The Annals of Statistics
2015, Vol. 43, No. 4, 1617–1646
DOI: 10.1214/15-AOS1318

BANDWIDTH SELECTION IN KERNEL EMPIRICAL RISK MINIMIZATION VIA THE GRADIENT

BY MICHAËL CHICHIGNOUD[1] AND SÉBASTIEN LOUSTAU

ETH Zürich and University of Angers

In this paper, we deal with the data-driven selection of multidimensional and possibly anisotropic bandwidths in the general framework of kernel empirical risk minimization. We propose a universal selection rule, which leads to optimal adaptive results in a large variety of statistical models such as nonparametric robust regression and statistical learning with errors in variables. These results are stated in the context of smooth loss functions, where the gradient of the risk appears as a good criterion to measure the performance of our estimators. The selection rule consists of a comparison of gradient empirical risks. It can be viewed as a nontrivial improvement of the so-called Goldenshluger–Lepski method to nonlinear estimators. Furthermore, one main advantage of our selection rule is the nondependency on the Hessian matrix of the risk, usually involved in standard adaptive procedures.

1. Introduction. We consider the minimization problem of an unknown risk function $R:\mathbb{R}^m \to \mathbb{R}$, where $m \geq 1$ is the dimension of the statistical model. We assume the existence of a risk minimizer

$$\theta^\star \in \arg\min_{\theta\in\mathbb{R}^m} R(\theta), \tag{1.1}$$

where the risk function corresponds to the expectation of an appropriate loss function w.r.t. an unknown distribution. In empirical risk minimization, this quantity is usually estimated by its empirical version from an i.i.d. sample. However, in many problems such as local M-estimation or errors-in-variables models, a nuisance parameter can be involved in the empirical version. This parameter most often coincides with some bandwidth related to a kernel that gives rise to "kernel empirical risk minimization." One typically deals with this issue in pointwise estimation, as, for example, in Polzehl and Spokoiny [41] with localized likelihoods or in Chichignoud and Lederer [9] with local M-estimators. In learning theory, many authors have recently investigated supervised and unsupervised learning with errors in variables. As a rule, such matters require one to plug deconvolution kernels into the empirical risk, as Loustau and Marteau [32] in noisy discriminant analysis

Received January 2014; revised January 2015.

[1]Supported in part as member of the German–Swiss Research Group FOR916 (Statistical Regularization and Qualitative Constraints) with grant number 20PA20E-134495/1.

MSC2010 subject classifications. Primary 62G05, 62G20; secondary 62G08, 62H30.

Key words and phrases. Adaptivity, bandwidth selection, ERM, robust regression, statistical learning, errors-in-variables.

or Hall and Lahiri [17] in quantile and moment estimation; see also Dattner, Reiss and Trabs [12].

In the above papers, the authors studied the theoretical properties of kernel empirical risk minimizers and proposed deterministic choices of bandwidths to deduce optimal minimax results. As usual, these optimal bandwidths are related to the smoothness of the target function or the underlying density and are not achievable in practice. Adaptivity is therefore one of the biggest challenges. In this respect, data-driven bandwidth selections have been already proposed in [9, 10, 12, 41], which are all based on Lepski-type procedures.

Lepski-type procedures are rather appropriate to construct data-driven bandwidths involved in kernels; for further details, see, for example, [21, 28, 29]. It is well known that they suffer from the restriction to isotropic bandwidths with multidimensional data, which is the consideration of nested neighborhoods (hyper-cube). Many improvements were made by Kerkyacharian, Lepski and Picard [24] and more recently by Goldenshluger and Lepski [15] to select anisotropic bandwidths (hyper-rectangle). Nevertheless, their approach still does not provide anisotropic bandwidth selection for nonlinear estimators, which is the scope of this paper. The only work we can mention is [9] in a restrictive case, which is pointwise estimation in nonparametric regression. Therefore, the study of data-driven selection of anisotropic bandwidths is still an open issue. Moreover, this field is of great interest in practice, especially in image denoising; see, for example, [2, 22].

The main contribution of our paper is to bring new insights to the problem of bandwidth selection in kernel empirical risk minimization in a possible anisotropic framework. To this end, we first introduce a new criterion called *gradient excess risk*, which makes the anisotropic bandwidth selection possible. We then provide a novel data-driven selection based on the comparison of "Gradient empirical risks." That can be viewed as an extension of the so-called Goldenshluger–Lepski method (GL method; see [15]) and of the empirical risk comparison method (ERC method; see [10]). Eventually, we derive an upper bound for the gradient excess risk (called gradient inequality) and optimal results in many settings, such as pointwise and global estimation in nonparametric regression and clustering with errors in variables.

Note that we consider the risk minimization over the finite dimensional set $\mathbb{R}^m$. In statistical learning or nonparametric estimation, one usually aims at estimating a functional object belonging to some Hilbert space. However, in many examples, the target function can be approximated by a finite object, thanks, for instance, to a suitable decomposition in a basis of the Hilbert space. This is typically the case in local M-estimation, where the target function is assumed to be locally polynomial (and even constant in many cases). Moreover, in statistical learning, one is often interested in the estimation of a finite number of parameters, as in clustering. The extension to the infinite-dimensional case is discussed in Section 5.

The structure of this paper is as follows: the main ideas behind the gradient excess risk are introduced in the remainder of this section. An upper bound for the gradient excess risk of the data-driven procedure is presented in Section 2. This procedure is applied to clustering in Section 3 and to robust nonparametric regression in Section 4. Additionally, a discussion of our assumptions and an outlook are given in Section 5, and Section 6 illustrates the behavior of the method with numerical results. The proofs are finally conducted in the Appendix.

1.1. *The gradient excess risk approach.* In the literature, such as in statistical learning, the excess risk $R(\widehat{\theta}) - R(\theta^\star)$ is the main criterion to measure the performance of some estimator $\widehat{\theta}$. Originally, Vapnik and Chervonenkis [46] proposed to control this quantity via the empirical process theory, which gives rise to slow rates $\mathcal{O}(n^{-1/2})$ for the excess risk; see also [45]. In the last decade, many authors have improved such a bound by giving fast rates $\mathcal{O}(n^{-1})$ using the so-called localization technique; see [4, 26, 34, 36, 37, 43] and Boucheron, Bousquet and Lugosi [5] for an overview in classification. This technique consists of studying the increments of an empirical process in the neighborhood of the target $\theta^\star$. In particular, it requires a variance-risk correspondence, equivalent to the eminent margin assumption. As far as we know, this complicated modus operandi is the major obstacle to the anisotropic bandwidth selection issue. In what follows, we introduce an alternative criterion to solve this issue, namely the gradient excess risk (G-excess risk, for short, in the sequel). This quantity is defined as

$$|G(\widehat{\theta}, \theta^\star)|_2 := |G(\widehat{\theta}) - G(\theta^\star)|_2 \qquad \text{where } G := \nabla R, \tag{1.2}$$

whereas $|\cdot|_2$ denotes the Euclidean norm on $\mathbb{R}^m$ and $\nabla R : \mathbb{R}^m \to \mathbb{R}^m$ denotes the gradient of the risk R. With a slight abuse of notation, G denotes the gradient, whereas $G(\cdot, \theta^\star)$ denotes the G-excess risk. Under regularity assumptions on $R(\cdot)$, the G-excess risk is linked with the excess risk, thanks to the following lemma.

LEMMA 1. *Let $\theta^\star$, defined as in (1.1), and U be the Euclidean ball of $\mathbb{R}^m$ centered at $\theta^\star$, with radius $\delta > 0$. Assume $\theta \mapsto R(\theta)$ is $\mathcal{C}^2(U)$, each second partial derivative of R is bounded on U by a constant κ_1 and the Hessian matrix $H_R(\cdot)$ is positive definite at $\theta^\star$. Then, for $\delta > 0$ small enough, we have*

$$\sqrt{R(\theta) - R(\theta^\star)} \le 2\frac{\sqrt{m\kappa_1}}{\lambda_{\min}} |G(\theta, \theta^\star)|_2 \qquad \forall \theta \in U,$$

where $\lambda_{\min}$ is the smallest eigenvalue of $H_R(\theta^\star)$.

The proof is based on the inverse function theorem and a Taylor expansion of the function $R(\cdot)$. Let us explain how this lemma, together with standard probabilistic tools, leads to fast rates for the excess risk. In this section, $\widehat{R}$ denotes the usual empirical risk with associated gradient $\widehat{G} := \nabla \widehat{R}$ and associated ERM $\widehat{\theta}$ for ease of exposition. Under the assumptions of Lemma 1, $G(\theta^\star) = \widehat{G}(\widehat{\theta}) = (0, \dots, 0)^\top$, and we have the following heuristic:

$$\sqrt{R(\widehat{\theta}) - R(\theta^\star)} \lesssim \big|G(\widehat{\theta}, \theta^\star)\big|_2 = \big|G(\widehat{\theta}) - \widehat{G}(\widehat{\theta})\big|_2 \\ \leq \sup_{\theta \in \mathbb{R}^m} \big|G(\theta) - \widehat{G}(\theta)\big|_2 \lesssim n^{-1/2}, \tag{1.3}$$

where $\lesssim$ denotes the inequality up to some positive constant. The last bound only requires a concentration inequality applied to the empirical process $\widehat{G}(\cdot) - G(\cdot)$. Therefore, this heuristic provides fast rates for the excess risk without any localization technique. Furthermore, similar bounds can be obtained for the ℓ_2-norm $|\widehat{\theta} - \theta^\star|_2$ using the same path. Indeed, under the same assumptions, the assertion of Lemma 1 holds, replacing the square root of the excess risk by $|\widehat{\theta} - \theta^\star|_2$ (see the proof of Lemma 1), and then optimal rates are deduced.

From the model selection point of view, standard penalization techniques—based on localization—suffer from the dependency on parameters involved in the margin assumption. More precisely, in the strong margin assumption framework, the construction of the penalty requires the knowledge of $\lambda_{\min}$, related to the Hessian matrix of the risk. Although many authors have recently investigated the adaptivity w.r.t. these parameters, by proposing "margin-adaptive" procedures (see [41] for the propagation method, [27] for aggregation and [3] for the slope heuristic), the theory is not completed and remains a hard issue; see the related discussion in Section 5. As an alternative, our data-driven procedure does not suffer from the dependency on $\lambda_{\min}$ since we focus on a gradient inequality in Section 2.

1.2. *Kernel empirical risk minimization.* In this section, the kernel empirical risk minimization is properly defined and illustrated with two examples: local M-estimators and deconvolution k-means. For some $p \in \mathbb{N}^\star$, consider a $\mathbb{R}^p$-random variable Z distributed according to P, absolutely continuous w.r.t. the Lebesgue measure. In what follows, we observe a sample $\mathcal{Z}_n := \{Z_1, \dots, Z_n\}$ of independent and identically distributed (i.i.d.) random variables according to P. Moreover, we call a kernel of order $r \in \mathbb{N}^\star$ a symmetric function $K : \mathbb{R}^d \to \mathbb{R}$, $d \geq 1$, which satisfies the following properties:

- $\int_{\mathbb{R}^d} K(x)\,dx = 1$,
- $\int_{\mathbb{R}^d} K(x) x_j^k\,dx = 0\ \forall k \leq r, \forall j \in \{1, \dots, d\}$,
- $\int_{\mathbb{R}^d} |K(x)||x_j|^r\,dx < \infty, \forall j \in \{1, \dots, d\}$.

For any $h \in \mathcal{H} \subset \mathbb{R}_+^d$, the dilation K_h is defined as

$$K_h(x) = \Pi_h^{-1} K(x_1/h_1, \dots, x_d/h_d) \qquad \forall x \in \mathbb{R}^d,$$

where $\Pi_h := \prod_{j=1}^d h_j$. For a given kernel K, we define the kernel empirical risk indexed by an anisotropic bandwidth $h \in \mathcal{H} \subset (0, 1]^d$ as

$$\widehat{R}_h(\theta) := \frac{1}{n} \sum_{i=1}^n \ell_{K_h}(Z_i, \theta), \tag{1.4}$$

and an associated kernel empirical risk minimizer (kernel ERM) as

$$\widehat{\theta}_h \in \arg\min_{\theta \in \mathbb{R}^m} \widehat{R}_h(\theta). \tag{1.5}$$

The function $\ell_{K_h} : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}_+$ is a loss function associated to a kernel K_h such that $\theta \mapsto \ell_{K_h}(Z, \theta)$ is twice differentiable P-almost surely and such that the limit of its expectation coincides with the risk, that is,

$$\lim_{h \to (0,\dots,0)} \mathbb{E}\widehat{R}_h(\theta) = R(\theta) \qquad \forall \theta \in \mathbb{R}^m, \tag{1.6}$$

where $\mathbb{E}$ denotes the expectation w.r.t. the distribution of the sample $\mathcal{Z}_n$.

