A Robust, Fully Adaptive M-estimator for Pointwise Estimation in Heteroscedastic Regression

MICHAËL CHICHIGNOUD[1,*], JOHANNES LEDERER[1,*]

[1] *Seminar for Statistics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich*
E-mail: chichignoud@stat.math.ethz.ch; lederer@stat.math.ethz.ch

Abstract We introduce a robust and fully adaptive method for pointwise estimation in heteroscedastic regression. We allow for noise and design distributions that are unknown and fulfill very weak assumptions only. In particular, we do not impose moment conditions on the noise distribution, and we allow for zero noise. Moreover, we do not require a strictly positive density for the design distribution. In a first step, we fix a bandwidth and construct M-estimators that consist of a contrast and a kernel. We then choose the contrast and the kernel that minimize the empirical variance and demonstrate that the corresponding M-estimator is adaptive with respect to the noise and design distributions and adaptive (Huber) minimax for contamination models. In a second step, we additionally choose a data-driven bandwidth via Lepski's method. This leads to an M-estimator that is adaptive with respect to the noise and design distributions and, additionally, adaptive with respect to the smoothness of an isotropic, locally polynomial target function. These results are also extended to anisotropic, locally constant target functions. Our data-driven approach provides, in particular, a level of robustness that adapts to the noise, contamination, and outliers. We finally conclude with a detailed discussion of our assumptions and an outlook on possible extensions.

Keywords: Adaptation, Huber contrast, Lepski's method, M-estimation, minimax estimation, nonparametric regression, pointwise estimation, robust estimation.

AMS 2000 Subject Classification: Primary 62G08; secondary 62G20, 62G35.

1. Introduction

We introduce a new method for pointwise estimation in heteroscedastic regression that is adaptive with respect to the model. The new method is, in particular, adaptive with respect to the noise and design distributions (D-adaptive) and with respect to the smoothness of the regression function (S-adaptive).

Let us briefly review the related literature. The asymptotic normality of M-estimators for the location parameter in regular models is proved in the pioneering paper [12]. Later, minimax results in nonparametric regression were derived in the series of papers [26–29]. More recently, a block median method is used in [7] to prove the asymptotic equivalence between Gaussian regression and homoscedastic regression for deterministic designs and possibly heavy-tailed noises. Together with a blockwise Stein's Method with wavelets, this leads to an estimator that is adaptive optimal over Besov spaces with respect to the L_2-risk and adaptive optimal over isotropic Hölder classes with respect to the punctual risk. This estimator is thus S-adaptive. Additionally, the noise density at 0 is estimated, and a D-adaptive estimator is then found with a plug-in method. However, in contrast to this paper, only homoscedastic regression is considered and multivariate regression functions, in particular anisotropic functions, are not allowed for. We finally mention [24], where a modified version of Lepski's method is applied in homoscedastic regression.

[*]. The authors acknowledge partial financial support as member of the German-Swiss Research Group FOR916 (Statistical Regularization and Qualitative Constraints) with grant number 20PA20E-134495/1.

What is the main idea behind our approach? Consider the estimation of $t^0 \in \mathbb{R}$ in the translation model $\mathcal{Y} \sim g(\cdot - t^0)$ for a probability density g. The M-estimator $\hat{t}$ of t^0 corresponding to a constrast ρ and a sample $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ of $\mathcal{Y}$ is then

$$\hat{t} := \arg\min_t \sum_{i=1}^n \rho(\mathcal{Y}_i - t).$$

It holds that (see [12–14])

$$\sqrt{n}(\hat{t} - t^0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{AV}), \quad \mathrm{AV} := \frac{\int (\rho')^2 dG}{\left(\int \rho'' dG\right)^2}, \tag{1.1}$$

where G is the distribution of $\mathcal{Y}$, ρ' and ρ'' are the first and second derivatives of the contrast ρ, and $\mathcal{L}$ indicates convergence in law. In other words, $\hat{t}$ is asymptotically normal with asymptotic variance AV. This result suggests that an optimal estimator is obtained minimizing the asymptotic variance. This is the main idea behind our approach. To support this idea further, we recall that (see [12])

$$\inf_\rho \frac{\int (\rho')^2 dG}{\left(\int \rho'' dG\right)^2} = \big(I(G)\big)^{-1}, \tag{1.2}$$

where $I(G)$ is the Fisher information for the true distribution G and the infimum is taken over all twice differentiable contrasts. This implies, together with the Crámer-Rao Inequality, that an efficient M-estimator exists. Huber proposed in [12, Proposal 3] to minimize an estimate of the above asymptotic variance (since the the distribution G is not available in practice) over the family of Huber contrasts (see below). He also conjectured that the corresponding estimator is minimax for certain contamination models. More recently, an M-estimator with a contrast that minimizes an estimate of the asymptotic variance was introduced for the parametric model, and its asymptotic normality was proved (see [1]). As examples, Huber contrasts indexed by their scale and a family of ℓ_p losses are treated. In this paper, we consider local M-estimators consisting of a contrast and a kernel such that an estimate of the nonasymptotic variance is minimized. We present, in particular, a nonasymptotic result which shows that the corresponding estimator mimics the oracle, that is, the function that minimizes the true variance. An advantage of our approach is, for example, that a data-driven selection of the scale of the Huber contrast provides an adaptive robustness with respect to outliers. Additionally, a suitable choice of the support of the kernel can take a maximal number of points around x_0 into account (cf. [10]). In particular, noncentered or even nonconvex supports can be considered. Finally, we show that our estimator is D-adaptive for various sets of contrasts and kernels with finite entropy.

We finally study the problem of S-adaptation. Our main goal is to find a simultaneously D- and S-adaptive pointwise estimator for anisotropic target functions. However, this is not straightforward since the standard Lepski's method (see [19, 21]) only applies to isotropic functions. Therefore, we restrict ourselves to these functions in a first step. We use Lepski's method for the S-adaptation plugging-in an estimate of the minimal variance for the D-adaptation (this is also the case in the context of model selection, see [4], or the Lasso, see [6]).This way, we obtain the first estimator in heteroscedastic regression with random design and noise distributions with heavy tails which is simultaneously D- and S-adaptive and optimal in a sense describe later. Additionally, we allow for zero noise (which is detected by the estimator). Furthermore, we note that the application of Lepski's method for nonlinear estimators is still nonstandard, and only very few examples can be found in the literature ([8, 23, 24]). In a next step, we extend our results to anisotropic target functions. For this, we have to restrict ourselves to locally constant target functions. We apply a modification of Lepski's methods given in [16, 20] and construct an optimal, simultaneously S-and

D-adaptive estimator. This is the first application of such a method to nonlinear estimators for anisotropic target functions. Of great interest is, in particular, the corresponding selection of an anisotropic bandwidth for applications in the context of image denoising (cf. [2]). Moreover, our methods can be applied to establish robust, adaptive confidence bands (cf. [11]).

The structure of this paper is as follows: In the following section, we first introduce a nonasymptotic variance that resembles the asymptotic variance (see Theorem 1) and then provide a choice for the contrast and the kernel (see Theorem 2). We show, in particular, that the corresponding estimator is Huber minimax (see Section 2.3). Then, we provide a choice for the bandwidth for isotropic, locally polynomial target functions (see Theorem 3) and for anisotropic, locally constant target functions (see Theorem 4). After this, we give a discussion on our assumptions and an outlook in Section 4. The proofs are finally conducted in Section 5 and in the Appendix, and some sample entropy calculations are presented in Section 6.1.

2. A D-adaptive Estimator for Fixed Bandwidths

In this section, we consider pointwise estimation in heteroscedastic regression for fixed bandwidths. In the first part, we define an estimator with a local polynomial approach for a fixed kernel and a fixed contrast. In the second part, we additionally allow for the selection of the kernel and the contrast via a minimization of the variance of the estimator. Finally, we elaborate on the parametric model and relate to important classical results.

Let us specify the model beforehand. We assume the observations $Z^{(n)} := (X_i, Y_i)_{i=1,\ldots,n}$, $n \in \mathbb{N}^*$, to be distributed according to P and to satisfy the set of equations

$$Y_i = f^*(X_i) + \sigma(X_i)\,\xi_i, \quad i = 1, \ldots, n. \tag{2.1}$$

We aim at estimating the target function $f^* : [0,1]^d \to \mathbb{R}$ at a given point x_0 on $(0,1)^d$. The target function is assumed to be smooth, more specifically, it is assumed to belong to a Hölder class (see Definition 4 below). The target function is obscured by the second part of the above model, the noise. The noise variables $(\xi_i)_{i\in 1,\ldots,n}$ are assumed to be distributed independently according to the densities $g_i(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}$. The noise densities may be unknown, but we assume that $\sum_i g_i(\cdot)$ is symmetric and that there exist $A \in]0,1]$ and $\gamma_{\min} > 0$ such that

$$\int_{-\gamma_{\min}\|\sigma\|_\infty^{-1}}^{\gamma_{\min}\|\sigma\|_\infty^{-1}} n^{-1} \sum_i g_i\,(z)\,dz \geq A. \tag{2.2}$$

The latter assumption is trivially satisfied with $A = 1$ and $\gamma_{\min} = 1$ if $\|\sigma\|_\infty = 0$ (invoking the convention $1/0 = \infty$). We stress that we do not impose, unlike in the literature on the median (cf. [7]), any moment assumptions on the noise, and we do not require that the noise densities are positive at 0. Indeed, Assumption (2.2) imposes that the density $\sum_i g_i(\cdot)$ has enough mass on the interval $[-\gamma_{\min}, \gamma_{\min}]$ (We refer to Section 4 for a more detailed discussion on the assumptions.) The noise level $\sigma : [0,1]^d \to \mathbb{R}_+^*$ is assumed to be bounded, but may also be unknown. Usually, the noise level is the variance of the noise, however, this is not the case if the noise distribution does not have any moments, for example. Finally, the design points $(X_i)_{i\in 1,\ldots,n}$ are assumed to be distributed independently and identically according to $\mu(\cdot)$. For ease of exposition, we also assume that $(X_i)_{i\in 1,\ldots,n}$ and $(\xi_i)_{i\in 1,\ldots,n}$ are mutually independent.

2.1. Definitions and First Results

In this part, we introduce an estimator of $f^*(x_0)$ with a local polynomial approach for a fixed bandwidth, a fixed kernel, and a fixed contrast. The properties of this estimator are highlighted in Theorem 1.

As a first step, we set the framework for the local polynomial approach (LPA), described for example in [15] and in [30, Chapter 1]. The key idea of the LPA is to approximate the function in the neighborhood of the point in question by a polynomial. To start, we consider a hyperrectangle, not necessarily centered neighborhood $V_h \subseteq [0,1]^d$ of the point in question $x_0 \in (0,1)^d$ such that $\int_{V_h} dx = \prod_j h_j$, where h_j is the jth component of a fixed bandwidth $h \in \mathcal{H} := [h_{\min}, h_{\max}]^d \subseteq (0,1)^d$. The minimal and maximal bandwidths are given by

$$h_{\min} := \left(\frac{C[\ln(n)]^6}{n}\right)^{1/d} \quad \text{and} \quad h_{\max} := [\ln(n)]^{\text{-}1/(2b+d)}, \tag{2.3}$$

where C is a constant large enough such that Conditions 1, 2, and 3 in Section 5.1 are satisfied. Additionally, we define for a fixed $b \in \mathbb{N}$ the set $\mathcal{P} := \{p = (p_1, \dots, p_d) \in \mathbb{N}^d : \ 0 \leq |p| \leq b\}$ with $|p| = p_1 + \cdots + p_d$ and denote its cardinality by $|\mathcal{P}|$. For any multi-indexed vector $t^\top = \big(t_{p_1,\dots,p_d} \in \mathbb{R} \ : \ p \in \mathcal{P}\big) \in \mathbb{R}^{|\mathcal{P}|}$ and for any $x \in [0,1]^d$, we then define the desired polynomial as

$$\mathrm{P}_t(x) := \sum_{p \in \mathcal{P}} t_p \left(\frac{x - x_0}{h}\right)^p \mathbb{1}_{V_h}(x).$$

Here, $\mathbb{1}$ is the indicator function, $z^p := z_1^{p_1} \cdots z_d^{p_d}$ for all $z \in \mathbb{R}^d$, and the division by h is understood coordinatewise. Finally, for a fixed $M > 0$, we define the set of all polynomials of degree b as $\mathcal{F} := \left\{\mathrm{P}_t \ : \ t \in [\text{-}M, M]^{|\mathcal{P}|}\right\}$.

We now introduce the desired estimator of $f^*(x_0)$. To this end, we first specify what we mean by a kernel and a contrast. A kernel (function) $K : \mathbb{R}^d \to \mathbb{R}$ is a nonnegative function with a compact support included in $[\text{-}1/2, 1/2]^d$, $\|K\|_\infty \leq \mathcal{K}_{\max}$ (for a given constant $\mathcal{K}_{\max} \geq 1$), and $\int K(x)dx = 1$. For ease of exposition, we use the notation $K_h(x) := K\left((x - x_0)/h\right) / \prod_j h_j$ at some points. Next, we specify what we mean by a contrast (function):

Definition 1. *A function ρ is called contrast (function) if it has the following properties:*

1. *$\rho : \mathbb{R} \to \mathbb{R}_+$ is a convex and symmetric function and $\rho(0) = 0$;*
2. *the derivative ρ' of ρ is 1-Lipschitz on $\mathbb{R}$ and bounded: $\|\rho'\|_\infty < \gamma_{\max}$, for a given constant $\gamma_{\max} \geq 1$;*
3. *the second derivative ρ'' of ρ is defined almost everywhere and is $L_{\rho''}$-Lipschitz with respect to the measure P for some $L_{\rho''} > 0$. Moreover, we assume that $\|\rho''\|_\infty \leq 1$ (without loss of generality) and*
$$\rho''_{\min} := \inf_{z \in [\text{-}\gamma_{\min}, \gamma_{\min}]} \rho''(z) > 0,$$
where $\gamma_{\min} > 0$ is defined in (2.2).

Note that Assumption *3* implies that contrasts are strictly convex on the interval $[\text{-}\gamma_{\min}, \gamma_{\min}]$. Moreover, for a given $A > 0$, $\gamma_{\min}$ implicitly depends on the noise distribution via Assumption (2.2), and we assume that it is known; its estimation is discussed in Section 4. Well-known contrasts are, for any scale $\gamma > 0$, the Huber contrast (see [12])

$$\rho_{\mathrm{H},\gamma}(z) := \begin{cases} z^2/2 & \text{if } \ |z| \le \gamma \\ \\ \gamma(|z| - \gamma/2) & \text{otherwise} \end{cases}$$

and the contrast induced by the arctan function (see [26])

$$\rho_{\mathrm{arc},\gamma}(z) := \gamma z \arctan(z/\gamma) - \frac{\gamma^2}{2} \log(1 + z^2/\gamma^2).$$

Note that the absolute loss (cf. Assumption *3*) and quadratic loss (cf. Assumption *2*) do not satisfy the above conditions. However, they can be mimicked by the Huber contrast with γ small (median) and big (mean). We can now combine a kernel and a contrast to obtain the local λ-criterion for any $f \in \mathcal{F}$:

$$P_n\lambda(f) := n^{\text{-}1} \sum_{i=1}^{n} \lambda(X_i, Y_i, f), \quad \text{where} \quad \lambda(x, y, f) := \rho\big(y - f(x)\big)\, K_h(x), \quad \text{for all } x, y \in \mathbb{R}. \tag{2.4}$$

The λ-LPA estimator $\hat{f}_\lambda(x_0)$ of $f^*(x_0)$ is finally defined as

$$\hat{f}_\lambda := \arg\min_{f\in\mathcal{F}} P_n\lambda(f). \tag{2.5}$$

The coefficients of the estimated polynomial can be considered as estimators of the derivatives of the function f^* at x_0. In this paper, however, we focus on the estimation of $f^*(x_0)$.

The variance of the estimator is crucial for the following. To state it explicitly, we need to introduce some more notation: First, let λ' and λ'' be the first and second derivative of the function $\lambda(x, y, \cdot)$ and set $\Pi_h := \prod_{j=1}^d h_j$, $P\zeta := \mathbb{E}_{(X,Y)\sim P}\, P_n\zeta(X, Y)$, and $\lambda'_\infty := \sup_{x,y,f} \Pi_h|\lambda'(x, y, f)| = \|\rho'\|_\infty\|K\|_\infty$. We then introduce the crucial quantity

$$\mathrm{V}(\lambda) := \left(\frac{\sqrt{\Pi_h}\sqrt{P\,[\lambda'(f^*)]^2 + \lambda'_\infty (n\Pi_h)^{-1/4}}}{P\lambda''(f^*)} \right)^2 . \tag{2.6}$$

We call it nonasymptotic variance since it plays the role of the variance in the risk bounds in the theorems below. The explicit expressions of the numerator and the denominator can be deduced from

$$P[\lambda'(f^*)]^2 = \int \mu(x) K_h^2(x) \int \big[\rho'\big(\sigma(x)z\big)\big]^2\, n^{\text{-}1} \sum_i g_i\,(z)\, dz\, dx \tag{2.7}$$

$$\text{and} \qquad P\lambda''(f^*) = \int \mu(x) K_h(x) \int \rho''\big(\sigma(x)z\big) n^{\text{-}1} \sum_i g_i\,(z)\, dz\, dx. \tag{2.8}$$

The variance $\mathrm{V}(\lambda)$ depends on h, but one can show that this dependence is weak. From Assumption (2.2), the strict convexity of ρ on $[\text{-}\gamma_{\min}, \gamma_{\min}]$, and the boundedness assumption on ρ' in Definition 1, we conclude that $\mathrm{V}(\lambda) < \infty$. In the particular case $h = (1, \dots, 1)$ (see the parametric case below), the nonasymptotic variance $\mathrm{V}\,(\lambda)$ tends towards the asymptotic variance $\mathrm{AV}(\lambda)$ defined in (1.1) as $n \to +\infty$.

At this point, we can give a first result for the above estimator. To this end, we define the bias term of the estimator as

$$b_h := \inf_{f\in\mathcal{F}} \sup_{x\in V_h} |f(x) - f^*(x)|, \tag{2.9}$$

and we introduce the entropy term for all $\varepsilon > 0$ as

$$\breve{B}_\varepsilon := 27\int_0^1 \sqrt{H_{\mathcal{F},\nu}(u) \wedge n} du + 2\left(\frac{1}{(nh_{\min}^d)^{1/4}} + \frac{1}{\sqrt{n}}\right) H_{\mathcal{F},\nu}(1) + \varepsilon. \tag{2.10}$$

$H_{\mathcal{F},\nu}(\cdot)$ is the metric entropy of the set $\mathcal{F}$ with respect to the pseudometric

$$\nu(f_1, f_2) := \sqrt{\Pi_h P\left[\lambda'(f_1) - \lambda'(f_2)\right]^2} \qquad f_1, f_2 \in \mathcal{F}.$$

Then, the follwoing result holds:

Theorem 1. *If n is sufficiently large (according to Condition 1 in Section 5.1), it holds that for any $\lambda \in \Lambda$, any $h \in \mathcal{H}$, and for all $q \geq 1$*

$$\mathbb{E}_{f^*}\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right|^q$$
$$\leq C_q\left(b_h + \breve{B}_0\frac{\sqrt{\mathrm{V}(\lambda)}}{\sqrt{n\Pi_h}}\right)^q + 2|\mathcal{P}|((1+|\mathcal{P}|)M)^q \exp\left(-\frac{n\Pi_h/(4\ln^2 n)}{98\gamma_{\max}^2\mathcal{K}_{\max}^2 + 4\mathcal{K}_{\max}\gamma_{\max}}\right).$$

For a constant C_q ($C_q = 4q|\mathcal{P}| \cdot 68^q\,\mathrm{Gamma}(q)$ works, where $\mathrm{Gamma}(\cdot)$ is the classical Gamma function).

