}_{h^{(\mathrm{int},t_{j_0-1})}}\right| \\
&\leq r\mathrm{C}_{\bar{B}} {C_A}^s \left(\log n\right)^{As} n^{-\frac{s}{2s+r}} + \sum_{j_0=j_A}^{r} \sum_{\substack{j\in\mathcal{R} \\ t_j\leq t_{j_0}}} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \\
&\leq r\mathrm{C}_{\bar{B}} {C_A}^s \left(\log n\right)^{As} n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \sum_{j_0=j_A}^{j} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \\
&\leq r\mathrm{C}_{\bar{B}} {C_A}^s \left(\log n\right)^{As} n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| h_j^{(t_j)}, \qquad (5.24)
\end{aligned}$$

since the sum is telescoping, and by noticing that: $h_j^{(\mathrm{int},t_j)} = h_j^{(t_j)}$.

Now, it remains to control $\left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$ for $j \in j_A : r$ using the test at the iteration t_j on $Z_{h^{(t_j)},j}$:

$$\begin{aligned}
\mathbb{1}_{\mathcal{E}_{\mathrm{hp}}\cap\{\hat{h}=h\}} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \leq \mathbb{1}_{\hat{h}=h} \left|Z_{h^{(t_j)},j}\right| + \mathbb{1}_{\widetilde{\mathcal{A}}_n\cap\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h^{(t_j)},j)} \left|Z_{h^{(t_j)},j} - \bar{Z}_{h^{(t_j)},j}\right| \\
+ \mathbb{1}_{\mathcal{B}\mathrm{ern}_{\bar{Z}}(h^{(t_j)},j)} \left|\bar{Z}_{h^{(t_j)},j} - \mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|
\end{aligned}$$

By construction of the CDRodeo procedure, if $\hat{h} = h$, then j is deactivated at iteration t_j, in other words:

$$\mathbb{1}_{\mathcal{E}_{\mathrm{hp}}\cap\{\hat{h}=h\}} \left|Z_{h^{(t_j)},j}\right| \leq \lambda_{h^{(t_j)},j}.$$

We also apply:

- the definition of $\mathcal{B}\mathrm{ern}_{\bar{Z}}(h^{(t_j)}, j)$:

$$\mathbb{1}_{\mathcal{B}\mathrm{ern}_{\bar{Z}}(h^{(t_j)},j)} \left|\bar{Z}_{h^{(t_j)},j} - \mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \leq \frac{1}{2}\lambda_{h^{(t_j)},j},$$

- 1. of Lemma 3 (note in particular $\prod_{k=1}^{d} h_k^{(t_j)} \leq 1$):

$$\mathbb{1}_{\widetilde{\mathcal{A}}_n\cap\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h^{(t_j)},j)} \left|Z_{h^{(t_j)},j} - \bar{Z}_{h^{(t_j)},j}\right| = \mathbb{1}_{\widetilde{\mathcal{A}}_n\cap\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h^{(t_j)},j)} \left|\Delta_{Z,h^{(t_j)}j}\right| \leq \frac{1}{4}\lambda_{h^{(t_j)},j}.$$

Therefore:

$$\mathbb{1}_{\mathcal{E}_{\mathrm{hp}}\cap\{\hat{h}=h\}} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \leq \mathbb{1}_{\mathcal{E}_{\mathrm{hp}}\cap\{\hat{h}=h\}} \frac{7}{4}\lambda_{h^{(t_j)},j}.$$

Hence:

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}\cap\{\hat{h}=h\}}\left|\bar{B}_h\right| \leq \mathbb{1}_{\{\hat{h}=h\}}\left(r\mathrm{C}_{\bar{B}}{C_A}^s\left(\log n\right)^{As}n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r}\frac{7}{4}\lambda_{h^{(t_j)},j}\times h_j^{(t_j)}\right),$$
$$\leq \mathbb{1}_{\{\hat{h}=h\}}\left(r\mathrm{C}_{\bar{B}}{C_A}^s\left(\log n\right)^{As}n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r}\frac{7\mathrm{C}_\lambda(\log n)^{a/2}}{4\left(n\prod_{k=1}^{d}h_k^{(t_j)}\right)^{1/2}}\right). \tag{5.25}$$

Then we control $\prod_{k=1}^{d}h_k^{(t_j)}$ using the same disjunction of subcases as above:

Subcase (C.a) $t_j \geq 0$. At the iteration $t_j \geq 0$, the Direct Step has begun, thus the Reverse Step is over. Since $h \in \mathcal{H}_{\text{hp}}$, the irrelevant components have already their final value: for all $k \in \mathcal{R}^c$,

$$1 \geq h_k^{(t_j)} = h_k = h_{\text{irr}} > \beta.$$

Moreover, during the Direct Step, at iteration t_j, all components are lower bounded by the current active bandwidth value $\beta^{t_j}h_0$, *i.e.*: for any $k \in \mathcal{R}$,

$$h_k^{(t_j)} \geq \beta^{t_j}h_0.$$

Recall that $j \geq j_A$, thus:

$$t_j \leq t_{j_A} \leq t(A, C_A).$$

It follows:

$$h_k^{(t_j)} \geq \beta^{t(A,C_A)}h_0 = C_A\left(\log n\right)^A n^{-\frac{1}{2s+r}}.$$

Therefore:

$$\prod_{k=1}^{d}h_k^{(t_j)} \geq \beta^{d-r}\left(C_A\left(\log n\right)^A n^{-\frac{1}{2s+r}}\right)^r.$$

Then the upper bound in Equation (5.25) becomes:

$$\frac{7\mathrm{C}_\lambda(\log n)^{a/2}}{4\left(n\prod_{k=1}^{d}h_k^{(t_j)}\right)^{1/2}} \leq \frac{7\mathrm{C}_\lambda}{4\beta^{\frac{d-r}{2}}{C_A}^{\frac{r}{2}}}\left(\log n\right)^{\frac{a-Ar}{2}}n^{-\frac{1}{2}\left(1-\frac{r}{2s+r}\right)}$$
$$= \frac{7\mathrm{C}_\lambda}{4\beta^{\frac{d-r}{2}}{C_A}^{\frac{r}{2}}}\left(\log n\right)^{\frac{a-Ar}{2}}n^{-\frac{s}{2s+r}}.$$

Subcase (C.b) $t_j < 0$. At iteration t_j, only iterations of the Reverse Step have been performed. Thus, the current bandwidth has only been increased. Therefore:

$$\frac{7\mathrm{C}_\lambda(\log n)^{a/2}}{4\left(n\prod_{k=1}^{d}h_k^{(t_j)}\right)^{1/2}} \leq \frac{7\mathrm{C}_\lambda(\log n)^{a/2}}{4\left(nh_0^d\right)^{1/2}}.$$

Remark that the lower bound on h_0 (3.1) is exactly defined so, we have

$$\frac{7\mathrm{C}_\lambda(\log n)^{a/2}}{4\left(nh_0^d\right)^{1/2}} \leq \frac{7}{4}\left(\frac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}}.$$

Note that $n^{-\frac{p}{2p+1}}$ is smaller than the minimax optimal rate for any regularity and any sparsity structure (except for the degenerate case where $r = 0$ and which is solved separately: cf (Case A)):

$$n^{-\frac{p}{2p+1}} = \min_{\substack{1\le r'\le d\\ 1\le s'\le p}} \left(n^{-\frac{s'}{2s'+r'}} \right).$$

When we reunite the two subcases, Inequality (5.25) becomes:

$$\mathbb{1}_{\mathcal{E}_{\rm hp}\cap\{\hat{h}=h\}} \left|\bar{B}_h\right| \le r\mathrm{C}_{\bar{B}} {C_A}^s (\log n)^{As}\, n^{-\frac{s}{2s+r}} + r \times \max\left(\frac{7\mathrm{C}_\lambda}{4\beta^{\frac{d-r}{2}} {C_A}^{\frac{r}{2}}} \frac{(\log n)^{\frac{a-Ar}{2}}}{n^{\frac{s}{2s+r}}}, \frac{7}{4}\left(\frac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}} \right),$$

which concludes the proof of Inequality (5.6)

5.5.2 Proof of Inequality (5.7)

Let us now prove the second inequality (5.7). By definition: $\mathcal{E}_{\rm hp} \subset \mathcal{B}\mathrm{ern}_{\bar{f}}(h)$. Thus, we have

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| \le \sigma_h := \mathrm{C}_\sigma \sqrt{\frac{(\log n)^a}{n\prod_{k=1}^d h_k}}.$$

Two cases occur: in the first case, the deviation is controlled by a concentration inequality; in the second case, we control the deviation by $\mathbb{E}Z_{hj}$ thanks to the tests on the Z_{hj}'s.