The agenda is the data-driven selection of the "best" estimator in the family $\{\widehat{\theta}_h, h \in \mathcal{H}\}$. This issue arises in many examples, such as local fitted likelihood (Polzehl and Spokoiny [41]), image denoising (Astola et al. [22]) and robust nonparametric regression; see Chichignoud and Lederer [9]. In such a framework, we observe a sample of i.i.d. pairs $Z_i = (W_i, Y_i)_{i=1}^n$, and the kernel empirical risk has the following general form:

$$\frac{1}{n} \sum_{i=1}^n \ell_{K_h}(Z_i, \theta) = \frac{1}{n} \sum_{i=1}^n \rho(Z_i, \theta) K_h(W_i - x_0),$$

where $\rho(\cdot, \cdot)$ is some likelihood and $x_0 \in \mathbb{R}^d$. Another example arises when we observe a contaminated sample $Z_i = X_i + \varepsilon_i$, $i = 1, \dots, n$ in the problem of clustering. In this case, the kernel empirical risk is defined according to

$$\frac{1}{n} \sum_{i=1}^n \ell_{K_h}(Z_i, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} \min_{j=1,\dots,k} |x - c_j|_2^2 \widetilde{K}_h(Z_i - x)\, dx,$$

where $\widetilde{K}_h(\cdot)$ is a deconvolution kernel and $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk}$ is a codebook.

In the next section, we present the bandwidth selection rule in the general context of kernel empirical risk minimization. We especially deal with clustering with errors in variables and robust nonparametric regression in Sections 3 and 4, respectively.

2. Selection rule and gradient inequality. The anisotropic bandwidth selection issue has been recently investigated in Goldenshluger and Lepski [15] (GL method) in density estimation; see also [11] for deconvolution estimation and [13, 14] for the white noise model. This method, based on the comparison of estimators, requires some "linearity" property, which is trivially satisfied by kernel estimators. However, kernel ERMs are usually nonlinear (except for the least square estimator), and the GL method cannot be directly applied to such estimators.

To tackle this issue, we introduce a new selection rule based on the comparison of gradient empirical risks instead of estimators (i.e., kernel ERM). To that end, we first introduce some notations. For any $h \in \mathcal{H}$ and any $\theta \in \mathbb{R}^m$, the gradient empirical risk (G-empirical risk) is defined as

$$\widehat{G}_h(\theta) := \frac{1}{n}\sum_{i=1}^n \nabla \ell_{K_h}(Z_i,\theta) = \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial \theta_j}\ell_{K_h}(Z_i,\theta)\right)_{j=1,\dots,m}. \tag{2.1}$$

Note that we have coarsely $\widehat{G}_h(\widehat{\theta}_h) = (0,\dots,0)^\top$ since $\ell_{K_h}(Z_i,\cdot)$ is twice differentiable almost surely. According to (1.6), we also notice that the limit of the expectation of the G-empirical risk coincides with the gradient of the risk.

Following Goldenshluger and Lepski [15], we introduce an auxiliary G-empirical risk in the comparison. For any couple of bandwidths $(h,\eta) \in \mathcal{H}^2$ and any $\theta \in \mathbb{R}^m$, the auxiliary G-empirical risk is defined as

$$\widehat{G}_{h,\eta}(\theta) := \frac{1}{n}\sum_{i=1}^n \nabla \ell_{K_h * K_\eta}(Z_i,\theta), \tag{2.2}$$

where $K_h * K_\eta(\cdot) := \int_{\mathbb{R}^d} K_h(\cdot - x)K_\eta(x)\,dx$ stands for the convolution between K_h and K_η. The gradient inequality stated in Theorem 1 is based on the control of some random processes as follows.

DEFINITION 1 (Majorant). For any integer $l > 0$, we call *majorant* a function $\mathcal{M}_l : \mathcal{H}^2 \to \mathbb{R}_+$ such that

$$\mathbb{P}\Big(\sup_{\lambda,\eta\in\mathcal{H}}\{|\widehat{G}_{\lambda,\eta} - \mathbb{E}\widehat{G}_{\lambda,\eta}|_{2,\infty} + |\widehat{G}_\eta - \mathbb{E}\widehat{G}_\eta|_{2,\infty} - \mathcal{M}_l(\lambda,\eta)\}_+ > 0\Big) \le n^{-l},$$

where $|T|_{2,\infty} := \sup_{\theta\in\mathbb{R}^m}|T(\theta)|_2$ for all $T : \mathbb{R}^m \to \mathbb{R}^m$ with $|\cdot|_2$ the Euclidean norm on $\mathbb{R}^m$, and $\mathbb{E}$ is understood coordinatewise.

The main issue for applications is to compute right order majorants. It follows from classical tools such as Talagrand's inequalities (Talagrand [42], Boucheron, Lugosi and Massart [6], Bousquet [7]; see also [16]). In Sections 3 and 4 such majorant functions are computed in clustering and in robust nonparametric regression.

We are now ready to define the selection rule as

$$\widehat{h} \in \arg\min_{h\in\mathcal{H}} \widehat{\mathrm{BV}}(h), \tag{2.3}$$

where $\widehat{\mathrm{BV}}(h)$ is an estimate of the bias–variance decomposition at a given bandwidth $h \in \mathcal{H}$. It is explicitly defined as

$$\widehat{\mathrm{BV}}(h) := \sup_{\eta\in\mathcal{H}}\{|\widehat{G}_{h,\eta} - \widehat{G}_\eta|_{2,\infty} - \mathcal{M}_l(h,\eta)\} + \mathcal{M}_l^\infty(h)$$

$$\text{with } \mathcal{M}_l^\infty(h) := \sup_{\lambda\in\mathcal{H}} \mathcal{M}_l(\lambda,h).$$

The kernel ERM $\widehat{\theta}_{\widehat{h}}$, defined in (1.5), with bandwidth $\widehat{h}$, selected in (2.3), satisfies the following bound.

THEOREM 1 (Gradient inequality). *For any $n \in \mathbb{N}^\star$ and for any $l \in \mathbb{N}^\star$, we have with probability* $1 - n^{-l}$,

$$\left|G\left(\widehat{\theta}_{\widehat{h}}, \theta^\star\right)\right|_2 \leq 3 \inf_{h \in \mathcal{H}} \left\{B(h) + \mathcal{M}_l^\infty(h)\right\},$$

where $B : \mathcal{H} \to \mathbb{R}_+$ *is a bias function defined as*

$$B(h) := \max\left(|\mathbb{E}\widehat{G}_h - G|_{2,\infty}, \sup_{\eta \in \mathcal{H}} |\mathbb{E}\widehat{G}_{h,\eta} - \mathbb{E}\widehat{G}_\eta|_{2,\infty}\right) \qquad \forall h \in \mathcal{H}. \tag{2.4}$$

Theorem 1 is the main result of this paper. The G-excess risk of the data-driven estimator $\widehat{\theta}_{\widehat{h}}$ is bounded with high probability. The RHS in the gradient inequality can be viewed as the minimization of a usual bias–variance trade-off. Indeed, the bias term $B(h)$ is deterministic and tends to 0 as $h \to (0, \dots, 0)$. The majorant $\mathcal{M}_l^\infty(h)$ upper bounds the stochastic part of the G-empirical risk and can be viewed as a variance term.

The gradient inequality of Theorem 1 is sufficient to establish adaptive fast rates in noisy clustering and adaptive minimax rates in nonparametric estimation; see Sections 3 and 4. Moreover, the construction of the selection rule (2.3), as well as the upper bound in Theorem 1, does not suffer from the dependency on $\lambda_{\min}$ related to the smallest eigenvalue of the Hessian matrix of the risk; see Lemma 1. In other words, the method is robust w.r.t. this parameter, which is a major improvement in comparison with other adaptive or model selection methods of the literature cited in the Introduction.

PROOF OF THEOREM 1. For some $h \in \mathcal{H}$, we start with the following decomposition:

$$\begin{aligned} \left|G\left(\widehat{\theta}_{\widehat{h}}, \theta^\star\right)\right|_2 &= \left|(\widehat{G}_{\widehat{h}} - G)(\widehat{\theta}_{\widehat{h}})\right|_2 \leq |\widehat{G}_{\widehat{h}} - G|_{2,\infty} \\ &\leq |\widehat{G}_{\widehat{h}} - \widehat{G}_{\widehat{h},h}|_{2,\infty} + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} + |\widehat{G}_h - G|_{2,\infty}. \end{aligned} \tag{2.5}$$

By definition of $\widehat{h}$ in (2.3), the first two terms in the RHS of (2.5) are bounded as follows:

$$\begin{aligned} &|\widehat{G}_{\widehat{h}} - \widehat{G}_{\widehat{h},h}|_{2,\infty} + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} \\ &\quad = |\widehat{G}_{h,\widehat{h}} - \widehat{G}_{\widehat{h}}|_{2,\infty} - \mathcal{M}_\ell(h, \widehat{h}) + \mathcal{M}_\ell(h, \widehat{h}) \\ &\qquad + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} - \mathcal{M}_\ell(\widehat{h}, h) + \mathcal{M}_\ell(\widehat{h}, h) \\ &\quad \leq \sup_{\eta \in \mathcal{H}} \left\{|\widehat{G}_{h,\eta} - \widehat{G}_\eta|_{2,\infty} - \mathcal{M}_\ell(h, \eta)\right\} + \mathcal{M}_\ell^\infty(h) \\ &\qquad + \sup_{\eta \in \mathcal{H}} \left\{|\widehat{G}_{\widehat{h},\eta} - \widehat{G}_\eta|_{2,\infty} - \mathcal{M}_\ell(\widehat{h}, \eta)\right\} + \mathcal{M}_\ell^\infty(\widehat{h}) \\ &\quad = \widehat{\mathrm{BV}}(h) + \widehat{\mathrm{BV}}(\widehat{h}) \leq 2\widehat{\mathrm{BV}}(h). \end{aligned} \tag{2.6}$$

Besides, the last term in (2.5) is controlled as follows:

$$\begin{aligned}|\widehat{G}_h - G|_{2,\infty} &\le |\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty} + |\mathbb{E}\widehat{G}_h - G|_{2,\infty} \\ &\le |\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty} - \mathcal{M}_l(\lambda, h) + \mathcal{M}_l(\lambda, h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty} \\ &\le \sup_{\lambda,\eta}\{|\widehat{G}_{\lambda,\eta} - \mathbb{E}\widehat{G}_{\lambda,\eta}|_{2,\infty} + |\widehat{G}_\eta - \mathbb{E}\widehat{G}_\eta|_{2,\infty} - \mathcal{M}_l(\lambda, \eta)\} \\ &\quad + \mathcal{M}_l^\infty(h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty} \\ &=: \zeta + \mathcal{M}_l^\infty(h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty}.\end{aligned}$$

Using (2.5) and (2.6), together with the last inequality, we have for all $h \in \mathcal{H}$,

$$\left|G(\widehat{\theta}_{\widehat{h}}, \theta^\star)\right|_2 \le 2\widehat{\mathrm{BV}}(h) + \zeta + \mathcal{M}_l^\infty(h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty}. \tag{2.7}$$

It then remains to control the term $\widehat{\mathrm{BV}}(h)$. We have

$$\begin{aligned}&\widehat{\mathrm{BV}}(h) - \mathcal{M}_l^\infty(h) \\ &\quad\le \sup_{\lambda,\eta}\{|\widehat{G}_{\lambda,\eta} - \mathbb{E}\widehat{G}_{\lambda,\eta}|_{2,\infty} + |\widehat{G}_\eta - \mathbb{E}\widehat{G}_\eta|_{2,\infty} - \mathcal{M}_l(\lambda, \eta)\} \\ &\qquad + \sup_\eta |\mathbb{E}\widehat{G}_{h,\eta} - \mathbb{E}\widehat{G}_\eta|_{2,\infty} \\ &\quad= \zeta + \sup_\eta |\mathbb{E}\widehat{G}_{h,\eta} - \mathbb{E}\widehat{G}_\eta|_{2,\infty}.\end{aligned}$$

The gradient inequality follows directly from (2.7), Definition 1 and the definition of ζ. □

3. Application to noisy clustering. Let us consider an integer $k \ge 1$ and a $\mathbb{R}^d$-random variable X with law P with density f w.r.t. the Lebesgue measure on $\mathbb{R}^d$ satisfying $\mathbb{E}_P|X|_2^2 < \infty$, where $|\cdot|_2$ stands for the Euclidean norm in $\mathbb{R}^d$. Moreover, we restrict the study to the compact set $[0, 1]^d$, assuming that $X \in [0, 1]^d$ almost surely. We want to construct k centroids minimizing some distortion,

$$\mathcal{W}(\mathbf{c}) := \mathbb{E}_P w(\mathbf{c}, X), \tag{3.1}$$

where $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{d\times k}$ is a candidate codebook of k centroids. For ease of exposition, we study this quantization problem with the Euclidean distance, by choosing the standard k-means loss function, namely,

$$w(\mathbf{c}, x) = \min_{j=1,\dots,k} |x - c_j|_2^2, \qquad x \in \mathbb{R}^d.$$

In this section, we are interested in the inverse statistical learning context (see [31]), which corresponds to the minimization of (3.1), thanks to a noisy set of observations,

$$Z_i = X_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

where $(\varepsilon_i)_{i=1}^n$ are i.i.d. with density g w.r.t. the Lebesgue measure on $\mathbb{R}^d$ and mutually independent of the original sample $(X_i)_{i=1}^n$. This topic was first considered in [8], where general oracle inequalities are proposed. Let us fix a kernel K_h of order $r \in \mathbb{N}^\star$ with $h \in \mathcal{H}$ and consider $\widetilde{K}_h$ a deconvolution kernel defined such that $\mathcal{F}[\widetilde{K}_h] = \mathcal{F}[K_h]/\mathcal{F}[g]$, where $\mathcal{F}$ stands for the usual Fourier transform. As introduced in Section 1.2, we have at our disposal the family of kernel ERM defined as

$$(3.2) \qquad \widehat{\mathbf{c}}_h \in \arg\min_{\mathbf{c}\in\mathbb{R}^{dk}} \widehat{\mathcal{W}}_h(\mathbf{c}) \qquad \text{where } \widehat{\mathcal{W}}_h(\mathbf{c}) := \frac{1}{n}\sum_{i=1}^n w(\mathbf{c},\cdot) * \widetilde{K}_h(Z_i - \cdot),$$

where $f * g(\cdot) := \int_{[0,1]^d} f(x)g(\cdot - x)\,dx$ stands for the convolution product (restricted to the compact $[0,1]^d$ for simplicity). From an adaptive point of view, Chichignoud and Loustau [10] have recently investigated the problem of choosing the bandwidth in (3.2). They established fast rates of convergence—up to a logarithmic term—for a data-driven selection of h, based on a comparison of kernel empirical risks. However, their approach is restricted to isotropic bandwidth selection and depends on the parameters involved in the margin assumption (in particular on $\lambda_{\min}$ in Lemma 1).

