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,
Université Paris-Saclay, France

Claire Lacour,
Université Paris-Est Marne-la-Vallée, France

Vincent Rivoirard
Université Paris-Dauphine, France

March 31, 2019

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 $(X_i, Y_i) \in \mathbb{R}^d$, $i = 1,\dots,n$. 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. Its computational complexity is only $O(dn\log n)$.

Keywords: *conditional density, high dimension, minimax rates, kernel density estimators, greedy algorithm, sparsity, nonparametric inference.*

1 Introduction

1.1 Motivations

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

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

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. Let us define $f : \mathbb{R}^d \to \mathbb{R}_+$ such as for any $x \in \mathbb{R}^{d_1}$, $f(x,\cdot)$ is the conditional density of Y conditionally to $X = x$:

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

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

The issue of estimating a conditional density may arise as soon as we observe a (possibly multidimensional) response Y associated with a (possibly multidimensional) covariate X. We often study the regression function $\mathbb{E}(Y|X = x)$, but this information is restrictive, and the entire distribution is more informative than the mean (think in particular to the case of an asymetric or multimodal distribution). Thus the problem of estimating the conditional distribution is found 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 method, 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 the single-index model, as Fan et al. (2009) or Bouaziz and Lopez (2010), 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 contain no 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 $r \in [0, d]$ 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 $\mathcal{R}^c = \{1, \dots, d\} \setminus \mathcal{R}$. 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. 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 (2.3) and (2.5). 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 significance 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. Furthermore, tuning RevDir CDRODEO is very easy (see Section 3.2) and we show that the total worst-case complexity of 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 notations

The plan of the paper is the following. First we describe in Section 2 the estimation procedure. We give heuristic ideas based on the oracle approach and explain why some modifications of the Direct CDRodeo procedure are necessary. Then a detailed presentation of our algorithm is provided in Section 2.2.4. Next the main result is stated in Section 3. The complexity of the algorithm is computed in Section 3.4. The proofs are gathered in Section 4.

In the sequel, we denote by $\star$ the convolution product. For a function $g : (u_1, \dots, u_d) \mapsto g(u_1, \dots, u_d)$, we denote $\partial_j g$ the partial derivative $\frac{\partial}{\partial u_j} g$ when there is no ambiguity. We introduce the following partial order on the bandwidths:

$$h \preceq h' \quad \Leftrightarrow \quad \forall k \in \{1, \dots, d\} \quad h_k \leq h'_k.$$

2 Estimation procedure

2.1 Kernel rule

Our estimation procedure of the conditional density f is based on a kernel rule, namely the kernel estimator introduced in (Bertin et al., 2016). So, let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, namely K satisfies $\int_{\mathbb{R}} K(t)dt = 1$. Then, for any bandwidth $h = (h_j)_{j=1,\dots,d} \in (\mathbb{R}_+^*)^d$, the estimator of f associated with K and h is defined for any $w \in \mathbb{R}^d$, 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 the problem studied in this paper. 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 classical kernel density estimator extensively studied in the literature.*

Since f can be expressed as the ratio

$$f(x, y) = \frac{f_{XY}(x, y)}{\mathrm{f}_X(x)},$$

the class of rules defined as the ratio of two density estimates has intensively been studied. The estimate $\hat{f}_h(w)$ does not belong to this class. Actually, our goal is to take into account the specific nature of the conditional density f, not the nature of f_{XY} and f_X. In particular, a relevant component both for the joint density f_{XY} and the marginal density f_X may be irrelevant for the conditional density; this occurs if a component of X is independent of Y and in this case relevance may be not detected by a ratio of two density estimates. Similarly, the smoothness of f can be different from the smoothness of the functions f_{XY} and f_X. Remark that if we could take $\tilde{\mathrm{f}}_X = \mathrm{f}_X$, then

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

which ensures that $\mathbb{E}[\hat{f}_h(w)]$ is a good approximation of f when h is small enough under mild assumptions on K and f. These arguments justify the introduction of $\hat{f}_h(w)$. The choice of $\tilde{\mathrm{f}}_X$ is essential and will be discussed in Section 3.3. Equality (2.2) shows that the selection of h will be essentially dictated by the intrinsic properties of the conditional density f.

Now, as explained in Introduction, the principal issue is to choose an appropriate bandwidth h which adapts simultaneously to the unknown sparsity and smoothness of f. In particular, large values of the components of the bandwidths will correspond to irrelevant components of f, namely $\mathcal{R}^c = \{1, \ldots, d\} \setminus \mathcal{R}$. Several estimation kernel procedures based on optimization over an exhaustive grid of bandwidths have been proposed in the literature. But the larger the class of bandwidths, the larger the computational time. So, most of them have to face with large running times, leading to intractable procedures, even for moderately large dimensions. Furthermore, as explained in Introduction, very few are able to deal with the two-fold adaptive objective.

These are the reasons why, unlike classical methods involving criteria minimization over a large class of smoothing parameters, we propose an algorithm generating an iterative smooth path through the set of bandwidths in the same spirit as Wasserman and Lafferty (2006) and Lafferty and Wasserman (2008) for nonparametric regression, Liu et al. (2007) for density estimation and Nguyen (2018) for non-adaptive conditional density estimation. The greediness of our procedure, which is presented in the next paragraph, leans on the selection of this path of bandwidths. It enables us to address adaptive conditional density estimation in high dimensions.

2.2 From the Direct CDRodeo procedure to the RevDir CDRodeo procedure

In the sequel, to describe our algorithm, we fix $w = (x, y)$, the estimation point, and we assume that K is of class $\mathcal{C}^1$.