1. $\max_{k\in\mathcal{R}} t_k \le t(A, C_A)$. Then, $\forall k \in \mathcal{R}$:

$$h_k = \beta^{t_k} h_0 > \beta^{t(A,C_A)} h_0 = C_A (\log n)^A n^{-\frac{1}{2s+r}}.$$

Besides, for $k \in \mathcal{R}^c$:

$$h_k = h_{\rm irr} > \beta.$$

Therefore:

$$\sigma_h \le \mathrm{C}_\sigma \sqrt{\frac{(\log n)^a}{n\beta^{d-r}\left(C_A(\log n)^A n^{-\frac{1}{2s+r}}\right)^r}} = \frac{\mathrm{C}_\sigma}{\beta^{(d-r)/2}{C_A}^{r/2}} (\log n)^{(a-Ar)/2} n^{-\frac{s}{2s+r}}.$$

2. $\max_{k\in\mathcal{R}} t_k > t(A, C_A)$. First remark that for any $k \in 1:d$,

$$\sigma_h = \frac{\mathrm{C}_\sigma}{\mathrm{C}_\lambda} h_k\, \lambda_{hk}.$$

Hence, it suffices to control the threshold in order to bound the deviation. Let us consider $j_0 \in \arg\max_{k\in\mathcal{R}} t_k$ (actually assuming (5.16) means that $j_0 = 1$). In particular, when $\hat{h} = h$, the component j_0 is deactivated during the last iteration, and during the Direct

Step (recall that $t(A, C_A) > 0$). Let us consider the penultimate iteration, i.e. Iteration $t_{j_0} - 1$. At this iteration, j_0 is not deactivated, *i.e.*:

$$\mathbb{1}_{\hat{h}=h} \left| Z_{h^{(t_{j_0}-1)}j_0} \right| > \mathbb{1}_{\hat{h}=h} \lambda_{h^{(t_{j_0}-1)}j_0}.$$

Then we use 1. of Lemma 3. Note that $\prod_{k=1}^{d} h_k^{(t_{j_0}-1)} \leq 1$, thus:

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}} \left| \Delta_{Z,h^{(t_{j_0}-1)}j_0} \right| \leq \frac{1}{4} \lambda_{h^{(t_{j_0}-1)}j_0}.$$

Remember the definition of $\mathcal{B}\text{ern}_{\bar{Z}}(h, j)$, thus

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}} \left| \bar{Z}_{h^{(t_{j_0}-1)}j_0} - \mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}j_0} \right] \right| \leq \frac{1}{2} \lambda_{h^{(t_{j_0}-1)}j_0}.$$

Therefore:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \left| \mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}j_0} \right] \right| > \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \frac{1}{4} \lambda_{h^{(t_{j_0}-1)}j_0}. \tag{5.26}$$

Let us compare $h^{(t_{j_0}-1)}$ to h. Recall $h = h^{(t_{j_0})}$, since t_{j_0} is the final iteration of our algorithm. We have:

- for $k \in \mathcal{R}^c$, $h_k^{(t_{j_0}-1)} = h_k$. Indeed, $t_k < 0$, hence the components k have been deactivated before Iteration $t_{j_0} - 1$, and have the same value for the last two iterations.
- for $k \in \mathcal{R}$, $h_k \geq \beta h_k^{(t_{j_0}-1)}$. Indeed, at worst, the component k was active during Iteration $t_{j_0} - 1$ and have been multiplied by β.

Therefore:

$$\prod_{k=1}^{d} h_k \geq \beta^r \prod_{k=1}^{d} h_k^{(t_{j_0}-1)}$$

and

$$h_{j_0} \lambda_{hj_0} = \mathrm{C}_\lambda \sqrt{\frac{(\log n)^a}{n \prod_{k=1}^{d} h_k}} \leq \beta^{-\frac{r}{2}} h_{j_0}^{(t_{j_0}-1)} \lambda_{h^{(t_{j_0}-1)}j_0}.$$

To summarize, we have

$$\begin{aligned}
\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \left| \bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w) \right] \right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \sigma_h = \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \frac{\mathrm{C}_\sigma}{\mathrm{C}_\lambda} h_{j_0} \lambda_{hj_0} \\
&\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \beta^{-\frac{r}{2}} \frac{\mathrm{C}_\sigma}{\mathrm{C}_\lambda} h_{j_0}^{(t_{j_0}-1)} \lambda_{h^{(t_{j_0}-1)}j_0} \\
&\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} 4\beta^{-\frac{r}{2}} \frac{\mathrm{C}_\sigma}{\mathrm{C}_\lambda} h_{j_0}^{(t_{j_0}-1)} \left| \mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}j_0} \right] \right|.
\end{aligned}$$

Then we apply 2. of Lemma 2:

$$\left| \mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}j_0} \right] \right| \leq \mathrm{C}_{E\bar{Z}} \left(h_{j_0}^{(t_{j_0}-1)} \right)^{s-1}.$$

Therefore:

$$\begin{aligned}
\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathrm{hp}}}\left|\bar{f}_h(w)-\mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathrm{hp}}} 4\beta^{-\frac{r}{2}}\frac{\mathrm{C}_\sigma}{\mathrm{C}_\lambda}h_{j_0}^{(t_{j_0}-1)}\times \mathrm{C}_{E\bar{Z}}\left(h_{j_0}^{(t_{j_0}-1)}\right)^{s-1}\\
&\leq \frac{4\mathrm{C}_{E\bar{Z}}\mathrm{C}_\sigma\beta^{-\frac{r}{2}}}{\mathrm{C}_\lambda}\left(\beta^{t_{j_0}-1}h_0\right)^s = \frac{4\mathrm{C}_{E\bar{Z}}\mathrm{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathrm{C}_\lambda}\left(\beta^{t_{j_0}}h_0\right)^s\\
&\leq \frac{4\mathrm{C}_{E\bar{Z}}\mathrm{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathrm{C}_\lambda}\left(\beta^{t(A,C_A)}h_0\right)^s = \frac{4\mathrm{C}_{E\bar{Z}}\mathrm{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathrm{C}_\lambda}\left(C_A(\log n)^A n^{-\frac{1}{2s+r}}\right)^s\\
&= \frac{4{C_A}^s\mathrm{C}_{E\bar{Z}}\mathrm{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathrm{C}_\lambda}(\log n)^{sA}n^{-\frac{s}{2s+r}}.
\end{aligned}$$

Reuniting the two cases, we obtain Inequality (5.7):

$$\begin{aligned}
\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathrm{hp}}}\left|\bar{f}_h(w)-\mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathrm{hp}}}\sigma_h\\
&\leq \max\left(\frac{\mathrm{C}_\sigma}{\beta^{(d-r)/2}{C_A}^{r/2}}(\log n)^{(a-Ar)/2}, \frac{4{C_A}^s\mathrm{C}_{E\bar{Z}}\mathrm{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathrm{C}_\lambda}(\log n)^{sA}\right)n^{-\frac{s}{2s+r}}.
\end{aligned}$$

5.6 Proof of Proposition 2

Let us evaluate the number of operations of our procedure. During the Reverse Step, each bandwidth of $\mathcal{A}\mathrm{ct}^{(-1)}$ can be multiplied by β^{-1} several times until the loop condition is achieved:

$$(\mathcal{A}\mathrm{ct}^{(t)}\neq\emptyset)\&(\max\hat{h}_k^{(t)}\leq\beta).$$

In particular, $\max\hat{h}_k^{(t)}\leq 1$. Since $\hat{h}_k^{(t)}=h_0\beta^{-|t_k|}$,

$$|t_k| = \log\left(\frac{\hat{h}_k^{(t)}}{h_0}\right)/\log\left(\beta^{-1}\right)\leq\frac{\log(h_0^{-1})}{\log(\beta^{-1})} = O\left(\frac{\log(n)}{d(2p+1)}\right)$$

using the lower bound on h_0 (3.1). Thus, during this Reverse Step, note that only $|\mathcal{A}\mathrm{ct}^{(-1)}|$ components are updated and:

- the number of updates of the Z_{hj}'s is of order $\frac{|\mathcal{A}\mathrm{ct}^{(-1)}|}{d(2p+1)}\log(n)$ given the above remark,
- the computation of the Z_{hj}'s and the comparison to the threshold cost $\mathcal{O}(|\mathcal{A}\mathrm{ct}^{(-1)}|n)$ operations.

Therefore at worst, there are $\mathcal{O}\left(\frac{|\mathcal{A}\mathrm{ct}^{(-1)}|^2}{d}\log(n)n\right)$ operation during the Reverse Step.