In the following, adaptive fast rates of convergence for the excess risk are obtained via the gradient approach. For this purpose, we assume that the Hessian matrix $H_{\mathcal{W}}$ is positive definite. This assumption was considered for the first time in Pollard [39] and is often referred as Pollard's regularity assumptions; see also [30]. Under these assumptions, we can state the same kind of result as Lemma 1 in the framework of clustering with k-means.

LEMMA 2. *Let $\mathbf{c}^\star$ be a minimizer of* (3.1), *and assume f is continuous and $H_{\mathcal{W}}(\mathbf{c}^\star)$ is positive definite. Let U be the Euclidean ball center at $\mathbf{c}^\star$ with radius $\delta > 0$. Then, for δ sufficiently small*,

$$\sqrt{\mathcal{W}(\mathbf{c}) - \mathcal{W}(\mathbf{c}^\star)} \leq C\left|\nabla\mathcal{W}(\mathbf{c}) - \nabla\mathcal{W}(\mathbf{c}^\star)\right|_2 \qquad \forall \mathbf{c} \in U,$$

where $C > 0$ is a constant which depends on $H_{\mathcal{W}}(\mathbf{c}^\star)$, k and d.

We have at our disposal a family of kernel ERM $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$ with associated kernel empirical risk $\widehat{\mathcal{W}}_h(\cdot)$ defined in (3.2). We propose to apply the selection rule (2.3) to choose the bandwidth $h \in \mathcal{H}$. In this problem as well, we first consider the G-excess risk approach to establish adaptive fast rates of convergence for the excess risk. For any $h \in \mathcal{H}$, the G-empirical risk vector of $\mathbb{R}^{dk}$ is given by

$$\begin{aligned}\nabla\widehat{\mathcal{W}}_h(\mathbf{c}) &:= \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial c_j^u}\int_{[0,1]^d} w(\mathbf{c},x)\widetilde{K}_h(Z_i - x)\,dx\right)_{(u,j)\in\{1,\dots,d\}\times\{1,\dots,k\}} \\ &= \left(-\frac{1}{n}\sum_{i=1}^n 2\int_{V_j(\mathbf{c})}(x^u - c_j^u)\widetilde{K}_h(Z_i - x)\,dx\right)_{(u,j)\in\{1,\dots,d\}\times\{1,\dots,k\}},\end{aligned}$$

where x^u denotes the uth coordinate of $x \in \mathbb{R}^d$ and $V_j(\mathbf{c})$, $j = 1, \dots, k$ are open Voronoï cells associated to $\mathbf{c}$, defined as $V_j(\mathbf{c}) = \{x \in [0, 1]^d : \forall u \neq j, |x - c_j|_2 < |x - c_u|_2\}$. Note that $\nabla \widehat{\mathcal{W}}_h(\widehat{\mathbf{c}}_h) = (0, \dots, 0)^\top$ a.s. by smoothness. The construction of the rule follows the general case of Section 2, which requires the introduction of an auxiliary G-empirical risk. For any couple of bandwidths $(h, \eta) \in \mathcal{H}^2$, the auxiliary G-empirical risk is defined as

$$\nabla \widehat{\mathcal{W}}_{h,\eta}(\mathbf{c}) := \left(-\frac{1}{n} \sum_{i=1}^n 2 \int_{V_j} (x^u - c_j^u) \widetilde{K}_{h,\eta}(Z_i - x)\, dx \right)_{(u,j)\in\{1,\dots,d\}\times\{1,\dots,k\}} \in \mathbb{R}^{dk},$$

where $\widetilde{K}_{h,\eta} = \widetilde{K_h * K_\eta}$ is the auxiliary deconvolution kernel as in Comte and Lacour [11].

The statement of the oracle inequality is based on the computation of a majorant function. For this purpose, we need the following additional assumptions on the kernel $K \in \mathbb{L}_2(\mathbb{R}^d)$.

(**K1**) There exists $S = (S_1, \dots, S_d) \in \mathbb{R}_+^d$ such that the kernel K satisfies

$$\operatorname{supp} \mathcal{F}[K] \subset [-S, S] \quad \text{and} \quad \sup_{t \in \mathbb{R}^d} \left| \mathcal{F}[K](t) \right| < \infty,$$

where $\operatorname{supp} g = \{x : g(x) \neq 0\}$ and $[-S, S] = \bigotimes_{v=1}^d [-S_v, S_v]$.

This assumption is standard in deconvolution estimation and is satisfied by many standard kernels, such as the *sinc* kernel.

We also consider a kernel K of order $r \in \mathbb{N}^\star$, according to the definition of Section 1.2. Kernels of order r satisfying (**K1**) could be constructed by using the so-called Meyer wavelet; see [33]. Additionally, we need an assumption on the noise distribution g:

Noise assumption NA(ρ, β). There exist a vector $\beta = (\beta_1, \dots, \beta_d) \in (0, \infty)^d$ and a positive constant ρ such that for all $t \in \mathbb{R}^d$,

$$\left| \mathcal{F}[g](t) \right| \geq \rho \prod_{j=1}^d \left(\frac{t_j^2 + 1}{2} \right)^{-\beta_j/2}.$$

NA(ρ, β**)** deals with a polynomial behavior of the Fourier transform of the noise density g. An exponential decreasing of the characteristic function of g is not considered in this paper for simplicity; see [11] in multivariate deconvolution for such a study.

We are now ready to compute some majorant functions in our context. For some $s^+ > 0$, let $\mathcal{H} := [h_-, h^+]^d$ be the bandwidth set such that $0 < h_- < h^+ < 1$,

$$h_- := \left(\frac{\log^6(n)}{n} \right)^{1/(2 \vee 2 \sum_{j=1}^d \beta_j)} \quad \text{and} \quad h^+ := \left(1/\log(n)\right)^{1/(2s^+)}. \tag{3.3}$$

LEMMA 3. *Assume* (**K1**) *and* **NA**(ρ,β) *hold for some* $\rho>0$ *and some* $\beta\in\mathbb{R}_+^d$. *Let* $a\in(0,1)$, *and consider* $\mathcal{H}_a:=\{(h_-,\dots,h_-)\}\cup\{h\in\mathcal{H}:\forall j=1,\dots,d\ \exists m_j\in\mathbb{N}:h_j=h^+a^{m_j}\}$ *an exponential net of* $\mathcal{H}=[h_-,h^+]^d$, *such that* $|\mathcal{H}_a|\le n$. *For any integer* $l>0$, *let us introduce the function* $\mathcal{M}_l^{\mathrm{k}}:\mathcal{H}^2\to\mathbb{R}_+$ *defined as*

$$\mathcal{M}_l^{\mathrm{k}}(h,\eta):=b_1'\sqrt{kd}\left(\frac{\prod_{i=1}^d\eta_i^{-\beta_i}}{\sqrt{n}}+\frac{\prod_{i=1}^d(h_i\vee\eta_i)^{-\beta_i}}{\sqrt{n}}\right),$$

where $b_1':=b_1'(l)>0$ *is linear in* l *and independent of* n; *see the Appendix for details.*

Then, for n sufficiently large, the function $\mathcal{M}_l^{\mathrm{k}}(\cdot,\cdot)$ *is a majorant, that is,*

$$\mathbb{P}\Big(\sup_{h,\eta\in\mathcal{H}_a}\{|\nabla\widehat{\mathcal{W}}_{h,\eta}-\mathbb{E}\nabla\widehat{\mathcal{W}}_{h,\eta}|_{2,\infty}+|\nabla\widehat{\mathcal{W}}_{\eta}-\mathbb{E}\nabla\widehat{\mathcal{W}}_{\eta}|_{2,\infty}-\mathcal{M}_l^{\mathrm{k}}(h,\eta)\}_+>0\Big)\le n^{-l},$$

where $\mathbb{E}$ *denotes the expectation w.r.t. to the sample and* $|T|_{2,\infty}=\sup_{\mathbf{c}\in[0,1]^{dk}}|T(\mathbf{c})|_2$ *for all* $T:\mathbb{R}^{dk}\to\mathbb{R}^{dk}$ *with* $|\cdot|_2$ *the Euclidean norm on* $\mathbb{R}^{dk}$.

The proof is based on a Talagrand inequality; see the Appendix. This lemma is the cornerstone and gives the order of the variance term in such a problem.

We are now ready to define the selection rule in this setting as

$$\widehat{h}\in\arg\min_{h\in\mathcal{H}_a}\Big\{\sup_{\eta\in\mathcal{H}_a}\{|\nabla\widehat{\mathcal{W}}_{h,\eta}-\nabla\widehat{\mathcal{W}}_{\eta}|_{2,\infty}-\mathcal{M}_l^{\mathrm{k}}(h,\eta)\}+\mathcal{M}_l^{\mathrm{k},\infty}(h)\Big\},\tag{3.4}$$

where $\mathcal{M}_l^{\mathrm{k},\infty}(h):=\sup_{\lambda\in\mathcal{H}_a}\mathcal{M}_l^{\mathrm{k}}(\lambda,h)$ and $\mathcal{H}_a$ is defined in Lemma 3. Eventually, we need an additional assumption on the regularity of the density f to control the bias term in Theorem 2. The regularity is expressed in terms of anisotropic Nikol'skii class.

DEFINITION 2 (Anisotropic Nikol'skii space). Let $s=(s_1,s_2,\dots,s_d)\in\mathbb{R}_+^d$, $q\in[1,\infty[$ and $L>0$ be fixed. We say that $f:[0,1]^d\to[-L,L]$ belongs to the anisotropic Nikol'skii class $\mathcal{N}_{q,d}(s,L)$ if for all $j=1,\dots,d$, $z\in\mathbb{R}$ and for all $x\in(0,1]^d$,

$$\left(\int\left|\frac{\partial^{\lfloor s_j\rfloor}}{\partial x_j^{\lfloor s_j\rfloor}}f(x_1,\dots,x_j+z,\dots,x_d)-\frac{\partial^{\lfloor s_j\rfloor}}{\partial x_j^{\lfloor s_j\rfloor}}f(x_1,\dots,x_j,\dots,x_d)\right|^q dx\right)^{1/q}\le L|z|^{s_j-\lfloor s_j\rfloor},$$

and $\|\frac{\partial^l}{\partial x_j^l}f\|_q\le L$, for any $l=0,\dots,\lfloor s_j\rfloor$, where $\lfloor s_j\rfloor$ is the largest integer strictly less than s_j.

Nikol'skii classes were introduced in approximation theory by Nikol'skii; see [38], for example. We also refer to [15, 24] where the problem of adaptive estimation has been treated for the Gaussian white noise model and for density estimation, respectively.

In the sequel, we assume that the multivariate density f belongs to the anisotropic Nikol'skii class $\mathcal{N}_{2,d}(s,L)$, for some $s \in \mathbb{R}_+^d$ and some $L > 0$. In other words, the density has possible different regularities in all directions. The statement of a nonadaptive upper bound for the excess risk in the anisotropic case has been already investigated in [10]. In the following theorem, we propose the adaptive version of the previous cited result, where the bandwidth $\widehat{h}$ is chosen via the selection rule (3.4).

