Bernoulli **20**(3), 2014, 1560–1599
DOI: 10.3150/13-BEJ533

A robust, adaptive M-estimator for pointwise estimation in heteroscedastic regression

MICHAËL CHICHIGNOUD[*] and JOHANNES LEDERER[**]

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

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. Moreover, we do not require a positive density for the design distribution. In a first step, we study the consistency of locally polynomial M-estimators that consist of a contrast and a kernel. Afterwards, minimax results are established over unidimensional Hölder spaces for degenerate design. We then choose the contrast and the kernel that minimize an empirical variance term 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, multivariate, 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.

Keywords: adaptation; Huber contrast; Lepski's method; M-estimation; minimax estimation; nonparametric regression; pointwise estimation; robust estimation

1. Introduction

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

Let us first briefly summarize the related literature. First, the seminal paper [14] contains a proof of the asymptotic normality of M-estimators for the location parameter in regular models. Furthermore, the series of papers [30–33] provide minimax results for nonparametric regression. More recently, a block median method was used in [5] to prove the asymptotic equivalence between Gaussian regression and homoscedastic regression for deterministic designs and possibly heavy-tailed noises. Using a blockwise Stein's Method with wavelets, this leads to an S-adaptive 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. Moreover, using an estimate of the noise density at 0 and a plug-in method, this also leads to a D-adaptive estimator. However, in contrast to this paper, only homoscedastic regression is considered and multivariate regression functions, in particular anisotropic functions, are not allowed for. Next, a modified version of Lepski's method was applied for homoscedastic regression in [27]. Finally, local M-estimators, also for regression models with degenerate designs, were intensively studied in the

1350-7265 © 2014 ISI/BS

case of Gaussian regression: S-adaptivity results of a local least squares estimator were derived in [8], sup-norm S-minimax results were established in [9], and the effect of degenerate designs on the L_2-norm was investigated with wavelet-type estimators in [1]. However, in contrast to this paper, pointwise estimation with random, possibly degenerate designs and heteroscedastic, possibly heavy-tailed noises has not been included.

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 the contrast $\rho(\cdot)$ and the 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 [14–16])

$$\sqrt{n}\big(\hat{t} - t^0\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{AV}), \qquad \text{where } \mathrm{AV} := \frac{\int (\rho')^2 \,\mathrm{d}G}{(\int \rho'' \,\mathrm{d}G)^2}, \tag{1.1}$$

G is the distribution of $\mathcal{Y} - t^0$, $\rho'(\cdot)$ and $\rho''(\cdot)$ are the first and second derivatives of the contrast $\rho(\cdot)$, 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 by minimizing the asymptotic variance. Moreover, the Crámer–Rao Inequality and (see [14])

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

where $I(\cdot)$ is the Fisher information and the infimum is taken over all twice differentiable contrasts, imply that this M-estimator is efficient. Huber proposed in [14], Proposal 3, to minimize an estimate of the above asymptotic variance (since the distribution G is not available in practice) over the family of Huber contrasts (their definition is given below). He also conjectured that the corresponding estimator is minimax for certain contamination models (for more details, see Section A.1 in the arXiv version). More recently, in [2], an M-estimator with a contrast that minimizes an estimate of the asymptotic variance was introduced for the parametric model, its asymptotic normality was proved, and especially Huber contrasts indexed by their scale and a family of ℓ_p losses were considered.

In a first step, we derive general properties of M-estimators such as pointwise risk bounds. This includes, in particular, S-minimax results for degenerate designs and allows us to recover results in [7] (see Theorem 1 and Remark 1). In a second step, we then consider a local M-estimator that consists of a contrast and a kernel that minimize an estimate of the variance and show, in particular, that this estimator mimics the oracle, which minimizes the true variance. Our data-driven approach can be used, for example, for the selection of the scale of the Huber contrast with an adaptive robustness with respect to outliers or for the selection of a suitable (even noncentered or nonconvex) support that takes a maximal number of points around x_0 into account (cf. [12] for the latter objective). Finally, we show that our estimator is, under some restrictions on the design and the noise level (see Condition 3), D-adaptive for various sets of contrasts and kernels with finite entropy.

We finally study simultaneous D- and S-adaptation for anisotropic target functions. In a first step, we study the case of isotropic target functions, where the standard Lepski's method (see [23, 24]) can be applied. To this end, we assume that the variance of the estimator is decreasing with respect to the bandwidth and plug-in an estimate of the minimal variance for the D-adaptation to apply Lepski's method for the S-adaptation (see Section 4.1). This yields the first estimator in heteroscedastic regression with random designs and heavy-tailed noise distributions that is simultaneously D- and S-adaptive and optimal in a sense describe later. Furthermore, we note that applications of Lepski's method to nonlinear estimators are still nonstandard and can only be found in a small number of examples in the literature [6,26,27]. In a next step, we extend our results to anisotropic target functions. For this, we restrict ourselves to locally constant target functions and homoscedastic regression with uniform design and apply a modification of Lepski's method given in [19,22] to construct an optimal, simultaneously S- and D-adaptive estimator. This is the first application of Lepski's method to nonlinear estimators of anisotropic target functions and yields a selection of an anisotropic bandwidth which is of great interest for applications in the context of image denoising (cf. [17]), for example.

Although we consider estimation problems, our approach may also be useful for inference, for example, for the construction of confidence bands. While confidence bands for parametric estimation are derived from central limit theorems (see (1.1)), confidence bands for nonparametric regression are especially desired to be adaptive with respect to the smoothness of the target function. The construction of such S-adaptive confidence bands is more difficult than in the parametric case (see [13]), but since Lepski-type procedures have already been used in this context, see [10], Theorem 1 and Corollary 1, we expect that our approach may be useful for the construction of S-adaptive confidence bands for regression with possibly heavy-tailed noises (see Section 5 for a discussion of some technical aspects). Eventually, if for example the smoothness is known, our approach may be used, plugging an estimate of the variance in the confidence band, to obtain D-adaptive confidence bands, which are, in particular, adaptive with respect to the design and the noise distributions.

The structure of this paper is as follows: In the following section, we first introduce an estimator which satisfies a risk bound (see Theorem 1). So, S-minimax results are deduced over Hölder spaces (see Corollaries 1, 2 and 3). We then provide a choice for the contrast and the kernel (see Theorem 2) via the minimization of a nonasymptotic variance. 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 5. The proofs are finally conducted in Section 6 and in the Appendix. For conciseness, only the crucial proofs are presented here. For the remaining proofs and more details, in particular, on the parametric model and on a comparison to classical results, we refer to the longer version available on arXiv and the webpages of the authors.

2. Preliminary definitions and results

In this section, we give some preliminary definitions and results. After specifying the model, we introduce a first estimator and then, we present a risk bound and S-minimax properties of this estimator.

Let us first specify the model. The observations $(X_i, Y_i)_{i=1,\dots,n}$ satisfy the set of equations

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

and are distributed according to the probability measure $\mathbb{P} := \mathbb{P}^{(n)}_{f^*}$ with associated expectation $\mathbb{E} := \mathbb{E}^{(n)}_{f^*}$. We aim at estimating the target function $f^* : [0, 1]^d \to [-M, M]$ (for $M > 0$) 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 $(\xi_i)_{i \in 1,\dots,n}$ 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 $g_i(\cdot)$ may be unknown but are assumed to be symmetric. We stress that we do not impose, unlike in the literature on the median (cf. [5]), any moment assumptions on the noise, and we do not require that the noise densities are positive at 0. We postpone the detailed discussion on the assumptions to the end of the next section. The noise level $\sigma : [0, 1]^d \to [0, \infty)$ 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 distributions do not have any moments, for example. Finally, the design points $(X_i)_{i \in 1,\dots,n}$ are assumed to be distributed independently and identically according to the density $\mu(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}$. We assume that $\mu(\cdot)$ vanishes at most finitely many points. For ease of exposition, we also assume that $(X_i)_{i \in 1,\dots,n}$ and $(\xi_i)_{i \in 1,\dots,n}$ are mutually independent.

Next, we introduce an estimator of $f^*(x_0)$ with a local polynomial approach (LPA) for a fixed bandwidth, a fixed kernel, and a fixed contrast. The key idea of the LPA, as described for example in [18] or in [34], Chapter 1, is to approximate the target function in a neighborhood of size $h \in (0, 1]^d$ of a given point x_0 by a polynomial. To start, we define for a fixed $m \in \mathbb{N}$ the set $\mathcal{P} := \{p = (p_1, \dots, p_d)^\top \in \mathbb{N}^d : 0 \le |p| \le m\}$ with $|p| = p_1 + \cdots + p_d$ and denote its cardinality by $|\mathcal{P}|$. The cardinality $|\mathcal{P}|$ is exponential in d and enters the bounds derived below as a factor. For any multi-indexed column vector $t = (t_{p_1,\dots,p_d} \in \mathbb{R} : p \in \mathcal{P}) \in \mathbb{R}^{|\mathcal{P}|}$ and for any $x \in [0, 1]^d$, we then define the desired polynomial as

$$\mathrm{P}_t(x) := t^\top U\left(\frac{x - x_0}{h}\right) := \sum_{p \in \mathcal{P}} t_p \left(\frac{x - x_0}{h}\right)^p.$$

Here, $z^p := z_1^{p_1} \cdots z_d^{p_d}$ for all $z \in \mathbb{R}^d$, and the division by h is understood coordinate wise. Next, for $M > 0$, we define $\mathcal{F} := \{\mathrm{P}_t : t \in [-M, M]^{|\mathcal{P}|}\}$ as a set of polynomials of degree at most m. We now specify what we mean by a kernel and a contrast:

Definition 1. *A function $K : \mathbb{R}^d \to [0, \infty)$ is called kernel (function) if it has the following properties:*

1. *$K(\cdot)$ has a (not necessarily symmetric) support which is a hypercube having edge length one and contains the origin;*
2. *$\|K\|_\infty < \infty$ and $\int K(x)\,\mathrm{d}x = 1$.*

For ease of exposition, we set $\Pi_h := \prod_{j=1}^d h_j$ and use the notation $K_h(\cdot) := K((\cdot - x_0)/h)/\Pi_h$ at some points. Moreover, we define the neighborhood of x_0 of size h as $V_h :=$

$\{x \in \mathbb{R}^d : K_h(x) > 0\}$ and assume for simplicity that the kernel is chosen such that $V_h \subseteq [0,1]^d$. Next, we specify what we mean by a contrast:

Definition 2. *A function $\rho : \mathbb{R} \to [0,\infty)$ is called contrast (function) if it has the following properties:*

1. *$\rho(\cdot)$ is convex, symmetric and $\rho(0) = 0$;*
2. *the derivative $\rho'(\cdot)$ of $\rho(\cdot)$ is 1-Lipschitz and bounded;*
3. *the second derivative $\rho''(\cdot)$ of $\rho(\cdot)$ is defined Lebesgue almost everywhere and is 1-Lipschitz with respect to the measure $\mathbb{P}$. Moreover, $\|\rho''\|_\infty \le 1$.*

The constants in the Lipschitz condition and the boundedness condition in the last definition are set to 1 for ease of exposition only. Well-known contrasts are the Huber contrast (see [14]), for any scale $\gamma > 0$ and $z \in \mathbb{R}$,

$$\rho_{\mathrm{H},\gamma}(z) := \begin{cases} z^2/2, & \text{if } |z| \le \gamma, \\ \gamma\bigl(|z| - \gamma/2\bigr), & \text{otherwise,} \end{cases} \tag{2.2}$$

and the contrast induced by the arctan function (see [30])

$$\rho_{\mathrm{arc},\gamma}(z) := \gamma z \arctan(z/\gamma) - \frac{\gamma^2}{2}\ln\bigl(1 + z^2/\gamma^2\bigr). \tag{2.3}$$

Note that the square loss and the absolute loss do not satisfy the above definition. However, they can be mimicked by the Huber contrast with γ small (median) and γ large (mean). Let us define, for any function ζ, the empirical measure as $P_n\zeta := \frac{1}{n}\sum_{i=1}^n \zeta(X_i, Y_i)$. We can now combine a kernel and a contrast to obtain the λ-LPA estimator $\hat{f}_\lambda(x_0)$ of $f^*(x_0)$ defined as:

$$\begin{aligned} \hat{f}_\lambda := \arg\min_{f \in \mathcal{F}} P_n\lambda(f), \qquad & \text{where } \lambda(f)(x,y) := \rho\bigl(y - f(x)\bigr)K_h(x) \\ & \text{for } x \in [0,1]^d \text{ and } y \in \mathbb{R}. \end{aligned} \tag{2.4}$$

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 $f^*(x_0)$.

2.1. A first risk bound

In this section, we present a risk bound for the estimator introduced above. This estimator involves, in particular, fixed contrasts, kernels and bandwidths.

To ease the presentation, we introduce some additional definitions. First, we define the best approximation of the target f^* in $\mathcal{F}$ as

$$f^0 := \arg\min\Bigl\{\sup_{x \in V_h}\bigl|f(x) - f^*(x)\bigr| : f \in \mathcal{F}, f(x_0) = f^*(x_0)\Bigr\} \tag{2.5}$$

and the associated bias term as

$$b_h(\mathcal{F}) := \sup_{x \in V_h} \bigl| f^0(x) - f^*(x) \bigr|. \tag{2.6}$$

The minimum is not necessarily unique, but all minimizers work for our derivations. We then fix a multi-indexed vector $t^0 = (t^0_{p_1,\dots,p_d})_{p \in \mathcal{P}}$ such that $\mathrm{P}_{t^0} = f^0$. We recall that the entropy with bracketing of a set of functions $\mathcal{A}$ for a given radius $u > 0$ with respect to a (pseudo)metric Δ is the logarithm of the minimal number of pairs of functions $(f_1^{(j)}, f_2^{(j)}) \in \mathcal{A} \times \mathcal{A}$ such that for any $f \in \mathcal{A}$, there is a couple $(f_1^{(j)}, f_2^{(j)})$ such that $f_1^{(j)} \leq f \leq f_2^{(j)}$ and $\Delta(f_1^{(j)}, f_2^{(j)}) \leq u$. Here, in particular, $H_{\mathcal{F}}(\cdot)$ denotes the entropy with bracketing of $\mathcal{F}$ with respect to the pseudometric $\sqrt{\Pi_h \mathbb{E} P_n [\lambda'(f_1) - \lambda'(f_2)]^2}$, $f_1, f_2 \in \mathcal{F}$, where

$$\lambda'(f)(x, y) := \rho'\bigl(y - f(x)\bigr) K_h(x) \qquad \text{for } x \in [0,1]^d \text{ and } y \in \mathbb{R}. \tag{2.7}$$

The entropy $H_{\mathcal{F}}(\cdot)$ cannot be calculated if the probability law is unknown. However, it can be upper bounded invoking an upper bound for the pseudometric. For this, one may use that

$$\sqrt{\Pi_h \mathbb{E} P_n \bigl[\lambda'(f_1) - \lambda'(f_2)\bigr]^2} \leq \|K\|_\infty \bigl\| t^{(1)} - t^{(2)} \bigr\|_1$$

due to the continuity of $\rho'(\cdot)$ and the definition of $\mathcal{F}$. Here, $t^{(1)}$ and $t^{(2)} \in [-M, M]^{|\mathcal{P}|}$ are such that $\mathrm{P}_{t^{(1)}} = f_1$ and $\mathrm{P}_{t^{(2)}} = f_2$, respectively. Therefore, the entropy $H_{\mathcal{F}}(\cdot)$ can be bounded by $|\mathcal{P}|$ times the entropy of $[-M, M]$ with respect to the Euclidean distance multiplied by $\|K\|_\infty$ (this is, in particular, independent of n).

As a next step, we introduce the condition under which we derive the risk bound.

Condition 1. *Let $\rho(\cdot)$ be a contrast, $K(\cdot)$ a kernel, $n \in \{1, 2, \dots\}$, and $h \in (0, 1]^d$. We say that Condition 1 is satisfied if the smallest eigenvalue Φ_h of the matrix*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}\biggl[U\Bigl(\frac{X_i - x_0}{h}\Bigr) U^\top \Bigl(\frac{X_i - x_0}{h}\Bigr) \rho''\bigl(\sigma(X_i)\xi_i\bigr) K_h(X)\biggr]$$

is positive, $n\Pi_h \geq 1$, and, defining

$$\delta_h := \frac{2|\mathcal{P}|^2}{\Phi_h} \biggl[\mathbb{E}\bigl[K_h(X)\bigr] b_h(\mathcal{F}) + \frac{54\|\rho'\|_\infty(\sqrt{\mathbb{E}[\Pi_h K_h^2(X)]} + \|K\|_\infty/\sqrt{n\Pi_h})}{\sqrt{n\Pi_h}(\ln(n|\mathcal{P}|) + \int_0^1 H_{\mathcal{F}}^{1/2}(u)\,\mathrm{d}u + H_{\mathcal{F}}(1))^{-1}}\biggr],$$

that

$$4\bigl(b_h(\mathcal{F}) + \delta_h\bigr) \leq \inf_{x \in V_h} \frac{1}{n} \sum_{i=1}^n \mathbb{E}\bigl[\rho''\bigl(\sigma(x)\xi_i\bigr)\bigr]. \tag{2.8}$$

Condition 1 can be interpreted in the following sense: n must be sufficiently large and h appropriate for the setting under consideration. In particular, h, as a function of n, is usually chosen such that $h \to 0$ and $n\Pi_h \to \infty$ as $n \to \infty$ to satisfy Condition 1. We postpone a detailed discussion of Condition 1 to after the main result of this section.

