Anisotropic oracle inequalities in noisy quantization

Sébastien Loustau

Anisotropic oracle inequalities in noisy quantization

Sébastien Loustau loustau@math.univ-angers.fr
LAREMA
Université d'Angers
2 Boulevard Lavoisier,
49045 Angers Cedex, France

Editor:

Abstract

The effect of errors in variables in quantization is investigated. We prove general exact and non-exact oracle inequalities with fast rates for an empirical minimization based on a noisy sample $Z_i = X_i + \epsilon_i, i = 1, \ldots, n$, where X_i are i.i.d. with density f and ϵ_i are i.i.d. with density η. These rates depend on the geometry of the density f and the asymptotic behaviour of the characteristic function of η.

This general study can be applied to the problem of k-means clustering with noisy data. For this purpose, we introduce a deconvolution k-means stochastic minimization which reaches fast rates of convergence under standard Pollard's regularity assumptions.
Keywords: Quantization, Deconvolution, Fast rates, Margin assumption, k-means clustering.

1. Introduction

The goal of empirical vector quantization (Graf and Luschgy (2000)) or clustering (Hartigan (1975)) is to replace data by an efficient and compact representation, which allows one to reconstruct the original observations with a certain accuracy. The problem was originated in signal processing and has many applications in cluster analysis or information theory. The statistical model could be described as follows. Given independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$, with unknown law P with density f on $\mathbb{R}^d$ with respect to the Lebesgue measure, we want to choose a quantizer (or classifier) $g \in \mathcal{G}$, where $\mathcal{G}$ is the set of all possible quantizers (or classifiers). The measure of the accuracy of g will be evaluate thanks to a distortion or risk given by, for some loss function ℓ:

$$R(g) = \mathbb{E}_P \ell(g, X) = \int_{\mathbb{R}^d} \ell(g, x) f(x) dx. \tag{1}$$

The most investigated example of such a framework is probably cluster analysis, where given some integer $k \geq 2$, we want to build k clusters of the set of observations $X_1, \ldots, X_n$. In this framework, a classifier $g \in \mathcal{G}$ assigns cluster $g(x) \in \{1, \ldots, k\}$ to an observation $x \in \mathbb{R}^d$.

However, in many real-life situations, direct data $X_1, \ldots, X_n$ are not available and measurement errors occur. Then, we observe only a corrupted sample $Z_i = X_i + \epsilon_i, i = 1, \ldots n$

with noisy distribution $\tilde{P}$, where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. independent of $X_1, \ldots, X_n$ with density η. The problem of noisy empirical vector quantization or noisy clustering is to represent compactly and efficiently the measure P when a contaminated empirical version $Z_1, \ldots, Z_n$ is observed. This problem is a particular case of inverse statistical learning (see Loustau (2012)), and is known to be an inverse problem. To our best knowledge, it has not been yet considered in the literature. This paper tries to fill this gap by giving a theoretical study of this problem. The construction of an algorithm to deal with clustering from a noisy dataset will be the core of a future paper.

A quiet natural habit in statistical learning is to endow clustering or empirical vector quantization into the general and extensively studied problem of empirical risk minimization (see Vapnik (2000),Bartlett and Mendelson (2006),Koltchinskii (2006)). This is exactly the guiding thread of this contribution. For this purpose, given a class of classifier or quantizer $\mathcal{G}$ (possibly infinite-dimensional space), let us consider a loss function $\ell : \mathcal{G} \times \mathbb{R}^d$ where $\ell(g, x)$ measures the loss of g at point x. In such a framework, given data $X_1, \ldots, X_n$, it is extremely standard to consider an empirical risk minimizer (ERM) defined as:

$$\hat{g}_n \in \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \ell(g, X_i). \tag{2}$$

Since the pioneer's work of Vapnik, many authors have investigated the statistical performances of (2) in such a generality. We describe below two possible examples that fall into the specific problem of clustering or empirical quantization.

Example 1 (The k-means clustering problem) *The finite dimensional clustering problem deals with the construction of a vector $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^{dk}$ to represent efficiently with $k \geq 1$ centers a set of observations $X_1, \ldots, X_n \in \mathbb{R}^d$. For this purpose, it is standard to consider the loss function $\gamma : \mathbb{R}^{dk} \times \mathbb{R}^d$ defined as:*

$$\gamma(\mathbf{c}, x) := \min_{j=1,\ldots k} \|x - c_j\|^2.$$

In this case, the empirical risk minimizer is given by $\hat{c}_n = \arg\min \sum_{i=1}^n \min_{j=1,\ldots k} \|X_i - c_j\|^2$ and is known as the popular k-means (Pollard (1981),Pollard (1982)).