2.2.1 The Direct CDRodeo procedure

To select the bandwidth, we would like to use local variations of f. Indeed, heuristically, the larger the local variations of f, the smaller the bandwidth. So, we naturally rely on partial derivatives of f, which are, of course, not observed. So, as a proxy of $\frac{\partial}{\partial w_j} f$, we consider Z_{hj}, the partial derivatives of the estimator with respect to the components of the bandwidths, defined for $h \in (\mathbb{R}_+^*)^d$ and $j \in \{1, \ldots, d\}$ by:

$$Z_{hj} := \frac{\partial}{\partial h_j} \hat{f}_h(w). \tag{2.3}$$

Denoting $J : t \mapsto K(t) + tK'(t)$, Z_{hj} can be easily expressed, which constitutes a key step to obtain algorithms with low computational time. We obtain:

$$Z_{hj} = \frac{-1}{nh_j^2} \sum_{i=1}^{n} \frac{1}{\tilde{\mathrm{f}}_X(X_i)} J\left(\tfrac{w_j - W_{ij}}{h_j}\right) \prod_{k \neq j}^{d} h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right). \tag{2.4}$$

The CDRodeo procedure proposed by Nguyen (2018), called the *Direct* CDRodeo procedure in the sequel, involves the Z_{hj}'s as follows:

1. We start from a bandwidth $h = (h_1, \ldots, h_d)$ whose components h_j are all equal to $h_0 > 0$ quite large (typically, h_0 is close to 1).

2. At each step, for all j, if j is not deactivated, we compare $|Z_{hj}|$ to a threshold λ_{hj}, where

$$\lambda_{hj} := \mathrm{C}_\lambda \sqrt{\frac{(\log n)^a}{n h_j^2 \prod_{k=1}^d h_k}}, \tag{2.5}$$

   with $\mathrm{C}_\lambda = 4\|J\|_2 \|K\|_2^{d-1}$ and $a > 1$ a tuning parameter. Observe that λ_{hj}^2 is a good proxy of $\mathrm{Var}(Z_{hj})$ up to the logarithmic term.

   - If $|Z_{hj}| > \lambda_{hj}$, then $h_j \leftarrow \beta h_j$ for $\beta \in (0,1)$ a constant fixed in advance, and j is still active.
   - If $|Z_{hj}| \leq \lambda_{hj}$, j is deactivated and h_j remains unchanged for the next steps of the path.

3. We stop when all components are deactivated or if $\prod_{j=1}^n h_j < \frac{\log n}{n}$.

The next paragraph provides heuristic arguments explaining why such an algorithm is able, simultaneously, to detect irrelevant components and provide suitable bandwidths for relevant components.

2.2.2 Heuristic arguments

Introducing

$$\bar{Z}_{hj} = \frac{-1}{n h_j^2} \sum_{i=1}^n \frac{1}{\mathrm{f}_X(X_i)} J\left(\tfrac{w_j - W_{ij}}{h_j}\right) \prod_{k \neq j}^d h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right), \tag{2.6}$$

which is close to Z_{hj} if $\tilde{\mathrm{f}}_X$ is a good estimate of f_X, we easily obtain that $\mathbb{E}[\bar{Z}_{hj}] = 0$ if $j \in \mathcal{R}^c$, which means that, with high probability, j is rapidly deactivated by the Direct CDRodeo procedure. Indeed, λ_{hj} is tuned (via the Bernstein concentration inequality) so that with high probability, $|\bar{Z}_{hj} - \mathbb{E}[\bar{Z}_{hj}]| \leq \lambda_{hj}$. We then obtain large smoothing parameters for irrelevant components.

To explain heuristically why the Direct CDRodeo procedure is suitable for relevant components, we use the oracle approach. For the sake of simplicity, we assume that $\tilde{\mathrm{f}}_X = \mathrm{f}_X$. Given a bandwidth h, we have:

$$\mathbb{E}[(\hat{f}_h(w) - f(w))^2] = B^2(h) + \mathrm{Var}(\hat{f}_h(w)),$$

where $B(h) := \mathbb{E}[\hat{f}_h(w)] - f(w)$ is the bias term and

$$\mathrm{Var}(\hat{f}_h(w)) = \frac{1}{n}\mathrm{Var}\left(\frac{\mathrm{K}_h(w - W_1)}{\mathrm{f}_X(X_1)}\right) \approx \frac{1}{n}\|\mathrm{K}_h\|^2 \approx \frac{1}{n} \times \prod_{j=1}^d \frac{1}{h_j}, \tag{2.7}$$

where previous approximations are justified if f is bounded from above and f_X bounded from below in the neighborhood of w. Then, the ideal bandwidth should be a global minimizer of the function

$$h \mapsto \tilde{R}(h) := B^2(h) + \frac{1}{n} \times \prod_{j=1}^d \frac{1}{h_j}.$$

Denoting h^* such a global minimizer, we assume that the sign of B is constant in the neighborhood of h^*. Without loss of generality, we then assume that B is positive in the neighborhood of h^*. So h^* will be a minimizer of

$$h \mapsto R(h) := B(h) + \frac{1}{\sqrt{n \times \prod_{j=1}^{d} h_j}}. \tag{2.8}$$

Then, if B is of class $\mathcal{C}^1$, h^* should satisfy for any j,

$$\frac{\partial}{\partial h_j} B(h^*) = \frac{1}{2} \sqrt{\frac{1}{n (h_j^*)^2 \prod_{k=1}^{d} h_k^*}}.$$

Ideally, a good algorithm would select a bandwidth satisfying this property. Of course, partial derivatives of the bias are unknown but for any h, under mild assumptions,

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

so Z_{hj} is an unbiased estimate of $\frac{\partial}{\partial h_j} B(h)$. Finally, heuristically, an ideal bandwidth should satisfy

$$Z_{h^* j} \approx \sqrt{\frac{1}{n (h_j^*)^2 \prod_{k=1}^{d} h_k^*}},$$

which is the case for the Direct CDRodeo procedure up to a logarithmic term, since CDRodeo stops as soon as $|Z_{hj}| = \lambda_{hj}$ (observe that similar arguments can be used if B remains negative in the neighborhood of h^* and in this case, we have to replace Z_{hj} with $-Z_{hj}$).