The variance term of the estimator is crucial for the following. To state it explicitly, we need to introduce some more notation: First, we introduce λ'' (similarly as λ' in (2.7)) as

$$\lambda''(f)(x, y) := \rho''\big(y - f(x)\big)K_h(x) \qquad \text{for } x \in [0, 1]^d \text{ and } y \in \mathbb{R}.$$

We then introduce the crucial quantity

$$\mathrm{V}(\lambda) := \left(\frac{\sqrt{\Pi_h \mathbb{E}P_n[\lambda'(f^*)]^2} + \|\rho'\|_\infty \|K\|_\infty \ln^2(n)/\sqrt{n\Pi_h}}{\mathbb{E}P_n\lambda''(f^*)}\right)^2. \tag{2.9}$$

We call it nonasymptotic variance, since it plays the role of the variance in the risk bounds in the theorems below. From Condition 1 and Definitions 1 and 2, we conclude that $\mathrm{V}(\lambda) < \infty$. The term $\|\rho'\|_\infty \|K\|_\infty \frac{\ln^2(n)}{\sqrt{n\Pi_h}}$ depends on h and n. However, the bandwidth is typically chosen such that $n\Pi_h \to \infty$ for $n \to \infty$ so that this term vanishes asymptotically. Additionally, besides the normalization $\sqrt{\Pi_h}$ in front of the first term, a dependence on h is given through λ. We will discuss this after giving the main result of this section. If $h = (1, \dots, 1)^\top$ (parametric case), the nonasymptotic variance $\mathrm{V}(\lambda)$ tends towards the asymptotic variance $\mathrm{AV}(\lambda)$ defined in (1.1) as $n \to \infty$.

The main result of this section reads as the following.

Theorem 1. *Let λ be as in (2.4), $n \in \{1, 2, \dots\}$, and $h \in (0, 1]^d$ such that Condition 1 is satisfied. Then, for all $q \geq 1$,*

$$\begin{aligned}&\mathbb{E}\big|\hat{f}_\lambda(x_0) - f^*(x_0)\big|^q \\ &\quad\leq C_q\left(b_h(\mathcal{F}) + \left[27\int_0^1 H_{\mathcal{F}}^{1/2}(u)\,\mathrm{d}u + \frac{4H_{\mathcal{F}}(1)}{\ln^2(n)} + 1\right]\frac{\sqrt{\mathrm{V}(\lambda)}}{\sqrt{n\Pi_h}}\right)^q + 2^q\frac{M^q}{n^2}\end{aligned}$$

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

The proof can be easily deduced integrating the result of Proposition 3 and using Proposition 2 (the propositions can be found in the Appendix, see also the more detailed arXiv version).

Remark 1. In contrast to Huber's asymptotic results (see [14] and also [2,30–33]), the above theorem holds for finite (but sufficiently large) sample sizes n. We note that the desired variance term $\mathrm{V}(\lambda)$ is found up to constants, which are of minor interest for this paper. Moreover, a wide range of designs (including degenerate designs, e.g.) and noise levels (including zero noise, e.g.) is covered. Let us compare this result to [7]: assume that $d = 1$ and the noise $(\sigma(X_i)\xi_i)_i$ is identically and independently normal distributed with variance $\sigma > 0$, and consider the local Huber estimator with $\rho(\cdot) = \rho_{\mathrm{H},\ln(n)}(\cdot)$ (2.2), where $\gamma = \ln(n)$, and the indicator kernel $K(\cdot) = \mathbb{1}_{[-1/2,1/2]}(\cdot)$. As we mentioned above, the Huber estimator, with a large parameter γ, mimics the local least squares estimator. Indeed it holds

$$\frac{\sqrt{\mathrm{V}(\lambda)}}{\sqrt{n\Pi_h}} \asymp \frac{\sigma}{\sqrt{n\int_{x_0-h/2}^{x_0+h/2}\mu(x)\,\mathrm{d}x}}.$$

The term on the right-hand side is the classical standard deviation of the local least squares estimator. Theorem 1 then implies the results of [7], Theorem 1 and Proposition 1, in the Gaussian case and extends them to heteroscedastic, heavy-tailed noises.

Remark 2. While the above bound is – to the best of our knowledge – already a new result, the final goal is to provide a specific λ that minimizes this bound since the second term $2^q M^q/n^2$ is neglectable and since the bias term $b_h(\mathcal{F})$ is independent of λ. However, the bandwidth h, which accounts for the smoothness of the target function, is included in $\mathrm{V}(\lambda)$. This makes simultaneous D- and S-adaptation difficult. The specific dependences of the numerator and the denominator on h can be deduced from

$$\Pi_h \mathbb{E} P_n\big[\lambda'(f^*)\big]^2 = \Pi_h \int \mu(x)K_h^2(x)\int\big[\rho'\big(\sigma(x)z\big)\big]^2 n^{-1}\sum_i g_i(z)\,\mathrm{d}z\,\mathrm{d}x \tag{2.10}$$

and

$$\mathbb{E} P_n \lambda''(f^*) = \int \mu(x)K_h(x)\int \rho''\big(\sigma(x)z\big)n^{-1}\sum_i g_i(z)\,\mathrm{d}z\,\mathrm{d}x. \tag{2.11}$$

We study this in detail in the following section for three examples.

Discussion of Condition 1. *The condition $\Phi_h > 0$ is fulfilled in many examples. Indeed, with a change of variables and by the definition of $K_h(\cdot)$, we obtain*

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\bigg[U\Big(\frac{X_i - x_0}{h}\Big)U^\top\Big(\frac{X_i - x_0}{h}\Big)\rho''\big(\sigma(X_i)\xi_i\big)K_h(X)\bigg]\\ &\quad= \int U(x)U^\top(x)\mu(x_0+hx)K(x)\int \rho''\big(\sigma(x_0+hx)z\big)\frac{1}{n}\sum_{i=1}^n g_i(z)\,\mathrm{d}z\,\mathrm{d}x.\end{aligned}$$

According to [34], *Lemma* 1.6, *a sufficient condition for $\Phi_h > 0$ is thus that*

$$\mu(x_0+hx)K(x)\int \rho''\big(\sigma(x_0+hx)z\big)\frac{1}{n}\sum_{i=1}^n g_i(z)\,\mathrm{d}z > 0 \tag{2.12}$$

for all x in some set in the kernel support with positive Lebesgue measure. Recall that $\mu(x_0+h\cdot)$ is positive almost everywhere in the support of $K(\cdot)$ since $\mu(\cdot)$ vanishes only at finitely many points. The condition $\Phi_h > 0$ is thus fulfilled if

$$\inf_{x\in V_h}\int \rho''\big(\sigma(x)z\big)\frac{1}{n}\sum_{i=1}^n g_i(z)\,\mathrm{d}z > 0. \tag{2.13}$$

This condition is satisfied, for example, for all densities $g_i(\cdot)$ and bounded $\sigma(\cdot)$ if the contrast function is strictly convex. This holds true for $\rho_{\mathrm{arc},\gamma}(\cdot)$ (see (2.3)). The Huber contrast $\rho_{\mathrm{H},\gamma}(\cdot)$ (see (2.2)), however, is strictly convex on the interval $(-\gamma,\gamma)$ only. It holds that $\rho''_{\mathrm{H},\gamma}(\cdot)=\mathbb{1}_{[-\gamma,\gamma]}(\cdot)$; therefore, the densities $g_i(\cdot)$ have to satisfy the additional constraint

$$\inf_{x\in V_h}\int \mathbb{1}_{[-\gamma,\gamma]}\bigl(\sigma(x)z\bigr)\frac{1}{n}\sum_{i=1}^{n} g_i(z)\,\mathrm{d}z>0$$

to ensure $\Phi_h>0$ in this case. If we assume, for simplicity, that the noise level is constant $\sigma(\cdot)\equiv\sigma>0$, the last constraint simplifies to $\int_{-\gamma/\sigma}^{\gamma/\sigma}\frac{1}{n}\sum_{i=1}^{n} g_i(z)\,\mathrm{d}z>0$. So even for the Huber contrast with a fixed $\gamma>0$, the assumption $\Phi_h>0$ is weaker than the standard assumption in the literature of $g_i(\cdot)$ being positive and continuous in the origin for all $i\in\{1,2,\dots,n\}$.

For the other crucial part of the condition, we first note that for $h\to 0$, the quantity on the right-hand side of (2.8) tends to a positive constant if $\sigma(\cdot)$ is continuous in x_0. Indeed, since ρ'' is $\mathbb{P}$-continuous, it holds that

$$\inf_{x\in V_h}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\bigl[\rho''\bigl(\sigma(x)\xi_i\bigr)\bigr]\longrightarrow\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\bigl[\rho''\bigl(\sigma(x_0)\xi_i\bigr)\bigr]\tag{2.14}$$

as $h\to 0$. Similarly as above, the quantity on the right-hand side can be lower bounded by a positive constant for many contrasts and noise densities. We now give the rate for the quantity δ_h. It holds that

$$\delta_h\asymp\Phi_h^{-1}\biggl[\mathbb{E}\bigl[K_h(X)\bigr]b_h(\mathcal{F})+\frac{\sqrt{\mathbb{E}\Pi_h K_h^2(X)}}{\sqrt{n\Pi_h}}\ln(n)+\frac{\|K\|_\infty}{n\Pi_h}\ln(n)\biggr],$$

which should be (cf. (2.8)) bounded by a constant. Here, "$\asymp$" indicates the asymptotic dependence on n. The above display corresponds to a so-called bias–variance decomposition up to a factor $\ln(n)$. We also note that if f^* is continuous as assumed in the standard literature, the bias term tends to zero as $h\to 0$. In the literature, one typically chooses some couple of positive constants (α_1,α_2), and $\sigma(\cdot)$ and some bandwidth $h=(h_1,\dots,h_d)$ such that

$$n^{-\alpha_1/d}\bigl(\ln(n)\bigr)^{\alpha_2}\le h_j\le\bigl(\ln(n)\bigr)^{-1}\qquad\text{for all } j=1,\dots,d,\tag{2.15}$$

where we assume that n is sufficiently large such that the above inequalities can hold. For appropriate (α_1,α_2), Condition 1 is then satisfied for n sufficiently large in many examples.

Example 1. If, for example, the design is uniform ($\mu(\cdot)\equiv 1$) and the noise level homoscedastic ($\sigma(\cdot)\equiv\sigma>0$), it holds that $\Phi_h\asymp$ *const* and thus $\delta_h\asymp b_h(\mathcal{F})+\ln(n)/\sqrt{n\Pi_h}$. Choosing a bandwidth $h=h_n$ as in (2.15) with $\alpha_1=1$ and $\alpha_2=4$, Condition (2.8) is satisfied for n sufficiently large.

Example 2. For degenerated designs, however, it is possible that $\Phi_h \to 0$ as $h \to 0$. For example, let $d = 1$, the noise level be homoscedastic ($\sigma(\cdot) \equiv \sigma > 0$), and

$$\mu(\cdot) = \frac{s+1}{x_0^{s+1} + (1-x_0)^{s+1}} |\cdot - x_0|^s \mathbb{1}_{[0,1]}(\cdot) \tag{2.16}$$

with $s > -1$ and $x_0 \in [0, 1]$ (see [7]). The density explodes (for $s < 0$) or vanishes (for $s > 0$) at x_0, so that one will either have a lot or very little observations in the vicinity of x_0. This is reflected in δ_h (recall that $d = 1$ and thus $h \in (0, 1]$):

$$\delta_h \asymp b_h(\mathcal{F}) + \frac{\ln(n)}{\sqrt{nh^{s+1}}}.$$

So, similarly as above, one may choose a bandwidth like in (2.15) with $\alpha_1 = 1/(s+1)$ and $\alpha_2 = 4/(s+1)$.

We finally note that the concrete form of Condition 1 is due to the application of deviation inequalities for bounded empirical processes. Similarly, we could relax the boundedness condition on the empirical processes involved to Bernstein conditions (see, e.g., [35]). This allows to incorporate unbounded contrasts such as the least squares contrast and the factor $\|\rho'\|_\infty$ in Condition 1 should be replaced by the factor $\sqrt{\mathbb{E}[\rho'(\sigma(X)\xi)]^2}$, where ξ would be a sub-Gaussian random variable.

2.2. S-minimax results

In this section, we deduce some corollaries adapted to simple examples from the above results.

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

$$R_{n,q}[\tilde{f}, \mathcal{S}] := \sup_{f^* \in \mathcal{S}} \mathbb{E}\big|\tilde{f}(x_0) - f^*(x_0)\big|^q \quad \text{and} \quad R_{n,q}[\mathcal{S}] := \inf_{\bar{f}} R_{n,q}[\bar{f}, \mathcal{S}], \tag{2.17}$$

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 3. *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*

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

We can give some simple examples for one dimensional target functions, that is, $d = 1$. We call $\mathbb{H}_1(\beta, L, M)$ Hölder space, with parameters $\beta, L, M > 0$, the set of $\lfloor\beta\rfloor$-times differentiable functions $f : [0, 1] \to \mathbb{R}$ such that $\|f^{(j)}\|_\infty \le M$ for all $j \in \{0, 1, \dots, \lfloor\beta\rfloor\}$ and satisfied the Hölder continuity $|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)| \le L|x-y|^{\beta-\lfloor\beta\rfloor}$ for all $x, y \in [0, 1]^d$.

The following corollary can now be easily deduced from Theorem 1.

Corollary 1. *Consider the model in Example 1, that is, uniform design* ($\mu(\cdot) \equiv 1$) *and homoscedastic noise level* ($\sigma(\cdot) \equiv \sigma > 0$). *Let* β, L *and* M *be positive parameters. Moreover, let* $\hat{f}_\lambda$ *be defined as in* (2.4) *with* $m = \lfloor \beta \rfloor$, $h \asymp n^{-1/(2\beta+1)}$, $\rho(\cdot) = \rho_{\text{arc},1}(\cdot)$ *as in* (2.3), *and* $K(\cdot) := \mathbb{1}_{[-1/2,1/2]}(\cdot)$. *Then, it holds that*

$$\Pi_h \mathbb{E} P_n \big[\lambda'(f^*)\big]^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\rho'_{\text{arc},1}(\sigma \xi_i)\big]^2,$$

$$\mathbb{E} P_n \lambda''(f^*) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\rho''_{\text{arc},1}(\sigma \xi_i)$$

and

$$\limsup_{n\to\infty} n^{q\beta/(2\beta+1)} \, R_{n,q}\big(\hat{f}_\lambda, \mathbb{H}_1(\beta, L, M)\big) < \infty.$$

The rate $n^{-\beta/(2\beta+1)}$ is a standard S-minimax rate in the context Gaussian noise (see [34], Chapter 2). Here, however, this rate is achieved for a large class of noise distributions.

Similarly, one can deduce the next corollary.

Corollary 2. *Consider the model in Example 2, that is, a degenerate design as in* (2.16) *with* $s > -1$ *and a homoscedastic noise level* ($\sigma(\cdot) \equiv \sigma > 0$). *Let* β, L, *and* M *be positive parameters. Moreover, let* $\hat{f}_\lambda$ *be defined as in* (2.4) *with* $m = \lfloor \beta \rfloor$, $h \asymp n^{-1/(2\beta+s+1)}$, $\rho(\cdot) = \rho_{\text{arc},1}(\cdot)$ *as in* (2.3), *and* $K(\cdot) := \mathbb{1}_{\{-1/2,1/2\}}(\cdot)$. *Then, it holds that*

$$\Pi_h \mathbb{E} P_n \big[\lambda'(f^*)\big]^2 = \frac{h^s}{x_0^{s+1} + (1-x_0)^{s+1}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\rho'_{\text{arc},1}(\sigma \xi_i)\big]^2,$$

$$\mathbb{E} P_n \lambda''(f^*) = \frac{h^s}{x_0^{s+1} + (1-x_0)^{s+1}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\rho''_{\text{arc},1}(\sigma \xi_i)$$

and

$$\limsup_{n\to\infty} n^{q\beta/(2\beta+s+1)} \, R_{n,q}\big(\hat{f}_\lambda, \mathbb{H}_1(\beta, L, M)\big) < \infty.$$

Thus, the rate $n^{-\beta/(2\beta+s+1)}$ is achieved. This rate is S-minimax in the nonparametric regression with homoscedastic Gaussian noise (see [7]). Note that we have only considered examples with homoscedastic noises here. For heteroscedastic noises, the dependence on h can be very involved for some contrast functions (cf. equations (2.10) and (2.11)). But, as highlighted by the next example, this is not always the case.