THEOREM 2. *Assume* (**K1**) *and* $\mathbf{NA}(\rho, \beta)$ *hold for some* $\rho > 0$ *and some* $\beta \in \mathbb{R}_+^d$. *Assume the Hessian matrix of* $\mathcal{W}$ *is positive definite for any* $\mathbf{c}^\star \in \mathcal{M}$. *Then, for any* $s \in (0, s^+]^d$, *any* $L > 0$, *we have*

$$\limsup_{n \to \infty} n^{1/(1+\sum_{j=1}^d \beta_j/s_j)} \sup_{f \in \mathcal{N}_{2,d}(s,L)} \big[\mathbb{E}\mathcal{W}(\widehat{\mathbf{c}}_{\widehat{h}}) - \mathcal{W}(\mathbf{c}^\star)\big] < \infty,$$

where $\widehat{h}$ *is driven in* (3.4).

This theorem is a direct application of Theorem 1, Lemma 2 and the majorant construction. It gives adaptive fast rates of convergence for the excess risk of $\widehat{\mathbf{c}}_{\widehat{h}}$ and significantly improves the result stated in [10] for two reasons: first, the selection rule allows the extension to the anisotropic case; besides, there is no logarithmic term in the adaptive rate. In our opinion, the localization technique used in [10] seems to be the major obstacle to avoid the extra $\log n$ term.

4. Application to robust nonparametric regression. In this section, we apply the gradient inequality to the framework of local M-estimation in nonparametric robust regression. It will give adaptive minimax results for nonlinear estimators for both pointwise and global estimation.

Let us specify the model beforehand. For some $n \in \mathbb{N}^\star$, we observe a training set $\mathcal{Z}_n := \{(W_i, Y_i), i = 1, \dots, n\}$ of i.i.d. pairs, distributed according to the probability measure P on $[0,1]^d \times \mathbb{R}$ satisfying the set of equations

$$Y_i = f^\star(W_i) + \xi_i, \qquad i = 1, \dots, n. \tag{4.1}$$

We aim at estimating the target function $f^\star : [0,1]^d \to [-B, B]$, $B > 0$. The noise variables $(\xi_i)_{i \in 1,\dots,n}$ are assumed to be i.i.d. with symmetric density g_ξ w.r.t. the Lebesgue measure. We also assume g_ξ is continuous at 0 and $g_\xi(0) > 0$. For simplicity, the design points $(W_i)_{i=1}^n$ are assumed to be i.i.d. according to the uniform law on $[0,1]^d$ (extension to a more general design is straightforward), and we assume that $(W_i)_{i=1}^n$ and $(\xi_i)_{i=1}^n$ are mutually independent for ease of exposition.

Eventually, we restrict the estimation of $f^\star$ to the closed set $\mathcal{T} \subset [0,1]^d$ to avoid discussion on boundary effects. We will consider a point $x_0 \in \mathcal{T}$ for pointwise estimation and the $\mathbb{L}_q(\mathcal{T})$-risk for global estimation, for $q \in [1, +\infty)$.

Next, we introduce an estimate of $f^\star(x_0)$ at any $x_0 \in \mathcal{T}$ with the local constant approach (LCA). The key idea of LCA, as described, for example, in [44], Chapter 1, is to approximate the target function by a constant in a neighborhood of size $h \in (0,1)^d$ of a given point x_0, which corresponds to a model of dimension $m = 1$. To deal with heavy-tailed noises, we especially employ the Huber loss (see [19]) defined as follows. For any scale $\gamma > 0$ and $z \in \mathbb{R}$,

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

The parameter γ selects the level of robustness of the Huber loss between the square loss (large value of γ) and the absolute loss (small value of γ). Let $\mathcal{H} := [h_-, h^+]^d$ be the bandwidth set such that $0 < h_- < h^+ < 1$,

$$h_- := \frac{\log^{6/d}(n)}{n^{1/d}} \quad \text{and} \quad h^+ := \frac{1}{\log^2(n)}.$$

For any $x_0 \in \mathcal{T}$, the local estimator $\widehat{f}_h(x_0)$ of $f^\star(x_0)$ is defined as

$$\widehat{f}_h(x_0) := \operatorname*{arg\,min}_{t \in [-B,B]} \widehat{R}_h^{\text{loc}}(t), \qquad h \in \mathcal{H}, \tag{4.2}$$

where $\widehat{R}_h^{\text{loc}}(\cdot) := \frac{1}{n}\sum_{i=1}^n \rho_\gamma(Y_i - \cdot) K_h(W_i - x_0)$ is the local empirical risk, and K_h is a 1-Lipschitz kernel of order 1. We notice that the local empirical risk estimates the local risk $R^{\text{loc}}(\cdot) := \mathbb{E}_{Y|W=x_0} \rho_\gamma(Y - \cdot)$ whose $f^\star(x_0)$ is its unique minimizer.

In nonparametric estimation, one is usually interested in pointwise or global risk instead of excess risk. Since Theorem 1 controls the G-excess risk of the adaptive estimator, we present the following lemma that links the pointwise risk with the G-excess risk.

LEMMA 4. *Assume that* $\sup_{h\in\mathcal{H}} |\widehat{f}_h(x_0) - f^\star(x_0)| \le \mathbb{E}\rho''_\gamma(\xi_1)/4$ *holds. Then, for all* $h \in \mathcal{H}$,

$$|\widehat{f}_h(x_0) - f^\star(x_0)| \le \frac{2}{\mathbb{E}\rho''_\gamma(\xi_1)} \big|G^{\text{loc}}(\widehat{f}_h(x_0)) - G^{\text{loc}}(f^\star(x_0))\big|,$$

where G^{loc} (*and*, *resp*., ρ''_γ) *denotes the derivative of* R^{loc} (*resp*., *the second derivative of* ρ_γ).

The proof is given in the Appendix. We can also deduce the same inequality with the $\mathbb{L}_q(\mathcal{T})$-norm. The assumption $\sup_{h\in\mathcal{H}} |\widehat{f}_h(x_0) - f^\star(x_0)| \le \mathbb{E}\rho''_\gamma(\xi_1)/4$ is necessary to use the theory of differential calculus and can be satisfied by using the consistency of $\widehat{f}_h$. In this respect, the definitions of h_- and h^+ above imply the consistency of all estimators $\widehat{f}_h$, $h \in \mathcal{H}$; for further details, see below as well as [9], Theorem 1.

4.1. *The selection rule in pointwise estimation.* We now present the application of the selection rule for pointwise estimation. To compute the procedure, we define the G-empirical risk as

$$\widehat{G}_h^{\mathrm{loc}}(t) := \frac{\partial \widehat{R}_h^{\mathrm{loc}}}{\partial t}(t) = -\frac{1}{n}\sum_{i=1}^n \rho_\gamma'(Y_i - t) K_h(W_i - x_0). \tag{4.3}$$

For two bandwidths h, λ, we introduce the auxiliary G-empirical risk as

$$\widehat{G}_{h,\eta}^{\mathrm{loc}}(t) := -\frac{1}{n}\sum_{i=1}^n \rho_\gamma'(Y_i - t) K_{h,\eta}(W_i - x_0),$$

where $K_{h,\eta} := K_h * K_\eta$, as before.

To apply the results of Section 2, we need to compute optimal majorants of the associated empirical processes. The construction of such bounds for the pointwise case has already received attention in the literature; see [9], Proposition 2. For any integer $l \in \mathbb{N}^\star$, let us introduce the function $\mathcal{M}_l^{\mathrm{loc}} : \mathcal{H}^2 \to \mathbb{R}_+$ defined as

$$\mathcal{M}_l^{\mathrm{loc}}(h, \eta) := C_0 \|K\|_2 \sqrt{\mathbb{E}[\rho_\gamma'(\xi_1)]^2}\left(\sqrt{\frac{l \log(n)}{n \prod_{j=1}^d h_j \vee \eta_j}} + \sqrt{\frac{l \log(n)}{n \prod_{j=1}^d \eta_j}}\right),$$

where $C_0 > 0$ is an absolute constant which does not depend on the model. Then if we set $\mathcal{H}_a := \{(h_-, \dots, h_-)\} \cup \{h \in \mathcal{H} : \forall j = 1, \dots, d\ \exists m_j \in \mathbb{N} : h_j = h^+ a^{m_j}\}$, $a \in (0, 1)$, an exponential net of $\mathcal{H} = [h_-, h^+]^d$, such that $|\mathcal{H}_a| \le n$, for any $l > 0$, the function $\mathcal{M}_l^{\mathrm{loc}}(\cdot, \cdot)$ is a majorant according to Definition 1.

Eventually, we introduce the data-driven bandwidth following the schema of the selection rule in Section 2,

$$\widehat{h}^{\mathrm{loc}} \in \arg\min_{h \in \mathcal{H}_a}\Big\{\sup_{\eta \in \mathcal{H}_a}\{|\widehat{G}_{h,\eta}^{\mathrm{loc}} - \widehat{G}_\eta^{\mathrm{loc}}|_\infty - \mathcal{M}_l^{\mathrm{loc}}(h, \eta)\} + \mathcal{M}_l^{\mathrm{loc},\infty}(h)\Big\}, \tag{4.4}$$

where $\mathcal{M}_l^{\mathrm{loc},\infty}(h) := \sup_{h' \in \mathcal{H}_a} \mathcal{M}_l^{\mathrm{loc}}(h', h)$. To derive minimax adaptive rates for local estimation, we start with the definition of the anisotropic Hölder class.

DEFINITION 3 (Anisotropic Hölder class). Let $s = (s_1, s_2, \dots, s_d) \in \mathbb{R}_+^d$ and $L > 0$ be fixed. We say that $f : [0, 1]^d \to [-L, L]$ belongs to the anisotropic Hölder class $\Sigma(s, L)$ of functions if for all $j = 1, \dots, d$ and for all $x \in (0, 1]^d$,

$$\left|\frac{\partial^{\lfloor s_j \rfloor}}{\partial x_j^{\lfloor s_j \rfloor}} f(x_1, \dots, x_j + z, \dots, x_d) - \frac{\partial^{\lfloor s_j \rfloor}}{\partial x_j^{\lfloor s_j \rfloor}} f(x_1, \dots, x_j, \dots, x_d)\right| \le L|z|^{s_j - \lfloor s_j \rfloor} \qquad \forall z \in \mathbb{R},$$

and

$$\sup_{x \in [0,1]^d}\left|\frac{\partial^l}{\partial x_j^l} f(x)\right| \le L \qquad \forall l = 0, \dots, \lfloor s_j \rfloor,$$

where $\lfloor s_j \rfloor$ is the largest integer strictly less than s_j.

THEOREM 3. *For any $s \in (0,1]^d$, any $L > 0$ and any $q \geq 1$, it holds for all $x_0 \in \mathcal{T}$,*

$$\limsup_{n\to\infty} (n/\log(n))^{q\bar{s}/(2\bar{s}+1)} \sup_{f\in\Sigma(s,L)} \mathbb{E}\big|\widehat{f}_{\widehat{h}^{\mathrm{loc}}}(x_0) - f^\star(x_0)\big|^q < \infty,$$

where $\bar{s} := (\sum_{j=1}^d s_j^{-1})^{-1}$ denotes the harmonic average.

The proposed estimator $\widehat{f}_{\widehat{h}}$ is then adaptive minimax over anisotropic Hölder classes in pointwise estimation. The minimax optimality of this rate [with the $\log(n)$ factor] has been stated by [25] in the white noise model for pointwise estimation; see also [13]. For simplicity, we did not study the case of locally polynomial functions [i.e., $s \in (0,\infty)^d$].

Chichignoud and Lederer [9], Theorem 2, have shown that the variance of local M-estimators is of order $\mathbb{E}[\rho'_\gamma(\xi_1)]^2/n(\mathbb{E}\rho''_\gamma(\xi_1))^2$, and therefore their Lepski-type procedure depends on this quantity. Thanks to the gradient approach, we obtain the same result without the dependency on the parameter $\mathbb{E}\rho''_\gamma(\xi_1)$, which corresponds to $\lambda_{\min}$ in the general setting. The selection rule is therefore robust w.r.t. to the fluctuations of this parameter, in particular when γ is small (median estimator).