Remark 1. *We note that we could replace in (2.2) the global quantity $\|\sigma\|_\infty$ by the local one $\sup_{x\in V_h}|\sigma(x)|$. Moreover, if we additionally impose Condition 3 on n, the second term on the right hand side of the above bound is of order $o(1/n)$ and thus negligible. However, we stress that the above result is nonasymptotic - in contrast to the classical results of Huber (cf. [12] and also [26–29], [1]). Moreover, we also stress that we do not impose conditions on the design and the noise level except for its boundedness. In particular, we allow for degenerate designs and vanishing noise. (A more detailed discussion is given in Section 4.) For the proof, we use Bernstein's inequality and chaining arguments, in particular, we use deviation inequalities in [22] that rely on Dudley's entropy integral. With this, we can recover the shape of the variance, but we obtain an additional (large) factor $C_q\breve{B}_0^q$. The reduction of these factors is of minor interest for this paper. Finally, for further implications of the above result, we refer to Section 2.3.*

Remark 2. *The above bound is, to the best of our knowledge, a new result in nonparametric regression. However, the next step is to choose a λ that minimizes the right hand side. If we neglect the second term, this reduces to a minimization of the variance term $\breve{B}_0\sqrt{\mathrm{V}(\lambda)}/\sqrt{n\Pi_h}$ since the bias term b_h does not depend on λ. Note that V does not depend on the target function, and, in particular, not on the smoothness of the target function. This allows for a wide range of applications in various models, for example, in high dimensional settings (see Section 4). The adaptation with respect to the smoothness of the target function is finally done via the selection of a suitable bandwidth. The simultaneous D- and S-adaptation is difficult since the variance V depends on h. We detail this in Section 3.*

2.2. Selection of the Kernel and the Contrast for Fixed Bandwidths (D-adaptation)

How should the combined function $\lambda \in \Lambda$, that is, the kernel and the contrast, be selected? We introduce an oracle that minimizes the bound in Theorem 1 above and then propose a selection that mimics this oracle. We first introduce, for a given set of contrasts Υ, a given set of kernels $\mathcal{K}$, and a bandwidth $h > 0$, the set of possible combined functions λ:

$$\Lambda := \left\{\lambda : \lambda(x, y, f) = \rho\big(y - f(x)\big)\, K_h(x),\ \rho \in \Upsilon,\ K \in \mathcal{K}\right\}. \tag{2.11}$$

We then note that the bias term b_h in Theorem 1 is of importance for the choice of the bandwidth later. For a fixed bandwidth, however, we can concentrate on the second term only and introduce the oracle as

$$\lambda^* := \arg\min_{\lambda\in\Lambda} \mathrm{V}(\lambda). \tag{2.12}$$

To mimic the oracle λ^*, we then define the estimator $\widehat{\lambda}$

$$\widehat{\lambda} := \arg\min_{\lambda\in\Lambda} \widehat{\mathrm{V}}(\lambda), \quad \text{where} \quad \widehat{\mathrm{V}}(\lambda) := \left(\frac{\sqrt{\Pi_h}\sqrt{P_n\left[\lambda'(\hat{f}_\lambda)\right]^2} + \lambda'_\infty (n\Pi_h)^{-1/4}}{P_n\lambda''(\hat{f}_\lambda)}\right)^2. \tag{2.13}$$

Note that we estimate $P\left[\lambda'\left(f^*\right)\right]^2$ and $P\lambda''\left(f^*\right)$ by their empirical versions $P_n\left[\lambda'\left(\hat{f}_\lambda\right)\right]^2$ and $P_n\lambda''\left(\hat{f}_\lambda\right)$, and that estimate f^* by $\hat{f}_\lambda$ and that the explicit expressions for the numerator and the denominator are given by

$$P_n[\lambda'(\hat{f}_\lambda)]^2 = \frac{1}{n}\sum_{i=1}^{n} K_h^2(X_i)\left[\rho'\big(Y_i - \hat{f}_\lambda(X_i)\big)\right]^2$$

$$\text{and} \qquad P_n\lambda''(\hat{f}_\lambda) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i)\rho''\big(Y_i - \hat{f}_\lambda(X_i)\big).$$

We now show that the estimator that results from (2.5) and (2.13) performs - up to constants - as well as the oracle. For this, we define for all $z > 0$

$$B_z := \left(1 \vee 27\int_0^1 \sqrt{H_{\mathcal{F}\cup\Lambda,\omega}(u) \wedge n du}\right) + 2\left(\frac{1}{(nh_{\min}^d)^{1/4}} + \frac{1}{\sqrt{n}}\right) H_{\mathcal{F}\cup\Lambda,\omega}(1) + 10\sqrt{z} + \frac{2z}{(n\Pi_h)^{1/4}}, \tag{2.14}$$

where $H_{\mathcal{F}\cup\Lambda,\omega}(\cdot)$ is the metric entropy of $\mathcal{F}\cup\Lambda$ with respect to the pseudometric

$$\omega\left((f_1, \lambda_1), (f_2, \lambda_2)\right) := \nu(f_1, f_2) \ \vee\ \sqrt{\Pi_h P\left[\kappa(f_1, \lambda_1) - \kappa(f_2, \lambda_2)\right]^2} \ \vee\ \sqrt{\Pi_h P\left[\lambda_1''(f_1) - \lambda_2''(f_2)\right]^2} \tag{2.15}$$

for any $f_1, f_2 \in \mathcal{F},\ \lambda_1, \lambda_2 \in \Lambda$,

$$\kappa(f, \lambda) := \frac{\lambda'(f)}{\sqrt{\Pi_h P[\lambda'(f)]^2} + \lambda'_\infty/(n\Pi_h)^{\frac{1}{4}}},$$

and $\nu(\cdot,\cdot)$ is defined above Theorem 1. For example, we give, in the Appendix, the computation of this entropy for the family of Huber contrasts indexed by the scale. Then, we have the following result:

Theorem 2. *If n is sufficiently large (according to Conditions 1 and 2 in Section 5.1), then, for any $h \in \mathcal{H}$ and for all $q \geq 1$, it holds that*

$$\mathbb{E}_f \left| \hat{f}_{\widehat{\lambda}}(x_0) - f^*(x_0) \right|^q \leq 2C_q \left(b_h + B_0 \frac{\sqrt{\mathrm{V}(\lambda^*)}}{\sqrt{n\Pi_h}} \right)^q + o(1/n).$$

Remark 3. *We stress that our estimator does not depend on the densities $(g_i)_i$ and $(\mu_i)_i$ and the noise level σ, and we observe that it achieves - up to constants- the optimal variance $\mathrm{V}(\lambda^*)$. We thus call $\hat{f}_{\widehat{\lambda}}$ D-adaptive optimal. This notion of optimality, however, depends on the family Λ under consideration.*

Remark 4. *Via a bias/variance trade-off, we can obtain S-minimax results (with minimal variance) with respect to the Hölder smoothness β of the target function (see Definition 4 below). Indeed, we can obtain the usual S-minimax rate $n^{-\bar{\beta}/(2\bar{\beta}+1)}$, where $\bar{\beta}$ is the harmonic average of $\vec{\beta}$.*

2.3. Parametric Case and Huber Minimaxity

We finally elaborate on the special case of parametric estimation, that is, we assume $f^* = t^0$, $t^0 \in [\text{-}M, M]$, and consider the model $\mathcal{Y} \sim g(\cdot - t^0)$ for a symmetric density g. In parametric estimation, we set the kernel equal to 1 and thus consider the estimator

$$\widehat{t}_\rho := \arg \min_{t \in [\text{-}M,M]} \frac{1}{n} \sum_{i=1}^{n} \rho\big(\mathcal{Y}_i - t\big) \tag{2.16}$$

of the scalar t^0.

From the above results, we can now deduce the following corollary:

Corollary 1. *Let ρ^* and $\widehat{\rho}$ be constructed according to (2.12) and (2.13) with $h := (1, \ldots, 1)$ and $\lambda(y,t) := \rho\big(y - t\big)$ for all $y \in \mathbb{R}$ and $t \in [\text{-}M, M]$. Then, if n is sufficiently large (according to Conditions 1 and 2 in Section 4.1), it holds that*

$$\mathbb{E}_{t^0} \left| \widehat{t}_{\widehat{\rho}} - t^0 \right|_1 \leq 2C_1 B_0 \frac{\sqrt{\mathrm{V}(\rho^*)}}{\sqrt{n}} + o(1/n).$$

We note that the constant M does only appear in the residual term and does not play a major role in the following.

Let us relate our results to the Huber minimaxity. For this, we define the set of r-contaminated normal distributions for a contamination level $r \in [0, 1[$ as

$$\mathcal{G}_r := \left\{ G \,:\, G = (1-r)N + rT,\ T \in \Xi \right\},$$

where N is the standard normal distribution and Ξ is the set of all symmetric real distributions. The “minimax” variance over this set of distribution is then as follows:

Lemma 1. *Let the distribution G_0 be the minimizer of the Fisher information $I(G)$ over $\mathcal{G}_r$. Then, for any $r \in [0,1[$*

$$\inf_{\rho} \sup_{G \in \mathcal{G}_r} \mathrm{AV}(\rho, G) \geq \sup_{G \in \mathcal{G}_r} I^{\text{-}1}(G) = I^{\text{-}1}(G_0),$$

where the infimum is taken over all twice differentiable and convex contrasts and AV *is defined in* (1.1)*. Moreover, the expression of the density of the distribution G_0 is*

$$g_0(z) = \begin{cases} \frac{1-r}{\sqrt{2\pi}} \exp\left(\gamma_r t + \gamma_r^2/2\right) & \textit{if} \;\; t \leq \text{-}\gamma_r \\ \\ \frac{1-r}{\sqrt{2\pi}} \exp\left(-t^2/2\right) & \textit{if} \;\; \text{-}\gamma_r \leq t \leq \gamma_r \\ \\ \frac{1-r}{\sqrt{2\pi}} \exp\left(-\gamma_r t + \gamma_r^2/2\right) & \textit{if} \;\; t \geq \gamma_r \end{cases},$$

where γ_r is the solution of

$$(1-r)^{\text{-}1} = 2\int_0^{\gamma_r} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz + \frac{\sqrt{2}}{\gamma_r \sqrt{\pi}} e^{-\gamma_r^2/2}.$$

The first claim follows from (1.2) and the second one from [12, Theorem 2]. Lemma 1 shows that $I^{\text{-}1}(G_0)$ is a lower bound for the asymptotic variance in the worst case. This asymptotic variance can be achieved, as we see in the following result:

Lemma 2. *For any $r \in [0,1[$ and the Huber contrast $\rho_{\mathrm{H},\gamma_r}$ as defined in the previous section, it holds that the Huber corresponds to the maximum likelihood estimator for the distribution G_0, $\rho_{\mathrm{H},\gamma_r}(\cdot) = \text{-}\ln(g_0(\cdot))$ and*

$$\sup_{G \in \mathcal{G}_r} \mathrm{AV}(\rho_{\mathrm{H},\gamma_r}, G) \leq I^{\text{-}1}(G_0).$$

This is a corollary of [12, Theorem 2]. It means that the estimator constructed with $\rho_{\mathrm{H},\gamma_r}$ has minimal asymptotic variance for the worst distribution G_0 in $\mathcal{G}_r$. We may say that $I^{\text{-}1}(G_0)$ is the asymptotic minimax variance and the estimator constructed with $\rho_{\mathrm{H},\gamma_r}$ is asymptotic minimax.

Usually, a minimax estimator is desired for an unknown contamination level r. We show that it can be constructed with Corollary 1: Set $\Upsilon_{\mathrm{H}} := \{\rho_{\mathrm{H},\gamma} : \gamma \in [\gamma_{\min}, \gamma_{\max}]\}$ such that $\gamma_r \in [\gamma_{\min}, \gamma_{\max}]$ for all $r \in [0,1[$. Then, define $\widehat{\gamma}$ as the minimizer of $\widehat{\mathrm{V}}(\rho_{\mathrm{H},\gamma})$ (see (2.13)) over $[\gamma_{\min}, \gamma_{\max}]$. Finally, define $\widehat{t_{\widehat{\gamma}}}$ according to (2.16) with $\rho = \rho_{\mathrm{H},\widehat{\gamma}}$. The resulting estimator $\widehat{t_{\widehat{\gamma}}}$ has then the following property:

Corollary 2. *For any $r \in [0,1[$, it holds that*

$$\sup_{G \in \mathcal{G}_r} \mathbb{E}_{t^0} \left| \widehat{t_{\widehat{\gamma}}} - t^0 \right| \leq \frac{2C_1 B_0}{\sqrt{nI(G_0)}} + o(1/n).$$

The estimator $\widehat{t_{\widehat{\gamma}}}$ is thus adaptive with respect to the contamination level r and is (up to constants) asymptotic minimax in the above sense. This corollary is deduced from Corollary 1 and the definition of $\mathrm{V}(\rho^*)$. In the following, we then focus to find upper bounds with the minimal value of the variance as Theorem 2, that is, the optimality for us.

3. A D-adaptive and S-adaptive Estimator

In this section, we introduce an estimator of $f^*(x_0)$ that is simultaneously S- and D-adaptive. For this, we apply the data-driven procedure introduced above to select the contrast and the kernel, and we apply the data-driven Lepski's method to select the bandwidth. Afterwards, we present adaptive S-minimax results for this D-adaptive estimator.

Let us introduce the necessary definitions first. To start, we recall the notion of S-minimaxity. To this end, let $\tilde{f}_n(x_0)$ be an estimator of $f^*(x_0)$ and $\mathcal{S}$ a set of functions. For any $q > 0$, we then define the maximal risk and the *S-minimax risk* of $\tilde{f}_n$ for x_0 and $\mathcal{S}$ as

$$R_{n,q}\big[\tilde{f}_n, \mathcal{S}\big] := \sup_{f^* \in \mathcal{S}} \mathbb{E}_f \big|\tilde{f}_n(x_0) - f^*(x_0)\big|^q \text{ and } R_{n,q}\big[\mathcal{S}\big] := \inf_{\tilde{f}} R_{n,q}\big[\tilde{f}, \mathcal{S}\big], \tag{3.1}$$

respectively. The infimum on the right hand side is taken over all estimators. We can now define the *S-minimax rates of convergence* and the *(asymptotic) S-minimax estimators*:

Definition 2. *A sequence ϕ_n is an S-minimax rate of convergence and the estimator $\hat{f}$ is an (asymptotic) S-minimax estimator with respect to the set $\mathcal{S}$ if*

$$0 < \liminf_{n\to\infty} \phi_n^{-q}\, R_{n,q}\big[\mathcal{S}\big] \le \limsup_{n\to\infty} \phi_n^{-q}\, R_{n,q}\big[\hat{f}, \mathcal{S}\big] < \infty.$$

Usually, the set $\mathcal{S}$ is unknown. In our case, for example, it depends the smoothness $\vec{\beta}$. More generally, $\mathcal{S} = \mathcal{S}_m$, $m \in \mathcal{M}$, for a set of parameters $\mathcal{M}$. It is then desirable to have an estimator that is adaptive with respect to $\mathcal{M}$. This motivates the following definition, where $\Psi := \{\psi_n(m)\}_{m\in\mathcal{M}}$ is a given family of normalizations:

Definition 3. *The family Ψ is called admissible if there exist an estimator $\hat{f}_n$ such that*

$$\limsup_{n\to\infty} \sup_{m\in\mathcal{M}} \psi_n^{-q}(m)\, R_{n,q}\big(\hat{f}_n, \mathcal{S}_m\big) < \infty.$$

The estimator $\hat{f}_n$ is then called Ψ-adaptive in the S-minimax sense.

The LPA is designed for functions that can be locally approximated by polynomials. This is, for example, the case for Hölder classes. Similarly as in [3], we define:

Definition 4. *Let $\vec{\beta} := (\beta_1, \ldots, \beta_d) \in]0, +\infty[^d$ such that $\lfloor\beta_1\rfloor = \ldots = \lfloor\beta_d\rfloor =: \lfloor\beta\rfloor$, and let $L, M > 0$. The function $s : [0,1]^d \to$ [-M, M] belongs to the anisotropic Hölder Class $\mathbb{H}_d(\vec{\beta}, L, M)$ if for all $x, x_0 \in [0,1]^d$*

$$\begin{aligned} |s(x) - P(s)(x - x_0)| &\le L \sum_{j=1}^{d} |x_j - x_{0,j}|^{\beta_j} \text{ and} \\ \sum_{p\in\mathcal{S}_{\lfloor\beta\rfloor}} \sup_{x\in[0,1]^d} \left| \frac{\partial^{|p|} s(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right| &\le M, \end{aligned}$$

where $P(s)(x - x_0)$ is the Taylor polynomial of s of order $\lfloor\beta\rfloor$ at x_0, and x_j and $x_{0,j}$ are the jth components of x and x_0, respectively.

We distinguish two cases in the following: First, we consider the special case of isotropic Hölder classes, that is, $\beta_1 = \ldots = \beta_d$. These classes require only one common bandwidth for all dimensions that is chosen with the standard version of Lepski's method (see [19] and [21]). Afterwards, we allow for anisotropic Hölder classes. These classes necessitate a separate bandwidth for every dimension of the domain under consideration. The standard version of Lepski's Method is not applicable because it requires a monotonous bias. We circumvent this problem using a modified version of Lepski's method as described in [16] and [20].

3.1. A Fully Adaptive Estimator for Isotropic, Locally Polynomial Functions

We first allow for functions that can be approximated locally by polynomials but restrict ourselves to isotropic Hölder classes. Therefore, only one bandwidth $h_{\text{iso}} = h_1 = \ldots = h_d > 0$ has to be selected. Geometrically, this means that we select a hypercube in $\mathbb{R}^d$ with edge length h_{iso} as domain of interest (in contrast to the anisotropic case where we select a hyperrectangle with edge lengths $h_1, \ldots, h_d$).

A major issue is the choice of the bandwidth. Unfortunately, we cannot apply Lepski's method directly since the variance $\mathrm{V}(\lambda_{h_{\text{iso}}})/(nh_{\text{iso}}^d)$ for (cf. Definition 2.4)

$$\lambda_{h_{\text{iso}}}(x,y,f) := \lambda(x,y,f) := \rho\big(y - f(x)\big)\, K_{h_{\text{iso}}}(x) \quad \text{for all } x, y \in \mathbb{R}$$

(or an estimate of it as, for example, in (2.13)) is not necessarily monotonous with respect to the bandwidth (see also the next section and Section 4). We can circumvent this problem with a redefinition of the variance term. For this, we introduce the set of bandwidths $\mathcal{H}^{\text{iso}} := [h_{\min}, h_{\max}]$, where $h_{\min}$ and $h_{\max}$ are defined in (2.3), and we introduce the maximal variance for any $\rho \in \Upsilon$ and $K \in \mathcal{K}$ (see (2.11))

$$\mathrm{V}_{\max}(\rho, K) := \sup_{h_{\text{iso}} \in \mathcal{H}^{\text{iso}}} \mathrm{V}(\lambda_{h_{\text{iso}}}). \tag{3.2}$$

The variance V is defined in (2.6) and $\lambda_{h_{\text{iso}}} := \lambda$ is defined according to (2.4) with $h := (h_{\text{iso}}, \ldots, h_{\text{iso}})$. The modified variance term $\mathrm{V}_{\max}(\rho, K)$ does not depend on h_{iso}. On the one hand, we may lose considerably taking the supremum with respect to h_{iso}, on the other hand, this allows us to avoid more restrictive assumptions on the design and the noise. This is detailed in Section 4. We now define, for any $\rho \in \Upsilon$, $K \in \mathcal{K}$, and λ, the new oracle as

$$(\bar\rho^*, \bar K^*) := \arg \min_{\rho \in \Upsilon,\, K \in \mathcal{K}} \mathrm{V}_{\max}(\rho, K) \tag{3.3}$$

and the estimator of the variance as

$$\widehat{\mathrm{V}}_{\max}(\rho, K) := \sup_{h_{\text{iso}} \in \mathcal{H}^{\text{iso}}} \widehat{\mathrm{V}}(\lambda_{h_{\text{iso}}}), \tag{3.4}$$

where $\widehat{\mathrm{V}}(\lambda_{h_{\text{iso}}})$ is defined in (2.13). We then select a contrast and a kernel according to

$$(\bar\rho, \bar K) := \arg \min_{\rho \in \Upsilon,\, K \in \mathcal{K}} \widehat{\mathrm{V}}_{\max}(\rho, K) \tag{3.5}$$

and introduce the isotropic M-estimator as

$$\hat f_{\text{iso}}^{h_{\text{iso}}} := \arg \min_{f \in \mathcal{F}} n^{-1} \sum_i \bar\rho(Y_i - f(X_i)) \bar K_{(h_{\text{iso}}, \ldots, h_{\text{iso}})}(X_i). \tag{3.6}$$

Eventually, we set $\mathcal{H}_\epsilon^{\text{iso}} := \left\{h_{\text{iso}} \in \mathcal{H}^{\text{iso}},\ \exists m \in \mathbb{N} \,:\, h_{\text{iso}} = h_{\max}\epsilon^m\right\}$, $\epsilon \in]0,1[$, a net on the set of bandwidths $\mathcal{H}^{\text{iso}}$ such that $|\mathcal{H}_\epsilon^{\text{iso}}| \leq n$ and apply Lepski's method for isotropic functions (see [19] and [21]) to define the data-driven bandwidth $\hat{h}_{\text{iso}}$:

$$\hat{h}_{\text{iso}} := \max\left\{ h_{\text{iso}} \in \mathcal{H}_\epsilon^{\text{iso}} \ :\ \left|\hat{f}_{\text{iso}}^{h_{\text{iso}}}(x_0) - \hat{f}_{\text{iso}}^{h'_{\text{iso}}}(x_0)\right| \leq 20(B_0 + \text{iso}_\epsilon(n))\frac{\sqrt{\widehat{\mathrm{V}}_{\max}(\bar{\rho},\bar{K})}}{\sqrt{n(h'_{\text{iso}})^d}}, \right.$$
$$\left. \text{for all } h'_{\text{iso}} \in \mathcal{H}_\epsilon^{\text{iso}} \text{ such that } h'_{\text{iso}} \leq h_{\text{iso}} \right\}, \qquad (3.7)$$

where $\text{iso}_\epsilon(n) := 11\sqrt{\ln(n|\mathcal{H}_\epsilon^{\text{iso}}|)}$.

We now obtain on isotropic Hölder classes

$$\mathbb{H}_d^{\text{iso}}(\beta, L, M) := \mathbb{H}_d((\beta,\dots,\beta), L, M), \quad \text{for all } \beta, L, M > 0 \qquad (3.8)$$

the following result:

Theorem 3. *For n sufficiently large (according to Conditions 1, 2, and 3 in Section 5.1), $x_0 \in (0,1)^d$, $\beta \in [0,b]$, and $L > 0$, we have*

$$\mathcal{R}_{n,q}\left[\hat{f}_{\text{iso}}^{\hat{h_{\text{iso}}}}(x_0), \mathbb{H}_d^{\text{iso}}(\beta, L, M)\right] \leq \mathcal{C}_q^{\text{iso}} \inf_{h_{\text{iso}} \in \mathcal{H}^{\text{iso}}} \left\{ Ld\, h_{\text{iso}}^\beta + (B_0 + \text{iso}_\epsilon(n))\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*, \bar{K}^*)}}{\sqrt{nh_{\text{iso}}^d}} \right\}^q + o(1/n),$$

for a constant $\mathcal{C}_q^{\text{iso}}$ ($\mathcal{C}_q^{\text{iso}} = \frac{2^{q-1}}{\epsilon^{d/2}}\left(\frac{2\beta}{d} \vee \frac{d}{2\beta}\right)[40^q + 2C_q]$ works).