Corollary 3. *Consider the model* (2.1) *with* $d = 1$, *a degenerate design as in* (2.16) *with* $s > -1$, *a heteroscedastic noise level* $\sigma(\cdot) \equiv |\cdot - x_0|^\alpha$, $0 \le \alpha \le s/2$, *and a noise* $(\xi_i)_i$ *with finite variance.*

Let β, L and M be positive parameters. Moreover, let $\hat{f}_\lambda$ be defined as in (2.4) with $m = \lfloor \beta \rfloor$, $h \asymp n^{-1/(2\beta+s-2\alpha+1)}$, $\rho(\cdot) = \rho_{\mathrm{H},\ln(n)}(\cdot)$ as in (2.2), and $K(\cdot) := \mathbb{1}_{\{-1/2,1/2\}}(\cdot)$. Then, it holds that

$$\Pi_h \mathbb{E} P_n\big[\lambda'\big(f^*\big)\big]^2 \asymp h^{s+2\alpha}, \qquad \mathbb{E} P_n \lambda''\big(f^*\big) \asymp h^s$$

and

$$\limsup_{n\to\infty} n^{q\beta/(2\beta+s-2\alpha+1)} R_{n,q}\big(\hat{f}_\lambda, \mathbb{H}_1(\beta, L, M)\big) < \infty.$$

This result illustrates the effect of small noise levels on the rate and the possible compensations to degenerate (unfavorable) designs. In particular, if $\alpha = s/2$, we get the standard minimax rate $n^{-\beta/(2\beta+1)}$ as in Corollary 1. We assume that α is smaller than $s/2$, since otherwise the noise level is very small and the bandwidth chosen is thus as small as possible. We also recall that the noise level is assumed to be bounded so that we only consider the case $\alpha \geq 0$.

3. A D-adaptive estimator for fixed bandwidths

In this section, we discuss the selection of the combined function λ, that is, of the kernel and the contrast. For this, we introduce an oracle that minimizes the bound in Theorem 1 above and then provide an estimator that mimics this oracle. This estimator is then D-adaptive, that is, adaptive with respect to the noise and the design distributions.

To this end, we first introduce $\Lambda := \Upsilon \times \mathcal{K}$ as the set of possible combined functions λ as in (2.4) for a given set of contrasts Υ, a given set of kernels $\mathcal{K}$, and a fixed bandwidth $h \in (0,1]^d$. For example, one may consider a subset of the set of Huber functions indexed by the scale $\gamma > 0$ as set of contrasts $\Upsilon := \{\rho_{\mathrm{H},\gamma}(\cdot) : \gamma > 0\}$. An example for the set of kernels is the set of indicator functions with different supports as

$$\begin{aligned}\mathcal{K} &:= \big\{\mathbb{1}_{S(u)}(\cdot) : u \in [-1/2, 1/2]^d\big\} \\ &\quad \text{for } S(u) := [-1/2 + u_1, 1/2 + u_1] \times \cdots \times [-1/2 + u_d, 1/2 + u_d].\end{aligned}$$

This contains, in particular, the symmetric indicator kernel $\mathbb{1}_{S(0)}(\cdot)$. In this section, the bandwidth h is fixed so that the bias term $b_h(\mathcal{F})$ in Theorem 1 is of minor importance; we then introduce the oracle as the minimizer of the variance (2.9)

$$\lambda^* := \arg\min_{\lambda\in\Lambda} \mathrm{V}(\lambda). \tag{3.1}$$

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

$$\begin{aligned}&\widehat{\lambda} := \arg\min_{\lambda\in\Lambda} \widehat{\mathrm{V}}(\lambda), \\ &\quad \text{where } \widehat{\mathrm{V}}(\lambda) := \left(\frac{\sqrt{\Pi_h P_n[\lambda'(\hat{f}_\lambda)]^2} + \|\rho'\|_\infty \|K\|_\infty \ln^2(n)/\sqrt{n\Pi_h}}{P_n \lambda''(\hat{f}_\lambda)}\right)^2.\end{aligned} \tag{3.2}$$

Note that we estimate the target function f^* by $\hat{f}_\lambda$ and $\mathbb{E}P_n[\lambda'(f^*)]^2$ and $\mathbb{E}P_n\lambda''(f^*)$ by their empirical versions $P_n[\lambda'(\hat{f}_\lambda)]^2$ and $P_n\lambda''(\hat{f}_\lambda)$, respectively. The explicit expressions for the numerator and the denominator can be obtained using

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

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

We now show that the estimator $\hat{f}_{\hat{\lambda}}$ that results from (2.4) and (3.2) performs – up to constants – as well as the oracle $\hat{f}_{\lambda^*}$. For this, we define $H_{\mathcal{F}\times\Lambda}(\cdot)$ as the entropy with bracketing of $\mathcal{F}\times\Lambda$ with respect to the (pseudo)metric

$$\sqrt{\Pi_h\mathbb{E}P_n\big[\kappa(f_1,\lambda_1) - \kappa(f_2,\lambda_2)\big]^2} \vee \sqrt{\Pi_h\mathbb{E}P_n\big[\lambda_1''(f_1) - \lambda_2''(f_2)\big]^2} \tag{3.3}$$

for any $f_1, f_2 \in \mathcal{F}, \lambda_1, \lambda_2 \in \Lambda$, where $\kappa(f,\lambda) := \lambda'(f)/(\sqrt{\Pi_h\mathbb{E}P_n[\lambda'(f^*)]^2} + \|\rho'\|_\infty\|K\|_\infty \times \ln^2(n)/\sqrt{n\Pi_h})$. We compute in the Appendix a bound for this entropy for the set of Huer contrasts indexed by the scale.

Before giving the main result of this section, we give the necessary assumptions.

Condition 2. *Let $\Lambda = \Upsilon \times \mathcal{K}$ be a set of functions as in (2.4) where Υ is a set of contrasts as in Definition 2 and $\mathcal{K}$ is a set of kernels as in Definition 1, $n \in \{1, 2, \ldots\}$, and $h \in (0,1]^d$. We say that Condition 2 is satisfied if the smallest eigenvalue Φ_h (defined in Condition 1) is positive, $n\Pi_h \geq \ln^4(n)$, and, defining for any $\lambda \in \Lambda$*

$$\delta_h^*(\lambda) := \frac{2|\mathcal{P}|^2}{\Phi_h}\left[\mathbb{E}\big[K_h(X)\big]b_h(\mathcal{F}) + \frac{54\|\rho'\|_\infty(\sqrt{\mathbb{E}[\Pi_h K_h^2(X)]} + \|K\|_\infty/\sqrt{n\Pi_h})}{\sqrt{n\Pi_h}(\ln(n|\mathcal{P}|) + \int_0^1 H_{\mathcal{F}\times\Lambda}^{1/2}(u)\,\mathrm{d}u + H_{\mathcal{F}\times\Lambda}(1))^{-1}}\right],$$

it holds for all $\lambda \in \Lambda$

$$4\big(b_h(\mathcal{F}) + \delta_h^*(\lambda)\big) \leq \inf_{x\in V_h}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\rho''\big(\sigma(x)\xi_i\big)\big]. \tag{3.4}$$

Condition 3. *Additionally, we say that Condition 3 is satisfied if, defining*

$$\begin{aligned} s_h(\lambda) :=& \big(1 \vee 2\|K\|_\infty\big)\big[\delta_h^*(\lambda) + b_h(\mathcal{F})\big] \\ &+ 27\frac{(1 \vee \|K\|_\infty\|\rho'\|_\infty^2)}{\sqrt{n\Pi_h}}\left(\ln\big(n|\mathcal{P}|\big) + \int_0^1 H_{\mathcal{F}\times\Lambda}^{1/2}(u)\,\mathrm{d}u + H_{\mathcal{F}\times\Lambda}(1)\right), \end{aligned}$$

it holds for all $\lambda \in \Lambda$

$$s_h(\lambda) \leq \frac{1}{2\|K\|_\infty}\min\big\{\mathbb{E}P_n\lambda''\big(f^*\big), \Pi_h\mathbb{E}P_n\big[\lambda'\big(f^*\big)\big]^2\big\}. \tag{3.5}$$

We discuss the above conditions after the following result.

Theorem 2. *Let* Λ *be a set of functions as in* (2.4), $n \in \{1, 2, \ldots\}$, *and* $h \in (0, 1]^d$ *such that Conditions* 2 *and* 3 *are satisfied. Then, for all* $q \geq 1$,

$$\mathbb{E}\big|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\big|^q$$
$$\leq T_q\left(b_h(\mathcal{F}) + \left[27\int_0^1 H^{1/2}_{\mathcal{F}\times\Lambda}(u)\,\mathrm{d}u + \frac{4H_{\mathcal{F}\times\Lambda}(1)}{\ln^2(n)} + 1\right]\frac{\sqrt{\mathrm{V}(\lambda^*)}}{\sqrt{n\Pi_h}}\right)^q + \frac{5(2M)^q}{n^2}$$

for a constant T_q ($T_q = 2q|\mathcal{P}|117^q\,\mathrm{Gamma}(q)$ *works*, *where* $\mathrm{Gamma}(\cdot)$ *is the classical Gamma function*).

Remark 3. Apart from the given assumptions, the estimator $\hat{f}_{\hat{\lambda}}(x_0)$ does not premise knowledge about the noise level $\sigma(\cdot)$ and the densities $g_i(\cdot)$ and $\mu(\cdot)$ but achieves – up to constants – the optimal variance term $\mathrm{V}(\lambda^*)$ for all such functions. The estimator is thus called D-adaptive optimal (with respect to the set Λ). For example, for the Huber contrast (2.2) indexed by the scale γ, $\Upsilon := \{\rho_{\mathrm{H},\gamma}(\cdot) : \gamma > 0\}$, the estimator is D-adaptive minimax (Huber minimax) for the set of contamination models, see Section A.1 in the arXiv version. Finally, we mention that appropriate choices of the bandwidth h in the above result lead to S-minimax results.

Discussion of Conditions 2 and 3. *Condition* 2 *limits the possible sets of combined functions* Λ *and thus, in particular, the sets of possible contrast functions* Υ. *It demands that all possible combined functions* $\lambda \in \Lambda$ *fulfill Condition* 1, *which then leads to consistent estimators* (*see Proposition* 1) *and to sets of contrast with finite entropy. Condition* 2 *demands, in particular, that the right-hand side of* (3.4) *is positive and, since the right-hand side of* (3.4) *is upper bounded by* 1, *that* $\sup_{\rho\in\Upsilon}\|\rho'\|_\infty$ *does not increase too rapidly with* n.

In the following, we illustrate these restrictions with an example. We consider a homoscedastic model ($\sigma(\cdot) \equiv \sigma > 0$) *and* Υ *equal to a set of Huber contrasts* $\rho_{\mathrm{H},\gamma}(\cdot)$ *as in* (2.2) *with scale parameter* $\gamma \in [\gamma_-, \gamma^+]$, $\gamma^+ \geq \gamma_- > 0$. *It holds that* $\sup_{\rho\in\Upsilon}\|\rho'_{\mathrm{H},\gamma}\|_\infty = \gamma^+$. *This implies that* γ^+ *must not increase too rapidly with* n. *Moreover, it must hold that*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[\rho''_{\mathrm{H},\gamma}(\sigma\xi_i)\big] = \int_{-\gamma/\sigma}^{\gamma/\sigma}\frac{1}{n}\sum_{i=1}^n g_i(z)\,\mathrm{d}z \geq \int_{-\gamma_-/\sigma}^{\gamma_-/\sigma}\frac{1}{n}\sum_{i=1}^n g_i(z)\,\mathrm{d}z > 0.$$

For noise densities that are positive and continuous in the origin, this condition is verified for all $\gamma_- > 0$. *For more involved noise densities* (*vanished at the origin*), *however,* γ_- *has to be chosen sufficiently large.*

Condition 3 *is similar to Condition* 2 *since* $s_h(\lambda) \asymp \delta^*_h(\lambda)$. *However, the terms in the minimum on the right-hand side of* (3.5), *can be small for a certain design and noise level. The second term, vanishes if* $\sigma(\cdot) \equiv 0$ *since* $\rho'(0) = 0$. *Moreover, if the design degenerates* (*as in* (2.16)) *with a large* s, $\mathbb{E}P_n\lambda''(f^*)$ *and* $\Pi_h\mathbb{E}P_n[\lambda'(f^*)]^2$ *then tend to zero faster than* $s_h(\lambda)$ *as* $n \to \infty$ (*cf.* (2.10) *and* (2.11)). *This is due to the estimation of* $\mathbb{E}P_n\lambda''(f^*)$ *and* $\Pi_h\mathbb{E}P_n[\lambda'(f^*)]^2$: *if* $\sigma(\cdot) \equiv 0$ *or if the design degenerates, the above terms are small* (*cf.* (2.9)), *and thus, the estimation error of them* (*which is related to* $s_h(\cdot)$) *obstructs their behavior.*

4. 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 a modification of the data-driven Lepski's method to select the bandwidth. In the first part, we consider isotropic, locally polynomial target functions, in the second part anisotropic, locally constant functions. To simplify the exposition, we present asymptotic results only.

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

Definition 4. *Let $\vec{\beta} := (\beta_1, \dots, \beta_d) \in]0, +\infty[^d$ such that $\lfloor \beta_1 \rfloor = \cdots = \lfloor \beta_d \rfloor =: \lfloor \beta \rfloor$, 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 $x, x_0 \in [0,1]^d$*

$$\bigl|s(x) - \mathrm{P}(s)(x - x_0)\bigr| \leq L \sum_{j=1}^{d} |x_j - x_{0,j}|^{\beta_j} \quad \textit{and}$$

$$\sum_{p \in \mathcal{S}_{\lfloor \beta \rfloor}} \sup_{x \in [0,1]^d} \left| \frac{\partial^{|p|} s(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right| \leq M,$$

where $\mathrm{P}(s)(x - x_0)$ 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.

The parameter $\vec{\beta}$ is usually unknown; thus, it is desirable to have an estimator that is adaptive with respect to $\vec{\beta}$. This motivates the following definition, where $\Psi := \{\psi_n(\vec{\beta})\}_{\vec{\beta} \in \mathcal{M}}$ is a given family of normalizations for a set of parameters $\mathcal{M}$:

Definition 5. *The family Ψ is called admissible if there exists an estimator $\hat{f}_n$ such that*

$$\limsup_{n \to \infty} \sup_{\vec{\beta} \in \mathcal{M}} \psi_n^{-q}(\vec{\beta}) R_{n,q}\bigl(\hat{f}_n, \mathbb{H}_d(\vec{\beta}, L, M)\bigr) < \infty.$$

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

We distinguish two cases in the following: First, we consider the special case of isotropic Hölder classes, that is, $\beta_1 = \cdots = \beta_d$. These classes only require a common bandwidth for all dimensions that is chosen with the standard version of Lepski's method (see [24] and [23]). 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 in this case, because it requires a monotonous bias. We circumvent this problem using a modified version of Lepski's method as described in [19] and [22].

4.1. A fully adaptive estimator for isotropic, locally polynomial functions

Here, we consider isotropic Hölder classes with $\beta \in (0, m+1]$, where m is the degree of the estimator $\hat{f}_\lambda$ and may be chosen arbitrarily large. Therefore, only one bandwidth $h_{\text{iso}} = h_1 = \cdots = h_d > 0$ 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. In the following, we assume that the variance term $\mathrm{V}(\lambda_{h_{\text{iso}}})/(nh_{\text{iso}}^d)$ for (see Definitions (2.4) and (2.9))

$$\lambda_{h_{\text{iso}}}(f)(x,y) := \rho\big(y - f(x)\big) K_{h_{\text{iso}}}(x) \qquad \text{for all } x \in [0,1]^d, y \in \mathbb{R},$$

is *decreasing* in the bandwidth so that we can apply Lepski's method. This imposes an additional restriction on the design and the noise. After the main result of this section, we give some examples for designs and noises that fulfill this restriction. Next, we introduce the set of bandwidths $\mathcal{H}^{\text{iso}} := [h_-, h^+]$, where $0 < h_- < h^+ < 1$ are defined as (cf. (2.15))

$$h_- := \frac{\ln^{6/d}(n)}{n^{1/d}} \quad \text{and} \quad h^+ := \frac{1}{\ln(n)}. \tag{4.1}$$

Since the inequality $h_- < h^+$ has to be satisfied, n is required to be large enough. We then introduce the isotropic M-estimator for any $h_{\text{iso}} \in \mathcal{H}^{\text{iso}}$ as

$$\hat{f}_{\text{iso}}^{h_{\text{iso}}} := \arg\min_{f \in \mathcal{F}} P_n \widehat{\lambda}_{h_{\text{iso}}}(f),$$

where

$$\widehat{\lambda}_{h_{\text{iso}}} = \arg\min_{\lambda_{h_{\text{iso}}} \in \Lambda} \widehat{\mathrm{V}}(\lambda_{h_{\text{iso}}})$$

and $\widehat{\mathrm{V}}(\cdot)$ is defined in (2.9). Eventually, we introduce a net $\mathcal{H}_\epsilon^{\text{iso}} := \{h_{\text{iso}} \in \mathcal{H}^{\text{iso}}, \exists m \in \mathbb{N} : h_{\text{iso}} = h^+\epsilon^m\}$, $\epsilon \in (0,1)$, such that $1 \le |\mathcal{H}_\epsilon^{\text{iso}}| \le n$ and then apply Lepski's method for isotropic functions (see [24] and [23]) to define the data-driven bandwidth $\hat{h}_{\text{iso}}$:

$$\begin{aligned} \hat{h}_{\text{iso}} := \max\Bigg\{ & h_{\text{iso}} \in \mathcal{H}_\epsilon^{\text{iso}} : \big|\hat{f}_{\text{iso}}^{h_{\text{iso}}}(x_0) - \hat{f}_{\text{iso}}^{h'_{\text{iso}}}(x_0)\big| \le 15\sqrt{2}\big(B + \text{iso}_\epsilon(n)\big)\sqrt{\frac{\widehat{\mathrm{V}}(\widehat{\lambda}_{h'_{\text{iso}}})}{n(h'_{\text{iso}})^d}}, \\ & \text{for all } h'_{\text{iso}} \in \mathcal{H}_\epsilon^{\text{iso}} \text{ such that } h'_{\text{iso}} \le h_{\text{iso}} \Bigg\}, \end{aligned} \tag{4.2}$$

where $\text{iso}_\epsilon(n) := 11\sqrt{\ln(n|\mathcal{H}_\epsilon^{\text{iso}}|)}$ and $B := 27\int_0^1 H_{\mathcal{F}\times\Lambda}^{1/2}(u)\,\mathrm{d}u + \frac{4H_{\mathcal{F}\times\Lambda}(1)}{\ln^2(n)}$.

