

Adaptive Noisy Clustering

Michael Chichignoud, Sébastien Loustau

▶ **To cite this version:**

Michael Chichignoud, Sébastien Loustau. Adaptive Noisy Clustering. 2013. hal-00831672

HAL Id: hal-00831672
https://hal.science/hal-00831672

Preprint submitted on 7 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Adaptive Noisy Clustering

Michaël CHICHIGNOUD* and Sébastien LOUSTAU†

June 7, 2013

Abstract

The problem of adaptive noisy clustering is investigated. Given a set of noisy observations $Z_i = X_i + \epsilon_i$, $i = 1, \dots, n$, the goal is to design clusters associated with the law of X_i's, with unknown density f with respect to the Lebesgue measure. Since we observe a corrupted sample, a direct approach as the popular k-means is not suitable in this case. In this paper, we propose a noisy k-means minimization, which is based on the k-means loss function and a deconvolution estimator of the density f. In particular, this approach suffers from the dependance on a bandwidth involved in the deconvolution kernel. Fast rates of convergence for the excess risk are proposed for a particular choice of the bandwidth, which depends on the smoothness of the density f.

Then, we turn out into the main issue of the paper: the data-driven choice of the bandwidth. We state an adaptive upper bound for a new selection rule, called ERC (Empirical Risk Comparison). This selection rule is based on the Lepski's principle, where empirical risks associated with different bandwidths are compared. Finally, we illustrate that this adaptive rule can be used in many statistical problems of M-estimation where the empirical risk depends on a nuisance parameter.

1 Introduction

Motivation. Nonparametric procedures of estimation contain some nuissance parameter(s) whose optimal selection is not obvious. From the minimax point of view, these methods reach optimal rates of convergence, based on a regularity assumption over the unknown function to estimate. As a consequence, optimal parameters depend on some unknown smoothness index $s > 0$ of some functional space (e.g. the Hölder smoothness). In density estimation, the most popular technique of kernel estimators (see Rosenblatt [1956] or Parzen [1962]) suffers from the dependence on a bandwidth parameter $\lambda > 0$. In deconvolution as well, kernel deconvolution estimators are rather popular to estimate a density from a sequence of independent and identically distributed (i.i.d.) contaminated observations:

$$Z_i = X_i + \epsilon_i, \text{ for } i = 1, \dots, n, \tag{1.1}$$

where f denotes the unknown density of the i.i.d. sequence $X_1, X_2, ..., X_n$ and η is the known density of the i.i.d. random variables $\epsilon_1, \epsilon_2, ..., \epsilon_n$, independent of the X_i's. In this framework, a kernel deconvolution estimator of f is given by:

$$\hat{f}_\lambda(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{K}_\lambda \left(Z_i - x \right), \tag{1.2}$$

where $\mathcal{K}_\lambda$ is a deconvolution kernel and $\lambda > 0$ is a bandwidth. Estimators of the form (1.2) are of first interest in this paper. Fan [1991] has proved *minimax rates of convergence* for (1.2) for an optimal value of the bandwidth. This deterministic choice trades off a bias term and a variance term and depends on unknown parameters, such as the smoothness index $s > 0$ of the density f.

From the adaptive point of view, the aim is the data-driven selection of an estimator from a given family which has some *adaptive optimal properties*: the selected estimator reaches the minimax rate for any function in a vast range of regularities $s \in]0, s^+[$, $s^+ > 0$. In this case, the proposed estimator

*ETH Zürich, Seminar für Statistik, HG G11 Rämistrasse 101, 8092 Zürich, SWITZERLAND
†LAREMA, Université d'Angers, 2 Bvd Lavoisier 49045 Angers Cedex, FRANCE

does not depend on the exact smoothness index $s > 0$ of the target function but only on the upper bound s^+. It reaches minimax adaptivity with respect to the unknown smoothness. One of the most popular method for choosing the bandwidth is suggested by Lepski, Mammen, and Spokoiny [1997] in a gaussian white noise model. It is based on the Lepski's principle (Lepski [1990]). The idea is to test several estimators (by comparison) for different values of the bandwidth. This work is at the origin of a quantity of theoretical papers dealing with minimax adaptive bounds in nonparametric estimation (see for instance Goldenshluger and Nemirovski [1997], Mathé [2006], Chichignoud [2012]). In the context of deconvolution, Comte and Lacour [2013] gets adaptive optimal results (for pointwise and global risks) using an improvement of the standard Lepski's principle (see Goldenshluger and Lepski [2011]). From the practical point of view, Lepski's method has also received further development, such as the intersection of confidence intervals (ICI) rule (see Katkovnik [1999]). This algorithm reveals computational advantages in comparison to the traditional Lepski's procedure, or even traditional cross-validation techniques since it does not require to compute all the estimators in the family. This algorithm was originally design for a problem of gaussian filtering. It is at the core of many applications in image processing (see Kervrann and Boulanger [2006], Astola, Egiazarian, Foi, and Katkovnik [2010] and references therein).

Noisy data. In the present paper, we deal with the problem of clustering with noisy observations (2.1). Classical results in the presence of noisy observations are given in a deconvolution framework (see the references above), or alternatively in regression with errors-in-variables. Fan and Truong [1993] gives for the first time the minimax rates in the model of regression with errors-in-variables. Delaigle, Hall, and Meister [2008] studies both density deconvolution and regression with noisy data when the noise density η is unknown. We mention the monograph of Meister [2009] for a complete survey, including cross-validation techniques to choose the bandwidth in regression with noisy measurements. In statistical learning, Koltchinskii [2000] proposes to study different geometric characteristics of a multivariate distribution, such as the entropy dimension or the number of clusters, thanks to noisy data of the form (2.1). In a quiet related framework, minimax results in Hausdorff distance are stated in Genovese, Perone-Pacifico, Verdinelli, and Wasserman [2010] for manifold estimation of the support of a distribution thanks to noisy data.

More recently, Loustau and Marteau [2012] proposes to study a model of classification with noisy data, by giving for the first time minimax rates in binary classification with errors in variables. This paper is at the origin of other works in statistical learning with noisy data (see Loustau [2012] for supervised classification and Loustau [2013] for unsupervised problems). In these problems, the use of a deconvolution kernel estimator (1.2) is necessary to derive *excess risk bounds*. Loustau and Marteau [2012] suggests a deterministic bandwidth choice to get minimax *fast* convergence rates (i.e. faster than $n^{-1/2}$). Unfortunately, as usually, this choice depends on the unknown smoothness of the density f.

Outlines. The aim of this contribution is to get adaptive fast rates for the excess risk via a new Lepski-type procedure. To the best of our knowledge, standard adaptive procedures such as cross-validation, model selection or aggregation cannot be directly applied in our particular context (see Section 4 for details). Moreover, the Lepski's principle, which is usually used by comparing estimators for a given pointwise or global risk, cannot be directly applied to get excess risk bounds. In this contribution, we design a new selection rule based on the Lepski's principle with a comparison of *empirical risks* with different nuisance parameters. This method, called *Empirical Risk Comparison* (ERC), allows us to derive adaptive results in the context of clustering with noisy data. It could be applied to the general setting of M-estimation to derive adaptive results in other statistical problems.

The paper is organized as follows. In Section 2, we describe the model and the empirical risk minimization called noisy k-means, which proposes to use a collection of deconvolution estimators (1.2) to deal with noisy data. In Section 3, we state a non-adaptive risk bound, under a smoothness assumption over the density f, and an ill-posedness assumption over the noise distribution. These rates are reached by the noisy k-means procedure, where the bandwidth $\lambda > 0$ is chosen to trade off a bias-variance decomposition. In Section 4, we present the adaptive procedure ERC to choose the bandwidth automatically. The theoretical results guarantee the same rates of convergence, modulo an extra-log term which seems to be optimal. Finally, we conclude in Section 5 by a generalization of the ERC selection rule to other M-estimation problems, such as binary classification, local M-estimation or quantile estimation. Section

6 concludes the paper whereas Section 7-8 are dedicated to the proofs of the main results.

2 Noisy Clustering

2.1 The problem

Isolate meaningfull groups from the data is an interesting topic in data analysis with applications in many fields, such as biology or social sciences. However, in many real-life situations, direct data are not available and measurement errors occur. Then, we observe a corrupted sample of i.i.d. observations:

$$Z_i = X_i + \epsilon_i, i = 1, \dots n, \tag{2.1}$$

where f denotes the unknown density of the i.i.d. sequence $X_1, X_2, ..., X_n$ and η is the known density of the i.i.d. random variables $\epsilon_1, \epsilon_2, ..., \epsilon_n$, independent of the X_i's. In the sequel, the law of the sequence $(X_i)_{i=1,...,n}$ is denoted as P_X, with density f with respect to the Lebesgue measure on $\mathbb{R}^d$. We also assume that X_i, $i = 1, \dots, n$ are countained in $\mathcal{B}(0,1)$, the unit Euclidean ball of $\mathbb{R}^d$. This assumption is rather classical in clustering (see Bartlett, Linder, and Lugosi [1998], Levrard [2012]) and known as the peak power constraint (extension to $\mathcal{B}(0,M)$ with $M > 1$ is straightforward). Given some integer $k \geq 1$, the problem of noisy clustering is to learn k clusters from P_X when a contaminated empirical version $Z_1, \dots, Z_n$ is observed. This problem is a particular case of inverse statistical learning and is known to be an inverse problem (see Loustau [2012]). It has been studied recently in Loustau [2013], where non-adaptive results are proposed.

For this purpose, we introduce a set of codebooks $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk}$, and the standard *k-means* loss function $\gamma(\mathbf{c}, X) := \min_{j=1,...,k} \|X - c_j\|^2$, where $\|\cdot\|$ stands for the Euclidean norm on $\mathbb{R}^d$. The corresponding clustering risk of a codebook $\mathbf{c}$ is given by:

$$R(\mathbf{c}) := \mathbb{E}_{P_X} \gamma(\mathbf{c}, X) = \int_{\mathbb{R}^d} \gamma(\mathbf{c}, x) f(x) dx. \tag{2.2}$$

Given (2.2), we measure the performance of the latter codebook $\mathbf{c}$ in terms of excess risk, defined as:

$$R(\mathbf{c}, \mathbf{c}^*) := R(\mathbf{c}) - R(\mathbf{c}^*), \tag{2.3}$$

where $\mathbf{c}^* \in \arg\min_{\mathbf{c} \in \mathbb{R}^{dk}} R(\mathbf{c})$ is called an *oracle*. The oracle set is denoted as $\mathcal{M}$ and we assume in the rest of the paper that the number $|\mathcal{M}|$ of oracles is finite. This assumption is satisfied in the context of Pollard's regularity assumptions (see Pollard [1982]), i.e. when f has a continous density with respect to Lebesgue such that the Hessian matrix of $\mathbf{c} \mapsto R(\mathbf{c})$ is positive definite. In the direct case, the problem of minimizing (2.3) has been investigated in a variety of areas. For a given number of clusters $k \geq 1$, the most popular technique is the k-means procedure. It consists in partitioning the dataset $X_1, \dots, X_n$ into k clusters by minimizing the empirical risk:

$$R_n(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,...,k} \|X_i - c_j\|^2,$$

where $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk}$ is a set of centers. A cluster is associated to each observation by giving its nearest center c_j, $j = 1, \dots, k$. The k-means clustering minimization has been widely studied in the literature. Since the early work of Pollard (Pollard [1981],Pollard [1982]), consistency and rates of convergence have been considered by many authors. Biau, Devroye, and Lugosi [2008] suggests rates of convergence of the form $\mathcal{O}(1/\sqrt{n})$ whereas Bartlett, Linder, and Lugosi [1998] proposes a complete minimax study. More recently, Levrard [2012] states fast rates of the form $\mathcal{O}(1/n)$ under Pollard's regularity assumptions. It improves a previous result of Antos, Györfi, and György [2005].

However, in this paper, the problem is the knowledge of $X_1, \dots, X_n$ since we deal with a noisy dataset (2.1). For this reason, we introduce a deconvolution step in the stochastic minimization of the k-means procedure.

2.2 The noisy k-means minimization

Following Loustau [2013], the idea is to plug a deconvolution kernel estimator of the form (1.2) into the true risk (2.2). For this purpose, let us introduce the following notations. We denote by $\mathcal{F}[g]$ the Fourier transform of an integrable function g, whereas $\mathcal{F}^{-1}$ stands for the inverse Fourier transform. Let $\mathcal{K}$ be a kernel in $L_2(\mathbb{R}^d)$ such that $\mathcal{F}[\mathcal{K}]$ exists. Then, provided that $\mathcal{F}[\eta]$ exists and is strictly positive, we can introduce a deconvolution kernel $\mathcal{K}_\lambda$ as follows:

$$\begin{aligned} \mathcal{K}_\lambda \ &: \mathbb{R}^d \to \mathbb{R} \\ &t \mapsto \lambda^{-d} \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)}\right](t/\lambda), \end{aligned} \tag{2.4}$$

where $\lambda > 0$ is called the bandwidth. The kernel $\mathcal{K}$ in (2.4) is a kernel with particular properties (see Section 3.1). Note that with a slight abuse of notations, we write t/λ for the vector $(t_1/\lambda, \dots, t_d/\lambda)$. Moreover, (2.4) depends explicitely on the density η of the noise which is supposed to be known. In practice, this knowledge could be avoided using repeated measurements (see for instance Delaigle, Hall, and Meister [2008]).