This result has the flavor of an oracle inequality: the first term on the right hand side is supposed to be a bound of the smallest possible pointwise risk, whereas the second term $o(1/n)$ is, at least asymptotically, insignificant. The latter is justified by the following corollary:

Corollary 3. *Under Conditions of the previous theorem and if $\mathrm{V}_{\max}(\bar{\rho}^*, \bar{K}^*) > 0$, we have*

$$\limsup_{n\to\infty} \sup_{\beta>0,\, L>0} \left(\frac{n}{(B_0 + \text{iso}_\epsilon(n))\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*, \bar{K}^*)}}\right)^{\beta/(2\beta+d)} \mathcal{R}_{n,q}\left[\hat{f}_{\text{iso}}^{\hat{h}_{\text{iso}}}(x_0), \mathbb{H}_d^{\text{iso}}(\beta, L, M)\right] < \infty.$$

This corollary can be deduced minimizing the first term on the right hand side of the last theorem by the usual bias/variance trade-off.

Remark 5. *This corollary shows that our estimator is simultaneously S- and D-adaptive. We note that this result generalizes results in [7] (that rely on the asymptotic equivalence of the block median method) to heteroscedastic regression with random design. We also stress that our estimator does not require positive noise densities at their median and thus allows for more general noise densities. Additionally, the choice of the contrast is Huber minimax (see Corollary 2 and [12]). We also note that Lepski's method has been used for locally M-estimators in [24], but not to locally polynomial M-estimators as it is done here. We can finally deduce the rate $(\ln(n)/n)^{\beta/(2\beta+1)}$ in the above result from the entropy calculations in Section 6.1. This rate is asymptotically nearly optimal (see [5] and [19]); the additional factor $\ln(n)$ is the usual price to pay in pointwise adaptive estimation. This is discussed in more detail in Section 4.*

Remark 6. *Note that our estimator detects the presence of noise or not. Indeed, the maximal variance* (3.2) *vanishes when the noise level is zero. The threshold term, in Lepski's procedure* (3.7)*, also vanishes and the procedure then selects a small bandwidth (maybe the smallest one). Our estimator thus has a small bias and novariance, that is, only a simple approximation of the target.*

3.2. A Fully Adaptive Estimator for Anisotropic, Locally Constant Functions

In this part, we allow for anisotropic Hölder classes and for (possibly) separate bandwidths for each dimension. In return (see Section 4), we restrict ourselves to locally constant functions, that is, $b = 0$ (and thus $|\mathcal{P}| = 1$) and $\mathcal{F} = [\text{-}M, M]$, and we restrict ourselves to the uniform design $\mu(\cdot) \equiv 1$ with a homoscedastic noise $\sigma(\cdot) \equiv \sigma \geq 0$. We introduce an S- and D-adaptive estimator of $f^*(x_0)$ in this setting and give its main properties in Theorem 4. The results are, in particular, applicable to linear estimators, or more generally, to M-estimators with two times differentiable contrasts.

We first introduce an estimator for each $h \in \mathcal{H}$. For this, we define the variance

$$\mathrm{V}(\rho, K) := \left(\frac{\sqrt{\int \left[\rho'(\sigma z)\right]^2 \frac{1}{n}\sum_{i=1}^n g_i(z)dz} + \|\rho'\|_\infty \|K\|_\infty (nh_{\min}^d)^{-1/4}}{\int \rho''(\sigma z)\frac{1}{n}\sum_{i=1}^n g_i(z)dz} \right)^2, \tag{3.9}$$

which is independent of the bandwidth h. As above, we then introduce the oracle for a set of contrasts Υ and a set of kernels $\mathcal{K}$ as

$$(\rho^*, K^*) := \arg \min_{\rho \in \Upsilon,\, K \in \mathcal{K}} \mathrm{V}(\rho, K). \tag{3.10}$$

Next, we introduce an estimator of the variance as

$$\widehat{\mathrm{V}}(\rho, K) := \widehat{\mathrm{V}}(\lambda_{h_{\max}}), \tag{3.11}$$

where $\widehat{V}(\lambda)$ is defined in (2.13) and $\lambda_{h_{\max}}(x, y, f) := \rho\big(y - f(x)\big) K_{h_{\max}}(x)$. The data-driven selection of the contrast and the kernel is finally

$$(\hat{\rho}, \hat{K}) := \arg \min_{\rho \in \Upsilon,\, K \in \mathcal{K}} \widehat{\mathrm{V}}(\rho, K), \tag{3.12}$$

and, similarly to (2.4) and (2.5), the estimator is

$$\hat{f}^h := \arg\min_{f \in \mathcal{F}} n^{\text{-}1} \sum_i \hat{\rho}(Y_i - f(X_i))\hat{K}_h(X_i) \tag{3.13}$$

for all $h \in (0,1)^d$. It is again necessary that $(\hat{\rho}, \hat{K})$ does not depend on the bandwidth h; we discuss this point in Section 4.

Eventually, we can describe the choice of the bandwidth h with Lepski's method. For this, we define for all $a, b \in \mathbb{R}$ the scalar $a \vee b := \max(a, b)$ and for all $h, h' \in (0,1)^d \times (0,1)^d$ the vector $h \vee h' := (h_1 \vee h'_1, \ldots, h_d \vee h'_d)$. We then consider the two families of Locally Constant Approximation (LCA) estimators

$$\left\{\hat{f}^h\right\}_{h \in (0,1)^d} \quad \text{and} \quad \left\{\hat{f}^{h,h'} := \hat{f}^{h \vee h'}\right\}_{h,h' \in (0,1)^d \times (0,1)^d},$$

where $\hat{f}^h$ is defined in (3.13). Note that $\hat{f}^{h,h'} = \hat{f}^{h',h}$ by symmetry. Recall the definition of the set of bandwidths $\mathcal{H} := [h_{\min}, h_{\max}]^d$, where $h_{\min}$ and $h_{\max}$ are defined in (2.3). Additionally, we introduce an order $\preceq$ on $\mathcal{H}$ such that

$$h \preceq h' \quad \Leftrightarrow \quad \prod_{j=1}^{d} h_j \leq \prod_{j=1}^{d} h'_j.$$

In particular, the variance is decreasing on this order. We finally introduce a net $\mathcal{H}_\epsilon := \{h_{\min}\} \cup \{h \in \mathcal{H} \,:\, \forall j = 1, \ldots, d,\ \exists m_j \in \mathbb{N} \,:\, h_j = h_{\max}\epsilon^{m_j}\},\ \epsilon \in]0,1[$ (where we assume that $|\mathcal{H}_\epsilon| \leq n$), set $\mathrm{ani}_\epsilon(n) := 11\sqrt{\ln(n|\mathcal{H}_\epsilon|)}$, and select the bandwidth according to

$$\hat{h} := \max_{\preceq} \left\{ h \in \mathcal{H}_\epsilon \,:\, \left| \hat{f}^{h,h'}(x_0) - \hat{f}^{h'}(x_0) \right| \leq 16(B_0 + \mathrm{ani}_\epsilon(n)) \frac{\sqrt{\widehat{\mathrm{V}}(\hat{\rho}, \hat{K})}}{\sqrt{n\Pi_{h'}}}, \right.$$
$$\left. \text{for all } h' \in \mathcal{H}_\epsilon \text{ such that } h' \preceq h \right\}, \tag{3.14}$$

where the maximum is taken with respect to the order $\preceq$, $\widehat{\mathrm{V}}_{\max}(\cdot,\cdot)$ and $(\hat{\rho}, \hat{K})$ are defined in (3.4) and (3.12), and B_z is defined in (2.14).

We can now give the following result for the estimator $\hat{f}^{\hat{h}}$:

Theorem 4. *If n is sufficiently large (according to Conditions 1, 2, and 3 in Section 5.1), $x_0 \in (0,1)^d$, $\vec{\beta} \in (0,1]^d$, and $L > 0$, then, it holds that*

$$\mathcal{R}_{n,q}\big[\hat{f}^{\hat{h}}(x_0), \mathbb{H}_d(\vec{\beta}, L, M)\big] \leq \mathcal{C}_q \inf_{h \in \mathcal{H}} \left\{ L \sum_{j=1}^{d} h_j^{\beta_j} + (B_0 + \mathrm{ani}_\epsilon(n)) \frac{\sqrt{\mathrm{V}(\rho^*, K^*)}}{\sqrt{n\Pi_h}} \right\}^q + o(1/n)$$

for a constant $\mathcal{C}_q$ ($\mathcal{C}_q = \frac{2d\epsilon^{-d/2}}{\min_j \beta_j}$ $[5q\,\mathrm{Gamma(q)}\,1152^q]$ works).

We can also derive the following corollary from Theorem 4 via a bias/variance trade-off:

Corollary 4. *Let $\bar{\beta} := \left(\sum_j 1/\beta_j\right)^{-1}$ be the harmonic average. Under the conditions of the previous theorem and if $\mathrm{V}(\rho^*, K^*) > 0$, it holds that*

$$\limsup_{n\to\infty} \sup_{\vec{\beta}\in(0,1]^d,\, L>0} \left(\frac{n}{(B_0 + \mathrm{ani}_\epsilon(n))\sqrt{\mathrm{V}(\rho^*, K^*)}} \right)^{\bar{\beta}/(2\bar{\beta}+1)} \mathcal{R}_n\big[\hat{f}^{\hat{h}}(x_0), \mathbb{H}_d(\vec{\beta}, L, M)\big] < \infty.$$

This corollary can be deduced minimizing the first term on the right hand side of the last theorem by the usual bias/variance trade-off.

Remark 7. *This transfers the results of the previous section to anisotropic Hölder classes. However, as opposed to the previous results, the above corollary only allows for locally constant functions. Moreover, we note that this is, to the best of our knowledge, the first application of [20]'s Method to select an anisotropic bandwidth for nonlinear M-estimators. We discuss this in Section 4. Finally, we refer to the remarks after Theorem 3. The adaptive S-minimax rate $(\ln(n)/n)^{\bar{\beta}/(2\bar{\beta}+1)}$ follows from the definition of is nearly optimal. The optimal rate is given by [17] in the white noise model for anisotropic Hölder functions.*

4. Discussion

Let us detail on the assumptions and restrictions and highlight some open problems:

1. The symmetry assumption on our model (2.1) (cf. [12], [25]) leads to $\mathbb{E}_{f^*}(\sum_i \rho'(\xi_i)) = 0$. We stress that we only assume that the sum $\sum_i g_i(\cdot)$ is symmetric. This is satisfied, of course, if all densities $g_i(\cdot)$ are symmetric, but this may not be the case. The symmetry assumption can be replaced in the proof of Proposition 1 (control of the deviations of M-estimators) if the expectation stays very small, that is, $\mathbb{E}_{f^*}(\rho'(\xi)) < n^{\text{-}1}$. To ensure small expectations for asymmetric sums of densities, we expect that an asymmetric contrast has to be chosen. This seems to be an interesting but hard problem.
2. It is well-known that the median is very sensitive to the noise density at 0. Indeed, its variance is $1/(4g^2(0))$. The value of $g(0)$ is estimated in [7], for example, but in practice, this requires many observations near the location. On the contrary, contrasts as in Definition 1 (the Huber contrast with a scale γ, for example) depend on the mass of the noise density on the interval $[\text{-}\gamma, \gamma]$ (denominator of the variance (2.6)). Moreover, note that the term in assumption (2.2) is, up to ρ''_{min}, a lower bound of the denominator of the variance in (2.6). Therefore, the parameter $\gamma_{\min}$ can be estimated for a given A similarly as the denominator of the variance. The mentioned assumption guarantees the consistence of M-estimators with a contrast strictly convex on the interval $[\text{-}\gamma_{\min}, \gamma_{\min}]$. Additionally, if the parameter $\gamma_{\min}$ is chosen as a function of A such that there is a sufficiently large mass is on the appropriate interval, it guarantees a good estimation of the variance for all $\gamma \geq \gamma_{\min}$. However, we note that A is expected to require a calibration in practice.
3. Conditions 2 and 3 on n in the following section are only introduced to simplify the residual terms in the proofs. However, the first assumption in Condition 1 is crucial. We recall that $b_{h_{\max}}$ is the bias and $A\rho''_{\min}$ is a lower bound of the denominator of the variance, that is, the mass of the noise density on $[\text{-}\gamma_{\min}, \gamma_{\min}]$. Condition 1 thus means that this mass has to be larger than the bias. This ensures that the denominator of the variance is not too small and thus that the estimator is consistent (cf. Lemma 7).
4. To estimate the variance of M-estimators (2.6), we use its empirical version but the residuals stays unknown. To solve this problem, we notice $Y_i - \hat{f}_\lambda$ is an estimate of $\sigma(X_i)\xi_i$ if and only if $\hat{f}_\lambda$ is a consistent estimator of f^*. The assumption (2.2) is assumed to guarantee the consistence of all of estimators in Λ. However, a pre-estimator could be used (for example the contrast associated to the arctan function as defined below Definition 1) and thus a more general family of estimators could be considered (with some of them nonconsistent).
5. We do not assume any conditions on the design and the noise level except for the boundedness of the noise level. The design density and the noise level could be zero or explode at x_0. This can be detected via the variance (2.6) if the rate of convergence is influenced (see [9] for degenerate design). However, the design and the noise level could compensate each other such that no effect is visible the variance term. This is a very interesting point and could be studied in the future.
6. Lepski's method is very sensitive to outliers (see [24]). In this paper, however, we chose the robustness via the minimization of the variance. This could be interesting for many applications.
7. In Section 3.2, we present anisotropic results for pointwise estimation in heteroscedastic regression with heavy tailed noises and random designs. To the best of this knowledge, this is the first result of this kind for nonlinear M-estimators in our framework. We note, however, that we have to restrict ourselves to locally constant M-estimators because of the bias term (cf. Lemma 12).

8. We allow in this paper for a selection of the contrast and the kernel from large families. Additionally, we can extend the family of contrast allowing for a selection the support of the kernel (not necessarily centered at x_0). This could be interesting (cf. [2, 10]) especially for applications. Furthermore, we could add, for example, the selection of the tail of the contrast. Such extensions are only limited by the required convexity of the contrast and the complexity of the selection. Indeed, we need that contrast is convex and strictly convex around 0 to ensure that the denominator of the variance (2.6) is positive. We think that this has to be studied further.
9. We obtain the desired variance in Theorems 1 and 2 up to the constants $\breve{B}_0$ and B_0, respectively (cf. Remark 1). These constants are mostly due to Dudley's integral that is a part of the deviation inequalities from [22] we use. We expect that these constants can be reduced with a refined analysis.
10. The variance and the choice of the contrast and the kernel do not depend on the bias term (see Theorem 1, (2.13), and Remark 2) and, more generally, do not depend on what we estimate. This is an interesting point because this allows for a treatment of other problems as in high dimensional settings. In [18], for example, the Huber loss with an ℓ_1 penalization is studied. They show that the shape of the tuning parameter is similar to the variance of M-estimators (cf. (1.1)). We thus expect that our results on the choice of the contrast can be applied in high dimensional settings.
11. The simultaneous D- and S-adaptation is a hard problem especially because the variance (2.6) depends on the bandwidth which is the parameter of interest in S-adaptation (see Section 3). Lepski's method requires a decreasing variance with respect to the bandwidth, but unfortunately, this is not always the case in heteroscedastic regression. For example, the noise level could be zero in a neighborhood V_h of x_0 and huge on the set $V_{h'} \setminus V_h$, where $V_{h'}$ is a bigger neighborhood of x_0. This would imply that the variance increases. To avoid such problems, we propose to maximize the variance with respect to h (see Section 3), but this is a very conservative approach. Models with a homoscedastic noise and a uniform designs do not have these issues (cf. Section 3.2). It may also happen that the design and the noise level are such that the variance is decreasing and thus Lepski's method is applicable without problems.
12. From the computation of the entropy (in Section 6.1) and the definition of $\mathrm{iso}_\epsilon(n)$, the shape of $20(B_0 + \mathrm{iso}_\epsilon(n))$ in the threshold term in (3.7) is $C\ln(n)$ where C is a positive and known constant but large. An appropriate value for applications is rather between 1 and 2 (see [21]). Usually, such quantities are calibrated with cross-validation or similar methods. Moreover, we showed in Corollary 3 that our estimator achieves the minimax rate up to a factor $\ln(n)$. As mentioned in Remark 5, this is due to the threshold term in the selection rule (3.7) and is nearly optimal. Indeed, the optimal factor is $(b-\beta)\ln(n)$ in a certain sense (see [17]). To achieve this optimality, the term $\mathrm{iso}_\epsilon(n)$, in (3.7), has to be proportional to $\ln(h_{\max}/h_{\mathrm{iso}})$ (see [21]). The same remark applies to the anisotropic rate in Corollary 4 (see [17]). The optimality of these rates is only proved in the white noise model (see [5, 17, 19]), but we conjecture that they are nearly S-minimax optimal in more general settings (for all of models where the Fisher information exists, for example).

5. Proofs of the Main Results

Let us introduce some additional notation to simplify the exposition. First, we introduce the best approximation of the target f^* in $\mathcal{F}$:

$$f^0 := \arg\min \left\{ \sup_{x \in V_h} \left| f(x) - f^*(x) \right| \ : \ f \in \mathcal{F}, f(x_0) = f^*(x_0) \right\}. \tag{5.1}$$

The minimum is not necessarily unique, but all minimizers work for our derivations. We then set $t^0 := t^0(f^*, x_0, h) := \{t_p^0 : \ p \in \mathcal{P}\}$ and $f^0 := \mathrm{P}_{t^0}$. Next, we denote the vector of the monomials $(x - x_0)^p/h^p$ of order smaller or equal than b by $\mathbb{X}$ and the smallest eigenvalue of the matrix $\int \mathbb{X}^\top \mathbb{X}\mu(x)K_h(x)dx$ by ν. This allows us to define the set

$$\mathcal{F}_{\delta_n} := \left\{f = \mathrm{P}_t \ : \ \|t - t^0\|_{\ell_1} \leq \delta_n\right\}, \tag{5.2}$$

where

$$\delta_n := 2|\mathcal{P}|^{3/2}(A\rho''_{\min}\nu)^{-1}\left((\ln n)^{-1} + b_h \int K_h(x)\mu(x)dx\right).$$

Furthermore, we denote the vector of partial derivatives of the λ-criterion $P_n\lambda(\cdot)$ (defined in (2.4)) by

$$\tilde{D}_\lambda(\cdot) := \left(-\frac{\partial}{\partial t_p}\, P_n\lambda(\cdot)\right)^\top_{p\in\mathcal{P}} \tag{5.3}$$

and the corresponding expectation and the "parametric" expectation with respect to the distribution P^0 of $(X, f^0(X) + \sigma(X)\xi)$ by

$$P\left[\tilde{D}_\lambda(\cdot)\right] \quad \text{and} \quad P^0\left[\tilde{D}_\lambda(\cdot)\right], \tag{5.4}$$

respectively. Next, we introduce the *Jacobian matrix* J_D of $P^0\left[\tilde{D}_\lambda\right]$ as

$$\left(J_D(\cdot)\right)_{p,q\in\mathcal{P}} := \left(\frac{\partial}{\partial t_q}P^0\left[\tilde{D}^p_\lambda(\cdot)\right]\right)_{p,q\in\mathcal{P}} = \left(\frac{\partial}{\partial t_q}P^0\left[-\frac{\partial}{\partial t_p}\, P_n\lambda(\cdot)\right]\right)_{p,q\in\mathcal{P}}, \tag{5.5}$$

where $\tilde{D}^p_\lambda(\cdot)$ is the pth component of $\tilde{D}_\lambda(\cdot)$. The Jacobian matrix exists according to Definition 1 and Fubini's Theorem. Furthermore, the sup-norm on $\mathbb{R}^{|\mathcal{P}|}$ is denoted by $\|\cdot\|_{\ell_\infty}$, and the vector of coefficients of the estimated polynomial $\hat{f}_\lambda$ is denoted by $\hat{t}_\lambda$. Finally, we set $\tilde{a}_n := \max\left\{\sqrt{b_{h_{\max}}}, (\ln n)^{-1}\right\}$ and

$$a_n := \frac{(1 + \tilde{a}_n)\sqrt{1 + \tilde{a}_n}}{(1 - \tilde{a}_n)\sqrt{1 - \tilde{a}_n}}, \tag{5.6}$$

and

$$c_\lambda := P\lambda''(f^*). \tag{5.7}$$

By Definition 1, it holds that $\inf_{\lambda\in\Lambda} c_\lambda > 0$.

5.1. Conditions on n

Condition 1: We assume that n is sufficiently large such that for all $\lambda \in \Lambda$

$$\sqrt{b_{h_{\max}} + \delta_n} \leq \frac{1}{2} \wedge \frac{A\rho''_{\min}}{2L_{\rho''}} \quad \text{and} \quad \gamma_{\max}\mathcal{K}_{\max}B_0 \leq \sqrt{nh^d_{\min}}(2\ln n)^{-1}.$$

Condition 2: We assume that n is sufficiently large such that for all $\lambda \in \Lambda$ and all $h \in \mathcal{H}$

$$(2\vee L_{\rho''})\gamma_{\max}\mathcal{K}_{\max}(\delta_n{+}b_h){+}\frac{\gamma^2_{\max}\mathcal{K}^2_{\max}B_{\ln(n)}}{\sqrt{nh^d_{\min}}} \leq \tilde{a}_n \max\left\{A\rho''_{\min}\int K_h(x)\mu(x)dx, \Pi_h \inf_{\lambda\in\Lambda} P\left[\lambda'(f^*)\right]^2\right\}$$

and $\tilde{a}_n < 1/3$, where $\tilde{a}_n$ is defined in (5.6).

Condition 3: We assume that n is sufficiently large such that

$$2\ln(n) \leq \frac{nh_{\min}^d/(4\ln^2 n)}{98\gamma_{\max}^2\mathcal{K}_{\max}^2 + 4\gamma_{\max}\mathcal{K}_{\max}}, \quad \frac{11\sqrt{2\ln(n)}}{(nh_{\min}^d)^{1/4}} \leq 1, \quad \text{and} \quad \frac{2\gamma_{\max}\mathcal{K}_{\max}}{A\rho''_{\min}(nh_{\min}^d)^{1/4}} \leq 1.$$

5.2. An Auxilliary Result

Proposition 1. *Let Λ be a set of functions as in (2.4) such that $\mathcal{H}_{\mathcal{F}\cup\Lambda,\omega} < \infty$, and let n be sufficiently large (according to Condition 1 above). Then, for any $z > 0$ and any $h \in \mathcal{H}$,*

$$\mathbb{P}_{f^*}\left(\left\{\sup_{\lambda\in\Lambda}\left[\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right| - 2\frac{\sqrt{\mathrm{V}(\lambda)}\,B_z}{\sqrt{n\Pi_h}}\right] \geq 3b_h\right\} \cap \bigcap_{\lambda\in\Lambda}\left\{\hat{f}_\lambda \in \mathcal{F}_{\delta_n}\right\}\right) \leq 2|\mathcal{P}|\exp(\text{-}z),$$

where B_z is defined in (2.14).