For the Direct Step, the stopping condition is $\left(\prod_{k=1}^d\hat{h}_k^{(t)}>\frac{(\log n)^{1+a}}{n}\right)$, which is satisfied for the penultimate iteration, hence:

$$\prod_{k=1}^d\hat{h}_k>\beta^d\frac{(\log n)^{1+a}}{n}.$$

We denote t_k the deactivation times of $\hat{h}$, then

$$h_0^d\beta^{\sum_{k=1}^d t_k}>\beta^d\frac{(\log n)^{1+a}}{n},$$

which gives

$$\sum_{k=1}^{d} t_k < \frac{\log(\beta^{-d}(\log n)^{-(1+a)} n h_0^d)}{\log(1/\beta)}.$$

Thus, during the Direct Step, note that only $|\mathcal{A}\text{ct}^{(0)}|$ components are updated and

- the total number of updates of the Z_{hj}'s is of order $\log_{\frac{1}{\beta}}(n)$ given the above remark,
- the computation of the Z_{hj}'s and the comparison to the threshold cost $\mathcal{O}(|\mathcal{A}\text{ct}^{(0)}|n)$ operations.

Therefore at worst, there are $\mathcal{O}(|\mathcal{A}\text{ct}^{(-1)}|\log(n)n)$ operations during the Direct Step. Using $|\mathcal{A}\text{ct}^{(-1)}| + |\mathcal{A}\text{ct}^{(0)}| \leq d$, the sum of these two steps leads to the proposition.

6 Appendix

6.1 Lemmas

The following lemmas are mainly proved in Nguyen (2018). Note that some adjustments have been made from their initial versions. In particular, we have refined points 2. of Lemma 1 and of Lemma 2 to take into account the extension of our results to Hölder smoothness. In the sequel, we only prove results of subsequent lemmas which were not established in Nguyen (2018).

Lemma 1 (Lemma 5 of Nguyen (2018): $\bar{f}_h(w)$ behaviour)**.** *Under Assumption $\boldsymbol{\mathcal{L}_X}$, for any bandwidth $h \in (0,1]^d$, and any $i \in 1:n$,*

1. *Let $C_{\bar{E}} := \|f\|_{\infty,\,\mathcal{U}} \|K\|_1^d$. Then*

$$\left|\mathbb{E}\bar{f}_{h1}(w)\right| \leq \mathbb{E}\left|\bar{f}_{h1}(w)\right| \leq C_{\bar{E}}.$$

2. *If f has only r relevant components $\mathcal{R}$ and belongs to $\mathcal{H}_d(s,L)$ and if the order p of the kernel K is larger than or equal to s,*

$$\left|\bar{B}_h\right| \leq C_{\bar{B}} \sum_{k\in\mathcal{R}} h_k^s, \tag{6.1}$$

with $C_{\bar{B}} > 0$ a constant only depending on L, s and K.

3. *Let $\mathcal{B}ern_{\bar{f}}(h) := \{|\bar{f}_h(w) - \mathbb{E}[\bar{f}_h(w)]| \leq \sigma_h\}$, where $\sigma_h := C_\sigma \sqrt{\frac{(\log n)^a}{n \prod_{k=1}^{d} h_k}}$ with $C_\sigma = \frac{2\|K\|_2^d \|f\|_{\infty,\,\mathcal{U}}^{\frac{1}{2}}}{\delta^{\frac{1}{2}}}$. If Cond(h): $\prod_{k=1}^{d} h_k \geq \frac{4\|K\|_\infty^{2d}}{9\delta^2 C_\sigma^2} \frac{(\log n)^a}{n}$ is satisfied, then:*

$$\mathbb{P}\left(\mathcal{B}ern_{\bar{f}}(h)^c\right) \leq 2e^{-(\log n)^a}.$$

4. *Let $\mathcal{B}ern_{|\bar{f}|}(h) := \{\left|\frac{1}{n}\sum_{i=1}^{n} |\bar{f}_{hi}(w)| - \mathbb{E}[|\bar{f}_h(w)|]\right| \leq C_{\bar{E}}\}$. Then*

$$\mathbb{P}\left(\mathcal{B}ern_{|\bar{f}|}(h)^c\right) \leq 2e^{-C_{\gamma|f|} n \prod_{k=1}^{d} h_k},$$

with $C_{\gamma|f|} := \min\left(\frac{C_{\bar{E}}{}^2}{C_\sigma{}^2}; \frac{3\delta C_{\bar{E}}}{4\|K\|_\infty^d}\right)$.

Lemma 2 (Lemma 6 of Nguyen (2018): $\bar{Z}_{hj}$ behaviour)**.** *If K is chosen as in Section 3.1, and under Assumption $\boldsymbol{\mathcal{L}_X}$, for any $j \in 1, \dots, d$ and any bandwidth $h \in (0, h_0]^d$, we have the following results.*

1. *Let $C_{E|\bar{Z}|} := \|f\|_{\infty,\, \mathcal{U}} \|J\|_1 \|K\|_1^{d-1}$. We have*
$$\mathbb{E}|\bar{Z}_{h1j}| \leq C_{E|\bar{Z}|} h_j^{-1}.$$
2. *If f has only r relevant components $\mathcal{R}$, for $j \notin \mathcal{R}$:*
$$\mathbb{E}\bar{Z}_{hj} = 0,$$
and if in addition f belongs to $\mathcal{H}_d(s, L)$, for $j \in \mathcal{R}$:
$$|\mathbb{E}[\bar{Z}_{h,j}]| \leq C_{E\bar{Z}} h_j^{s-1}, \tag{6.2}$$
where $C_{E\bar{Z}} := \left(\int |z^s K(z)| dz\right) \frac{\|K\|_1^{r-1} L}{(s-1)!}$ denoting $(s-1)! := (s-q+1)(s-q+2)\dots(s-1)$.
3. *Let $\mathcal{B}ern_{\bar{Z}}(h, j) := \{|\bar{Z}_{hj} - \mathbb{E}\bar{Z}_{hj}| \leq \frac{1}{2}\lambda_{hj}\}$. If the bandwidth satisfies:*

$Cond_{\bar{Z}}(h)$: $\prod_{k=1}^d h_k \geq cond_{\bar{Z}} \frac{(\log n)^a}{n}$, *with* $cond_{\bar{Z}} := \frac{4\|J\|_\infty^2 \|K\|_\infty^{2(d-1)}}{3^2 \|f\|_{\infty,\, \mathcal{U}} \|J\|_2^2 \|K\|_2^{2(d-1)}}$,

then:
$$\mathbb{P}\left(\mathcal{B}ern_{\bar{Z}}(h, j)^c\right) \leq 2e^{-\frac{\delta}{\|f\|_{\infty,\, \mathcal{U}}} (\log n)^a}.$$
4. *Let $\mathcal{B}ern_{|\bar{Z}|}(h, j) := \{|\frac{1}{n}\sum_{i=1}^n |\bar{Z}_{hij}| - \mathbb{E}|\bar{Z}_{h1j}|| \leq C_{E|\bar{Z}|} h_j^{-1}\}$. Then,*
$$\mathbb{P}\left(\mathcal{B}ern_{|\bar{Z}|}(h, j)^c\right) \leq 2e^{-C_{\gamma|\bar{Z}|} n \prod_{k=1}^d h_k},$$
with $C_{\gamma|\bar{Z}|} := \min\left(\frac{\delta C_{E|\bar{Z}|}^2}{4\|f\|_{\infty,\, \mathcal{U}} \|J\|_2^2 \|K\|_2^{2(d-1)}}; \frac{3\delta C_{E|\bar{Z}|}}{4\|K\|_\infty^{d-1} \|J\|_\infty}\right)$.

Lemma 3. *For any $h \in \mathcal{H}_{hp}^{Rev} \cup \mathcal{H}_{hp}^{Dir}$ and any component $j \in 1:d$, under Assumptions $\boldsymbol{\mathcal{L}_X}$ and $\boldsymbol{\mathcal{E}\mathbf{f}_X}$, if $\sqrt{\prod_{k=1}^d h_k} \leq 1$, then*

1. *we have:*
$$\mathbb{1}_{\mathcal{B}ern_{|\bar{Z}|}(hj) \cap \widetilde{\mathcal{A}}_n} |\Delta_{Z,hj}| \leq \frac{1}{4}\lambda_{hj}$$
2. *for $C_{M\Delta} := \frac{4M_X C_{\bar{E}}}{\delta C_\sigma}$:*
$$\mathbb{1}_{\widetilde{\mathcal{A}}_n \cap \mathcal{B}ern_{|\bar{f}|}(h)} |\Delta_h| \leq C_{M\Delta} \sigma_h.$$

Lemma 4 (Taylor's theorem)**.** *Let $g : [0, 1] \to \mathbb{R}$ be a function of class C^q. Then we have:*
$$g(1) - g(0) = \sum_{l=1}^q \frac{g^{(l)}(0)}{l!} + \int_{t_1=0}^1 \int_{t_2=0}^{t_1} \cdots \int_{t_q=0}^{t_{q-1}} (g^{(q)}(t_q) - g^{(q)}(0)) dt_q dt_{q-1} \dots dt_1.$$