Moreover, let $\mathcal{C} := \left\{\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk} \ : \ c_j \in \mathcal{B}(0,1),\ j = 1, \dots, k\right\}$ be the set of possible centers in the unit ball $\mathcal{B}(0,1)$ of the Euclidean space $\mathbb{R}^d$. We then introduce the following collection of noisy k-means minimizers:

$$\hat{\mathbf{c}}_\lambda := \arg\min_{\mathbf{c} \in \mathcal{C}} R_n^\lambda(\mathbf{c}), \quad \lambda > 0, \tag{2.5}$$

where $R_n^\lambda(\mathbf{c})$ is called the deconvolution empirical risk. This quantity is defined as:

$$R_n^\lambda(\mathbf{c}) = \int_{\mathbb{R}^d} \gamma(\mathbf{c}, x) \hat{f}_\lambda(x) dx = \frac{1}{n} \sum_{i=1}^n \gamma_\lambda(\mathbf{c}, Z_i), \tag{2.6}$$

where $\gamma_\lambda(\mathbf{c}, Z)$ is the deconvolution product:

$$\gamma_\lambda(\mathbf{c}, Z) := \left[\mathcal{K}_\lambda * \gamma(\mathbf{c}, \cdot)\right](Z) = \int_{\mathcal{B}(0,1)} \mathcal{K}_\lambda\left(Z - x\right) \gamma(\mathbf{c}, x) dx, \quad \mathbf{c} = (c_1, \dots, c_k) \in \mathcal{C}.$$

Note that the restriction to the closed unit ball $\mathcal{B}(0,1)$ appears only for technicalities, and using any compact set is possible.

Parameter λ in (2.5) is of great interest in this paper. In particular, an appropriate choice of the bandwidth allows us to get fast rates (Section 3) and adaptive results (Section 4). In minimax nonparametric estimation, the standard choice of λ trades off a *bias-variance decomposition*, that is an upper bound of the measurement error (see Tsybakov [2008] for an overview). Thanks to Loustau [2013], we can expect the same kind of upper bounds for the excess risk as follows:

$$\begin{aligned} R(\hat{\mathbf{c}}_\lambda, \mathbf{c}^*) \leq (R - R_n^\lambda)(\hat{\mathbf{c}}_\lambda, \mathbf{c}^*) \ &\leq \ (R - R^\lambda)(\hat{\mathbf{c}}_\lambda, \mathbf{c}^*) + (R^\lambda - R_n^\lambda)(\hat{\mathbf{c}}_\lambda, \mathbf{c}^*) \\ &=: \ \text{bias}(\lambda) + \text{var}(\lambda), \end{aligned} \tag{2.7}$$

where in the sequel, for any fixed $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$, $R^\lambda(\mathbf{c}, \mathbf{c}') := \mathbb{E}\left[R_n^\lambda(\mathbf{c}) - R_n^\lambda(\mathbf{c}')\right]$ and $\mathbb{E}$ is the expectation w.r.t. $P_Z^{\otimes n}$. The first part of the decomposition is called a *bias* term. It depends on the unknown smoothness $s > 0$ of the density f and on the deconvolution kernel (see Proposition 1 for details). The second term of this decomposition is called the *variance* term. It is the stochastic error of the empirical risk minimization. It depends on a standard complexity parameter and on the noise assumption (see below). This term could be controlled using empirical process theory in the spirit of Blanchard, Bousquet, and Massart [2008] (see Proposition 2). Therefore, as a first step, we derive in Section 3 fast rates of convergence from an optimal bandwidth $\bar{\lambda} := \bar{\lambda}(s)$ minimizing the latter bias-variance trade-off (see Theorem 1).

3 Fast rates for noisy clustering

In this section, we propose to give a non-adaptive excess risk bound for the noisy k-means procedure (2.5). This result is obtained under classical assumptions from both the statistical inverse problem literature and the area of fast rates. We recall and discuss these assumptions for completeness.

3.1 Main assumptions

First of all, as in standard deconvolution problems, the use of a deconvolution kernel requires some additional assumptions on the kernel $\mathcal{K} \in L_2(\mathbb{R}^d)$ in (2.4).

(K1) There exist $S = (S_1, \ldots, S_d) \in \mathbb{R}_d^+$, $K_1 > 0$ such that kernel $\mathcal{K}$ satisfies

$$\text{supp}\mathcal{F}[\mathcal{K}] \subset [-S, S] \text{ and } \sup_{t \in \mathbb{R}^d} |\mathcal{F}[\mathcal{K}](t)| \leq K_1,$$

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

This assumption is trivially satisfied for different standard kernels, such as the *sinc* kernel. This assumption arises for technicalities in the proofs and can be relaxed using a finer algebra. Moreover, in the sequel, we consider a kernel $\mathcal{K}$ of order $m \in \mathbb{N}$ as follows:

- $\int_{\mathbb{R}^d} \mathcal{K}(x)dx = 1$,
- $\int_{\mathbb{R}^d} \mathcal{K}(x)x_v^k dx = 0$, $\forall k \leq m$, $\forall v \in \{1, \ldots, d\}$,
- $\int_{\mathbb{R}^d} |\mathcal{K}(x)||x_v|^m dx < \infty$, $\forall v \in \{1, \ldots, d\}$.

For the construction of kernels of order m, the univariate case is presented in Tsybakov [2008]. Comte and Lacour [2013] have detailed the multivariate case in an anisotropic framework, where the kernel can have a different order in each direction. The construction of kernels of order m satisfying **(K1)** could be managed using for instance the so-called Meyer wavelet (see Mallat [2000]).

Moreover, we need an additional assumption on the regularity of the density f to control the bias term. In this paper, this regularity in expressed in terms of isotropic Hölder spaces.

Definition 1. *Fix $s > 0$ and $L > 0$, and let $\lfloor s \rfloor$ be the largest integer strictly less than s. The* isotropic Hölder class $\Sigma_d(s, L)$ *is the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ having on $\mathbb{R}^d$ all partial derivatives of order $\lfloor s \rfloor$ and such that for any $x, y \in \mathbb{R}^d$:*

$$\left| \frac{\partial^{|p|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} - \frac{\partial^{|p|} f(y)}{\partial y_1^{p_1} \cdots \partial y_d^{p_d}} \right| \leq L \sum_{v=1}^d |x_v - y_v|^{s - \lfloor s \rfloor}, \quad \forall\, p \in \mathbb{N}^d \ : \ |p| := p_1 + \cdots + p_d = \lfloor s \rfloor;$$

$$\sum_{m=0}^{\lfloor s \rfloor} \sum_{|p|=m} \sup_{x \in \mathbb{R}^d} \left| \frac{\partial^{|p|} f(x)}{\partial x_1^{p_1} \cdots \partial x_d^{p_d}} \right| \leq L,$$

where x_v and y_v are the v^{th} components of x and y.

In the sequel, we assume that the multivariate density f of the law P_X belongs to the isotropic Hölder class $\Sigma_d(s, L)$, for some $s, L > 0$. It means that the density f has a similar regularity in any direction. An extension to the anisotropic Hölder class is given in Loustau [2013], which states fast rates in this case. As in standard density estimation or deconvolution, the bandwidth choice is more nasty and depends explicitely on the direction (see also Comte and Lacour [2013]). It is out of the scope of the present paper.

We also need an assumption on the noise distribution η as follows:

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

$$|\mathcal{F}[\eta](t)| \geq \rho \prod_{v=1}^d \left(\frac{t_v^2 + 1}{2} \right)^{-\beta_v/2}.$$

NA(ρ, β) deals with a lower bound on the behaviour of the characteristic function of the noise density η. This lower bound is a sufficient condition to get excess risk bounds. However, to study the optimality in the minimax sense, we need an upper bound of the same order for the characteristic funtion. This is not the purpose of this paper. Moreover, this noise assumption is related with a polynomial behaviour of

the Fourier transform of η. This case is called the mildly ill-posed case in the deconvolution or statistical inverse problem literature (see Meister [2009]). 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 Comte and Lacour [2013] in multivariate deconvolution).

Finally, to reach fast rates of convergence, we need to introduce a margin assumption. This type of assumption is now standard in classification since the work of Tsybakov (Mammen and Tsybakov [1999] or Tsybakov [2004]). In clustering, we use the following version of the well-known margin assumption (see Bartlett and Mendelson [2006] for a related point of view):

Margin Assumption MA(κ): For any $\mathbf{c} \in \mathcal{C}$, there exists some positive constant κ such that:

$$|||\mathbf{c} - \mathbf{c}^*(\mathbf{c})|||^2 \leq \kappa \big[R(\mathbf{c}) - \inf_{\mathbf{c}' \in \mathbb{R}^{dk}} R(\mathbf{c}')\big],$$

where $\mathbf{c}^*(\mathbf{c}) \in \mathcal{M}$ is the nearest optimal cluster associated to $\mathbf{c}$ and $|||\cdot|||$ stands for the Euclidean norm in $\mathbb{R}^{dk}$.

The margin assumption proposes a control of the Euclidean norm by the excess risk. Since we restrict the study to a compact set, it is easy to see that $\mathbf{MA}(\kappa)$ allows us to write:

$$\|\gamma(\mathbf{c}, X) - \gamma(\mathbf{c}^*(\mathbf{c}), X)\|^2_{L_2(P_X)} \leq C_1 |||\mathbf{c} - \mathbf{c}^*(\mathbf{c})|||^2 \leq C_1 \kappa \mathbb{E}_{P_X}\big[\gamma(\mathbf{c}, X) - \gamma(\mathbf{c}^*(\mathbf{c}), X)\big].$$

As a result, we can use a localization principle and reach fast rates of convergence.

The introduction of a margin assumption in clustering is actually not a novelty. It is strongly related with some well-known regularity assumptions involved in the study of the consistency of the k-means procedure (see Pollard [1982],Antos, Györfi, and György [2005]). Indeed, as shown in Antos, Györfi, and György [2005], $\mathbf{MA}(\kappa)$ is satisfied if f is continuous and the Hessian matrix of the mapping $\mathbf{c} \longmapsto R(\mathbf{c})$ is positive definite at any point $\mathbf{c}^* \in \mathcal{M}$. In this case, the constant κ is related with the smallest eigenvalue of the Hessian matrix. These conditions have been introduced by Pollard to get limit theorems for the k-means.

Finally, Levrard [2012] has interpreted Pollard's regularity assumption in terms of well-separated classes as follows. For any $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{C}$, we associate to each center c_i, $i = 1, \ldots, k$ the Voronoï cell $V_i(\mathbf{c})$ defined as:

$$V_i(\mathbf{c}) = \{x \in \mathbb{R}^d : \min_{j=1,\ldots,k} \|x - c_j\| = \|x - c_i\|\}.$$

Let $\partial V_i(\mathbf{c})$ be the boundary of the Voronoi cell $V_i(\mathbf{c})$ associated with c_i, for $i = 1, \ldots, k$. Then, a sufficient condition to have a continuous density f and a positive definite Hessian matrix is to control the sup-norm of f on the union of all possible $|\mathcal{M}|$ boundaries $\partial V(\mathbf{c}^*) = \cup_{i=1}^k \partial V_i(\mathbf{c}^*)$, associated with $c^* \in \mathcal{M}$ as follows:

$$\|f_{|\cup_{\mathbf{c}^* \in \mathcal{M}} \partial V(\mathbf{c}^*)}\|_\infty \leq T(d) \inf_{\mathbf{c}^* \in \mathcal{M}, i=1,\ldots k} P_X(V_i(\mathbf{c}^*)),$$

where $T(d)$ is a constant depending on the dimension d. As a result, the margin assumption $\mathbf{MA}(\kappa)$ is guaranteed when the source distribution P_X is well concentrated around its optimal clusters, which is related to well-separated classes.

3.2 A first excess risk bound

We now present an excess risk bound for the collection of estimators introduced in (2.5), under the previous assumptions.

Theorem 1. *Assume that* $\mathbf{NA}(\rho, \beta)$ *and* $\mathbf{MA}(\kappa)$ *are satisfied for some* $\beta \in (1/2, \infty)^d$, $\rho, \kappa > 0$. *Suppose* $\eta_\infty := \|\eta\|_\infty < \infty$ *and* $f \in \Sigma_d(s, L)$ *with* $s, L > 0$. *Then, denoting by* $\hat{\mathbf{c}}_n^{\overline{\lambda}}$ *a solution of* (2.5) *with:*

$$\overline{\lambda} = n^{-1/(2s+2\bar{\beta})},$$

there exists a universal constant C_1 *depending on* $\kappa, w, L, d, s, \beta, \rho, k, \eta_\infty$ *and* $|\mathcal{M}|$, *and an integer* $n_0 \in \mathbb{N}^*$ *such that for any* $\mathbf{c}^* \in \mathcal{M}$ *and any* $n \geq n_0$*:*

$$\mathbb{E} R(\hat{\mathbf{c}}_{\overline{\lambda}}, \mathbf{c}^*) \leq C_1 n^{-s/(s+\bar{\beta})},$$

where $\bar{\beta} = \sum_{v=1}^d \beta_v$.

