Bandwidth selection in kernel empirical risk minimization via the gradient

Michaël Chichignoud, Sébastien Loustau

BANDWIDTH SELECTION IN KERNEL EMPIRICAL RISK MINIMIZATION VIA THE GRADIENT

By Michaël Chichignoud[*] and Sébastien Loustau

ETH Zürich and University of Angers

Abstract: In this paper, we deal with the data-driven selection of multidimensional and (possibly) anisotropic bandwidths in the general problem 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 regression or 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. This turns out to be helpful to derive excess risk bounds - with fast rates of convergence - in noisy clustering as well as adaptive minimax results for pointwise and global estimation in robust nonparametric regression. The selection rule consists of a comparison of the gradient empirical risks. It can be viewed as a non-trivial improvement of the so-called GL method (see Goldenshluger and Lepski [16]) to non-linear estimators. Another main advantage of our selection rule is the non-dependency on the smallest eigenvalue of the Hessian matrix of the risk, which is a changing and unknown parameter determined by the underlying model.

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 have at hand. Assume there exists a minimizer called "oracle":

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

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 corresponds to some bandwidth related to a kernel which gives rise to the problem of "kernel empirical risk minimization". One typically deals with this issue in pointwise estimation as e.g. in Polzehl and Spokoiny [46] with localized likelihoods or in Chichignoud and Lederer [10] in the setting of robust estimation with local M-estimators. In learning theory, many authors have recently investigated supervised and unsupervised learning with errors in variables. As a rule, such issues (viewed

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

MSC 2010 subject classifications: Primary 62G05, 62G20; secondary 62G08, 62H30

Keywords and phrases: Adaptivity, Bandwidth Selection, ERM, Errors-in-variables, Robust Regression, Statistical Learning

as an inverse problem) require to plug-in deconvolution kernels in the empirical risk such as Hall and Lahiri [20] in quantile and moment estimation, Loustau and Marteau [36] in noisy discriminant analysis, Loustau [34] in noisy learning, Chichignoud and Loustau [11] in noisy clustering and Dattner, Reiß and Trabs [13] in quantile estimation (see Section 2.2 for further details about these examples).

All the above papers study the theoretical properties of such kernel empirical risk minimizers and propose 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. The adaptivity is therefore one of the biggest challenges. In this respect, data-driven bandwidth selections and optimal adaptive results have been already proposed in [10, 11, 13, 46], which are all based on Lepski-type procedures.

Before describing these procedures and their interest, let us briefly explain why other popular data-driven methods are not suitable in our context. Model selection procedures have been introduced to select the hypothesis space over a sequence of nested models (e.g. finite dimension models) with a fixed empirical risk. Unfortunately, the bandwidth parameter affects the kernel empirical risk and a model selection technique cannot be directly applied in our setting. Another popular candidate is cross-validation. This useful technique is based on the following two-step procedure. First of all, a family of estimators is constructed from a subset of the observations called the training set. The rest of the sample, called the test set, is used to select the best estimator in the previous family. However, in errors-in-variables models, this test set is not available since we observe contaminated observations. Aggregation methods suffer from the same handicap. Nevertheless, we can mention Meister [41] in deconvolution estimation who is able to estimate the $\mathbb{L}_2$-risk via a Fourier analysis, but this is not directly applicable in our general context.

Lepski-type procedures are rather appropriate to construct data-driven bandwidths involved in kernels (for further details, see e.g. [24, 31, 32]). It is well-known that these procedures suffer from the restriction to isotropic bandwidths with multidimensional data, which is the consideration of nested neighborhoods (hyper-cube). Many improvements have been made by Kerkyacharian, Lepski and Picard [26] and more recently by Goldenshluger and Lepski [16] to select anisotropic bandwidths (hyper-rectangle). Nevertheless, their approach still does not provide anisotropic bandwidth selection for non-linear estimators as in our purpose. The only work we can mention is Chichignoud and Lederer [10] in a restrictive case which is pointwise estimation in nonparametric regression. Therefore, the study of data-driven selection of anisotropic bandwidths deserves some clarifications. Moreover, this field is of first interest in practice, especially in image denoising (see e.g. [2, 4]).

The main contribution of our paper is to solve this issue in the framework of kernel empirical risk minimization. To this end, we provide a novel universal data-driven selection of anisotropic bandwidths suitable for our large context of models (see Section 3 for a proper definition). This method can be viewed as a generalization of the so-called Goldenshluger-Lepski method (GL method, see [16]) and of the Empirical Risk Comparison method (ERC method, see [11]). This will enable us to construct estimators which have adaptive optimal properties. We especially derive an oracle inequality for the "Gradient excess risk" (described below), which leads to adaptive optimal results in many settings such as pointwise and global estimation in nonparametric regression and clustering with errors-in-variables.

Along the present paper, we deal with smooth loss functions, where the smoothness is related to the differentiability of the associated risk function. Under this restriction, we propose a new criterion to measure the performance of an estimator $\widehat{\theta}$, namely the Gradient excess risk (G-excess risk for short in the sequel). This quantity is defined as:

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

where $|\cdot|_2$ 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. The use of a smooth loss function, together with (1.1), leads to $G(\theta^\star) = (0,\ldots,0)^\top \in \mathbb{R}^m$ and the G-excess risk $|G(\theta,\theta^\star)|_2$ corresponds to $|G(\theta)|_2$. The main idea behind this criterion is summarized in Lemma 1 (see Section 2.1), which gives the following inequality:

$$\sqrt{R(\widehat{\theta}) - R(\theta^\star)} \lesssim \lambda_{\min}^{-1} |G(\widehat{\theta},\theta^\star)|_2,$$

where $a \lesssim b$ ($a, b \in \mathbb{R}$) means that $\exists c > 0$ such that $a \leq cb$, θ lies in a neighborhood of $\theta^\star$ and $\lambda_{\min}$ is the smallest eigenvalue of the definite positive Hessian matrix of the risk function R at $\theta^\star$. This quantity especially coincides with the usual Fisher information in maximum likelihood estimation.

With such an inequality, we can deduce "fast" rates of convergence $\mathcal{O}(n^{-1})$ for the usual excess risk $R(\widehat{\theta}) - R(\theta^\star)$ if we have at our disposal "slow" rates of convergence $\mathcal{O}(n^{-1/2})$ for the G-excess risk (see Section 2.1 for further details). One of the contributions of our paper consists in stating fast rates for the excess risk - in the presence of smooth loss functions - without using the so-called localization technique (see Mammen and Tsybakov [38], Koltchinskii [28], Blanchard, Bousquet and Massart [7], Bartlett and Mendelson [6]). In Section 4, we illustrate this phenomenon in clustering, where fast rates have been recently proposed using localization (see [33, 11]).

From an adaptive point of view, the introduction of the G-excess risk (1.2) has some interesting properties. In standard excess risk bounds, the use of localization techniques has an important drawback: any model selection or adaptive procedure suffers from the knowledge of the parameters involved in the so-called "margin assumption" (see e.g. Tsybakov [47], Koltchinskii [28]), such as the smallest eigenvalue $\lambda_{\min}$ of the Hessian matrix. Due to the G-excess risk approach, an important contribution of our paper is the non-dependency of our data-driven procedure on $\lambda_{\min}$. We give further comments on this point in the sequel.

In this paper, we consider the risk minimization (1.1) over a finite dimensional parameter of $\mathbb{R}^m$. 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 to a suitable decomposition in a basis of the Hilbert space for instance. 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 (see other examples in Section 2.2). The extension to the infinite dimensional case is discussed in Section 6.

The structure of this paper is as follows: the next section describes the main ideas behind our approach and states the first notations. In Section 3, we state an oracle inequality for the G-excess risk of the data-driven procedure. We then apply this procedure to the unsupervised learning problem of clustering in Section 4 and to robust nonparametric regression in Section 5. Additionally, we give a discussion in Section 6. The proofs are finally conducted in Section 7.

2. Main ideas and first notations.

In this section, we present the main ideas of this contribution, namely the gradient excess risk approach and the heuristic of our data-driven selection rule. We also present some examples where a bandwidth is involved in empirical risk minimization, from both local M-estimation or errors-in-variables problems.

2.1. *The gradient excess risk approach.*

As mentioned above, we suggest to work with the "G-excess risk" defined in (1.2). The most important fact with (1.2) is the following: with smooth loss functions, slow rates $\mathcal{O}(n^{-1/2})$ for the G-excess risk $|G(\widehat{\theta},\theta^\star)|_2$ lead to fast rates $\mathcal{O}(n^{-1})$ for the usual excess risk $R(\widehat{\theta})-R(\theta^\star)$ thanks to the following lemma.

LEMMA 1. *Let $\theta^\star$ satisfy (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 $\mathcal{C}^2(U)$, all of second partial derivatives of R are 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:*

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

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

The proof, given in Section 7, uses some standard tools from differential calculus applied to the multivariate risk function $R\in\mathcal{C}^2(U)$ at a neighborhood of $\theta^\star$. The constant two appearing in the RHS can be arbitrarily close to one, depending on the size of this neighborhood. In the sequel, we use Lemma 1 to a consistent family of estimators.

Let us explain how the previous lemma, together with standard probabilistic tools, allows us to establish 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 empirical risk minimizer (ERM) $\widehat{\theta}$ for ease of exposition. Thanks to the smoothness of the loss function, $G(\theta^\star)=\widehat{G}(\widehat{\theta})=(0,\dots,0)^\top$ and we lead to the following heuristic:

$$\sqrt{R(\widehat{\theta})-R(\theta^\star)}\lesssim|G(\widehat{\theta},\theta^\star)|_2=|G(\widehat{\theta})-\widehat{G}(\widehat{\theta})|_2\le\sup_{\theta\in\mathbb{R}^m}|G(\theta)-\widehat{G}(\theta)|_2\lesssim n^{-1/2}. \tag{2.1}$$

The last inequality comes from the application of a concentration inequality to the empirical process $\widehat{G}(\cdot)$, which requires no localization technique. Somehow, Lemma 1 guarantees that for a smooth loss function, fast rates occur when the Hessian matrix of the risk is positive definite at $\theta^\star$.

Now, let us compare our approach to the literature on excess risk bounds. Vapnik and Chervonenkis [50] have originally proposed to control the excess risk via the theory of empirical processes. It gives rise to slow rates $\mathcal{O}(n^{-1/2})$ for the excess risk (see also [49]). In the last decade, many authors have improved such a bound by giving fast rates using the so-called localization technique (see [7, 28, 38, 40, 42, 47] and the references therein). This field has been especially studied in classification (see Boucheron, Bousquet and Lugosi [8] for a nice survey). The principle of localization is to study the increments of an empirical process in the neighborhood of the target $\theta^\star$. Using a uniform Bernstein-type inequality, this random quantity can be bounded by its expectation with high probability. This complex technique requires a variance-risk correspondence, that is to say a control of the variance term appearing in the concentration inequality by the excess risk. This is equivalent to the so-called margin assumption. Interestingly, the next lemma suggests to link the margin assumption with some smoothness conditions on the loss function as follows.

Lemma 2. *Let $\mathcal{X}$ be a $\mathbb{R}^p$-random variable with law $P_{\mathcal{X}}$ and assume there exists a loss function $\ell : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}_+$ such that $R(\cdot) = \mathbb{E}_{P_{\mathcal{X}}}\ell(\mathcal{X}, \cdot)$. Let us consider an oracle defined in (1.1) and let U be the Euclidean ball of center $\theta^\star$ and radius $\delta > 0$ such that:*

- *$\theta \mapsto \ell(\mathcal{X}, \theta)$ is twice differentiable on U, $P_{\mathcal{X}}$-almost surely;*
- *$R(\cdot) = \mathbb{E}\ell(\mathcal{X}, \cdot)$ is three times differentiable on U and the partial derivatives of third order are bounded;*
- *the Hessian matrix $H_R(\theta^\star)$ is positive definite.*

Then, for δ sufficiently small, we have:

$$\mathbb{E}_{P_{\mathcal{X}}}\left[\ell(\mathcal{X},\theta) - \ell(\mathcal{X},\theta^\star)\right]^2 \leq 3\kappa_2\lambda_{\min}^{-1}\left[R(\theta) - R(\theta^\star)\right], \ \forall\theta \in U,$$

where $\kappa_2 = \mathbb{E}_{P_{\mathcal{X}}} \sup_{\theta\in U} |\nabla\ell(\mathcal{X},\theta)|_2^2$ and $\lambda_{\min}$ is the smallest eigenvalue of $H_R(\theta^\star)$.

The proof is given in Section 7 and uses a Taylor expansion at $\theta^\star$. Note that the regularity of the loss function implies a strong margin assumption, i.e. a power of the excess risk equals to 1. Weaker margin assumptions - where the power of the excess risk is less than 1 - have been considered in the literature (see Tsybakov [47], Koltchinskii [28], Bartlett and Mendelson [6]) and allow them to obtain fast rates of convergence for the excess risk between $\mathcal{O}(n^{-1/2})$ and $\mathcal{O}(n^{-1})$. However, to the best of our knowledge, these weaker margin assumptions are very often related to non-smooth loss functions, such as the hinge loss or the hard loss in the specific context of binary classification.