We now obtain on isotropic Hölder classes $\mathbb{H}_d^{\text{iso}}(\beta, L, M) := \mathbb{H}_d((\beta, \ldots, \beta), L, M)$, for all $\beta, L, M > 0$ the following result:

Theorem 3. *Let Λ be a set of combined functions as in* (2.4) *and $n \in \{1, 2, \dots\}$ such that Conditions* 2 *and* 3 *are satisfied for all $h_{\text{iso}} \in \mathcal{H}^{\text{iso}}$. Then, for any $x_0 \in (0,1)^d$, any $\beta \in (0, m+1]$, and any $L > 0$, there exists a universal positive constant $C > 0$ such that*

$$R_{n,q}\big[\hat{f}_{\text{iso}}^{\hat{h}_{\text{iso}}}(x_0), \mathbb{H}_d^{\text{iso}}(\beta, L, M)\big] \leq C \inf_{h_{\text{iso}} \in \mathcal{H}^{\text{iso}}} \left\{ Ldh_{\text{iso}}^{\beta} + \text{iso}_{\epsilon}(n) \sqrt{\frac{\mathrm{V}(\lambda^*_{h_{\text{iso}}})}{nh_{\text{iso}}^d}} \right\}^q \qquad \textit{as } n \to \infty.$$

Remark 4. This oracle inequality like result shows the simultaneous S- and D-adaptation of the estimator. It generalizes results in [5], which rely on the asymptotic equivalence of the block median method, in two important aspects: First, it allows for heteroscedastic regression models with random designs. Second, it does not require that the noise densities are positive at their median and thus allows for a wider range densities. Finally, we note that Lepski's method has been used for locally constant M-estimators in [27] but – to the best of our knowledge – never to locally polynomial M-estimators as it is done here.

Remark 5. If only S-adaptation is considered, the conditions on n can be considerably relaxed. In fact, assuming that $\mathrm{V}(\lambda_.)$ is known, the estimator $\hat{f}_{\lambda_{\tilde{h}_{\text{iso}}}}$ of (2.4) can be applied instead of $\hat{f}_{\text{iso}}^{\hat{h}_{\text{iso}}}$, where $\tilde{h}_{\text{iso}}$ is selected from (4.2) replacing $\widehat{\mathrm{V}}(\widehat{\lambda}_{h'_{\text{iso}}})$ by $\mathrm{V}(\lambda_{h'_{\text{iso}}})$. The Conditions 2 and 3 can then be replaced by Condition 1.

Remark 6. The variance term is decreasing for settings with indicator kernels and homoscedastic noise levels (as one can check easily starting from (2.9)); for settings with indicator kernels, Huber contrasts, $\sigma(\cdot) = 1 + |\cdot - x_0|^{\alpha}$ for a $\alpha \in [0, \leq 1/2]$, and $d = 1$; and for many other settings. On the contrary, the variance term can be increasing, for example, if the noise level is symmetric in x_0 and convex.

Corollary 4. *Consider the model in Example* 1 *in the previous section with $\mu(\cdot) \equiv 1$ (uniform design) and $\sigma(\cdot) \equiv 1$ (homoscdastic noise level). For any $\beta \in (0, m+1]$ and any $L > 0$, it holds that*

$$\limsup_{n \to \infty} \left(\frac{n}{\ln(n)} \right)^{q\beta/(2\beta+d)} R_{n,q}\big[\hat{f}_{\text{iso}}^{\hat{h}_{\text{iso}}}(x_0), \mathbb{H}_d^{\text{iso}}(\beta, L, M)\big] < \infty.$$

This corollary can be deduced minimizing the term on the right-hand side of the last theorem with a standard bias/variance trade-off.

Remark 7. The rate $(\ln(n)/n)^{\beta/(2\beta+1)}$ in the above corollary is admissible (cf. Definition 5) over isotropic Hölder spaces and is asymptotically optimal (see [4] and [24]) up to the logarithm $\ln(n)$, which is the usual price for the adapativity (see Section 5 for more details). Moreover, the approach used to deduce the above corollary presumes uniform designs and homescedastic noises; however, more elaborate approaches, perhaps similar to the ones in [8], may lead to comparable results for degenerate designs.

4.2. A fully adaptive estimator for anisotropic, locally constant functions

In this part, we allow for anisotropic Hölder classes and bandwidths. In return, we restrict ourselves to locally constant functions, that is, $m = 0$ (and thus $|\mathcal{P}| = 1$) and $\mathcal{F} = [-M, M]$. Moreover, we restrict ourselves to uniform designs ($\mu(\cdot) \equiv 1$) and homoscedastic ($\sigma(\cdot) \equiv \sigma \geq 0$) and identically distributed noise ($g_i(\cdot) \equiv g(\cdot)$ for all $i = 1, \dots, n$). For this setting, we introduce an S- and D-adaptive estimator of $f^*(x_0)$. The main properties of this estimator are given in Theorem 4.

We introduce an estimator for each bandwidth in the set $\mathcal{H} := [h_-, h^+]^d$, where h_- and h^+ are defined in the previous section. For this, we define the variance term as

$$\mathrm{V}(\rho, K) := \left(\frac{\sqrt{\int [\rho'(\sigma z)]^2 g(z)\, \mathrm{d}z} + \|\rho'\|_\infty \|K\|_\infty \ln^2(n)/\sqrt{nh_-^d}}{\int \rho''(\sigma z) g(z)\, \mathrm{d}z} \right)^2, \tag{4.3}$$

and the oracle for a set of contrasts Υ and a set of kernels $\mathcal{K}$ as

$$\left(\rho^*, K^*\right) := \arg \min_{\rho \in \Upsilon, K \in \mathcal{K}} \mathrm{V}(\rho, K). \tag{4.4}$$

Next, we introduce an estimator of the variance term as

$$\widehat{\mathrm{V}}(\rho, K) := \left(\frac{\sqrt{(1/n) \sum_{i=1}^n [\rho'(Y_i - \hat{f}_{\lambda_{h^+}}(X_i))]^2} + \|\rho'\|_\infty \|K\|_\infty \ln^2(n)/\sqrt{nh_-^d}}{(1/n) \sum_{i=1}^n \rho''(Y_i - \hat{f}_{\lambda_{h^+}}(X_i))} \right)^2, \tag{4.5}$$

where $\hat{f}_{\lambda_{h^+}}$ is defined in (2.4) with $\lambda = \lambda_{h^+}(f)(x, y) := \rho(y - f(x)) K_{h^+}(x)$, and an estimator of the oracle as

$$(\hat{\rho}, \hat{K}) := \arg \min_{\rho \in \Upsilon, K \in \mathcal{K}} \widehat{\mathrm{V}}(\rho, K). \tag{4.6}$$

We stress that the variance term V, the oracle (ρ^*, K^*), and their estimators $\widehat{V}$ and $(\hat{\rho}, \hat{K})$ are independent of the bandwidth. We can finally introduce the desired estimator $\hat{f}^h$ for all $h \in \mathcal{H}$:

$$\hat{f}^h := \arg \min_{f \in \mathcal{F}} n^{-1} \sum_i \hat{\rho}\big(Y_i - f(X_i)\big) \hat{K}_h(X_i). \tag{4.7}$$

The crucial step is now the choice of the bandwidth with a modified version of Lepski's method (see [19] and [20]). First, we define for all $a, b \in \mathbb{R}$ the scalar $a \vee b := \max(a, b)$ and for all $h, h' \in \mathcal{H}$ the vector $h \vee h' := (h_1 \vee h'_1, \dots, h_d \vee h'_d)$. We then consider the two families of Locally Constant Approximation (LCA) estimators (provoked by (4.7))

$$\{\hat{f}^h\}_{h \in \mathcal{H}} \quad \text{and} \quad \{\hat{f}^{h,h'} := \hat{f}^{h \vee h'}\}_{h,h' \in \mathcal{H}^2}.$$

Note that $\hat{f}^{h,h'} = \hat{f}^{h',h}$ (commutativity). Similarly as above, we then introduce a net $\mathcal{H}_\epsilon := \{(h_-, \dots, h_-)\} \cup \{h \in \mathcal{H} : \forall j = 1, \dots, d\ \exists m_j \in \mathbb{N} : h_j = h^+ \epsilon^{m_j}\}$, $\epsilon \in (0, 1)$, such that $|\mathcal{H}_\epsilon| \leq n$

and set $\mathrm{ani}_\epsilon(n) := 11\sqrt{\ln(n|\mathcal{H}_\epsilon|)}$. We finally select the bandwidth according to

$$\hat{h} := \max_{\preceq}\Bigg\{h \in \mathcal{H}_\epsilon : \big|\hat{f}^{h,h'}(x_0) - \hat{f}^{h'}(x_0)\big| \le 16\big(B + \mathrm{ani}_\epsilon(n)\big)\sqrt{\frac{\widehat{\mathrm{V}}(\hat{\rho}, \hat{K})}{n\Pi_{h'}}} \\ \text{for all } h' \in \mathcal{H}_\epsilon \text{ such that } h' \preceq h\Bigg\}. \tag{4.8}$$

The maximum is taken with respect to the order $\preceq$ which we define as $h \preceq h' \Leftrightarrow \prod_{j=1}^d h_j \le \prod_{j=1}^d h'_j$. Note, in particular, that the right-hand side of (4.8) is decreasing with respect to this order.

The above choice of the bandwidth leads to the estimator $\hat{f}^{\hat{h}}$ with the following properties:

Theorem 4. *Let Λ be a set of combined functions as in (2.4) and let $n \in \{1, 2, \ldots\}$ such that Conditions 2 and 3 are satisfied for all $h \in \mathcal{H}$. Then, for any $x_0 \in (0, 1)^d$, any $\vec{\beta} \in (0, 1]^d$, and any $L > 0$, there exists a universal constant C such that*

$$R_{n,q}\big[\hat{f}^{\hat{h}}(x_0), \mathbb{H}_d(\vec{\beta}, L, M)\big] \le C \inf_{h\in\mathcal{H}}\Bigg\{L\sum_{j=1}^d h_j^{\beta_j} + \mathrm{ani}_\epsilon(n)\sqrt{\frac{\mathrm{V}(\rho^*, K^*)}{n\Pi_h}}\Bigg\}^q.$$

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

Corollary 5. *For any $\vec{\beta} \in (0, 1]^d$ and any $L > 0$, it holds that*

$$\limsup_{n\to\infty}\left(\frac{n}{\ln(n)}\right)^{q\bar{\beta}/(2\bar{\beta}+1)} R_{n,q}\big[\hat{f}^{\hat{h}}(x_0), \mathbb{H}_d(\vec{\beta}, L, M)\big] < \infty,$$

where $\bar{\beta} := (\sum_j 1/\beta_j)^{-1}$ is the harmonic average.

Remark 8. In contrast to the previous part, only locally constant functions are considered here, which is due to the bias term (cf. Lemma 6). To the best of our knowledge, the presented choice of the bandwidth is the first application of the *anisotropic* Lepski's principle ([22], see also [11,19, 20]) for the selection of an anisotropic bandwidth for nonlinear M-estimators. We also note that, comparing the adaptive rate $(\ln(n)/n)^{\bar{\beta}/(2\bar{\beta}+1)}$ with the optimal rate in the white noise model (see [20]), for example, one finds that this rate is nearly optimal. We finally refer to the remarks after Theorem 3.

5. Discussion

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

1. Instead of assuming that the densities $g_i(\cdot)$ are symmetric (cf. [14,29]), it is sufficient that the sum $\sum_i g_i(\cdot)$ is symmetric. We are, however, not aware of examples where this generalization is relevant.

2. The variance of the median estimator is $1/(4g^2(0))$ which implies a strong sensitive to the noise density at 0. Moreover, the estimation of $g(0)$ (see [5], e.g.) requires many observations near $f(x_0)$ in practice. On the contrary, Huber contrast with scale γ (allowed by our approach), the denominator of the variance term (2.9) depends on the mass of the noise density on the interval $[-\gamma, \gamma]$ instead of the mass at 0.
3. To estimate the variance term (2.9), we plug an estimate $Y_i - \hat{f}_\lambda$ of the residuals. Condition 2 ensures the consistency of all estimators in Λ. This is considerably restrictive on the initial family Λ. This problem can be circumvented using a pre-estimator (e.g., with the contrast (2.3)) instead of $\hat{f}_\lambda$ for the estimation of the variance.
4. Lepski's method is very sensitive to outliers (see [27]). To complement it with the adaptive robustness of the estimator via the minimization of the variance term can thus be interesting for many applications.
5. The variance term and its empirical version do not depend on the bias term (see Theorem 1, Definition (3.2) and Remark 2) and, more generally, not on the specific model. The procedure presented in this paper may thus be interesting for other models, such as high dimensional settings (cf. [21]), for example.
6. The quantity $15\sqrt{2}(B + \text{iso}_\epsilon(n))$ in the threshold term in (4.2) contains the factor $\ln(n)$ and known but large constants. For applications, it should usually be chosen considerably smaller (see [23]) and can probably be tuned with the *propagation* method [28], for example.
7. As mentioned in the Introduction, Lepski-type procedures are also useful to get S-adaptive confident bands (see [10] and references therein). This requires deviation inequalities that can be derived along the presented lines (see Proposition 3) but also a lower bound for the bias term of the estimator (cf. [10], Condition 3, Section 3.2 and Section 3.5 for Discussion), which seems not to be available here, since robust M-estimators – and thus the bias term – do not have explicit expressions. For our purposes, we circumvent this issue by using the bias term of the criterions's derivative as an estimator of the expected criterion's derivative, see Lemmas 1 and 2. However, this way, we only obtain an upper bound. We therefore suggest to establish first S-adaptive confidence bands for the criterion's derivative viewed as an estimator and then, using the smoothness of the contrast, confidence bands with respect to a pointwise semi-norm or sup-norm.

6. Proofs of the main results

Let us introduce some additional notation to simplify the exposition. For this, we introduce

$$\mathcal{F}_\delta := \{ f = \mathrm{P}_t \in \mathcal{F} : \| t - t^0 \|_{\ell_1} \leq \delta \} \tag{6.1}$$

as a ball in $\mathcal{F}$ with radius $\delta > 0$ centered at f^0. Furthermore, we denote the column vector of partial derivatives of the criterion $P_n\lambda(\cdot)$ (defined in (2.4)) by

$$\tilde{D}_\lambda(\mathrm{P}_t) := \left(-\frac{\partial}{\partial t_p} P_n \lambda(\mathrm{P}_t) \right)_{p \in \mathcal{P}} \qquad \text{for all } t \in \mathbb{R}^{|\mathcal{P}|}, \tag{6.2}$$

and the "parametric" expectation with respect to the distribution $\mathbb{E}^0$ of $(X, f^0(X)+\sigma(X)\xi)$ by

$$\mathbb{E}^0\big[\tilde{D}_\lambda(\cdot)\big]. \tag{6.3}$$

Next, for all $t\in\mathbb{R}^{|\mathcal{P}|}$, we introduce the *Jacobian matrix* J_D of $\mathbb{E}^0[\tilde{D}_\lambda]$ as

$$\big(J_D(\mathrm{P}_t)\big)_{p,q\in\mathcal{P}} := \left(\frac{\partial}{\partial t_q}\mathbb{E}^0\big[\tilde{D}^p_\lambda(\mathrm{P}_t)\big]\right)_{p,q\in\mathcal{P}} = \left(\frac{\partial}{\partial t_q}\mathbb{E}^0\left[-\frac{\partial}{\partial t_p}P_n\lambda(\mathrm{P}_t)\right]\right)_{p,q\in\mathcal{P}}, \tag{6.4}$$

where $\tilde{D}^p_\lambda(\cdot)$ is the pth component of $\tilde{D}_\lambda(\cdot)$. The Jacobian matrix exists according to Definition 2 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 $\hat{t}_\lambda$. Moreover, we set

$$c_\lambda := \mathbb{E}P_n\lambda''\big(f^*\big) \tag{6.5}$$

and $b_h := b_h(\mathcal{F})$. We finally define $\lambda'_\infty := \|\rho'\|_\infty\|K\|_\infty$ and for any $z\geq 0$

$$B_z := 27\int_0^1 H^{1/2}_{\mathcal{F}\times\Lambda}(u)\,\mathrm{d}u + \frac{4H_{\mathcal{F}\times\Lambda}(1)}{\ln^2(n)} + 7\sqrt{2z} + \frac{2z}{\ln^2(n)}. \tag{6.6}$$

6.1. Auxiliary results

The following propositions are basic for the proofs of the main results. The proofs of the propositions are given in the Appendix.