4.2. *The selection rule in global estimation.* The aim of this section is to derive adaptive minimax results for $\widehat{f}_h$ for the $\mathbb{L}_q$-risk. To this end, we need to modify the selection rule (4.4) including a global ($\mathbb{L}_q$-norm) comparison of G-empirical risks. For this purpose, for all $t \in \mathbb{R}$, we denote the G-empirical risks at a given point $x_0 \in \mathcal{T}$ as

$$\widehat{G}_h^{\mathrm{loc}}(t, x_0) = -\frac{1}{n}\sum_{i=1}^n \rho'_\gamma(Y_i - t) K_h(W_i - x_0)$$

and

$$\widehat{G}_{h,\eta}^{\mathrm{loc}}(t, x_0) = -\frac{1}{n}\sum_{i=1}^n \rho'_\gamma(Y_i - t) K_{h,\eta}(W_i - x_0),$$

where the dependence in x_0 is explicitly written. Then we define, for $q \in [1,\infty[$ and for any function $\omega : \mathbb{R}\times\mathcal{T} \to \mathbb{R}$, the $\mathbb{L}_q$-norm and $\mathbb{L}_{q,\infty}$-semi-norm

$$\|\omega(t,\cdot)\|_q := \left(\int_{\mathcal{T}} |\omega(t,x)|^q\,dx\right)^{1/q} \quad \text{and} \quad \|\omega\|_{q,\infty} := \sup_{t\in[-B,B]} \|\omega(t,\cdot)\|_q.$$

The construction of majorants is based on uniform bounds for $\mathbb{L}_q$-norms of empirical processes. Recently, Goldenshluger and Lepski investigated this topic [16], Theorem 2. For any integer $l \in \mathbb{N}^\star$, let us introduce the function $\Gamma_{l,q} : \mathcal{H} \to$

$\mathbb{R}_+$ defined as

$$\Gamma_{l,q}(h) := C_q \|\rho'_\gamma\|_\infty \sqrt{1+l} \times \begin{cases} 4\|K\|_q \left(n \prod_{j=1}^d h_j\right)^{-(q-1)/q}, & \text{if } q \in [1,2[, \\ \frac{30q}{\log(q)} (\|K\|_2 \vee \|K\|_q) \left(n \prod_{j=1}^d h_j\right)^{-1/2}, & \text{if } q \in [2,\infty[, \end{cases}$$

where $C_q > 0$ is an absolute constant which does not depend on n. Then, for any $l > 0$, the function $\mathcal{M}^{\mathrm{glo}}_{l,q}(\lambda,\eta) := \Gamma_{l,q}(\lambda \vee \eta) + \Gamma_{l,q}(\eta)$ is a majorant according to Definition 1.

We finally select the bandwidth according to

$$\widehat{h}^{\mathrm{glo}}_q \in \arg\min_{h\in\mathcal{H}} \Big\{ \sup_{\eta\in\mathcal{H}} \{ \|\widehat{G}^{\mathrm{loc}}_{h,\eta} - \widehat{G}^{\mathrm{loc}}_\eta\|_{q,\infty} - \mathcal{M}^{\mathrm{glo}}_{l,q}(h,\eta)\} + 2\Gamma_{l,q}(h)\Big\}.$$

The above choice of the bandwidth leads to the estimator $\widehat{f}_{\widehat{h}^{\mathrm{glo}}_q}$ with the following adaptive minimax properties for the $\mathbb{L}_q$-risk over anisotropic Nikol'skii classes; see Definition 2.

THEOREM 4. *For any $s \in (0,1]^d$, any $L > 0$ and any $q \geq 1$, it holds that*

$$\limsup_{n\to\infty} \psi^{-1}_{n,q}(s) \sup_{f\in\mathcal{N}_{q,d}(s,L)} \mathbb{E}\|\widehat{f}_{\widehat{h}^{\mathrm{glo}}_q} - f\|^q_q < \infty,$$

where $\bar{s} := (\sum_{j=1}^d s_j^{-1})^{-1}$ denotes the harmonic average and

$$\psi_{n,q}(s) = \begin{cases} (1/n)^{q(q-1)\bar{s}/(q\bar{s}+q-1)}, & \text{if } q \in [1,2[, \\ (1/n)^{q\bar{s}/(2\bar{s}+1)}, & \text{if } q \geq 2. \end{cases}$$

We refer to [18, 20] for the minimax optimality of these rates over Nikol'skii classes. The proposed estimate $\widehat{f}_{\widehat{h}^{\mathrm{glo}}_q}$ is then adaptive minimax. To the best of our knowledge, the minimax adaptivity over anisotropic Nikol'skii classes has never been studied in regression with possible heavy-tailed noises. We finally refer to the remarks after Theorem 3.

5. Discussion. Our paper solves the general bandwidth selection issue in kernel ERM by using a novel selection rule, based on the minimization of an estimate of the bias–variance decomposition of the gradient excess risk. This new criterion simultaneously upper bounds the estimation error (ℓ_2-norm) and the prediction error (excess risk) with optimal rates.

One of the key messages we would like to highlight is the following: if we consider smooth loss functions and a family of consistent ERM, fast rates of convergence are automatically reached, provided that the Hessian matrix of the risk function is positive definite. This statement is based on the key Lemma 1 in Section 1.1, where the square root of the excess risk is controlled by the G-excess risk.

From an adaptive point of view, one can take another look at Lemma 1. On the RHS of Lemma 1, the G-excess risk is multiplied by the constant $\lambda_{\min}^{-1}$, that is, the smallest eigenvalue of the Hessian matrix at $\theta^\star$. This parameter is also involved in the margin assumption. As a result, our selection rule does not depend on this parameter since the margin assumption is not required to obtain slow rates for the G-excess risk. This fact partially solves an issue highlighted by Massart [35], Section 8.5.2, in the model selection framework:

> *It is indeed a really hard work in this context to design margin adaptive penalties. Of course recent works on the topic, involving local Rademacher penalties, for instance, provide at least some theoretical solution to the problem but still if one carefully looks at the penalties which are proposed in these works, they systematically involve constants which are typically unknown. In some cases, these constants are absolute constants which should nevertheless considered as unknown just because the numerical values coming from the theory are obviously over pessimistic. In some other cases, it is even worse since they also depend on nuisance parameters related to the unknown distribution.*

In Section 6 below, we also illustrate the robustness of the method with numerical results. An interesting and challenging open problem would be to employ the gradient approach in the model selection framework in order to propose a more robust penalization technique (i.e., which does not depend on the parameter $\lambda_{\min}$).

The gradient approach requires two main ingredients: the first one concerns the smoothness of the loss function in terms of differentiability; the second one affects the dimension of the statistical model that we have at hand, which has to be parametric, that is, of finite dimension $m \in \mathbb{N}^\star$. From our point of view, the smoothness of the loss function is not a restriction, since modern algorithms are usually based—in order to reduce computational complexity—on some kind of gradient descent methods in practice. On the other hand, the second ingredient might be more restrictive from the model selection point of view. An interesting open problem would be to employ the same path when the dimension $m \geq 1$ is possibly larger than n, that is, in a high-dimensional setting.

6. Numerical results. For completeness, we illustrate the performance of our selection rule in the context of clustering with errors in variables, and compare it to the most recent bandwidth selection procedure in that framework: ERC method, recently evolved in [10]. This method has both theoretical and computational advantages (see also [23]); however, it only provides isotropic bandwidth selection. For this reason, our anisotropic selection rule may outperform ERC method.

The computation of the selection rule (3.4) requires many optimization steps. We first compute a family of codebooks $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$ according to (3.2), by using a noisy version of the vanilla k-means algorithm. This technique gives an approximation of the optimal solution (3.2) thanks to an iterative procedure based on

Newton optimization. More theoretical foundations are detailed in [8]. Second, we use parallel execution in order to compute the comparison of gradient empirical risks.

Experiments. We generate an i.i.d. noisy sample $\mathcal{D}_n = \{Z_1, \dots, Z_n\}$ such that for any $i = 1, \dots, n$,

$$Z_i = \begin{cases} X_i^{(1)} + \varepsilon_i(u), & \text{if } Y_i = 1, \\ X_i^{(2)} + \varepsilon_i(u), & \text{if } Y_i = 2, \end{cases} \tag{6.1}$$

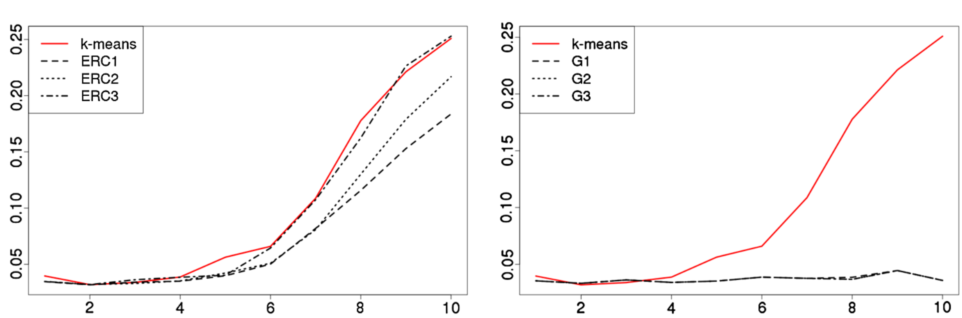
where $(X_i^{(1)})_{i=1}^n$ [resp., $(X_i^{(2)})_{i=1}^n$] are i.i.d. Gaussian with density $f_{\mathcal{N}(0_2, I_2)}$ (resp., $f_{\mathcal{N}((5,0)^T, I_2)}$) and $(Y_i)_{i=1}^n$ are i.i.d. such that $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = 2) = 1/2$. Here, $(\varepsilon_i(u))_{i=1}^n$ are i.i.d. with Gaussian noise with zero mean $(0, 0)^T$ and covariance matrix $\Sigma(u) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & u \end{smallmatrix}\right)$ for $u \in \{1, \dots, 10\}$. In this setting, we compare both adaptive procedures [our selection rule (3.2) and ERC method] to the standard k-means with Lloyd's algorithm by computing the empirical clustering error according to

$$\mathcal{I}_n(\widehat{c}_1, \widehat{c}_2) := \min_{\widehat{\mathbf{c}} = (\widehat{c}_1, \widehat{c}_2), (\widehat{c}_2, \widehat{c}_1)} \frac{1}{n} \sum_{i=1}^n \mathbf{1}\big(Y_i \neq f_{\widehat{\mathbf{c}}}(X_i)\big), \tag{6.2}$$

where $f_{\widehat{\mathbf{c}}}(x) \in \arg\min_{j=1,2} |x - \widehat{\mathbf{c}}_j|_2^2$ and $Y_i \in \{1, 2\}$, $i = 1, \dots, n$ correspond to the latent class labels defined in (6.1).

Similar to many adaptive methods, Lepki-type procedures suffer from a dependency on a tuning parameter. In particular, in ERC method, a constant governs the variance threshold (see [21] or [10]), and in our selection rule as well, a constant $b_1' > 0$ appears in the majorant function of Lemma 3. As discussed earlier, the choice of this constant remains an hard issue for application. In the sequel, we illustrate the behavior of both adaptive methods w.r.t. 3 constants: 0.1, 1 and 10.

Figure 1(a)–(b) illustrates the evolution of the clustering risk (6.2) when $u \in \{1, \dots, 10\}$ in model (6.1) for k-means (red curve) versus both adaptive procedures.



(a) k-means vs ERC method (b) k-means vs Gradient

FIG. 1. *Clustering risk averaged over* 100 *replications with* $n = 200$ *for k-means versus ERC* (a) *and the gradient* (b).

In Figure 1(a), we compare the clustering risk (6.2) of k-means (red curve) with ERC with 3 different constants (ERC1, ERC2 and ERC3). The methods are comparable, and we observe that ERC performance is sensitive to the choice of the constant. Nevertheless, a good calibration of this constant gives slightly better results than k-means. In Figure 1(b), the gradient approach with three different constants (G1, G2 and G3) gives a clustering risk less than 5% for any $u \in \{1, \dots, 10\}$. In comparison, standard k-means completely fails when u is increasing. As a conclusion, our selection rule significantly outperforms k-means and ERC for any constant. This highlights the importance in practice to choose two different bandwidths in each direction in this model, that is, an anisotropic bandwidth. Our selection rule is also robust to the choice of the constant, which confirms the theoretical study.