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 needs 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 [46] for the propagation method, [30] for aggregation and [3] for the slope heuristic), the theory is not completed and remains a hard issue (see the related discussion in Section 6). As mentioned above, it is surprising to note that our data-driven procedure does not suffer from the dependency on $\lambda_{\min}$ since we focus on stating slow rates for the G-excess risk and we do not need any margin assumption.

2.2. *Kernel empirical risk minimization and examples.* Let us fix some notations. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and for some $p \in \mathbb{N}^\star$, consider a $\mathbb{R}^p$-random variable Z on $(\Omega, \mathcal{F}, \mathbb{P})$ with law 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 with law P. The expectation w.r.t. the law of $\mathcal{Z}_n$ is denoted by $\mathbb{E}$.

In this paper, we are primarily interested in the kernel empirical risk minimization problem, where a bandwidth is involved in the empirical risk. In the sequel, 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:

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

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

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

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

$$\widehat{R}_h(\theta) := \frac{1}{n} \sum_{i=1}^n \ell_{K_h}(Z_i, \theta), \tag{2.2}$$

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

$$\widehat{\theta}_h \in \arg\min_{\theta \in \mathbb{R}^m} \widehat{R}_h(\theta). \tag{2.3}$$

Along the paper, 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 $\widehat{R}_h$ is an asymptotically unbiased estimator of the true risk R, i.e.

$$\lim_{h \to (0,\dots,0)} \mathbb{E}\widehat{R}_h(\theta) = R(\theta), \ \forall \theta \in \mathbb{R}^m. \tag{2.4}$$

We recall that the aim of the paper is the data-driven selection of the "best" kernel ERM in the family $\{\widehat{\theta}_h, h \in \mathcal{H}\}$. In the sequel, we list many examples of kernel empirical risk minimizations over finite dimensional spaces.

We start with local M-estimation which is usually employed in pointwise estimation. The key idea, as described for example in [48, Chapter 1], is to approximate the target function in a neighborhood of size h of a given point x_0 by a polynomial. An estimation of this polynomial can be then derived by minimizing an appropriate kernel empirical risk as in the examples below.

- **Local Fitted Likelihood - Polzehl and Spokoiny [46]**

  Let us introduce a sample of independent random variables $(W_i, Y_i) \in [0,1] \times \mathbb{R}$, $i = 1, \dots, n$, where Y_i has a probability density $g(\cdot, f_i^\star)$ with parameter $f_i^\star = f^\star(W_i)$. The aim is to estimate the quantity $f^\star(x_0) = \theta^\star$ at a given point x_0. This model contains standard nonparametric problems such as Gaussian regression, binary classification model, inhomogeneous exponential and Poisson models. In such a case, one usually applies the local version of the well-known likelihood method. It gives rise to the minimization of the localized negative log-likelihood as:

  $$\frac{1}{n}\sum_{i=1}^{n} -\log\left(g(Y_i, t)\right) K_h\left(W_i - x_0\right) \longrightarrow \min_{t\in\mathbb{R}}.$$

  In this framework, Polzehl and Spokoiny [46] have stated adaptive minimax rates in the isotropic case for the Kullback divergence via a comparison of localized log-likelihoods (see also [25]).

- **Image denoising - Astola et al. [4]**

  Let $(W_i, Y_i) \in [0,1]^2 \times \mathbb{R}$, $i = 1, \dots, n$ be the data associated to the Gaussian regression $Y_i = f(W_i) + \varepsilon_i$. In image denoising, the design W_i corresponds to the pixel location, whereas Y_i corresponds to the pixel color. To denoise a pixel (pointwise estimation), the authors consider the local least square estimate as:

  $$\frac{1}{n}\sum_{i=1}^{n} \left[Y_i - f_t(W_i)\right]^2 K_h\left(W_i - x_0\right) \longrightarrow \min_{t\in\mathbb{R}^m},$$

  where f_t is a polynomial of order $m-1$ with coefficients t. The adaptive theoretical properties of such an estimator have been investigated in the isotropic case in [18].

- **Robust nonparametric regression - Chichignoud and Lederer [10]**

  We consider the regression model $Z_i = (W_i, Y_i)$ such that $Y_i = f(W_i) + \xi_i$, where $W_1, \dots, W_n$ are independent and uniformly distributed on $[0,1]^d$, and $\xi_1, \dots, \xi_n$ are i.i.d. with possibly heavy-tailed density g_ξ. We introduce the local empirical risk :

  $$\frac{1}{n}\sum_{i=1}^{n} \rho(Y_i - t)\ K_h(W_i - x_0) \longrightarrow \min_{t\in\mathbb{R}}$$

  where $x_0 \in (0,1)^d$ and $\rho : \mathbb{R} \to \mathbb{R}_+$ is a convex, twice differentiable loss function with a bounded derivative such as the so-called Huber loss. For pointwise estimation, the authors have obtained adaptive results in the anisotropic case. This example is explicitly developed in Section 5 and generalized to global estimation.

Now, we turn out into errors-in-variables models, where a deconvolution kernel is involved in the empirical risk. Suppose we observe an i.i.d. sequence:

$$Z_i = X_i + \epsilon_i,\ \ i = 1, \dots, n, \tag{2.5}$$

where the X_i's have density f and ϵ_i's are independent to the X_i's with known density g. In this model, many statistical issues have been investigated.

- **Moment estimation - Hall and Lahiri [20]**

  From the observations $Z_i,\ i=1,\dots,n$ defined in (2.5), Hall and Lahiri [20] considered the estimation of the r-th moment of the density f, for any $r \in \mathbb{N}^\star$. This requires the use of a deconvolution kernel $\widetilde{K}_h$ (constructed from K_h) in their estimation procedure. They especially estimate the r-th moment by calculating the r-th moment of the deconvolution kernel estimator. However, this issue can be viewed as the following kernel empirical risk minimization:

  $$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}}(x^r-\mu)^2\widetilde{K}_h(Z_i-x)dx \longrightarrow \min_{\mu\in\mathbb{R}}.$$

  Hall and Lahiri [20] propose a complete minimax study of this problem and also consider quantile estimation as in the next example.

- **Quantile estimation - Dattner, Reiß and Trabs [13]**

  Given noisy data $Z_i,\ i=1,\dots,n$ as in (2.5), the goal is to estimate a τ-quantile q_τ of the density f, for any $\tau\in(0,1)$. Dattner, Reiß and Trabs [13] minimize the following kernel empirical risk:

  $$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}}(x-\eta)(\tau-\mathbb{1}_{x\le\eta})\ \widetilde{K}_h(Z_i-x)dx \longrightarrow \min_{\eta\in\mathbb{R}}.$$

  Minimax rates for this problem have been stated in [20]. Dattner, Reiß and Trabs [13] have investigated the adaptive minimax issue via a standard Lepski-type procedure. However, this is suitable to select an isotropic bandwidth, only.

- **Noisy Clustering - Chichignoud and Loustau [11]**

  Let us consider an integer $k\ge 1$. In the problem of clustering with noisy inputs $Z_i = X_i+\epsilon_i,\ i=1,\dots,n$, one wants to estimate k cluster centers $\mathbf{c}^\star=(c_1^\star,\dots,c_k^\star)\in(\mathbb{R}^d)^k$ of the density f minimizing some distortion. To this end, Chichignoud and Loustau [11] combine a deconvolution kernel and the well-known k-means distortion to give the following kernel empirical risk minimization:

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

  The authors have investigated the problem of selecting the bandwidth. They prove adaptive fast rates -up to a logarithmic term- for a data-driven selection of h, based on a comparison of empirical risks. However, as above, this paper only deals with an isotropic bandwidth. The anisotropic issue is especially studied for such a problem in Section 4.

In the sequel, we will present the selection rule of the bandwidth in the general context of kernel empirical risk minimization including all of the previous examples. We especially deal with the noisy clustering and the robust nonparametric regression in Sections 4 and 5, respectively.

2.3. *Heuristic of the selection rule.* From an adaptive point of view, we aim at selecting a kernel ERM into the family $\{\widehat{\theta}_h, h \in \mathcal{H}\}$ defined in (2.3), where $\mathcal{H} \subset \mathbb{R}_+^d$ is a set of anisotropic bandwidths. The anisotropic issue has been recently investigated in Goldenshluger and Lepski [16] (GL method) in density estimation (see also [12] in deconvolution estimation and [14, 15] for the white noise model). This method, based on the comparison of estimators, requires some "linearity" property, which is trivially satisfied for kernel estimators in density estimation. However, kernel ERM are usually non-linear (except for the least square estimator), and the GL method cannot be directly applied.

A first trail would be to compare the empirical risks (2.2) - viewed as estimators - instead of kernel ERM. This comparison has been already employed by Chichignoud and Loustau [11] with the ERC method, which is only suitable for isotropic bandwidths (see also [46]). Unfortunately, as far as we know, the GL method cannot be performed by using this comparison. More precisely, the requirement of the localization argument seems to be the main obstacle to the GL method.

To tackle this impasse, we introduce a new selection rule based on the comparison of gradient empirical risks instead of empirical risks or kernel ERM themselves. For any $h \in \mathcal{H}$ and any $\theta \in \mathbb{R}^m$, the gradient empirical risk is defined as:

$$\widehat{G}_h(\theta) := \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_{K_h}(Z_i, \theta) = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \ell_{K_h}(Z_i, \theta)\right)_{j=1,\ldots,m}. \tag{2.6}$$

Note that we have coarsely $\widehat{G}_h(\widehat{\theta}_h) = (0, \ldots, 0)^\top$ since $\ell_{K_h}(Z_i, \cdot)$ is twice differentiable almost surely. According to (2.4), we also notice that the G-empirical risk is an asymptotically unbiased estimator of the gradient of the risk.

We are now ready to describe the main ideas of our data-driven procedure. Since a nuisance bandwidth is involved in our empirical risk, we can give a bias-variance decomposition of the G-excess risk thanks to (2.1) as follows:

$$|G(\widehat{\theta}_h, \theta^\star)|_2 \leq |G - \widehat{G}_h|_{2,\infty} \leq |\mathbb{E}\widehat{G}_h - G|_{2,\infty} + |\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty}, \tag{2.7}$$

where the expectation $\mathbb{E}$ is understood coordinatewise and $|T|_{2,\infty} := \sup_{\theta \in \mathbb{R}^m} |T(\theta)|_2$ for all functions $T : \mathbb{R}^m \to \mathbb{R}^m$. The selection rule is constructed in a way that the selected bandwidth mimics the oracle bandwidth $h^\star$, which trades off the bias-variance decomposition (2.7). For this purpose, we introduce the data-driven bandwidth:

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

where $\widehat{\mathrm{BV}}(\cdot)$ is an estimate of the bias-variance decomposition and satisfies with high probability:

$$\sup_{h \in \mathcal{H}} \left\{ |\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty} + |\mathbb{E}\widehat{G}_h - G|_{2,\infty} - \widehat{\mathrm{BV}}(h) \right\} \leq 0.$$

The construction of $\widehat{\mathrm{BV}}$ consists of two steps: we first apply a Talagrand's inequality to control the variance (stochastic) term $|\widehat{G}_h - \mathbb{E}\widehat{G}_h|_{2,\infty}$, whereas the second step is to estimate the bias term $|\mathbb{E}\widehat{G}_h - G|_{2,\infty}$. This requires the comparison between G-empirical risks and an auxiliary G-empirical risk $\widehat{G}_{h,\eta}$ ($\eta \in \mathcal{H}$) associated to some convoluted kernel as in Goldenshluger and Lepski [16] (see next section for further details).

3. Selection rule and oracle inequality. In this section, we describe in details the selection rule and give the main oracle inequality. More precisely, Theorem 1 gives an upper bound for the G-excess risk (1.2) of the kernel ERM $\widehat{\theta}_{\widehat{h}}$, where $\widehat{h}$ is chosen by the selection rule described below. Gathering with Lemma 1 in Section 2, we derive excess risk bounds as well as ℓ_2-risk bounds.

The construction of the selection rule is based on the comparison of the G-empirical risks (2.6) defined in the previous section. As mentioned above, we need to 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:

$$\widehat{G}_{h,\eta}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\nabla\ell_{K_h*K_\eta}(Z_i,\theta), \tag{3.1}$$

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 statement of the main oracle inequality needs a control of the deviation of some random processes depending on the auxiliary G-empirical risk. This control is given by the next definition.

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

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

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 $\mathbb{R}^m$.

The main issue for applications is to compute right order majorants. This could be done thanks to the theory of empirical processes, such as Talagrand's inequalities (see for instance [9, 17]). In Sections 4 and 5, such majorant functions are computed in noisy clustering and in robust nonparametric regression.