Recall that λ depends on the bandwidth h, which is fixed here. We also note that the constants 2 and 3 can be replaced by o(1). Finally, if only one fixed function $\lambda \in \Lambda$ is considered, the expressions simplify considerably as we show in the following lemma:

Lemma 3. *Let $\lambda \in \Lambda$ be fixed, $\mathcal{H}_{\mathcal{F},\nu} < \infty$, and let n be sufficiently large (according to Condition 1 above). Then, for any $\varepsilon > 0$ and any $h \in \mathcal{H}$,*

$$\mathbb{P}_{f^*}\left(\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right| \geq 2\frac{\sqrt{\mathrm{V}(\lambda)}\,\breve{B}_\varepsilon}{\sqrt{n\Pi_h}} + 3b_h,\ \hat{f}_\lambda \in \mathcal{F}_{\delta_n}\right) \leq 2|\mathcal{P}|\exp\left(-\frac{\varepsilon^2}{100 + \frac{4\varepsilon}{(n\Pi_h)^{1/4}}}\right),$$

where $\breve{B}_\varepsilon$ is defined in (2.10).

This claim can be deduced similarly as Proposition 1, but one has to choose z such that $\varepsilon = 10\sqrt{z} + \frac{2z}{(n\Pi_h)^{1/4}}$.

5.3. Proof of Theorem 1

First, we recall that $\sup_{f\in\mathcal{F}}\|f\|_\infty \leq |\mathcal{P}|M$ and set

$$\Omega := \left\{\forall\lambda \in \Lambda,\ \hat{f}_\lambda \in \mathcal{F}_{\delta_n}\right\} \quad \text{and} \quad \Omega^c := \left\{\exists\lambda \in \Lambda,\ \hat{f}_\lambda \notin \mathcal{F}_{\delta_n}\right\}. \tag{5.8}$$

Then, since $\hat{f}_\lambda \in \mathcal{F}$ and $\|f^*\|_\infty \leq M$, the risk can be bounded by

$$\begin{aligned}\mathbb{E}_{f^*}\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right|^q &= \mathbb{E}_{f^*}\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right|^q\mathbb{1}_\Omega + \mathbb{E}_{f^*}\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right|^q\mathbb{1}_{\Omega^c}\\ &\leq \mathbb{E}_{f^*}\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right|^q\mathbb{1}_\Omega + ((1+|\mathcal{P}|)M)^q\mathbb{P}_{f^*}(\Omega^c).\end{aligned}$$

Using Lemma 7, Lemma 8, the last inequality, and simple computations, we obtain

$$\begin{aligned}&\mathbb{E}_{f^*}\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right|^q\\ &\quad\leq \mathbb{E}_{f^*}\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right|^q\mathbb{1}_\Omega + ((1+|\mathcal{P}|)M)^q 2|\mathcal{P}|\exp\left(-\frac{n\Pi_h/(4\ln^2 n)}{98\gamma_{\max}^2\mathcal{K}_{\max}^2 + 4\gamma_{\max}\mathcal{K}_{\max}}\right)\\ &\quad\leq 2^q\mathbb{E}_{f^*}\left(\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right| - 3b_h - \frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_0}{\sqrt{n\Pi_h}}\right)_+^q\mathbb{1}_\Omega \quad + 2^q\left(3b_h + \frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_0}{\sqrt{n\Pi_h}}\right)^q\\ &\qquad + ((1+|\mathcal{P}|)M)^q 2|\mathcal{P}|\exp\left(-\frac{n\Pi_h/(4\ln^2 n)}{98\gamma_{\max}^2\mathcal{K}_{\max}^2 + 4\gamma_{\max}\mathcal{K}_{\max}}\right).\end{aligned} \tag{5.9}$$

Let us now bound the first term on the right hand side of the last inequality. To do so, we use simple computations to obtain

$$\begin{aligned}
&\mathbb{E}_{f^*}\left(\left|\hat{f}_\lambda(x_0)-f^*(x_0)\right|-3b_h-\frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_0}{\sqrt{n\Pi_h}}\right)_+^q \mathbb{1}_\Omega \\
&\quad = q\int_0^\infty z^{q-1}\mathbb{P}_{f^*}\left(\left|\hat{f}_\lambda(x_0)-f^*(x_0)\right|-3b_h-\frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_0}{\sqrt{n\Pi_h}}\geq z,\,\Omega\right)dz.
\end{aligned}$$

Setting $z=\frac{2\sqrt{\mathrm{V}(\lambda)}}{\sqrt{n\Pi_h}}\varepsilon$ in the last inequality, using the definition of $\breve{B}_\varepsilon$, and Corollary 3, we get

$$\begin{aligned}
&\mathbb{E}_{f^*}\left(\left|\hat{f}_\lambda(x_0)-f^*(x_0)\right|-3b_h-\frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_0}{\sqrt{n\Pi_h}}\right)_+^q \mathbb{1}_\Omega \\
&\quad = q\left(\frac{2\sqrt{\mathrm{V}(\lambda)}}{\sqrt{n\Pi_h}}\right)^q\int_0^\infty \varepsilon^{q-1}\mathbb{P}_{f^*}\left(\left|\hat{f}_\lambda(x_0)-f^*(x_0)\right|\geq 3b_h+\frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_\varepsilon}{\sqrt{n\Pi_h}},\Omega\right)d\varepsilon \\
&\quad \leq 2q|\mathcal{P}|\left(\frac{2\sqrt{\mathrm{V}(\lambda)}}{\sqrt{n\Pi_h}}\right)^q\int_0^\infty \varepsilon^{q-1}\exp\left(-\frac{\varepsilon^2}{100+4\varepsilon}\right)d\varepsilon \\
&\quad \leq 2q|\mathcal{P}|\left(3b_h+\frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_0}{\sqrt{n\Pi_h}}\right)^q\int_0^\infty \varepsilon^{q-1}\exp\left(-\frac{\varepsilon^2}{100+4\varepsilon}\right)d\varepsilon.
\end{aligned}$$

One may then check that for any $a,b>0$ and any $q\geq 1$

$$\int_0^\infty \epsilon^{q-1}e^{-\frac{\epsilon^2}{a+b\epsilon}}\,\mathrm{d}\,\epsilon \leq 1+(a+b)^{q/2}\,\mathrm{Gamma}(q), \tag{5.10}$$

so that

$$\mathbb{E}_{f^*}\left(\left|\hat{f}_\lambda(x_0)-f^*(x_0)\right|-3b_h-\frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_0}{\sqrt{n\Pi_h}}\right)_+^q \mathbb{1}_\Omega \leq 2q|\mathcal{P}|\left(3b_h+\frac{2\sqrt{\mathrm{V}(\lambda)}\breve{B}_0}{\sqrt{n\Pi_h}}\right)^q(11.2)^q\,\mathrm{Gamma}(q)$$

where $\mathrm{Gamma}(\cdot)$ is the usual Gamma function. From (5.9) and the last inequalities, the theorem can be deduced. ■

5.4. Proof of Theorem 2

First, we set for all $h\in\mathcal{H}$

$$\Delta := \bigcap_{\lambda\in\Lambda}\left\{\sqrt{\widehat{\mathrm{V}}(\lambda)}\in\left[\frac{\sqrt{1-\tilde{a}_n}}{1+\tilde{a}_n}\sqrt{\mathrm{V}(\lambda)},\frac{\sqrt{1+\tilde{a}_n}}{1-\tilde{a}_n}\sqrt{\mathrm{V}(\lambda)}\right]\right\}. \tag{5.11}$$

Then, we observe that, since $\hat{f}_{\hat{\lambda}}\in\mathcal{F}$ and $\|\hat{f}_{\hat{\lambda}}\|_\infty,\|f^*\|_\infty\leq M$ and $\sup_{f\in\mathcal{F}}\|f\|_\infty\leq|\mathcal{P}|M$, the risk can be bounded by

$$\begin{aligned}
\mathbb{E}_{f^*}\left|\hat{f}_{\hat{\lambda}}(x_0)-f^*(x_0)\right|^q &= \mathbb{E}_{f^*}\left|\hat{f}_{\hat{\lambda}}(x_0)-f^*(x_0)\right|^q\mathbb{1}_\Delta+\mathbb{E}_{f^*}\left|\hat{f}_{\hat{\lambda}}(x_0)-f^*(x_0)\right|^q\mathbb{1}_{\Delta^c} \\
&\leq \mathbb{E}_{f^*}\left|\hat{f}_{\hat{\lambda}}(x_0)-f^*(x_0)\right|^q\mathbb{1}_\Delta+((1+|\mathcal{P}|)M)^q\mathbb{P}_{f^*}(\Delta^c).
\end{aligned}$$

Using Lemma 9, Lemma 8, the last inequality, and simple computations, we obtain

$$\begin{aligned}
&\mathbb{E}_{f^*}\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big|^q \\
&\quad\leq \mathbb{E}_{f^*}\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big|^q \mathbb{1}_\Delta + ((1+|\mathcal{P}|)M)^q\big(2n^{-2} + \mathbb{P}_{f^*}(\Omega^c)\big) \\
&\quad\leq 2^q \mathbb{E}_{f^*}\left(\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big| - 3b_h - \frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)_+^q \mathbb{1}_\Delta \\
&\qquad + 2^q\left(3b_h + \frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)^q + ((1+|\mathcal{P}|)M)^q\big(2/n^2 + \mathbb{P}_{f^*}(\Omega^c)\big). \qquad (5.12)
\end{aligned}$$

Let us now bound the first term on the right hand side of the last inequality. To do so, we use simple computations to obtain

$$\begin{aligned}
&\mathbb{E}_{f^*}\left(\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big| - 3b_h - \frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)_+^q \mathbb{1}_\Delta \\
&\quad = q\int_0^\infty (z')^{q-1}\mathbb{P}_{f^*}\left(\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big| - 3b_h - \frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}} \geq z', \Delta\right)dz'. \qquad (5.13)
\end{aligned}$$

On the event Δ and by definition of a_n in (5.6), it yields

$$\sqrt{\mathrm{V}(\lambda^*)} \geq \frac{1-\tilde{a}_n}{\sqrt{1+\tilde{a}_n}}\sqrt{\widehat{\mathrm{V}}(\lambda^*)} \geq \frac{1-\tilde{a}_n}{\sqrt{1+\tilde{a}_n}}\sqrt{\widehat{\mathrm{V}}(\hat{\lambda})} \geq \frac{(1-\tilde{a}_n)\sqrt{1-\tilde{a}_n}}{(1+\tilde{a}_n)\sqrt{1+\tilde{a}_n}}\sqrt{\mathrm{V}(\hat{\lambda})} = a_n^{-1}\sqrt{\mathrm{V}(\hat{\lambda})} \quad (5.14)$$

Setting $z' = \frac{2a_n\sqrt{\mathrm{V}(\lambda)}}{\sqrt{n\Pi_h}}\varepsilon$ in (5.13), defining $\overline{B}_\epsilon := B_0 + \epsilon$, using the definition of B_ε, the last inequality, and Proposition 1 with $\varepsilon = 10\sqrt{z} + \frac{2z}{(n\Pi_h)^{1/4}}$, we get

$$\begin{aligned}
&\mathbb{E}_{f^*}\left(\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big| - 3b_h - \frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)_+^q \mathbb{1}_\Delta \\
&\quad = q\left(\frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}}{\sqrt{n\Pi_h}}\right)^q \int_0^\infty \varepsilon^{q-1}\mathbb{P}_{f^*}\left(\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big| \geq 3b_h + \frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}\overline{B}_\varepsilon}{\sqrt{n\Pi_h}}, \Delta\right)d\varepsilon \\
&\quad \leq q\left(\frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}}{\sqrt{n\Pi_h}}\right)^q \int_0^\infty \varepsilon^{q-1}\mathbb{P}_{f^*}\left(\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big| \geq 3b_h + \frac{2\sqrt{\mathrm{V}(\hat{\lambda})}\overline{B}_\varepsilon}{\sqrt{n\Pi_h}}, \Omega\right)d\varepsilon \\
&\quad \leq q\left(\frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}}{\sqrt{n\Pi_h}}\right)^q \int_0^\infty \varepsilon^{q-1}\mathbb{P}_{f^*}\left(\sup_{\lambda\in\Lambda}\left[\big|\hat{f}_{\lambda}(x_0) - f^*(x_0)\big| - 3b_h - \frac{2\sqrt{\mathrm{V}(\lambda)}\overline{B}_\varepsilon}{\sqrt{n\Pi_h}}\right] \geq 0, \Omega\right)d\varepsilon \\
&\quad \leq 2q|\mathcal{P}|\left(\frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}}{\sqrt{n\Pi_h}}\right)^q \int_0^\infty \varepsilon^{q-1}\exp\left(-\frac{\varepsilon^2}{100+4\varepsilon}\right)d\varepsilon \\
&\quad \leq 2q|\mathcal{P}|\left(3b_h + \frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)^q \int_0^\infty \varepsilon^{q-1}\exp\left(-\frac{\varepsilon^2}{100+4\varepsilon}\right)d\varepsilon \\
&\quad \leq 2q|\mathcal{P}|\left(3b_h + \frac{2a_n\sqrt{\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)^q * (11.2)^q\,\mathrm{Gamma}(q).
\end{aligned}$$

The last inequality is obtained from (5.10). From (5.12) and the last inequality, the theorem can be deduced. ∎

5.5. Proof of Theorem 3

For ease of exposition, we set $k := h_{\text{iso}}$ and $\hat{k} := \hat{h}_{\text{iso}}$. Then, one may verify that the *oracle bandwidth*

$$k^* := \arg\min_{k \in \mathcal{H}^{\text{iso}}} \left\{ Ld\beta^{-1}k^\beta + 2\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*, \bar{K}^*)}(B_0 + \text{iso}_\epsilon(n))}{d\sqrt{nk^d}} \right\}$$

is well defined. Moreover, let us introduce the element k_ϵ^* of the net $\mathcal{H}_\epsilon^{\text{iso}}$ such that $k_\epsilon^* \le k^* \le \epsilon^{-1}k_\epsilon^*$. Furthermore, from Condition 3 on n and Lemmas 7 and 9 with $h = (k, \dots, k)$, it follows that

$$\mathbb{P}_{f^*}\left(\exists k \in \mathcal{H}_\epsilon^{\text{iso}} \,:\, \hat{f}_{\text{iso}}^k \notin \mathcal{F}_{\delta_n}\right) \le 2|\mathcal{P}| \sum_{k \in \mathcal{H}_\epsilon^{\text{iso}}} \exp\left(-\frac{nh_{\min}^d/(4\ln^2 n)}{98\gamma_{\max}^2\mathcal{K}_{\max}^2 + 4\gamma_{\max}\mathcal{K}_{\max}}\right) \le 2|\mathcal{P}|n^{-1} \quad (5.15)$$

and

$$\begin{aligned}\sum_{k \in \mathcal{H}_\epsilon^{\text{iso}}} \mathbb{P}_{f^*}(\Delta^c) \le& \sum_{k \in \mathcal{H}_\epsilon^{\text{iso}}} \frac{2}{n|\mathcal{H}_\epsilon^{\text{iso}}|} + 2|\mathcal{P}| \sum_{k \in \mathcal{H}_\epsilon^{\text{iso}}} \exp\left(-\frac{nh_{\min}^d/(4\ln^2 n)}{98\gamma_{\max}^2\mathcal{K}_{\max}^2 + 4\gamma_{\max}\mathcal{K}_{\max}}\right)\\ \le& 4|\mathcal{P}|n^{-1}, \end{aligned} \quad (5.16)$$

where $h_{\min}$ and Δ are defined in (2.3) and (5.11), respectively. Thus, we may restrict our considerations in the following to the event $\bigcap_{k \in \mathcal{H}_\epsilon^{\text{iso}}} \left\{\hat{f}_{\text{iso}}^k \in \mathcal{F}_{\delta_n}\right\} \cap \Delta$.

Control of the risk on the event $\{k_\epsilon^* \le \hat{k}\}$. With the triangular inequality and Lemma 8, we obtain

$$\begin{aligned}&\mathbb{E}_{f^*}\left[\left|\hat{f}_{\text{iso}}^{\hat{k}}(x_0) - f^*(x_0)\right|^q \mathbb{1}_{k_\epsilon^* \le \hat{k}}\right]\\ &\le 2^{q-1}\mathbb{E}_{f^*}\left[\left|\hat{f}_{\text{iso}}^{\hat{k}}(x_0) - \hat{f}_{\text{iso}}^{k_\epsilon^*}(x_0)\right|^q \mathbb{1}_{k_\epsilon^* \le \hat{k}}\right] + 2^{q-1}\mathbb{E}_{f^*}\left|\hat{f}_{\text{iso}}^{k_\epsilon^*}(x_0) - f^*(x_0)\right|^q. \end{aligned} \quad (5.17)$$

The first term on the right hand side of the last inequality is controlled by the construction of the procedure (3.7), and thus

$$\mathbb{E}_{f^*}\left[\left|\hat{f}_{\text{iso}}^{\hat{k}}(x_0) - \hat{f}_{\text{iso}}^{k_\epsilon^*}(x_0)\right|^q \mathbb{1}_{k_\epsilon^* \le \hat{k}}\right] \le \mathbb{E}_{f^*}\left[20\frac{\sqrt{\widehat{\mathrm{V}}_{\max}(\bar{\rho}, \bar{K})}(B_0 + \text{iso}_\epsilon(n))}{\sqrt{n(k_\epsilon^*)^d}}\right]^q.$$

On the event $\bigcap_{k \in \mathcal{H}_\epsilon^{\text{iso}}} \Delta$, we get similarly as in (5.14)

$$\mathbb{E}_{f^*}\left[\left|\hat{f}_{\text{iso}}^{\hat{k}}(x_0) - \hat{f}_{\text{iso}}^{k_\epsilon^*}(x_0)\right|^q \mathbb{1}_{k_\epsilon^* \le \hat{k}}\right] \le \left(20\frac{\sqrt{1+\tilde{a}_n}}{1-\tilde{a}_n}\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*, \bar{K}^*)}(B_0 + \text{iso}_\epsilon(n))}{\sqrt{n(k_\epsilon^*)^d}}\right)^q,$$

where $\tilde{a}_n$ is defined above (5.6). Recall that, by the definitions of the Hölder classes (Definition 4), we can control the bias for any $\beta \in]0, b]$ and any $k > 0$ by

$$b_k \le \sup_{x \in V_k} |\mathrm{P}(f^*)(x - x_0) - f^*(x)| \le Ldk^\beta, \quad (5.18)$$

where $\mathrm{P}(f^*)(x-x_0)$ is the Taylor Polynomial of f^* at x_0. So we can deduce finally from Theorem 2 with $h=(k,\dots,k)$ and $b_h=b_k$ a bound for the second term in (5.17):

$$\mathbb{E}_{f^*}\left|\hat{f}_{\mathrm{iso}}^{k_\epsilon^*}(x_0)-f^*(x_0)\right|^q \leq 2C_q\left(Ld(k_\epsilon^*)^\beta+\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*,\bar{K}^*)}B_0}{\sqrt{n(k_\epsilon^*)^d}}\right)^q+o(1/n).$$

Using (5.17) and the last two inequalities, and invoking Condition 2 in Section 5.1, we have a control of the risk on the event $\{k_\epsilon^*\leq\hat{k}\}$:

$$\begin{aligned}&\mathbb{E}_{f^*}\left[\left|\hat{f}_{\mathrm{iso}}^{\hat{k}}(x_0)-f^*(x_0)\right|^q\mathbb{1}_{k_\epsilon^*\leq\hat{k}}\right]\\&\quad\leq 2^{q-1}\left[40^q+2C_q\right]\left(Ld(k_\epsilon^*)^\beta+\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*,\bar{K}^*)}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k_\epsilon^*)^d}}\right)^q+o(1/n).\end{aligned}\tag{5.19}$$

Control of the risk on the event $\{k_\epsilon^*>\hat{k}\}$. In order to control the risk on the complementary event, we observe that

$$\mathbb{E}_{f^*}\left[\left|\hat{f}_{\mathrm{iso}}^{\hat{k}}(x_0)-f^*(x_0)\right|^q\mathbb{1}_{k_\epsilon^*>\hat{k}}\right]\leq((1+|\mathcal{P}|)M)^q\,\mathbb{P}_{f^*}(k_\epsilon^*>\hat{k}).\tag{5.20}$$

We now show that the probability $\mathbb{P}_{f^*}(k_\epsilon^*>\hat{k})$ is small. According to the construction of the procedure (3.7), we have

$$\begin{aligned}\mathbb{P}_{f^*}(k_\epsilon^*>\hat{k})&\leq\mathbb{P}_{f^*}\left(\exists k'\in\mathcal{H},\ k'<k_\epsilon^*\ :\ \left|\hat{f}_{\mathrm{iso}}^{k_\epsilon^*}(x_0)-\hat{f}_{\mathrm{iso}}^{k'}(x_0)\right|>20\frac{\sqrt{\widehat{\mathrm{V}}_{\max}(\bar{\rho},\bar{K})}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}}\right)\\&\leq 2\sum_{k'\in\mathcal{H}_\epsilon^{\mathrm{iso}}\,:\,k'\leq k_\epsilon^*}\mathbb{P}_{f^*}\left(\left|\hat{f}_{\mathrm{iso}}^{k'}(x_0)-f^*(x_0)\right|>\frac{20}{2}\frac{\sqrt{\widehat{\mathrm{V}}_{\max}(\bar{\rho},\bar{K})}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}}\right).\end{aligned}$$