The proof is an application of a localization approach in the spirit Massart [2007], applied to the noisy set-up. As in Loustau [2012], the decomposition (2.7) allows us to control the excess risk. More precisely, the variance can be controlled by mixing empirical process argues as in Blanchard, Bousquet, and Massart [2008], gathering with the noise assumption $\mathbf{NA}(\rho, \beta)$. The bias term is bounded using both the smoothness of f and the margin assumption $\mathbf{MA}(\kappa)$.

Theorem 1 improves the previous result of Loustau [2013] in the particular case of finite dimensional clustering, where a $\sqrt{\log\log(n)}$ term appears in the RHS (see Theorem 3 in Loustau [2013]). Rates of convergence of Theorem 1 are fast rates when $\bar{\beta} < s$. It generalizes the result of Levrard [2012] to the errors-in-variables case since we can see coarsely that rates to the order $\mathcal{O}(1/n)$ are reached when $\epsilon = 0$. Here, the prize to pay for the inverse problem is the quantity $\sum_{i=1}^{d} \beta_i$, related to the tail behavior of the characteristic function of the noise distribution η in $\mathbf{NA}(\rho, \beta)$.

An open problem is to derive the optimality of Theorem 1 in the minimax sense, under the margin assumption $\mathbf{MA}(\kappa)$ and the noise assumption $\mathbf{NA}(\rho, \beta)$. In this direction, Loustau and Marteau [2012] proposes a complete minimax study in classification with error-in-variables by using a comparable estimation procedure. We then conjecture that the rate $n^{-s/(s+\bar{\beta})}$ is minimax over Hölder spaces.

4 Bandwidth Selection

In this section, we turn out into the main issue of this paper: the data-driven choice of the bandwidth $\lambda > 0$ in the collection of estimators $\{\hat{\mathbf{c}}_\lambda, \lambda > 0\}$ defined in 2.5. The goal is to reach adaptive excess risk bound similar to Theorem 1 for a choice of λ which does not depend on the smoothness of f.

In supervised learning (such as regression or binary classification), it is standard to choose a bandwidth - or a tuning - parameter using a decomposition of the set of observations. A training set is used to construct a family of candidate estimators, each one associated with a different value of the bandwidth. Then, a test set allows to estimate the generalization performances of each candidate. It gives rise to the family of cross-validation methods, or aggregation procedures. Unfortunately, in unsupervised tasks, this simple estimation is not possible. The lack of efficiency of cross-validation methods in clustering has been illustrated in Hastie, Tibshirani, and Friedman [2002] for the problem of choosing k in the k-means. Moreover, in the presence of errors in variables, such as in deconvolution, it is also quiet standard to perform cross-validation to choose the bandwidth of a deconvolution estimator. As described in Meister [2009], it is possible to estimate the squared risk $\|\hat{f}_\lambda - f\|^2$ with Plancherel theorem, leading to the estimation of the Fourier transform of the unknown density. However, in our framework, this method seems hopeless since the optimal value of λ does not minimize a squared risk but an excess risk of the form (2.3). Eventually, model selection was introduced for selecting the hypothesis space over a sequence of nested models (e.g. finite dimension models) with a fixed empirical risk. Penalization methods are also suitable to choose smoothing parameters of well-known statistical methods such as splines, SVM or Tikhonov regularization methods. The idea is to replace the choice of the smoothing parameter by the choice of the radius into a suitable ellipsoid. Unfortunately, here, the nuisance parameter λ affects directly the empirical risk (2.6), and a model selection method can not be directly applied in this context.

Theorem 1 below motivates the use of a comparison method based on the Lepski's principle (Lepski [1990]). Indeed, the non-adaptive choice of $\bar{\lambda} = n^{-1/(2s+2\bar{\beta})}$ in Theorem 1 trades off a bias-variance decomposition of the excess risk and allows to get fast rates of convergence. As a result, the Lepski's principle appears as the most commonly tool to propose an adaptive estimator $\hat{\mathbf{c}}_{\hat{\lambda}}$, where $\hat{\lambda}$ mimics the oracle $\bar{\lambda}$ of Theorem 1. The built estimator $\hat{\mathbf{c}}_{\hat{\lambda}}$ will be called adaptive since it does not depend on the smoothness s.

To define the selection rule, we first remind some definitions and notations. Given a kernel $\mathcal{K}$ satisfying the previous assumptions, we note $\|\mathcal{K}\|_1$ the L_1-norm of the kernel on $\mathbb{R}^d$. The constant $\eta_\infty := \|\eta\|_\infty$ is the sup-norm of the noise density η, whereas $\rho > 0$ and $\bar{\beta} = \sum_{v=1}^{d} \beta_v$ are parameters involved in the noise assumption $\mathbf{NA}(\rho, \beta)$. Moreover, κ is the constant in the margin assumption $\mathbf{MA}(\kappa)$. In the sequel, $\mathcal{V}(d) = \pi^{d/2}/\Gamma(d/2+1)$, where $\Gamma(\cdot)$ stands for the Gamma function.
Define the threshold term:

$$\delta_\lambda := \frac{2^{10}\sqrt{2}\mathcal{V}(d)\|\mathcal{K}\|_1^2 \kappa \eta_\infty}{\rho^2} \frac{\lambda^{-2\bar{\beta}} \log(n)}{n}, \tag{4.1}$$

where λ belongs to the bandwidth set $\Lambda := [\lambda_{\min}, \lambda_{\max}]$ with

$$\lambda_{\min} := \frac{\log^{1/\bar{\beta}}(n)}{n^{1/2\bar{\beta}}} \text{ and } \lambda_{\max} := \big(1/\log(n)\big)^{1/(2s^+ + 2\bar{\beta})},$$

where $s^+ > 0$ is an upper bound on the regularity index of f. In this section, we take n sufficiently large such that $n^{-1/(2s+2\bar{\beta})} \in \Lambda$. Moreover, for some constant $a \in (0,1)$, we set:

$$\Lambda_a := \{\lambda \in \Lambda : \exists m \in \mathbb{N}\,,\ \lambda = \lambda_{\max} a^m\}\,,$$

a discrete exponential net on the bandwidth set with cardinal $|\Lambda_a|$.

We are ready to introduce the adaptive bandwidth choice, called ERC (Empirical Risk Comparison):

$$\hat{\lambda} = \max\left\{\lambda \in \Lambda_a \ :\ R_n^{\lambda'}(\hat{\mathbf{c}}_\lambda) - R_n^{\lambda'}(\hat{\mathbf{c}}_{\lambda'}) \le 3\delta_{\lambda'},\ \forall \lambda' \le \lambda\right\}. \tag{4.2}$$

The noisy k-means estimator (2.5) with bandwidth $\hat{\lambda}$ chosen from ERC rule (4.2) has the following property.

Theorem 2. *Assume that* $\mathbf{NA}(\rho,\beta)$ *and* $\mathbf{MA}(\kappa)$ *are satisfied for some* $\beta \in (1/2,\infty)^d$, $\rho,\kappa > 0$. *Suppose* $\eta_\infty := \|\eta\|_\infty < \infty$ *and* $f \in \Sigma(s,L)$, *where* $s \in [0, s^+)$ *and* $L > 0$. *Then, there exists a universal constant* C_2 *depending on* $\kappa, w, L, d, s, \beta, \rho, k, \eta_\infty$, $|\mathcal{M}|$, *and* $n_1 \in \mathbb{N}$ *such that for any* $\mathbf{c}^* \in \mathcal{M}$ *and any* $n \ge n_1$, *estimator* $\hat{\mathbf{c}}_n^{\hat{\lambda}}$ *with* $\hat{\lambda}$ *selected by ERC* (4.2) *satisfies:*

$$\mathbb{E}R(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) \le C_2 \left(\frac{\log(n)}{n}\right)^{s/(s+\bar{\beta})},$$

where $\bar{\beta} = \sum_{v=1}^d \beta_v$.

Theorem 2 is an adaptive upper bound for the estimator $\hat{\mathbf{c}}^{\hat{\lambda}}$, where $\hat{\lambda}$ is chosen from the ERC selection rule (4.2). The estimator $\hat{\mathbf{c}}^{\hat{\lambda}}$ is then adaptive w.r.t. the smoothness s. This adaptive excess risk bound coincides with the non-adaptive previous result of Theorem 1, up to an extra log term. This is the prize to pay for the data-driven property of the procedure.

Let us remind that it is standard to pay a $\log(n)$ factor in pointwise estimation (see Lepski [1990] and Brown and Low [1996]). However, it is well-known that there is no prize to pay for adaptation in global estimation (e.g. L_p-norm). In the problem of noisy clustering, the choice of λ concerns the estimation of the density f. This estimation is used in the procedure of noisy k-means, where we plug $\hat{f}_\lambda$ into the true risk. At the first glance, we can conjecture that a global estimation of f is sufficient. However, a closer look at the problem of noisy clustering, or more generally noisy classification, clearly highlights the nature of the estimation problem we have at hand: a pointwise problem of estimation of a density f thanks to noisy or corrupted measurements (see Loustau [2012], Theorem 3). As a result, we can conjecture that the result of Theorem 2 is optimal, in the adaptive minimax sense, as in standard pointwise estimation.

The threshold term δ_λ - which comes from the control of the stochastic part of the excess risk, see Proposition 2 - has the following form (see (4.1)):

$$\delta_\lambda = C_{\text{adapt}} \frac{\lambda^{-2\bar{\beta}} \log(n)}{n},$$

where the (large) constant $C_{\text{adapt}} > 0$ depends on the margin constant κ which is possibly unknown. Indeed, by definition, it depends on the underlying density f. In practice, we recommend a painstaking calibration of this constant. From the theoretical point of view, this constant could be chosen from the *propagation* method suggested by Spokoiny and Vial [2009].

The proof of Theorem 2, given in Section 7, is based on the standard Lepski's principle subject to major modifications. This rule traditionally uses a comparison between estimators (possibly defined as minimizers) indexed by a nuisance parameter to get the best one from the considered family. In our context, we have proposed to compare empirical risks indexed by the nuisance parameter (bandwidth) to control the performance of the selected risk minimizer in terms of excess risk. Note that this idea was

suggested by Polzehl and Spokoiny [2006] via local likelihood Comparison with the Kullback divergence as measurement error. To the best of our knowledge, it is the first time that the Lepski's principle is applied in statistical learning on this way.

In the following section, we give sufficient conditions to apply the ERC method in the more general context of M-estimation depending on a nuisance parameter. Many examples could be considered in future works.

5 ERC's extension

In this section, we propose an extension of the ERC selection rule (4.2) to a more general context of M-estimation that we illustrate by some examples. To this end, let us introduce the random variable $\mathcal{Z}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with law $\mathbb{P}_{\mathcal{Z}}$ defined on $\mathcal{X}$. Given a sample of i.i.d. random variables $\mathcal{Z}_1, \dots, \mathcal{Z}_n \in \mathcal{X}$ with probability law $\mathbb{P}_{\mathcal{Z}}$, we construct an empirical risk denoted by $\mathcal{R}_n^\lambda(\cdot)$, where $\lambda > 0$ is a bandwidth to choose via ERC method. This parameter can be the bandwidth of a standard kernel in a localization approach (see Example 4. and 5.), or the bandwidth of a deconvolution kernel in error-in-variables models (see Example 1., 2. and 3.). We consider the collection of M-estimators:

$$\hat{g}_n^\lambda \in \arg\min_{g \in \mathcal{G}} \mathcal{R}_n^\lambda(g), \quad \lambda > 0, \tag{5.1}$$

where $\mathcal{G}$ is a fixed family of candidates that depends on the considered problem. Given the collection of estimators (5.1), we focus on the selection of the nuisance parameter $\lambda > 0$ appearing in the empirical risk. Without loss of generality, we assume that λ belongs to the set $\Delta_n = [\lambda_-, \lambda^+]$, where $\lambda_- \leq \lambda^+$ and $\lambda_-, \lambda^+ \to 0$ as $n \to \infty$.

The risk of the estimators (5.1) is measured thanks to a quantity $\mathcal{R}(\hat{g}_n^\lambda)$, called the true risk. To obtain good properties for $\hat{g}_n^\lambda$, the first condition on $\mathcal{R}_n^\lambda(\cdot)$ is that it has to be an asymptotically unbiased estimator of the true risk $\mathcal{R}(\cdot)$. To this end, for any fixed estimator $g \in \mathcal{G}$, we introduce the expectation of the empirical risk as $\mathcal{R}^\lambda(g) = \mathbb{E}\mathcal{R}_n^\lambda(g)$, where $\mathbb{E}$ denotes the expectation w.r.t. the product probability of the training set $(\mathcal{Z}_1, \dots, \mathcal{Z}_n)$. Then, the empirical risk has the following property:

$$\lim_{n \to \infty} \mathbb{E}\mathcal{R}_n^\lambda(g) = \mathcal{R}(g), \text{ for any fixed } g \in \mathcal{G}.$$

We also denote as g^* the best possible decision rule in $\mathcal{G}$, called an oracle, defined as:

$$g^* \in \arg\min_{g \in \mathcal{G}} \mathcal{R}(g).$$

Eventually, the performance of the estimate $\hat{g}_n^\lambda$ is measured via the excess risk $\mathcal{R}(\hat{g}_n^\lambda, g^*) := \mathcal{R}(\hat{g}_n^{\hat{\lambda}}) - \mathcal{R}(g^*)$.