Proposition 1. *Let $\Lambda=\Upsilon\times\mathcal{K}$ be a set of functions as in* (2.4) *where Υ is a set of contrasts as in Definition* 2 *and $\mathcal{K}$ is a set of kernels as in Definition* 1. *Let $n\in\{1,2,\dots\}$ and $h\in(0,1]^d$ be such that Condition* 2 *is satisfied. Then, $\mathbb{P}(\bigcap_{\lambda\in\Lambda}\{\hat{f}_\lambda\in\mathcal{F}_{\delta^*_h(\lambda)}\})\geq 1-n^{-2}$, where $\delta^*_h(\cdot)$ is defined in Condition* 2.

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

Proposition 2. *For any $z\geq 0$, it holds that*

$$\mathbb{P}\left(\sup_{\lambda\in\Lambda}\sup_{f\in\mathcal{F}_{\delta^*_h(\lambda)}}\frac{\|\tilde{D}_\lambda(f)-\mathbb{E}[\tilde{D}_\lambda(f)]\|_{\ell_\infty}}{\sqrt{\Pi_h\mathbb{E}P_n[\lambda'(f^*)]^2}+\lambda'_\infty\ln^2(n)/\sqrt{n\Pi_h}}\geq\frac{B_z}{\sqrt{n\Pi_h}}\right)\leq 2|\mathcal{P}|\exp(-z),$$

where B_z is defined in (6.6).

This proposition is directly deduced from Massart's Inequality (see the arXiv version for details).

Proposition 3. *Let Λ be a set of functions as in (2.4), $n \in \{1, 2, \dots\}$, and $h \in (0, 1]^d$ be such that Condition 2 is satisfied. Then, for any $z \geq 0$, it holds that*

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

We note that the constants 2 and 3 can be replaced by o(1).

Proposition 4. *Let $\Lambda = \Upsilon \times \mathcal{K}$ be a set of functions as in (2.4) where Υ is a set of contrasts as in Definition 2 and $\mathcal{K}$ is a set of kernels as in Definition 1. Let $n \in \{1, 2, \dots\}$ and $h \in (0, 1]^d$ be such that Condition 2 is satisfied.*

Then, $\mathbb{P}(\Delta) \geq 1 - 5/n^2$, where $\Delta := \bigcap_{\lambda\in\Lambda}\{\sqrt{\widehat{\mathrm{V}}(\lambda)} \in [\frac{\sqrt{2}}{3}\sqrt{\mathrm{V}(\lambda)}, \sqrt{6}\sqrt{\mathrm{V}(\lambda)}]\}$.

We note that the constants $\sqrt{2}/3$ and $\sqrt{6}$ can be replaced by o(1).

6.2. Proof of Theorem 2

First, we set $\Delta := \bigcap_{\lambda\in\Lambda}\{\sqrt{\widehat{\mathrm{V}}(\lambda)} \in [\frac{\sqrt{2}}{3}\sqrt{\mathrm{V}(\lambda)}, \sqrt{6}\sqrt{\mathrm{V}(\lambda)}]\}$. Then, we observe that, since $\hat{f}_{\hat{\lambda}} \in \mathcal{F}$, $\sup_{f\in\mathcal{F}}|f(x_0)| \leq M$, and $|f^*(x_0)| \leq M$, the risk can be bounded by

$$\begin{aligned}\mathbb{E}\left|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\right|^q &= \mathbb{E}\left|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\right|^q\mathbb{1}_\Delta + \mathbb{E}\left|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\right|^q\mathbb{1}_{\Delta^c}\\ &\leq \mathbb{E}\left|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\right|^q\mathbb{1}_\Delta + (2M)^q\mathbb{P}(\Delta^c).\end{aligned}$$

Using Proposition 4, Lemma 3, the last inequality, and simple computations, we obtain

$$\begin{aligned}\mathbb{E}\left|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\right|^q &\leq \mathbb{E}\left|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\right|^q\mathbb{1}_\Delta + 5(2M)^q/n^2\\ &\leq 2^q\mathbb{E}\left(\left|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\right| - 3b_h - \frac{6\sqrt{3\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)_+^q\mathbb{1}_\Delta \qquad (6.7)\\ &\quad + 2^q\left(3b_h + \frac{6\sqrt{3\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)^q + 5(2M)^q/n^2.\end{aligned}$$

Let us now bound the first term on the right-hand side of the last inequality. To do so, we note that on the event Δ

$$\sqrt{\mathrm{V}(\lambda^*)} \geq \sqrt{\frac{\widehat{\mathrm{V}}(\lambda^*)}{6}} \geq \sqrt{\frac{\widehat{\mathrm{V}}(\hat{\lambda})}{6}} \geq \sqrt{\frac{\mathrm{V}(\hat{\lambda})}{27}}. \qquad (6.8)$$

Using the last inequality and integrating the result of Proposition 3 with $\varepsilon = 10\sqrt{z} + \frac{2z}{\ln^2(n)}$, we get (for more details see the arXiv version)

$$\mathbb{E}\left(\left|\hat{f}_{\hat{\lambda}}(x_0) - f^*(x_0)\right| - 3b_h - \frac{6\sqrt{3\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)_+^q\mathbb{1}_\Delta \leq T_q\left(b_h + \frac{\sqrt{\mathrm{V}(\lambda^*)}B_0}{\sqrt{n\Pi_h}}\right)^q.$$

From (6.7) and the last inequality, the theorem can be deduced.

6.3. Proof of Theorem 3

For ease of exposition, we set $B_0 = B$ (cf. (6.6)), $k := h_{\mathrm{iso}}$, and $\hat{k} := \hat{h}_{\mathrm{iso}}$. Then, one may verify that the *oracle bandwidth*

$$k^* := \arg\min_{k\in\mathcal{H}^{\mathrm{iso}}} \left\{ Ldk^\beta + c\big(B_0 + \mathrm{iso}_\epsilon(n)\big)\sqrt{\frac{\mathrm{V}(\lambda_k^*)}{nk^d}} \right\}$$

is well defined, where c is a constant chosen such that both terms are equal at the point k^*. Next, from Propositions 1 and 4 with $h = (k, \dots, k)$, it follows that

$$\mathbb{P}\big(\exists k \in \mathcal{H}_\epsilon^{\mathrm{iso}}, \exists \lambda_k \in \Lambda : \hat{f}_{\mathrm{iso}}^k \notin \mathcal{F}_{\delta_k^*(\lambda_k)}\big) \le \sum_{k\in\mathcal{H}_\epsilon^{\mathrm{iso}}} n^{-2} \le n^{-1} \tag{6.9}$$

and

$$\sum_{k\in\mathcal{H}_\epsilon^{\mathrm{iso}}} \mathbb{P}\big(\Delta_k^c\big) \le \sum_{k\in\mathcal{H}_\epsilon^{\mathrm{iso}}} \frac{5}{n^2} \le 5n^{-1}, \tag{6.10}$$

where $\Delta_k := \Delta$ is defined in Proposition 4. Thus, we may restrict our considerations to the event $\bigcap_{k\in\mathcal{H}_\epsilon^{\mathrm{iso}},\lambda_k\in\Lambda}\{\hat{f}_{\mathrm{iso}}^k \in \mathcal{F}_{\delta_k^*(\lambda_k)}\} \cap \Delta_k$, since we are only interested in the asymptotic behavior. We now introduce $k_\epsilon^* \in \mathcal{H}_\epsilon^{\mathrm{iso}}$ such that $k_\epsilon^* \le k^* \le \epsilon^{-1}k_\epsilon^*$.

Control of the risk on the event $\{k_\epsilon^* \le \hat{k}\}$

With the triangular inequality and Lemma 3, we obtain

$$\begin{aligned} &\big|\hat{f}_{\mathrm{iso}}^{\hat{k}}(x_0) - f^*(x_0)\big|^q \mathbb{1}_{k_\epsilon^*\le\hat{k}} \\ &\quad\le 2^{q-1}\big(\big|\hat{f}_{\mathrm{iso}}^{\hat{k}}(x_0) - \hat{f}_{\mathrm{iso}}^{k_\epsilon^*}(x_0)\big|^q \mathbb{1}_{k_\epsilon^*\le\hat{k}} + \big|\hat{f}_{\mathrm{iso}}^{k_\epsilon^*}(x_0) - f^*(x_0)\big|^q\big). \end{aligned} \tag{6.11}$$

The first term on the right-hand side of the last inequality is controlled using the procedure (4.2) to obtain

$$\mathbb{E}\big[\big|\hat{f}_{\mathrm{iso}}^{\hat{k}}(x_0) - \hat{f}_{\mathrm{iso}}^{k_\epsilon^*}(x_0)\big|^q \mathbb{1}_{k_\epsilon^*\le\hat{k}}\big] \le \mathbb{E}\left[15\sqrt{2}\frac{\sqrt{\widehat{\mathrm{V}}(\widehat{\lambda}_{k_\epsilon^*})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k_\epsilon^*)^d}}\right]^q.$$

On the event $\bigcap_{k\in\mathcal{H}_\epsilon^{\mathrm{iso}}} \Delta_k$, we get similarly as in (6.8)

$$\mathbb{E}\big[\big|\hat{f}_{\mathrm{iso}}^{\hat{k}}(x_0) - \hat{f}_{\mathrm{iso}}^{k_\epsilon^*}(x_0)\big|^q \mathbb{1}_{k_\epsilon^*\le\hat{k}}\big] \le \left(45\sqrt{6}\frac{\sqrt{\mathrm{V}(\lambda_{k_\epsilon^*}^*)}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k_\epsilon^*)^d}}\right)^q.$$

Recall that, by the definitions of the Hölder classes (Definition 4), we can control the bias for any $\beta \in (0, m+1]$ and any $k > 0$ by

$$b_k \leq \sup_{x \in V_k} \left| \mathrm{P}\big(f^*\big)(x - x_0) - f^*(x) \right| \leq L d k^{\beta}, \tag{6.12}$$

where $\mathrm{P}(f^*)(x - x_0)$ is the Taylor Polynomial of f^* at x_0. So we can finally deduce from Theorem 2 with $h = (k, \dots, k)$ and $b_h = b_k$ a bound for the second term in (6.11) for n sufficiently large:

$$\mathbb{E}\big| \hat{f}_{\mathrm{iso}}^{k_\epsilon^*}(x_0) - f^*(x_0) \big|^q \leq \mathcal{C}_1 \left(Ld\big(k_\epsilon^*\big)^{\beta} + \sqrt{\frac{\mathrm{V}(\lambda_{k_\epsilon^*}^*)}{n(k_\epsilon^*)^d}} \right)^q,$$

where $\mathcal{C}_1$ is a universal constant. Using (6.11) and the above inequalities, we have a control of the risk on the event $\{k_\epsilon^* \leq \hat{k}\}$:

$$\mathbb{E}\big[\big| \hat{f}_{\mathrm{iso}}^{\hat{k}}(x_0) - f^*(x_0) \big|^q \mathbb{1}_{k_\epsilon^* \leq \hat{k}} \big] \leq \mathcal{C}_2 \left(Ld\big(k_\epsilon^*\big)^{\beta} + \big(B_0 + \mathrm{iso}_\epsilon(n)\big) \sqrt{\frac{\mathrm{V}(\lambda_{k_\epsilon^*}^*)}{n(k_\epsilon^*)^d}} \right)^q, \tag{6.13}$$

where $\mathcal{C}_1$ is also a universal constant.

Control of the risk on the event $\{k_\epsilon^* > \hat{k}\}$

In order to control the risk on the complementary event, we observe that

$$\mathbb{E}\big[\big| \hat{f}_{\mathrm{iso}}^{\hat{k}}(x_0) - f^*(x_0) \big|^q \mathbb{1}_{k_\epsilon^* > \hat{k}} \big] \leq (2M)^q \mathbb{P}\big(k_\epsilon^* > \hat{k}\big). \tag{6.14}$$

We now show that the probability $\mathbb{P}(k_\epsilon^* > \hat{k})$ is small. According to the procedure (4.2), we have

$$\begin{aligned} \mathbb{P}\big(k_\epsilon^* > \hat{k}\big) &\leq \mathbb{P}\left(\exists k' \in \mathcal{H}, k' < k_\epsilon^* : \big| \hat{f}_{\mathrm{iso}}^{k_\epsilon^*}(x_0) - \hat{f}_{\mathrm{iso}}^{k'}(x_0) \big| > 15\sqrt{2} \frac{\sqrt{\widehat{\mathrm{V}}(\widehat{\lambda}_{k'})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}} \right) \\ &\leq 2 \sum_{k' \in \mathcal{H}_\epsilon^{\mathrm{iso}} : k' \leq k_\epsilon^*} \mathbb{P}\left(\big| \hat{f}_{\mathrm{iso}}^{k'}(x_0) - f^*(x_0) \big| > \frac{15}{\sqrt{2}} \frac{\sqrt{\widehat{\mathrm{V}}(\widehat{\lambda}_{k'})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}} \right). \end{aligned}$$

On the event $\bigcap_{k \in \mathcal{H}_\epsilon^{\mathrm{iso}}} \Delta_k$, we get similarly as in (6.8)

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

Consequently,

$$\mathbb{P}\big(k_\epsilon^* > \hat{k}\big) \leq 2 \sum_{k' \in \mathcal{H}_\epsilon^{\mathrm{iso}} : k' \leq k_\epsilon^*} \mathbb{P}\left(\big| \hat{f}_{\mathrm{iso}}^{k'}(x_0) - f^*(x_0) \big| > 5 \frac{\sqrt{\mathrm{V}(\widehat{\lambda}_{k'})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}} \right). \tag{6.15}$$

By definition, the oracle bandwidth k^* is the one which gives the best trade-off. Thus, that the variance is decreasing, we obtain for all $k' \le k^*_\epsilon \le k^*$

$$\begin{aligned} Ld\big(k'\big)^\beta \le Ld\big(k^*_\epsilon\big)^\beta \le Ld\big(k^*\big)^\beta &= \frac{\sqrt{\mathrm{V}(\lambda^*_{k^*})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k^*)^d}} \\ &\le \frac{\sqrt{\mathrm{V}(\lambda^*_{k^*_\epsilon})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k^*_\epsilon)^d}} \le \frac{\sqrt{\mathrm{V}(\lambda^*_{k'})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}} \le \frac{\sqrt{\mathrm{V}(\widehat{\lambda}_{k'})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}}. \end{aligned}$$

From (6.12), (6.15), and the last inequality, we get

$$\begin{aligned} \mathbb{P}\big(k^*_\epsilon > \hat{k}\big) &\le 2 \sum_{k' \in \mathcal{H}^{\mathrm{iso}}_\epsilon : k' \le k^*_\epsilon} \mathbb{P}\left(\big|\hat{f}^{k'}_{\mathrm{iso}}(x_0) - f^*(x_0)\big| > 2\frac{\sqrt{\mathrm{V}(\widehat{\lambda}_{k'})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}} + 3b_{k'}\right) \\ &\le 2 \sum_{k' \in \mathcal{H}^{\mathrm{iso}}_\epsilon : k' \le k^*_\epsilon} \mathbb{P}\left(\sup_{\lambda_{k'} \in \Lambda}\left[\big|\hat{f}^{k'}_{\mathrm{iso}}(x_0) - f^*(x_0)\big| - 2\frac{\sqrt{\mathrm{V}(\lambda_{k'})}(B_0 + \mathrm{iso}_\epsilon(n))}{\sqrt{n(k')^d}}\right] > 3b_{k'}\right). \end{aligned}$$

Since $\mathrm{iso}_\epsilon(n)/\ln^2(n) \le 1$ for n sufficiently large, using the definition of $\mathrm{iso}_\epsilon(n)$, Proposition 3 with $h = (k', \dots, k')$, $\lambda = \lambda_{k'}$, and z such that $B_z = (B_0 + \mathrm{iso}_\epsilon(n))$, we obtain