We are now ready to define the selection rule as:

$$\widehat{h} = \arg\min_{h\in\mathcal{H}}\widehat{\mathrm{BV}}(h), \tag{3.2}$$

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:

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

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

THEOREM 1. *Let $\mathcal{M}_l(\cdot,\cdot)$ be a majorant according to Definition 1. For any $n \in \mathbb{N}^\star$ and for any $l \in \mathbb{N}^\star$, we have with probability $1-n^{-l}$:*

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

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

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

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. Of course, a bound in expectation can be deduced coarsely. The proof of Theorem 1, postponed at the end of the section, is based on the definition of $\widehat{h}$ in (3.2). The first step is to decompose the G-excess risk by using the auxiliary G-empirical risk (3.1). Then, Definition 1 completes the proof.

The RHS in the oracle inequality can be viewed as the minimization of an usual bias-variance trade-off. Indeed, the bias term $B(h)$ is deterministic and tends to 0 as $h \to (0,\dots,0)$. The sup-majorant $\mathcal{M}_l^\infty(h)$ upper bounds the stochastic part of the G-empirical risk and is viewed as a variance term.

We call the result of Theorem 1 "oracle inequality" since minimizing the bias-variance trade-off in the RHS can be viewed as minimizing the G-excess-risk $|G(\widehat{\theta}_h, \theta^\star)|_2$. Note that a rigorous proof of this claim still remains an open problem. However, this bound is sufficient to establish adaptive fast rates in noisy clustering and adaptive minimax rates in nonparametric estimation (see Sections 4 and 5).

In order to show the power of the G-excess risk, we simultaneously deduce a control of the estimation error $|\widehat{\theta}_{\widehat{h}} - \theta^\star|_2$ as well as a bound for the excess risk $R(\widehat{\theta}_{\widehat{h}}) - R(\theta^\star)$. In the presence of smooth loss functions, Lemma 1 is at the origin of the corollary below.

COROLLARY 1. *Suppose the assumptions of Lemma 1 are satisfied and for all $h \in \mathcal{H}$, the estimator $\widehat{\theta}_h$ of $\theta^\star$ is consistent. Then, for n sufficiently large, for any $l \in \mathbb{N}^\star$, with probability $1-n^{-l}$, it holds:*

$$R(\widehat{\theta}_{\widehat{h}}) - R(\theta^\star) \leq 36 \frac{m\kappa_1}{\lambda_{\min}^2} \inf_{h\in\mathcal{H}} \left\{B(h) + \mathcal{M}_l^\infty(h)\right\}^2,$$

and

$$|\widehat{\theta}_{\widehat{h}} - \theta^\star|_2 \leq 6 \frac{\sqrt{m\kappa_1}}{\lambda_{\min}} \inf_{h\in\mathcal{H}} \left\{B(h) + \mathcal{M}_l^\infty(h)\right\},$$

where $\kappa_1, \lambda_{\min}$ are positive constants defined in Lemma 1.

We highlight that the consistency of all estimators $\{\widehat{\theta}_h,\ h \in \mathcal{H}\}$ is necessary in order to apply Lemma 1. This usually implies restrictions on the bandwidth set (see Sections 4 and 5 for further details).

The first inequality of Corollary 1 will be used in Section 4 in the setting of clustering with errors-in-variables. In this case, we are interested in excess risk bounds and the statement of fast rates of convergence. The second inequality of Corollary 1 is the main tool to establish minimax rates for both pointwise and global risks in the context of robust nonparametric regression (see Section 5).

The construction of the selection rule (3.2), as well as the upper bound in Theorem 1, does not suffer from the dependency of $\lambda_{\min}$ 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:

$$\begin{aligned} |G(\widehat{\theta}_{\widehat{h}}, \theta^\star)|_2 &= \left|(\widehat{G}_{\widehat{h}} - G)(\widehat{\theta}_{\widehat{h}})\right|_2 \leq |\widehat{G}_{\widehat{h}} - G|_{2,\infty} \\ &\leq |\widehat{G}_{\widehat{h}} - \widehat{G}_{\widehat{h},h}|_{2,\infty} + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} + |\widehat{G}_h - G|_{2,\infty}. \end{aligned} \tag{3.3}$$

By definition of $\widehat{h}$ in (3.2), the first two terms in the RHS of (3.3) are bounded as follows:

$$\begin{aligned} |\widehat{G}_{\widehat{h}} - \widehat{G}_{\widehat{h},h}|_{2,\infty} + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} &= |\widehat{G}_{h,\widehat{h}} - \widehat{G}_{\widehat{h}}|_{2,\infty} - \mathcal{M}_\ell(h,\widehat{h}) + \mathcal{M}_\ell(\widehat{h},h) \\ &\quad + |\widehat{G}_{\widehat{h},h} - \widehat{G}_h|_{2,\infty} - \mathcal{M}_\ell(\widehat{h},h) + \mathcal{M}_\ell(h,\widehat{h}) \\ &\leq \sup_{\eta\in\mathcal{H}} \left\{ |\widehat{G}_{h,\eta} - \widehat{G}_\eta|_{2,\infty} - \mathcal{M}_\ell(h,\eta) \right\} + \mathcal{M}_\ell^\infty(h) \\ &\quad + \sup_{\eta\in\mathcal{H}} \left\{ |\widehat{G}_{\widehat{h},\eta} - \widehat{G}_\eta|_{2,\infty} - \mathcal{M}_\ell(\widehat{h},\eta) \right\} + \mathcal{M}_\ell^\infty(\widehat{h}) \\ &= \widehat{\mathrm{BV}}(h) + \widehat{\mathrm{BV}}(\widehat{h}) \leq 2\widehat{\mathrm{BV}}(h). \end{aligned} \tag{3.4}$$

Besides, the last term in (3.3) is controlled as follows:

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

Using (3.3) and (3.4), gathering with the last inequality, we have for all $h \in \mathcal{H}$:

$$|G(\widehat{\theta}_{\widehat{h}}, \theta^\star)|_2 \leq 2\widehat{\mathrm{BV}}(h) + \zeta + \mathcal{M}_l^\infty(h) + |\mathbb{E}\widehat{G}_h - G|_{2,\infty}. \tag{3.5}$$

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

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

The oracle inequality follows directly from (3.5), Definition 1 and the definition of ζ. ∎

4. Application to noisy clustering.

In this section, we are interested in the statistical learning problem of clustering. Let us consider an integer $k \geq 1$ 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 $[0,1]^d$, assuming that $X \in [0,1]^d$ almost surely. In the sequel, we denote by $\mathbf{c} = (c_1, \dots, c_k) \in (\mathbb{R}^d)^k$ a set of k cluster's centers, often called a codebook in the literature of clustering. Then, we want to construct a codebook $\mathbf{c}$ minimizing some risk or distortion:

$$\mathcal{W}(\mathbf{c}) := \mathbb{E}_P w(\mathbf{c}, X), \tag{4.1}$$

where $w(\mathbf{c}, x)$ measures the loss of the codebook $\mathbf{c}$ at point x. For ease of exposition, we study the risk minimization of (4.1) based on the Euclidean distance, by choosing a loss function related to the standard k-means loss function, namely:

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

The existence of a minimizer $\mathbf{c}^\star$ of (4.1) is proved in [19] when $\mathbb{E}_P|X|_2^2 < \infty$ (as well as for the ERM $\widehat{\mathbf{c}}$ defined below). In the standard vector quantization context studied in Section 4.1, we have at our disposal an i.i.d. sample $(X_1, \dots, X_n)$ with law P and an associated ERM:

$$\widehat{\mathbf{c}} \in \arg\min_{\mathbf{c} \in \mathbb{R}^{dk}} \widehat{\mathcal{W}}(\mathbf{c}), \text{ where } \widehat{\mathcal{W}}(\mathbf{c}) := \frac{1}{n} \sum_{i=1}^n w(\mathbf{c}, X_i). \tag{4.2}$$

Several authors have studied the statistical properties of $\widehat{\mathbf{c}}$. Pollard has proved strong consistency and central limit theorem (see [44, 45]), whereas Bartlett, Linder and Lugosi [5] have investigated minimax rates of convergence $\mathcal{O}(n^{-1/2})$ for the excess risk $\mathcal{W}(\widehat{\mathbf{c}}) - \mathcal{W}(\mathbf{c}^\star)$. More recently, Levrard [33] has proved fast rates of convergence $\mathcal{O}(n^{-1})$ under a margin assumption.

In this section, we are also interested in the inverse statistical learning context (see [35]), which corresponds to the minimization of (4.1) thanks to a noisy set of observations:

$$Z_i = X_i + \epsilon_i, \; i = 1, \dots, n,$$

where $(\epsilon_i)_{i=1}^n$ are i.i.d. with density g w.r.t. the Lebesgue measure on $\mathbb{R}^d$ and independent of the original sample $(X_i)_{i=1}^n$. This problem was first considered in [34], where general oracle inequalities are proposed. Let us fix a kernel K_h of order $r \in \mathbb{N}^\star$ with $h \in \mathcal{H}$ (see the definition in Section 2.2) 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 2, in this setting, we have at our disposal the family of kernel ERM defined as:

$$\widehat{\mathbf{c}}_h \in \arg\min_{\mathbf{c} \in \mathbb{R}^{dk}} \widehat{\mathcal{W}}_h(\mathbf{c}), \text{ where } \widehat{\mathcal{W}}_h(\mathbf{c}) := \frac{1}{n} \sum_{i=1}^n w(\mathbf{c}, \cdot) * \widetilde{K}_h(Z_i - \cdot), \tag{4.3}$$

with $f * g(\cdot) := \int_{[0,1]^d} f(x) g(\cdot - x) dx$ stands for the convolution product restricted to the compact $[0,1]^d$. Note that we restrict ourselves to the compact $[0,1]^d$ for simplicity, whereas any other compact could be considered. Recently, Chichignoud and Loustau [11] have investigated the problem of choosing the bandwidth in (4.3). They prove fast rates -up to a logarithmic term- for a data-driven selection of h, based on a comparison of kernel empirical risks. However, this paper only deals with a hyper-cube bandwidth h. Furthermore, the method explicitly depends on the parameters involved in the margin assumption and in particular on $\lambda_{\min}$ in Lemma 1.

In this section, the aim is twofold. At first, we give fast rates for the excess risk of $\widehat{\mathbf{c}}$ in (4.2) without any localization technique. The proof is extremely simple and illustrate rather well the power of the G-excess risk approach in this standard statistical learning context. Secondly, we apply the selection rule (3.2) to choose the anisotropic bandwidth in (4.3) from noisy data. We then establish adaptive minimax rates for the excess risk. In this problem as well, the use of the G-excess risk is crucial and allows us to construct a more robust data-driven procedure (i.e. which does not depend on the parameter $\lambda_{\min}$).

4.1. *Fast rates in the direct case.* The statement of fast rates for the excess risk $\mathcal{W}(\widehat{\mathbf{c}})-\mathcal{W}(\mathbf{c}^\star)$ is based on the gradient approach (see (2.1)). For this purpose, we assume that the Hessian matrix $H_{\mathcal{W}}$ is positive definite at each oracle $\mathbf{c}^\star$. This assumption has been considered for the first time in Pollard [44] and is often referred as the Pollard's regularity assumptions. Under these assumptions, we can state the same kind of result as Lemma 1 in the framework of clustering with k-means.

Lemma 3. *Let $\boldsymbol{c}^\star$ be a minimizer of (4.1) and assume $H_{\mathcal{W}}(\boldsymbol{c}^\star)$ is positive definite. Let us consider $\mathbb{C} := \{\boldsymbol{c} = (c_1,\dots,c_k) \in [0,1]^{dk} : \forall i \neq j \in \{1,\dots,k\}, c_i \neq c_j\}$. Then:*

- *$\forall x \in \mathbb{R}^d$, $\boldsymbol{c} \mapsto w(\boldsymbol{c},x)$ is infinitely differentiable on $\mathbb{C}\setminus\Delta_x$, where $\Delta_x = \{c \in [0,1]^{dk} : x \in \partial V(\boldsymbol{c})\}$ and $\partial V(\boldsymbol{c}) = \{x \in \mathbb{R}^d : \exists i \neq j$ such that $|x-c_i|_2 = |x-c_j|_2\}$;*
- *Let U be the Euclidean ball center at $\boldsymbol{c}^\star$ with radius $\delta > 0$. Then, for δ sufficiently small:*
$$\sqrt{\mathcal{W}(\boldsymbol{c})-\mathcal{W}(\boldsymbol{c}^\star)} \leq 2\frac{\sqrt{2kd}}{\lambda_{\min}}|\nabla\mathcal{W}(\boldsymbol{c},\boldsymbol{c}^\star)|_2,\ \forall \boldsymbol{c} \in U,$$
where $\lambda_{\min} > 0$ is the smallest eigenvalue of $H_{\mathcal{W}}(\boldsymbol{c}^\star)$.

As mentioned above, we need the consistency - in terms of Euclidean distance - of the ERM $\widehat{\mathbf{c}}$ defined in (4.2) in order to obtain the inequality of Lemma 3 with $\mathbf{c} = \widehat{\mathbf{c}}$. Pollard [44] has especially studied the consistency of $\widehat{\mathbf{c}}$ and allows us to satisfy our needs (see the proof of Theorem 2 for further details).