6.2 Proof of Inequality (6.1) in Lemma 1

We recall that the notation $\cdot$ means the multiplication term by term of two vectors, then we have:

$$\bar{B}_h = \mathbb{E}\bar{f}_h(w) - f(w) = \int_{u\in\mathbb{R}^d} \left(\prod_{k=1}^d \frac{K(h_k^{-1}(w_k-u_k))}{h_k}\right) f(u)du - f(w)$$
$$= \int_{z\in\mathbb{R}^d} \left(\prod_{k=1}^d K(z_k)\right)(f(w-h\cdot z)-f(w))dz.$$

For any $z\in\mathbb{R}^d$, let us introduce the notations $\overline{z}_0 := w$ and for $k\in 1,\dots,d$, $\overline{z}_k := w - \sum_{j=1}^k h_j z_j e_j$, where $\{e_j\}_{j=1}^d$ is the canonical basis of $\mathbb{R}^d$. Then, we write:

$$f(w-h.z)-f(w) = \sum_{k=1}^d f(\overline{z}_k) - f(\overline{z}_{k-1}) = \sum_{k\in\mathcal{R}} f(\overline{z}_k) - f(\overline{z}_{k-1}),$$

since for $k\notin\mathcal{R}$, $f(\overline{z}_k)-f(\overline{z}_{k-1})=0$. We apply Taylor's theorem (cf Lemma 4) to the functions $g_k : t\in[0,1]\mapsto f(\overline{z}_{k-1}-th_kz_ke_k)$, $k\in\mathcal{R}$:

$$f(\overline{z}_k)-f(\overline{z}_{k-1}) = g_k(1)-g_k(0) = \sum_{l=1}^q \frac{(\text{-}z_kh_k)^l}{l!}\partial_k^l f(\overline{z}_{k-1}) + J_k,$$

where we recall that q is the largest integer smaller than s and with

$$J_k := \int_{0\le t_q\le\cdots\le t_1\le 1} \left(g_k^{(q)}(t_q) - g_k^{(q)}(0)\right) dt_{1:q}$$
$$= (\text{-}h_kz_k)^q \int_{0\le t_q\le\cdots\le t_1\le 1} \left(\partial_k^q f(\overline{z}_{k-1}-t_qh_kz_ke_k) - \partial_k^q f(\overline{z}_{k-1})\right) dt_{1:q}.$$

We denote $\mathrm{I}_k := \int_{z\in\mathbb{R}^d}\left(\prod_{k'=1}^d K(z_{k'})\right)J_kdz$ and for any $z\in\mathbb{R}^d$, we denote $z_{-k}\in\mathbb{R}^{d-1}$ the vector z without its k^{th} variable, then we obtain:

$$\bar{B}_h = \sum_{k\in\mathcal{R}}\int_{z\in\mathbb{R}^d}\left(\prod_{k'=1}^d K(z_{k'})\right)\left(J_k + \sum_{l=1}^q \frac{(\text{-}h_k)^l}{l!}\partial_k^l f(\overline{z}_{k-1})z_k^l\right)dz$$
$$= \sum_{k\in\mathcal{R}}\left(\mathrm{I}_k + \sum_{l=1}^q \mathrm{II}_{k,l}\right),$$

where

$$\mathrm{II}_{k,l} := \int_{z_{-k}\in\mathbb{R}^{d-1}}\left(\prod_{k'\neq k} K(z_{k'})\right)\frac{(\text{-}h_k)^l}{l!}\partial_k^l f(\overline{z}_{k-1})\int_{z_k\in\mathbb{R}} z_k^l K(z_k)dz_kdz_{-k}$$
$$= \frac{(\text{-}h_k)^l}{l!}\int_{z_{-k}\in\mathbb{R}^{d-1}}\partial_k^l f(\overline{z}_{k-1})\left(\prod_{k'\neq k}K(z_{k'})\right)dz_{-k}\times\int_{t\in\mathbb{R}} t^lK(t)dt = 0,$$

since K is of order $p \geq s > q$. So,

$$\bar{B}_h = \sum_{k \in \mathcal{R}} \mathrm{I}_k.$$

Now we control $|J_k|$:

$$\begin{aligned} |J_k| &\leq |h_k z_k|^q \left| \int_{0 \leq t_q \leq \cdots \leq t_1 \leq 1} \left[\partial_k^q f(\overline{z}_{k-1} - t_q h_k z_k e_k) - \partial_k^q f(\overline{z}_{k-1}) \right] dt_{1:q} \right| \\ &\leq |h_k z_k|^q \int_{0 \leq t_q \leq \cdots \leq t_1 \leq 1} L |t_q h_k z_k|^{s-q} dt_{1:q} = \frac{L(h_k |z_k|)^s}{s(s-1)\ldots(s-q)}. \end{aligned}$$

So:

$$|\mathrm{I}_k| = \left| \int_{z \in \mathbb{R}^d} \left(\prod_{k'=1}^{d} K(z_{k'}) \right) J_k dz \right| \leq \frac{L \|K\|_1^{d-1} \|(\cdot)^s K(\cdot)\|_1}{s(s-1)\ldots(s-q)} {h_k}^s.$$

Finally,

$$\left| \bar{B}_h \right| \leq \mathrm{C}_{\bar{B}} \sum_{k \in \mathcal{R}} h_k^s, \tag{6.3}$$

with $\mathrm{C}_{\bar{B}} := \frac{L \|K\|_1^{d-1} \|(\cdot)^s K(\cdot)\|_1}{s(s-1)\ldots(s-q)}$.

6.3 Proof of Inequality (6.2) in Lemma 2

Let $j \in \mathcal{R}$. Denoting $J : \mathbb{R} \to \mathbb{R}$ the function $t \mapsto tK'(t) + K(t)$, we can write

$$\bar{Z}_{h,j} = \frac{1}{n} \sum_{i=1}^{n} \frac{-J(\frac{w_j - W_{ij}}{h_j}) \prod_{k \neq j} K(\frac{w_k - W_{ik}}{h_k})}{\mathrm{f}_X(X_i) h_j \prod_{k=1}^d h_k}.$$

Then, taking the expectation,

$$\mathbb{E}[\bar{Z}_{hj}] = -\frac{1}{h_j} \int_{\mathbb{R}^d} J(z_j) \left(\prod_{k \neq j} K(z_k) \right) f(w - h \cdot z) dz.$$

To simplify the notations, we assume $\mathcal{R} = 1 : r$. Then, by integration by part

$$\begin{aligned} \mathbb{E}[\bar{Z}_{h,j}] &= \int_{\mathbb{R}^d} (z_j K(z_j)) \left(\prod_{k \neq j} K(z_k) \right) \partial_j f(w - h \cdot z) dz \\ &= \int_{\mathbb{R}^r} \left(\prod_{k \in \mathcal{R}} K(z_k) \right) z_j \partial_j f_{\mathcal{R}}(w_{1:r} - (h.z)_{1:r}) dz_{1:r}, \end{aligned} \tag{6.4}$$

where $f_{\mathcal{R}}$ is the restriction of f to the first r components (remember that for any $u \in \mathbb{R}^r$ and any $v \in \mathbb{R}^{d-r}$ $f_{\mathcal{R}}(u) := f_{\mathcal{R}}(u, v)$ does not depend on v). Let us denote by $G_{j,z,h} : [0,1] \to \mathbb{R}$ the function

$$t \mapsto \partial_j f_{\mathcal{R}}(w_1 - h_1 z_1, \ldots, w_j - t h_j z_j, \ldots, w_r - h_r z_r).$$

Then

$$\mathbb{E}[\bar{Z}_{h,j}] = \int_{\mathbb{R}^r} \left(\prod_{k \in \mathcal{R}} K(z_k) \right) z_j G_{j,z,h}(1) dz_{1:r}$$
$$= \int_{\mathbb{R}^r} \left(\prod_{k \in \mathcal{R}} K(z_k) \right) z_j \{ G_{j,z,h}(1) - G_{j,z,h}(0) \} dz_{1:r},$$

since the order p of K satisfies: $p \geq s > q \geq 1$. Next we use the Taylor expansion given by Lemma 4:

$$G_{j,z,h}(1) - G_{j,z,h}(0) = \sum_{l=1}^{q-1} \frac{G^{(l)}_{j,z,h}(0)}{l!} + R'_{j,z,h,q-1}, \tag{6.5}$$

where $R'_{j,z,h,q-1} := \int_{t_1=0}^{1} \int_{t_2=0}^{t_1} \cdots \int_{t_{q-1}=0}^{t_{q-2}} (G^{(q-1)}_{j,z,h}(t_{q-1}) - G^{(q-1)}_{j,z,h}(0)) dt_{q-1} dt_{q-2} \ldots dt_1$. But

$$G^{(l)}_{j,z,h}(t) = (-h_j z_j)^l \partial_j^{l+1} f_{\mathcal{R}}(w_1 - h_1 z_1, \ldots, w_j - t h_j z_j, \ldots, w_r - h_r z_r).$$