On the event $\bigcap_{k\in\mathcal{H}_\epsilon^{\mathrm{iso}}}\Delta$, we get similarly as in (5.14)

$$\mathbb{P}_{f^*}(k_\epsilon^*>\hat{k})\leq 2\sum_{k'\in\mathcal{H}_\epsilon^{\mathrm{iso}}\,:\,k'\leq k_\epsilon^*}\mathbb{P}_{f^*}\left(\left|\hat{f}_{\mathrm{iso}}^{k'}(x_0)-f^*(x_0)\right|>10\frac{\sqrt{1-\tilde{a}_n}}{1+\tilde{a}_n}\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho},\bar{K})}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}}\right),$$

where $\tilde{a}_n$ is defined above (5.6). According to Condition 2 in Section 5.1, we have $\tilde{a}_n\leq 1/3$. Consequently,

$$\mathbb{P}_{f^*}(k_\epsilon^*>\hat{k})\leq 2\sum_{k'\in\mathcal{H}_\epsilon^{\mathrm{iso}}\,:\,k'\leq k_\epsilon^*}\mathbb{P}_{f^*}\left(\left|\hat{f}_{\mathrm{iso}}^{k'}(x_0)-f^*(x_0)\right|>5\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho},\bar{K})}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}}\right).\tag{5.21}$$

By definition, the oracle bandwidth k^* is the one which gives the best trade-off, so that for all $k'\leq k_\epsilon^*\leq k^*$

$$\begin{aligned}Ld(k^*)^\beta\leq Ld(k_\epsilon^*)^\beta=\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*,\bar{K}^*)}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k_\epsilon^*)^d}}&\leq\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*,\bar{K}^*)}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k^*)^d}}\\&\leq\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*,\bar{K}^*)}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}}\\&\leq\frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho},\bar{K})}(B_0+\mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}}.\end{aligned}$$

From (5.18), (5.21) and the last inequality, we get

$$\begin{aligned}
&\mathbb{P}_{f^*}(k^*_\epsilon > \hat{k}) \\
&\quad \leq 2 \sum_{k' \in \mathcal{H}^{\mathrm{iso}}_\epsilon \,:\, k' \leq k^*_\epsilon} \mathbb{P}_{f^*}\left(\left| \hat{f}^{k'}_{\mathrm{iso}}(x_0) - f^*(x_0) \right| > 2 \frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}, \bar{K})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}} + 3b_k \right) \\
&\quad \leq 2 \sum_{k' \in \mathcal{H}^{\mathrm{iso}}_\epsilon \,:\, k' \leq k^*_\epsilon} \mathbb{P}_{f^*}\left(\sup_{\rho, K} \left[\left| \hat{f}^{k'}_{\mathrm{iso}}(x_0) - f^*(x_0) \right| - 2 \frac{\sqrt{\mathrm{V}_{\max}(\rho, K)}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}} \right] > 3b_k \right).
\end{aligned}$$

Using $\mathrm{iso}_\epsilon(n)/(n(k')^d)^{1/4} \leq 1$ (see Condition 3 on n), the definition of $\mathrm{iso}_\epsilon(n)$, Proposition 1 with $h = (k', \ldots, k')$, $\mathrm{V}(\lambda) = \mathrm{V}_{\max}(\rho, K)$ and z such that $B_z = (B_0 + \mathrm{iso}_\epsilon(n))$, we obtain

$$\mathbb{P}_{f^*}(k^*_\epsilon > \hat{k}) \leq 4|\mathcal{P}| \sum_{k' \in \mathcal{H}^{\mathrm{iso}}_\epsilon \,:\, k' \leq k^*_\epsilon} \exp\left(- \frac{(\mathrm{iso}_\epsilon(n))^2}{100 + 4\,\mathrm{iso}_\epsilon(n)/(n(k')^d)^{1/4}} \right) \leq 4|\mathcal{P}| n^{-1}.$$

Then, in view to the last inequality, (5.15), (5.16), (5.19), and (5.20), we conclude

$$\mathbb{E}_{f^*} \left| \hat{f}^{\hat{h}}(x_0) - f^*(x_0) \right|^q \leq 2^{q-1} \left[40^q + 2C_q \right] \left(Ld(k^*_\epsilon)^\beta + \frac{\sqrt{\mathrm{V}_{\max}(\bar{\rho}^*, \bar{K}^*)}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k^*_\epsilon)^d}} \right)^q + o(1/n).$$

By definition of k^* and k^*_ϵ in the beginning of the proof, the theorem is proved.

∎

5.6. Proof of Theorem 4

One may verify that the *oracle bandwidth*

$$h^* := \arg\min_{h \in \mathcal{H}} \left\{ L \sum_{j=1}^{d} \beta_j^{-1} (h_j)^{\beta_j} + 2 \frac{\sqrt{\mathrm{V}(\rho^*, K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{d\sqrt{n\Pi_h}} \right\}$$

is well defined. Moreover, define the element h^*_ϵ of $\mathcal{H}_\epsilon$ such that for all $j = 1, \ldots, d$, $h^*_{\epsilon,j} \leq h^*_j \leq \epsilon^{-1} h^*_{\epsilon,j}$. We then note that the estimator $\hat{f}^h$ is a constant function and $f^0 \equiv f^*(x_0)$ since we only consider locally constant functions ($|\mathcal{P}| = 1$). To stress the importance of the bandwidth, we set for any $h \in \mathcal{H}$

$$\tilde{\mathcal{D}}_h(\cdot) := \tilde{D}_{\tilde{\lambda}_h}(\cdot) = n^{-1} \sum_i \hat{\rho}'(Y_i - \cdot)\hat{K}_h(X_i)$$

and

$$\mathcal{D}_h(\cdot) := P\left[\tilde{D}_{\tilde{\lambda}_h}(\cdot) \right] = \int \hat{K}_h(x) \int \hat{\rho}'(\sigma z + f^*(x) - \cdot) \mathbb{G}(z) dz dx. \tag{5.22}$$

Here, $\tilde{\lambda}_h(x, y, f) := \hat{\rho}(y - f(x))\hat{K}_h(x)$, $\mathbb{G}(\cdot) := \frac{1}{n}\sum_{i=1}^n g_i(\cdot)$, and $(\hat{\rho}, \hat{K})$ and $\tilde{D}_\lambda(\cdot)$ are defined in (3.12) and (5.3), respectively. Next, for uniform designs and homoscedastic noise levels, the quantity c_{λ_h} simplifies for any λ_h to

$$c_{\lambda_h} = c_\rho := \int \rho''(\sigma z) \mathbb{G}(z) dz. \tag{5.23}$$

Moreover, according to Lemma 10, we have for any two constant functions $f, \tilde{f} \in \mathcal{F}_{\delta_n}$

$$|f - \tilde{f}| \leq (1 + 2\sqrt{b_{h_{\max}} + \delta_n}) \inf_{h \in \mathcal{H}} c_{\hat{\rho}}^{-1} |\mathcal{D}_h(f) - \mathcal{D}_h(\tilde{f})|. \tag{5.24}$$

Furthermore, from Condition 3 on n, Lemma 7, and Lemma 9, it follows that

$$\mathbb{P}_{f^*}\left(\exists h \in \mathcal{H}_\epsilon \ : \ \hat{f}^h \notin \mathcal{F}_{\delta_n}\right) \leq 2 \sum_{h \in \mathcal{H}_\epsilon} \exp\left(-\frac{n\Pi_h/(4\ln^2 n)}{98\gamma_{\max}^2 \mathcal{K}_{\max}^2 + 4\gamma_{\max}\mathcal{K}_{\max}}\right) \leq 2n^{-1} \tag{5.25}$$

and

$$\mathbb{P}_{f^*}(\Delta^c) \leq \frac{2}{n^2} + 2\exp\left(-\frac{nh_{\min}^d/(4\ln^2 n)}{98\gamma_{\max}^2 \mathcal{K}_{\max}^2 + 4\gamma_{\max}\mathcal{K}_{\max}}\right) \leq 4n^{-1}, \tag{5.26}$$

where Δ is defined in (5.11). Thus, we restrict our considerations in the following to the event $\left\{\hat{f}^h \in \mathcal{F}_{\delta_n} \text{ for all } h \in \mathcal{H}_\epsilon\right\} \cap \Delta$. Moreover, we work on the event $\mathcal{A} := \{h_\epsilon^* \preceq \hat{h}\}$ and its complement $\mathcal{A}^c$ separately. For this, we decompose the risk into $R_{\mathcal{A}}\left(\hat{f}^h, f^*\right) := \mathbb{E}_{f^*}\left[\left|\hat{f}^h(x_0) - f^*(x_0)\right|^q \mathbb{1}\{\mathcal{A}\}\right]$ and $R_{\mathcal{A}^c}\left(\hat{f}^h, f^*\right) := \mathbb{E}_{f^*}\left[\left|\hat{f}^h(x_0) - f^*(x_0)\right|^q \mathbb{1}\{\mathcal{A}^c\}\right]$.

Control of the risk on the event $\mathcal{A}$. With the triangular inequality and Lemma 8, we obtain

$$R_{\mathcal{A}}\left(\hat{f}^{\hat{h}}, f^*\right) \leq 3^{q-1}\left[R_{\mathcal{A}}\left(\hat{f}^{h_\epsilon^*, \hat{h}}, \hat{f}^{\hat{h}}\right) + R_{\mathcal{A}}\left(\hat{f}^{\hat{h}, h_\epsilon^*}, \hat{f}^{h_\epsilon^*}\right) + R_{\mathcal{A}}\left(\hat{f}^{h_\epsilon^*}, f^*\right)\right]. \tag{5.27}$$

Let us now control the first term on the right hand side of the last inequality. First, we observe that

$$R_{\mathcal{A}}\left(\hat{f}^{h_\epsilon^*, \hat{h}}, \hat{f}^{\hat{h}}\right) \leq \mathbb{E}_{f^*} \sup_{h \in \mathcal{H} \,:\, h \succeq h_\epsilon^*} \left|\hat{f}^{h_\epsilon^*, h}(x_0) - \hat{f}^h(x_0)\right|^q \mathbb{1}_{\mathcal{A}}. \tag{5.28}$$

To simplify the presentation, we introduce the notation $\tau_n := (1 + 2\sqrt{b_{h_{\max}} + \delta_n})$. Using (5.24) and taking $f = \hat{f}^{h_\epsilon^*, h}$ and $\tilde{f} = \hat{f}^h$, we then have

$$\left|\hat{f}^{h_\epsilon^*, h}(x_0) - \hat{f}^h(x_0)\right| \leq \tau_n c_{\hat{\rho}}^{-1} \left|\mathcal{D}_h\left(\hat{f}^{h_\epsilon^*, h}\right) - \mathcal{D}_h\left(\hat{f}^h\right)\right|.$$

Recall that, by definition, $\tilde{\mathcal{D}}_h(\hat{f}^h) = 0$ for all $h \in \mathcal{H}$. We then obtain from the last inequality for any $h \in \mathcal{H}$

$$\begin{aligned}\left|\hat{f}^{h_\epsilon^*, h}(x_0) - \hat{f}^h(x_0)\right| \leq \tau_n c_{\hat{\rho}}^{-1} \Big(&\left|\mathcal{D}_h\left(\hat{f}^{h_\epsilon^*, h}\right) - \mathcal{D}_{h_\epsilon^* \vee h}\left(\hat{f}^{h_\epsilon^*, h}\right)\right| \\ &+ \left|\mathcal{D}_{h_\epsilon^* \vee h}\left(\hat{f}^{h_\epsilon^*, h}\right) - \tilde{\mathcal{D}}_{h_\epsilon^* \vee h}\left(\hat{f}^{h_\epsilon^*, h}\right)\right| + \left|\tilde{\mathcal{D}}_h\left(\hat{f}^h\right) - \mathcal{D}_h\left(\hat{f}^h\right)\right|\Big).\end{aligned} \tag{5.29}$$

Using the last inequality and (5.28), we have

$$\begin{aligned}R_{\mathcal{A}}\left(\hat{f}^{\hat{h}, h_\epsilon^*}, \hat{f}^{\hat{h}}\right) \leq& 2^{q-1}\tau_n^q \mathbb{E}_{f^*} c_{\hat{\rho}}^{-q} \sup_{h \in \mathcal{H}_\epsilon} \sup_{f \in \mathcal{F}_{\delta_n}} \left|\mathcal{D}_h(f) - \mathcal{D}_{h_\epsilon^* \vee h}(f)\right|^q \\ &+ 2^q \tau_n^q \sum_{h \in \mathcal{H}_\epsilon \,:\, h \succeq h_\epsilon^*} \mathbb{E}_{f^*} c_{\hat{\rho}}^{-q} \sup_{f \in \mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f)\right|^q.\end{aligned}$$

Using Lemma 11 and Lemma 12 with $h' = h_\epsilon^*$ and $\rho = \hat{\rho}$, we get

$$R_{\mathcal{A}}\left(\hat{f}^{\hat{h}, h_\epsilon^*}, \hat{f}^{\hat{h}}\right) \leq 2^{q-1}\tau_n^{2q}\left(L\sum_{j=1}^d (h_{\epsilon,j}^*)^{\beta_j}\right)^q + 2^q \tau_n^q \bar{C}_q \left(\frac{a_n \sqrt{\mathrm{V}(\rho^*, K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h_\epsilon^*}}}\right)^q.$$

According to Conditions 1 and 2 in Section 5.1, we have $\tau_n \le 2$ and $a_n \le 2\sqrt{2}$. Thus,

$$R_{\mathcal{A}}\left(\hat{f}^{\hat{h},h^*_\epsilon}, \hat{f}^{\hat{h}}\right) \le 2^{4q}\bar{C}_q\left(L\sum_{j=1}^{d}(h^*_{\epsilon,j})^{\beta_j} + \frac{\sqrt{\mathrm{V}(\rho^*,K^*)}(B_0+\mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h^*_\epsilon}}}\right)^q. \tag{5.30}$$

The second term on the right hand side of (5.27) is controlled by the construction of the procedure (3.14), thus

$$R_{\mathcal{A}}\left(\hat{f}^{\hat{h},h^*_\epsilon}, \hat{f}^{h^*_\epsilon}\right) \le \mathbb{E}_{f^*}\left[16\frac{\sqrt{\widehat{\mathrm{V}}(\hat{\rho},\hat{K})}(B_0+\mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h^*_\epsilon}}}\right]^q \mathbb{1}_{\mathcal{A}}.$$

On the event Δ,

$$R_{\mathcal{A}}\left(\hat{f}^{\hat{h},h^*_\epsilon}, \hat{f}^{h^*_\epsilon}\right) \le \left(16\frac{\sqrt{1+\tilde{a}_n}}{1-\tilde{a}_n}\frac{\sqrt{\mathrm{V}(\rho^*,K^*)}(B_0+\mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h^*_\epsilon}}}\right)^q, \tag{5.31}$$

where $\tilde{a}_n$ is defined above (5.6). By the definition of the Hölder class (Definition 4) and b_h (Definition (2.9)), we can control the bias for any $\vec{\beta} : \lfloor\beta\rfloor \le b$ and for any $h \in \mathcal{H}$:

$$b_h \le \sup_{x\in V_h}|\mathrm{P}(f^*)(x-x_0) - f^*(x)| \le L\sum_{j=1}^{d} h_j^{\beta_j},$$

where $\mathrm{P}(f^*)(x-x_0)$ is the Taylor Polynomial of f^* at x_0. Finally, with Theorem 2, we can bound the third term in (5.27):

$$R_{\mathcal{A}}\left(\hat{f}^{h^*_\epsilon}, f^*\right) \le 2C_q\left(L\sum_{j=1}^{d}(h^*_{\epsilon,j})^{\beta_j} + \frac{\sqrt{\mathrm{V}(\rho^*,K^*)}B_0}{\sqrt{n\Pi_{h^*_\epsilon}}}\right)^q + o(1/n).$$

Using (5.27), (5.30), (5.31), and the last inequality, and invoking to Condition 2 in Section 5.1, we have a control of the risk on the event $\mathcal{A}$:

$$\begin{aligned}&R_{\mathcal{A}}\left(\hat{f}^{\hat{h}}, f^*\right)\\ &\le 3^{q-1}\left[2^{4q}\bar{C}_q + 32^q + 2C_q\right]\left(L\sum_{j=1}^{d}(h^*_{\epsilon,j})^{\beta_j} + \frac{\sqrt{\mathrm{V}(\rho^*,K^*)}(B_0+\mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h^*_\epsilon}}}\right)^q + o(1/n).\end{aligned} \tag{5.32}$$

Control of the risk on the event $\mathcal{A}^c$. In order to control the risk on the complementary event $\mathcal{A}^c$, we observe that

$$R_{\mathcal{A}^c}\left(\hat{f}^{\hat{h}}, f^*\right) \le (2M)^q\mathbb{P}_{f^*}(\mathcal{A}^c). \tag{5.33}$$

We now show that the probability $\mathbb{P}_{f^*}(\mathcal{A}^c)$ is small. According to the construction of the procedure (3.14), the event $\mathcal{A}^c$ implies that there exists a $h' \in \mathcal{H}$ such that $h' \preceq h^*_\epsilon$ and

$$\left|\hat{f}^{h^*_\epsilon,h'}(x_0) - \hat{f}^{h'}(x_0)\right| > 16\frac{\sqrt{\widehat{\mathrm{V}}(\hat{\rho},\hat{K})}(B_0+\mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}.$$

Using (5.24) and taking $f = \hat{f}^{h^*_\epsilon,h'}$ and $\tilde{f} = f^{h'}$, we have on the event $\mathcal{A}^c$

$$\tau_n c_{\hat{\rho}}^{-1}\left|\mathcal{D}_{h'}\left(\hat{f}^{h^*_\epsilon,h'}\right) - \mathcal{D}_{h'}\left(\hat{f}^{h'}\right)\right| > 16\frac{\sqrt{\widehat{\mathrm{V}}(\hat{\rho},\hat{K})}(B_0+\mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}.$$

From the last inequality, we obtain (cf. (5.29))

$$\tau_n c_{\hat\rho}^{-1} \sup_{f\in\mathcal{F}_{\delta_n}} \left|\mathcal{D}_{h'}(f) - \mathcal{D}_{h^*_\epsilon \vee h'}(f)\right| + 2\tau_n c_{\hat\rho}^{-1} \sup_{f\in\mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\right| > 16 \frac{\sqrt{\widehat{\mathrm{V}}(\hat\rho,\hat K)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}.$$

Together with Lemma 12, this yields

$$\tau_n^2 L \sum_{j=1}^{d} (h^*_{\epsilon,j})^{\beta_j} + 2\tau_n c_{\hat\rho}^{-1} \sup_{f\in\mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\right| > 16 \frac{\sqrt{\widehat{\mathrm{V}}(\hat\rho,\hat K)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}.$$

On the event $\bigcap_{h\in\mathcal{H}_\epsilon}\Delta$, we get similarly as in (5.14)

$$\tau_n^2 L \sum_{j=1}^{d} (h^*_{\epsilon,j})^{\beta_j} + 2\tau_n c_{\hat\rho}^{-1} \sup_{f\in\mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\right| > 16 \frac{\sqrt{1-\tilde a_n}}{1+\tilde a_n} \frac{\sqrt{\mathrm{V}(\hat\rho,\hat K)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}},$$

where $\tilde a_n$ is defined above (5.6). According to Conditions 1 and 2 in Section 5.1, we have $\tau_n \le 2$ and $\tilde a_n \le 1/3$. Consequently,

$$c_{\hat\rho}^{-1} \sup_{f\in\mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\right| > \frac{16}{8} \frac{\sqrt{\mathrm{V}(\hat\rho,\hat K)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}} - L\sum_{j=1}^{d} (h^*_{\epsilon,j})^{\beta_j}.$$

By definition, the oracle bandwidth h^*_ϵ is the one which gives the best trade-off, then for all $h' \preceq h^*_\epsilon \preceq h^*$

$$\begin{aligned} L\sum_{j=1}^{d} (h^*_j)^{\beta_j} \le L\sum_{j=1}^{d} (h^*_{\epsilon,j})^{\beta_j} = \frac{\sqrt{\mathrm{V}(\rho^*,K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h^*_\epsilon}}} &\le \frac{\sqrt{\mathrm{V}(\rho^*,K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h^*}}} \\ &\le \frac{\sqrt{\mathrm{V}(\rho^*,K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}} \\ &\le \frac{\sqrt{\mathrm{V}(\hat\rho,\hat K)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}. \end{aligned}$$

From last two inequalities, we obtain on the event $\mathcal{A}^c$

$$c_{\hat\rho}^{-1} \sup_{f\in\mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\right| > \frac{\sqrt{\mathrm{V}(\hat\rho,\hat K)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}.$$

Then, we have a control of the following probability

$$\mathbb{P}_{f^*}(\mathcal{A}^c) \le \sum_{h'\in\mathcal{H}_\epsilon\,:\,h'\preceq h^*_\epsilon} \mathbb{P}_{f^*}\left(\sup_{\rho,K} \sup_{f\in\mathcal{F}_{\delta_n}} \frac{\left|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\right|}{c_\rho\sqrt{\mathrm{V}(\rho,K)}} > \frac{B_0 + \mathrm{ani}_\epsilon(n)}{\sqrt{n\Pi_{h'}}} \right).$$

Using $\mathrm{ani}_\epsilon(n)/(n\Pi_{h'})^{1/4} \le 1$ (see Condition 3 on n) Lemma 6 with z such that $B_z = B_0 + \mathrm{ani}_\epsilon(n)$, we deduce that

$$\begin{aligned} \mathbb{P}_{f^*}(\mathcal{A}^c) &\le \sum_{h'\in\mathcal{H}_\epsilon\,:\,h'\preceq h^*_\epsilon} \exp\left(-\frac{(\mathrm{ani}_\epsilon(n))^2}{100 + 4\,\mathrm{ani}_\epsilon(n)/(n\Pi_{h'})^{1/4}} \right) \\ &\le n^{-1}. \end{aligned}$$