This fact allows us to control the excess risk of $\widehat{\mathbf{c}}$ as follows.

Theorem 2. *Suppose the assumptions of Lemma 3 hold. Then, for n sufficiently large, the ERM $\widehat{\boldsymbol{c}}$ defined in (4.2) satisfies:*

$$\mathbb{E}\mathcal{W}(\widehat{\boldsymbol{c}})-\mathcal{W}(\boldsymbol{c}^\star) \leq \frac{8b_1^2kd\lambda_{\min}^{-2}}{n},$$

where $b_1 > 0$ is explicitly given in the proof.

The proof is a direct application of the heuristic (2.1) presented in Section 2. In particular, the study of a derivate empirical process leads to slow rates $\mathcal{O}(n^{-1/2})$ for the G-excess risk. Lemma 3 concludes the proof.

Contrary to the results in [33], we establish fast rates for the excess risk without assuming the continuity of the underlying density f of X. This improvement is due to the G-excess risk approach, which does not require any localization technique. Indeed, we do not need that $|\mathbf{c}-\mathbf{c}^\star|_2 \lesssim \mathcal{W}(\mathbf{c}) - \mathcal{W}(\mathbf{c}^\star)$, for $\mathbf{c}$ in a neighborhood of $\mathbf{c}^\star$, which holds for instance when the density f is continuous (see Antos, Györfi and György [1]).

4.2. *Adaptive fast rates in noisy clustering.* We have at our disposal a family of kernel ERM $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$ defined in (4.3) with associated kernel empirical risk $\widehat{\mathcal{W}}_h(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^n w(\mathbf{c},\cdot) * \widetilde{K}_h(Z_i - \cdot)$, with $\widetilde{K}_h$ a deconvolution kernel. We propose to apply the selection rule (3.2) to choose the bandwidth $h \in \mathcal{H}$. In this problem as well, the use of the G-excess risk approach is of first interest to establish adaptive fast rates for the excess risk. For any $h \in \mathcal{H}$, the G-empirical risk vector is defined as:

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

where for any $j = 1,\dots,k$, $V_j(\mathbf{c}) := \{x \in [0,1]^d : \arg\min_{a=1,\dots,k}|x - c_a|_2 = j\}$ is the Voronoï cells associated to $\mathbf{c}$, and x_u denotes the u^{th} coordinate of $x \in \mathbb{R}^d$. Note that $\nabla\widehat{\mathcal{W}}_h(\widehat{\mathbf{c}}_h) = (0,\dots,0)^\top$ by smoothness. The construction of the rule follows exactly the general case of Section 3, which is based on 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:

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

where $\widetilde{K}_{h,\eta} = \widetilde{K_h * K_\eta}$ is the deconvolution kernel as in Comte and Lacour [12].

The statement of the oracle inequality is based on the computation of a majorant function. For this purpose, we need the following additional assumptions. First of all, as in standard deconvolution problems, the use of a deconvolution kernel requires some additional assumptions on the kernel K of order $r \in \mathbb{N}^\star$, according to the definition of Section 2.2.

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

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

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

This assumption is standard in deconvolution estimation and is satisfied for many standard kernels, such as the *sinc* kernel. Moreover, the construction of kernels of order r satisfying **(K1)** could be managed by using the so-called Meyer wavelet (see [37]).

Additionally, we need an assumption on the noise distribution g:

Noise Assumption $\mathbf{NA}(\rho,\beta)$. There exists some vector $\beta=(\beta_1,\dots,\beta_d)\in(0,\infty)^d$ and some positive constant ρ such that $\forall t\in\mathbb{R}^d$:

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

$\mathbf{NA}(\rho,\beta)$ deals with a lower bound on the behavior of the characteristic function of the noise density g. This lower bound is a sufficient condition to obtain excess risk bounds. However, to study the optimality in the minimax sense (see [36]), we need an upper bound of the same order for the characteristic function. This is not the purpose of this paper. Additionally, this noise assumption is related to a polynomial behavior of the Fourier transform of g. This case is called the mildly ill-posed case in the deconvolution or statistical inverse problem literature (see [41]). The severely ill-posed case corresponds to an exponential decreasing of the characteristic function in $\mathbf{NA}(\rho,\beta)$, such as a Gaussian measurement error. This case is not considered in this paper for simplicity (see [12] in multivariate deconvolution).

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

$$h_-:=\left(\frac{\log^6(n)}{n}\right)^{1/\max(2,2\sum_{j=1}^d\beta_j)}\quad\text{and }h^+:=\big(1/\log(n)\big)^{1/(2(r+1))}. \tag{4.4}$$

Lemma 4. *Assume* $(\mathbf{K1})$ *and* $\mathbf{NA}(\rho,\beta)$ *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* $|\mathcal{H}_a|\leq n$*. For any integer* $l>0$*, let us introduce the function* $\mathcal{M}_l^{\mathrm{k}}:\mathcal{H}^2\to\mathbb{R}_+$ *defined as:*

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

where $b_1'>0$*. Then, for* n *sufficiently large, the function* $\mathcal{M}_l^{\mathrm{k}}$ *is a majorant, i.e.*

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

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

The proof is based on a chaining argument and a uniform Talagrand's inequality (see Section 7). This lemma is the cornerstone of the oracle inequality below, 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:

$$\widehat{h} = \arg\min_{h\in\mathcal{H}_a}\left\{\sup_{\eta\in\mathcal{H}_a}\left\{|\nabla\widehat{\mathcal{W}}_{h,\eta} - \nabla\widehat{\mathcal{W}}_{\eta}|_{2,\infty} - \mathcal{M}_l^{\mathrm{k}}(h,\eta)\right\} + \mathcal{M}_l^{\mathrm{k},\infty}(h)\right\}, \tag{4.5}$$

where $\mathcal{M}_l^{\mathrm{k},\infty}(h) := \sup_{\lambda\in\mathcal{H}_a}\mathcal{M}_l^{\mathrm{k}}(\lambda,h)$. The next theorem gives the main result of this section, namely a control of the G-excess risk of the kernel ERM $\widehat{\mathbf{c}}_{\widehat{h}}$.

Theorem 3. *Assume* (**K1**) *and* $\mathbf{NA}(\rho,\beta)$ *hold for some* $\rho > 0$ *and some* $\beta \in \mathbb{R}_+^d$. *Then, for* n *large enough, With probability* $1 - n^{-l}$, *it holds:*

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

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

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

The proof of Theorem 3, given in Section 7, is an application of Theorem 1 gathering with Lemma 4. Note that the infimum in the RHS is restricted over the net $\mathcal{H}_a$. However, as shown in Theorem 4 below, this is sufficient to obtain adaptive optimal fast rates.

As mentioned in the previous section, we can deduce fast rates for the excess risk as an important contribution. For this purpose, we need an additional assumption on the regularity of the density f to control the bias function in Theorem 3. This regularity is expressed in terms of anisotropic Nikol'skii spaces.

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

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

and $\|D_j^l f\|_q \le L$, $\forall l = 0, \ldots, \lfloor s_j\rfloor$, *where* $D_j^l f$ *denotes the* l*-th order partial derivative of* f *w.r.t. the variable* x_j *and* $\lfloor s_j\rfloor$ *is the largest integer strictly less than* s_j.

The Nikol'skii spaces have been considered in approximation theory by Nikol'skii (see [43] for example). We also refer to [16, 26] where the problem of adaptive estimation over a scale s has been treated for the Gaussian white noise model and for density estimation, respectively.

In the sequel, we assume that the multivariate density f of the law P_X belongs to the anisotropic Nikol'skii class $\mathcal{N}_2(s, L)$, for some $s \in \mathbb{R}_+^d$ and some $L > 0$. It means that the density f has possible different regularities in all directions. The statement of a non-adaptive upper bound for the excess risk in this framework has been already investigated in [34]. 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 (4.5).

Theorem 4. *Assume* (**K1**) *and* $\mathbf{NA}(\rho,\beta)$ *hold for some* $\rho > 0$ *and some* $\beta \in \mathbb{R}_+^d$*. Assume* P *has a continuous density* $f \in \mathcal{N}_2(s,L)$ *for some* $s \in (0, r+1)^d$*, the Hessian matrix of* $\mathcal{W}$ *is definite positive for any* $\mathbf{c}^\star \in \mathcal{M}$*. Then, we have:*

$$\limsup_{n\to\infty} n^{1/(1+\sum_{j=1}^d \beta_j/s_j)} \sup_{f\in\mathcal{N}_2(s,L)} \mathbb{E}\left[\mathcal{W}(\widehat{\mathbf{c}}_{\widehat{h}}) - \mathcal{W}(\mathbf{c}^\star)\right] < \infty,$$

where $\widehat{h}$ *is chosen in* (4.5)*.*

This theorem uses Theorem 3 and Lemma 3, gathering with the consistency of the family of kernel ERM $\{\widehat{\mathbf{c}}_h, h \in \mathcal{H}\}$. In this respect, the definitions of h_- and h^+ in (4.4), gathering with the continuity of the density f, imply the consistency of our family (see Lemma 10 in Section 7).

This result gives adaptive fast rates for the excess risk of $\widehat{\mathbf{c}}_{\widehat{h}}$. It significantly improves the result stated in [11] for two main reasons. First of all, the selection rule allows the extension to the anisotropic case. Besides, there is no logarithmic term in the adaptive rate, which can be explained as follows. The localization technique used in Chichignoud and Loustau [11] seems the main obstacle to avoid the extra-log term. The use of G-excess risk approach allows us to avoid the localization technique and therefore the extra-log term in the adaptive fast rates. The result of Theorem 4 also extend the result to Nikol'skii spaces instead of Hölder spaces as in [11].

5. Application to robust nonparametric regression.

In this section, we will apply the result of Theorem 1 to the framework of local M-estimation, which leads to standard results in nonparametric regression. Indeed, oracle inequalities for the G-excess risk will give us adaptive minimax results for both pointwise and global estimation.

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

$$Y_i = f^\star(W_i) + \xi_i, \quad i = 1, \dots, n, \tag{5.1}$$

where the noise variables $(\xi_i)_{i=1,\dots,n}$ are i.i.d. with symmetric density g_ξ w.r.t. the Lebesgue measure. We aim at estimating the target function $f^\star : [0,1]^d \to [\text{-}B, B],\ B > 0$. Moreover, we also assume that g_ξ is continuous at 0 and $g_\xi(0) > 0$. For simplicity, in the sequel, the design points $(W_i)_{i=1,\dots,n}$ are i.i.d. according to the uniform law on $[0,1]^d$ (extension to a more general design is straightforward) and we suppose that $(W_i)_{i=1,\dots,n}$ and $(\xi_i)_{i=1,\dots,n}$ 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 the point $x_0 \in \mathcal{T}$ for pointwise estimation and the $\mathbb{L}_q(\mathcal{T})$-risk for global estimation.

Next, we introduce an estimate of $f^\star(x_0)$ at any $x_0 \in \mathcal{T}$ with the local constant approach (LCA) with a fixed bandwidth. The key idea of the LCA, as described for example in [48, 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 constant, which corresponds to a model of dimension $m = 1$. To deal with heavy-tailed noises, we especially employ the popular Huber loss (see [23]) defined as follows. For any scale $\gamma > 0$ and $z \in \mathbb{R}$,

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

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:

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

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, non-negative kernel of order 1 (see the definition in Section 2.2). As in (2.4), the expectation of the local empirical risk has a limit denoted by $R^{\text{loc}}(\cdot) := \mathbb{E}_{Y|W=x_0} \rho_\gamma(Y - \cdot)$ whose its unique minimizer is $f^\star(x_0)$.

In this section, we are interested in the bandwidth selection problem in the family $\{\widehat{f}_h, h \in \mathcal{H}\}$, where $\mathcal{H}$ is defined above. We want to state minimax adaptive results for both pointwise and global risks. Since Theorem 1 controls the G-excess risk of the adaptive estimator, we present the following lemma that gives rive to a control of the pointwise risk. A same inequality can be deduced with the $\mathbb{L}_q(\mathcal{T})$-norm.

Lemma 5. *Assume that* $\sup_{h \in \mathcal{H}} |\widehat{f}_h(x_0) - f^\star(x_0)| \leq \mathbb{E}\rho_\gamma''(\xi_1)/4$. *Then, for all* $h \in \mathcal{H}$,

$$|\widehat{f}_h(x_0) - f^\star(x_0)| \leq \frac{2}{\mathbb{E}\rho_\gamma''(\xi_1)} \left| G^{\text{loc}}\big(\widehat{f}_h(x_0)\big) - G^{\text{loc}}\big(f^\star(x_0)\big) \right|,$$

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 Section 7. The assumption $\sup_{h \in \mathcal{H}} |\widehat{f}_h(x_0) - f^\star(x_0)| \leq \mathbb{E}\rho_\gamma''(\xi_1)/4$ is necessary to use the theory of differential calculus and can be satisfied by using the consistency of $\widehat{f}_h$. In this direction, the definitions of h_- and h^+ above imply the consistency of all estimators $\widehat{f}_h, h \in \mathcal{H}$ (see [10, Theorem 1] for further details). This lemma allows us to link the local G-excess risk and the pointwise semi-norm.