Then, the first $q-1$ terms in the r.h.s. of (6.5) vanish since $\int z_j^{l+1} K(z_j) dz_j = 0$. Now, we will bound the integral remainder of (6.5). Using that f belongs to $\mathcal{H}_d(s, L)$, for all $t \in [0,1]$,

$$\left| G^{(q-1)}_{j,z,h}(t) - G^{(q-1)}_{j,z,h}(0) \right| \leq |h_j z_j|^{q-1} L |t h_j z_j|^{s-q},$$

since $w - h \cdot z + (1-t) h_j z_j e_j \in \mathcal{U}$. Hence

$$|R'_{j,z,h,q-1}| \leq \int_{t_1=0}^{1} \int_{t_2=0}^{t_1} \cdots \int_{t_{q-1}=0}^{t_{q-2}} \left| G^{(q-1)}_{j,z,h}(t_{q-1}) - G^{(q-1)}_{j,z,h}(0) \right| dt_{q-1} dt_{q-2} \ldots dt_1$$
$$\leq L (h_j |z_j|)^{s-1} \int_{t_1=0}^{1} \int_{t_2=0}^{t_1} \cdots \int_{t_{q-1}=0}^{t_{q-2}} t_{q-1}^{s-q} dt_{q-1} dt_{q-2} \ldots dt_1 = \frac{L (h_j |z_j|)^{s-1}}{(s-1)!},$$

denoting $(s-1)! := (s-q+1)(s-q+2)\ldots(s-1)$. Finally,

$$|\mathbb{E}[\bar{Z}_{h,j}]| = \left| \int_{\mathbb{R}^r} \left(\prod_{k \in \mathcal{R}} K(z_k) \right) z_j R'_{j,z,h,q-1} dz_{1:r} \right| \leq \int_{\mathbb{R}^r} \left(\prod_{k \in \mathcal{R}} |K(z_k)| \right) |z_j| \frac{L (h_j |z_j|)^{s-1}}{(s-1)!} dz_{1:r}$$
$$\leq \frac{L h_j^{s-1}}{(s-1)!} \left(\prod_{k \in \mathcal{R} \setminus \{j\}} \|K\|_1 \right) \int_{\mathbb{R}} |z_j|^s |K(z_j)| dz_{1:r} \leq \mathrm{C}_{E\bar{Z}} h_j^{s-1},$$

denoting $\mathrm{C}_{E\bar{Z}} := \left(\int_{\mathbb{R}} |z|^s |K(z)| dz \right) \|K\|_1^{r-1} L/(s-1)!$.

6.4 Proof of Lemma 3

Before establishing the upper bounds, let us control $\mathbb{1}_{\widetilde{\mathcal{A}}_n} \left\| \frac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X} \right\|_{\infty,\, \mathcal{U}_1}$. First, using Assumption $\boldsymbol{\mathcal{L}_X}$:

$$\delta := \inf_{u \in \mathcal{U}_1} \mathrm{f}_X(u) > 0,$$

remark that: for any $u \in \mathcal{U}_1$,

$$\begin{aligned}
\mathbb{1}_{\widetilde{\mathcal{A}}_n} \tilde{\mathrm{f}}_X(u) &\geq \mathbb{1}_{\widetilde{\mathcal{A}}_n} \left(\mathrm{f}_X(u) - \|\mathrm{f}_X - \tilde{\mathrm{f}}_X\|_{\infty,\mathcal{U}_1} \right) \\
&\geq \mathbb{1}_{\widetilde{\mathcal{A}}_n} \left(\delta - M_X \frac{(\log n)^{\frac{a}{2}}}{\sqrt{n}} \right) \quad \text{by Condition (ii)}, \\
&\geq \mathbb{1}_{\widetilde{\mathcal{A}}_n} \frac{\delta}{2} \quad \text{(for } n \text{ large enough)}.
\end{aligned}$$

Therefore:

$$\tilde{\delta}_X := \inf_{u \in \mathcal{U}_1} \tilde{\mathrm{f}}_X(u) \geq \mathbb{1}_{\widetilde{\mathcal{A}}_n} \frac{\delta}{2},$$

which leads to:

$$\begin{aligned}
\mathbb{1}_{\widetilde{\mathcal{A}}_n} \left\| \frac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X} \right\|_{\infty,\, \mathcal{U}_1} &\leq \mathbb{1}_{\widetilde{\mathcal{A}}_n} \frac{\left\| \mathrm{f}_X - \tilde{\mathrm{f}}_X \right\|_{\infty,\, \mathcal{U}_1}}{\tilde{\delta}_X} \\
&\leq \frac{2M_X}{\delta} \frac{(\log n)^{a/2}}{n^{1/2}}. \qquad (6.6)
\end{aligned}$$

Let us now prove the first upper bound.

1. We still denote, for any bandwidth h, any component k and any observation i,

$$\bar{Z}_{hik} := \frac{\partial}{\partial h_k} \left(\frac{\mathrm{K}_h(w - W_i)}{\mathrm{f}_X(X_i)} \right),$$

such that $\bar{Z}_{hk} = \frac{1}{n} \sum_{i=1}^{n} \bar{Z}_{hik}$, with $\{\bar{Z}_{hik}\}_{i=1}^n$ i.i.d.. Then we can write:

$$\Delta_{Z,hk} := Z_{hk} - \bar{Z}_{hk} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathrm{f}_X}{\tilde{\mathrm{f}}_X}(X_i) - 1 \right) \bar{Z}_{hik} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X}(X_i) \right) \bar{Z}_{hik}.$$

Note that since K is compactly supported, if $X_i \notin \mathcal{U}_1$,

$$\bar{Z}_{hik} = 0.$$

Hence:

$$\begin{aligned}
|\Delta_{Z,hk}| &\leq \left\| \tfrac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X} \right\|_{\infty,\, \mathcal{U}_1} \times \frac{1}{n} \sum_{i=1}^{n} |\bar{Z}_{hik}| \\
&\leq \left\| \tfrac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X} \right\|_{\infty,\, \mathcal{U}_1} \times \left(\mathbb{E}\left[|\bar{Z}_{h1k}| \right] + \frac{1}{n} \sum_{i=1}^{n} |\bar{Z}_{hik}| - \mathbb{E}\left[|\bar{Z}_{hik}| \right] \right).
\end{aligned}$$

Using the above Inequality (6.6) and the upper bounds 1. and 4. of Lemma 2:

$$\begin{aligned}
\mathbb{1}_{\widetilde{\mathcal{A}}_n \cap \mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h,k)} |\Delta_{Z,hk}| &\leq \left(\frac{2M_X}{\delta} \frac{(\log n)^{a/2}}{n^{1/2}} \right) \times 2\mathrm{C}_{E|\bar{Z}|} h_k^{-1} \\
&\leq \frac{1}{4} \lambda_{h,k} := \frac{\mathrm{C}_\lambda}{4} \frac{(\log n)^{a/2}}{n^{1/2} h_k \left(\prod_{k'=1}^{d} h_{k'} \right)^{1/2}},
\end{aligned}$$

if $\left(\prod_{k'=1}^{d} h_{k'}\right)^{1/2} \leq \frac{\delta \mathrm{C}_\lambda}{16 M_X \mathrm{C}_{E|\bar{Z}|}}$. Note that M_X is determined in order to satisfy:

$$\frac{\delta \mathrm{C}_\lambda}{16 M_X \mathrm{C}_{E|\bar{Z}|}} = 1.$$

Hence the condition on the bandwidth becomes:

$$\left(\prod_{k'=1}^{d} h_{k'}\right)^{1/2} \leq 1.$$

2. We still denote, for any bandwidth h and any observation i,

$$\bar{f}_{hi}(w) := \frac{\mathrm{K}_h(w - W_i)}{\mathrm{f}_X(X_i)},$$

such that $\bar{f}_h(w) = \frac{1}{n}\sum_{i=1}^{n} \bar{f}_{hi}(w)$, with $\{\bar{f}_{hi}(w)\}_{i=1}^n$ i.i.d. Then we can write:

$$\Delta_h := \hat{f}_h(w) - \bar{f}_h(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{f}_X}{\tilde{\mathrm{f}}_X}(X_i) - 1\right)\bar{f}_{hi}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X}(X_i)\right)\bar{f}_{hi}(w).$$

Note that since K is compactly supported, if $X_i \notin \mathcal{U}_1$,

$$\bar{f}_{hi}(w) = 0.$$

Hence:

$$\begin{aligned}
|\Delta_h| &\leq \left\| \frac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X} \right\|_{\infty,\, \mathcal{U}_1} \times \frac{1}{n}\sum_{i=1}^{n} |\bar{f}_{hi}(w)| \\
&\leq \left\| \frac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X} \right\|_{\infty,\, \mathcal{U}_1} \times \left(\mathbb{E}\left[|\bar{f}_{h1}(w)|\right] + \frac{1}{n}\sum_{i=1}^{n} |\bar{f}_{hi}(w)| - \mathbb{E}\left[|\bar{f}_{hi}(w)|\right] \right).
\end{aligned}$$