APPENDIX

A.1. Proof of Lemma 1. The proof is based on standard tools from differential calculus applied to the multivariate risk function $R \in \mathcal{C}^2(U)$, where U is an open ball centered at $\theta^\star$. The first step is to apply a Taylor expansion of first order which gives, for all $\theta \in U$,

$$\begin{aligned} R(\theta) - R(\theta^\star) &= (\theta - \theta^\star)^\top \nabla R(\theta^\star) \\ &\quad + \sum_{k\in\mathbb{N}^m : |k|=2} \frac{2(\theta-\theta^\star)^k}{k_1!\cdots k_m!} \int_0^1 (1-t) \frac{\partial^2}{\partial\theta^k} R\big(\theta^\star + t(\theta-\theta^\star)\big)\,dt, \end{aligned}$$

where $\frac{\partial^2}{\partial\theta^k} R = \frac{\partial^2}{\partial\theta_1^{k_1}\cdots\partial\theta_m^{k_m}} R$, $|k| = k_1 + \cdots + k_m$ and $(\theta-\theta^\star)^k = \prod_{j=1}^m (\theta_j - \theta_j^\star)^{k_j}$. Now, by the property $\nabla R(\theta^\star) = 0$ and the boundedness of the second partial derivatives, we can write

$$\begin{aligned} R(\theta) - R(\theta^\star) &\le \kappa_1 \sum_{k\in\mathbb{N}^m : |k|=2} |\theta - \theta^\star|^k \le \kappa_1 \sum_{i,j=1}^m |\theta_i - \theta_i^\star| \times |\theta_j - \theta_j^\star| \\ &\le m\kappa_1 |\theta - \theta^\star|_2^2. \end{aligned}$$

It then remains to show the inequality

$$|\theta - \theta^\star|_2 \le 2\big|G(\theta, \theta^\star)\big|_2/\lambda_{\min}, \tag{A.1}$$

where $\lambda_{\min}$ is defined in the lemma. This can be done by using standard inverse function theorem and the mean value theorem for multi-dimensional functions. Indeed, since the Hessian matrix of R—also viewed as the Jacobian matrix of G—is positive definite at $\theta^\star$, and since $R \in \mathcal{C}^2(U)$, the inverse function theorem shows the existence of a bijective function $G^{-1} \in \mathcal{C}^1(G(U))$ such that

$$|\theta - \theta^\star|_2 = \big|G^{-1} \circ G(\theta) - G^{-1} \circ G(\theta^\star)\big|_2 \qquad \text{for any } \theta \in U.$$

We can then apply a vector-valued version of the mean value theorem to obtain

$$\begin{aligned} |\theta-\theta^\star|_2 &\le \sup_{u\in[G(\theta),G(\theta^\star)]} \|\| J_{G^{-1}}(u)\|\|_2 |G(\theta^\star)-G(\theta)|_2 \\ &\qquad \text{for any } \theta\in U, \end{aligned} \tag{A.2}$$

where $[G(\theta), G(\theta^\star)]$ denotes the multi-dimensional bracket between $G(\theta)$ and $G(\theta^\star)$, and $\|\|\cdot\|\|_2$ denotes the operator norm associated to the Euclidean norm $|\cdot|_2$. Since $|\theta-\theta^\star|_2 \le \delta$ and G is continuous, we now have

$$\lim_{\delta\to 0} \sup_{u\in[G(\theta),G(\theta^\star)]} \|\| J_{G^{-1}}(u)\|\|_2 = \|\| J_{G^{-1}}(G(\theta^\star))\|\|_2.$$

Then, for $\delta > 0$ small enough, we have with (A.2)

$$\begin{aligned} |\theta-\theta^\star|_2 &\le 2\|\| J_{G^{-1}}(G(\theta^\star))\|\|_2 |G(\theta^\star)-G(\theta)|_2 \\ &= 2\|\| J_G^{-1}(\theta^\star)\|\|_2 |G(\theta^\star)-G(\theta)|_2 \\ &= 2\|\| H_R^{-1}(\theta^\star)\|\|_2 |G(\theta^\star)-G(\theta)|_2, \end{aligned}$$

where H_R is the Hessian matrix of R. (A.1) follows easily, and the proof is complete.

A.2. Proofs of Section 3.

PROOF OF LEMMA 2. The Hessian matrix of $\mathcal{W}(\cdot)$ involves integrals over faces of the Voronoï diagram $(V_j(\mathbf{c}))_{j=1}^k$. For $i \ne j$, let us denote the face (possibly empty) common to $V_i(\mathbf{c})$ and $V_j(\mathbf{c})$ as F_{ij}. Moreover, denote $\sigma(\cdot)$ the $(d-1)$-dimensional Lebesgue measure. Then, since f is continuous and $X \in [0,1]^d$, uniform continuity arguments ensure that the integral $\int_{F_{ij}} |x-m|_2^2 f(x)\sigma(dx)$ exists and depends continuously on the location of the center m, for any i, j and for any $m \in \mathbb{R}^d$. Then we can use the following lemma due to [40].

LEMMA 5 ([40]). *Suppose $\mathbb{E}_P|X|_2 < \infty$ and P has a continuous density f w.r.t. Lebesgue measure. Assume integral $\int_{F_{ij}} |x-m|_2^2 f(x)\sigma(dx)$ exists and depends continuously on the location of the centers, for any i, j and for any $m \in \mathbb{R}^d$. Then if centers c_i, $i = 1, \dots, d$ are all distinct, $\mathcal{W}(\cdot)$ has a Hessian matrix $H_{\mathcal{W}}(\cdot)$ made up of $d \times d$ blocks,*

$$H_{\mathcal{W}}(\mathbf{c})(i,j) = \begin{cases} 2\mathbb{P}(X\in V_i(\mathbf{c})) - 2\sum_{u\ne i} \delta_{iu}^{-1} \int_{F_{iu}} f(x)|x-c_i|_2^2 \sigma(dx), & \text{if } i=j, \\ -2\delta_{ij}^{-1} \int_{F_{ij}} f(x)(x-c_i)(x-c_j)^\top \sigma(dx), & \text{otherwise}, \end{cases}$$

where $\delta_{ij} = |c_i - c_j|_2$ and $\mathbf{c} \in \mathbb{R}^{dk}$.

Hence there exists $\delta > 0$ such that $\mathcal{W}(\cdot) \in \mathcal{C}^2(U)$, and Lemma 1 with $R = \mathcal{W}$ completes the proof. $\square$

PROOF OF LEMMA 3. We start with the study of $|\nabla\widehat{\mathcal{W}}_h - \mathbb{E}\nabla\widehat{\mathcal{W}}_h|_{2,\infty}$. For ease of exposition, we denote by P_n^Z the empirical measure with respect to Z_i, $i = 1, \ldots, n$ and by P^Z the expectation w.r.t. the law of Z. Then we have

$$\begin{aligned}
&|\nabla\widehat{\mathcal{W}}_h - \mathbb{E}\nabla\widehat{\mathcal{W}}_h|_{2,\infty} \\
&\quad = \sup_{\mathbf{c}\in[0,1]^{dk}} \big|\nabla\widehat{\mathcal{W}}_h(\mathbf{c}) - \mathbb{E}\nabla\widehat{\mathcal{W}}_h(\mathbf{c})\big|_2 \\
&\quad \le \sqrt{kd}\sup_{\mathbf{c},i,j}\left|(P_n^Z - P^Z)\left(\int_{V_j} 2(x^i - c_j^i)\widetilde{K}_h(Z-x)\,dx\right)\right|.
\end{aligned} \tag{A.3}$$

The cornerstone of the proof is to apply a concentration inequality to this supremum of empirical process. We use in the sequel the following Talagrand-type inequality; see, for example, [11].

LEMMA 6. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be i.i.d. random variables, and let $\mathcal{S}$ be a countable subset of $\mathbb{R}^m$. Consider the random variable*

$$U_n(\mathcal{S}) := \sup_{\mathbf{c}\in\mathcal{S}}\left|\frac{1}{n}\sum_{l=1}^n \psi_{\mathbf{c}}(\mathcal{X}_l) - \mathbb{E}\psi_{\mathbf{c}}(\mathcal{X}_l)\right|,$$

where $\psi_{\mathbf{c}}$ is such that $\sup_{\mathbf{c}\in\mathcal{S}}|\psi_{\mathbf{c}}|_\infty \le M$, $\mathbb{E}U_n(\mathcal{S}) \le E$ and $\sup_{\mathbf{c}\in\mathcal{S}}\mathbb{E}[\psi_{\mathbf{c}}(Z)^2] \le v$. Then, for any $\delta > 0$, we have

$$\mathbb{P}\big(U_n(\mathcal{S}) \ge (1+2\delta)E\big) \le \exp\left(-\frac{\delta^2 nE}{6v}\right) \vee \exp\left(-\frac{(\delta\wedge 1)\delta nE}{21M}\right).$$

The proof of Lemma 6 is omitted; see [11]. We hence have to compile the quantities E, v and M associated with the random variable

$$\widetilde{\zeta}_n = \sup_{\mathbf{c},i,j}\left|(P_n^Z - P^Z)\left(\int_{V_j} 2(x^i - c_j^i)\widetilde{K}_h(Z-x)\,dx\right)\right|.$$

The compilation of $E := E(h) > 0$ uses the same path as [10], Lemma 3. More precisely, we can apply a chaining argument to the function $\int_{V_j} 2(x^i - u)\widetilde{K}_h(Z - x)\,dx$, for any $u \in (0, 1)$. Then we have, together with a maximum inequality due to [35], Chapter 6,

$$\mathbb{E}\widetilde{\zeta}_n \le \frac{b_3}{2\sqrt{n}\Pi_h(\beta)} + \frac{b_4}{2\sqrt{n}\Pi_h(\beta+1/2)} \le \frac{b_5}{\sqrt{n}\Pi_h(\beta)} := E(h), \tag{A.4}$$

where $\Pi_h(\beta) := \prod_{i=1}^d h_i^{\beta_i}$ for $\beta \in \mathbb{R}_+^d$ provided that $\prod_{i=1}^d h_i^{-1/2} \ge b_1/b_1'$ (thanks to the definition of $\mathcal{H}_a$ and n sufficiently large). The constant $b_3, b_4, b_5 > 0$ can be explicitly computed. This calculation is omitted for simplicity. Besides, using [10], Lemma 1, with $\psi_{\mathbf{c},i,j}(Z) := \int_{V_j} 2(x^i - c^i_j)\widetilde{K}_h(Z-x)\,dx$, we have

$$\sup_{\mathbf{c},i,j} \mathbb{E}\big[\psi_{\mathbf{c},i,j}(Z)^2\big] \le \frac{b_6}{\Pi_h(2\beta)} := v(h), \tag{A.5}$$

whereas [10], Lemma 2, allows us to write

$$\sup_{\mathbf{c},i,j} |\psi_{\mathbf{c},i,j}|_\infty \le \frac{b_7}{\Pi_h(\beta+1/2)} := M(h), \tag{A.6}$$

where b_6, b_7 are absolute constants. Hence, Lemma 6, together with (A.3)–(A.6), gives us, for all $\delta > 0$,

$$\mathbb{P}\big(|\nabla\widehat{\mathcal{W}}_h - \mathbb{E}\nabla\widehat{\mathcal{W}}_h|_{2,\infty} \ge \sqrt{kd}(1+2\delta)E(h)\big)$$
$$\le \exp\biggl(-\frac{\delta^2 nE(h)}{6v(h)}\biggr) \vee \exp\biggl(-\frac{(\delta\wedge 1)\delta nE(h)}{21M(h)}\biggr).$$

Moreover, note that from the previous calculations, we have $nE(h)/v(h) = c\sqrt{n}/\Pi_h(\beta)$ and $nE(h)/M(h) = c'\sqrt{n}\sqrt{\Pi_h(1/2)}$, where $c, c' > 0$ depend on b_5, b_6 and b_5, b_7, respectively. Provided that $\sqrt{n}(c\Pi_h(\beta) \wedge c'\sqrt{\Pi_h(1/2)}) \ge (\log n)^2$ (thanks to the definition of $\mathcal{H}_a$ and n sufficiently large), we come up with

$$\mathbb{P}\big(|\nabla\widehat{\mathcal{W}}_h - \mathbb{E}\nabla\widehat{\mathcal{W}}_h|_{2,\infty} \ge \sqrt{kd}(1+2\delta)E(h)\big)$$
$$\le \exp\biggl\{-\biggl(\frac{\delta^2}{6} \wedge \frac{(\delta\wedge 1)\delta}{21}\biggr)(\log n)^2\biggr\}.$$

This gives us the first part of the majorant of Lemma 3.

The last step is to show a similar bound for the auxiliary empirical process $|\nabla\widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla\widehat{\mathcal{W}}_{h,\eta}|_{2,\infty}$. This can be easily done by using Lemma 6 together with the previous results. Then we have for any $h, \eta \in \mathcal{H}_a$,

$$\mathbb{P}\big(|\nabla\widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla\widehat{\mathcal{W}}_{h,\eta}|_{2,\infty} \ge \sqrt{kd}(1+2\delta)E(h\vee\eta)\big)$$
$$\le \exp\biggl\{-\biggl(\frac{\delta^2}{6} \wedge \frac{(\delta\wedge 1)\delta}{21}\biggr)(\log n)^2\biggr\},$$

where with a slight abuse of notation, the maximum $\vee$ is understood coordinate-wise. Using the union bound, the definition of $\mathcal{M}^{\mathrm{k}}_l(\cdot,\cdot)$ allows us to write

$$\mathbb{P}\Bigl(\sup_{h,\eta}\{|\nabla\widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla\widehat{\mathcal{W}}_{h,\eta}|_{2,\infty} + |\nabla\widehat{\mathcal{W}}_h - \mathbb{E}\nabla\widehat{\mathcal{W}}_h|_{2,\infty} - \mathcal{M}^{\mathrm{k}}_l(h,\eta)\} > 0\Bigr)$$
$$\le (\operatorname{card}\mathcal{H}_a)^2 \sup_{h,\eta}\mathbb{P}\big(|\nabla\widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla\widehat{\mathcal{W}}_{h,\eta}|_{2,\infty}$$
$$+ |\nabla\widehat{\mathcal{W}}_h - \mathbb{E}\nabla\widehat{\mathcal{W}}_h|_{2,\infty} - \mathcal{M}^{\mathrm{k}}_l(h,\eta) > 0\big)$$

$$\leq (\text{card}\,\mathcal{H}_a)^2 \sup_{h,\eta}\{\mathbb{P}(|\nabla\widehat{\mathcal{W}}_h - \mathbb{E}\nabla\widehat{\mathcal{W}}_h|_{2,\infty} - \sqrt{kd}(1+2\delta)E(h) > 0)$$
$$+ \mathbb{P}(|\nabla\widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla\widehat{\mathcal{W}}_{h,\eta}|_{2,\infty} - \sqrt{kd}(1+2\delta)E(h\vee\eta) > 0)\}$$
$$\leq 2(\text{card}\,\mathcal{H}_a)^2 \exp\left(-\frac{\delta^2}{6}\wedge\frac{(\delta\wedge 1)\delta}{21}(\log n)^2\right) \leq n^{-l},$$

where we choose $b_1' = b_5(1+2\delta)$ with $\delta := \delta(l) = 1 \vee (21(l+2)/(\log n))$. $\square$