Note that if previous arguments are only heuristic ones, several issues can be pointed out:

1. Some singular points of the risk function R (defined in (2.8)) can correspond to non-global minimizers. In particular, the larger the distance between the initial bandwidth of the algorithm and the minimizer of R, the larger the probability to stop at a local minimizer of R. To circumvent this problem, we can take a small value for h_0. But taking a too small value for h_0 may be inappropriate for irrelevant components.

2. If $\mathrm{card}(\mathcal{R})$ is large, many components of the optimal bandwidth are small, which leads to many steps of the *Direct* CDRodeo procedure and then to a larger computational cost.

The first point shows that initialization appears as a key point of the algorithm. In view of these issues, it is natural to consider some variations of the *Direct* CDRodeo procedure. They are described in the next paragraph.

2.2.3 The Reverse CDRodeo procedure

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). 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. 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, $h_1^* = 1/n$ (up to a logarithmic term). So, since $\mathcal{R}$ is unknown, the initialization of the bandwidth must be not larger than $h_{0,\text{rev}} = (1/n, \dots, 1/n)$. However, such a small bandwidth leads to instability problem. In particular, the variance of $\hat{f}_{h_{0,\text{rev}}}(w)$ is of order n^{d-1} (see Equation (2.7)).

2.2.4 Our procedure: The RevDir CDRodeo procedure

Previous arguments show that to circumvent previous issues, we have to combine Direct and Reverse CDRodeo procedures, leading to the *RevDir* CDRodeo procedure. This new procedure, precisely described by Algorithm 1, comprises two steps after fixing the initial bandwidth whose components are all equal to h_0, where h_0 is assumed to be larger than all relevant components of the optimal bandwidth.

1. The first step is the Reverse CDRodeo algorithm with a sequence of non-decreasing bandwidths to estimate $\mathcal{R}^c$.

2. The second step, which concerns only components j such that after the Reverse Step $h_j = h_0$, is the Direct CDRodeo algorithm. Its goal is to deal with components associated with $\mathcal{R}$.

The output bandwidth of the algorithm is denoted $\hat{h}$. The function f is finally estimated by $\hat{f} := \hat{f}_{\hat{h}}$. Figure 1 illustrates the bandwidth path for two components. If the component belongs to $\mathcal{A}\text{ct}^{(-1)}$ (resp. $\mathcal{A}\text{ct}^{(0)}$), it is deactivated during the Reverse Step (resp. the Direct Step) and is 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, note that the tuning of h_0, as well as of the parameters a and β, needs a careful attention, which is discussed in the next section.

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 $\mathcal{C}^1$, with compact support denoted $\text{supp}(K)$. We shall also assume that K is of order p, *i.e.*: for $\ell = 1, \dots, p-1$, $\int_{\mathbb{R}} t^\ell K(t)dt = 0$. 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:

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

In the sequel, we denote

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

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* $k = 1:d$, $H_k^{(0)} \leftarrow h_0$.

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

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

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 $\hat{f}_{\hat{h}}(w)$).

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.*

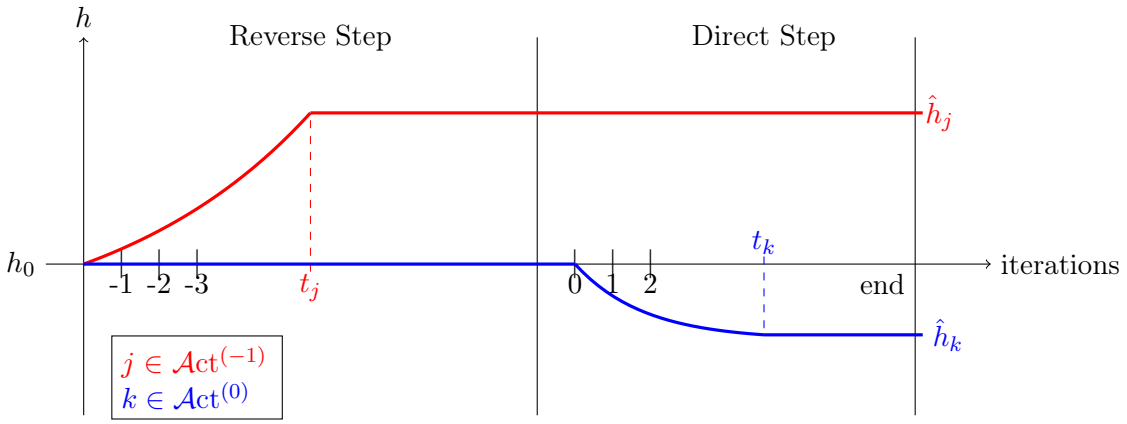

Figure 1: The two patterns of bandwidth path: the components $j \in \mathcal{A}\text{ct}^{(-1)}$ with a deactivation time $t_j \leq 0$ in red, and in blue the components $k \in \mathcal{A}\text{ct}^{(0)}$ with a deactivation time $t_k \geq 0$.