From (5.33) and the last inequality, we obtain a control of the risk on the event $\mathcal{A}^c$:

$$R_{\mathcal{A}^c}\left(\hat{f}^{\hat{h}}, f^*\right) \leq (2M)^q n^{-1}.$$

Then, in view of the last inequality, (5.25), (5.26), and (5.32), we conclude that

$$\begin{aligned}&\mathbb{E}_{f^*}\left|\hat{f}^{\hat{h}}(x_0) - f^*(x_0)\right|^q \\ &\leq 3^{q-1}\left[2^{4q}\bar{C}_q + 32^q + 2C_q\right]\left(L\sum_{j=1}^{d}(h^*_{\epsilon,j})^{\beta_j} + \frac{\sqrt{\mathrm{V}(\rho^*,K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h^*_\epsilon}}}\right)^q + o(1/n).\end{aligned}$$

With the definition of h^* and h^*_ϵ in the beginning of the proof, the theorem can be deduced. ■

6. Appendix

6.1. A Entropy Calculations

First, let us give a bound for the entropy $H_{\mathcal{F},\nu}$ (defined below (2.10)) and its Dudley's integral. For this, we recall that the metric entropy of a set is the logarithm of the minimal number of balls (with respect to the corresponding metric) needed to cover the set (see, for example, [31]). For any $v \in]0,1]$, we then have

$$H_{\mathcal{F},\nu}(v) \leq |\mathcal{P}|\ln\left(\frac{2M}{v^2}\right) \quad \text{and} \quad \int_0^1 \sqrt{H_{\mathcal{F},\nu}(u) \wedge n du} \leq \sqrt{|\mathcal{P}|\ln(2M)} + \sqrt{|\mathcal{P}|}\int_1^\infty \sqrt{\ln(v)/(2v^3)}dv.$$

We now give a bound for the entropy $H_{\mathcal{F}\cup\Lambda,\omega}$, that is, (defined below (2.14)) for the special set of Huber contrasts indexed by the scale γ,$\Upsilon_{\mathrm{H}} := \{\rho = \rho_{\mathrm{H},\gamma} \ : \ \gamma \in [\gamma_{\min}, \gamma_{\max}]\}$. Here, the positive constants are chosen such that $\gamma_{\min} \leq 1$ and $\gamma_{\max} \geq 1$. In this example, we do not consider the choice of the kernel, we just take the indicator function as kernel and $\mathcal{K} = \left\{\mathbb{1}_{[-1/2,1/2]^d}(\cdot)\right\}$. In this case, we have $\Lambda = \Upsilon_{\mathrm{H}}$. For $v \in]0,1]$, we finally give the following bound

$$H_{\mathcal{F}\cup\Upsilon_{\mathrm{H}},\omega}(v) \leq (1 + |\mathcal{P}|)\ln\left(16\left[12 \vee g_\infty\right]\frac{\left[2M \vee \gamma_{\max}\right]\gamma^2_{\max}}{\gamma^4_{\min}v^2}\right)$$

and

$$\begin{aligned}\int_0^1 \sqrt{H_{\mathcal{F}\cup\Upsilon_{\mathrm{H}},\nu}(u) \wedge n du} \leq &\sqrt{(1 + |\mathcal{P}|)\ln\left(16\left[12 \vee g_\infty\right]\left[2M \vee \gamma_{\max}\right]\frac{\gamma^2_{\max}}{\gamma^4_{\min}}\right)} \\ &+ \sqrt{1 + |\mathcal{P}|}\int_1^\infty \sqrt{\ln(v)/(2v^3)}dv.\end{aligned}$$

Here, $g_\infty := \sup_{i=1,\ldots,n}\|g_i\|_\infty$ where $(g_i)_i$ are the noise densities in the model (2.1).

6.2. B Proof of the Auxilliary Result

Proof of Proposition 1. The definitions of $\hat{f}_\lambda$ and f^0 (see (2.5) and (5.1), respectively) imply that

$$\left|\hat{f}_\lambda(x_0) - f^*(x_0)\right| = \left|\left(\hat{t}_\lambda\right)_{0,\ldots,0} - t^0_{0,\ldots,0}\right| \leq \left\|\hat{t}_\lambda - t^0\right\|_{\ell_\infty}.$$

Using $\hat{f}_\lambda \in \mathcal{F}_{\delta_n}$, Lemma 4, and the last inequality, we have

$$\big|\hat{f}_\lambda(x_0) - f^*(x_0)\big| \le (1-\sqrt{\delta_n})^{-1} c_\lambda^{-1} \|P^0\big[\tilde{D}_\lambda(\hat{f}_\lambda)\big] - P^0\big[\tilde{D}_\lambda(f^0)\big]\|_{\ell_\infty}.$$

Recall that by definition $\tilde{D}_\lambda(\hat{f}_\lambda) = 0$ and $P^0\big[\tilde{D}_\lambda(f^0)\big] = 0$. Thus, for all $\lambda \in \Lambda$ such that $\hat{f}_\lambda \in \mathcal{F}_{\delta_n}$, the last inequality implies

$$\big|\hat{f}_\lambda(x_0) - f^*(x_0)\big| \quad\le\quad (1-\sqrt{\delta_n})^{-1} c_\lambda^{-1} \Big(\|\tilde{D}_\lambda(\hat{f}_\lambda) - P\big[\tilde{D}_\lambda(\hat{f}_\lambda)\big]\|_{\ell_\infty} + \|P\big[\tilde{D}_\lambda(\hat{f}_\lambda)\big] - P^0\big[\tilde{D}_\lambda(\hat{f}_\lambda)\big]\|_{\ell_\infty}\Big).$$

From Lemma 5 and the last display, we obtain

$$\begin{aligned}\big|\hat{f}_\lambda(x_0) - f^*(x_0)\big| \quad&\le\quad (1-\sqrt{\delta_n})^{-1} c_\lambda^{-1} \Big(\|\tilde{D}_\lambda(\hat{f}_\lambda) - P\big[\tilde{D}_\lambda(\hat{f}_\lambda)\big]\|_{\ell_\infty} + (1+\sqrt{b_h+\delta_n}) c_\lambda b_h\Big)\\ &\le\quad \frac{1+\sqrt{b_h+\delta_n}}{1-\sqrt{\delta_n}} b_h + (1-\sqrt{\delta_n})^{-1} \sup_{f\in\mathcal{F}_{\delta_n},\lambda\in\Lambda} c_\lambda^{-1}\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\|_{\ell_\infty}.\end{aligned}$$

As $\sqrt{b_h+\delta_n} \le 1/2$ according to Condition 1, this yields

$$\big|\hat{f}_\lambda(x_0) - f^*(x_0)\big| \le 3b_h + 2 \sup_{f\in\mathcal{F}_{\delta_n},\lambda\in\Lambda} c_\lambda^{-1}\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\|_{\ell_\infty}.$$

From the last inequality and the definitions of $\mathrm{V}(\cdot)$ and c_λ introduced in (2.6) and (5.7), respectively, we deduce

$$\begin{aligned}&\mathbb{P}_{f^*}\left(\left\{\sup_{\lambda\in\Lambda}\left[\big|\hat{f}_\lambda(x_0) - f^*(x_0)\big| - 2\frac{\sqrt{\mathrm{V}(\lambda)}\,B_z}{\sqrt{n\Pi_h}}\right] \ge 3b_h\right\} \cap \bigcap_{\lambda\in\Lambda}\left\{\hat{f}_\lambda \in \mathcal{F}_{\delta_n}\right\}\right)\\ &\le \mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F}_{\delta_n},\lambda\in\Lambda}\left[2c_\lambda^{-1}\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\|_{\ell_\infty} - 2\frac{\sqrt{\mathrm{V}(\lambda)}\,B_z}{\sqrt{n\Pi_h}}\right] \ge 0\right)\\ &\le \mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F}_{\delta_n},\lambda\in\Lambda}\frac{\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\|_{\ell_\infty}}{\sqrt{\Pi_h}\sqrt{P\left[\lambda'(f^*)\right]^2 + \lambda'_\infty/(n\Pi_h)^{1/4}}} \ge \frac{B_z}{\sqrt{n\Pi_h}}\right).\end{aligned}$$

Using Lemma 6 and the last inequality, we finally obtain

$$\mathbb{P}_{f^*}\left(\left\{\sup_{\lambda\in\Lambda}\left[\big|\hat{f}_\lambda(x_0) - f^*(x_0)\big| - 2\frac{\sqrt{\mathrm{V}(\lambda)}\,B_z}{\sqrt{n\Pi_h}}\right] \ge 3b_h\right\} \cap \bigcap_{\lambda\in\Lambda}\left\{\hat{f}_\lambda \in \mathcal{F}_{\delta_n}\right\}\right) \le 2|\mathcal{P}|e^{-z}.$$

$\blacksquare$

6.3. C Technical Lemmas

We first give a result for the deterministic criterion $P^0\left[\tilde{D}_\lambda(\cdot)\right]$ defined in (5.4):

Lemma 4. *For any $\lambda \in \Lambda$ and any $h \in \mathcal{H}$, and for n sufficiently large (see Condition 2 in Section 5.1), the following holds:*

1. $P^0\left[\tilde{D}_\lambda(f^0)\right] = 0$, *and the function* $P^0\left[\tilde{D}_\lambda(f)\right]$ *is bijective as function of* $\mathcal{F}_{\delta_n}$ *(see Definition (5.2)) on the corresponding image.*

2. *For any* $f, \tilde{f} \in \mathcal{F}_{\delta_n}$,

$$\|t - \tilde{t}\|_{\ell_\infty} \leq (1 - \sqrt{\delta_n})^{-1} c_\lambda^{-1} \|P^0\big[\tilde{D}_\lambda(f)\big] - P^0\big[\tilde{D}_\lambda(\tilde{f})\big]\|_{\ell_\infty},$$

where $f = \mathrm{P}_t$ *and* $\tilde{f} = \mathrm{P}_{\tilde{t}}$.

Next, we consider the bias:

Lemma 5. *For any* $h \in \mathcal{H}$ *and any* $\lambda \in \Lambda$, *if* n *is sufficiently large (see Condition 2 in Section 5.1), it holds that*

$$\sup_{f \in \mathcal{F}_{\delta_n}} \big\|P^0\big[\tilde{D}_\lambda(f)\big] - P\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_\infty} \leq (1 + \sqrt{b_h + \delta_n}) c_\lambda b_h.$$

The following lemma allows us to control the deviations of the process $\tilde{D}_\lambda(\cdot)$:

Lemma 6. *For any* $h \in \mathcal{H}$, *it holds that*

$$\mathbb{P}_{f^*}\left(\sup_{f \in \mathcal{F}_{\delta_n}, \lambda \in \Lambda} \frac{\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\|_{\ell_\infty}}{\sqrt{\Pi_h}\sqrt{P\left[\lambda'(f^*)\right]^2 + \lambda'_\infty/(n\Pi_h)^{1/4}}} \geq \frac{B_z}{\sqrt{n\Pi_h}}\right) \leq 2|\mathcal{P}| \exp(\text{-}z).$$

Now, we bound the probability of the event "the λ-LPA estimator does not belong to the ball centered on t^0 with radius δ_n".

Lemma 7. *For any* $h \in \mathcal{H}$, *if* n *is sufficiently large (according to Condition 1 in Section 5.1), it holds that*

$$\mathbb{P}_{f^*}\left(\Omega^c\right) \leq 2|\mathcal{P}| \exp\left(-\frac{n\Pi_h/(4\ln^2 n)}{98\gamma_{\max}^2\mathcal{K}_{\max}^2 + 4\gamma_{\max}\mathcal{K}_{\max}}\right),$$

where $\Omega^c := \left\{\exists \lambda \in \Lambda,\ \hat{f}_\lambda \notin \mathcal{F}_{\delta_n}\right\}$, *and* $\gamma_{\max}$ *and* $\mathcal{K}_{\max}$ *are defined in Section 2.*

Next, we do some simple algebra.

Lemma 8. *For any* $x, y \in \mathbb{R}_0^+$, *it holds that*

$$x^q \leq 2^q[x-y]_+^q + 2^q y^q.$$

Moreover, for any $l, q \in \mathbb{N}^*$ *and* $x_1, \dots, x_l \geq 0$, *it holds that*

$$\left(\sum_{i=1}^{l} x_i\right)^q \leq l^{q-1}\left(\sum_{i=1}^{l} x_i^q\right).$$

The following lemma allows us to get our hands on the estimator $\widehat{\mathrm{V}}(\cdot)$.

Lemma 9. *For any* $h \in \mathcal{H}$, *if* n *is sufficiently large according to Condition 2 in Section 5.1, it holds that*

$$\mathbb{P}_{f^*}\left(\Delta\right) \geq 1 - 2/n^2 - \mathbb{P}_{f^*}(\Omega^c),$$

where $\Delta := \bigcap_{\lambda \in \Lambda}\left\{\sqrt{\widehat{\mathrm{V}}(\lambda)} \in \left[\frac{\sqrt{1-\tilde{a}_n}}{1+\tilde{a}_n}\sqrt{\mathrm{V}(\lambda)}, \frac{\sqrt{1+\tilde{a}_n}}{1-\tilde{a}_n}\sqrt{\mathrm{V}(\lambda)}\right]\right\}$, Ω^c *is defined in Lemma 7, and* $\tilde{a}_n$ *is defined in (5.6).*

We now consider functions near the target f^*.

Lemma 10. *Let $\mathcal{D}_h(\cdot) : [\text{-}M, M] \to \mathbb{R}$ and $c_{\hat\rho}$ be as defined in the proof of Theorem 4 and assume $f^* \in \mathbb{H}_d(\vec\beta, L, M)$ and n sufficiently large (according to Condition 1 in Section 5.1). Then, for any $t, \tilde t \in [f^*(x_0) - \delta_n, f^*(x_0) + \delta_n]$, it holds that*

$$|t - \tilde t| \le (1 + 2\sqrt{b_{h_{\max}} + \delta_n}) \inf_{h \in \mathcal{H}} c_{\hat\rho}^{\text{-}1} |\mathcal{D}_h(t) - \mathcal{D}_h(\tilde t)|.$$

Next, we controll the distance of $\tilde{\mathcal{D}}_h(f)$ to $\mathcal{D}_h(f)$ for appropirate bandwidth h and functions f:

Lemma 11. *For n sufficiently large (see Condition 3 in Section 5.1), it holds for any $h \in \mathcal{H}$ that*

$$\mathbb{E}_{f^*} c_{\hat\rho}^{\text{-}q} \sup_{f \in \mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f)\right|^q \le \frac{\bar C_q}{n|H_\epsilon|} \left(\frac{a_n \sqrt{\mathrm{V}(\rho^*, K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n \Pi_h}} \right)^q$$

for a constant $\bar C_q$ ($\bar C_q = 4q24^q\,\mathrm{Gamma}(q)$ works). The functionals $\tilde{\mathcal{D}}$ and $\mathcal{D}$ are defined in the proof of Theorem 4, Gamma(q) is the classical Gamma function, $\mathrm{V}(\rho^*, K^*)$ is defined in (3.9) and (3.10), $\mathrm{ani}_\epsilon(n)$ is defined in Section (3.2), and a_n is defined in (5.6).

Eventually, we look at the distance to $\mathcal{D}_{h' \vee h}(f)$ to $\mathcal{D}_h(f)$ for appropirate bandwidths h and h' and functions f:

Lemma 12. *For any $h' \in \mathcal{H}$, any $f^* \in \mathbb{H}_d(\vec\beta, L, M)$ such that $\vec\beta \in]0, 1]^d$, and for n sufficiently large (according to Condition 2 in Section 5.1), it holds that*

$$\sup_{h \in \mathcal{H}} \sup_{f \in \mathcal{F}_{\delta_n}} \left|\mathcal{D}_{h' \vee h}(f) - \mathcal{D}_h(f)\right| \le (1 + \sqrt{\delta_n + b_{h_{\max}}}) c_{\hat\rho} L \sum_{j=1}^{d} (h'_j)^{\beta_j},$$

where $\mathcal{D}_h$ and $c_{\hat\rho}$ are defined in (5.22) and (5.23) in the proof of Theorem 4.

6.4. D Proofs of the Technical Lemmas

Proof of Lemma 4. Let us proof the first claim. For this, we note that the components of $P^0\big[\tilde D_\lambda(f)\big]$ are given by

$$P^0\big[\tilde D_\lambda^p(f)\big] = \int \left(\frac{x - x_0}{h}\right)^p \mu(x) K_h(x) \int \rho'\big(\sigma(x)z + f^0(x) - f(x)\big) \frac{1}{n} \sum_{i=1}^{n} g_i(z) dz\, dx.$$

Since ρ and $\sum_i g_i(\cdot)$ are symmetric, it holds that $\int \rho'(z) \sum_i g_i(z) dz = 0$ and $P^0\big[\tilde D_\lambda^p(f^0)\big] = 0$. We now show that $P^0\big[\tilde D_\lambda^p(\cdot)\big]$ is injective on the image of $\mathcal{F}_{\delta_n}$ exploiting further the symmetry of $\rho(\cdot)$ and $\sum_i g_i(\cdot)$. Consider $f, \tilde f \in \mathcal{F}_{\delta_n}$ such that $P^0\big[\tilde D_\lambda(f)\big] = P^0\big[\tilde D_\lambda(\tilde f)\big]$. We have to show that $f = \tilde f$. For this, we first note that

$$\sum_{p \in \mathcal{P}} (t_p - \tilde t_p) \left(P^0\big[\tilde D_\lambda^p(\mathrm{P}_t)\big] - P^0\big[\tilde D_\lambda^p(\mathrm{P}_{\tilde t})\big] \right) = 0,$$

where t and $\tilde t$ are such that $\mathrm{P}_t = f$ and $\mathrm{P}_{\tilde t} = \tilde f$. To simplify the presentation, we introduce the notation $u(\cdot) := (f - f^0)(\cdot)$, $\tilde u(\cdot) := (\tilde f - f^0)(\cdot)$, and $\mathbb{G}(\cdot) := n^{\text{-}1} \sum_{i=1}^{n} g_i(\cdot)$. Since $\mathbb{G}(\cdot)$ is symmetric,

K is nonnegative, and ρ' is odd and positive on $\mathbb{R}_+^*$, the last display implies

$$\begin{aligned} &\int K_h(x)\mu(x)\big[u(x)-\tilde u(x)\big]\int \big[\rho'\big(\sigma(x)z-u(x)\big)-\rho'\big(\sigma(x)z-\tilde u(x)\big)\big]\ \mathbb{G}(z)\ dz\ dx=0\\ \Leftrightarrow\quad &\int K_h(x)\mu(x)\big|u(x)-\tilde u(x)\big|\int \big|\rho'\big(\sigma(x)z-u(x)\big)-\rho'\big(\sigma(x)z-\tilde u(x)\big)\big|\ \mathbb{G}(z)\ dz\ dx=0. \end{aligned}$$

As $f,\tilde f\in\mathcal{F}_{\delta_n}$, it holds that $\sup_{x\in V_h}|u(x)|\vee|\tilde u(x)|\le\delta_n$. Moreover, using the mean value theorem, the P-continuity of ρ'', Assumption (2.2), $\inf_{z\in[\text{-}\gamma_{\min},\gamma_{\min}]}\rho''(z)>0$ and $\sqrt{\delta_n}\le\frac{1}{2}\wedge\frac{A\rho''_{\min}}{2L_{\rho''}}$, we obtain

$$\begin{aligned} &\int K_h(x)\mu(x)\big|u(x)-\tilde u(x)\big|\int\big|\rho'\big(\sigma(x)z-u(x)\big)-\rho'\big(\sigma(x)z-\tilde u(x)\big)\big|\ \mathbb{G}(z)\ dz\ dx\\ &\ge\int K_h(x)\mu(x)\big|u(x)-\tilde u(x)\big|^2\inf_{s:|s|\le\delta_n}\int\rho''\big(\sigma(x)z-s\big)\,\mathbb{G}(z)dzdx\\ &\ge\int K_h(x)\mu(x)\big|u(x)-\tilde u(x)\big|^2\inf_{s:|s|\le\delta_n}\int\rho''\big(\sigma(x)z-s\big)\,\mathbb{G}(z)dzdx\\ &\ge\int K_h(x)\mu(x)\big|u(x)-\tilde u(x)\big|^2dx\left[\int\rho''\big(\sigma(x)z\big)\,\mathbb{G}(z)dz-\delta_nL_{\rho''}\right]\\ &\ge\frac{A\rho''_{\min}}{2}\int K_h(x)\mu(x)\big|u(x)-\tilde u(x)\big|^2dx, \end{aligned}$$

where $A,\gamma_{\min}>0$ are introduced in Assumption (2.2) and $\rho''_{\min}$ in Definition 1. The last display and the positivity of K over its support yield $\sup_{x\in\mathcal{V}}\big|u(x)-\tilde u(x)\big|=0$. As u and $\tilde u$ are polynomials with finite degree, we finally obtain that $f=\tilde f$, and the first claim is proved.