PROOF OF THEOREM 2. The proof of Theorem 2 is a direct application of Theorem 1 and Lemma 3. Indeed, for any $l \in \mathbb{N}^\star$, for n large enough, we have with probability $1-n^{-l}$,

$$|\nabla\mathcal{W}(\widehat{\mathbf{c}}_{\widehat{h}}, \mathbf{c}^\star)|_2 \leq 3 \inf_{h\in\mathcal{H}_a}\{B(h) + \mathcal{M}_l^{\mathrm{k},\infty}(h)\},$$

where $B(h)$ is defined as

$$B(h) := \max\left(|\mathbb{E}\nabla\widehat{\mathcal{W}}_h - \nabla\mathcal{W}|_{2,\infty}, \sup_\eta |\mathbb{E}\nabla\widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla\widehat{\mathcal{W}}_\eta|_{2,\infty}\right) \qquad \forall h \in \mathcal{H}_a.$$

The control of the bias function is as follows:

$$\begin{aligned}
&|\mathbb{E}\nabla\widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla\widehat{\mathcal{W}}_\eta|_{2,\infty}^2 \\
&\quad= \sup_{\mathbf{c}\in[0,1]^{dk}} \sum_{i,j}\left\{\int_{V_j} 2(x^i - c_j^i)\big(\mathbb{E}_{PZ}\widetilde{K}_{h,\eta}(Z-x) - \mathbb{E}_{PZ}\widetilde{K}_\eta(Z-x)\big)\,dx\right\}^2 \\
&\quad= \sup_{\mathbf{c}\in[0,1]^{dk}} \sum_{i,j}\left\{\int_{V_j} 2(x^i - c_j^i)\big(\mathbb{E}_{PX}K_{h,\eta}(X-x) - \mathbb{E}_{PX}K_\eta(X-x)\big)\,dx\right\}^2 \\
&\quad\leq 4 \sup_{\mathbf{c}\in[0,1]^{dk}} \sum_{i,j}\int_{V_j}(x^i - c_j^i)^2\,dx|K_\eta * (K_h * f - f)|_2^2 \\
&\quad\leq 4k|\mathcal{F}[K]|_\infty |f_h - f|_2^2,
\end{aligned}$$

where $|f_h - f|_2 := |K_h * f - f|_2$ is the usual nonparametric bias term in deconvolution estimation. Besides, note that

$$\begin{aligned}
&|\mathbb{E}\nabla\widehat{\mathcal{W}}_h - \nabla\mathcal{W}|_{2,\infty}^2 \\
&\quad= \sup_{\mathbf{c}\in[0,1]^{dk}} \sum_{i,j}\left\{\int_{V_j} 2(x^i - c_j^i)\big(\mathbb{E}_{PX}K_h(X-x) - f(x)\big)\,dx\right\}^2 \\
&\quad\leq 4 \sup_{\mathbf{c}\in[0,1]^{dk}} \sum_{i,j}\int_{V_j}(x^i - c_j^i)^2\,dx|K_h * f - f|_2^2.
\end{aligned}$$

Then we need a control of the bias function,

$$B^{\mathrm{k}}(h) := 2\sqrt{k}\big(1 \vee \big|\mathcal{F}[K]\big|_\infty\big)|K_h * f - f|_2 \qquad \forall h \in \mathcal{H}.$$

By using Comte and Lacour [11], Proposition 3, we directly have for all $f \in \mathcal{N}_{2,d}(s, L)$,

$$B^{\mathrm{k}}(h) \le 2\sqrt{k}\big(1 \vee \big|\mathcal{F}[K]\big|_\infty\big)L \sum_{j=1}^{d} h_j^{s_j} \qquad \forall h \in \mathcal{H}. \tag{A.7}$$

Now, we have to use a result such as Lemma 2, for our family of estimators $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}_a\}$. In other words, we need to check that this family of estimators is consistent with respect to the Euclidean norm in $\mathbb{R}^{dk}$.

LEMMA 7. *Assume f is continuous, $X \in [0,1]^d$ a.s. and the Hessian matrix of $\mathcal{W}$ is positive definite on $\mathcal{M}$. Consider the family $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}_a\}$ with $\mathcal{H}_a$ defined in Lemma 3. Then, for any $\delta > 0$, for any $l \in \mathbb{N}^\star$, for any $\widehat{\mathbf{c}}_h \in \mathcal{H}_a$, there exists $\mathbf{c}^\star \in \mathcal{M}$ such that for n great enough, with probability $1 - n^{-l}$,*

$$\big|\widehat{\mathbf{c}}_h - \mathbf{c}^\star\big|_2 \le \delta.$$

PROOF. Using [1], the positive definiteness of the Hessian matrix on $\mathcal{M}$ and the continuity of f, we have, for any $\widehat{\mathbf{c}}_h \in \mathcal{H}_a$, for some constant $A_1 > 0$, $|\widehat{\mathbf{c}}_h - \mathbf{c}^\star|_2 \le A_1(\mathcal{W}(\widehat{\mathbf{c}}_h) - \mathcal{W}(\mathbf{c}^\star))$, where $\mathbf{c}^\star \in \arg\min_{\mathbf{c} \in \mathcal{M}} |\widehat{\mathbf{c}}_h - \mathbf{c}|_2$. It remains to show that by definition of $\mathcal{H}_a$ in Lemma 3, with high probability, $\mathcal{W}(\widehat{\mathbf{c}}_h) - \mathcal{W}(\mathbf{c}^\star) \to 0$ as n tends to infinity. This can be seen easily from Chichignoud and Loustau [10], which gives the order of the bias term and the variance term for such a problem. At this stage, we can notice that localization is used in [10], and appears to be necessary here. However, using a global approach (i.e., a simple Hoeffding inequality to the family of kernel ERM), we can have, for any $l \in \mathbb{N}^\star$, with probability $1 - n^{-l}$,

$$\mathcal{W}(\widehat{\mathbf{c}}_h) - \mathcal{W}(\mathbf{c}^\star) \lesssim \frac{\Pi_h(-\beta)}{\sqrt{n}} + \sum_{j=1}^{d} h_j^{s_j} \qquad \forall h \in \mathcal{H}_a.$$

By definition of $\mathcal{H}_a$, the RHS tends to zero as $n \to \infty$, and then for n great enough, this term is controlled by δ. □

Then, for any $h \in \mathcal{H}_a$ and n great enough, Lemma 2 allows us to write with probability $1 - n^{-l}$,

$$\sqrt{\mathcal{W}(\widehat{\mathbf{c}}_h) - \mathcal{W}(\mathbf{c}^\star)} \le 2\frac{\sqrt{kd}}{\lambda_{\min}}\big|\nabla\mathcal{W}(\widehat{\mathbf{c}}_h, \mathbf{c}^\star)\big|_2.$$

Using Theorem 1 with $l = q$, bias control (A.7) and the last inequality, there exists an absolute constant $b_8 > 0$ such that

$$\sup_{f \in \mathcal{N}_2(s,L)} \mathbb{E}\big[\mathcal{W}(\widehat{\mathbf{c}}_{\widehat{h}}) - \mathcal{W}(\mathbf{c}^\star)\big] \le b_8 \inf_{h \in \mathcal{H}_a}\left\{\sum_{j=1}^{d} h_j^{s_j} + \frac{\Pi_h(-\beta)}{n}\right\}^2 + b_8 n^{-q}.$$

Let $h^\star$ denote the oracle bandwidth as $h^\star := \arg\inf_{h\in\mathcal{H}}\{\sum_{j=1}^d h_j^{s_j} + \frac{\Pi_h(-\beta)}{n}\}$, and define the oracle bandwidth $h_a^\star$ on the net $\mathcal{H}_a$ such that $ah_{a,j}^\star \le h_j^\star \le h_{a,j}^\star$, for all $j = 1, \dots, d$. Eventually, we have

$$\sup_{f\in\mathcal{N}_2(s,L)} \mathbb{E}\big[\mathcal{W}(\widehat{\mathbf{c}}_{\widehat{h}}) - \mathcal{W}(\mathbf{c}^\star)\big] \le b_8 a^{-qd/2} \inf_{h\in\mathcal{H}} \left\{ \sum_{j=1}^d h_j^{s_j} + \frac{\Pi_h(-\beta)}{n} \right\}^2 + b_8 n^{-q}.$$

By a standard bias variance trade-off, we obtain the assertion of the theorem, provided that $q \ge 1$. □

A.3. Proofs of Section 4.

PROOF OF LEMMA 4. By definition, we first note that

$$\big|G^{\mathrm{loc}}\big(\widehat{f}_h(x_0)\big) - G^{\mathrm{loc}}\big(f^\star(x_0)\big)\big| = \big|\mathbb{E}\rho_\gamma'\big(\xi_1 + f^\star(x_0) - \widehat{f}_h(x_0)\big) - \mathbb{E}\rho_\gamma'(\xi_1)\big|.$$

Using the mean value theorem and the assumption $\sup_{h\in\mathcal{H}} |\widehat{f}_h(x_0) - f^\star(x_0)| \le \mathbb{E}\rho_\gamma''(\xi)/4$, there exists $c \in [-\mathbb{E}\rho_\gamma''(\xi_1)/4, \mathbb{E}\rho_\gamma''(\xi_1)/4]$ such that

$$\big|G^{\mathrm{loc}}\big(\widehat{f}_h(x_0)\big) - G^{\mathrm{loc}}\big(f^\star(x_0)\big)\big| = \mathbb{E}\rho_\gamma''(\xi_1 + c)\big|f^\star(x_0) - \widehat{f}_h(x_0)\big|.$$

Since $\mathbb{E}\rho_\gamma''(\xi_1 + \cdot)$ is a 2-Lipschitz function, it yields

$$\big|G^{\mathrm{loc}}\big(\widehat{f}_h(x_0)\big) - G^{\mathrm{loc}}\big(f^\star(x_0)\big)\big| \ge \frac{\mathbb{E}\rho_\gamma''(\xi_1)}{2}\big|f^\star(x_0) - \widehat{f}_h(x_0)\big|.$$

The proof is complete. □

PROOF OF THEOREM 3. From [9], Theorem 1, we notice that all estimators $\{\widehat{f}_h(x_0), h \in \mathcal{H}\}$ are consistent, and thus, for n sufficiently large, the assumption of Lemma 4 holds for all $x_0 \in \mathcal{T}$. Using Theorem 1 with $l > 0$ and Lemma 4, we get

$$\big|\widehat{f}_{\widehat{h}^{\mathrm{loc}}}(x_0) - f^\star(x_0)\big| \le \frac{6}{\mathbb{E}\rho_\gamma''(\xi_1)} \inf_{h\in\mathcal{H}_a} \big\{B(h) + 2\mathcal{M}_l^{\mathrm{loc},\infty}(h)\big\},$$

with $B(h) = \max(|\mathbb{E}\widehat{G}_h^{\mathrm{loc}} - G^{\mathrm{loc}}|_\infty, \sup_{\eta\in\mathcal{H}} |\mathbb{E}\widehat{G}_{h,\eta}^{\mathrm{loc}} - \mathbb{E}\widehat{G}_\eta^{\mathrm{loc}}|_\infty)$. The control of $B(\cdot)$ over Hölder classes is based on the same schema as in [13], applied to the function $F_t(\cdot) := \mathbb{E}\rho_\gamma'(f^\star(\cdot) - t + \xi_1)$. For any $f \in \Sigma(s, L)$ and any $h \in \mathcal{H}$, we then want to show

$$\begin{aligned} B^{\mathrm{loc}}(h) &\le \sup_{t\in[-B,B]} \sup_{y\in\mathcal{T}} \left| \int K_h(x-y)\big[F_t(x) - F_t(y)\big]\,dx \right| \\ &\le L|K|_\infty \sum_{j=1}^d h_j^{s_j}. \end{aligned} \tag{A.8}$$

By definition, we see that $|\mathbb{E}\widehat{G}_h^{\mathrm{loc}} - G^{\mathrm{loc}}|_\infty = \sup_{t\in[-B,B]} |\mathbb{E}K_h(W - x_0)[F_t(W) - F_t(x_0)]|$ and by definition of $\mathbb{E}\widehat{G}_{h,\eta}^{\mathrm{loc}}$ and F_t, we have

$$\begin{aligned}
-\mathbb{E}\widehat{G}_{h,\eta}^{\mathrm{loc}}(t) &= \int F_t(x) K_{h,\eta}(x - x_0)\,dx \\
&= \int F_t(x)\Bigl(\int K_h(x-y)K_\eta(y-x_0)\,dy\Bigr)\,dx.
\end{aligned}$$