5.1. *The selection rule in pointwise estimation.* To compute the selection procedure in pointwise estimation, we define the G-empirical risk as:

$$\widehat{G}_h^{\mathrm{loc}}(t) := \frac{\partial \widehat{R}_h^{\mathrm{loc}}}{\partial t}(t) = -\frac{1}{n}\sum_{i=1}^n \rho'_\gamma\big(Y_i - t\big)\, K_h(W_i - x_0). \tag{5.2}$$

For any couple of bandwidths $(h,\lambda) \in \mathcal{H}^2$, we introduce the auxiliary G-empirical risk as:

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

where $K_{h,\lambda} := K_h * K_\lambda$ as above.

To apply the results of Section 3, we need to compute optimal majorants of the associated empirical processes. The construction of such bounds for the pointwise case has already deserved some interests. The next lemma is a direct application of [10, Proposition 2].

Lemma 6. *For any integer $l \in \mathbb{N}^\star$, let us introduce the function $\Gamma_l^{\mathrm{loc}} : \mathcal{H} \to \mathbb{R}_+$ defined as:*

$$\Gamma_l^{\mathrm{loc}}(h) := C_0 \|K\|_2 \sqrt{\mathbb{E}[\rho'_\gamma(\xi_1)]^2}\sqrt{\frac{l\log(n)}{n\Pi_h}},$$

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

$$\mathbb{P}\left(\sup_{h,\eta\in\mathcal{H}_a}\left\{|\widehat{G}_{h,\eta}^{\mathrm{loc}} - \mathbb{E}\widehat{G}_{h,\eta}^{\mathrm{loc}}|_\infty + |\widehat{G}_{\eta}^{\mathrm{loc}} - \mathbb{E}\widehat{G}_{\eta}^{\mathrm{loc}}|_\infty - \mathcal{M}_l^{\mathrm{loc}}(h,\eta)\right\}_+ > 0\right) \le n^{-l},$$

where $|T|_\infty := \sup_{t\in[-B,B]} |T(t)|$ for all $T : \mathbb{R} \to \mathbb{R}$ and $h \vee \eta = (h_1 \vee \eta_1, \dots, h_d \vee \eta_d)$.

The proof is given in Section 7 as an application of [10, Proposition 2]. We notice that, unlike Definition 1, $|\cdot|_{2,\infty}$ is replaced by $|\cdot|_\infty$ since the G-empirical risk (5.2) is unidimensional.

Eventually, we introduce the data-driven bandwidth following the schema of the selection rule in Section 3:

$$\widehat{h}^{\mathrm{loc}} := \arg\min_{h\in\mathcal{H}_a}\left\{\sup_{\eta\in\mathcal{H}_a}\left\{|\widehat{G}_{h,\eta}^{\mathrm{loc}} - \widehat{G}_{\eta}^{\mathrm{loc}}|_\infty - \mathcal{M}_l^{\mathrm{loc}}(h,\eta)\right\} + 2\Gamma_l^{\mathrm{loc}}(h)\right\}. \tag{5.3}$$

We are now ready to give the oracle inequality for the pointwise risk:

Theorem 5 (Local Oracle Inequality). *Consider the model* (5.1) *and assume that* n *is great enough. Then, for any* $l > 0$*, with probability* $1 - n^{-l}$*, we have:*

$$|\widehat{f}_{\widehat{h}^{\mathrm{loc}}}(x_0) - f^\star(x_0)| \leq \frac{6}{\mathbb{E}\rho''_\gamma(\xi_1)} \inf_{h\in\mathcal{H}_a} \left\{ B^{\mathrm{loc}}(h) + 2\Gamma_l^{\mathrm{loc}}(h) \right\},$$

where $B^{\mathrm{loc}}(h)$ *denotes the bias term* $B^{\mathrm{loc}}(h) := \int K_h(x-y)\,|f^\star(x) - f^\star(y)|\,dx$.

The proof is a direct application of Theorem 1 and Lemma 5, since $G^{\mathrm{loc}}(f^\star(x_0)) = 0$ and

$$\sup_{\eta\in\mathcal{H}} \Gamma_l^{\mathrm{loc}}(h\vee\eta) = \Gamma_l^{\mathrm{loc}}(h).$$

Note that the infimum in the RHS of Theorem 5 is restricted to the net $\mathcal{H}_a$. However, as shown in Theorem 6 below, this is sufficient to obtain minimax adaptive results.

Chichignoud and Lederer [10, Theorem 2] have shown that the variance of local M-estimators is of order $\mathbb{E}[\rho'_\gamma(\xi_1)]^2/n(\mathbb{E}\rho''_\gamma(\xi_1))^2$. Therefore, their Lepski-type procedure depends on this quantity. Here, 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 - thanks to the gradient approach. The selection rule is therefore robust w.r.t. to the fluctuations of this parameter, in particular when γ is small (median estimator).

Now, we focus on the minimax issue for pointwise estimation and we start with the definition of the anisotropic Hölder class.

Definition 3 (Anisotropic Hölder Class). *Let* $s = (s_1, s_2, \ldots, s_d) \in (0,1]^d$ *and* $L > 0$*. We say that* $f : [0,1]^d \to [-L, L]$ *belongs to the anisotropic Hölder space* $\Sigma(s, L)$ *of functions if for all* $j = 1, ..., d$ *and for all* $z \in \mathbb{R}$*:*

$$\sup_{x\in[0,1]^d} |f(x_1, \ldots, x_j + z, \ldots, x_d) - f(x_1, \ldots, x_j, \ldots, x_d)| \leq L|z|^{s_j},$$

We then give the main result of this subsection.

Theorem 6. *For any* $s \in (0,1]^d$*, any* $L > 0$ *and any* $q \geq 1$*, it holds for all* $x_0 \in \mathcal{T}$*:*

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

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

The proposed estimator $\widehat{f}_{\widehat{h}}$ is then adaptive minimax over anisotropic Hölder spaces in pointwise estimation. The minimax optimality of this rate (with the $\log(n)$ factor) has been stated by [27] in the white noise model for pointwise estimation (see also [14]). We did not study the case of locally polynomial functions, which is further complicated to study in nonparametric regression. In this case, we could consider smoother functions $f^\star \in \Sigma(s, L)$, with $s \in (0, s^+)^d,\ s^+ > 1$.

5.2. *The selection rule in global estimation.* The aim of this section is to derive adaptive minimax results for $\widehat{f}_h$ in $\mathbb{L}_q$-risk. To this end, we need to modify the selection rule (5.3) 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:

$$\widehat{G}_h^{\text{glo}}(t,x_0) = -\frac{1}{n}\sum_{i=1}^n \rho'_\gamma\big(Y_i - t\big)K_h(W_i - x_0) \text{ and } \widehat{G}_{h,\eta}^{\text{glo}}(t,x_0) = -\frac{1}{n}\sum_{i=1}^n \rho'_\gamma\big(Y_i - t\big)K_{h,\eta}(W_i - x_0),$$

where the dependence in x_0 is explicitly written. We then 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. This topic has been recently investigated in [17] and gives the following lemma.

LEMMA 7. *For any $l \in \mathbb{N}^\star$, let us introduce the function $\Gamma_{l,q}^{\text{glo}} : \mathcal{H} \to \mathbb{R}_+$ defined as:*

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

where $\Pi_h = \prod_{j=1}^d h_j$ and $C_q > 0$ is an absolute constant which does not depend on n. Then, for any $l > 0$, the function $\mathcal{M}_{l,q}^{\text{glo}}(\lambda,\eta) := \Gamma_{l,q}^{\text{glo}}(\lambda\vee\eta) + \Gamma_{l,q}^{\text{glo}}(\eta)$ is a majorant, i.e.

$$\mathbb{P}\left(\sup_{h,\eta\in\mathcal{H}}\left\{\|\widehat{G}_{h,\eta}^{\text{glo}} - \mathbb{E}\widehat{G}_{h,\eta}^{\text{glo}}\|_{q,\infty} + \|\widehat{G}_{\eta}^{\text{glo}} - \mathbb{E}\widehat{G}_{\eta}^{\text{glo}}\|_{q,\infty} - \mathcal{M}_{l,q}^{\text{glo}}(h,\eta)\right\}_+ > 0\right) \le n^{-l}.$$

The proof is a direct application of [17, Theorem 2]. The constant C_q can be explicitly given from this theorem. Note that their approach does not allow us to obtain the term $\sqrt{\mathbb{E}[\rho'_\gamma(\xi_1)]^2}$ in the majorant's expression as in pointwise estimation but only the term $\|\rho'_\gamma\|_\infty$, which is a bound of it.

We finally select the bandwidth according to the selection rule in Section 3:

$$\widehat{h}_q^{\text{glo}} := \arg\min_{h\in\mathcal{H}}\left\{\sup_{\eta\in\mathcal{H}}\left\{\|\widehat{G}_{h,\eta}^{\text{glo}} - \widehat{G}_{\eta}^{\text{glo}}\|_{q,\infty} - \mathcal{M}_{l,q}^{\text{glo}}(h,\eta)\right\} + 2\Gamma_{l,q}^{\text{glo}}(h)\right\}.$$

THEOREM 7 (Global Oracle Inequality). *Consider the model (5.1) and assume that n is great enough. For any $l > 0$, we then have with probability $1 - n^{-l}$:*

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

where $B_q^{\text{glo}}(h) := \big\|\int K_h(x-\cdot)\big|f^\star(x) - f^\star(\cdot)\big|dx\big\|_q$ is called the global bias term.

We note that there is no restriction about the infimum over $\mathcal{H}$ - compared to the local oracle inequality - which is due to the construction of majorant. The proof is based on the same scheme as the proof of Theorem 1, by adding the $\mathbb{L}_q$-norm. Gathering with a global version of Lemma 5 (i.e. a control of the $\mathbb{L}_q$-norm instead of the pointwise semi-norm), we get the result. The proof is omitted for concision.

The above choice of the bandwidth leads to the estimator $\widehat{f}_{\widehat{h}_q^{\text{glo}}}$ with the following adaptive minimax properties for the $\mathbb{L}_q$-risk over anisotropic Nikol'skii spaces (see Definition 2 in Section 4).

THEOREM 8. *For any $s \in (0,1]^d$, any $L > 0$ and any $q \geq 1$, it holds:*

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

where $\bar{s} := \left(\sum_{j=1}^d s_j^{-1}\right)^{-1}$ 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 [21, 22] for the minimax optimality of these rates over Nikol'skii spaces. The proposed estimate $\widehat{f}_{\widehat{h}_q^{\text{glo}}}$ is then adaptive minimax. To the best of our knowledge, the minimax adaptivity over anisotropic Nikol'skii spaces has never been done in regression with possible heavy-tailed noises. As in pointwise estimation, this result could be extend to the case of local polynomial functions of order $k \geq 1$.

6. Discussion.

This paper deals with the bandwidth selection problem in kernel empirical risk minimization. We propose a new criterion called the gradient excess risk (1.2), which allows us to derive optimal fast rates of convergence for the excess risk as well as adaptive minimax rates for global and pointwise risks.

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 2, where the square root of the excess risk is controlled by the G-excess risk.

From an adaptive point of view, another look at Lemma 1 can be done. In the RHS of Lemma 1, the G-excess risk is multiplied by the constant $\lambda_{\min}^{-1}$, i.e. the smallest eigenvalue of the Hessian matrix at $\theta^\star$. This parameter is also involved in the margin assumption (see Lemma 2). 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 [39, 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."*

We can also mention the work of Koltchinskii [29], who has studied the general margin assumption. In this context, a "link function" $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ describes the relationship between the excess risk and the variance term, i.e.

$$\varphi\left(\sqrt{\mathbb{E}_{P_\mathcal{X}}\left[\ell(\mathcal{X},\theta)-\ell(\mathcal{X},\theta^\star)\right]^2}\right) \leq R(\theta)-R(\theta^\star),$$

for all θ belongs to a ball of $\theta^\star$. In our context, with smooth loss functions, the link function corresponds to the square function: $\varphi(x) = Cx^2,\ \forall x \in \mathbb{R}_+$ with $C = \lambda_{\min}/(3\kappa_2)$ (see Lemma 2). Koltchinskii [29, Section 6.3] has highlighted the issue of the adaptivity w.r.t. the link function as follows:

> *"It happens that the link function is involved in a rather natural way in the construction of complexity penalties that provide optimal convergence rates in many problems. Since the link function is generally distribution dependent, the development of adaptive penalization methods of model selection is a challenge, for instance, in classification setting."*

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 $\lambda_{\min}$).

Our paper solves the general bandwidth selection issue in ERM by using a new universal selection rule, based on the minimization of an estimate of the bias-variance decomposition of the gradient excess risk. It allows us to extend to non-linear estimators the anisotropic issue in bandwidth selection. However, this 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, i.e. of finite dimension $m \in \mathbb{N}^*$. 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. An interesting open problem would be to employ the same path when the risk functional R is measured over a functional set. For this purpose, we should consider some functional derivative in order to apply the gradient approach.