As illustrated in Section 3 for the particular case of noisy clustering, the performances of $\hat{g}_n^\lambda$ in terms of excess risk is strongly related with the behaviour of $\mathcal{R}_n^\lambda(g)$ as an estimator of the true risk $\mathcal{R}(g)$. The message of this section is the following : if the behaviour of $\mathcal{R}_n^\lambda(g)$ in (5.1) can be decomposed into a standard bias-variance decomposition, an adaptive choice $\hat{\lambda}$ can be proposed using the ERC rule. It leads to the following assumptions that are sufficient conditions to propose an optimal adaptive procedure:

Bias/Variance Conditions.

1. *There exists an increasing function denoted by* $\boldsymbol{Bias}(\cdot)$ *such that:*

$$\left|(\mathcal{R}^\lambda - \mathcal{R})(g, g^*)\right| \leq \boldsymbol{Bias}(\lambda) + \frac{1}{4}\mathcal{R}(g, g^*), \quad \textit{for all } g \in \mathcal{G} \textit{ and } \lambda \in \Delta.$$

2. *There exists a decreasing function denoted by* $\boldsymbol{Var}_t(\cdot)$ *($t \geq 0$) such that:*

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} \left\{ \left|(\mathcal{R}_n^\lambda - \mathcal{R}^\lambda)(g, g^*)\right| - \frac{1}{4}\mathcal{R}(g, g^*) \right\} > \boldsymbol{Var}_t(\lambda)\right) \leq e^{-t}, \quad \textit{for all } \lambda \in \Delta \textit{ and } t \geq 0.$$

The above conditions ensure a control of the excess risk using the bias-variance decomposition (2.7). It allows to find a quantity λ^* which trades off both terms $\mathbf{Bias}(\lambda)$ and $\mathbf{Var}_t(\lambda)$. In particular, using (2.7), Bias/Variance conditions *1.* and *2.* lead to the following exponential inequality:

$$\mathbb{P}\Big(\mathcal{R}(\hat{g}_n^\lambda) - \mathcal{R}(g^*) \geq 2\mathbf{Bias}(\lambda) + 2\mathbf{Var}_t(\lambda)\Big) \leq e^{-t}, \quad \text{for all } t \geq 0.$$

A close inspection of the proof of Theorem 2 highlights that Conditions *1.* and *2.* are sufficient to show the adaptive property of the ERC rule in this general context as follows. Let Δ_a be a discrete exponential net on the set Δ. Then, the general ERC rule is given by:

$$\hat{\lambda} = \max\Big\{\lambda \in \Delta_a \ : \ \mathcal{R}_n^{\lambda'}(\hat{g}_n^\lambda) - \mathcal{R}_n^{\lambda'}(\hat{g}_n^{\lambda'}) \leq 8\mathbf{Var}_t(\lambda),\ \forall \lambda' \leq \lambda\Big\}. \tag{5.2}$$

Thus, using the latter Bias/Variance conditions, one can establish oracle inequatilies for the selection rule (5.2) as follows:

Theorem 3. *Suppose the Bias/Variance Conditions 1. and 2. holds. Consider the family* $\{\hat{g}_n^\lambda, \lambda > 0\}$ *defined in* (5.1). *Then, there exists a universal constant* C_3 *such that*

$$\mathbb{E}\mathcal{R}(\hat{g}_n^{\hat{\lambda}}, g^*) \leq C_3\left(\inf_{\lambda\in\Delta}\Big\{\boldsymbol{Bias}(\lambda) + \boldsymbol{Var}_t(\lambda)\Big\} + e^{-t}\right), \quad \textit{for all } t \geq 0,$$

where $\hat{\lambda}$ *is chosen in* (5.2).

We omit the proof of Theorem 3 since it can be deduced coarselly from the proof of Theorem 2. This result allows to get a control of the expected excess risk of $\hat{g}_n^{\hat{\lambda}}$ via a data-driven parameter $\hat{\lambda}$. The adaptive estimator performs as well as the best one, i.e. the one minimizing the Bias/Variance trade-off in the family $\{\hat{g}_n^\lambda,\ \lambda \in \Lambda\}$. This result could be of great interest in many statistical learning context. We now give some examples in the context of noisy data, and then we turn out into the problem of local M-estimation.

Example 1: Noisy Clustering. The framework of Section 3 exactly falls into the general model of this section. Indeed, in the problem of clustering with noisy inputs (2.1), the empirical risk is defined as in (2.6):

$$\mathcal{R}_n^\lambda(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^n \int \mathcal{K}_\lambda\left(Z_i - x\right) \min_{j=1,\dots,k} \|x - c_j\|_2^2 dx$$

where $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^{dk}$ and $\lambda > 0$ is the bandwidth of the deconvolution estimator. Conditions *1.* and *2.* correspond to Propositions 1 and 2 (see Section 7).

Example 2: Discriminant Analysis (Loustau and Marteau [2012]). The model of discriminant analysis is the supervised counterpart of the clustering problem. Suppose we have at hand two samples $Z_1^1, \dots, Z_n^1$ and $Z_1^2, \dots, Z_m^2$ of the form (2.1), with density $p * \eta$ and $q * \eta$. The aim is to design a decision set $G \subset \mathbb{R}^d$ such that if $x \in G$, x is associated to the density p, and q otherwise. In this context, Loustau and Marteau [2012] states minimax fast rates of convergence for the excess risk of classsification $\mathcal{R}(G) - \mathcal{R}(G^*)$, where $\mathcal{R}(G)$ is defined as:

$$\mathcal{R}(G) = \frac{1}{2}\left[\int_{G^C} p(x)dx + \int_G q(x)dx\right].$$

Using a deconvolution kernel of the form (2.4), we can introduce an asymptotically unbiased estimator of $\mathcal{R}(G)$ given by:

$$\mathcal{R}_n^\lambda(G) = \frac{1}{2}\left[\int_{G^C} \hat{p}_\lambda(x)dx + \int_G \hat{q}_\lambda(x)dx\right],$$

where $\hat{p}_\lambda$ and $\hat{q}_\lambda$ are deconvolution kernel estimators of p and q with given bandwidth $\lambda > 0$. In this context, Loustau and Marteau [2012] have proved minimax fast rates of convergence for the excess risk $\mathcal{R}(\hat{G}_n^\lambda) - \mathcal{R}(G^*)$, using a bias-variance decomposition as in (2.7). The generalization of ERC in this particular minimax framework illustrates that a minimax adaptive excess risk bounds could be stated by using the ERC rule (5.2), up to an extra log term.

Example 3: Quantile estimation (Dattner, Reiß, and Trabs [2013]). Given noisy data (2.1), the objective of quantile estimation is to estimate a τ-quantile of the distribution P_X, which is given by:

$$q_\tau := \arg\min_{\eta\in\mathbb{R}} \int_{-\infty}^{+\infty} (x-\eta)(\tau - \mathbb{1}_{x\leq\eta}) f(x)dx,$$

where $\tau \in (0,1)$. In this problem, q_τ can be seen as the oracle associated with the following risk:

$$\mathcal{R}(\eta) := \int_{-\infty}^{+\infty} (x-\eta)(\tau - \mathbb{1}_{x\leq\eta}) f(x)dx.$$

Dattner, Reiß, and Trabs [2013] proposes to estimate q_τ using a deconvolution kernel estimator of the form (2.4). The proposed estimator is defined as:

$$\hat{q}_\tau^\lambda = \arg\min_{\eta\in\mathbb{R}} \int_{-\infty}^{+\infty} (x-\eta)(\tau - \mathbb{1}_{x\leq\eta}) \hat{f}_n^\lambda(x)dx,$$

where $\hat{f}_\lambda(\cdot)$ is given in (1.2) with bandwidth $\lambda > 0$. Interestingly, they prove convergence rates for the quantity $|q_\tau - \hat{q}_\tau^\lambda|$, (provided that $f \in \Sigma(s, L)$) for an optimal bandwidth $\lambda := \bar{\lambda}(s)$ which trades off a bias-variance decomposition similar to (2.7). As usually, the choice is non-adaptive, and the authors use the standard Lepski's procedure to give adaptive upper bounds for the measurement error $|q_\tau - \hat{q}_\tau^\lambda|$.

We conjecture that Theorem 3 could be stated from the previous mentioned work. Thus, ERC method (5.2) could be used to obtain sharper risk bounds, i.e. adaptive upper bounds for the excess risk $\mathcal{R}(\hat{q}_\tau^{\hat{\lambda}}) - \mathcal{R}(\eta)$. In particular, this result could recover Dattner, Reiß, and Trabs [2013].

Example 4: Local Fitted Likelihood (Polzehl and Spokoiny [2006]). Let us introduce a sample of independent random variables $\mathcal{Z}_i = (X_i, Y_i) \in [0,1]^n \times \mathbb{R}^n$, $i = 1, \dots, n$, where P_{θ_i} denotes the distribution of Y_i with given probability density $p(\cdot, \theta_i)$, with a parameter $\theta_i = \theta(X_i)$. In Polzehl and Spokoiny [2006], the aim is to estimate the quantity $\theta(x)$ at a given point x (pointwise estimation). This model contains standard nonparametric problems such as, for example:

- Gaussian regression where $Y_i = \theta(X_i) + \epsilon_i$, $i = 1, \dots, n$, with gaussian errors ϵ_i;
- Binary classification model where $\theta(x) = \mathbb{P}(Y_i = 1 | X = x)$;
- Inhomogeneous exponential model $p(\cdot, \theta_i) = e^{-\cdot/\theta_i}/\theta_i$;
- Inhomogeneous Poisson $\mathbb{P}(Y_i = k | X_i) = \theta_i^k e^{-k/\theta_i}/k!$, $k \in \mathbb{N}$.

In this problem, one usually applies the local version of the well-known likelihood method. It gives rise to the minimization of a localized log-likelihood as follows:

$$\hat{\theta}^\lambda \in \arg\min_{t\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n -\log\left(p(Y_i, t)\right) \frac{1}{\lambda}\mathcal{K}\left(\frac{X_i - x}{\lambda}\right),$$

where $\mathcal{K}(\cdot)$ is a kernel function and $\lambda > 0$ plays the role of a bandwidth. In this context, the localized likelihood depends on a bandwidth. Moreover, in such a framework, the accuracy of a given $\hat{\theta}^\lambda$ can be measured thanks to the Kullback-Leibler divergence between the two probability P_θ and $P_{\hat{\theta}^\lambda}$, given by

$$\mathcal{K}(\theta, \hat{\theta}^\lambda) = \mathbb{E}_{P_\theta} \log \frac{p(Y, \theta)}{p(Y, \hat{\theta}^\lambda)}.$$

Polzehl and Spokoiny [2006] proposes a data-driven selection of λ using Lepski's principle comparing localized likelihoods. This rule is similar to ERC method (5.2) where the empirical risk corresponds to the localized likelihood, which is an asymptotically unbiased estimator of the risk $\mathcal{R}(\cdot) := -\mathbb{E}_{P_\theta} \log p(Y, \cdot)$. In this framework, the excess risk $\mathcal{R}(\hat{\theta}^\lambda) - \mathcal{R}(\theta)$ associated to the likelihood is the Kullback-Liebler divergence:

$$\mathcal{R}(\hat{\theta}^\lambda) - \mathcal{R}(\theta) = \mathcal{K}(\theta, \hat{\theta}^\lambda).$$

Eventually, we note that Bias/Variance Conditions can be established thanks to a margin assumption $\mathcal{K}(\theta, \hat{\theta}^\lambda) \sim I^{-1}|\theta, \hat{\theta}^\lambda|$, where I denotes the Fisher information. Then, ERC method is relevant and coincides with the adaptive method introduced by Polzehl and Spokoiny [2006].