Example 2 (Learning principal curves) *Another possible example is to consider quantization with principal curves (see Biau and Fisher (2012)). In the definition of Kégl et al. (2000), a principal curve can be defined as the minimizer of the least-square distortion:*

$$W(g) = \mathbb{E}_P \inf_t \|X - g(t)\|^2,$$

over a collection of parameterized curves $g : t \mapsto (g_1(t), \ldots, g_d(t))$. Principal curves can be useful in a wide range of statistical learning or data mining problems, such as speech recognition, social sciences or geology (see Biau and Fisher (2012) and the references therein). As in (2), we can minimize the empirical least-square distortion $W_n(g)$, namely the distortion integrated with respect to the empirical measure.

In this paper, we propose to adopt a comparable strategy in the presence of noisy measurements. Since we observe a corrupted sample $Z_i = X_i + \epsilon_i$, $i = 1, \ldots, n$, the empirical risk minimization (2) is not available. However, we can introduce a deconvolution step in the estimation procedure by constructing a kernel deconvolution estimator of the density f of the form:

$$\hat{f}_\lambda(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right), \tag{3}$$

where $\mathcal{K}_\eta$ is a deconvolution kernel and $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}_d^+$ is a regularization parameter (see Section 2 for details). With a slight abuse of notations, we write in (3), for any $x = (x_1, \ldots, x_d), Z_i = (Z_{1,i}, \ldots, Z_{d,i}) \in \mathbb{R}^d$:

$$\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right) = \frac{1}{\Pi_{i=1}^d \lambda_i}\mathcal{K}_\eta\left(\frac{Z_{1,i} - x_1}{\lambda_1}, \ldots, \frac{Z_{d,i} - x_d}{\lambda_d}\right).$$

Given this estimator, we construct an empirical risk by plugging (3) into the true risk (1) to get a so-called deconvolution empirical risk minimization. The idea was originated in Loustau and Marteau (2012) for discriminant analysis. To fix some notations, in this paper, a solution of this stochastic minimization can be written:

$$\hat{g}_n^\lambda \in \arg\min_{g \in \mathcal{G}} R_n^\lambda(g), \text{ where } R_n^\lambda(g) = \frac{1}{n}\sum_{i=1}^{n} \ell_\lambda(g, Z_i). \tag{4}$$

Section 2 is devoted to the detailled construction of the deconvolution empirical risk $R_n^\lambda(\cdot)$, throught the loss $\ell_\lambda(g, \cdot)$.

The purpose of this work is to study the statistical performances of $\hat{g}_n^\lambda$ in (4) in terms of oracle inequalities. On the one hand, we study the theoretical performances of $\hat{g}_n^\lambda$ thanks to exact oracle inequalities. An exact oracle inequality states that with high probability:

$$R(\hat{g}_n^\lambda) \leq \inf_{g \in \mathcal{G}} R(g) + r_{n,f,\eta}(\mathcal{G}), \tag{5}$$

where $r_{n,f,\eta}(\mathcal{G}) \longrightarrow 0$ as $n \to \infty$. The residual term $r_{n,f,\eta}(\mathcal{G})$ is called the rate of convergence. It is a function of the complexity of $\mathcal{G}$, the behaviour of the density f, and the density of the noise η. In this paper, the behaviour of f depends on two different assumptions : a margin assumption and a regularity assumption. The margin assumption is related to the difficulty of the problem whereas the regularity assumption will be expressed in terms of anisotropic Hölder spaces.
On the other hand, we propose non-exact oracle inequalities, i.e. the existence of a constant $\epsilon > 0$, such that with high probability:

$$R(\hat{g}_n^\lambda) \leq (1 + \epsilon) \inf_{g \in \mathcal{G}} R(g) + r^\star_{n,f,\eta}(\mathcal{G}). \tag{6}$$

The main difference between (5) and (6) resides in the residuals which appears in the Right Hand Sides (RHS). As in Lecué and Mendelson (2012), one of the message of this paper is to highlight the presence of faster rates of convergence (i.e. $r^\star_{n,f,\eta} = o(r_{n,f,\eta})$ as $n \to \infty$) for non-exact oracle inequalities. The cornerstone idea of these results resides in a bias-variance decomposition of the risk $R(\hat{g}_n^\lambda)$ as in Loustau (2012). However, in comparison to Loustau (2012), this work extend the previous results to unsupervised learning, non-exact oracle inequalities and to an anisotropic class of densities f.

The paper is organized as follows. In Section 2, we present the method and the main assumptions on the density η (noise assumption), the kernel in (3) and the density f (regularity and margin assumptions). We state the main theoretical results in Section 3, which consists in exact and non-exact oracle inequalities with fast rates of convergence. It allows to recover recent results in the area of fast rates. These results are applied in Section 4 for the problem of finite dimensional clustering with k-means. Section 5 concludes the paper with a discussion whereas Section 6-7 give detailled proofs of the main results.

2. Deconvolution ERM

2.1 Construction of the estimator

The deconvolution ERM introduced in this paper is originally due to Loustau and Marteau (2012) in discriminant analysis (see also Loustau (2012) for such a generality in supervised classification). The main idea of the construction is to estimate the true risk (1) thanks to a deconvolution kernel as follows.