Let us now turn to the second claim. We set $D(\cdot):=P^0\big[\tilde D_\lambda(\cdot)\big]$ and note that $D(\cdot)$ is differentiable and injective on $\mathcal{F}_{\delta_n}$ (the latter according to the first claim). We can consequently find an inverse of the function D on the image of D on $\mathcal{F}_{\delta_n}$. We then obtain, denoting the matrix ℓ_∞-norm by $|||\cdot|||_\infty$ and the inverse of D by $D^{\text{-}1}$, for all $f\in\mathcal{F}_{\delta_n}$

$$|||J_{D^{-1}}(f)|||_\infty=|||J_D^{-1}(f)|||_\infty=|||J_D(f)|||_\infty^{-1}\le[J_D(f)]_{0,0}^{-1}=\big[P\lambda''(f)\big]^{-1}\le(1-\sqrt{\delta_n})^{-1}c_\lambda^{-1}.$$

The constant c_λ is defined in (5.7) and the last inequality is obtained by the P-continuity of ρ'' and the condition on δ_n. The mean value theorem and the last inequality then imply for any $f,\tilde f\in\mathcal{F}_{\delta_n}$ and the associated coefficients t and $\tilde t$

$$\|t-\tilde t\|_{\ell_\infty}=\left\|D^{-1}\circ D(f)-D^{-1}\circ D(\tilde f)\right\|_{\ell_\infty}\le(1-\sqrt{\delta_n})^{-1}c_\lambda^{-1}\left\|D(f)-D(\tilde f)\right\|_{\ell_\infty}.$$

This proves the second claim. ■

Proof of Lemma 5. By the definitions of $P\big[\tilde D_\lambda^p(\cdot)\big]$ and $P^0\big[\tilde D_\lambda^p(\cdot)\big]$ in (5.4), we have for any $f\in\mathcal{F}_{\delta_n}$, any $\lambda\in\Lambda$, and any $p\in\mathcal{P}$

$$\begin{aligned} &\big|P^0\big[\tilde D_\lambda^p(f)\big]-P\big[\tilde D_\lambda^p(f)\big]\big|\\ &\quad\le\int\mu(x)K_h(x)\int\big|\rho'\big(\sigma(x)z+f^0(x)-f(x)\big)-\rho'\big(\sigma(x)z+f^*(x)-f(x)\big)\big|\ \mathbb{G}(z)dz\ dx(6.1) \end{aligned}$$

It additionally holds for all $f \in \mathcal{F}_{\delta_n}$ that $\sup_{x\in V_h} |f^0(x) - f(x)| \leq \delta_n$. Together with the definition of f^0 in (5.1), this implies for any $f \in \mathcal{F}_{\delta_n}$

$$\sup_{x\in V_h} |f^*(x) - f(x)| \leq \sup_{x\in V_h} |f^*(x) - f^0(x)| + \sup_{x\in V_h} |f^0(x) - f(x)| \leq b_h + \delta_n.$$

This implies, since ρ'' exists and is continuous with respect to the measure P (see Definition 1) and due to the mean value theorem, that for all $h \in \mathcal{H}$, all $\lambda \in \Lambda$ and for all $x \in V_h$ there is a $u_x \in \mathbb{R} : |u_x| \leq b_h + \delta_n$ such that

$$\begin{aligned}
&\left|\rho'\big(\sigma(x)z + f^0(x) - f(x)\big) - \rho'\big(\sigma(x)z + f^*(x) - f(x)\big)\right| \\
&\quad \leq |f^*(x) - f^0(x)|\rho''\big(\sigma(x)z + u_x\big) \\
&\quad \leq |f^*(x) - f^0(x)|\left(\rho''\big(\sigma(x)z\big) + 2L_{\rho''}(b_h + \delta_n)\right).
\end{aligned}$$

Using $\sqrt{b_h + \delta_n} \leq A\rho''_{\min}/(2L_{\rho''})$, (6.1), the last inequality, and the definitions of A, b_h, and c_λ defined in (2.2), (2.9) and (5.7) respectively, we obtain for any $\lambda \in \Lambda$

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_{\delta_n}} \left\|P^0\big[\tilde{D}_\lambda(f)\big] - P\big[\tilde{D}_\lambda(f)\big]\right\|_{\ell_\infty} \\
&\quad \leq \int \mu(x)K_h(x)|f^*(x) - f^0(x)| \int \left[\rho''\big(\sigma(x)z\big) + A\rho''_{\min}\sqrt{b_h + \delta_n}\right] \mathbb{G}(z)dz\,dx \\
&\quad \leq (1 + \sqrt{b_h + \delta_n})c_\lambda b_h.
\end{aligned}$$

■

Proof of Lemma 6. In this proof, we use a special case of a deviation inequality derived in [22, Corollary 6.9]. Adapted to our needs, this deviation inequality reads as follows:

Massart's Deviations Inequality: Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent, real valued random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Define $S_n(\pi) := \sum_{i=1}^n (\pi(\mathcal{X}_i) - \mathbb{E}\pi(\mathcal{X}_i))$ for a set of integrable, real valued functions $\pi \in \Pi$. If for some positive constants $\tilde{\sigma}$ and b

$$\sup_{\pi\in\Pi} n^{-1} \sum_{i=1}^n \mathbb{E}\big[\pi^2(\mathcal{X}_i)\big] \leq \tilde{\sigma}^2 \quad \text{and} \quad \sup_{\pi\in\Pi} \|\pi(\cdot)\|_\infty \leq \tilde{b}, \tag{6.2}$$

it holds for any $\epsilon \in (0, 1]$ and all $z > 0$

$$\mathbb{P}\left(\sup_{\pi\in\Pi} S_n(\pi) \geq E + 7\tilde{\sigma}\sqrt{2nz} + 2\tilde{b}z\right) \leq \exp(\text{-}z), \tag{6.3}$$

where

$$E := 27\sqrt{n} \int_0^{\tilde{\sigma}} \sqrt{H(u) \wedge n} du + 2(\tilde{b} + \tilde{\sigma})H(\tilde{\sigma}),$$

and $H(\cdot)$ is the $L_2(P)$-entropy with bracketing of Π.

Recall that the distance $\omega(\cdot, \cdot)$ is defined in (2.15). We now apply Massart's Inequality (6.3) with

$$\pi(\mathcal{X}_i) = \frac{\frac{1}{\sqrt{\Pi_h}}\rho'(Y_i - f(X_i))\,\left(\frac{X_i - x_0}{h}\right)^p K\left(\frac{X_i - x_0}{h}\right)}{\sqrt{\Pi_h}\sqrt{P\left[\lambda'(f^*)\right]^2 + \lambda'_\infty (n\Pi_h)^{\text{-}1/4}}},$$

$(P\left[\lambda'(f^*)\right]^2$ is given in (2.7)) and

$$\tilde{\sigma} = 1, \quad \tilde{b} = \frac{(n\Pi_h)^{1/4}}{\sqrt{\Pi_h}}, \quad H(\cdot) = H_{\mathcal{F}_{\delta_n} \cup \Lambda, \omega}(\cdot), \quad \text{and} \quad S_n(\pi) = \frac{n\sqrt{\Pi_h}\left(\tilde{D}^p_\lambda(f) - P\big[\tilde{D}^p_\lambda(f)\big]\right)}{\sqrt{\Pi_h}\sqrt{P\left[\lambda'(f^*)\right]^2 + \lambda'_\infty (n\Pi_h)^{-1/4}}}$$

to obtain

$$E = 27\sqrt{n}\int_0^1 \sqrt{H_{\mathcal{F}_{\delta_n,\omega} \cup \Lambda}(u)) \wedge n du} + 2\left(\frac{(n\Pi_h)^{1/4}}{\sqrt{\Pi_h}} + 1\right) H_{\mathcal{F}_{\delta_n} \cup \Lambda, \omega}(1),$$

and

$$\begin{aligned}
&\mathbb{P}_{f^*}\left(\sup_{f \in \mathcal{F}_{\delta_n}, \lambda \in \Lambda} \frac{\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\|_{\ell_\infty}}{\sqrt{\Pi_h}\sqrt{P\left[\lambda'(f^*)\right]^2 + \lambda'_\infty/(n\Pi_h)^{1/4}}} \geq \frac{E}{n\sqrt{\Pi_h}} + 7\sqrt{\frac{2z}{n\Pi_h}} + \frac{2z}{(n\Pi_h)^{3/4}}\right) \\
&\leq \sum_{p \in \mathcal{P}} \mathbb{P}_{f^*}\left(\sup_{f \in \mathcal{F}_{\delta_n}, \lambda \in \Lambda} \frac{n\sqrt{\Pi_h}\big|\tilde{D}^p_\lambda(f) - P\big[\tilde{D}^p_\lambda(f)\big]\big|}{\sqrt{\Pi_h}\sqrt{P\left[\lambda'(f^*)\right]^2 + \lambda'_\infty/(n\Pi_h)^{1/4}}} \geq E + 7\tilde{\sigma}\sqrt{2nz} + 2\tilde{b}z\right) \\
&\leq 2|\mathcal{P}|\exp(\text{-}z).
\end{aligned}$$

Note that the factor 2 in the last inequality appears because we need to control deviations of the absolute value of the empirical process. The claim is now deduced with simple calculations from the last display noticing that $B_z \geq E/\sqrt{n} + 7\sqrt{2z} + 2z(n\Pi_h)^{-1/4}$.
■

Proof of Lemma 7. First, we show that f^0 is the unique solution of the equality $P^0\big[\tilde{D}_\lambda(f)\big] = 0$ on $\mathcal{F}$. For this, we consider $f \in \mathcal{F}$ such that $P^0\big[\tilde{D}_\lambda(f)\big] = 0$. We then observe that

$$\sum_{p \in \mathcal{P}} (t^0_p - t_p)\, P^0\big[\tilde{D}^p_\lambda(f)\big] = 0,$$

where t is such that $\mathrm{P}_t = f$ and t^0 is defined in (5.1). Since $\mathbb{G}(\cdot) := n^{-1}\sum_{i=1}^n g_i(\cdot)$ is symmetric, K and μ are nonnegative, and ρ' is odd, the last equality implies

$$\begin{aligned}
&\int K_h(x)\mu(x)\big[f^0(x) - f(x)\big]\int \rho'\big(\sigma(x)z + f^0(x) - f(x)\big)\,\mathbb{G}(z)\,dz\,dx = 0 \\
\Leftrightarrow\quad &\int K_h(x)\mu(x)\big|f^0(x) - f(x)\big|\int \rho'\big(\sigma(x)z + |f^0(x) - f(x)|\big)\,\mathbb{G}(z)\,dz\,dx = 0 \\
\Leftrightarrow\quad &K_h(x)\mu(x)\big|f^0(x) - f(x)\big|\int \rho'\big(\sigma(x)z + |f^0(x) - f(x)|\big)\,\mathbb{G}(z)\,dz = 0 \quad \text{for all } x \in V_h.
\end{aligned}$$

Thus, if $f \neq f^0$, there exists an open, nonempty set $\widetilde{\mathcal{V}} \subseteq V_h$ such that

$$\sup_{x \in \widetilde{\mathcal{V}}} \int \rho'\big(\sigma(x)z + |f^0(x) - f(x)|\big)\,\mathbb{G}(z)\,dz = 0,$$

since f and f^0 are continuous. Recall that for any x, $\int \rho'(\sigma(x)z)\mathbb{G}(z)dz = 0$. Since $\mathbb{G}(z)$ is a density and therefore not translation invariant and since $\int \rho''(\sigma(x)z)\mathbb{G}(z)dz > 0$, this yields

$$\sup_{x \in \widetilde{\mathcal{V}}} \int \rho'\big(\sigma(x)z + |f^0(x) - f(x)|\big)\,\mathbb{G}(z)\,dz = 0 \Longrightarrow \sup_{x \in \widetilde{\mathcal{V}}} |f^0(x) - f(x)| = 0.$$

This contradicts $f \neq f^0$ because f and f_0 are polynomials of finite degree. In other words, f^0 is the unique solution of $P^0\big[\tilde{D}_\lambda(f)\big] = 0$ on $\mathcal{F}$.

We now look at the event $\big\{\hat{f}_\lambda \in \mathcal{F}_{\delta_n} \text{ for all } \lambda \in \Lambda\big\}$. To this end, we recall that $\hat{f}_\lambda$ is the solution of equation $\tilde{D}_\lambda(\cdot) = 0$ and the following inclusions hold:

$$\begin{aligned}
&\big\{\exists\lambda \,:\, \hat{f}_\lambda \notin \mathcal{F}_{\delta_n}\big\}\\
&\subseteq \left\{\exists\lambda \,:\, \sup_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_n}} \big\|\tilde{D}_\lambda(f) - P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1} \geq \inf_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_n}} \big\|P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1}\right\}\\
&\subseteq \left\{\sup_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_n},\lambda\in\Lambda} \big\|\tilde{D}_\lambda(f) - P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1} \geq \inf_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_n},\lambda\in\Lambda} \big\|P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1}\right\}. \qquad (6.4)
\end{aligned}$$

Next, for any $\lambda \in \Lambda$, it holds that

$$\begin{aligned}
&\big\|\tilde{D}_\lambda(f) - P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1}\\
&\quad\leq \big\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1} + \big\|P\big[\tilde{D}_\lambda(f)\big] - P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1}\\
&\quad\leq |\mathcal{P}|\,\big\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_\infty} + |\mathcal{P}|\,\big\|P\big[\tilde{D}_\lambda(f)\big] - P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_\infty}. \qquad (6.5)
\end{aligned}$$

We then set $\vartheta_h := \int \mu(x)K_h(x)dx$ and use the continuity of ρ' ot derive, similiarly as in Lemma 5, for any $\lambda \in \Lambda$

$$\sup_{f\in\mathcal{F}} \big\|P\big[\tilde{D}_\lambda(f)\big] - P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_\infty} \leq \vartheta_h b_h. \qquad (6.6)$$

To control the stochastic term, we can then apply Massart's Inequality (6.3) with $\pi = f$,

$$\tilde{\sigma} = 1, \quad \tilde{b} = \frac{1}{\sqrt{\Pi_h}}, \quad H(\cdot) = H_{\mathcal{F}\cup\Lambda,\omega}(\cdot), \quad \textit{and} \quad S_n(\pi) = \frac{n\sqrt{\Pi_h}}{\gamma_{\max}\mathcal{K}_{\max}}\left(\tilde{D}^p_\lambda(f) - P\big[\tilde{D}^p_\lambda(f)\big]\right)$$

to obtain

$$E = 27\sqrt{n}\int_0^1 \sqrt{H_{\mathcal{F}\cup\Lambda,\omega}(u) \wedge n du} + 2\left(\frac{1}{\sqrt{\Pi_h}} + 1\right) H_{\mathcal{F}}(1),$$

and

$$\begin{aligned}
&\mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F},\lambda\in\Lambda} \big\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_\infty} \geq \frac{\gamma_{\max}\mathcal{K}_{\max}}{n\sqrt{\Pi_h}}E + 7\gamma_{\max}\mathcal{K}_{\max}\sqrt{\frac{2z}{n\Pi_h}} + \frac{2\gamma_{\max}\mathcal{K}_{\max}z}{n\Pi_h}\right)\\
&\leq \sum_{p\in\mathcal{P}} \mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F},\lambda\in\Lambda} \frac{n\sqrt{\Pi_h}}{\gamma_{\max}\mathcal{K}_{\max}}\big|\tilde{D}^p_\lambda(f) - P\big[\tilde{D}^p_\lambda(f)\big]\big| \geq E + 7\tilde{\sigma}\sqrt{2nz} + 2\tilde{b}z\right)\\
&\leq 2|\mathcal{P}|\exp(\text{-}z).
\end{aligned}$$

Setting $\varepsilon := E' + 7\gamma_{\max}\mathcal{K}_{\max}\sqrt{\frac{2z}{n\Pi_h}} + \frac{2\lambda'_\infty z}{n\Pi_h}$ and $E' := \frac{\gamma_{\max}\mathcal{K}_{\max}}{n\sqrt{\Pi_h}}E$, we can rewrite the last inequality to get

$$\mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F},\lambda\in\Lambda} \big\|\tilde{D}_\lambda(f) - P\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_\infty} \geq \varepsilon\right) \leq 2|\mathcal{P}|\exp\left(-\frac{n\Pi_h(\varepsilon - E')^2}{98\gamma^2_{\max}\mathcal{K}^2_{\max} + 4\gamma_{\max}\mathcal{K}_{\max}(\varepsilon - E')}\right).$$

Using (6.5), (6.6), and the last inequality, we then obtain for all $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_n},\lambda\in\Lambda} \big\|\tilde{D}_\lambda(f) - P^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1} \geq \varepsilon\,|\mathcal{P}|\right)\\
&\qquad\leq 2|\mathcal{P}|\exp\left(-\frac{n\Pi_h(\varepsilon - E' - \vartheta_h b_h)^2}{98\gamma^2_{\max}\mathcal{K}^2_{\max} + 4\gamma_{\max}\mathcal{K}_{\max}(\varepsilon - E' - \vartheta_h b_h)}\right). \qquad (6.7)
\end{aligned}$$

Now, let us look at $\inf_{f\in\mathcal{F}\backslash\mathcal{F}_{\delta_n},\lambda\in\Lambda}\left\|P^0\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_1}$ in (6.4). By the definition of $\tilde{D}_\lambda(\cdot)$, we have for any $f\in\mathcal{F}\backslash\mathcal{F}_{\delta_n}$ and any $\lambda\in\Lambda$

$$\begin{aligned}\left\|P^0\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_1} &= \sum_{p\in\mathcal{P}}\left|\int\left(\frac{x-x_0}{h}\right)^p\mu(x)K_h(x)\int\rho'\left(\sigma(x)z+f^0(x)-f(x)\right)\mathbb{G}(z)dz\,dx\right|\\ &\geq \sum_{p\in\mathcal{P}}\frac{|t_p^0-t_p|}{\|t^0-t\|_{\ell_1}}\left|\int\left(\frac{x-x_0}{h}\right)^p\mu(x)K_h(x)\int\rho'\left(\sigma(x)z+f^0(x)-f(x)\right)\mathbb{G}(z)dz\,dx\right|\\ &\geq \left|\int\frac{f^0(x)-f(x)}{\|t^0-t\|_{\ell_1}}\mu(x)K_h(x)\int\rho'\left(\sigma(x)z+f^0(x)-f(x)\right)\mathbb{G}(z)dz\,dx\right|,\end{aligned}$$

where t is such that $f=\mathrm{P}_t$. Since $\mathbb{G}(\cdot)$ is symmetric, $\rho'(\cdot)$ increasing (because of the convexity of ρ), K is nonnegative, and ρ' is odd (ρ is symmetric) and positive on $\mathbb{R}_+^*$ (because of $\rho'(0)=0$, the convexity of ρ and the strict convexity around 0), the last equality implies for all $f\in\mathcal{F}\backslash\mathcal{F}_{\delta_n}$

$$\begin{aligned}\left\|P^0\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_1} &\geq \int\frac{\left|f^0(x)-f(x)\right|}{\|t^0-t\|_{\ell_1}}\mu(x)K_h(x)\int\rho'\left(\sigma(x)z+\left|f^0(x)-f(x)\right|\right)\mathbb{G}(z)dz\,dx\\ &\geq \int\frac{\left|f^0(x)-f(x)\right|}{\|t^0-t\|_{\ell_1}}\mu(x)K_h(x)\int\rho'\left(\sigma(x)z+\delta_n\frac{\left|f^0(x)-f(x)\right|}{\|t^0-t\|_{\ell_1}}\right)\mathbb{G}(z)dz\,dx.\end{aligned}$$

Since $\left|f^0(x)-f(x)\right|\|t^0-t\|_{\ell_1}^{-1}\leq 1$, we obtain with the mean value theorem for all $f\in\mathcal{F}\backslash\mathcal{F}_{\delta_n}$

$$\begin{aligned}\left\|P^0\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_1} &\geq \delta_n\int\frac{\left|f^0(x)-f(x)\right|^2}{\|t^0-t\|_{\ell_1}^2}\mu(x)K_h(x)\inf_{u\in[0,\delta_n]}\int\rho''\left(\sigma(x)z+u\right)\mathbb{G}(z)dz\,dx\\ &\geq \delta_n\inf_{t:\|t\|_{\ell_1}\geq\delta_n}\int\frac{\left|\mathrm{P}_t(x)\right|^2}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)\inf_{u\in[0,\delta_n]}\int\rho''\left(\sigma(x)z+u\right)\mathbb{G}(z)dz\,dx.\end{aligned}$$

We then derive, using that $\sqrt{\delta_n}\leq\frac{1}{2}\wedge\frac{A\rho''_{\min}}{2L_{\rho''}}$, and ρ'' is P-continuous,

$$\inf_{f\in\mathcal{F}\backslash\mathcal{F}_{\delta_n}}\left\|P^0\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_1}\geq\frac{\delta_nA\rho''_{\min}}{2}\inf_{t:\|t\|_{\ell_1}\geq\delta_n}\int\frac{\left|\mathrm{P}_t(x)\right|^2}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)\,dx.$$

We then observe that $\mathrm{P}_t(x)=t\mathbb{X}^\top$ and thus

$$\int\frac{\left|\mathrm{P}_t(x)\right|^2}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)dx=t\left[\int\frac{\mathbb{X}^\top\mathbb{X}}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)dx\right]t^\top.$$

The matrix $\int\mathbb{X}^\top\mathbb{X}\mu(x)K_h(x)dx$ is positive definite (this follows from standart results, see, for instance, [30], Lemma 1.6). We can thus write

$$t\left[\int\frac{\mathbb{X}^\top\mathbb{X}}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)dx\right]t^\top\geq\nu/|\mathcal{P}|,$$

where ν is the smallest eigenvalue of the matrix $\int\mathbb{X}^\top\mathbb{X}\mu(x)K_h(x)dx$. In summary, we have

$$\inf_{f\in\mathcal{F}\backslash\mathcal{F}_{\delta_n},\lambda\in\Lambda}\left\|P^0\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_1}\geq\frac{A\rho''_{\min}\nu\delta_n}{2|\mathcal{P}|}.$$

With $\delta_n = 2|\mathcal{P}|^2((\ln n)^{\text{-}1} + \vartheta_h b_h)/(A\rho''_{\min}\nu)$, Inequalities (6.4) and (6.7), and the last inequality, we obtain

$$\begin{aligned}\mathbb{P}_{f^*}\left\{\exists\lambda\in\Lambda\ :\ \hat{f}_\lambda\notin\mathcal{F}_{\delta_n}\right\} &\leq \mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F}\backslash\mathcal{F}_{\delta_n},\lambda\in\Lambda}\left\|\tilde{D}_\lambda(f)-P^0\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_1}\geq\frac{A\rho''_{\min}\nu\delta_n}{2|\mathcal{P}|}\right)\\ &\leq 2|\mathcal{P}|\exp\left(-\frac{n\Pi_h((\ln n)^{\text{-}1}-E')^2}{98\gamma^2_{\max}\mathcal{K}^2_{\max}+4\gamma_{\max}\mathcal{K}_{\max}((\ln n)^{\text{-}1}-E')}\right).\end{aligned}$$