Example 5: Robust M-estimation (Chichignoud and Lederer [2013]). Let us consider non-parametric regression where we observe a sequence of i.i.d. random variables $\mathcal{Z}_i = (X_i, Y_i) \in [0,1]^n \times \mathbb{R}^n$, $i = 1, \dots, n$ satisfying:

$$Y_i = f(X_i) + \xi_i,$$

where the design X_i is uniformly distributed on $[0,1]$ and the noise ξ_i is symmetric, possibly heavy-tailed and such that $\mathbb{P}(\text{-}1 \leq \xi_i \leq 1) > 0$. We are interested into the pointwise estimation of the regression function $f : [0,1] \to \mathbb{R}$ at a given x_0. In such a framework, localized approaches (e.g. with a kernel function) are usually used. In particular, we can focus on the local Huber estimator as follows:

$$\hat{f}_\lambda(x_0) := \arg\min_{t \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho_H(Y_i - t) \frac{1}{\lambda} \mathcal{K}\left(\frac{X_i - x_0}{\lambda}\right),$$

where $\rho_H(u) = (u^2/2)\mathbb{1}_{|u| \leq 1} + (|u| - 1/2)\mathbb{1}_{|u|>1}$ is the Huber contrast (cf. Huber [1964]), $\mathcal{K}(\cdot)$ is a kernel function and $\lambda > 0$ is the so-called bandwidth. The estimator was recently investigated by Chichignoud and Lederer [2013].

Following the approach of Theorem 3, we can use the general ERC rule to get adaptive upper bounds for the mean square error of $\hat{f}_{\hat{\lambda}}(x_0)$. Indeed, as a first step, since the Huber contrast is smooth enough, a margin assumption holds:

$$\mathcal{R}(t, f(x_0)) \sim \frac{|t - f(x_0)|}{\mathbb{E}\rho''(\xi)},$$

for t closed to $f(x_0)$. Then, from a careful look at the proof of Chichignoud and Lederer [2013], there is nice hope that the Bias/Variance Conditions are satisfied. Eventually, a direct application of Theorem 3 is possible.

6 Conclusion

This paper could be seen as a first step into the study of adaptive noisy clustering. Several problems remain open and could be the core of future works.

Firstly, we obtain in Theorem 1 a non-adaptive excess risk bound in clustering with noisy data for a collection of deconvolution empirical risk minimizers. This bound highlights the presence of fast rates of convergence, which improves the previous result stated in Loustau [2013] using Koltchinskii's localization approach (see Koltchinskii [2006]). Here, fast rates of the form $\mathcal{O}(n^{-s/(\bar{\beta}+s)})$ are obtained, where $s > 0$ is the Hölder regularity of the density f and $\bar{\beta}$ deals with the asymptotic behaviour of the characteristic function of the noise. These rates are reached for a non-adaptive bandwidth choice $\bar{\lambda} = n^{-1/2(\bar{\beta}+s)}$. Then, we turn out into the main issue of the paper: the adaptive choice of the bandwidth of the estimator. We introduce a new selection rule based on Lepski's heuristic, where empirical risks are compared instead of estimators. This rule, called ERC, allows us to get an adaptive excess risk bound which coincides with the previous upper bound, up to a $\log(n)$ term. This prize to pay for the adaptivity seems to be optimal, as discussed at the end of Theorem 2.

The introduction of the ERC rule in noisy clustering leads to several open problems. First of all, the proposed selection rule suffers from the dependence on the threshold term δ_λ in (4.1), which depends on unknown constants, such as the margin constant κ in $\mathbf{MA}(\kappa)$. An interesting but challenging open problem is to investigate the adaptivity with respect to the margin assumption. Moreover, this threshold is not realistic and a precise calibration of this term is of practical interest. In this direction, it could be interesting to develop the propagation method presented in Spokoiny and Vial [2009]. Another interesting direction is to extend the result of this paper to the anisotropic case. From Loustau [2013], we know the presence of fast rates of convergence for the collection of noisy k-means, where the density f has an anisotropic Hölder regularity. These results are quiet similar to the isotropic case, except that the choice of the bandwidth is more nasty. Exactly as in a deconvolution framework, in the anisotropic case, the optimal bandwidth is not the same in each direction. As a result, the problem of adaptation is more difficult in this case, since we have to consider a d-dimensional grid of parameters λ_j, $j = 1, \dots, d$. For this purpose, the application of a ERC based on Goldenshluger and Lepski [2011] is a challenging open problem.

The construction of an algorithm to compute the ERC rule is also of first interest. This could be done thanks to the recent developments stated in Brunet and Loustau [2013], where a noisy k-means algorithm is proposed. Then, it could be interesting to test over simulated as well as real datasets the problem of choosing the bandwidth in the algorithm. An implementation of the ICI algorithm will be efficient to avoid the calculation of all the estimators in the collection of noisy k-means.

Finally, in this contribution, the Lepski's heuristic is introduced for the first time to get adaptive excess risk bounds in statistical learning. As discussed in Section 5, the data-driven choice of the bandwidth proposed in this paper can be applied to a more general context of M-estimation, where a nuisance parameter appears in the empirical risk. In this general context, ERC appears to be useful to get adaptive excess risk bounds, provided that a suitable bias-variance decomposition is available for the empirical risk (see Theorem 3). We conjecture that the guiding thread of this paper could be use in a variety of statistical models, where parameter selection is involved.

7 Proofs

7.1 Basic Results

Proofs of Theorem 1-2 are based on the following two basic propositions.

Proposition 1 (Bias control)**.** *Suppose $f \in \Sigma_d(s, L)$. Let $\mathcal{K}(\cdot)$ a kernel of order $\lfloor s \rfloor$ with respect to ν, and assume that $\mathbf{MA}(\kappa)$ is satisfied for some positive constant κ. For any $\epsilon > 0$ and any $\lambda > 0$, it holds:*

$$\left|(R - R^\lambda)(\mathbf{c}, \mathbf{c}^*)\right| \leq \zeta_1 \kappa \epsilon \lambda^{2s} + \frac{1}{2\epsilon} R(\mathbf{c}, \mathbf{c}^*), \quad \textit{for all } \mathbf{c} \in \mathcal{C},$$

where $\zeta_1 = 16[\mathcal{V}(d)]^2 \left(\int \left|\mathcal{K}(u)\right| \frac{Ld|u|^s}{l!} du \right)^2$ with $\mathcal{V}(d) = \pi^{d/2}/\Gamma(d/2+1)$ and $\Gamma(\cdot)$ stands for the Gamma function.

Proof. We consider the case $d = 1$ for simplicity. Using the elementary property $\mathbb{E}_{P_Z} \mathcal{K}_\lambda (Z - x) = 1/\lambda \mathbb{E}_{P_X} \mathcal{K}\left((X - x)/\lambda\right)$, gathering with Fubini, we can write:

$$(R^\lambda - R)(\mathbf{c}, \mathbf{c}^*) = \int_{\mathcal{B}(0,1)} \big(\gamma(\mathbf{c}, x) - \gamma(\mathbf{c}^*, x)\big) \int_{\mathbb{R}} \mathcal{K}(u) \big(f(x + \lambda u) - f(x)\big) du dx,$$

Now, since f has $l = \lfloor s \rfloor$ derivatives and $\mathcal{K}(\cdot)$ a kernel of order l, there exists $\tau \in]0, 1[$ such that:

$$\begin{aligned}
\int_{\mathbb{R}} \mathcal{K}(u) \left(f(x + \lambda u) - f(x)\right) du &\leq \int_{\mathbb{R}} \mathcal{K}(u) \left(\sum_{k=1}^{l-1} \frac{f^{(k)}(x)}{k!} (\lambda u)^k + \frac{f^{(l)}(x + \tau \lambda u)}{l!} (\lambda u)^l \right) du \\
&\leq \int_{\mathbb{R}} \mathcal{K}(u) \left(\frac{(\lambda u)^l}{l!} (f^{(l)}(x + \tau \lambda u) - f^{(l)}(x)) \right) du \\
&\leq \lambda^s \int_{\mathbb{R}} \left|\mathcal{K}(u)\right| \frac{L|u|^s}{l!} du,
\end{aligned}$$

where we use in last line the Hölder smoothness of f. From (8.2) and $\mathbf{MA}(\kappa)$, we have that $|\gamma(\mathbf{c}, X) - \gamma(\mathbf{c}^*, X)| \leq 4 \|\mathbf{c} - \mathbf{c}^*\| \leq 4\sqrt{\kappa R(\mathbf{c}, \mathbf{c}^*)}$ for any $\mathbf{c} \in \mathcal{C}$. We then have for any $\epsilon > 0$:

$$\begin{aligned}
(R^\lambda - R)(\mathbf{c}, \mathbf{c}^*) &\leq \int_{\mathbb{R}} \left|\mathcal{K}(u)\right| \frac{L|u|^s}{l!} du \ \lambda^s \int_{\mathcal{B}(0,1)} |\gamma(\mathbf{c}, x) - \gamma(\mathbf{c}^*, x)| dx \\
&\leq 8 \int_{\mathbb{R}} \left|\mathcal{K}(u)\right| \frac{L|u|^s}{l!} du \ \lambda^s \sqrt{\kappa R(\mathbf{c}, \mathbf{c}^*)} \\
&\leq \left(8 \int_{\mathbb{R}} \left|\mathcal{K}(u)\right| \frac{L|u|^s}{l!} du \right)^2 \epsilon \kappa \lambda^{2s} + \frac{1}{2\epsilon} R(\mathbf{c}, \mathbf{c}^*),
\end{aligned}$$

where the last inequality comes from Young's inequality:

$$xy^a \leq x^{1/(1-a)} + ay, \ \forall a < 1, \ \forall x, y > 0,$$

with $a = \frac{1}{2}$.

The same algebra in the d-dimensional case leads to the definition of constant ζ_1. □

Proposition 2 (Variance control)**.** *Assume that* $\eta_\infty := \|\eta\|_\infty < \infty$*, and* $\mathbf{NA}(\rho, \beta)$ *and* $\mathbf{MA}(\kappa)$ *are satisfied for some* $\beta \in (0, \infty)^d$*,* $\rho > 0$*, and for some positive constant* κ*, respectively. Then, for any* $t, A > 0$*, we have with probability* $1 - e^{-t}$*:*

$$\left|R_n^\lambda - R^\lambda\right|(\mathbf{c}, \mathbf{c}^*) \leq A^{-1}\left[R(\mathbf{c}, \mathbf{c}^*) + r_\lambda^*(A, t)\right], \ \text{for all } \mathbf{c} \in \mathcal{C},$$

where $r_\lambda^*(A, t)$ *satisfies the following fixed point equation:*

$$\sqrt{r} A \frac{\lambda^{-\bar{\beta}}}{\sqrt{n}} \left[\zeta_2 \sqrt{\eta_\infty}\sqrt{2\kappa t} + 6\left(1 + \frac{1}{\sqrt{n}\lambda^{d/2}}\right)\zeta_3 + \frac{2t\zeta_2}{3\sqrt{n}\lambda^{d/2}}\right] + \frac{2t}{3n} = r, \tag{7.1}$$

with $\bar{\beta} = \sum_{v=1}^d \beta_v$ *and:*

$$\zeta_2 := 4\sqrt{\mathcal{V}(d)}\rho^{-1}\|w\|_1 \text{ and } \zeta_3 := \frac{8\sqrt{\mathcal{V}(d)}}{\rho}\|w\|_1 \left(\frac{1}{3} \vee \sqrt{2\eta_\infty}\right)\left[\log(|\mathcal{M}|) + kd\big(\log(kd) + 6\log(2)\big)\right],$$

and $\mathcal{V}(d) = \pi^{d/2}/\Gamma(d/2 + 1)$.

Proof. The proof of Proposition 2 is based on the Bousquet's version of Talagrand concentration inequality (see Bousquet [2002]) applied to the following weighted random variable:

$$Z_\lambda := A \sup_{\mathbf{c} \in \mathcal{C}} \frac{r}{r + R(\mathbf{c}, \mathbf{c}^*)} \left|\frac{1}{n}\sum_{i=1}^n \gamma_\lambda(\mathbf{c}, Z_i) - \gamma_\lambda(\mathbf{c}^*, Z_i) - \mathbb{E}\big[\gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}^*, Z)\big]\right|. \tag{7.2}$$

More precisely, let us introduce the following notations. For any $\lambda \in \Lambda$, define

$$\sigma_\lambda^2 := A^2 \sup_{\mathbf{c} \in \mathcal{C}} \mathbb{E}\left[\frac{r\big(\gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}^*, Z)\big)}{r + R(\mathbf{c}, \mathbf{c}^*)}\right]^2, \text{ and } T_\lambda := A \sup_{\mathbf{c} \in \mathcal{C}} \left\|\frac{r\big[\gamma_\lambda(\mathbf{c}, \cdot) - \gamma_\lambda(\mathbf{c}^*, \cdot)\big]}{r + R(\mathbf{c}, \mathbf{c}^*)}\right\|_\infty. \tag{7.3}$$

Then, under separability condition over $\mathcal{C}$, for any $t > 0$, the classical Bousquet's inequality claims that:

$$\mathbb{P}\left(Z_\lambda \geq \mathbb{E}Z_\lambda + \sqrt{\frac{2t}{n}\left(\sigma_\lambda^2 + [1 + T_\lambda]\,\mathbb{E}Z_\lambda\right)} + \frac{t}{3n}\right) \leq e^{-t}.$$

In the sequel, we use a simpler version of Bousquet's inequality as follows. By simple algebra, we have:

$$\begin{aligned}
\mathbb{E}Z_\lambda + \sqrt{\frac{2t}{n}\left(\sigma_\lambda^2 + [1 + T_\lambda]\,\mathbb{E}Z_\lambda\right)} + \frac{t}{3n} &\leq \mathbb{E}Z_\lambda + \sigma_\lambda\sqrt{\frac{2t}{n}} + \sqrt{\frac{2t}{n}(1 + T_\lambda)\,\mathbb{E}Z_\lambda} + \frac{t}{3n} \\
&\leq \sigma_\lambda\sqrt{\frac{2t}{n}} + \mathbb{E}Z_\lambda + 2\sqrt{\mathbb{E}Z_\lambda}\sqrt{\frac{t}{n}(1 + T_\lambda)} + \frac{t(1 + T_\lambda)}{3n} \\
&\leq \sigma_\lambda\sqrt{\frac{2t}{n}} + 2\mathbb{E}Z_\lambda + 2\frac{t(1 + T_\lambda)}{3n},
\end{aligned}$$

where we use $(a+b)^2 \leq 2(a^2 + b^2)$, $a, b \in \mathbb{R}$ to get the last inequality. We can then give a simpler version of Bousquet's inequality. For any $t > 0$, we have:

$$\mathbb{P}\left(Z_\lambda \geq 2\mathbb{E}Z_\lambda + \sigma_\lambda\sqrt{\frac{2t}{n}} + 2\frac{t(1 + T_\lambda)}{3n}\right) \leq e^{-t}. \tag{7.4}$$

We are now on time to give convenient upper bounds for the terms depending on λ in (7.4).