7. Appendix.

7.1. *Proofs of Section 1.*

Proof of Lemma 1. The proof of the lemma 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 $\theta \in U$:

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

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

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

It then remains to show the inequality

$$|\theta - \theta^\star|_2 \leq 2|G(\theta, \theta^\star)|_2 / \lambda_{\min}, \tag{7.1}$$

where $\lambda_{\min}$ is defined in the lemma. This could 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:

$$|\theta - \theta^\star|_2 = \left|G^{-1} \circ G(\theta) - G^{-1} \circ G(\theta^\star)\right|_2, \quad \text{for any } \theta \in U,$$

provided that $\delta > 0$ is chosen small enough. We can then apply a vector-valued version of the mean value theorem to obtain:

$$|\theta - \theta^\star|_2 \leq \sup_{u \in [G(\theta), G(\theta^\star)]} |||J_{G^{-1}}(u)|||_2 |G(\theta^\star) - G(\theta)|_2, , \quad \text{for any } \theta \in U, \tag{7.2}$$

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 $|\theta - \theta^\star|_2 \leq \delta$ 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 (7.2):

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

where H_R is the Hessian matrix of R. (7.1) follows easily and the proof is completed. ∎

Proof of Lemma 2. We first apply mean value Theorem to the function $\ell(x, \cdot)$ for all $x \in \mathbb{R}^p$. By integration, it yields

$$\mathbb{E}\left[\ell(\mathcal{X}, \theta) - \ell(\mathcal{X}, \theta^\star)\right]^2 \leq \kappa_2 |\theta - \theta^\star|_2^2.$$

Thanks to the smoothness of the risk, a Taylor expansion and the property $\nabla R(\theta^\star) = 0$ lead to:

$$\begin{aligned} R(\theta) - R(\theta^\star) &= (\theta - \theta^\star)^\top \nabla R(\theta^\star) + \frac{1}{2}(\theta - \theta^\star)^\top H_R\big(\theta^\star\big)(\theta - \theta^\star) + \mathcal{O}(|\theta - \theta^\star|_2^3), \\ &\geq \frac{\lambda_{\min}}{3}|\theta - \theta^\star|_2^2. \end{aligned}$$

where the last inequality is obtained choosing δ sufficiently small. This completes the proof. ∎

7.2. *Proofs of Section 4.*

Proof of Lemma 3. For a given $x \in \mathbb{R}^d$, it is easy to see that $\mathbf{c} \mapsto \min_{j=1,\dots,k} |x - c_j|_2^2$ is infinitely differentiable on $[0,1]^{dk} \setminus \Delta_x$. Then, if X admits a density w.r.t. the Lebesgue measure, we coarsely have $P_X(\partial V(\mathbf{c})) = 0$ and $\mathbf{c} \mapsto \min_{j=1,\dots,k} |x - c_j|_2^2$ is a.s. infinitely differentiable on $\mathbb{C} = [0,1]^{dk} \setminus \{c = (c_1, \dots, c_k) \in \mathbb{R}^{dk} : \exists i \neq j \text{ such that } c_i = c_j\}$. Hence, from Pollard's regularity conditions, using a dominated convergence theorem, $\mathbf{c} \mapsto \mathcal{W}(\mathbf{c})$ is infinitely differentiable on $\mathbb{C}$.

Let us consider $\mathbf{c}^\star$ an oracle, we can then show that $\mathbf{c}^\star \in \mathbb{C}$. Indeed, any optimal $\mathbf{c}^\star$ satisfies the centroid condition (see [19] for the definition). Therefore ([19, Theorem 4.2]), for any $i \neq j$, we have:

$$P_X(\{x \in \mathbb{R}^d : |x - c_i^\star|_2 = |x - c_j^\star|_2\}) = 0.$$

Besides, from the centroid condition, we have $P(V_i(\mathbf{c}^\star)) > 0$, for any $i \in \{1, \dots, k\}$. Suppose $\mathbf{c}_i^\star = \mathbf{c}_j^\star$ for some $i \neq j$. Hence, for any $x \in V_i(\mathbf{c}^\star)$, $|x - c_i^\star|_2 = |x - c_j^\star|_2$ and this leads to a contradiction since $P(V_i(\mathbf{c}^\star)) \leq P(\{x \in \mathbb{R}^d : |x - c_i^\star|_2 = |x - c_j^\star|_2\}) = 0$. Due to the centroid condition and the existence of a density w.r.t. the Lebesgue measure, we finally have $\mathbf{c}^\star \in (0,1)^{dk} \setminus \{c = (c_1, \dots, c_k) \in \mathbb{R}^{dk} : \exists i \neq j \text{ such that } c_i = c_j\}$ and the existence of an open set $U \subset \mathbb{C}$ containing $\mathbf{c}^\star$ is guaranteed.

To conclude Lemma 3, it is sufficient to apply Lemma 1 with $\kappa_1 = 2$ since $H_{\mathcal{W}}(\mathbf{c}^\star)$ is positive definite for any minimizer $\mathbf{c}^\star$. ∎

Proof of Theorem 2. The first assertion is based on the control of a supremum of an empirical process. For ease of exposition, we denote by P_n the empirical measure with respect to the sample X_i, $i = 1, \dots, n$ and by P the expectation w.r.t. the distribution P. Then, we obtain by the heuristic (2.1), for n great enough:

$$\begin{aligned} |\nabla \mathcal{W}(\widehat{\mathbf{c}}, \mathbf{c}^\star)|_2 &\leq \sup_{\mathbf{c} \in \mathbb{C}} |\nabla \widehat{\mathcal{W}}(\mathbf{c}) - \nabla \mathcal{W}(\mathbf{c})|_2 \\ &= \sup_{\mathbf{c} \in \mathbb{C}} \sqrt{\sum_{i,j=1}^{d} \left((P_n - P)(2(X^i - c_{ij})\mathbb{1}(X \in V_j(\mathbf{c}))\right)^2} \\ &\leq \sqrt{kd} \sup_{\mathbf{c} \in \mathbb{C}, i, j} \left|(P_n - P)(2(X^i - c_{ij})\mathbb{1}(X \in V_j(\mathbf{c}))\right|, \end{aligned}$$

where the supremum over i (and respectively j) is taken on $\{1,\dots,d\}$ (respectively $\{1,\dots,k\}$) and X^i is the i^{th} coordinate of X. We hence have to use a Talagrand's inequality to the random variable:

$$\zeta_n = \sup_{\mathbf{c}\in\mathbb{C},i,j} \left|(P_n - P)(2(X^i - c_{ij})\mathbb{1}(X\in V_j(\mathbf{c})))\right|.$$

With Bousquet's Inequality (see [9]), we have with probability $1-a$ ($a>0$) that:

$$|\nabla\mathcal{W}(\widehat{\mathbf{c}},\mathbf{c}^\star)|_2 \leq kd\left[\mathbb{E}\zeta_n + \sqrt{\frac{2\log a^{-1}}{n}[\sigma + (1+b)\mathbb{E}\zeta_n}\right] + \frac{\log(a^{-1})}{3n},$$

where $\sigma = \sup_{\mathbf{c}\in\mathbb{C},i,j}\mathbb{E}[2(X^i - c_{ij})\mathbb{1}(X\in V_j(\mathbf{c}))]^2$, $b = \sup_{\mathbf{c}\in\mathbb{C},i,j,x}|2(x^i - c_{ij})\mathbb{1}(x\in V_j(\mathbf{c}))|$. Firstly, it is easy to see that since $\mathbb{C}\subset[0,1]^{dk}$, $\sigma^2\leq 4$ and $b\leq 2$. Last step is then to control the quantity:

$$\mathbb{E}\zeta_n = \mathbb{E}\sup_{\mathbf{c}\in\mathbb{C},i,j}\left|(P_n - P)(2(X^i - c_{ij})\mathbb{1}(X\in V_j(\mathbf{c})))\right|.$$

For this purpose, we use a chaining argument. Let us consider, for any $v\in\mathbb{N}^\star$, Γ_v a a^v-net of $\mathbb{C}$ ($0<a<1$) w.r.t. the Euclidean distance. Let us denote $u_v(\mathbf{c}) := \arg\inf_{u\in\Gamma_v}|u-\mathbf{c}|_2$ and $u_0(\mathbf{c})$ an arbitrary point on $\mathbb{C}$. Thus, we have $u_v(\mathbf{c})\to\mathbf{c}$ a.s. and in $\mathbb{L}_1(P)$. By dominated convergence Theorem, for any couple (i,j), we have:

$$F_{ij}(\mathbf{c}) = F_{ij}(u_0(\mathbf{c})) + \sum_{v\in\mathbb{N}^\star}(F_{ij}(u_v(\mathbf{c})) - F(u_{v-1}(\mathbf{c}))),$$

where $F_{ij}(\mathbf{c}) = 2(X^i - c_{ij})\mathbb{1}(X\in V_j(\mathbf{c}))$. Then, we can write:

$$\begin{aligned}\mathbb{E}\zeta_n &= \mathbb{E}\sup_{i,j}|(P_n-P)(F_{ij}(u_0(\mathbf{c})))| \\ &\quad + \sum_{v\in\mathbb{N}^\star}\mathbb{E}\sup_{i,j}\sup_{u,u'\in\Gamma_v\times\Gamma_{v-1}:|u-u'|_2\leq a^v}\left|(P_n-P)(F_{ij}(u)-F_{ij}(u'))\right| \\ &:= A_1 + A_2.\end{aligned}$$

The control of A_1 and A_2 is based on a maximal inequality due to [39].

Lemma 8 (Maximal Inequality). *Let $\mathcal{X}_1,\dots,\mathcal{X}_n$ be a sequence of independent random variables. For any finite subset Φ of real functions, assume there exists some constants $\sigma, b>0$ such that for any $\phi\in\Phi$, $\frac{1}{n}\sum_{i=1}^n\mathbb{E}\phi^2(\mathcal{X}_i)\leq\sigma^2$ and $\|\phi\|_\infty\leq b$. Then:*

$$\mathbb{E}\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^n\phi(\mathcal{X}_i) - \mathbb{E}\phi(\mathcal{X}_i)\right| \leq \frac{2\sigma}{\sqrt{n}}\sqrt{2\log(|\Phi|)} + \frac{2b}{3n}\log(|\Phi|),$$

where $|\Phi|$ denotes the cardinal of the set Φ.

Then, using the previous lemma with $\sigma^2 = 4$ and $b = 2$, we have:

$$A_1 \leq \frac{4}{\sqrt{n}}\sqrt{2\log(kd)} + \frac{4}{3n}\log(kd|\mathcal{M}|).$$

For A_2, using the same path and the fact that $\mathbb{E}[(X^i - u_{ij})\mathbb{1}(X \in V_j(u)) - (X^i - u'_{ij})I(X \in V_j(u'))]^2 \leq |u - u'|^2 \leq a^v$, we have for any $v \in \mathbb{N}^\star$:

$$\begin{aligned}
&\mathbb{E}\sup_{i,j} \sup_{u,u' \in \Gamma_v \times \Gamma_{v-1} : |u'_u|_2 \leq a^v} \left|(P_n - P)(F_{ij}(u) - F_{ij}(u'))\right| \\
&\quad \leq \frac{2a^{v/2}}{\sqrt{n}}\sqrt{2\log(kd|\gamma_v \times \Gamma_{v-1}|)} + \frac{8}{3n}\log(kd|\gamma_v \times \Gamma_{v-1}|).
\end{aligned}$$

Now, it is easy to see that $|\Gamma_v \times \Gamma_{v-1}| \leq a^{kd(-2v+1)}$ by construction. Choosing $a = 1/4$, we hence have for A_2 by simple algebra:

$$\begin{aligned}
A_2 &\leq \sum_{v \in \mathbb{N}^\star}\left(\frac{2a^{v/2}}{\sqrt{n}}\sqrt{2\log(kda^{kd(-2v+1)})} + \frac{8}{3n}\log(kda^{kd(-2v+1)})\right) \\
&\leq \frac{4\sqrt{2}}{\sqrt{n}}\sqrt{\log(kd)} + 4\left(\frac{8\sqrt{2kd}}{\sqrt{n}}\sqrt{\log 4} + \frac{8kd}{3n}\log 4\right) + \frac{4}{3n}\log(kd).
\end{aligned}$$

Gathering with the previous inequalities, and integrating with respect to the sample, we arrive at:

$$\mathbb{E}|\nabla\mathcal{W}(\widehat{\mathbf{c}}, \mathbf{c}^\star)|_2 \leq b_1/\sqrt{n},$$

where $b_1 > 0$ can be explicit. Now, from Pollard [44], we have the a.s. convergence of the ERM $\widehat{\mathbf{c}}$ defined in (4.2) to the oracle $\mathbf{c}^\star$ w.r.t. the Euclidean distance. Eventually, from Lemma 3, $\widehat{\mathbf{c}}$ satisfies a.s. the inequality:

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

where $\lambda_{\min} > 0$ is the smallest eigenvalue of $H_{\mathcal{W}}(\mathbf{c}^\star)$. This concludes the proof. ∎

Proof of Lemma 4. 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:

$$\begin{aligned}
|\nabla\widehat{\mathcal{W}}_h - \mathbb{E}\nabla\widehat{\mathcal{W}}_h|_{2,\infty} &= \sup_{\mathbf{c} \in [0,]^{dk}} |\nabla\widehat{\mathcal{W}}_h(\mathbf{c}) - \mathbb{E}\nabla\widehat{\mathcal{W}}_h(\mathbf{c})|_2 \\
&\leq \sqrt{kd}\sup_{\mathbf{c},i,j}\left|\left(P_n^Z - P^Z\right)\left(\int_{V_j(\mathbf{c})} 2(x^i - c_{ij})\widetilde{K}_h(Z - x)dx\right)\right|. \qquad (7.3)
\end{aligned}$$

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 e.g. [12]).