Using the above Inequality (6.6) and the upper bounds 1. and 4. of Lemma 1:

$$\begin{aligned}
\mathbb{1}_{\widetilde{\mathcal{A}}_n \cap \mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)} |\Delta_h| &\leq \left(\frac{2M_X}{\delta} \frac{(\log n)^{a/2}}{n^{1/2}} \right) \times 2\mathrm{C}_{\bar{\mathrm{E}}} \\
&= \frac{4 M_X \mathrm{C}_{\bar{\mathrm{E}}}}{\delta \mathrm{C}_\sigma} \sigma_h \left(\prod_{k'=1}^{d} h_{k'} \right)^{1/2} \leq \mathrm{C}_{\mathrm{M}\Delta} \sigma_h.
\end{aligned}$$

6.5 Proof of Proposition 1

The proof is very similar to the Proposition 1 of (Nguyen, 2018). The main modification is due to the tighter log exponent in Condition (ii) and the enlarged neighborhood $\mathcal{U}_1$ of x. We introduce the classical kernel density estimator $\tilde{\mathrm{f}}_X^{\mathcal{K}}$: for any $u \in \mathbb{R}^{d_1}$ and a bandwidth $h_X \in \mathbb{R}_+^*$ to be specified later,

$$\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) := \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \prod_{j=1}^{d_1} \mathcal{K}\left(\frac{u_j - \widetilde{X}_{ij}}{h_X} \right), \tag{6.7}$$

where $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ is a kernel which is compactly supported, of class C^1 and of order $p_X \geq \frac{d_1}{2(c-1)}$, where we recall that $c > 1$ is defined by $n_X = n^c$. We first show that there exists $C_X > 0$ such that for any $\xi > 0$:

$$\mathbb{P}\left(\|\mathrm{f}_X - \tilde{\mathrm{f}}_X^{\mathcal{K}}\|_{\infty,\, \mathcal{U}_1} > C_X \frac{(\log n)^{\frac{1+\xi}{2}}}{\sqrt{n}} \right) \leq \mathcal{O}\left(n_X^{d_1+1} \exp\left(-(\log n)^{1+\xi} \right) \right). \tag{6.8}$$

Then we set

$$\tilde{\mathrm{f}}_X \equiv \tilde{\mathrm{f}}_X^{\mathcal{K}} \vee n^{-\frac{1}{2}},$$

and we shall prove that this estimator satisfies Condition (i) and Condition (ii) for $\tilde{\mathrm{f}}_X$.

Let us prove Inequality (6.8). Let us first explicit $\tilde{\mathrm{f}}_X^{\mathcal{K}}$'s behaviour. Following Lemma 5 gives a pointwise concentration inequality and a control of the bias of $\tilde{\mathrm{f}}_X^{\mathcal{K}}$ on $\mathcal{U}_1$. We introduce an enlarged neighborhood of $\mathcal{U}_1$:

$$\mathcal{U}_1' := \left\{ u' = u - h_X z : u \in \mathcal{U}_1, z \in \mathrm{supp}(\mathcal{K}) \right\}.$$

Lemma 5 ($\tilde{\mathrm{f}}_X^{\mathcal{K}}$ behaviour)**.** *The estimator $\tilde{\mathrm{f}}_X^{\mathcal{K}}$ satisfies the following results:*

1. *If there exists $q_X \in \mathbb{N}$ such that f_X is C^{q_X} on $\mathcal{U}_1'$ and such that $\mathcal{K}$ has $q_X - 1$ zero moments, then there exists a positive constant $\mathrm{C}'_{\mathrm{bias}_X}$ such that*

$$\left\| \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} \leq \mathrm{C}'_{\mathrm{bias}_X} h_X^{q_X}.$$

2. *For any $\xi > 0$, any $u \in \mathcal{U}_1$ and any $\lambda > 0$ such that:*

$$4\mathrm{C}_{\mathrm{var}_X} \frac{(\log n)^{1+\xi}}{n_X h_X^{d_1}} \leq \lambda^2 \leq \frac{9{\mathrm{C}_{\mathrm{var}_X}}^2}{\|\mathcal{K}\|_\infty^{2d_1}},$$

where $\mathrm{C}_{\mathrm{var}_X} := \|\mathcal{K}\|_2^{d_1} \|\mathrm{f}_X\|_{\infty,\, \mathcal{U}_1'}^{\frac{1}{2}}$,

$$\mathbb{P}\left(\left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) \right| > \lambda \right) \leq 2 \exp\left(-(\log n)^{1+\xi} \right).$$

This lemma is proved in Section 6.6. We define $p'_X = \min(p', p_X)$, so that: f_X is of class $C^{p'_X}$ and the first $p'_X - 1$ moments of $\mathcal{K}$ vanish. Therefore, we can apply 1. of Lemma 5:

$$\left\| \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} \leq \mathrm{C}'_{\mathrm{bias}_X} h_X^{p'_X}.$$

Therefore:

$$\begin{aligned} \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} &\leq \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} \right\|_{\infty,\, \mathcal{U}_1} + \left\| \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} \\ &\leq \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} \right\|_{\infty,\, \mathcal{U}_1} + \mathrm{C}'_{\mathrm{bias}_X} h_X^{p'_X}, \end{aligned}$$

and we have for any threshold λ:

$$\mathbb{P}\left(\left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} \geq \lambda \right) \leq \mathbb{P}\left(\left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} \right\|_{\infty,\, \mathcal{U}_1} \geq \lambda - \mathrm{C}'_{\mathrm{bias}_X} h_X^{p'_X} \right). \tag{6.9}$$

We have then reduced the problem to a concentration inequality of $\tilde{\mathrm{f}}_X^{\mathcal{K}}$ in sup norm. In order to move from a supremum on $\mathcal{U}_1$ to a maximum on a finite set of elements of $\mathcal{U}_1$, let us construct an ϵ-net $\{u_{(l)}\}_l$ of $\mathcal{U}_1$, in the meaning that for any $u \in \mathcal{U}_1$, there exists l such that $\|u - u_{(l)}\|_\infty := \max_{k=1:d_1} |u_k - u_{(l)k}| \leq \epsilon$. We denote $A > 0$ such that:

$$\mathrm{supp}(\mathcal{K}) \cup \mathrm{supp}(K) \subset \left[-\tfrac{A}{2}, \tfrac{A}{2}\right].$$

Set $N(\epsilon)$ is the smallest integer such that $2\epsilon N(\epsilon) \geq A$, and for $l \in (1 : N(\epsilon))^{d_1}$, $u_{(l)}$ such that its j-th component is equal to:

$$u_{(l)j} := x_j - \frac{A}{2} + (2l_j - 1)\epsilon.$$

Then $\{u_{(l)}\}_{l \in (1:N(\epsilon))^{d_1}}$ is an ϵ-net of $\mathcal{U}_1$. Therefore in order to obtain Inequality (6.8), we only need to obtain the concentration inequality for each point of $\{u_{(l)} : l \in (1 : N(\epsilon))^{d_1}\}$ and to control the difference of the function $\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}$ evaluated at the point u and at the nearest point of u in the ϵ-net. More formally, we have to control the following supremum

$$\sup_{u \in \mathcal{U}_1} \min_{l \in (1:N(\epsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) + \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right|.$$

For this purpose, we obtain (from Taylor's Inequality): for any $u, v \in \mathbb{R}^{d_1}$,

$$\left| \prod_{k=1}^{d_1} \mathcal{K}(u_k) - \prod_{k=1}^{d_1} \mathcal{K}(v_k) \right| \leq d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1 - 1} \|u - v\|_\infty.$$

Therefore, for any $u, v \in \mathcal{U}_1$:

$$\begin{aligned} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(v) \right| &\leq \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \left| \prod_{k=1}^{d_1} \mathcal{K}(\tfrac{u_k - \widetilde{X}_{ik}}{h_X}) - \prod_{k=1}^{d_1} \mathcal{K}(\tfrac{v_k - \widetilde{X}_{ik}}{h_X}) \right| \\ &\leq d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1 - 1} \frac{\|u - v\|_\infty}{h_X^{d_1 + 1}}. \end{aligned}$$

Since $\{u_{(l)} : l \in (1 : N(\epsilon))^{d_1}\}$ is an ϵ-net of $\mathcal{U}_1$:

$$\sup_{u \in \mathcal{U}_1} \min_{l \in (1:N(\epsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \leq d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1 - 1} \frac{\epsilon}{h_X^{d_1 + 1}},$$

and also:

$$\sup_{u \in \mathcal{U}_1} \min_{l \in (1:N(\epsilon))^{d_1}} \left| \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \leq d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1 - 1} \frac{\epsilon}{h_X^{d_1 + 1}}.$$

Therefore:

$$\sup_{u \in \mathcal{U}_1} \min_{l \in (1:N(\epsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) + \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \leq 2 d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1 - 1} \frac{\epsilon}{h_X^{d_1 + 1}}.$$