Control of σ_λ: The control of σ_λ is based on Lemma 1 and the margin assumption $\mathbf{MA}(\kappa)$. We have, from (7.3):

$$\sigma_\lambda^2 = A^2 \sup_{\mathbf{c}\in\mathcal{C}} \frac{r^2\,\mathbb{E}\big(\gamma_\lambda(\mathbf{c},Z)-\gamma_\lambda(\mathbf{c}^*,Z)\big)^2}{\big(R(\mathbf{c},\mathbf{c}^*)+r\big)^2} \le A^2 \sup_{\mathbf{c}\in\mathcal{C}} \frac{r^2\zeta_2^2\eta_\infty \lambda^{-2\bar\beta}\kappa R(\mathbf{c},\mathbf{c}^*)}{\big(R(\mathbf{c},\mathbf{c}^*)+r\big)^2} \le A^2\kappa\zeta_2^2\eta_\infty\lambda^{-2\bar\beta} r, \tag{7.5}$$

where the last inequality is obtained considering both cases $R(\mathbf{c},\mathbf{c}^*)\le r$ and $R(\mathbf{c},\mathbf{c}^*)>r$.

Control of T_λ: Using Lemma 2, and the margin assumption $\mathbf{MA}(\kappa)$, one gets immediately:

$$T_\lambda \le A\sqrt{\kappa r}\,\zeta_2\lambda^{-\bar\beta-d/2}. \tag{7.6}$$

Control of $\mathbb{E}Z_\lambda$: The control of $\mathbb{E}Z_\lambda$ needs Lemma 3 and a modified version of the so-called *peeling device* suggested by Levrard [2012]. Let $\mu>1$ be a real number. We then have:

$$Z_\lambda \le \sup_{\mathbf{c}\in\mathcal{C}\,:\,R(\mathbf{c},\mathbf{c}^*)\le r} \frac{Ar\,\big|R_n^\lambda - R^\lambda\big|(\mathbf{c},\mathbf{c}^*)}{R(\mathbf{c},\mathbf{c}^*)+r} + \sum_{k\ge 0} A \sup_{\mathbf{c}\in\mathcal{C}\,:\,r\mu^k\le R(\mathbf{c},\mathbf{c}^*)\le r\mu^{k+1}} \frac{r\,\big|R_n^\lambda - R^\lambda\big|(\mathbf{c},\mathbf{c}^*)}{R(\mathbf{c},\mathbf{c}^*)+r}$$

Taking the expectation on both sides and using Lemma 3 lead to:

$$\begin{aligned}\mathbb{E}\mathbb{Z}_\lambda &\le \frac{A\zeta_3\lambda^{-\bar\beta}\sqrt{r}}{\sqrt{n}}\left(1+\frac{1}{\sqrt{n}\lambda^{d/2}}\right) + \sum_{k\ge 0}\frac{A\zeta_3\lambda^{-\bar\beta}\sqrt{r}\mu^{(k+1)/2}}{\sqrt{n}(1+\mu^k)}\left(1+\frac{1}{\sqrt{n}\lambda^{d/2}}\right)\\ &\le \frac{A\zeta_3\lambda^{-\bar\beta}\sqrt{r}}{\sqrt{n}}\left(1+\frac{1}{\sqrt{n}\lambda^{d/2}}\right)\left(1+\sum_{k\ge 0}\frac{\mu^{(k+1)/2}}{(1+\mu^k)}\right) \le 3A\zeta_3\frac{\lambda^{-\bar\beta}\sqrt{r}}{\sqrt{n}}\left(1+\frac{1}{\sqrt{n}\lambda^{d/2}}\right),\end{aligned}$$

where the last inequality is obtained taking $\mu=4$. Thus, using the definition of $r=r_\lambda^*(A,t)$, the last inequality, (7.5) and (7.6), we have

$$\begin{aligned}\sigma_\lambda\sqrt{\frac{2t}{n}} + 2\mathbb{E}\mathbb{Z}_\lambda + 2\frac{t\,(1+T_\lambda)}{3n} &\le \sqrt{r}A\frac{\lambda^{-\bar\beta}}{\sqrt{n}}\left[\zeta_2\sqrt{\eta_\infty}\sqrt{2\kappa t} + 6\left(1+\frac{1}{\sqrt{n}\lambda^{d/2}}\right)\zeta_3 + \frac{2t\zeta_2}{3\sqrt{n}\lambda^{d/2}}\right] + \frac{2t}{3n}\\ &= r = r_\lambda^*(A,t)\,.\end{aligned}$$

Invoking Bousquet's inequality (7.4) and the last inequality, we obtain $\mathbb{P}\left(Z_\lambda \ge r_\lambda^*(A,t)\right)\le e^{-t}$. Then, the proof is complete by definition of Z_λ in (7.2). □

7.2 Proof of Theorem 1

As we mentioned above, the proof is based on the application of Propositions 1 and 2. Indeed, from the bias-variance decomposition (2.7), using Proposition 1 with $\epsilon=2$ and $\mathbf{c}=\hat{\mathbf{c}}_\lambda$, and Proposition 2 with $A=4$ and $\mathbf{c}=\hat{\mathbf{c}}_\lambda$, we have with probability $1-e^{-t}$

$$R(\hat{\mathbf{c}}_\lambda,\mathbf{c}^*) \le 2\zeta_1\kappa\lambda^{2s} + \frac{1}{2}R(\hat{\mathbf{c}}_\lambda,\mathbf{c}^*) + \frac{1}{4}r_{\lambda,n}^*(4,t)$$

Moreover, note that for any $\lambda>0$ such that $n\lambda^d\to\infty$ as $n\to\infty$, and for any $t>0$ and any $A\ge 1$, there exists $n_0\in\mathbb{N}$ such that for all $n\ge n_0$ it holds:

$$r_\lambda^*(A,t) \le 2A^2\frac{\lambda^{-2\bar\beta}}{n}\left[\zeta_2\sqrt{\eta_\infty}\sqrt{2\kappa t} + 6\left(1+\frac{1}{\sqrt{n}\lambda^{d/2}}\right)\zeta_3 + \frac{2t\zeta_2}{3\sqrt{n}\lambda^{d/2}}\right]^2, \tag{7.7}$$

looking at the solution of the fixed point equation (7.1). Using two last inequalities, we have for n such that $n\lambda^d\ge 1$, with probability $1-e^{-t}$

$$R(\hat{\mathbf{c}}_\lambda,\mathbf{c}^*) \le 4\zeta_1\kappa\lambda^{2s} + 64\frac{\lambda^{-2\bar\beta}}{n}\left[\zeta_2^2\eta_\infty 2\kappa t + 72\zeta_3^2 + \frac{4t^2\zeta_2^2}{9n\lambda^d}\right]. \tag{7.8}$$

By integration, the choice of λ in Theorem 1 gives:

$$\mathbb{E}R(\hat{\mathbf{c}}_\lambda,\mathbf{c}^*) \le \max\left(4\zeta_1\kappa,\, 48\left(2\zeta_2^2\eta_\infty\kappa + 12\zeta_3^2\right)\right) n^{-s/(s+\bar\beta)}.$$

■

7.3 Proof of Theorem 2

In the sequel, we write λ^* the element of Λ solution of the equation $12\zeta_1\kappa\lambda^{2s} = \delta_\lambda$, where ζ_1 is given in Proposition 1. Consequently, note that there exists a constant $\diamond > 0$ such that

$$\lambda^* = \diamond n^{-\frac{1}{2s+2\beta}}.$$

Moreover, we introduce λ_a^*, the element of Λ_a such that $\lambda_a^* \leq \lambda^* \leq a^{-1}\lambda_a^*$. Let us consider the event $\Omega = \left\{\lambda_a^* \leq \hat{\lambda}\right\}$. Firstly, by construction of $\hat{\lambda}$, we have on Ω:

$$R_n^{\lambda_a^*}(\hat{\mathbf{c}}_{\hat{\lambda}}, \hat{\mathbf{c}}_{\lambda_a^*}) = R_n^{\lambda_a^*}(\hat{\mathbf{c}}_{\hat{\lambda}}) - R_n^{\lambda_a^*}(\hat{\mathbf{c}}_{\lambda_a^*}) \leq 3\delta_{\lambda_a^*}. \tag{7.9}$$

By simple computations, one gets:

$$\begin{aligned} R(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) &= (R - R^{\lambda_a^*})(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) + (R^{\lambda_a^*} - R_n^{\lambda_a^*})(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) + R_n^{\lambda_a^*}(\hat{\mathbf{c}}_{\hat{\lambda}}, \hat{\mathbf{c}}_{\lambda_a^*}) + R_n^{\lambda_a^*}(\hat{\mathbf{c}}_{\lambda_a^*}, \mathbf{c}^*) \\ &\leq (R - R^{\lambda_a^*})(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) + (R^{\lambda_a^*} - R_n^{\lambda_a^*})(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) + R_n^{\lambda_a^*}(\hat{\mathbf{c}}_{\hat{\lambda}}, \hat{\mathbf{c}}_{\lambda_a^*}). \end{aligned}$$

Then, using Proposition 1 with some $\epsilon > 1/2$ (chosen later on) and $\mathbf{c} = \hat{\mathbf{c}}_{\hat{\lambda}}$, it yields on Ω using (7.9):

$$R(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) \leq \epsilon\zeta_1\kappa(\lambda_a^*)^{2s} + \frac{1}{2\epsilon}R(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) + (R^{\lambda_a^*} - R_n^{\lambda_a^*})(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) + 3\delta_{\lambda_a^*}.$$

Using Proposition 2 with $t = 2\log(n)$, and $\mathbf{c} = \hat{\mathbf{c}}_{\hat{\lambda}}$, it holds for any $A > 2\epsilon/(2\epsilon - 1)$, with probability at least $1 - n^{-2} - \mathbb{P}(\lambda_a^* \geq \hat{\lambda})$:

$$R(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) \leq \frac{2\epsilon A}{2\epsilon A - (2\epsilon + A)}\left(\epsilon\zeta_1\kappa(\lambda_a^*)^{2s} + \frac{1}{A}r_{\lambda_a^*}^*\left(A, 2\log(n)\right) + 3\delta_{\lambda_a^*}\right).$$

Choosing $\epsilon = 2$ and $A = \sqrt{2}$, we obtain:

$$R(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) \leq 2\left(2\zeta_1\kappa(\lambda_a^*)^{2s} + \frac{1}{\sqrt{2}}r_{\lambda_a^*}^*\left(\sqrt{2}, 2\log(n)\right) + 3\delta_{\lambda_a^*}\right).$$

From the definition of δ_λ and (7.7), there exists $n_1 \in \mathbb{N}$ such that $\delta_\lambda \geq r_\lambda^*\left(\sqrt{2}, 2\log(n)\right)$ for any $\lambda \in \Lambda$ and $n \geq n_1$. We then obtain:

$$R(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) \leq 4\zeta_1\kappa(\lambda_a^*)^{2s} + 8\delta_{\lambda_a^*}, \tag{7.10}$$

It remains to control the probability $\mathbb{P}(\hat{\lambda} \leq \lambda_a^*)$. Note that, using the definition of $\hat{\lambda}$, we have by union bound:

$$\mathbb{P}(\lambda_a^* \geq \hat{\lambda}) \leq \sum_{\lambda \leq \lambda_a^*} \mathbb{P}\left(R_n^\lambda(\hat{\mathbf{c}}^{\lambda_a^*}, \hat{\mathbf{c}}_{\hat{\lambda}}) > 3\delta_\lambda\right). \tag{7.11}$$