Lemma 9. *Let $\mathcal{X}_1, \dots, \mathcal{X}_n$ be i.i.d. random variables and let $\mathcal{S}$ be a countable subset of $\mathbb{R}^m$. Consider the random variable $U_n(\mathcal{S}) := \sup_{c\in\mathcal{S}} \left|\frac{1}{n}\sum_{l=1}^n \psi_{\boldsymbol{c}}(\mathcal{X}_l) - \mathbb{E}\psi_{\boldsymbol{c}}(\mathcal{X}_l)\right|$, where $\psi_{\boldsymbol{c}}$ is such that $\sup_{\boldsymbol{c}\in\mathcal{S}} |\psi_{\boldsymbol{c}}|_\infty \leq M$, $\mathbb{E}U_n(\mathcal{S}) \leq E$ and $\sup_{\boldsymbol{c}\in\mathcal{S}} \mathbb{E}\left[\psi_{\boldsymbol{c}}(Z)^2\right] \leq v$. Then, for any $\delta > 0$, we have:*

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

We hence have to compile the quantities E, v and M associated with the random variable:

$$\widetilde{\zeta}_n = \sup_{\mathbf{c},i,j}\left|\left(P_n^Z - P^Z\right)\left(\int_{V_j(\mathbf{c})} 2(x^i - c_{ij})\widetilde{K}_h(Z-x)dx\right)\right| := \sup_{\mathbf{c},i,j}\left|\frac{1}{n}\sum_{l=1}^n \psi_{\mathbf{c},i,j}(Z_l) - \mathbb{E}\psi_{\mathbf{c},i,j}(Z)\right|.$$

The compilation of $E := E(h) > 0$ uses the same path as in the proof of Theorem 2, gathering with [11, Lemma 3]. More precisely, we can use a chaining argument to the function:

$$\widetilde{F}_{ij}(u) = \int_{V_j(u)} 2(x^i - u_{ij})\widetilde{K}_h(Z-x)dx,$$

for any $u \in (0,1)^{dk}$. Then, we have, gathering with the maximum inequality of Lemma 8:

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

where $\Pi_h(\beta) := \Pi_{i=1}^d h_i^{\beta_i}$ for $\beta \in \mathbb{R}_+^d$ provided that $\Pi_{i=1}^d h_i^{-1/2} \geq b_1/b_1'$ (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 using the proof of Theorem 2 above and a precise look at [11, Lemma 3]. This calculation is omitted for simplicity. Besides, using [11, Lemma 1] and denote by $\psi_{\mathbf{c},i,j}(Z) = \int_{V_j(\mathbf{c})} 2(x^i - c_{ij})\widetilde{K}_h(Z-x)dx$, we have:

$$\sup_{\mathbf{c},i,j} \mathbb{E}\left[\psi_{\mathbf{c},i,j}(Z)^2\right] \leq \frac{b_6}{\Pi_h(2\beta)} := v(h),$$

whereas [11, Lemma 2] allows us to write:

$$\sup_{\mathbf{c},i,j} |\psi_{\mathbf{c},i,j}|_\infty \leq \frac{b_7}{\Pi_h(\beta+1/2)} := M(h), \tag{7.4}$$

where b_6, b_7 are absolute constants. Hence, Lemma 9 gathering with (7.3)-(7.4) gives us, for all $\delta > 0$,

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

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 $c, c' > 0$ depends on b_5, b_6 and b_5, b_7 respectively.

Provided that $\sqrt{n}(c\Pi_h(\beta) \wedge c'\sqrt{\Pi_h(1/2)}) \geq (\log n)^2$ (thanks to the definition of $\mathcal{H}_a$ and n sufficiently large), we come up with:

$$\mathbb{P}\left(|\nabla \widehat{\mathcal{W}}_h - \mathbb{E}\nabla \widehat{\mathcal{W}}_h|_{2,\infty} \geq \sqrt{kd}(1+2\delta)E(h) \right) \leq \exp\left\{ -\left(\frac{\delta^2}{6} \wedge \frac{(\delta \wedge 1)\delta}{21} \right) (\log n)^2 \right\}.$$

This gives us the first part of the majorant of Lemma 4.

Last step is to give the same kind of result 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 again Lemma 9 together with the previous results. Then, one obtains for any $h, \eta \in \mathcal{H}_a$:

$$\mathbb{P}\left(|\nabla \widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla \widehat{\mathcal{W}}_{h,\eta}|_{2,\infty} \geq \sqrt{kd}(1+2\delta)E(h \vee \eta) \right) \leq \exp\left\{ -\left(\frac{\delta^2}{6} \wedge \frac{(\delta \wedge 1)\delta}{21} \right) (\log n)^2 \right\},$$

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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{h,\eta} \left\{ |\nabla \widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla \widehat{\mathcal{W}}_{h,\eta}|_{2,\infty} + |\nabla \widehat{\mathcal{W}}_h - \mathbb{E}\nabla \widehat{\mathcal{W}}_h|_{2,\infty} - \mathcal{M}_l^{\mathrm{k}}(h,\eta) \right\} > 0 \right) \\
&\quad \leq (\mathrm{card}\mathcal{H}_a)^2 \sup_{h,\eta} \mathbb{P}\left(|\nabla \widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla \widehat{\mathcal{W}}_{h,\eta}|_{2,\infty} + |\nabla \widehat{\mathcal{W}}_h - \mathbb{E}\nabla \widehat{\mathcal{W}}_h|_{2,\infty} - \mathcal{M}_l^{\mathrm{k}}(h,\eta) > 0 \right) \\
&\quad \leq (\mathrm{card}\mathcal{H}_a)^2 \sup_{h,\eta} \left\{ \mathbb{P}\left(|\nabla \widehat{\mathcal{W}}_h - \mathbb{E}\nabla \widehat{\mathcal{W}}_h|_{2,\infty} - \sqrt{kd}(1+2\delta)E(h) > 0 \right) \right. \\
&\qquad\qquad \left. + \mathbb{P}\left(|\nabla \widehat{\mathcal{W}}_{h,\eta} - \mathbb{E}\nabla \widehat{\mathcal{W}}_{h,\eta}|_{2,\infty} - \sqrt{kd}(1+2\delta)E(h \vee \eta) > 0 \right) \right\} \\
&\quad \leq 2\,(\mathrm{card}\mathcal{H}_a)^2 \exp\left(-\frac{\delta^2}{6} \wedge \frac{(\delta \wedge 1)\delta}{21} (\log n)^2 \right) \leq n^{-l},
\end{aligned}$$

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

Proof of Theorem 3. The proof of Theorem 3 is a direct application of Theorem 1 and Lemma 4. Indeed, for any $l \in \mathbb{N}^\star$, for n large enough, we directly have with proba $1 - n^{-l}$:

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

where $B(h)$ is defined as:

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

The control of the bias function is as follows:

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

where $|f_h-f|_2=|K_h*f-f|_2=|\mathbb{E}_{PX}\widehat{f}_h-f|_2$ is the usual nonparametric bias term in deconvolution estimation. Besides, note that:

$$\begin{aligned}
|\mathbb{E}\nabla\widehat{\mathcal{W}}_{h}-\nabla\mathcal{W}|_{2,\infty}^2 &= \sum_{i,j}\left\{\int_{V_j}2(x^i-c_{ij})\left(\mathbb{E}_{PX}K_{h}(X-x)-f(x)\right)dx\right\}^2\\
&\leq 4\sum_{i,j}\int_{V_j}(x^i-c_{ij})^2dx|K_h*f-f)|_2^2.
\end{aligned}$$

The expression of B^{k} easily follows. ∎

Proof of Theorem 4. We start with a control of the bias function involved in Theorem 3, namely the quantity:

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

By using for instance Proposition 3 in Comte and Lacour [12], we directly have for all $f\in\mathcal{N}_2(s,L)$:

$$B^{\mathrm{k}}(h)\leq 2\sqrt{k}\,(1\vee|\mathcal{F}[K]|_\infty)\,L\sum_{j=1}^d h_j^{s_j},\quad\forall h\in\mathcal{H}.$$

Now, we have to use a result such as Lemma 3, for our family of kernel ERM $\{\widehat{\mathbf{c}}_h,\ h\in\mathcal{H}_a\}$. In other words, we need to check that this family is consistent w.r.t. the Euclidean norm in $\mathbb{R}^{dk}$.

LEMMA 10. *Assume f is continuous, $|X|_\infty\leq 1$ a.s. and the Hessian matrix of $\mathcal{W}$ is positive definite for any $\mathbf{c}^\star\in\mathcal{M}$. Consider the family $\{\widehat{\mathbf{c}}_h,\ h\in\mathcal{H}_a\}$ with $\mathcal{H}_a$ defined in Lemma 4. Then, for any $t>0$:*

$$\mathbb{P}\left(|\widehat{\mathbf{c}}_h-\mathbf{c}^\star(\widehat{\mathbf{c}}_h)|_2\to 0\right)\geq 1-e^{-t},$$

where $\mathbf{c}^\star(\widehat{\mathbf{c}}_h)=\arg\min_{\mathbf{c}^\star\in\mathcal{M}}|\widehat{\mathbf{c}}_h-\mathbf{c}^\star|_2$.

Proof of Lemma 10. Using [1] and the continuity of f, we have, for some constant $A_1 > 0$, $|\widehat{\mathbf{c}}_h - \mathbf{c}^\star(\widehat{\mathbf{c}}_h)|_2 \leq A_1(\mathcal{W}(\widehat{\mathbf{c}}_h) - \mathcal{W}(\mathbf{c}^\star(\widehat{\mathbf{c}}_h)))$. Moreover, by definition of $\mathcal{H}_a$ in Lemma 4, $\mathcal{W}(\widehat{\mathbf{c}}_h) - \mathcal{W}(\mathbf{c}^\star(\widehat{\mathbf{c}}_h)) \to 0$ as n tends to infinity. This could be seen easily from Loustau [34][Theorem 3], 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 [34], and then appears to be necessary here. However, using a global approach (i.e. a simple Hoeffding's inequality to our family of kernel ERM), we can have, for any $t > 0$, on an event Ω_t such that $\mathbb{P}(\Omega_t) \geq 1 - e^{-t}$:

$$\mathcal{W}(\widehat{\mathbf{c}}_h) - \mathcal{W}(\mathbf{c}^\star) \lesssim \frac{\Pi_h(-2\beta)}{\sqrt{n}} + \sum_{j=1}^d h_j^{s_j},\ \forall h \in \mathcal{H}_a.$$

It is finally clear that the RHS of this inequality tends to 0 as n tends to infinity, for any $h \in \mathcal{H}_a$. The proof of Lemma 10 is completed. ∎

Then, for any $h \in \mathcal{H}_a$, for any $t > 0$, and n great enough, Lemma 3 allows us to write on Ω_t:

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

Using Theorem 3 with $l = q$, the bias control (7.2) and the last inequality for a proper $t > 0$, there exists an absolute constant $b_8 > 0$ such that for n great enough:

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

Let $h^\star$ denote the oracle bandwidth as $h^\star := \arg\inf_{h\in\mathcal{H}}\left\{\sum_{j=1}^d h_j^{s_j} + \frac{\Pi_h(-2\beta)}{n}\right\}$, and define the oracle bandwidth $h_a^\star$ on the net $\mathcal{H}_a$ such that $ah_{a,j}^\star \leq h_j^\star \leq h_{a,j}^\star$, for all $j = 1, \ldots, d$. We finally get:

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

By a standard bias variance trade-off, we obtain the assertion of the theorem, provided that $q \geq 2$. ∎

7.3. *Proofs of Section 5.*

Proof of Lemma 5. By definition, we first note that:

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

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

$$\left|G^{\text{loc}}\big(\widehat{f}_h(x_0)\big)-G^{\text{loc}}\big(f^\star(x_0)\big)\right|=\mathbb{E}\rho''_\gamma(\xi_1+c)|f^\star(x_0)-\widehat{f}_h(x_0)|.$$