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

Definition 1. *We denote $\mathcal{R}$ the subset of $\{0, \ldots, d\}$ with cardinal r such that for any fixed $\{z_j\}_{j \in \mathcal{R}}$, the function $\{z_k\}_{k \in \mathcal{R}^c} \mapsto f(z_1, \ldots, z_d)$ 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. In the sequel, we consider the minimax point of view and 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 $\mathcal{C}^q$ 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.

In the sequel, we investigate adaptive results in terms of sparsity and smoothness properties on Hölder balls $\mathcal{H}_d(s, L)$. It means that our procedure will not depend on the knowledge of $\mathcal{R}$ and (s, L).

3.2 Tuning the RevDir CDRodeo procedure

The RevDir CDRodeo procedure depends on three tuning parameters, namely h_0, β and a.

In the sequel, we take $\beta \in (0, 1)$. Its value has no influence on rates of convergence. But of course, the larger β, the more accurate the procedure, but the larger the computational time. In practice, we set β close to 1.

The parameter a will be assumed to be larger than 1. Its value does not affect the polynomial rate of convergence but the smaller a, the smaller the exponent of the logarithmic factor of the rate. In practice, a will be larger but close to 1.

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 ideal 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 3.1, then h_0 is larger than all relevant components of the optimal bandwidth, as required by the RevDir CDRodeo procedure.

3.3 Assumptions and main result

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.

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 3. *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 shall have no or very few observations to estimate the distribution of Y given 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 4. *For the simpler problem of density estimation, since* $\mathrm{f}_X \equiv 1 \equiv \tilde{\mathrm{f}}_X$*, Assumption* $\boldsymbol{\mathcal{E}\mathbf{f}_X}$ *is obviously satisfied.*

The following proposition shows that conditions of Assumption $\boldsymbol{\mathcal{E}\mathbf{f}_X}$ are feasible if we have at hand a sample, with same distribution as X, whose size is large enough. 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* $\mathcal{C}^{p'}$ *with* $p' \geq \frac{d_1}{2(c-1)}$*, there exists an estimator* $\tilde{\mathrm{f}}_X$ *which satisfies Assumption* $\boldsymbol{\mathcal{E}\mathbf{f}_X}$*.*

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 $\beta = p'/(2p' + d_1)$). Then Condition (ii) of Assumption $\boldsymbol{\mathcal{E}\mathbf{f}_X}$ 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 $\mathcal{C}^\infty$, then we just need $c = 1$ and we can take $\tilde{X} = X$. If we know that f_X is at least of class $\mathcal{C}^1$ 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 $\boldsymbol{\mathcal{M}}$ *[Monotonicity]*
For all $j \in \mathcal{R}$*, for all* h *and* $h' \in (\mathbb{R}_+^*)^d$ *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* (2.3) *but with true* f_X *replacing* $\tilde{\mathrm{f}}_X$*.*

Let us comment Assumption $\boldsymbol{\mathcal{M}}$ that requires monotony of a specific bias term. Indeed, denoting M_j the pseudo-kernel defined by $M_j(z) = J(z_j) \prod_{k \neq j} K(z_k)$, we have

$$\mathbb{E}[\bar{Z}_{hj}] = \frac{\partial}{\partial h_j}(\mathrm{K}_h \star f - f)(w) = -\frac{1}{h_j} \int M_j(z)[f(w - h.z) - f(w)]dz,$$

which is, under mild assumptions, of order $\sum_{k=1}^d h_k^s h_j^{-1} \approx h_j^{s-1}$ if the smoothness of f at w is exactly s in each direction. In this case, Assumption $\boldsymbol{\mathcal{M}}$ is satisfied (we assume $s > 1$ subsequently). This assumption is needed to control the bias term $B(h) := (\mathrm{K}_h \star f - f)(w)$ to prevent the algorithm from stopping at bandwidths for which $\frac{\partial}{\partial h_j} B(h)$ vanishes. Remember that this term plays a key role for the RevDir CDRodeo procedure (see Section 2.2.2). It means that the RevDir CDRodeo procedure is not suitable for too irregular functions. Anyway, estimating non-smooth functions in large dimensions is a very intricate problem. Actually, this assumption is the price to pay for not exploring all possible bandwidths and only focusing on special paths and is the counterpart of the competitive computational time of the RevDir CDRodeo algorithm. Such conditions are shared by many iterative procedures. See the stopping time procedure proposed by Blanchard et al. (2016) and their Section 1.2 for instance or more generally, gradient descent algorithms that use convexity conditions. Observe that Assumption $\boldsymbol{\mathcal{M}}$ looks like a convexity condition.

Remark 5. *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* $\boldsymbol{\mathcal{M}}$ *is not required. See Nguyen (2018).*

We now derive the main result of our paper proved in Section 4 in which we show that $\hat{h}$ is closed to the ideal bandwidth h^* defined in Section 2.2.2.

Theorem 3.1. *We assume that f has only r relevant components with $r \in \{0, \dots, d\}$ and belongs to $\mathcal{H}_d(s, L)$ where $L > 0$ and $1 < s \leq p$. Then, under Assumptions $\boldsymbol{\mathcal{L}_X}$, $\boldsymbol{\mathcal{E}\mathrm{f}_X}$, $\boldsymbol{\mathcal{M}}$, 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,*

$$\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right]^{1/l} \leq C\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}} \tag{3.2}$$

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,

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

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 3.1 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.

Remark 6. *We need $s > 1$, which means that f has to be at least $\mathcal{C}^1$. 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.*

3.4 Algorithm complexity

We now discuss the complexity of CDRodeo without taking into account the pre-computation cost of $\tilde{\mathrm{f}}_X$ at the points X_i, $i = 1 : n$ (used for computing the Z_{hj}). 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 $O(d_1 n^2 \log n)$.