From simple computations and using twice Proposition 1 with $\epsilon > 0$ and $\mathbf{c} = \hat{\mathbf{c}}_\lambda$ or $\mathbf{c} = \hat{\mathbf{c}}_{\lambda_a^*}$, for any $\lambda \leq \lambda_a^*$, we have:

$$\begin{aligned} R_n^\lambda(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_{\hat{\lambda}}) &\leq (R_n^\lambda - R^\lambda)(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_{\hat{\lambda}}) + (R^\lambda - R)\left(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_\lambda\right) + R(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_\lambda) \\ &\leq (R_n^\lambda - R^\lambda)(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_\lambda) + 2\zeta_1\kappa\epsilon\lambda^{2s} + \frac{1}{2\epsilon}R(\hat{\mathbf{c}}_{\lambda_a^*}, \mathbf{c}^*) + \frac{1}{2\epsilon}R(\hat{\mathbf{c}}_\lambda, \mathbf{c}^*) + R(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_\lambda) \\ &\leq (R_n^\lambda - R^\lambda)(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_\lambda) + 2\zeta_1\kappa\epsilon\lambda^{2s} + \left(1 + \frac{1}{2\epsilon}\right)R(\hat{\mathbf{c}}_{\lambda_a^*}, \mathbf{c}^*) + \left(\frac{1}{2\epsilon} - 1\right)R(\hat{\mathbf{c}}_{\lambda_a^*}, \mathbf{c}^*). \end{aligned}$$

Note that $\delta_\lambda \geq \sqrt{2}r_\lambda^*\left(\sqrt{2}, 2\log(n)\right)$ for any $\lambda \in \Lambda$ and $n \geq n_1$. Then using twice Proposition 2 with $A = \sqrt{2}$ and $\mathbf{c} = \hat{\mathbf{c}}^{\lambda_a^*}$ or $\mathbf{c} = \hat{\mathbf{c}}^\lambda$, we get with probability $1 - 2n^{-2}$

$$R_n^\lambda(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_\lambda) \leq \delta_\lambda + 2\zeta_1\kappa\epsilon\lambda^{2s} + \left(1 + \frac{1}{2\epsilon} + \frac{1}{\sqrt{2}}\right)R(\hat{\mathbf{c}}_{\lambda_a^*}, \mathbf{c}^*) + \left(\frac{1}{2\epsilon} - 1 + \frac{1}{\sqrt{2}}\right)R(\hat{\mathbf{c}}_\lambda, \mathbf{c}^*).$$

We are now on time to choose $\epsilon = 1/(2-\sqrt{2})$ to get with probability $1-2n^{-2}$

$$R_n^{\lambda}(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_{\lambda}) \leq \delta_{\lambda} + 4\zeta_1 \kappa \lambda^{2s} + 2R(\hat{\mathbf{c}}_{\lambda_a^*}, \mathbf{c}^*).$$

Finally, using the inequality (7.8) with $\lambda \leq \lambda_a^*$ and $t = 2\log(n)$, we get with probability $1-3n^{-2}$

$$R_n^{\lambda}(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_{\lambda}) \leq 4\zeta_1 \kappa \lambda^{2s} + \delta_{\lambda} + 8\zeta_1 \kappa (\lambda_a^*)^{2s} + \delta_{\lambda_a^*}.$$

By definition of λ_a^* and λ^*, it holds that $\lambda \leq \lambda_a^* \leq \lambda^*$ and $12\zeta_1\kappa(\lambda_a^*)^{2s} \leq 12\zeta_1\kappa(\lambda^*)^{2s} = \delta_{\lambda^*} \leq \delta_{\lambda_a^*} \leq \delta_{\lambda}$, we thus get with probability $1-3n^{-2}$

$$R_n^{\lambda}(\hat{\mathbf{c}}_{\lambda_a^*}, \hat{\mathbf{c}}_{\lambda}) \leq 3\delta_{\lambda}.$$

Using (7.11) and the last inequality, we finally get:

$$\mathbb{P}(\lambda_a^* \geq \hat{\lambda}) \leq |\Lambda_a|/n^2 \leq 3/n.$$

Thus, using (7.10), the last inequality, and definitions of λ_a^* and λ^*, there exists a universal constant C_2 depending on $\kappa, w, L, d, s, \beta, \rho, k, \eta_\infty$, and $|\mathcal{M}|$ such that with probability greater than $1-4n^{-1}$:

$$R(\hat{\mathbf{c}}_{\hat{\lambda}}, \mathbf{c}^*) \leq 4\zeta_1\kappa(\lambda_a^*)^{2s} + 7\delta_{\lambda_a^*} \leq 4\zeta_1\kappa(\lambda^*)^{2s} + 7a^{-2\bar{\beta}}\delta_{\lambda^*} \leq C_2\left(\frac{\log(n)}{n}\right)^{s/(s+\bar{\beta})}.$$

This last assertion allows to complete the proof taking $n \geq n_1$ sufficiently large such that $4n^{-1}$ is negligible in comparison to $C_2\left(\frac{\log(n)}{n}\right)^{s/(s+\bar{\beta})}$. ∎

8 Appendix

In the proofs of Proposition 1 and Proposition 2, we use the following technical results.

Lemma 1. *Assume that* $\mathbf{NA}(\rho,\beta)$ *is satisfied for some* $\beta \in (\frac{1}{2},\infty)^d$ *and* $\rho > 0$*; and assume* $\eta_\infty := \|\eta\|_\infty < \infty$*. Then, for any* $\mathbf{c}, \mathbf{c}' \in \mathcal{C}^2$*, it holds*

$$\Big(\mathbb{E}\big[\gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}', Z)\big]^2\Big)^{1/2} \leq \zeta_2\sqrt{\eta_\infty}\lambda^{-\bar{\beta}}|||\mathbf{c}-\mathbf{c}'|||,$$

where $\zeta_2 = 4\sqrt{\mathcal{V}(d)}\rho^{-1}\|w\|_1$ *and* $\bar{\beta} = \sum_{v=1}^d \beta_v$.

Proof. For any $x \in \mathbb{R}^d$, let $m(x) = [\gamma(\mathbf{c}, x) - \gamma(\mathbf{c}', x)]\ \mathbb{1}_{\mathcal{B}(0,1)}(x)$. Using the Plancherel theorem and the convolution, it yields:

$$\begin{aligned}
\mathbb{E}\big(\gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}', Z)\big)^2 &= \mathbb{E}\big([\mathcal{K}_\lambda * m](Z)\big)^2 \\
&\leq \eta_\infty \int_{\mathbb{R}^d} \big([\mathcal{K}_\lambda * m](z)\big)^2 dz \\
&= \eta_\infty \int_{\mathbb{R}^d} \big|\mathcal{F}[\mathcal{K}_\lambda](t)\big|^2 \big|\mathcal{F}[m](t)\big|^2 dt.
\end{aligned}$$

Let us bound the Fourier transform of $\mathcal{K}_\lambda(\cdot)$; by definition of w, using $\mathbf{NA}(\rho,\beta)$ and Riemann-Lebesgue Theorem, we have:

$$\begin{aligned}
\big|\mathcal{F}[\mathcal{K}_\lambda](t)\big|^2 = \left|\frac{\mathcal{F}[w](\lambda t)}{\mathcal{F}[\eta](t)}\right|^2 &\leq \sup_{t\in[-\lambda^{-1},\lambda^{-1}]} \frac{\|w\|_1^2}{\big|\mathcal{F}[\eta](t)\big|^2} \\
&\leq \rho^{-2}\|w\|_1^2 \sup_{t\in[-\lambda^{-1},\lambda^{-1}]^d} \left(\frac{t^2+1}{2}\right)^{\beta} \leq \rho^{-2}\|w\|_1^2\lambda^{-2\bar{\beta}}. \qquad (8.1)
\end{aligned}$$

Two last inequalities then imply:

$$\mathbb{E}\big(\gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}', Z)\big)^2 \leq \eta_\infty\rho^{-2}\|w\|_1^2\lambda^{-2\bar{\beta}} \int \big|\mathcal{F}[m](t)\big|^2 dt = \eta_\infty\rho^{-2}\|w\|_1^2\lambda^{-2\bar{\beta}} \int |m(x)|^2 dx.$$

By definition of m, for any $\mathbf{c}, \mathbf{c}' \in \mathcal{C}^2$, we get

$$|m(x)| \leq \max_{j=1,\dots,k} \left| \|x - c_j\|^2 - \|x - c'_j\|^2 \right| \leq \max_{j=1,\dots,k} \|c_j - c'_j\| \left(\|x - c_j\| + \|x - c'_j\| \right) \leq 4 \|\mathbf{c} - \mathbf{c}'\|. \quad (8.2)$$

Therefore, (8.2) and (8.1) yield:

$$\mathbb{E}\big(\gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}', Z)\big)^2 \leq 16 \mathcal{V}(d) \eta_\infty \rho^{-2} \|w\|_1^2 \lambda^{-2\bar{\beta}} \|\mathbf{c} - \mathbf{c}'\|^2,$$

where $\mathcal{V}(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ is the Lebesgue measure on $\mathbb{R}^d$ of $\mathcal{B}(0,1)$. □

Lemma 2. *Assume that* $\mathbf{NA}(\rho, \beta)$ *is satisfied for some* $\beta \in (\frac{1}{2}, \infty)^d$ *and* $\rho > 0$. *Then, for any* $\mathbf{c}, \mathbf{c}' \in \mathcal{C}^2$, *it holds*

$$\|\gamma_\lambda(\mathbf{c}, \cdot) - \gamma_\lambda(\mathbf{c}', \cdot)\|_\infty \leq \zeta_2 \lambda^{-\bar{\beta} - d/2} \|\mathbf{c} - \mathbf{c}'\|,$$

where $\zeta_2 = 4\sqrt{\mathcal{V}(d)} \rho^{-1} \|w\|_1$ *and* $\bar{\beta} = \sum_{v=1}^d \beta_v$.

Proof. As in the proof of Lemma 1, let $m(\cdot)$ denote the function $[\gamma(\mathbf{c}, \cdot) - \gamma(\mathbf{c}', \cdot)]\, \mathbb{1}_{\mathcal{B}(0,1)}(\cdot)$. Using Cauchy-Schwarz inequality, it yields

$$\|\gamma_\lambda(\mathbf{c}, \cdot) - \gamma_\lambda(\mathbf{c}', \cdot)\|_\infty = \sup_{z \in \mathbb{R}^d} \left| \int \mathcal{K}_\lambda(z - x)\, m(x) dx \right| \leq \sup_{z \in \mathbb{R}^d} \sqrt{\int \big[\mathcal{K}_\lambda(z - x)\big]^2 dx \int m^2(x) dx}.$$

From (8.2), we have that $\int m^2(x) dx \leq 16 \mathcal{V}(d) \|\mathbf{c} - \mathbf{c}'\|^2$. Moreover, using the definition of $\mathcal{K}_\lambda(\cdot)$ and some change of variables, we get with $\mathbf{NA}(\rho, \beta)$:

$$\sup_{z \in \mathbb{R}^d} \int \big[\mathcal{K}_\lambda(z - x)\big]^2 dx = \int \big|\mathcal{F}[\mathcal{K}_\lambda](t)\big|^2 dt = \int \left| \frac{\mathcal{F}[w](\lambda t)}{\mathcal{F}[\eta](t)} \right|^2 dt \leq \rho^{-2} \|w\|_1^2 \lambda^{-2\bar{\beta} - d}.$$

Three last inequalities then imply

$$\|\gamma_\lambda(\mathbf{c}, \cdot) - \gamma_\lambda(\mathbf{c}', \cdot)\|_\infty \leq 4\sqrt{\mathcal{V}(d)} \rho^{-1} \|w\|_1 \lambda^{-\bar{\beta} - d/2} \|\mathbf{c} - \mathbf{c}'\|.$$

□

With this result, we also need to control the complexity involved in Section 3 thanks to the following lemma:

Lemma 3. *Assume that* $\mathbf{NA}(\rho, \beta)$ *are satisfied for some* $\beta \in (\frac{1}{2}, \infty)^d$ *and* $\rho > 0$. *Then,* $\forall \lambda, \delta > 0$, *we have:*

$$\mathbb{E} \sup_{(\mathbf{c}, \mathbf{c}^*) \in \mathcal{C} \times \mathcal{M}, \|\mathbf{c} - \mathbf{c}^*\|^2 \leq \delta} \left| R_n^\lambda - R^\lambda \right|(\mathbf{c}, \mathbf{c}^*) \leq \zeta_3 \frac{\sqrt{\delta}}{\sqrt{n} \lambda^{\bar{\beta}}} \left(1 + \frac{1}{\sqrt{n} \lambda^{d/2}} \right),$$

where

$$\zeta_3 := \frac{8\sqrt{\mathcal{V}(d)}}{\rho} \|w\|_1 \left(\frac{1}{3} \vee \sqrt{2\eta_\infty} \right) \big[\log(|\mathcal{M}|) + kd\big(\log(kd) + 6 \log(2) \big) \big].$$

Proof. The proof is based on the chaining argument, gathering with a maximal inequality. More precisely, we use a special case of a maximal inequality derived in [34, Lemma 6.6, Section 6.1]. Adapted to our needs, this maximal inequality reads as follows:

Lemma 4 (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 exist some constants* $\sigma, b > 0$ *such that for any* $\phi \in \Phi$

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}\phi^2(\mathcal{X}_i) \leq \sigma^2 \text{ and } \|\phi\|_\infty \leq b.$$

Then:

$$\mathbb{E} \sup_{\phi \in \Phi} \left| \frac{1}{n} \sum_{i=1}^n \phi(\mathcal{X}_i) - \mathbb{E}\phi(\mathcal{X}_i) \right| \leq \frac{2\sigma}{\sqrt{n}} \sqrt{2 \log(|\Phi|)} + \frac{2b}{3n} \log(|\Phi|), \quad (8.3)$$

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

Now, we start with the main part of the proof of the lemma. Consider $L : \mathbf{c} \in \mathcal{C} \mapsto L(\mathbf{c}) = \gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}^*, Z)$. We will use a chaining argument on the function L as follows. For some $\delta > 0$, let $\mathbf{c}^0 \in \mathcal{C}$ be fixed such that $\|\mathbf{c}^0 - \mathbf{c}^*\|^2 \leq \delta$. For some $0 < a < 1$, for any $v \in \mathbb{N}^*$, denote Γ_v a δa^v-net of $\mathcal{C}$. For $\mathbf{c} \in \mathcal{C}$, introduce the following notations:

$$u_0(\mathbf{c}) := \mathbf{c}^0, \quad u_v(\mathbf{c}) := \arg \inf_{u \in \Gamma_v} \|u - \mathbf{c}\|, \quad v \in \mathbb{N}^*.$$

We then deduce that $u_v(\mathbf{c}) \to \mathbf{c}$ for $v \to \infty$. The main ingredient of the chaining argument is the following decomposition (by continuity of L using dominated convergence theorem):

$$L(\mathbf{c}) = L(\mathbf{c}^0) + \sum_{v \in \mathbb{N}^*} L(u_v(\mathbf{c})) - L(u_{v-1}(\mathbf{c})).$$

Note that for simplicity, in the sequel, we write $\mathbf{c}^0$ instead of $\mathbf{c}^0(\mathbf{c}^*)$ since this vector depends on the minimum $\mathbf{c}^* \in \mathcal{M}$ where the cardinality of $\mathcal{M}$ could satisfy $|\mathcal{M}| \geq 2$. For easy of exposition, denote P_n the empirical measure and P the expectation w.r.t. the distribution P_Z, we then obtain:

$$\begin{aligned}
&\mathbb{E} \sup_{(\mathbf{c},\mathbf{c}^*) \in \mathcal{C} \times \mathcal{M}, \|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta} \left| R_n^\lambda - R^\lambda \right|(\mathbf{c}, \mathbf{c}^*) \\
&= \mathbb{E} \sup_{\mathbf{c}^* \in \mathcal{M}} \sup_{\mathbf{c} \in \mathcal{C} \,:\, \|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta} |(P_n - P)(\gamma_\lambda(\mathbf{c}^*, Z) - \gamma_\lambda(\mathbf{c}, Z))| \\
&\leq \mathbb{E} \sup_{\mathbf{c}^* \in \mathcal{M}} \left|(P_n - P)\big(\gamma_\lambda(\mathbf{c}^0, Z) - \gamma_\lambda(\mathbf{c}^*, Z)\big)\right| \\
&\quad + \mathbb{E} \sup_{\mathbf{c}^* \in \mathcal{M}} \sup_{\mathbf{c} \in \mathcal{C} \,:\, \|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta} \sum_{v \in \mathbb{N}^*} \left|(P_n - P)\big(\gamma_\lambda(u_v(\mathbf{c}), Z) - \gamma_\lambda(u_{v-1}(\mathbf{c}), Z)\big)\right| \\
&\leq \mathbb{E} \sup_{\mathbf{c}^* \in \mathcal{M}} \left|(P_n - P)\big(\gamma_\lambda(\mathbf{c}^0, Z) - \gamma_\lambda(\mathbf{c}^*, Z)\big)\right| \\
&\quad + \sum_{v \in \mathbb{N}^*} \mathbb{E} \sup_{(u,u') \in \Gamma_v \times \Gamma_{v-1} \,:\, \|u-u'\|^2 \leq \delta a^v} \left|(P_n - P)\big(\gamma_\lambda(u, Z) - \gamma_\lambda(u', Z)\big)\right| \\
&=: A_1 + A_2
\end{aligned} \tag{8.4}$$

We will now find bounds of A_1 and A_2 thanks to Lemma 4.

Bound of A_1: We first remind that $|\mathcal{M}|$ is finite. We can then apply Lemma 4, with $\phi(\mathcal{X}_i) = \gamma_\lambda(\mathbf{c}^0, Z_i) - \gamma_\lambda(\mathbf{c}^*, Z_i)$. Indeed, using the definition of $\mathbf{c}^0$ and Lemmas 1 and 2, we have $\sigma = \zeta_2 \sqrt{\eta_\infty} \lambda^{-\bar{\beta}} \sqrt{\delta}$ and $b = \zeta_2 \lambda^{-\bar{\beta}-d/2} \sqrt{\delta}$ in Lemma 4, it yields:

$$A_1 = \mathbb{E} \sup_{\mathbf{c}^* \in \mathcal{M}} \left|(P_n - P)\big(\gamma_\lambda(\mathbf{c}^0, Z) - \gamma_\lambda(\mathbf{c}^*, Z)\big)\right| \leq 2\zeta_2 \sqrt{\eta_\infty} \sqrt{2 \log(|\mathcal{M}|)} \frac{\lambda^{-\bar{\beta}} \sqrt{\delta}}{\sqrt{n}} + \frac{2\zeta_2 \log(|\mathcal{M}|)}{3} \frac{\lambda^{-\bar{\beta}-d/2} \sqrt{\delta}}{n}.$$

Bound of A_2: As previously, we use the maximal inequality with $\phi(\mathcal{X}_i) = \gamma_\lambda(u, Z_i) - \gamma_\lambda(u', Z_i)$ to the finite set $\Phi = \Gamma_v \times \Gamma_{v\text{-}1}$. According to Lemmas 1 and 2, we have $\sigma = \zeta_2 \sqrt{\eta_\infty} \lambda^{-\bar{\beta}} \sqrt{a^v \delta}$ and $b = \zeta_2 \lambda^{-\bar{\beta}-d/2} \sqrt{a^v \delta}$. We then have for any $v \in \mathbb{N}^*$

$$\begin{aligned}
&\mathbb{E} \sup_{(u,u') \in \Gamma_v \times \Gamma_{v-1} \,:\, \|u-u'\|^2 \leq \delta a^v} \left|(P_n - P)\big(\gamma_\lambda(u, Z) - \gamma_\lambda(u', Z)\big)\right| \\
&\leq 2\zeta_2 \sqrt{\eta_\infty} a^{v/2} \sqrt{2 \log(|\Gamma_v||\Gamma_{v\text{-}1}|)} \frac{\lambda^{-\bar{\beta}} \sqrt{\delta}}{\sqrt{n}} + \frac{2\zeta_2 a^{v/2} \log(|\Gamma_v||\Gamma_{v\text{-}1}|)}{3} \frac{\lambda^{-\bar{\beta}-d/2} \sqrt{\delta}}{n}.
\end{aligned}$$

We note that $|\Gamma_v| = \left(\sqrt{kd}/a^v\right)^{-kd}$ and taking $a = 1/4$, we obtain $\sum_{v \in \mathbb{N}^*} a^{v/2} \log(|\Gamma_v||\Gamma_{v\text{-}1}|) \leq kd[\log(kd) + 6 \log 2] =: \zeta_4$. We then have from the definition of A_2 and the last inequality:

$$A_2 \leq 2\sqrt{2}\zeta_2 \sqrt{\eta_\infty} \zeta_4 \frac{\lambda^{-\bar{\beta}} \sqrt{\delta}}{\sqrt{n}} + 2\zeta_2 \zeta_4 \frac{\lambda^{-\bar{\beta}-d/2} \sqrt{\delta}}{3n}.$$

From last bounds of A_1 and A_2, and the chaining decomposition (8.4), it yields

$$\mathbb{E} \sup_{\mathbf{c}^* \in \mathcal{M}} \sup_{\mathbf{c} \in \mathcal{C} \,:\, \|\mathbf{c}-\mathbf{c}^*\|^2 \le \delta} |P_n - P|(\gamma_\lambda(\mathbf{c}^*, Z) - \gamma_\lambda(\mathbf{c}, Z))$$
$$\le 2\sqrt{2}\zeta_2 \sqrt{\eta_\infty}\big(\sqrt{\log(|\mathcal{M}|)} + \zeta_4\big)\frac{\sqrt{\delta}}{\sqrt{n}\lambda^{\bar{\beta}}} + \frac{2\zeta_2}{3}\big(\log(|\mathcal{M}|) + \zeta_4\big)\frac{\sqrt{\delta}}{n\lambda^{\bar{\beta}+d/2}}$$

The proof is complete by definition of ζ_3. ∎

□

References

[1] A. Antos, L. Györfi, and A. György. Individual convergence rates in empirical vector quantizer design. *IEEE Trans. Inform. Theory*, 51 (11), 2005.

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

[3] P.L. Bartlett and S. Mendelson. Empirical minimization. *Probab. Theory and Related Fields*, 135 (3):311–334, 2006.

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

[5] G. Biau, L. Devroye, and G. Lugosi. On the performances of clustering in hilbert spaces. *IEEE Trans. Inform. Theory*, 54 (2), 2008.

[6] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36 (2):489–531, 2008.

[7] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[8] L. Brown and M. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, 24(6):2524–2535, 1996.

[9] C. Brunet and S. Loustau. The algorithm of noisy k-means. In preparation, 2013.

[10] M. Chichignoud. Minimax and minimax adaptive estimation in multiplicative regression: locally Bayesian approach. *Probab. Theory Related Fields*, 153(3-4):543–586, 2012.

[11] M. Chichignoud and Y. Lederer. A robust, fully adaptive m-estimator for pointwise estimation in heteroscedastic regression. to appear in Bernoulli, 2013.

[12] F. Comte and C. Lacour. Anisotropic adaptive kernel deconvolution. to appear in Annales de l'I. H. P., 2013.

[13] I. Dattner, M. Reiß, and M. Trabs. Adaptive quantile estimation in deconvolution with unknown error distribution. Submitted, 2013.

[14] A. Delaigle, P. Hall, and A. Meister. On deconvolution with repeated measurements. *Ann. Statist.*, 36 (2):665–685, 2008.

[15] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19:1257–1272, 1991.

[16] J. Fan and Y. Truong. Nonparametric regression with errors in variables. *Ann. Statist.*, 21:1900–1925, 1993.

[17] C. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. A. Wasserman. Minimax manifold estimation. *CoRR*, abs/1007.0549, 2010.

[18] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632, 2011.

[19] A. Goldenshluger and A. Nemirovski. On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.*, 6(2):135–170, 1997.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2002.

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

[22] V. Katkovnik. A new method for varying adaptive bandwidth selection. *IEEE Trans. Image Process.*, 47(9):2567–2571, 1999.

[23] Ch. Kervrann and J. Boulanger. Optimal spatial adaptation for patch-based image denoising. *IEEE*, 15(10):2866–2878, 2006.

[24] V. Koltchinskii. Local rademacher complexities and oracle inequalties in risk minimization. *Ann. Statist.*, 34 (6):2593–2656, 2006.

[25] V. Koltchinskii. Empirical geometry of multivariate data: A deconvolution approach. *Annals of Statistics*, 28 (2):591–629, 2000.

[26] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997.

[27] O.V. Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications*, 35(3):454–466, 1990.

[28] C. Levrard. Fast rates for empirical vector quantization. *hal.inria.fr/hal-00664068*, 2012.

[29] S. Loustau. Inverse statistical learning. In (minor) revision in Electronic Journal of Statistics, 2012.

[30] S. Loustau. Anisotropic oracle inequalities in noisy clustering. Submitted, 2013.

[31] S. Loustau and C. Marteau. Minimax fast rates for discriminant analysis with errors in variables. In (minor) revision in Bernoulli, 2012.

[32] S. Mallat. *Une exploration des signaux en ondelettes*. Ellipses, 2000.

[33] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27 (6):1808–1829, 1999.

[34] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

[35] P. Mathé. The Lepskii principle revisited. *Inverse Problems*, 22(3):L11–L15, 2006.

[36] A. Meister. *Deconvolution problems in nonparametric statistics*. Springer-Verlag, 2009.

[37] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[38] D. Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9 (1), 1981.

[39] D. Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10 (4), 1982.

[40] J. Polzehl and V. Spokoiny. Propagation-separation approach for local likelihood estimation. *Probab. Theory Related Fields*, 135(3):335–362, 2006.

[41] M. Rosenblatt. Remarks on som nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 23:832–837, 1956.

[42] V. Spokoiny and C. Vial. Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, 37(5B):2783–2807, 2009.

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

[44] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32 (1): 135–166, 2004.