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

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

The proof is completed. ■

Proof of Lemma 6. By simple algebra, we have:

$$\begin{aligned}&\mathbb{P}\left(\sup_{h,\eta\in\mathcal{H}_a}\left\{|\widehat{G}^{\text{loc}}_{h,\eta}-\mathbb{E}\widehat{G}^{\text{loc}}_{h,\eta}|_\infty+|\widehat{G}^{\text{loc}}_{\eta}-\mathbb{E}\widehat{G}^{\text{loc}}_{\eta}|_\infty-\mathcal{M}^{\text{loc}}_l(h,\eta)\right\}_+>0\right)\\&\quad\leq\sum_{\lambda,\eta\in\mathcal{H}_a}\mathbb{P}\left(|\widehat{G}^{\text{loc}}_{\lambda,\eta}-\mathbb{E}\widehat{G}^{\text{loc}}_{\lambda,\eta}|_\infty>\Gamma_l(\lambda\vee\eta)\right)+\sum_{\eta\in\mathcal{H}_a}\mathbb{P}\left(|\widehat{G}^{\text{loc}}_{\eta}-\mathbb{E}\widehat{G}^{\text{loc}}_{\eta}|_\infty>\Gamma^{\text{loc}}_l(\eta)\right).\end{aligned}$$

Using [10, Proposition 2] with $\Lambda=\{\rho_\gamma\}$ and $\mathcal{F}=[-B,B]$, it yields:

$$\begin{aligned}&\mathbb{P}\left(\sup_{h,\eta\in\mathcal{H}_a}\left\{|\widehat{G}^{\text{loc}}_{h,\eta}-\mathbb{E}\widehat{G}^{\text{loc}}_{h,\eta}|_\infty+|\widehat{G}^{\text{loc}}_{\eta}-\mathbb{E}\widehat{G}^{\text{loc}}_{\eta}|_\infty-\mathcal{M}^{\text{loc}}_l(h,\eta)\right\}_+>0\right)\\&\leq\sum_{\lambda,\eta\in\mathcal{H}_a}\exp\{-(l+2)\log(n)\}/2+\sum_{\eta\in\mathcal{H}_a}\exp\{-(l+2)\log(n)\}/2\\&\leq|\mathcal{H}_a|^2\exp\{-(l+2)\log(n)\}/2+|\mathcal{H}_a|\exp\{-(l+2)\log(n)\}/2\\&\leq n^{-l}\end{aligned}$$

■

Proof of Theorem 5. From [10, 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 5 holds for all $x_0\in\mathcal{T}$. Using Theorem 1 with $l>0$ and Lemma 5, it yields:

$$|\widehat{f}_{\widehat{h}^{\text{loc}}}(x_0)-f^\star(x_0)|\leq\frac{6}{\mathbb{E}\rho''_\gamma(\xi_1)}\inf_{h\in\mathcal{H}_a}\left\{B(h)+2\Gamma^{\text{loc}}_l(h)\right\},$$

with $B(h)=\max\left(|\mathbb{E}\widehat{G}^{\text{loc}}_h-G^{\text{loc}}|_\infty,\sup_{\eta\in\mathcal{H}}|\mathbb{E}\widehat{G}^{\text{loc}}_{h,\eta}-\mathbb{E}\widehat{G}^{\text{loc}}_{\eta}|_\infty\right)$ with the term $G^{\text{loc}}(\cdot):=\mathbb{E}_{Y|W=x_0}\rho'_\gamma(Y-\cdot)$. The control of the bias term is based on the same schema of [14] applied to the function $F_t(\cdot):=\mathbb{E}\rho'_\gamma(f^\star(\cdot)-t+\xi_1)$. For any $h\in\mathcal{H}$, it then remains to show:

$$\begin{aligned}B(h)&\leq\sup_{t\in[\text{-}B,B]}\sup_{y\in\mathcal{T}}\left|\int K_h(x-y)\big[F_t(x)-F_t(y)\big]dx\right|\\&\leq\sup_{y\in\mathcal{T}}\left|\int K_h(x-y)(f^\star(x)-f^\star(y))dx\right|.\end{aligned}\tag{7.5}$$

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

$$-\mathbb{E}\widehat{G}_{h,\eta}^{\mathrm{loc}}(t) = \int F_t(x)K_{h,\eta}(x-x_0)dx = \int F_t(x)\left(\int K_h(x-y)K_\eta(y-x_0)dy\right)dx.$$

Using Fubini theorem and the equation $\int K_h(x-y)dx = 1$ for all $y\in\mathcal{T}$, we get

$$-\mathbb{E}\widehat{G}_{h,\eta}^{\mathrm{loc}}(t) = \int K_\eta(y-x_0)F_t(y)dy + \int K_\eta(y-x_0)\int K_h(x-y)\big[F_t(x)-F_t(y)\big]dxdy.$$

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

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

We have then shown the first inequality in (7.5). Using the smoothness of ρ'_γ, we have:

$$\begin{aligned}\left|\int K_h(x-y)\big[F_t(x)-F_t(y)\big]dx\right| &= \left|\int K_h(x-y)\mathbb{E}\big[\rho'_\gamma\big(f^\star(x)-t+\xi_1\big)-\rho'_\gamma\big(f^\star(y)-t+\xi_1\big)\big]dx\right| \\ &\le \int K_h(x-y)\,|f^\star(x)-f^\star(y)|\,dx\end{aligned}$$

Therefore, (7.5) holds and the proof is completed. ∎

Proof of Theorem 6. For all $f\in\Sigma(s,L)$, we have :

$$\int K_h(x-y)\,|f^\star(x)-f^\star(y)|\,dx \le L\|K\|_\infty \sum_{j=1}^d h_j^{s_j}.$$

Using Theorem 5 with $l=q$, there exists an absolute constant $T_1>0$ such that

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

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

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

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

Proof of Theorem 7. Note that $\{\widehat{f}_h(x_0),\ h \in \mathcal{H}\}$ is a family of consistent estimators (see [10, Theorem 1]) and thus the assumption of Lemma 5 holds for n sufficiently large and for all $x_0 \in \mathcal{T}$. Gathering with Theorem 1 with $l > 0$ and adding the $\mathbb{L}_q$-norm, it gives:

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

where $B(h) = \max\left(\|\mathbb{E}\widehat{G}_h^{\text{glo}} - G^{\text{glo}}\|_{q,\infty}, \sup_{\eta\in\mathcal{H}} \|\mathbb{E}\widehat{G}_{h,\eta}^{\text{glo}} - \mathbb{E}\widehat{G}_\eta^{\text{glo}}\|_{q,\infty} \right)$ and $G^{\text{glo}}(\cdot, x_0) := \mathbb{E}_{Y|W=x_0}\rho'_\gamma(Y - \cdot)$, $\forall x_0 \in \mathcal{T}$. The control of the bias term is based on the schema of [16] for linear estimates: for any $h \in \mathcal{H}$, it remains to show:

$$B(h) \leq \left\| \int K_h(x - \cdot)\big|f^\star(x) - f^\star(\cdot)\big|dx \right\|_q \tag{7.6}$$

By definition, we see that $\|\mathbb{E}\widehat{G}_h^{\text{glo}} - G^{\text{loc}}\|_{q,\infty} = \sup_{t\in[\text{-}B,B]} \left\| \mathbb{E}K_h(W - \cdot)\big[F_t(W) - F_t(\cdot)\big] \right\|_q$, where we recall $F_t(x) := \mathbb{E}\rho'_\gamma(f^\star(x) - t + \xi_1)$. Moreover, in the proof of Theorem 6, we have shown that for any $x_0 \in \mathcal{T}$

$$\mathbb{E}\widehat{G}_\eta^{\text{glo}}(t, x_0) - \mathbb{E}\widehat{G}_{h,\eta}^{\text{glo}}(t, x_0) = \int K_\eta(y - x_0) \int K_h(x - y)\big[F_t(x) - F_t(y)\big]dxdy.$$

By Young inequality and the definition of the kernel in Section 2.2, it yields

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

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

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

Therefore, (7.6) holds and the proof is completed. ∎

Proof of Theorem 8. We first notice that for all $f \in \mathcal{N}_q(s, L)$, we have

$$\left\| \int K_h(x - \cdot)\big|f^\star(x) - f^\star(\cdot)\big|dx \right\|_q \leq L \sum_{j=1}^d h_j^{s_j}.$$

Using Theorem 7 with $l = q$, there exists an absolute constant $T_2 > 0$ such that

$$\sup_{f\in\mathcal{N}_{q,d}(s,L)} \mathbb{E}\|\widehat{f}_{\widehat{h}_q^{\text{glo}}} - f\|_q^q \leq T_2 \begin{cases} \inf_{h\in\mathcal{H}} \left\{ \sum_{j=1}^d h_j^{s_j} + (n\Pi_h)^{-(q-1)/q} \right\}^q + n^{-q} & \text{if } q \in [1, 2[\\ \inf_{h\in\mathcal{H}} \left\{ \sum_{j=1}^d h_j^{s_j} + (n\Pi_h)^{-1/2} \right\}^q + n^{-q} & \text{if } \ q \in [2, \infty[\end{cases} .$$

By a standard Bias/Variance trade-off, 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.

[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.

[3] Arlot, S. and Massart, P. (2009). Data-driven Calibration of Penalties for Least-Squares Regression. *Journal of Machine Learning Research* **10** 245–279.

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

[5] Bartlett, P. L., Linder, T. and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory* **44** 1802–1813.

[6] Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields* **135 (3)** 311-334.

[7] Blanchard, G., Bousquet, O. and Massart, P. (2008). Statistical performance of support vector machines. *Ann. Statist.* **36** 489–531.

[8] Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics* **9** 323-375.

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

[10] Chichignoud, M. and Lederer, J. (2013). A Robust, Fully Adaptive M-estimator for Pointwise Estimation in Heteroscedastic Regression. *To appear in Bernoulli, arXiv:1207.4447v3.*

[11] Chichignoud, M. and Loustau, S. (2013). Adaptive noisy clustering. *Submitted, arXiv:1306.2194.*

[12] Comte, F. and Lacour, C. (2013). Anisotropic adaptive kernel deconvolution. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 569–609.

[13] Dattner, I., Reiss, M. and Trabs, M. (2013). Adaptive quantile estimation in deconvolution with unknown error distribution. *arXiv: 1303.1698.*

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

[15] Goldenshluger, A. and Lepski, O. V. (2009). Structural adaptation via Lp-norm oracle inequalities. *Probab. Theory and Related Fields* **143** 41–71.

[16] Goldenshluger, A. and Lepski, O. (2011a). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632.

[17] Goldenshluger, A. and Lepski, O. (2011b). Uniform bounds for norms of sums of independent random functions. *Ann. Probab.* **39** 2318–2384.

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

[19] Graf, S. and Luschgy, H. (2000). *Foundation of quantization for probability distributions.* Springer-Verlag Lecture Notes in Mathematics, volume 1730.

[20] Hall, P. and Lahiri, S. N. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.* **36** 2110–2134.

[21] Hasminskii, R. and Ibragimov, I. On density estimation in the view of Kolmogorov's ideas in approximation theory. **3** 999–1010.

[22] Has'minskii, R. Z. and Ibragimov, I. A. (1981). *Statistical Estimation, Asymptotic Theory.* Springer-Verlag, Applications of Mathematics.

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

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

[25] Katkovnik, V. and Spokoiny, V. (2008). Spatially adaptive estimation via fitted local likelihood techniques. *IEEE Trans. Signal Process.* **56** 873–886.

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

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

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

[29] Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems. Lecture Notes in Mathematics* **2033**. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008.

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

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

[32] 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.

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

[34] Loustau, S. (2013a). Anisotropic Oracle Inequalities in Noisy Quantization. *Submitted*.

[35] Loustau, S. (2013b). Inverse statistical learning. *Electron. J. Stat.* **7** 2065–2097.

[36] Loustau, S. and Marteau, C. (2013). Minimax fast rates for discriminant analysis with errors in variables. *To appear in Bernoulli, arXiv:1201.3283v2*.

[37] Mallat, S. (2009). *A wavelet tour of signal processing*, Third ed. Elsevier/Academic Press, Amsterdam The sparse way, With contributions from Gabriel Peyré.

[38] Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829.

[39] Massart, P. (2007). *Concentration inequalities and model selection. Lecture Notes in Mathematics* **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour.

[40] Massart, P. and Nédélec, E. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326-2366.

[41] Meister, A. (2009). *Deconvolution problems in nonparametric statistics. Lecture Notes in Statistics* **193**. Springer-Verlag, Berlin.

[42] Mendelson, S. (2003). On the performance of kernel classes. *Journal of Machine Learning Research* **4** 759–771.

[43] Nikol'skii, S. M. (1975). *Approximation of functions of several variables and imbedding theorems.* Springer-Verlag, New York.

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

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

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

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

[48] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation. Springer Series in Statistics.* Springer, New York.

[49] Vapnik, V. N. (1998). *Statistical learning theory. Adaptive and Learning Systems for Signal Processing, Communications, and Control.* John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

[50] Vapnik, V. N. and Chervonenkis, A. Y. (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 Telemeh.* **2** 42-53.

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