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 $|\mathcal{A}\mathrm{ct}^{(-1)}| + |\mathcal{A}\mathrm{ct}^{(0)}| \leq d$, we obtain the following proposition. More details can be found in the proof (see Section 4.6).

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 order $dn|H|^d$, 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 $d = 5$ and $n = 10^5$, the ratio of complexities is already $\frac{dn(\log n)^d}{dn \log n} > 1.7 \times 10^4$.

4 Proofs

4.1 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';
- $l : m$ the set of consecutive integers from l to m;
- $v_{\mathcal{I}}$ 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 use 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 $\mathcal{A}\text{ct}^{(0)}$ 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 1). This gives the following formula: for any $k = 1 : d$, 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:

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

where $h_{\text{irr}} := \beta^{t_{\text{irr}}} h_0$ such that $t_{\text{irr}} \in \mathbb{Z}$ and $\beta < h_{\text{irr}} := \beta^{t_{\text{irr}}} h_0 \leq 1$. So $\beta^{t_{\text{irr}}}$ and h_{irr} are uniquely defined. 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 $\mathcal{H}_{\text{hp}}$:

$$\mathcal{H}_{\text{hp}}^{\text{Rev}} := \{h^{(t)} : h \in \mathcal{H}_{\text{hp}}, t < 0\} \tag{4.1}$$

$$\mathcal{H}_{\text{hp}}^{\text{Dir}} := \{h^{(t)} : h \in \mathcal{H}_{\text{hp}}, t \geq 0\}. \tag{4.2}$$

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

$$\mathcal{E}_{\text{hp}} := \widetilde{\mathcal{A}}_n \cap \bigcap_{h \in \mathcal{H}_{\text{hp}}} \left(\mathcal{B}\text{ern}_{\bar{f}}(h) \cap \mathcal{B}\text{ern}_{|\bar{f}|}(h)\right) \cap \bigcap_{h \in (\mathcal{H}_{\text{hp}}^{\text{Rev}} \cup \mathcal{H}_{\text{hp}}^{\text{Dir}})} \bigcap_{k=1}^{d} \left(\mathcal{B}\text{ern}_{\bar{Z}}(h,k) \cap \mathcal{B}\text{ern}_{|\bar{Z}|}(h,k)\right), \tag{4.3}$$

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

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

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:

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

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

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

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

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 $\mathrm{C}_{\bar{\mathrm{E}}}, \mathrm{C}_{E|\bar{Z}|}$.

4.2 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:*

$$\mathcal{E}_{hp} \subset \{\hat{h} \in \mathcal{H}_{hp}\} \tag{4.4}$$

and for n large enough:

$$\mathbb{P}\left(\mathcal{E}_{hp}^c\right) \leq 2e^{-(\log n)^{1+\frac{a-1}{2}}}. \tag{4.5}$$

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 $C_A > 0$:*

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} \left|\bar{B}_h\right| \leq r\, C_{\bar{B}} C_A{}^s \frac{(\log n)^{As}}{n^{\frac{s}{2s+r}}} + r \max\left(\frac{7C_\lambda}{4\beta^{\frac{d-r}{2}} C_A{}^{\frac{r}{2}}} \frac{(\log n)^{\frac{a-Ar}{2}}}{n^{\frac{s}{2s+r}}}, \frac{7}{4}\left(\frac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}} \right), \tag{4.6}$$

$$\begin{aligned}\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} \sigma_h \\ &\leq \max\left(\frac{C_\sigma}{\beta^{d-r} C_A{}^{r/2}} (\log n)^{(a-Ar)/2}, \frac{4C_A{}^s C_{E\bar{Z}} C_\sigma \beta^{-\frac{r}{2}-s}}{C_\lambda} (\log n)^{sA} \right) n^{-\frac{s}{2s+r}},\end{aligned} \tag{4.7}$$

where C_λ is the constant defined in (2.5) and $C_{\bar{B}}, C_\sigma, C_{E\bar{Z}}$ are constants defined in Lemmas 1 and 2 in Appendix.

4.3 Proof of Theorem 3.1

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

$$\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right] = \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\rm hp}^c}\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right] + \sum_{h\in\mathcal{H}_{\rm hp}} \mathbb{E}\left[\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}}\left|\hat{f}_h(w) - f(w)\right|^l\right]. \tag{4.8}$$

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

$$\hat{f}_h(w) - f(w) = \Delta_h + \left(\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right) + \bar{B}_h, \tag{4.9}$$

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 $\prod_{k=1}^d h_k \le 1$, since $h \in \mathcal{H}_{\rm hp}$. We apply 2. of Lemma 3 and 3. of Lemma 1: Since $\mathcal{E}_{\rm hp} \subset \left(\widetilde{\mathcal{A}}_n \cap \mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)\right) \cap \mathcal{B}\mathrm{ern}_{\bar{f}}(h)$:

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

and

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

Therefore:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}}\left|\hat{f}_h(w) - f(w)\right| \le \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}}\left((\mathrm{C}_{\mathrm{M}\Delta}+1)\sigma_h + \left|\bar{B}_h\right|\right). \tag{4.10}$$

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

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

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

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

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 $p/(2p+1) \ge s/(2s+r)$ (since $p \ge s$). Therefore, for n large enough:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\rm hp}}\left|\hat{f}_h(w) - f(w)\right| \le \mathrm{C}'\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}}. \tag{4.11}$$

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:

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

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:

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

Note that, since $a > 1$,

$$2l \log n + l \log(n^{\frac{1}{2}}) = o((\log n)^{1+\frac{a-1}{2}}), \tag{4.12}$$

therefore:

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

To conclude, we combine Equation (4.8) with the above upper bound and Inequality (4.11):

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

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

4.4 Proof of Proposition 3

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

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

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 $\mathcal{E}_{\mathrm{hp}}$, 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.*:

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

and in the case where $h_0 \leq \beta$, it suffices to prove that they remain active at all iterations $t = -1 : t_{\text{irr}}$. 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 $t = -1 : t_{\text{irr}}$, 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 $h \in \mathcal{H}_{\text{hp}}$, $t = -1 : t_{\text{irr}}$ 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 $t = -1 : t_{\text{irr}}$.We will then prove that for any $h \in \mathcal{H}_{\text{hp}}$, $t = -1 : t_{\text{irr}}$ and $k \in \mathcal{R}^c$,

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

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

$$Z_{h^{(t)}k} = \left(Z_{h^{(t)}k} - \bar{Z}_{h^{(t)}k}\right) + \left(\bar{Z}_{h^{(t)}k} - \mathbb{E}\bar{Z}_{h^{(t)}k}\right) + \mathbb{E}\bar{Z}_{h^{(t)}k}. \tag{4.13}$$

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 (4.4) 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:

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

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

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

thus, we have

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

Therefore:

$$|\mathcal{H}_{\text{hp}}| \leq \left(\log_{\frac{1}{\beta}} n\right)^r. \tag{4.14}$$

Let us also control the cardinal of $\mathcal{H}_{\text{hp}}^{\text{Rev}} \cup \mathcal{H}_{\text{hp}}^{\text{Dir}}$. 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

$$\left|\mathcal{H}_{\text{hp}}^{\text{Rev}} \cup \mathcal{H}_{\text{hp}}^{\text{Dir}}\right| \leq \left(\log_{\frac{1}{\beta}} n\right)^d. \tag{4.15}$$

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

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

note that:

- 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 for any $h \in \mathcal{H}_{\text{hp}}$ for n large enough (when $\log n \geq \frac{4\|K\|_\infty^{2d}}{9\beta^r \delta^2 C_\sigma^2}$). So, we have

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

- 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.

4.5 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 $k = 1 : d$,
$$h_k = \beta^{t_k} h_0.$$
For fixed A and C_A, 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 (3.1), observe that $t(A, C_A) > 0$ (for n large enough). To simplify the notations, we assume:
$$\mathcal{R} = 1 : r$$
and
$$t_1 \geq t_2 \geq \cdots \geq t_r. \tag{4.16}$$

4.5.1 Proof of Inequality (4.6)

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

$$\begin{aligned}
\bar{B}_h :&= \mathbb{E}\left[\bar{f}_h(w)\right] - f(w) \\
&= \int_{u\in\mathbb{R}^d} \mathrm{K}_h(u)\frac{f_{XY}(u)}{\mathrm{f}_X(u_{1:d_1})}du - f(w) \\
&= \int_{u\in\mathbb{R}^d} \mathrm{K}_h(u)f(u)du - f(w) \\
&= \int_{z\in\mathbb{R}^d} \left(\prod_{k=1}^{d} K(z_k)\right)\left[f(w-h\cdot z) - f(w)\right]dz \qquad (4.17)\\
&= \int_{z'\in\mathbb{R}^r} \left(\prod_{k=1}^{r} K(z'_k)\right)\left[f_{\mathcal{R}}\left(w_{1:r} - h_{1:r}\cdot z'\right) - f_{\mathcal{R}}(w_{1:r})\right]dz'.
\end{aligned}$$

We consider the following disjunction of cases:

(Case A) $\mathcal{R} = \emptyset$

(Case B) $\min_{j\in\mathcal{R}} t_j \geq t(A, C_A)$

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

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 (4.17):

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

(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 $j \geq j_A$,

$$h_j \geq C_A (\log n)^A n^{-\frac{1}{2s+r}}. \qquad (4.18)$$

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 $h^{(\mathrm{int},t_r)}=h$):

$$\begin{aligned}\bar{B}_h = \int_{z\in\mathbb{R}^d}\left(\prod_{k=1}^d K(z_k)\right)[&f(w-h^{(\mathrm{int},t(A,C_A))}\cdot z)-f(w)\\ &+f(w-h^{(\mathrm{int},t_{j_A})}\cdot z)-f(w-h^{(\mathrm{int},t(A,C_A))}\cdot z)\\ &+\sum_{j_0=j_A+1}^{r} f(w-h^{(\mathrm{int},t_{j_0})}\cdot z)-f(w-h^{(\mathrm{int},t_{j_0-1})}\cdot z)]dz\\ = \bar{B}_{h^{(\mathrm{int},t(A,C_A))}} + (\bar{B}_{h^{(\mathrm{int},t_{j_A})}} - \bar{B}_{h^{(\mathrm{int},t(A,C_A))}}) &+ \sum_{j_0=j_A+1}^{r}\left(\bar{B}_{h^{(\mathrm{int},t_{j_0})}} - \bar{B}_{h^{(\mathrm{int},t_{j_0-1})}}\right).\end{aligned} \tag{4.19}$$

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

$$\left|\bar{B}_{h^{(\mathrm{int},t(A,C_A))}}\right| \le r\mathrm{C}_{\bar{B}}{C_A}^s\,(\log n)^{As}\,n^{-\frac{s}{2s+r}}. \tag{4.20}$$