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

Then, in view of the last inequality, (6.9), (6.10), (6.13) and (6.14), we conclude that

$$\mathbb{E}\big|\hat{f}^{\hat{h}}(x_0) - f^*(x_0)\big|^q \le \mathcal{C}_2\left(Ld\big(k^*_\epsilon\big)^\beta + \big(B_0 + \mathrm{iso}_\epsilon(n)\big)\sqrt{\frac{\mathrm{V}(\lambda^*_{k^*_\epsilon})}{n(k^*_\epsilon)^d}}\right)^q \qquad \text{as } n \to \infty.$$

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

6.4. Proof of Theorem 4

We set $B = B_0$. One may then 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. Define now the element h^*_ϵ of $\mathcal{H}_\epsilon$ such that for all $j = 1, \dots, d$, $h^*_{\epsilon,j} \le h^*_j \le \epsilon^{-1} h^*_{\epsilon,j}$. 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

$$\mathcal{D}_h(\cdot) := \mathbb{E}\big[\tilde{D}_{\tilde{\lambda}_h}(\cdot)\big] = \int \hat{K}_h(x) \int \hat{\rho}'\big(\sigma z + f^*(x) - \cdot\big) g(z)\,\mathrm{d}z\,\mathrm{d}x. \tag{6.16}$$

Here, $\tilde{\lambda}_h(f)(x,y) := \hat{\rho}(y - f(x))\hat{K}_h(x)$ and $(\hat{\rho}, \hat{K})$ and $\tilde{D}_\lambda(\cdot)$ are defined in (4.6) and (6.2), respectively. Next, for uniform designs and homoscedastic noise levels, the quantity c_{λ_h}

$$c_{\lambda_h} = c_\rho := \int \rho''(\sigma z) g(z)\,\mathrm{d}z, \tag{6.17}$$

simplifies for any λ_h and does not depend on h. Moreover, according to Lemma 4, we have for any $h \in \mathcal{H}$, any $\lambda \in \Lambda$, and any two constant functions $f, \tilde{f} \in \mathcal{F}_{\delta_h^*(\lambda)}$

$$|f - \tilde{f}| \le \tfrac{4}{3} c_{\hat{\rho}}^{-1} \big|\mathcal{D}_h(f) - \mathcal{D}_h(\tilde{f})\big|. \tag{6.18}$$

Furthermore, from Propositions 1 and 4, it follows that

$$\mathbb{P}\big(\exists h \in \mathcal{H}_\epsilon, \exists \lambda_h \in \Lambda : \hat{f}^h \notin \mathcal{F}_{\delta_h^*(\lambda_h)}\big) \le \sum_{h \in \mathcal{H}_\epsilon} n^{-2} \le n^{-1} \tag{6.19}$$

and

$$\sum_{h \in \mathcal{H}_\epsilon} \mathbb{P}\big(\Delta_h^c\big) \le \sum_{h \in \mathcal{H}_\epsilon} \frac{5}{n^2} \le 5n^{-1}, \tag{6.20}$$

where $\Delta_h := \Delta$ is defined in Proposition 4. Thus, we may restrict our considerations to the event $\bigcap_{h \in \mathcal{H}_\epsilon, \lambda_h \in \Lambda} \{\hat{f}^h \in \mathcal{F}_{\delta_h^*(\lambda_h)}\} \cap \Delta_h$, since we are only interested on the asymptotic behavior. 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 $R_{\mathcal{A}}(\hat{f}^h, f^*) := \mathbb{E}[|\hat{f}^h(x_0) - f^*(x_0)|^q \mathbb{1}\{\mathcal{A}\}]$ and $R_{\mathcal{A}^c}(\hat{f}^h, f^*) := \mathbb{E}[|\hat{f}^h(x_0) - f^*(x_0)|^q \mathbb{1}\{\mathcal{A}^c\}]$.

Control of the risk on the event $\mathcal{A}$

With the triangular inequality and Lemma 3, we obtain

$$R_{\mathcal{A}}\big(\hat{f}^{\hat{h}}, f^*\big) \le 3^{q-1}\big[R_{\mathcal{A}}\big(\hat{f}^{h_\epsilon^*,\hat{h}}, \hat{f}^{\hat{h}}\big) + R_{\mathcal{A}}\big(\hat{f}^{\hat{h},h_\epsilon^*}, \hat{f}^{h_\epsilon^*}\big) + R_{\mathcal{A}}\big(\hat{f}^{h_\epsilon^*}, f^*\big)\big]. \tag{6.21}$$

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

$$R_{\mathcal{A}}\big(\hat{f}^{h_\epsilon^*,\hat{h}}, \hat{f}^{\hat{h}}\big) \le \mathbb{E} \sup_{h \in \mathcal{H}: h \succeq h_\epsilon^*} \big|\hat{f}^{h_\epsilon^*,h}(x_0) - \hat{f}^h(x_0)\big|^q. \tag{6.22}$$

Using (6.18) and taking $f = \hat{f}^{h_\epsilon^*,h}$ and $\tilde{f} = \hat{f}^h$, we then have

$$\big|\hat{f}^{h_\epsilon^*,h}(x_0) - \hat{f}^h(x_0)\big| \le 2c_{\hat{\rho}}^{-1}\big|\mathcal{D}_h\big(\hat{f}^{h_\epsilon^*,h}\big) - \mathcal{D}_h\big(\hat{f}^h\big)\big|.$$

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 $h \in \mathcal{H}$

$$\begin{aligned}
&\big|\hat{f}^{h_\epsilon^*,h}(x_0) - \hat{f}^h(x_0)\big| \\
&\quad \le 2c_{\hat{\rho}}^{-1}\big(\big|\mathcal{D}_h\big(\hat{f}^{h_\epsilon^*,h}\big) - \mathcal{D}_{h_\epsilon^*\vee h}\big(\hat{f}^{h_\epsilon^*,h}\big)\big| \\
&\qquad + \big|\mathcal{D}_{h_\epsilon^*\vee h}\big(\hat{f}^{h_\epsilon^*,h}\big) - \tilde{\mathcal{D}}_{h_\epsilon^*\vee h}\big(\hat{f}^{h_\epsilon^*,h}\big)\big| + \big|\tilde{\mathcal{D}}_h\big(\hat{f}^h\big) - \mathcal{D}_h\big(\hat{f}^h\big)\big|\big).
\end{aligned} \tag{6.23}$$

Denote by $\hat{\lambda}_h(f)(x,y) = \hat{\rho}(y - f(x))\hat{K}_h(x)$ and $\tilde{\delta}_h := \delta_h^*(\hat{\lambda}_h) \vee \delta_{h\vee h_\epsilon^*}^*(\hat{\lambda}_{h\vee h_\epsilon^*})$, using the last inequality and (6.22), we have

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

Using Lemma 5 and Lemma 6 with $h' = h_\epsilon^*$, there exists a universal positive constant $\mathcal{C}$ such that

$$R_{\mathcal{A}}\big(\hat{f}^{\hat{h},h_\epsilon^*}, \hat{f}^{\hat{h}}\big) \le \mathcal{C}\left(L\sum_{j=1}^d \big(h_{\epsilon,j}^*\big)^{\beta_j} + \frac{\sqrt{\mathrm{V}(\rho^*, K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h_\epsilon^*}}}\right)^q. \tag{6.24}$$

The second term on the right-hand side of (6.21) is controlled by the procedure (4.8), which implies

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

On the event $\bigcap_{h\in\mathcal{H}_\epsilon} \Delta_h$,

$$R_{\mathcal{A}}\big(\hat{f}^{\hat{h},h_\epsilon^*}, \hat{f}^{h_\epsilon^*}\big) \le \left(16\sqrt{6}\frac{\sqrt{\mathrm{V}(\rho^*, K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h_\epsilon^*}}}\right)^q. \tag{6.25}$$

By the definition of the Hölder class (Definition 4) and b_h (Definition (2.6)), we can control the bias for any $h \in \mathcal{H}$: $b_h \le \sup_{x\in V_h} |f^*(x_0) - f^*(x)| \le L\sum_{j=1}^d h_j^{\beta_j}$. Finally, with Theorem 2, we can bound the third term in (6.21): There exists a universal positive constant $\mathcal{C}$ such that

$$R_{\mathcal{A}}\big(\hat{f}^{h_\epsilon^*}, f^*\big) \asymp \mathcal{C}\left(L\sum_{j=1}^d \big(h_{\epsilon,j}^*\big)^{\beta_j} + \frac{\sqrt{\mathrm{V}(\rho^*, K^*)}B_0}{\sqrt{n\Pi_{h_\epsilon^*}}}\right)^q.$$

Using (6.21), (6.24), (6.25), and the last inequality, we have a control of the risk on the event $\mathcal{A}$ such that

$$R_{\mathcal{A}}\big(\hat{f}^{\hat{h}}, f^*\big) \le \mathcal{C}\left(L\sum_{j=1}^{d}\big(h^*_{\epsilon,j}\big)^{\beta_j} + \frac{\sqrt{\mathrm{V}(\rho^*, K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h^*_\epsilon}}}\right)^q \tag{6.26}$$

as $n \to \infty$ and for a universal positive constant $\mathcal{C}$.

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

$$R_{\mathcal{A}^c}\big(\hat{f}^{\hat{h}}, f^*\big) \le (2M)^q \mathbb{P}\big(\mathcal{A}^c\big). \tag{6.27}$$

We now show that the probability $\mathbb{P}(\mathcal{A}^c)$ is small. According to the construction of the procedure (4.8), the event $\mathcal{A}^c$ implies that there exists a $h' \in \mathcal{H}_\epsilon$ such that $h' \preceq h^*_\epsilon$ and

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

Using (6.18) and taking $f = \hat{f}^{h^*_\epsilon,h'}$ and $\tilde{f} = f^{h'}$, we have on the event $\mathcal{A}^c$

$$\frac{4}{3}c_{\hat{\rho}}^{-1}\big|\mathcal{D}_{h'}\big(\hat{f}^{h^*_\epsilon,h'}\big) - \mathcal{D}_{h'}\big(\hat{f}^{h'}\big)\big| > 16\frac{\sqrt{\widehat{\mathrm{V}}(\hat{\rho}, \hat{K})}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}.$$

From the last inequality, we obtain (cf. (6.23))

$$\begin{aligned}&\frac{4}{3}c_{\hat{\rho}}^{-1}\sup_{f\in\mathcal{F}_{\tilde{\delta}_{h'}}}\big|\mathcal{D}_{h'}(f) - \mathcal{D}_{h^*_\epsilon\vee h'}(f)\big| + \frac{8}{3}c_{\hat{\rho}}^{-1}\sup_{f\in\mathcal{F}_{\tilde{\delta}_{h'}}}\big|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\big|\\ &\quad > 16\frac{\sqrt{\widehat{\mathrm{V}}(\hat{\rho}, \hat{K})}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}.\end{aligned}$$

Together with Lemma 6, this yields

$$\frac{5}{3}L\sum_{j=1}^{d}\big(h^*_{\epsilon,j}\big)^{\beta_j} + \frac{8}{3}c_{\hat{\rho}}^{-1}\sup_{f\in\mathcal{F}_{\tilde{\delta}_{h'}}}\big|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\big| > 16\frac{\sqrt{\widehat{\mathrm{V}}(\hat{\rho}, \hat{K})}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}}.$$

On the event $\bigcap_{h\in\mathcal{H}_\epsilon}\Delta_h$, we get similarly as in (6.8)

$$\frac{5}{3}L\sum_{j=1}^{d}\big(h^*_{\epsilon,j}\big)^{\beta_j} + \frac{8}{3}c_{\hat{\rho}}^{-1}\sup_{f\in\mathcal{F}_{\tilde{\delta}_{h'}}}\big|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\big| > \frac{16\sqrt{2}}{3}\frac{\sqrt{\mathrm{V}(\hat{\rho}, \hat{K})}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}},$$

this implies

$$c_{\hat{\rho}}^{-1} \sup_{f\in\mathcal{F}_{\tilde{\delta}_{h'}}} \left|\tilde{\mathcal{D}}_{h'}(f) - \mathcal{D}_{h'}(f)\right| > \frac{16\sqrt{2}}{8} \frac{\sqrt{\mathrm{V}(\hat{\rho}, \hat{K})}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h'}}} - \frac{5}{8} L \sum_{j=1}^{d} \left(h_{\epsilon,j}^*\right)^{\beta_j}.$$

By definition, the oracle bandwidth h_ϵ^* is the one which gives the best trade-off. Thus by definition of h_ϵ^*, for all $h' \preceq h_\epsilon^* \preceq h^*$

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

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

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

Then, we have a control of the following probability

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

Using $\mathrm{ani}_\epsilon(n)/\ln^2(n) \le 1$ and Propostion 2 with z such that $B_z = B_0 + \mathrm{ani}_\epsilon(n)$, we deduce that

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

From (6.27) and the last inequality, we obtain on the event $\mathcal{A}^c$: $R_{\mathcal{A}^c}(\hat{f}^{\hat{h}}, f^*) \le (2M)^q n^{-1}$. Then, in view of the last inequality, (6.19), (6.20) and (6.26), we conclude that there exists a universal positive constant $\mathcal{C}$ such that

$$\mathbb{E}\left|\hat{f}^{\hat{h}}(x_0) - f^*(x_0)\right|^q \le \mathcal{C}\left(L \sum_{j=1}^{d} \left(h_{\epsilon,j}^*\right)^{\beta_j} + \frac{\sqrt{\mathrm{V}(\rho^*, K^*)}(B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n\Pi_{h_\epsilon^*}}}\right)^q.$$

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

Appendix

A.1. Proofs of the auxiliary results

Proof of Proposition 1. In this proof, we use a special case of a deviation inequality derived in [25], Corollary 6.9 (see the arXiv version for details). We recall that $\hat{f}_\lambda$ is the solution of the equation $\tilde{D}_\lambda(\cdot) = 0$, thanks to the continuity of $\rho'(\cdot)$, and we note that the following inclusion holds:

$$\begin{aligned}
&\bigcup_{\lambda\in\Lambda} \{\hat{f}_\lambda \notin \mathcal{F}_{\delta_h^*(\lambda)}\} \\
&\quad \subseteq \bigcup_{\lambda\in\Lambda} \Bigl\{ \sup_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_h^*(\lambda)}} \bigl\|\tilde{D}_\lambda(f) - \mathbb{E}^0\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_1} \geq \inf_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_h^*(\lambda)}} \bigl\|\mathbb{E}^0\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_1} \Bigr\} \qquad \text{(A.1)} \\
&\quad \subseteq \Bigl\{ \sup_{\lambda\in\Lambda} \Bigl[\sup_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_h^*(\lambda)}} \bigl\|\tilde{D}_\lambda(f) - \mathbb{E}^0\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_1} - \inf_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_h^*(\lambda)}} \bigl\|\mathbb{E}^0\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_1} \Bigr] \geq 0 \Bigr\}.
\end{aligned}$$

Next, it holds that

$$\begin{aligned}
\bigl\|\tilde{D}_\lambda(f) - \mathbb{E}^0\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_1} &\leq |\mathcal{P}| \bigl\|\tilde{D}_\lambda(f) - \mathbb{E}\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_\infty} \\
&\quad + |\mathcal{P}| \bigl\|\mathbb{E}\bigl[\tilde{D}_\lambda(f)\bigr] - \mathbb{E}^0\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_\infty}.
\end{aligned} \qquad \text{(A.2)}$$

By the definitions of $\mathbb{E}[\tilde{D}_\lambda^p(\cdot)]$ and $\mathbb{E}^0[\tilde{D}_\lambda^p(\cdot)]$ in (6.3), by change of variables and using that $\rho'(\cdot)$ is 1-Lipschitz we have for any $f \in \mathcal{F}$, and any $p \in \mathcal{P}$

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \bigl\|\mathbb{E}\bigl[\tilde{D}_\lambda(f)\bigr] - \mathbb{E}^0\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_\infty} \\
&\quad \leq \int \mu(x) K_h(x) \int \bigl|\rho'\bigl(\sigma(x)z + f^0(x) - f(x)\bigr) \\
&\qquad\qquad - \rho'\bigl(\sigma(x)z + f^*(x) - f(x)\bigr)\bigr| \mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x \qquad \text{(A.3)} \\
&\quad \leq \mathbb{E}\bigl[K_h(X)\bigr] b_h.
\end{aligned}$$

To control the stochastic term, we can then apply Massart's Inequality to get (see the arXiv version for details)

$$\begin{aligned}
&\mathbb{P}\Biggl(\sup_{\lambda\in\Lambda, f\in\mathcal{F}} \frac{\sqrt{n\Pi_h}\|\tilde{D}_\lambda(f) - \mathbb{E}[\tilde{D}_\lambda(f)]\|_{\ell_\infty}}{\|\rho'\|_\infty(\sqrt{\mathbb{E}[\Pi_h K_h^2(X)]} + \|K\|_\infty/\sqrt{n\Pi_h})} \\
&\qquad \geq 27 \int_0^1 H_{\mathcal{F}\times\Lambda}^{1/2}(u)\,\mathrm{d}u + 4H_{\mathcal{F}}(1) + 7\sqrt{2z} + 2z \Biggr) \\
&\quad \leq 2|\mathcal{P}| \exp(-z).
\end{aligned}$$