Invoking Condition 1 on n in Section 5.1 and the definition of E', the desired claim follows. ■

Proof of Lemma 8. For any $x, y \geq 0$, we have

$$\begin{aligned}x^q &= |x-y+y|^q\\ &= |[x-y]_+ + y|^q 1\{x\geq y\} + |y-[y-x]_+|^q 1\{x<y\}\\ &\leq (2^q[x-y]_+^q + 2^q y^q)1\{x\geq y\} + y^q 1\{x<y\}\\ &\leq 2^q[x-y]_+^q + 2^q y^q.\end{aligned}$$

For the second part, we set $x := (x_1, \ldots, x_l)^T$ and use Hölder's Inequality to derive

$$\|x\|_{\ell_1} \leq l^{1-1/q}\|x\|_{\ell_q}$$

from which the proof follows. ■

Proof of Lemma 9. We first recall by the definition of the estimator (2.13)

$$\sqrt{\widehat{\mathrm{V}}(\lambda)} = \frac{\sqrt{\Pi_h P_n\left[\lambda'(\hat{f}_\lambda)\right]^2} + \lambda'_\infty(n\Pi_h)^{-1/4}}{P_n\lambda''(\hat{f}_\lambda)},$$

where

$$\Pi_h P_n\left[\lambda'(\hat{f}_\lambda)\right]^2 = \sum_{i=1}^n \frac{1}{n\Pi_h}\left[\rho'(Y_i-\hat{f}_\lambda(X_i))\right]^2 K^2\left(\frac{X_i-x_0}{h}\right)$$

and

$$P_n\lambda''(\hat{f}_\lambda) = \sum_{i=1}^n \frac{1}{n\Pi_h}\rho''(Y_i-\hat{f}_\lambda(X_i))K\left(\frac{X_i-x_0}{h}\right).$$

Then, using Massart's Inequality (given in the proof of Lemma 6) with $\pi(\mathcal{X}_i) = \frac{1}{\gamma^2_{\max}\mathcal{K}^2_{\max}\sqrt{\Pi_h}}\left[\rho'(Y_i - f(X_i))\right]^2 K^2\left(\frac{X_i-x_0}{h}\right)$, $\tilde{\sigma} = 1$, $\quad \tilde{b} = \frac{1}{\sqrt{\Pi_h}}$,

$$H(\cdot) = H_{\mathcal{F}_{\delta_n}\cup\Lambda,\omega}(\cdot), \quad z = \ln n, \quad \text{and} \quad S_n(\pi) = \frac{\Pi_h^{3/2}}{\gamma^2_{\max}\mathcal{K}^2_{\max}}\left(P_n\left[\lambda'(f)\right]^2 - P\left[\lambda'(f)\right]^2\right),$$

we can control the deviations the process π as follows:

$$\mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F}_{\delta_n},\lambda\in\Lambda}\Pi_h\left|P_n\left[\lambda'(f)\right]^2 - P\left[\lambda'(f)\right]^2\right| \geq \frac{\gamma^2_{\max}\mathcal{K}^2_{\max}B_{\ln(n)}}{\sqrt{n\Pi_h}}\right) \leq 2/n^2, \tag{6.8}$$

where $\gamma_{\max}$, $\mathcal{K}_{\max}$ and B_z are defined in Section (2). Similarly, using Massart's Inequality with $\pi(\mathcal{X}_i) = \frac{1}{\gamma_{\max}\mathcal{K}_{\max}\sqrt{\Pi_h}}\rho''(Y_i - f(X_i))K\left(\frac{X_i - x_0}{h}\right)$, $\tilde{\sigma} = 1$, $\tilde{b} = \frac{1}{\sqrt{\Pi_h}}$,

$$H(\cdot) = H_{\mathcal{F}_{\delta_n}\cup\Lambda,\omega}(\cdot), \quad z = \mathrm{ani}_\epsilon(n), \quad \text{and} \quad S_n(\pi) = \frac{\sqrt{\Pi_h}}{\gamma_{\max}\mathcal{K}_{\max}}\left(P_n\lambda''(f) - P\lambda''(f)\right),$$

we control the deviations of π as follows:

$$\mathbb{P}_{f^*}\left(\sup_{f\in\mathcal{F}_{\delta_n},\lambda\in\Lambda}|P_n\lambda''(f) - P\lambda''(f)| \geq \frac{\gamma_{\max}\mathcal{K}_{\max}B_{\ln(n)}}{\sqrt{n\Pi_h}}\right) \leq 2/n^2. \tag{6.9}$$

Then, by the continuity of ρ' and ρ'' almost everywhere, $\rho' \leq \gamma_{\max}$, $\|\rho''\|_\infty \leq 1$, $\gamma_{\max} \geq 1$ (see Definition 1), $f, f^* \in \mathcal{F}_{\delta_n}$, and the mean value theorem, we have for all $f \in \mathcal{F}_{\delta_n}$

$$\begin{aligned}\Pi_h\left|P\left[\lambda'(f)\right]^2 - P\left[\lambda'(f^*)\right]^2\right| &\leq \frac{1}{n\Pi_n}\sum_{i=1}^n \mathbb{E}\left|\rho'(Y_i - f(X_i))^2 - \rho'(Y_i - f^*(X_i))^2\right|K^2\left(\frac{X_i - x_0}{h}\right) \\ &\leq 2\gamma_{\max}\mathcal{K}_{\max}(\delta_n + b_h).\end{aligned}$$

Similarly,

$$\sup_{f\in\mathcal{F}_{\delta_n}}|P\lambda''(f) - P\lambda''(f^*)| \leq L_{\rho''}\mathcal{K}_\infty(\delta_n + b_h) \leq L_{\rho''}\gamma_{\max}\mathcal{K}_{\max}(\delta_n + b_h).$$

Denote by $s_n := (2 \vee L_{\rho''})\gamma_{\max}\mathcal{K}_{\max}(\delta_n + b_h) + \frac{\gamma_{\max}^2\mathcal{K}_{\max}^2 B_{\ln(n)}}{\sqrt{n\Pi_h}}$. Moreover, we observe (under Condition 2 on n) that

$$s_n \leq \tilde{a}_n \max\left\{A\rho''_{\min}\int K_h(x)\mu(x)dx, \Pi_h \inf_{\lambda\in\Lambda} P\left[\lambda'(f^*)\right]^2\right\},$$

and thus

$$s_n \leq \tilde{a}_n \max\left\{\inf_{\lambda\in\Lambda} P\lambda''(f^*), \Pi_h \inf_{\lambda\in\Lambda} P\left[\lambda'(f^*)\right]^2\right\}.$$

Using this, (6.8), and (6.9), we obtain for $\lambda \in \Lambda$ with probability $1 - 2/n^2 - \mathbb{P}_{f^*}(\Omega^c)$

$$\sqrt{\widehat{\mathrm{V}}(\lambda)} \leq \frac{\sqrt{\Pi_h P\left[\lambda'(f^*)\right]^2 + s_n} + \lambda'_\infty(n\Pi_h)^{-1/4}}{P\lambda''(f^*) - s_n} \leq \frac{\sqrt{1 + \tilde{a}_n}}{1 - \tilde{a}_n}\sqrt{\mathrm{V}(\lambda)},$$

and

$$\sqrt{\widehat{\mathrm{V}}(\lambda)} \geq \frac{\sqrt{\Pi_h P\left[\lambda'(f^*)\right]^2 - s_n} + \lambda'_\infty(n\Pi_h)^{-1/4}}{P\lambda''(f^*) + s_n} \geq \frac{\sqrt{1 - \tilde{a}_n}}{1 + \tilde{a}_n}\sqrt{\mathrm{V}(\lambda)}.$$

This proves the claim. ■

Proof of Lemma 10. We recall that

$$\mathcal{D}_h(t) = \int \hat{K}_h(x)\int \hat{\rho}'\big(\sigma z + f^*(x) - t\big)\mathbb{G}(z)dzdx,$$

and thus, with the mean value theorem, there exists a $c \in [t, \tilde{t}]$ such that

$$\mathcal{D}_h(t) - \mathcal{D}_h(\tilde{t}) = (\tilde{t} - t)\int \hat{K}_h(x)\int \hat{\rho}''\big(\sigma z + f^*(x) - c\big)\mathbb{G}(z)dzdx.$$

As $t, \tilde{t} \in [f^*(x_0) - \delta_n, f^*(x_0) + \delta_n]$ and $f^* \in \mathbb{H}_d(\vec{\beta}, L, M)$, we have for any $x \in V_h$

$$|f^*(x) - c| \leq |f^*(x) - f^*(x_0)| + |f^*(x_0) - c| \leq b_{h_{\max}} + \delta_n. \tag{6.10}$$

Using $c_{\hat{\rho}} = \int \hat{\rho}''(\sigma z)\mathbb{G}(z)dz$ and the previous two inequalities, we obtain

$$\begin{aligned} &|\mathcal{D}_h(t) - \mathcal{D}_h(\tilde{t})| \\ &\quad = |t - \tilde{t}| \left| \int \hat{K}_h(x) \int \hat{\rho}''\big(\sigma z + f^*(x) - c\big)\mathbb{G}(z)dzdx \right| \\ &\quad = |t - \tilde{t}| \left| \int \hat{K}_h(x) \int \hat{\rho}''\big(\sigma z + f^*(x) - c\big)\mathbb{G}(z)dzdx - \int \hat{K}_h(x) \int \hat{\rho}''\big(\sigma z\big)\mathbb{G}(z)dzdx + c_{\hat{\rho}} \right|. \end{aligned}$$

As $\hat{\rho}''$ is $L_{\hat{\rho}''}$–Lipschitz, we obtain with (6.10) and Condition 1 in Section 5.1

$$\begin{aligned} &\left| \int \hat{K}_h(x) \int \hat{\rho}''\big(\sigma z + f^*(x) - c\big)\mathbb{G}(z)dzdx - \int \hat{K}_h(x) \int \hat{\rho}''\big(\sigma z\big)\mathbb{G}(z)dzdx \right| \\ &\leq L_{\hat{\rho}''} \int \hat{K}_h(x)|f^*(x) - c|dx \\ &\leq L_{\hat{\rho}''}(b_{h_{\max}} + \delta_n) \\ &\leq c_{\hat{\rho}}\sqrt{b_{h_{\max}} + \delta_n}. \end{aligned}$$

We then deduce from the last two displays that

$$|\mathcal{D}_h(t) - \mathcal{D}_h(\tilde{t})| \geq c_{\hat{\rho}}(1 - \sqrt{b_{h_{\max}} + \delta_n})|t - \tilde{t}|$$

and with Condition 1 finally

$$|t - \tilde{t}| \leq (1 + 2\sqrt{b_{h_{\max}} + \delta_n})c_{\hat{\rho}}^{-1}|\mathcal{D}_h(t) - \mathcal{D}_h(\tilde{t})|.$$

■

Proof of Lemma 11. Let us first set for any $h \in \mathcal{H}$

$$\tau_h := \frac{a_n\sqrt{\mathrm{V}(\rho^*, K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_h}}.$$

Thus, with Lemma 8, we obtain

$$\begin{aligned} &\mathbb{E}_{f^*} c_{\hat{\rho}}^{-q} \sup_{f \in \mathcal{F}_{\delta_n}} \left| \tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f) \right|^q \\ &\qquad \leq 2^q \mathbb{E}_{f^*} \left(c_{\hat{\rho}}^{-1} \sup_{f \in \mathcal{F}_{\delta_n}} \left| \tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f) \right| - \tau_h \right)_+^q + 2^q \tau_h^q. \end{aligned} \tag{6.11}$$

Next, note that $c_{\hat{\rho}}^{-1} \left| \tilde{\mathcal{D}}_h(\cdot) - \mathcal{D}_h(\cdot) \right| \leq 2\gamma_{\max}\mathcal{K}_{\max}(A\rho''_{\min})^{-1} =: T$. Consequently,

$$\begin{aligned} &\mathbb{E}_{f^*} \left(c_{\hat{\rho}}^{-1} \sup_{f \in \mathcal{F}_{\delta_n}} \left| \tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f) \right| - \tau_h \right)_+^q \\ &\qquad \leq q \int_0^T u^{q-1} \mathbb{P}_{f^*} \left(c_{\hat{\rho}}^{-1} \sup_{f \in \mathcal{F}_{\delta_n}} \left| \tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f) \right| - \tau_h \geq u \right) du. \end{aligned} \tag{6.12}$$

Similarly as in (5.14), we derive on the event $\{\Delta, \forall h \in \mathcal{H}_\epsilon\}$

$$\sqrt{\mathrm{V}(\rho^*, K^*)} \geq \sqrt{\mathrm{V}(\lambda_h^*)} \geq a_n^{-1}\sqrt{\mathrm{V}(\tilde{\lambda}_h)},$$

where $\lambda_h^*(x,y,f) := \rho^*\big(y - f(x)\big)\, K_h^*(x)$ and $\tilde{\lambda}_h(x,y,f) := \hat{\rho}\big(y - f(x)\big)\, \hat{K}_h(x)$ for all $x, y \in \mathbb{R}$. Setting $u = \frac{a_n\sqrt{\mathrm{V}(\rho^*,K^*)}}{\sqrt{n\Pi_h}}\varepsilon$ in (6.12), using the last inequality, and Lemma 6 with $z > 0$ such that $B_0 + \mathrm{ani}_\epsilon(n) + \varepsilon = B_z$, we get

$$\begin{aligned}
&\mathbb{E}_{f^*}\left(c_{\hat{\rho}}^{-1} \sup_{f \in \mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f)\right| - \tau_h\right)_+^q \\
&\quad \leq q\tau_h^q \int_0^T \varepsilon^{q-1}\mathbb{P}_{f^*}\left(c_{\hat{\rho}}^{-1} \sup_{f \in \mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f)\right| \geq \frac{\sqrt{\mathrm{V}(\tilde{\lambda}_h)}(B_0 + \mathrm{ani}_\epsilon(n) + \varepsilon)}{\sqrt{n\Pi_h}}\right) d\varepsilon \\
&\quad \leq q\tau_h^q \int_0^T \varepsilon^{q-1}\mathbb{P}_{f^*}\left(\sup_{\rho,K} \sup_{f \in \mathcal{F}_{\delta_n}} \frac{\left|\tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f)\right|}{c_\rho\sqrt{\mathrm{V}(\lambda_h)}} \geq \frac{B_0 + \mathrm{ani}_\epsilon(n) + \varepsilon}{\sqrt{n\Pi_h}}\right) d\varepsilon \\
&\quad \leq 2q\tau_h^q \int_0^T \varepsilon^{q-1} \exp\left(-\frac{(\varepsilon + \mathrm{ani}_\epsilon(n))^2}{100 + 4(\varepsilon + \mathrm{ani}_\epsilon(n))/(n\Pi_h)^{1/4}}\right) d\varepsilon
\end{aligned}$$

.

Using $\mathrm{ani}_\epsilon(n)/(n\Pi_h)^{1/4} \leq 1$, $(2\gamma_{\max}\mathcal{K}_{\max}(\rho''_{\min}A)^{-1})/(n\Pi_h)^{1/4} \leq 1$ (Condition 3 on n in Section 5.1), and (5.10) with $a = 104$ and $b = 4$, we get

$$\begin{aligned}
&\mathbb{E}_{f^*}\left(c_{\hat{\rho}}^{-1} \sup_{f \in \mathcal{F}_{\delta_n}} \left|\tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f)\right| - \tau_h\right)_+^q \\
&\quad \leq 2q\tau_h^q \exp\left(-\frac{(\mathrm{ani}_\epsilon(n))^2}{108}\right) \int_0^T \varepsilon^{q-1} \exp\left(-\frac{\varepsilon^2}{104 + 4\varepsilon}\right) d\varepsilon \\
&\quad \leq \frac{2q}{n|\mathcal{H}_\epsilon|}\tau_h^q (11.4)^q \,\mathrm{Gamma}(q).
\end{aligned}$$

From (6.11) and the last inequality, the lemma can be deduced. ∎

Proof of Lemma 12. Recall that we consider the uniform design and the homoscedastic noise level. By the definition of $\mathcal{D}_h$ and with a change of variables, we have

$$\begin{aligned}
&\sup_{f \in \mathcal{F}_{\delta_n}} \left|\mathcal{D}_{h' \vee h}(f) - \mathcal{D}_h(f)\right| \\
&= \sup_{f \in \mathcal{F}_{\delta_n}} \bigg|\int \hat{K}(x) \int \hat{\rho}'\big(\sigma z + f^*(x_0 + h \vee h'x) - f(x_0)\big)\mathbb{G}(z) dz\, dx \\
&\qquad - \int \hat{K}(x) \int \hat{\rho}'\big(\sigma z + f^*(x_0 + hx) - f(x_0)\big)\mathbb{G}(z) dz\, dx\bigg|,
\end{aligned}$$

where $\mathbb{G}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} g_i(\cdot)$. Using $f \in \mathbb{H}_d(\vec{\beta}, L, M)$, the L_1-continuity of ρ'', the last equality, and the mean value theorem, we obtain:

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_{\delta_n}} \big|\mathcal{D}_{h'\vee h}(f) - \mathcal{D}_h(f)\big| \\
&\qquad \le \sup_{|s|\le 2\delta_n + 2b_{h_{\max}}} \int \hat{\rho}''(\sigma z + s)\mathbb{G}(z)dz \int \hat{K}(x)\big|f^*(x_0 + h\vee h'x) - f^*(x_0 + hx)\big|dx \\
&\qquad \le \left(\int \hat{\rho}''(\sigma z)\mathbb{G}(z)dz + 2L_{\rho''}(\delta_n + b_{h_{\max}})\right) L\sum_{j=1}^{d} |h_j \vee {h_j}' - h_j|^{\beta_j}.
\end{aligned}$$

With Condition 1, this yields

$$\sup_{f\in\mathcal{F}_{\delta_n}} \big|\mathcal{D}_{h'\vee h}(f) - \mathcal{D}_h(f)\big| \le (1 + \sqrt{\delta_n + b_{h_{\max}}})c_{\hat{\rho}} L\sum_{j=1}^{d} (h'_j)^{\beta_j}.$$

■

Acknoledgements

We thank Oleg Lepski, Joseph Salmon, and Sara van de Geer for many helpful discussions.

References

[1] Arcones, M. A. (2005). Convergence of the optimal M-estimator over a parametric family of M-estimators. *Test* **14** 281–315.

[2] Astola, J., Egiazarian, K., Foi, A. and Katkovnik, V. (2010). From Local Kernel to Nonlocal Multiple-Model Image Denoising. *Int. J. Comput. Vision* **86** 1–32.

[3] Bertin, K. (2004). Estimation asymptotiquement exacte en norme sup de fonctions multidimensionnelles PhD thesis, Paris 6.

[4] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves : exponential bounds and rate of convergences. *Bernoulli* **4** 329-375.

[5] Brown, L. and Low, M. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535.

[6] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data. Springer Series in Statistics.* Springer, Heidelberg. Methods, theory and applications.

[7] Cai, T. T. and Zhou, H. H. (2009). Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.* **37** 3204–3235.

[8] Chichignoud, M. (2011). Minimax and minimax adaptive estimation in multiplicative regression : locally bayesian approach. *Probab. Theory Related Fields.* to appear.

[9] Gaïffas, S. (2005). Convergence rates for pointwise curve estimation with a degenerate design. *Math. Methods Statist.* **14** 1–27.

[10] Goldenshluger, A. and Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.* **6** 135–170.

[11] Hoffmann, M. and Nickl, R. (2011). Robust nonparametric estimation via wavelet median regression. *Ann. Statist.* **39** 2383–2409.

[12] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

[13] Huber, P. J. (1981). *Robust statistics.* Wiley, New York.

[14] Huber, P. and Ronchetti, E. (2009). *Robust statistics*, second ed. *Wiley Series in Probability and Statistics.* John Wiley & Sons Inc., Hoboken, NJ.

[15] Katkovnik, V. (1985). *Nonparametric identification and data smoothing.* "Nauka", Moscow (in Russian). The method of local approximation.

[16] Kerkyacharian, G., Lepski, O. V. and Picard, D. (2001). Non linear estimation in anisotropic multi-index denoising. *Probab. Theory and Related Fields* **121** 137-170.

[17] Klutchnikoff, N. (2005). *On the adaptive estimation of anisotropic functions.* Ph.D. thesis, Aix-Masrseille 1.

[18] Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Stat.* **5** 1015–1053.

[19] Lepski, O. V. (1990). On a Problem of Adaptive Estimation in Gaussian White Noise. *Theory of Probability and its Applications* **35** 454-466.

[20] Lepski, O. V. and Levit, B. Y. (1999). Adaptive nonparametric estimation of smooth multivariate functions. *Mathematicals methods of statistics.*

[21] Lepski, O. V., Mammen, E. and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947.

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

[23] Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation. *Probab. Theory Related Fields* **135** 335–362.

[24] Reiss, M., Rozenholc, Y. and Cuenod, C. (2011). Pointwise adaptive estimation for robust and quantile regression. Source: Arxiv.

[25] Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284.

[26] Tsybakov, A. B. (1982). Nonparametric signal estimation when there is incomplete information on the noise distribution. *Problems of Information Transmission* **18** 116–130.

[27] Tsybakov, A. B. (1982). Robust estimates of a function. *Problems of Information Transmission* **18** 39–52.

[28] Tsybakov, A. B. (1983). Convergence of nonparametric robust algorithms of reconstruction of functions. *Automation and Remote Control* 12 66–76.

[29] Tsybakov, A. B. (1986). Robust reconstruction of functions by a local approximation method. *Problems of Information Transmission* **22** 69–84.

[30] Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation.* Springer Publishing Company, Incorporated.

[31] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes. Springer Series in Statistics.* Springer-Verlag, New York. With applications to statistics.