Let us now control the other terms. The same arguments are used to control the second term $\bar{B}_{h^{(\mathrm{int},t_{j_A})}} - \bar{B}_{h^{(\mathrm{int},t(A,C_A))}}$ or the terms in the sum $\bar{B}_{h^{(\mathrm{int},t_{j_0})}} - \bar{B}_{h^{(\mathrm{int},t_{j_0-1})}}$ for $j_0 = (j_A+1):r$. To shorten the proof, the followings lines are applied to the control of the second term by identifying $h^{(\mathrm{int},t_{j_A-1})}$ to $h^{(\mathrm{int},t(A,C_A))}$ by a slight abuse of notation.
Then, let us fix $j_0\in\{j_A,\dots,r\}$. We 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 $j=1:d$,

$$h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})} \neq 0 \Rightarrow (j\in\mathcal{R} \text{ and } t_j\le t_{j_0}).$$

Indeed, given the definition of $h^{(\mathrm{int},t)}$ for all t, each irrelevant component j keeps the value h_j. For $j\in\mathcal{R}$, note that $\beta^{t_j\vee t_{j_0}}\neq\beta^{t_j\vee t_{j_0-1}}\Rightarrow t_j\le t_{j_0}$.
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

$$\begin{aligned}
\bar{B}_{h^{(\mathrm{int},t_{j_0})}} - \bar{B}_{h^{(\mathrm{int},t_{j_0-1})}} &= \int_{z\in\mathbb{R}^d} \left(\prod_{k=1}^{d} K(z_k)\right) [f(w - h^{(\mathrm{int},t_{j_0})}\cdot z) - f(w - h^{(\mathrm{int},t_{j_0-1})}\cdot z)]dz \\
&= \sum_{\substack{j\in\mathcal{R}\\ t_j\leq t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \int_{z\in\mathbb{R}^d} \left(\prod_{k=1}^{d} K(z_k)\right) z_j \partial_j f(w - h^{[j_0,u]}\cdot z)dz\ du \\
&= \sum_{\substack{j\in\mathcal{R}\\ t_j\leq t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathrm{int},t_{j_0})} - h_j^{(\mathrm{int},t_{j_0-1})}\right) \mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right] du, \qquad (4.21)
\end{aligned}$$

using Equation (5.4):

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

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 $\boldsymbol{\mathcal{M}}$ 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 $t_j \leq t_{j_0}$. 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) $t_j \geq 0$, *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 $t_k < 0 \leq t_j$, 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:

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

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

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

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 (5.4)).
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 $k \in \mathcal{R}$,

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

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

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

In each case (C.a and C.b), we have proved $\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| \leq \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$, then we apply this inequality in Equation (4.21):

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

Then, the previous decomposition of the bias (4.19) leads to:

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

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 $j = j_A : r$ 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:

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}\cap\{\hat{h}=h\}}\left|\bar{B}_h\right| \leq \mathbb{1}_{\{\hat{h}=h\}}\left(r\mathrm{C}_{\bar{B}}{C_A}^s\left(\log n\right)^{As}n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r}\frac{7}{4}\lambda_{h^{(t_j)},j}\times h_j^{(t_j)}\right),$$
$$\leq \mathbb{1}_{\{\hat{h}=h\}}\left(r\mathrm{C}_{\bar{B}}{C_A}^s\left(\log n\right)^{As}n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r}\frac{7\mathrm{C}_\lambda(\log n)^{a/2}}{4\left(n\prod_{k=1}^{d}h_k^{(t_j)}\right)^{1/2}}\right). \tag{4.22}$$

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 (4.22) 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 h_0's lower bound (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 (4.22) 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 (4.6)

4.5.2 Proof of Inequality (4.7)

Let us now prove the second inequality (4.7). 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 $k = 1 : d$,

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

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 (4.16) 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:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \left| \mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}j_0} \right] \right| > \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \frac{1}{4} \lambda_{h^{(t_{j_0}-1)}j_0}. \tag{4.23}$$

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 (4.7):

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

4.6 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.

5 Appendix

5.1 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 $i = 1:n$,*

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,*

$$\left|\bar{B}_h\right| \leq C_{\bar{B}} \sum_{k\in\mathcal{R}} h_k^s, \tag{5.1}$$

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 $j \in \mathcal{R}$:
$$|\mathbb{E}[\bar{Z}_{h,j}]| \leq C_{E\bar{Z}}h_j^{s-1}, \tag{5.2}$$
where $C_{E\bar{Z}} := \left(\int |z^sK(z)|dz\right)\frac{\|K\|_1^{r-1}L}{(s-1)!}$ 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 $C_{M\Delta} := \frac{4M_XC_{\bar{E}}}{\delta C_\sigma}$:*
$$\mathbb{1}_{\widetilde{\mathcal{A}}_n\cap\mathcal{B}ern_{|\bar{f}|}(h)}|\Delta_h| \leq C_{M\Delta}\sigma_h.$$

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

5.2 Proof of Inequality (5.1) 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 $k=1,\dots,d$, $\overline{z}_k := w - \sum_{j=1}^k h_j z_j e_j$, 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 $\mathrm{I}_k := \int_{z\in\mathbb{R}^d}\left(\prod_{k'=1}^d K(z_{k'})\right)J_k dz$ 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

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

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,

$$\left| \bar{B}_h \right| \leq \mathrm{C}_{\bar{B}} \sum_{k \in \mathcal{R}} h_k^s, \tag{5.3}$$

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

5.3 Proof of Inequality (5.2) 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

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

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 $\mathcal{R} = \{1, \ldots, r\}$. Then, by integration by part

$$\begin{aligned} \mathbb{E}[\bar{Z}_{h,j}] &= \int_{\mathbb{R}^d} (z_j K(z_j)) \left(\prod_{k \neq j} K(z_k) \right) \partial_j f(w - h \cdot z) dz \\ &= \int_{\mathbb{R}^r} \left(\prod_{k \in \mathcal{R}} K(z_k) \right) z_j \partial_j f_{\mathcal{R}}(w_{1:r} - (h.z)_{1:r}) dz_{1:r}, \end{aligned} \tag{5.4}$$