Note that the factor 2 in the last inequality appears because we need to control deviations of the absolute value of the empirical process. Using (A.2), (A.3), and the last inequality, we then obtain for all $z > 0$

$$\begin{aligned}\mathbb{P}\Bigg(&\sup_{\lambda\in\Lambda}\sup_{f\in\mathcal{F}\backslash\mathcal{F}_{\delta_h^*(\lambda)}}\frac{\sqrt{n\Pi_h}(\|\tilde{D}_\lambda(f)-\mathbb{E}^0[\tilde{D}_\lambda(f)]\|_{\ell_1}-|\mathcal{P}|\mathbb{E}[K_h(X)]b_h)}{\|\rho'\|_\infty(\sqrt{\mathbb{E}[\Pi_h K_h^2(X)]}+\|K\|_\infty/\sqrt{n\Pi_h})}\\ &\geq|\mathcal{P}|\bigg(27\int_0^1 H_{\mathcal{F}\times\Lambda}^{1/2}(u)\,\mathrm{d}u+4H_{\mathcal{F}}(1)+7\sqrt{2z}+2z\bigg)\Bigg)\leq 2|\mathcal{P}|e^{-z}.\end{aligned}\tag{A.4}$$

Now, let us have a look at $\inf_{f\in\mathcal{F}\backslash\mathcal{F}_{\delta_h^*(\lambda)}}\|\mathbb{E}^0[\tilde{D}_\lambda(f)]\|_{\ell_1}$ in (A.1). By the definition of $\tilde{D}_\lambda(\cdot)$ and using that $|t_p^0-t_p|\leq\|t^0-t\|_{\ell_1}$ for all $p\in\mathcal{P}$, we have for any $f\in\mathcal{F}\backslash\mathcal{F}_{\delta_h^*(\lambda)}$

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

where $\mathbb{G}(\cdot)=n^{-1}\sum_{i=1}^n g_i(\cdot)$ and t is such that $f=\mathrm{P}_t$. The last inequality is obtained using that $\sum_{p\in\mathcal{P}}(t_p^0-t_p)((x-x_0)/h)^p=f(x)-f^0(x)$ and the triangular inequality. Since $\mathbb{G}(\cdot)$ is symmetric, $\rho'(\cdot)$ increasing (because of the convexity of ρ), $K(\cdot)$ is nonnegative, and $\rho'(\cdot)$ is odd ($\rho(\cdot)$ is symmetric) and positive on $(0,\infty)$ (because of $\rho'(0)=0$, the convexity of $\rho(\cdot)$ and the strict convexity around 0), the last equality implies for all $f\in\mathcal{F}\backslash\mathcal{F}_{\delta_h^*(\lambda)}$

$$\begin{aligned}&\big\|\mathbb{E}^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1}\\ &\quad\geq\int\frac{|f^0(x)-f(x)|}{\|t^0-t\|_{\ell_1}}\mu(x)K_h(x)\int\rho'\big(\sigma(x)z+\big|f^0(x)-f(x)\big|\big)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x\\ &\quad\geq\int\frac{|f^0(x)-f(x)|}{\|t^0-t\|_{\ell_1}}\mu(x)K_h(x)\int\rho'\left(\sigma(x)z+\delta_h^*(\lambda)\frac{|f^0(x)-f(x)|}{\|t^0-t\|_{\ell_1}}\right)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x.\end{aligned}$$

Recall that for any x, $\int\rho'(\sigma(x)z)\mathbb{G}(z)\,\mathrm{d}z=0$ thanks to the symmetry of $\rho(\cdot)$ and $\mathbb{G}(\cdot)$. Since $|f^0(x)-f(x)|\|t^0-t\|_{\ell_1}^{-1}\leq 1$, we obtain with the mean value theorem for all $f\in\mathcal{F}\backslash\mathcal{F}_{\delta_h^*(\lambda)}$

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

We then derive, using that $2\delta_h^*(\lambda) \le \inf_{x\in V_h} \int \rho''(\sigma(x)z)\mathbb{G}(z)\,\mathrm{d}z$ for all $\lambda \in \Lambda$ (see Condition 2) and $\rho''(\cdot)$ is $\mathbb{P}$-continuous,

$$\inf_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_h^*(\lambda)}} \big\|\mathbb{E}^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1} \ge \frac{\delta_h^*(\lambda)}{2} \inf_{t:\|t\|_{\ell_1}\ge\delta_h^*(\lambda)} \int \frac{|\mathrm{P}_t(x)|^2}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)\int \rho''\big(\sigma(x)z\big)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x.$$

We then observe that $\mathrm{P}_t(x) = t^\top U(\frac{x-x_0}{h})$ and thus

$$\begin{aligned}&\int \frac{|\mathrm{P}_t(x)|^2}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)\int \rho''\big(\sigma(x)z\big)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x\\ &\quad = t^\top\biggl[\int \frac{U((x-x_0)/h)U^\top((x-x_0)/h)}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)\int \rho''\big(\sigma(x)z\big)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x\biggr]t.\end{aligned}$$

We can thus write by the definition of Φ_h in Condition 2

$$\begin{aligned}&t^\top\biggl[\int \frac{U((x-x_0)/h)U^\top((x-x_0)/h)}{\|t\|_{\ell_1}^2}\mu(x)K_h(x)\int \rho''\big(\sigma(x)z\big)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x\biggr]t\\ &\quad \ge \frac{\|t\|_{\ell_2}^2}{\|t\|_{\ell_1}^2}\Phi_h \ge \Phi_h/|\mathcal{P}|.\end{aligned}$$

In summary, we have for any $\lambda \in \Lambda$, $\inf_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_h^*(\lambda)}} \|\mathbb{E}^0[\tilde{D}_\lambda(f)]\|_{\ell_1} \ge \frac{\Phi_h\delta_h^*(\lambda)}{2|\mathcal{P}|}$. By the definition of $\delta_h^*(\lambda)$ in Condition 2 and as $n\Pi_h \ge 1$, it holds that

$$\delta_h^*(\lambda) > 2|\mathcal{P}|^2\frac{\|\rho'\|_\infty(\sqrt{\mathbb{E}[\Pi_h K_h^2(X)]} + \|K\|_\infty/\sqrt{n\Pi_h})}{\Phi_h\sqrt{n\Pi_h}(E^* + 7\sqrt{4\ln(2|\mathcal{P}|n)} + 4\ln(2|\mathcal{P}|n))^{-1}} + 2|\mathcal{P}|^2\mathbb{E}\big[K_h(X)\big]\frac{b_h}{\Phi_h}.$$

Using Inequalities (A.1) and (A.4) with $z = \ln(2|\mathcal{P}|n)$, and the last inequality, we obtain

$$\begin{aligned}\mathbb{P}\biggl(\bigcup_{\lambda\in\Lambda}\{\hat{f}_\lambda \notin \mathcal{F}_{\delta_h^*(\lambda)}\}\biggr) &\le \mathbb{P}\biggl(\sup_{\lambda\in\Lambda}\sup_{f\in\mathcal{F}\setminus\mathcal{F}_{\delta_h^*(\lambda)}}\biggl[\big\|\tilde{D}_\lambda(f) - \mathbb{E}^0\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_1} - \frac{\Phi_h\delta_h^*(\lambda)}{2|\mathcal{P}|}\biggr] \ge 0\biggr)\\ &\le 1/n^2.\end{aligned}$$

□

Proof of Proposition 3. The definitions of $\hat{f}_\lambda$ and f^0 (see (2.4) and (2.5), resp.) imply that $|\hat{f}_\lambda(x_0) - f^*(x_0)| = |(\hat{t}_\lambda)_{0,\dots,0} - t^0_{0,\dots,0}| \le \|\hat{t}_\lambda - t^0\|_{\ell_\infty}$. Using $\hat{f}_\lambda \in \mathcal{F}_{\delta_h^*(\lambda)}$, Lemma 1, and the last inequality, we have

$$\big|\hat{f}_\lambda(x_0) - f^*(x_0)\big| \le \tfrac{4}{3}c_\lambda^{-1}\big\|\mathbb{E}^0\big[\tilde{D}_\lambda(\hat{f}_\lambda)\big] - \mathbb{E}^0\big[\tilde{D}_\lambda\big(f^0\big)\big]\big\|_{\ell_\infty}.$$

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

$$\left|\hat{f}_\lambda(x_0)-f^*(x_0)\right|\le\tfrac{4}{3}c_\lambda^{-1}\left(\left\|\tilde{D}_\lambda(\hat{f}_\lambda)-\mathbb{E}\left[\tilde{D}_\lambda(\hat{f}_\lambda)\right]\right\|_{\ell_\infty}+\left\|\mathbb{E}\left[\tilde{D}_\lambda(\hat{f}_\lambda)\right]-\mathbb{E}^0\left[\tilde{D}_\lambda(\hat{f}_\lambda)\right]\right\|_{\ell_\infty}\right).$$

From Lemma 2 and the last display, we obtain

$$\begin{aligned}\left|\hat{f}_\lambda(x_0)-f^*(x_0)\right|&\le\tfrac{4}{3}c_\lambda^{-1}\left(\left\|\tilde{D}_\lambda(\hat{f}_\lambda)-\mathbb{E}\left[\tilde{D}_\lambda(\hat{f}_\lambda)\right]\right\|_{\ell_\infty}+\tfrac{5}{4}c_\lambda b_h\right)\\&\le\tfrac{5}{3}b_h+\tfrac{4}{3}\sup_{\lambda\in\Lambda}\sup_{f\in\mathcal{F}_{\delta_h^*(\lambda)}}c_\lambda^{-1}\left\|\tilde{D}_\lambda(f)-\mathbb{E}\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_\infty}.\end{aligned}$$

This yields

$$\left|\hat{f}_\lambda(x_0)-f^*(x_0)\right|\le 3b_h+2\sup_{\lambda\in\Lambda}\sup_{f\in\mathcal{F}_{\delta_h^*(\lambda)}}c_\lambda^{-1}\left\|\tilde{D}_\lambda(f)-\mathbb{E}\left[\tilde{D}_\lambda(f)\right]\right\|_{\ell_\infty}.$$

From the last inequality and the definitions of $\mathrm{V}(\cdot)$ and c_λ introduced in (2.9) and (6.5), respectively, we deduce

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

Using Proposition 2 and the last inequality, we finally obtain

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

□

Proof of Proposition 4. We first recall by the definition of the estimator (3.2)

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

where

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

and

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

In the following, we assume to be on the event $\bigcap_{\lambda\in\Lambda}\{\hat{f}_\lambda \in \mathcal{F}_{\delta_h^*(\lambda)}\}$, which is true with probability at least $1 - 1/n^2$ according to Proposition 1. Then, using Massart's Inequality (see the arXiv version for details) we can control the deviation of the process $\Pi_h P_n[\lambda'(\hat{f}_\lambda)]^2$ as follows:

$$\mathbb{P}\left(\sup_{\lambda\in\Lambda}\sup_{f\in\mathcal{F}_{\delta_h^*(\lambda)}} \left(\lambda'_\infty\right)^{-2}\left|\Pi_h P_n\left[\lambda'(f)\right]^2 - \Pi_h\mathbb{E}P_n\left[\lambda'(f)\right]^2\right| \geq \frac{B_{2\ln(n)}}{\sqrt{n\Pi_h}}\right) \leq 2/n^2, \tag{A.5}$$

where $B_.$ is defined in (6.6). Similarly, using again Massart's Inequality, we control the deviation of $P_n\lambda''(\hat{f}_\lambda)$ as follows:

$$\mathbb{P}\left(\sup_{\lambda\in\Lambda}\sup_{f\in\mathcal{F}_{\delta_h^*(\lambda)}} \|K\|_\infty^{-1}\left|P_n\lambda''(f) - \mathbb{E}P_n\lambda''(f)\right| \geq \frac{B_{2\ln(n)}}{\sqrt{n\Pi_h}}\right) \leq 2/n^2. \tag{A.6}$$

Then, for any $\lambda \in \Lambda$, by the continuity of ρ' and ρ'' almost everywhere, $\|\rho''\|_\infty \leq 1$, and the mean value theorem, we have for all $f \in \mathcal{F}_{\delta_h^*(\lambda)}$

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

Similarly, $\sup_{f\in\mathcal{F}_{\delta_h^*(\lambda)}} |\mathbb{E}P_n\lambda''(f) - \mathbb{E}P_n\lambda''(f^*)| \leq \|K\|_\infty(\delta_h^*(\lambda) + b_h)$. Note that for any $\lambda \in \Lambda$

$$s_n := \|K\|_\infty s_h(\lambda) \geq \big(1 \vee 2\|K\|_\infty\big)\|K\|_\infty\big[\delta_h^*(\lambda) + b_h\big] + \big[\left(\lambda'_\infty\right)^2 \vee \|K\|_\infty\big]\frac{B_{2\ln(n)}}{\sqrt{n\Pi_h}},$$

and we observe (under Condition 3) that $s_n \leq \frac{1}{2}\min\{\mathbb{E}P_n\lambda''(f^*), \Pi_h\mathbb{E}P_n[\lambda'(f^*)]^2\}$. Using this, (A.5), and (A.6), we obtain with probability $1 - 5/n^2$ for any $\lambda \in \Lambda$

$$\sqrt{\widehat{\mathrm{V}}(\lambda)} \leq \frac{\sqrt{\Pi_h\mathbb{E}P_n[\lambda'(f^*)]^2 + s_n} + \lambda'_\infty\ln^2(n)/\sqrt{n\Pi_h}}{\mathbb{E}P_n\lambda''(f^*) - s_n} \leq \sqrt{6}\sqrt{\mathrm{V}(\lambda)}$$

and

$$\sqrt{\widehat{\mathrm{V}}(\lambda)} \geq \frac{\sqrt{\Pi_h\mathbb{E}P_n[\lambda'(f^*)]^2 - s_n} + \lambda'_\infty\ln^2(n)/\sqrt{n\Pi_h}}{\mathbb{E}P_n\lambda''(f^*) + s_n} \geq \frac{\sqrt{2}}{3}\sqrt{\mathrm{V}(\lambda)}.$$

(Instead of the given factors in front of $\sqrt{\mathrm{V}(\lambda)}$, one could readily obtain factors that tend to one as $n \to \infty$. This is of minor interest here.) This proves the claim. □

A.2. Technical lemmas

We first give a result for the deterministic criterion $\mathbb{E}^0[\tilde{D}_\lambda(\cdot)]$ defined in (6.3):

Lemma 1. *Let λ be as in (2.4), $n \in \{1, 2, \ldots\}$, and $h \in (0, 1]^d$ such that Condition 2 is satisfied, the following holds*:

1. $\mathbb{E}^0[\tilde{D}_\lambda(f^0)] = 0$, *and the function $\mathbb{E}^0[\tilde{D}_\lambda(f)]$ is bijective as function of $\mathcal{F}_{\delta_h^*(\lambda)}$ (see Definition (6.1)) on the corresponding image.*
2. *For any $f, \tilde{f} \in \mathcal{F}_{\delta_h^*(\lambda)}$, $\|t - \tilde{t}\|_{\ell_\infty} \le \frac{4}{3} c_\lambda^{-1} \|\mathbb{E}^0[\tilde{D}_\lambda(f)] - \mathbb{E}^0[\tilde{D}_\lambda(\tilde{f})]\|_{\ell_\infty}$, where $\mathrm{P}_t = f$ and $\mathrm{P}_{\tilde{t}} = \tilde{f}$.*

Next, we consider the bias.

Lemma 2. *Let λ be as in (2.4), $n \in \{1, 2, \ldots\}$, and $h \in (0, 1]^d$ such that Condition 2 is satisfied, it holds that*

$$\sup_{f \in \mathcal{F}_{\delta_h^*(\lambda)}} \big\|\mathbb{E}^0\big[\tilde{D}_\lambda(f)\big] - \mathbb{E}\big[\tilde{D}_\lambda(f)\big]\big\|_{\ell_\infty} \le \tfrac{5}{4} c_\lambda b_h.$$

Next, we do some simple algebra.

Lemma 3. *For any $x, y \in [0, \infty)$, it holds that $x^q \le 2^q [x - y]_+^q + 2^q y^q$. Moreover, for any $l, q \in \{1, 2, \ldots\}$ and $x_1, \ldots, x_l \ge 0$, it holds, that $(\sum_{i=1}^l x_i)^q \le l^{q-1}(\sum_{i=1}^l x_i^q)$.*

The proof consists of simple algebra and is available in the arXiv version.

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

Lemma 4. *Let $\mathcal{D}_h(\cdot) : [-M, M] \to \mathbb{R}$ 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 such that Condition 2 is satisfied for all $h \in \mathcal{H}$. Then, for any $h \in \mathcal{H}$ and $t, \tilde{t} \in [f^*(x_0) - \delta_h^*(\lambda), f^*(x_0) + \delta_h^*(\lambda)]$, it holds that $|t - \tilde{t}| \le \frac{4}{3} c_{\hat{\rho}}^{-1} |\mathcal{D}_h(t) - \mathcal{D}_h(\tilde{t})|$.*

The proof of the lemma is similar to the one of Lemma 1 (see the arXiv version for details).