Using Fubini's theorem and the equation $\int K_h(x-y)\,dx = 1$ for all $y \in \mathcal{T}$, we get

$$\begin{aligned}
-\mathbb{E}\widehat{G}_{h,\eta}^{\mathrm{loc}}(t) &= \int K_\eta(y-x_0)F_t(y)\,dy \\
&\qquad + \int K_\eta(y-x_0)\Bigl(\int K_h(x-y)\bigl[F_t(x) - F_t(y)\bigr]\,dx\Bigr)\,dy \\
&= \int K_\eta(y-x_0)F_t(y)\,dy \\
&\qquad + \int K_\eta(y-x_0)\int K_h(x-y)\bigl[F_t(x) - F_t(y)\bigr]\,dx\,dy.
\end{aligned}$$

Then it holds for any $x_0 \in \mathcal{T}$,

$$\begin{aligned}
&\bigl|\mathbb{E}\widehat{G}_{h,\eta}^{\mathrm{loc}}(t) - \mathbb{E}\widehat{G}_{\eta}^{\mathrm{loc}}(t)\bigr| \\
&\quad = \left|\int K_\eta(y-x_0)\int K_h(x-y)\bigl[F_t(x) - F_t(y)\bigr]\,dx\,dy\right| \\
&\quad \le \bigl\|K_\eta(\cdot - x_0)\bigr\|_1 \sup_{y\in\mathcal{T}}\left|\int K_h(x-y)\bigl[F_t(x) - F_t(y)\bigr]\,dx\right| \\
&\quad = \sup_{y\in\mathcal{T}}\left|\int K_h(x-y)\bigl[F_t(x) - F_t(y)\bigr]\,dx\right|.
\end{aligned}$$

We have then shown the first inequality in (A.8). Using the smoothness of ρ'_γ, we have for all $f \in \Sigma(s, L)$,

$$\begin{aligned}
&\left|\int K_h(x-y)\bigl[F_t(x) - F_t(y)\bigr]\,dx\right| \\
&\quad = \left|\int K_h(x-y)\mathbb{E}\bigl[\rho'_\gamma\bigl(f(x) - t + \xi_1\bigr) - \rho'_\gamma\bigl(f(y) - t + \xi_1\bigr)\bigr]\,dx\right| \\
&\quad \le \left|\int K_h(x-y)\bigl(f(x) - f(y)\bigr)\,dx\right| \\
&\quad \le L\|K\|_\infty \sum_{j=1}^{d} h_j^{s_j}.
\end{aligned}$$

Therefore, (A.8) holds. Then, using Theorem 1 with $l = q$, Lemma 4 and (A.8), there exists an absolute constant $T_1 > 0$ such that

$$\sup_{f\in\Sigma(s,L)} \mathbb{E}|\widehat{f}_{\widehat{h}}(x_0) - f(x_0)|^q \le T_1 \inf_{h\in\mathcal{H}_a}\left\{\sum_{j=1}^d h_j^{s_j} + \sqrt{\frac{\log(n)}{n\Pi_h}}\right\}^q + T_1 n^{-q}.$$

Let $h^\star$ denote the oracle bandwidth as $h^\star := \arg\inf_{h\in\mathcal{H}}\{\sum_{j=1}^d h_j^{s_j} + \sqrt{\frac{\log(n)}{n\Pi_h}}\}$, and define the oracle bandwidth $h_a^\star$ such that $ah_{a,j}^\star \le h_j^\star \le h_{a,j}^\star$, for all $j = 1, \dots, d$. Then we get

$$\sup_{f\in\Sigma(s,L)} \mathbb{E}|\widehat{f}_{\widehat{h}}(x_0) - f(x_0)|^q \le T_1 a^{-qd/2} \inf_{h\in\mathcal{H}}\left\{\sum_{j=1}^d h_j^{s_j} + \sqrt{\frac{\log(n)}{n\Pi_h}}\right\}^q + T_1 n^{-q}.$$

By a standard bias variance trade-off, we obtain the assertion of the theorem. □

PROOF OF THEOREM 4. Here again, the assumption of Lemma 4 holds for n sufficiently large for all $x_0 \in \mathcal{T}$. Using Theorem 1 with $l > 0$ and adding the $\mathbb{L}_q$-norm, we have

$$\|\widehat{f}_{\widehat{h}_q^{\text{glo}}} - f\|_q \le \frac{6}{\mathbb{E}\rho_\gamma''(\xi_1)} \inf_{h\in\mathcal{H}}\{B(h) + 2\Gamma_{l,q}^{\text{glo}}(h)\},$$

where $B(h) = \max(\|\mathbb{E}\widehat{G}_h^{\text{loc}} - G^{\text{loc}}\|_{q,\infty}, \sup_{\eta\in\mathcal{H}} \|\mathbb{E}\widehat{G}_{h,\eta}^{\text{loc}} - \mathbb{E}\widehat{G}_\eta^{\text{loc}}\|_{q,\infty})$. The control of the bias term is based on the schema of [15] for linear estimates. For any $h \in \mathcal{H}$, we want to show that

$$B(h) \le \sup_{t\in[-B,B]} \left\| \int K_h(x - \cdot)[F_t(x) - F_t(\cdot)]\,dx \right\|_q \le L \sum_{j=1}^d h_j^{s_j}, \tag{A.9}$$

where we recall $F_t(x) := \mathbb{E}\rho_\gamma'(f(x) - f_t(x) + \xi_1)$. By definition, one has

$$\|\mathbb{E}\widehat{G}_h^{\text{loc}} - G^{\text{loc}}\|_{q,\infty} = \sup_{t\in[-B,B]} \|\mathbb{E}K_h(W - \cdot)[F_t(W) - F_t(\cdot)]\|_q.$$

Moreover, in the proof of Theorem 3, we have shown that for any $x_0 \in \mathcal{T}$,

$$\begin{aligned}\mathbb{E}\widehat{G}_\eta^{\text{loc}}&(t, x_0) - \mathbb{E}\widehat{G}_{h,\eta}^{\text{loc}}(t, x_0)\\ &= \int K_\eta(y - x_0) \int K_h(x - y)[F_t(x) - F_t(y)]\,dx\,dy.\end{aligned}$$

By Young's inequality and the definition of the kernel in Section 1.2, it yields

$$\begin{aligned}&\|\mathbb{E}\widehat{G}_\eta^{\text{loc}} - \mathbb{E}\widehat{G}_{h,\eta}^{\text{loc}}\|_{q,\infty}\\ &\quad= \sup_{t\in[-B,B]} \left\| \int K_\eta(y - \cdot) \int K_h(x - y)[F_t(x) - F_t(y)]\,dx\,dy \right\|_{q,\infty}\\ &\quad\le \sup_{t\in[-B,B]} \left\| \int K_h(x - \cdot)|F_t(x) - F_t(\cdot)|\,dx \right\|_{q,\infty}.\end{aligned}$$

Using the smoothness of ρ'_γ, we have for any $x, y \in \mathcal{T}$ and any $t \in [-B, B]$,

$$F_t(x) - F_t(y) = \mathbb{E}[\rho'_\gamma(f(x) - t + \xi_1) - \rho'_\gamma(f(y) - t + \xi_1)] \le |f(x) - f(y)|.$$

Therefore, (A.9) holds for all $f \in \mathcal{N}_{q,d}(s, L)$. Then, using Theorem 1 with $l = q$, Lemma 4 and (A.9), there exists an absolute constant $T_2 > 0$ such that

$$\sup_{f \in \mathcal{N}_{q,d}(s,L)} \mathbb{E}\|\widehat{f}_{\widehat{h}_q^{\mathrm{glo}}} - f\|_q^q$$

$$\le T_2 \times \begin{cases} \inf_{h \in \mathcal{H}} \left\{ \sum_{j=1}^{d} h_j^{s_j} + (n\Pi_h)^{-(q-1)/q} \right\}^q + n^{-q}, & \text{if } q \in [1, 2[, \\ \inf_{h \in \mathcal{H}} \left\{ \sum_{j=1}^{d} h_j^{s_j} + (n\Pi_h)^{-1/2} \right\}^q + n^{-q}, & \text{if } q \in [2, \infty[. \end{cases}$$

Computing these infimums, we obtain the assertion of the theorem. □

REFERENCES

[1] ANTOS, A., GYÖRFI, L. and GYÖRGY, A. (2005). Individual convergence rates in empirical vector quantizer design. *IEEE Trans. Inform. Theory* **51** 4013–4022. MR2239017

[2] ARIAS-CASTRO, E., SALMON, J. and WILLETT, R. (2012). Oracle inequalities and minimax rates for nonlocal means and related adaptive kernel-based methods. *SIAM J. Imaging Sci.* **5** 944–992. MR3022184

[3] ARLOT, S. and MASSART, P. (2009). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10** 245–279.

[4] BLANCHARD, G., BOUSQUET, O. and MASSART, P. (2008). Statistical performance of support vector machines. *Ann. Statist.* **36** 489–531. MR2396805

[5] BOUCHERON, S., BOUSQUET, O. and LUGOSI, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.* **9** 323–375. MR2182250

[6] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193

[7] BOUSQUET, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris* **334** 495–500. MR1890640

[8] BRUNET, C. and LOUSTAU, S. (2014). Noisy quantization: Theory and practice. Submitted.

[9] CHICHIGNOUD, M. and LEDERER, J. (2014). A robust, adaptive M-estimator for pointwise estimation in heteroscedastic regression. *Bernoulli* **20** 1560–1599. MR3217454

[10] CHICHIGNOUD, M. and LOUSTAU, S. (2014). Adaptive noisy clustering. *IEEE Trans. Inform. Theory* **60** 7279–7292. MR3273054

[11] COMTE, F. and LACOUR, C. (2013). Anisotropic adaptive kernel deconvolution. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 569–609. MR3088382

[12] DATTNER, I., REISS, M. and TRABS, M. (2013). Adaptive quantile estimation in deconvolution with unknown error distribution. Preprint. Available at arXiv:1303.1698.

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

[14] GOLDENSHLUGER, A. and LEPSKI, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143** 41–71. MR2449122

[15] GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. MR2850214

[16] GOLDENSHLUGER, A. and LEPSKI, O. (2011). Uniform bounds for norms of sums of independent random functions. *Ann. Probab.* **39** 2318–2384. MR2932670

[17] HALL, P. and LAHIRI, S. N. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.* **36** 2110–2134. MR2458181

[18] HASMINSKII, R. and IBRAGIMOV, I. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.* **18** 999–1010. MR1062695

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

[20] IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation: Asymptotic Theory. Applications of Mathematics* **16**. Springer, New York. MR0620321

[21] KATKOVNIK, V. (1999). A new method for varying adaptive bandwidth selection. *IEEE Trans. Image Process.* **47** 2567–2571.

[22] 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

[23] KATKOVNIK, V. and SPOKOINY, V. (2008). Spatially adaptive estimation via fitted local likelihood techniques. *IEEE Trans. Signal Process.* **56** 873–886. MR2518663

[24] KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121** 137–170. MR1863916

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

[26] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

[27] LECUÉ, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* **35** 1698–1721. MR2351102

[28] 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

[29] LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

[30] LEVRARD, C. (2013). Fast rates for empirical vector quantization. *Electron. J. Stat.* **7** 1716–1746. MR3080408

[31] LOUSTAU, S. (2013). Inverse statistical learning. *Electron. J. Stat.* **7** 2065–2097. MR3091617

[32] LOUSTAU, S. and MARTEAU, C. (2015). Minimax fast rates for discriminant analysis with errors in variables. *Bernoulli* **21** 176–208. MR3322316

[33] MALLAT, S. (2009). *A Wavelet Tour of Signal Processing*, 3rd ed. Elsevier/Academic Press, Amsterdam. MR2479996

[34] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618

[35] MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

[36] MASSART, P. and NÉDÉLEC, É. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366. MR2291502

[37] MENDELSON, S. (2004). On the performance of kernel classes. *J. Mach. Learn. Res.* **4** 759–771. MR2075996

[38] NIKOL'SKIĬ, S. M. (1975). *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, New York. MR0374877

[39] POLLARD, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.* **9** 135–140. MR0600539

[40] POLLARD, D. (1982). A central limit theorem for k-means clustering. *Ann. Probab.* **10** 919–926. MR0672292

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

[42] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006

[43] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[44] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[45] VAPNIK, V. N. (1998). *Statistical Learning Theory*. Wiley, New York. MR1641250

[46] VAPNIK, V. N. and CHERVONENKIS, A. YA. (1971). Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data. *Avtomat. i Telemekh.* **2** 42–53. MR0301855

SEMINAR FUER STATISTICS
ETH ZÜRICH
RÄMISTRASSE 101
CH-8092 ZÜRICH
SWITZERLAND
E-MAIL: chichignoud@stat.math.ethz.ch

LAREMA
UNIVERSITÉ D'ANGERS
2 BVD LAVOISIER
49045 ANGERS CEDEX
FRANCE
E-MAIL: loustau@math.univ-angers.fr