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:

$$G_{j,z,h}(1) - G_{j,z,h}(0) = \sum_{l=1}^{q-1} \frac{G^{(l)}_{j,z,h}(0)}{l!} + R'_{j,z,h,q-1}, \tag{5.5}$$

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

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

Then, the first $q-1$ terms in the r.h.s. of (5.5) vanish since $\int z_j^{l+1} K(z_j) dz_j = 0$. Now, we will bound the integral remainder of (5.5). 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 $\mathrm{C}_{E\bar{Z}} := \left(\int_{\mathbb{R}} |z|^s |K(z)| dz \right) \|K\|_1^{r-1} L/(s-1)!$.

5.4 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:

$$\begin{aligned}
\mathbb{1}_{\widetilde{\mathcal{A}}_n} \left\| \frac{\mathrm{f}_X - \tilde{\mathrm{f}}_X}{\tilde{\mathrm{f}}_X} \right\|_{\infty,\, \mathcal{U}_1} &\leq \mathbb{1}_{\widetilde{\mathcal{A}}_n} \frac{\left\| \mathrm{f}_X - \tilde{\mathrm{f}}_X \right\|_{\infty,\, \mathcal{U}_1}}{\tilde{\delta}_X} \\
&\leq \frac{2M_X}{\delta} \frac{(\log n)^{a/2}}{n^{1/2}}. \qquad (5.6)
\end{aligned}$$

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:

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

Using the above Inequality (5.6) 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:

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

Using the above Inequality (5.6) and the upper bounds 1. and 4. of Lemma 1:

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

5.5 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,

$$\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) := \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \prod_{j=1}^{d_1} \mathcal{K}\left(\frac{u_j - \widetilde{X}_{ij}}{h_X}\right), \tag{5.7}$$

where $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ is a kernel which is compactly supported, of class $\mathcal{C}^1$ 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 $\xi > 0$:

$$\mathbb{P}\left(\|\mathrm{f}_X - \tilde{\mathrm{f}}_X^{\mathcal{K}}\|_{\infty,\, \mathcal{U}_1} > C_X \frac{(\log n)^{\frac{1+\xi}{2}}}{\sqrt{n}} \right) \leq \mathcal{O}\left(n_X^{d_1+1} \exp\left(-(\log n)^{1+\xi} \right) \right). \tag{5.8}$$

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 (5.8). 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 $\mathcal{C}^{q_X}$ 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 $\mathrm{C}_{\mathrm{var}_X} := \|\mathcal{K}\|_2^{d_1} \|f_X\|_{\infty,\, \mathcal{U}_1'}^{\frac{1}{2}}$,

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

This lemma is proved in Section 5.6.

We define $p'_X = \min(p', p_X)$, so that: f_X is of class $\mathcal{C}^{p'_X}$ 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:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda\right) \leq \mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda - \mathrm{C}'_{\mathrm{bias}_X} h_X^{p'_X}\right). \tag{5.9}$$

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 (5.8), 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:

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

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 (5.9) becomes: for any threshold λ,

$$\begin{aligned}\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}}-\mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1}\geq\lambda\right)&\leq\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}}-\mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}\right\|_{\infty,\,\mathcal{U}_1}\geq\lambda-\mathrm{C}'_{\mathrm{bias}_X}h_X^{p'_X}\right)\\&\leq\mathbb{P}\left(\max_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})-\mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right|\geq\lambda-\mathrm{C}'_{\mathrm{bias}_X}h_X^{p'_X}-\mathrm{C}_{\mathrm{diff}}\frac{\epsilon}{h_X^{d_1+1}}\right)\\&\leq N(\epsilon)^{d_1}\max_{l\in(1:N(\epsilon))^{d_1}}\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})-\mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right|\geq\lambda-\mathrm{C}'_{\mathrm{bias}_X}h_X^{p'_X}-\mathrm{C}_{\mathrm{diff}}\frac{\epsilon}{h_X^{d_1+1}}\right).\end{aligned}\tag{5.10}$$

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

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 (5.10) becomes:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda\right) \leq N(\epsilon)^{d_1} \max_{l\in(1:N(\epsilon))^{d_1}} \mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right| \geq \lambda_X\right) \tag{5.11}$$

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 (5.11) becomes:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} \geq \lambda\right) \leq 2N(\epsilon)^{d_1}\exp\left(-(\log n)^{1+\xi}\right). \tag{5.12}$$

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 (5.8). 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:

$$\left\{\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} < \lambda\right\} \quad \Rightarrow \quad \left\{\left\|\tilde{\mathrm{f}}_X - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_1} < \lambda\right\}. \tag{5.13}$$

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 (5.13), and thus, from Equation (5.12), 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 $1 + \frac{a-1}{2} < 1 + \xi < a$,

$$\lambda = 4\sqrt{\mathrm{C}_{\mathrm{var}_X}}(\log n)^{\frac{1+\xi}{2}} n^{-\frac{1}{2}} \leq M_X(\log n)^{\frac{a}{2}} n^{-\frac{1}{2}} \text{ (for } n \text{ large enough).} \tag{5.14}$$

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).

5.6 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 $i = 1 : n$

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

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

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

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

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.

Blanchard, G., Hoffmann, M., and Reiß, M. (2016). Optimal adaptation for early stopping in statistical inverse problems . working paper or preprint.

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. *Submitted*.

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.

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*.

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

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.