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

Lemma 5. *For n sufficiently large n sufficiently large such that Conditions 2 and 3 are satisfied for all $h \in \mathcal{H}$. It holds that*

$$\mathbb{E} c_{\hat{\rho}}^{-q} \sup_{h \in \mathcal{H}: h \succeq h_\epsilon^*} \sup_{f \in \mathcal{F}_{\tilde{\delta}_h}} \big|\tilde{\mathcal{D}}_h(f) - \mathcal{D}_h(f)\big|^q \asymp 2^q \left(\frac{\sqrt{6 \mathrm{V}(\rho^*, K^*)} (B_0 + \mathrm{ani}_\epsilon(n))}{\sqrt{n \Pi_{h_\epsilon^*}}} \right)^q,$$

where $\tilde{\delta}_h$, h_ϵ^*, $\tilde{\mathcal{D}}$ 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 (4.3) and (4.4),* $\mathrm{ani}_\epsilon(n)$ *is defined in Section 4.2.*

The proof is an application of Proposition 2 (see the arXiv version for details).

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

Lemma 6. *For any $f^* \in \mathbb{H}_d(\vec{\beta}, L, M)$ such that $\vec{\beta} \in (0,1]^d$, and n sufficiently large such that Condition 2 is satisfied for all $h \in \mathcal{H}$, it holds that for any $h, h' \in \mathcal{H}$*

$$\sup_{f \in \mathcal{F}_{\bar{\delta}_h}} \left|\mathcal{D}_{h'\vee h}(f) - \mathcal{D}_h(f)\right| \le \frac{5}{4} c_{\hat{\rho}} L \sum_{j=1}^d \left(h'_j\right)^{\beta_j},$$

where $\mathcal{D}_h$ and $c_{\hat{\rho}}$ are defined in (6.16) and (6.17) in the proof of Theorem 4.

A.3. Proofs of the technical lemmas

Proof of Lemma 1. Let us proof the first claim. For this, we note that the components of $\mathbb{E}^0[\tilde{D}_\lambda(f)]$ are given by

$$\mathbb{E}^0\left[\tilde{D}_\lambda^p(f)\right] = \int \left(\frac{x - x_0}{h}\right)^p \mu(x) K_h(x) \int \rho'\left(\sigma(x)z + f^0(x) - f(x)\right) \frac{1}{n} \sum_{i=1}^n g_i(z)\, \mathrm{d}z\, \mathrm{d}x.$$

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

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

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 $\mathbb{G}(\cdot) := n^{-1}\sum_{i=1}^n g_i(\cdot)$. Since $\mathbb{G}(\cdot)$ is symmetric, $K(\cdot)$ is nonnegative, and $\rho'(\cdot)$ is odd and positive on $(0, \infty)$, the last display implies

$$\begin{aligned}
&\int K_h(x)\mu(x)\left[u(x) - \tilde{u}(x)\right] \\
&\quad \times \int \left[\rho'\left(\sigma(x)z - u(x)\right) - \rho'\left(\sigma(x)z - \tilde{u}(x)\right)\right]\mathbb{G}(z)\, \mathrm{d}z\, \mathrm{d}x = 0 \\
&\Leftrightarrow \quad \int K_h(x)\mu(x)\left|u(x) - \tilde{u}(x)\right| \\
&\qquad \times \int \left|\rho'\left(\sigma(x)z - u(x)\right) - \rho'\left(\sigma(x)z - \tilde{u}(x)\right)\right|\mathbb{G}(z)\, \mathrm{d}z\, \mathrm{d}x = 0.
\end{aligned}$$

As $f, \tilde{f} \in \mathcal{F}_{\delta_h^*(\lambda)}$, it holds that $\sup_{x \in V_h} |u(x)| \vee |\tilde{u}(x)| \le \delta_h^*(\lambda)$. Moreover, using the mean value theorem, the $\mathbb{P}$-continuity of ρ'' and Condition 2, we obtain

$$\begin{aligned}
&\int K_h(x)\mu(x)\big|u(x)-\tilde{u}(x)\big| \int \big|\rho'\big(\sigma(x)z-u(x)\big)-\rho'\big(\sigma(x)z-\tilde{u}(x)\big)\big|\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x \\
&\quad \ge \int K_h(x)\mu(x)\big|u(x)-\tilde{u}(x)\big|^2 \inf_{s:|s|\le\delta_h^*(\lambda)} \int \rho''\big(\sigma(x)z-s\big)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x \\
&\quad \ge \int K_h(x)\mu(x)\big|u(x)-\tilde{u}(x)\big|^2 \inf_{s:|s|\le\delta_h^*(\lambda)} \int \rho''\big(\sigma(x)z-s\big)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x \\
&\quad \ge \int K_h(x)\mu(x)\big|u(x)-\tilde{u}(x)\big|^2\bigg[\int \rho''\big(\sigma(x)z\big)\mathbb{G}(z)\,\mathrm{d}z-\delta_h^*(\lambda)\bigg]\,\mathrm{d}x \\
&\quad \ge \frac{1}{2}\int K_h(x)\mu(x)\big|u(x)-\tilde{u}(x)\big|^2 \int \rho''\big(\sigma(x)z\big)\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x.
\end{aligned}$$

The last display, Condition 2, and the nonnegativity of $K(\cdot)$ over its support yield that there exists an nonempty open set $\mathcal{V}$ such that $\sup_{x\in\mathcal{V}} |u(x)-\tilde{u}(x)| = 0$. 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 $D(\cdot) := \mathbb{E}^0[\tilde{D}_\lambda(\cdot)]$ and note that $D(\cdot)$ is differentiable and injective on $\mathcal{F}_{\delta_h^*(\lambda)}$ (the latter according to the first claim). We can consequently find an inverse of the function $D(\cdot)$ on the image of $D(\cdot)$ on $\mathcal{F}_{\delta_h^*(\lambda)}$. We then obtain, denoting the matrix ℓ_∞-norm by $|||\cdot|||_\infty$ and the inverse of $D(\cdot)$ by $D^{-1}(\cdot)$, for all $f \in \mathcal{F}_{\delta_h^*(\lambda)}$

$$\big|\big|\big|J_{D^{-1}}(f)\big|\big|\big|_\infty = \big|\big|\big|J_D^{-1}(f)\big|\big|\big|_\infty = \big|\big|\big|J_D(f)\big|\big|\big|_\infty^{-1} \le \big[J_D(f)\big]_{0,0}^{-1} = \big[\mathbb{E}P_n\lambda''(f)\big]^{-1} \le \tfrac{4}{3}c_\lambda^{-1}.$$

The constant c_λ is defined in (6.5) and the last inequality is obtained by the $\mathbb{P}$-continuity of $\rho''(\cdot)$ and Condition 2. The mean value theorem and the last inequality then imply for any $f, \tilde{f} \in \mathcal{F}_{\delta_h^*(\lambda)}$ and the associated coefficients t and $\tilde{t}$

$$\|t-\tilde{t}\|_{\ell_\infty} = \big\|D^{-1}\circ D(f) - D^{-1}\circ D(\tilde{f})\big\|_{\ell_\infty} \le \tfrac{4}{3}c_\lambda^{-1}\big\|D(f)-D(\tilde{f})\big\|_{\ell_\infty}.$$

This proves the second claim. □

Proof of Lemma 2. By the definitions of $\mathbb{E}[\tilde{D}_\lambda^p(\cdot)]$ and $\mathbb{E}^0[\tilde{D}_\lambda^p(\cdot)]$ in (6.3), we have for any $f \in \mathcal{F}_{\delta_h^*(\lambda)}$, any $\lambda \in \Lambda$, and any $p \in \mathcal{P}$

$$\begin{aligned}
&\big|\mathbb{E}^0\big[\tilde{D}_\lambda^p(f)\big] - \mathbb{E}\big[\tilde{D}_\lambda^p(f)\big]\big| \\
&\quad \le \int \mu(x)K_h(x) \qquad\qquad (A.7)\\
&\qquad \times \int \big|\rho'\big(\sigma(x)z+f^0(x)-f(x)\big)-\rho'\big(\sigma(x)z+f^*(x)-f(x)\big)\big|\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x.
\end{aligned}$$

It additionally holds for all $f \in \mathcal{F}_{\delta_h^*(\lambda)}$ that $\sup_{x\in V_h} |f^0(x) - f(x)| \le \delta_h^*(\lambda)$. Together with the definition of f^0 in (2.5), this implies for any $f \in \mathcal{F}_{\delta_h^*(\lambda)}$

$$\sup_{x\in V_h}\bigl|f^*(x) - f(x)\bigr| \le \sup_{x\in V_h}\bigl|f^*(x) - f^0(x)\bigr| + \sup_{x\in V_h}\bigl|f^0(x) - f(x)\bigr| \le b_h + \delta_h^*(\lambda).$$

This implies, due to the mean value theorem, that there is a $u_x \in \mathbb{R} : |u_x| \le b_h + \delta_h^*(\lambda)$ such that

$$\bigl|\rho'\bigl(\sigma(x)z + f^0(x) - f(x)\bigr) - \rho'\bigl(\sigma(x)z + f^*(x) - f(x)\bigr)\bigr| \le \bigl|f^*(x) - f^0(x)\bigr|\rho''\bigl(\sigma(x)z + u_x\bigr).$$

Using Condition 2, (A.7), the last inequality, and the definitions b_h, and c_λ defined in (2.6) and (6.5) respectively, we obtain for any $\lambda \in \Lambda$

$$\begin{aligned}&\sup_{f\in\mathcal{F}_{\delta_h^*(\lambda)}} \bigl\|\mathbb{E}^0\bigl[\tilde{D}_\lambda(f)\bigr] - \mathbb{E}\bigl[\tilde{D}_\lambda(f)\bigr]\bigr\|_{\ell_\infty}\\ &\quad\le \int \mu(x)K_h(x)\bigl|f^*(x) - f^0(x)\bigr| \int \bigl[\rho''\bigl(\sigma(x)z\bigr) + b_h + \delta_h^*(\lambda)\bigr]\mathbb{G}(z)\,\mathrm{d}z\,\mathrm{d}x \le \frac{5}{4}c_\lambda b_h.\end{aligned}$$ □

Proof of Lemma 6. 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

$$\begin{aligned}&\sup_{f\in\mathcal{F}_{\tilde{\delta}_h}} \bigl|\mathcal{D}_{h'\vee h}(f) - \mathcal{D}_h(f)\bigr|\\ &\quad= \sup_{f\in\mathcal{F}_{\tilde{\delta}_h}} \biggl|\int \hat{K}(x)\int \hat{\rho}'\bigl(\sigma z + f^*\bigl(x_0 + h\vee h'x\bigr) - f(x_0)\bigr)g(z)\,\mathrm{d}z\,\mathrm{d}x\\ &\qquad\qquad - \int \hat{K}(x)\int \hat{\rho}'\bigl(\sigma z + f^*(x_0 + hx) - f(x_0)\bigr)g(z)\,\mathrm{d}z\,\mathrm{d}x\biggr|.\end{aligned}$$

Using $f \in \mathbb{H}_d(\vec{\beta}, L, M)$, the $\mathbb{P}$-continuity of $\rho''(\cdot)$, the last equality, and the mean value theorem, we obtain:

$$\begin{aligned}&\sup_{f\in\mathcal{F}_{\tilde{\delta}_h}} \bigl|\mathcal{D}_{h'\vee h}(f) - \mathcal{D}_h(f)\bigr|\\ &\quad\le \sup_{|s|\le\tilde{\delta}_h + b_h}\int \hat{\rho}''(\sigma z + s)g(z)\,\mathrm{d}z\int \hat{K}(x)\bigl|f^*\bigl(x_0 + h\vee h'x\bigr) - f^*(x_0 + hx)\bigr|\,\mathrm{d}x\\ &\quad\le \biggl(\int \hat{\rho}''(\sigma z)g(z)\,\mathrm{d}z + \tilde{\delta}_h + b_h\biggr)L\sum_{j=1}^d \bigl|h_j \vee h_j' - h_j\bigr|^{\beta_j}.\end{aligned}$$

With Condition 2 and definition of $\tilde{\delta}_h$ in Proof of Theorem 4, this yields

$$\sup_{f\in\mathcal{F}_{\tilde{\delta}_h}} \bigl|\mathcal{D}_{h'\vee h}(f) - \mathcal{D}_h(f)\bigr| \le \frac{5}{4}c_{\hat{\rho}}L\sum_{j=1}^d \bigl(h_j'\bigr)^{\beta_j}.$$ □

Acknowledgements

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

We thank Oleg Lepski, Joseph Salmon, and Sara van de Geer for many helpful discussions. We also thank the associate editor and the referees for their valuable comments.

References

[1] Antoniadis, A., Pensky, M. and Sapatinas, T. (2014). Nonparametric regression estimation based on spatially inhomogeneous data: Minimax global convergence rates and adaptivity. *ESAIM Probab. Stat.* **18** 1–41.

[2] Arcones, M.A. (2005). Convergence of the optimal M-estimator over a parametric family of M-estimators. *Test* **14** 281–315. MR2203433

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

[4] Brown, L.D. and Low, M.G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535. MR1425965

[5] Cai, T.T. and Zhou, H.H. (2009). Asymptotic equivalence and adaptive estimation for robust nonparametric regression. *Ann. Statist.* **37** 3204–3235. MR2549558

[6] Chichignoud, M. (2012). Minimax and minimax adaptive estimation in multiplicative regression: Locally Bayesian approach. *Probab. Theory Related Fields* **153** 543–586. MR2948686

[7] Gaïffas, S. (2005). Convergence rates for pointwise curve estimation with a degenerate design. *Math. Methods Statist.* **14** 1–27. MR2158069

[8] Gaïffas, S. (2007). On pointwise adaptive curve estimation based on inhomogeneous data. *ESAIM Probab. Stat.* **11** 344–364 (electronic). MR2339297

[9] Gaïffas, S. (2007). Sharp estimation in sup norm with random design. *Statist. Probab. Lett.* **77** 782–794. MR2369683

[10] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707

[11] Goldenshluger, A. and Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150–1190. MR2543590

[12] Goldenshluger, A. and Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.* **6** 135–170. MR1466625

[13] Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872

[14] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

[15] Huber, P.J. (1981). *Robust Statistics*. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR0606374

[16] Huber, P.J. and Ronchetti, E.M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR2488795

[17] Katkovnik, V., Foi, A., Egiazarian, K. and Astola, J. (2010). From local kernel to nonlocal multiple-model image denoising. *Int. J. Comput. Vis.* **86** 1–32. MR2683762

[18] Katkovnik, V.Y. (1985). *Neparametricheskaya Identifikatsiya i Sglazhivanie Dannykh. Metod Lokalnoi Approksimatsii*. [*The Method of Local Approximation*]. *Teoreticheskie Osnovy Tekhnicheskoĭ Kibernetiki*. [*Theoretical Foundations of Engineering Cybernetics*]. Moscow: "Nauka". MR0874985

[19] Kerkyacharian, G., Lepski, O. and Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121** 137–170. MR1863916

[20] Klutchnikoff, N. (2005). Sur l'estimation adaptative de fonctions anisotropes. Ph.D. thesis, Aix-Marseille 1.

[21] Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Stat.* **5** 1015–1053. MR2836768

[22] Lepski, O.V. and Levit, B.Y. (1999). Adaptive nonparametric estimation of smooth multivariate functions. *Math. Methods Statist.* **8** 344–370. MR1735470

[23] 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. MR1447734

[24] Lepskiĭ, O.V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **35** 459–470. MR1091202

[25] Massart, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Berlin: Springer. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, with a foreword by Jean Picard. MR2319879

[26] Polzehl, J. and Spokoiny, V. (2006). Propagation–separation approach for local likelihood estimation. *Probab. Theory Related Fields* **135** 335–362. MR2240690

[27] Reiss, M., Rozenholc, Y. and Cuenod, C. (2009). Pointwise adaptive estimation for robust and quantile regression. Available at arXiv:0904.0543v1.

[28] Spokoiny, V. and Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.* **37** 2783–2807. MR2541447

[29] Stone, C.J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284. MR0362669

[30] Tsybakov, A.B. (1982). Nonparametric signal estimation when there is incomplete information on the noise distribution. *Probl. Inf. Transm.* **18** 44–60. MR0689339

[31] Tsybakov, A.B. (1982). Robust estimates of a function. *Problemy Peredachi Informatsii* **18** 39–52. MR0711899

[32] Tsybakov, A.B. (1983). Convergence of nonparametric robust algorithms of reconstruction of functions. *Avtomat. i Telemekh.* **12** 66–76. MR0749816

[33] Tsybakov, A.B. (1986). Robust reconstruction of functions by a local approximation method. *Problemy Peredachi Informatsii* **22** 69–84. MR0855002

[34] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. New York: Springer. Revised and extended from the 2004 French original, translated by Vladimir Zaiats. MR2724359

[35] van de Geer, S. and Lederer, J. (2013). The Bernstein–Orlicz norm and deviation inequalities. *Probab. Theory Related Fields* **157** 225–250.

Received January 2013 and revised May 2013