We denote $\mathrm{C}_{\mathrm{diff}} := 2d_1\|\mathcal{K}'\|_\infty\|\mathcal{K}\|_\infty^{d_1-1}$. We then obtain the following inequality:

$$\begin{aligned}\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}\right\|_{\infty,\,\mathcal{U}_1} &\leq \max_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right| \\ &\quad + \sup_{u\in\,\mathcal{U}_1}\min_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) + \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right| \\ &\leq \max_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right| + \mathrm{C}_{\mathrm{diff}}\frac{\epsilon}{h_X^{d_1+1}}.\end{aligned}$$

Then the inequality (6.9) becomes: for any threshold λ,

$$\begin{aligned}\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda\right) &\leq \mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda - \mathrm{C}'_{\mathrm{bias}_X}h_X^{p'_X}\right) \\ &\leq \mathbb{P}\left(\max_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right| \geq \lambda - \mathrm{C}'_{\mathrm{bias}_X}h_X^{p'_X} - \mathrm{C}_{\mathrm{diff}}\frac{\epsilon}{h_X^{d_1+1}}\right) \\ &\leq N(\epsilon)^{d_1}\max_{l\in(1:N(\epsilon))^{d_1}}\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right| \geq \lambda - \mathrm{C}'_{\mathrm{bias}_X}h_X^{p'_X} - \mathrm{C}_{\mathrm{diff}}\frac{\epsilon}{h_X^{d_1+1}}\right).\end{aligned} \tag{6.10}$$

It then remains to apply 2. of Lemma 5 for each $u_{(l)}$, $l \in (1 : N(\epsilon))^{d_1}$. We set the following settings:

- $h_X := n_X^{-\frac{c-1}{c.d_1}}$;
- $\epsilon := h_X^{1+\frac{d_1}{2}}n_X^{-\frac{1}{2}}$;
- $\lambda := 2\lambda_X$, where λ_X is defined by:

$$\lambda_X := 2\sqrt{\mathrm{C}_{\mathrm{var}_X}}(\log n)^{\frac{1+\xi}{2}}h_X^{-\frac{d_1}{2}}n_X^{-\frac{1}{2}} = 2\sqrt{\mathrm{C}_{\mathrm{var}_X}}(\log n)^{\frac{1+\xi}{2}}n_X^{-\frac{1}{2c}},$$

  where we recall that $\mathrm{C}_{\mathrm{var}_X} := \|\mathcal{K}\|_2^{d_1}\|\mathrm{f}_X\|_{\infty,\,\mathcal{U}_1'}^{\frac{1}{2}}$.

In particular, since we take $p_X \geq \frac{d_1}{2(c-1)}$ and we assume $p' \geq \frac{d_1}{2(c-1)}$, then $p'_X = \min(p', p_X) \geq \frac{d_1}{2(c-1)}$. Hence we obtain for n large enough:

$$\begin{aligned}\mathrm{C}'_{\mathrm{bias}_X}h_X^{p'_X} &= \mathrm{C}'_{\mathrm{bias}_X}n_X^{-\frac{p'_X(c-1)}{c.d_1}} \\ &\leq \mathrm{C}'_{\mathrm{bias}_X}n_X^{-\frac{1}{2c}} \\ &\leq \frac{1}{2}\lambda_X = \sqrt{\mathrm{C}_{\mathrm{var}_X}}(\log n)^{\frac{1+\xi}{2}}n_X^{-\frac{1}{2c}}\end{aligned}$$

and also, since $c > 1$:

$$\begin{aligned}\mathrm{C}_{\mathrm{diff}}\frac{\epsilon}{h_X^{d_1+1}} &= \mathrm{C}_{\mathrm{diff}}h_X^{-\frac{d_1}{2}}n_X^{-\frac{1}{2}} = \mathrm{C}_{\mathrm{diff}}\ n_X^{-\frac{1}{2c}} \\ &\leq \frac{1}{2}\lambda_X = \sqrt{\mathrm{C}_{\mathrm{var}_X}}(\log n)^{\frac{1+\xi}{2}}n_X^{-\frac{1}{2c}}.\end{aligned}$$

Hence, we have

$$\lambda - \mathrm{C}'_{\mathrm{bias}_X} h_X^{p'_X} - \mathrm{C}_{\mathrm{diff}} \frac{\epsilon}{h_X^{d_1+1}} \geq \lambda_X,$$

and the inequality (6.10) becomes:

$$\mathbb{P}\left(\left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} \geq \lambda \right) \leq N(\epsilon)^{d_1} \max_{l \in (1:N(\epsilon))^{d_1}} \mathbb{P}\left(\left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \geq \lambda_X \right). \tag{6.11}$$

We apply 2. of Lemma 5: we verify (since $n_X = n^c$)

$$\begin{aligned} 4\mathrm{C}_{\mathrm{var}_X} \frac{(\log n)^{1+\xi}}{n_X h_X^{d_1}} = \lambda_X^2 &= 4\mathrm{C}_{\mathrm{var}_X} (\log n)^{1+\xi} n^{-1} \\ &\leq \frac{9{\mathrm{C}_{\mathrm{var}_X}}^2}{\|\mathcal{K}\|_\infty^{2d_1}}, \quad \text{(for } n \text{ large enough)}, \end{aligned}$$

then we obtain

$$\mathbb{P}\left(\left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| > \lambda_X \right) \leq 2 \exp\left(-(\log n)^{1+\xi} \right).$$

Thus the inequality (6.11) becomes:

$$\mathbb{P}\left(\left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} \geq \lambda \right) \leq 2N(\epsilon)^{d_1} \exp\left(-(\log n)^{1+\xi} \right). \tag{6.12}$$

Let us control $2N(\epsilon)^{d_1}$:

$$2N(\epsilon)^{d_1} = 2\left\lceil \frac{A}{2\epsilon} \right\rceil^{d_1} = 2\left\lceil \frac{A}{2h_X^{1+\frac{d_1}{2}} n_X^{-\frac{1}{2}}} \right\rceil^{d_1} = o\left(n_X^{d_1+1} \right)$$

Therefore, we have obtained the desired concentration inequality (6.8). Now we consider $\tilde{\mathrm{f}}_X \equiv \tilde{\mathrm{f}}_X^{\mathcal{K}} \vee n^{-1/2}$, therefore $\tilde{\mathrm{f}}_X$ satisfies Condition (i). Let us show it also satisfies Condition (ii), for n large enough. We first show:

$$\left\{ \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} < \lambda \right\} \quad \Rightarrow \quad \left\{ \left\| \tilde{\mathrm{f}}_X - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} < \lambda \right\}. \tag{6.13}$$

Assume that for any $u \in \mathcal{U}_1$, $\left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathrm{f}_X(u) \right| < \lambda$. Let us fix $u \in \mathcal{U}_1$. Three cases occurs:

(a) When $\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) \geq n^{-\frac{1}{2}}$, then $\tilde{\mathrm{f}}_X(u) := \tilde{\mathrm{f}}_X^{\mathcal{K}}(u)$, and obviously:

$$\left| \tilde{\mathrm{f}}_X(u) - \mathrm{f}_X(u) \right| < \lambda.$$

(b) When $\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) < n^{-\frac{1}{2}}$ and $\mathrm{f}_X(u) \geq n^{-\frac{1}{2}}$, then since $\tilde{\mathrm{f}}_X(u) = n^{-\frac{1}{2}} > \tilde{\mathrm{f}}_X^{\mathcal{K}}(u)$,

$$\left| \tilde{\mathrm{f}}_X(u) - \mathrm{f}_X(u) \right| \leq \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathrm{f}_X(u) \right| < \lambda.$$

(c) When $\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) < n^{-\frac{1}{2}}$ and $\mathrm{f}_X(u) < n^{-\frac{1}{2}}$, then $\tilde{\mathrm{f}}_X(u) = n^{-\frac{1}{2}}$, so for n large enough:

$$\left|\tilde{\mathrm{f}}_X(u) - \mathrm{f}_X(u)\right| \leq n^{-\frac{1}{2}} < \lambda.$$

Therefore these three cases show Implication (6.13), and thus, from Equation (6.12), we obtain:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda\right) \leq \mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda\right) \leq 2N(\epsilon)^{d_1} \exp\left(-(\log n)^{1+\xi}\right).$$

Now, to obtain Condition (ii), for ξ such that $1 + \frac{a-1}{2} < 1 + \xi < a$,

$$\lambda = 4\sqrt{\mathrm{C}_{\mathrm{var}_X}}(\log n)^{\frac{1+\xi}{2}} n^{-\frac{1}{2}} \leq M_X(\log n)^{\frac{a}{2}} n^{-\frac{1}{2}} \text{ (for } n \text{ large enough).} \tag{6.14}$$

Therefore:

$$\begin{aligned}\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} \geq M_X(\log n)^{\frac{a}{2}} n^{-\frac{1}{2}}\right) &\leq \mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda\right) \\ &\leq 2N(\epsilon)^{d_1} \exp\left(-(\log n)^{1+\xi}\right) \\ &\leq \exp\left(-(\log n)^{1+\frac{a-1}{2}}\right),\end{aligned}$$

that is Condition (ii).

6.6 Proof of Lemma 5

The result 1. of Lemma 5 is proved in Lemma 4 of Nguyen (2018). To prove 2. of Lemma 5, let us fix $\xi > 0$. Then, we simply apply Bernstein's Inequality (see Lemma 10 in Nguyen (2018)). We define for any $u \in \mathcal{U}_1$ and for $i \in 1:n$

$$\tilde{\mathrm{f}}_{X,i}^{K}(u) := \frac{1}{h_X^{d_1}} \prod_{j=1}^{d_1} \mathcal{K}\left(\frac{u_j - \widetilde{X}_{ij}}{h_X}\right).$$

Observe that the $\tilde{\mathrm{f}}_{X,i}^{K}(u)$'s are *i.i.d.* Then we pick up the following bounds from (Nguyen, 2018, p. 23):

$$\begin{aligned}\left|\tilde{\mathrm{f}}_{X,1}^{K}(u)\right| &\leq \mathrm{M}_{h_X} := \|\mathcal{K}\|_\infty^{d_1} h_X^{-d_1}. \\ \mathrm{Var}\left(\tilde{\mathrm{f}}_{X,1}^{K}(u)\right) &\leq \mathrm{v}_{h_X} := \mathrm{C}_{\mathrm{var}_X} h_X^{-d_1},\end{aligned}$$

(we recall $\mathrm{C}_{\mathrm{var}_X} := \|\mathcal{K}\|_2^{2d_1} \|\mathrm{f}_X\|_{\infty,\,\mathcal{U}_1'}$). Therefore: for any $\lambda > 0$,

$$\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u)\right| > \lambda\right) \leq 2\exp\left(-\min\left(\frac{n_X\lambda^2}{4\mathrm{v}_{h_X}}, \frac{3n_X\lambda}{4\mathrm{M}_{h_X}}\right)\right).$$

Let us show that when

$$4\mathrm{C}_{\mathrm{var}_X}\frac{(\log n)^{1+\xi}}{n_X h_X^{d_1}} \leq \lambda^2 \leq \frac{9{\mathrm{C}_{\mathrm{var}_X}}^2}{\|\mathcal{K}\|_\infty^{2d_1}},$$

then, we have

$$(\log n)^{1+\xi} \leq \frac{n_X \lambda^2}{4\mathrm{v}_{h_X}} \leq \frac{3n_X\lambda}{4\mathrm{M}_{h_X}}.$$

Indeed,

$$\begin{aligned} \frac{n_X\lambda^2}{4\mathrm{v}_{h_X}} \leq \frac{3n_X\lambda}{4\mathrm{M}_{h_X}} \quad &\Leftrightarrow \quad \lambda \leq \frac{3\mathrm{v}_{h_X}}{\mathrm{M}_{h_X}} = \frac{3\mathrm{C}_{\mathrm{var}_X}}{\|\mathcal{K}\|_\infty^{d_1}} \\ &\Leftrightarrow \quad \lambda^2 \leq \frac{9\mathrm{C}_{\mathrm{var}_X}^2}{\|\mathcal{K}\|_\infty^{2d_1}} \end{aligned}$$

and

$$(\log n)^{1+\xi} \leq \frac{n_X\lambda^2}{4\mathrm{v}_{h_X}} \quad \Leftrightarrow \quad \frac{4\mathrm{C}_{\mathrm{var}_X}(\log n)^{1+\xi}}{n_X h_X^{d_1}} \leq \lambda^2.$$

Therefore when

$$4\mathrm{C}_{\mathrm{var}_X}\frac{(\log n)^{1+\xi}}{n_X h_X^{d_1}} \leq \lambda^2 \leq \frac{9{\mathrm{C}_{\mathrm{var}_X}}^2}{\|\mathcal{K}\|_\infty^{2d_1}},$$

$$\begin{aligned} \mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u)\right| > \lambda\right) &\leq 2\exp\left(-\min\left(\frac{n_X\lambda^2}{4\mathrm{v}_{h_X}}, \frac{3n_X\lambda}{4\mathrm{M}_{h_X}}\right)\right) = 2\exp\left(-\frac{n_X\lambda^2}{4\mathrm{v}_{h_X}}\right) \\ &\leq 2\exp\left(-(\log n)^{1+\xi}\right). \end{aligned}$$

Acknowledgements: We are very grateful to Benjamin Auder (Université Paris-Saclay) who helped us for parallelization of Rodeo algorithms.

References

Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.*, 36(3):279–298.

Bertin, K., Lacour, C., and Rivoirard, V. (2016). Adaptive pointwise estimation of conditional density function. *Ann. Inst. H. Poincaré Probab. Statist.*, 52(2):939–980.

Bouaziz, O. and Lopez, O. (2010). Conditional density estimation in a censored single-index regression model. *Bernoulli*, 16(2):514–542.

Brunel, E., Comte, F., and Lacour, C. (2007). Adaptive estimation of the conditional density in the presence of censoring. *Sankhyā*, 69(4):734–763.

Chagny, G. (2013). Warped bases for conditional density estimation. *Mathematical Methods of Statistics*, 22(4):253–282.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696.

De Gooijer, J. G. and Zerom, D. (2003). On conditional density estimation. *Statist. Neerlandica*, 57(2):159–176.

Delyon, B., Portier, F., et al. (2016). Integral approximation by kernel smoothing. *Bernoulli*, 22(4):2177–2208.

Efromovich, S. (2010). Dimension reduction and adaptation in conditional density estimation. *Journal of the American Statistical Association*, 105(490):761–774.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834.

Fan, J.-q., Peng, L., Yao, Q.-w., and Zhang, W.-y. (2009). Approximating conditional density functions using dimension reduction. *Acta Mathematicae Applicatae Sinica, English Series*, 25(3):445–456.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.*, 99(468):1015–1026.

Holmes, M. P., Gray, A. G., and Isbell, C. L. (2010). Fast kernel conditional density estimation: A dual-tree monte carlo approach. *Computational Statistics & Data Analysis*, 54(7):1707 – 1718.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336.

Ichimura, T. and Fukuda, D. (2010). A fast algorithm for computing least-squares cross-validations for nonparametric conditional kernel density functions. *Computational Statistics & Data Analysis*, 54(12):3404–3410.

Izbicki, R. and Lee, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4):1297–1316.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electron. J. Statist.*, 11(2):2800–2831.

Izbicki, R., Lee, A. B., and Pospisil, T. (2018). Abc-cde: Towards approximate bayesian computation with complex high-dimensional data and limited simulations. *arXiv preprint arXiv:1805.05480*.

Lafferty, J. and Wasserman, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1):28–63.

Le Pennec, E. and Cohen, S. (2013). Partition-based conditional density estimation. *ESAIM: Probability and Statistics*, eFirst.

Lincheng, Z. and Zhijun, L. (1985). Strong consistency of the kernel estimators of conditional density function. *Acta Mathematica Sinica*, 1(4):314–318.

Liu, H., Lafferty, J. D., and Wasserman, L. A. (2007). Sparse nonparametric density estimation in high dimensions using the rodeo. In *International Conference on Artificial Intelligence and Statistics*, pages 283–290.

Nguyen, M.-L. (2018). Nonparametric method for sparse conditional density estimation in moderately large dimensions. *arXiv:1801.06477*.

Nguyen, M.-L. (2019). *Estimation non paramétrique de densités conditionnelles : grande dimension, parcimonie et algorithmes gloutons*. PhD thesis, Université Paris-Saclay.

Nguyen, M.-L., Lacour, C., and Rivoirard, V. (2021). Supplementary material of adaptive greedy algorithm for moderately large dimensions in kernel conditional density estimation. *Submitted*.

Otneim, H. and Tjøstheim, D. (2018). Conditional density estimation using the local gaussian correlation. *Statistics and Computing*, 28(2):303–321.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2018). Abc random forests for bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728.

Rebelles, G. (2015). Pointwise adaptive estimation of a multivariate density under independence hypothesis. *Bernoulli*, 21(4):1984–2023.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York.

Sart, M. (2017). Estimating the conditional density by histogram type estimators and model selection. *ESAIM: Probability and Statistics*, 21:34–55.

Shiga, M., Tangkaratt, V., and Sugiyama, M. (2015). Direct conditional probability density estimation with sparse feature selection. *Machine Learning*, 100(2):161–182.

Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.*, 26(6):2420–2469.

Wasserman, L. and Lafferty, J. D. (2006). Rodeo: Sparse nonparametric regression in high dimensions. In *Advances in Neural Information Processing Systems*, pages 707–714.