Let us introduce $\mathcal{K} = \prod_{i=1}^d \mathcal{K}_j : \mathbb{R}^d \to \mathbb{R}$ a d-dimensional function defined as the product of d unidimensional function $\mathcal{K}_j$. Besides, $\mathcal{K}$ (and also η) belongs to $L_2(\mathbb{R}^d)$ and admits a Fourier transform. Then, if we denote by $\lambda = (\lambda_1, \ldots, \lambda_d)$ a set of (positive) bandwidths and by $\mathcal{F}[\cdot]$ the Fourier transform, we define $\mathcal{K}_\eta$ as:

$$\begin{aligned} \mathcal{K}_\eta \quad : \quad & \mathbb{R}^d \to \mathbb{R} \\ & t \mapsto \mathcal{K}_\eta(t) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)}\right](t). \end{aligned} \tag{7}$$

Given this deconvolution kernel, we construct an empirical risk by plugging (3) into the true risk $R(g)$ to get a so-called deconvolution empirical risk given by:

$$R_n^\lambda(g) = \frac{1}{n}\sum_{i=1}^n \ell_\lambda(g, Z_i) \text{ where } \ell_\lambda(g, Z_i) = \int_K \ell(g,x)\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right)dx. \tag{8}$$

Note that for technicalities, we restrict ourselves to a compact set $K \subset \mathbb{R}^d$ and study the risk minimization (1) only in K. Consequently, in this paper, we only provide a control of the true risk (1) restricted to K, namely the truncated risk:

$$R_K(g) = \int_K \ell(g,x) f(x)dx.$$

This restriction has been considered in Mammen and Tsybakov (1999) (or more recently in Loustau and Marteau (2012)). It is important to note that when f has compact support, we can see coarsely that $R_K(g) = R(g)$ for great enough K. In the sequel, for simplicity, we write $R(\cdot)$ for the restricted loss defined above. The choice of K is discussed in Section 3 and depends on the context.

2.2 Assumptions

For the sake of simplicity, we restrict ourselves to moderately or midly ill-posed inverse problem as follows. We introduce the following noise assumption **(NA)**:

(NA): There exist $(\beta_1, \ldots, \beta_d)' \in \mathbb{R}_+^d$ such that:

$$|\mathcal{F}[\eta](t)| \sim \Pi_{i=1}^d |t_i|^{-\beta_i}, \text{as } |t_i| \to +\infty, \forall i \in \{1, \ldots, d\}.$$

Moreover, we assume that $\mathcal{F}[\eta](t) \neq 0$ for all $t = (t_1, \ldots, t_d) \in \mathbb{R}^d$.

Assumption **(NA)** deals with the asymptotic behaviour of the characteristic function of the noise distribution. These kind of restrictions are standard in deconvolution problems for $d = 1$ (see Fan (1991); Meister (2009); Butucea (2007)). In this contribution, we only deal with d-dimensional mildly ill-posed deconvolution problems, which corresponds to a polynomial decreasing of $\mathcal{F}[\eta]$ in each direction. For the sake of brevity, we do not consider severely ill-posed inverse problems (exponential decreasing)or possible intermediates (e.g. a combination of polynomial and exponential decreasing functions). Recently, Comte and Lacour (2012) proposes such a study in the context of multivariate deconvolution. In our framework, the rates in these cases could be obtained through the same steps.

We also require the following assumptions on the kernel $\mathcal{K}$.

(K1) There exists $S = (S_1, \ldots, S_d) \in \mathbb{R}_d^+$, $K_1 > 0$ 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 $[-S, S] = \bigotimes_{i=1}^d [-S_i, S_i]$.

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 of order m, for a particular $m \in \mathbb{N}^d$.

K(m) The kernel $\mathcal{K}$ is of order $m = (m_1, \ldots, m_d) \in \mathbb{N}^d$, i.e.

- $\int_{\mathbb{R}^d} \mathcal{K}(x)dx = 1$
- $\int_{\mathbb{R}^d} \mathcal{K}(x)x_j^k dx = 0$, $\forall k \leq m_j$, $\forall j \in \{1, \ldots, d\}$.
- $\int_{\mathbb{R}^d} |\mathcal{K}(x)||x_j|^{m_j} dx < K_2$, $\forall j \in \{1, \ldots, d\}$.

The construction of kernels satisfying **K(m)** could be managed as in Tsybakov (2004a). This property is standard in nonparametric kernel estimation and allows to get satisfying approximations using the following assumption over the regularity of the density f.

Definition 1 *For some $s = (s_1, \ldots, s_d) \in \mathbb{R}_d^+$, $L > 0$, we say that f belongs to the anisotropic Hölder space $\mathcal{H}(s, L)$ if the following holds:*

- *the function f admits derivatives with respect to x_j up to order $\lfloor s_j \rfloor$, where $\lfloor s_j \rfloor$ denotes the largest integer less than s_j.*
- $\forall j = 1, \dots, d$, $\forall x \in \mathbb{R}^d$, $\forall x'_j \in \mathbb{R}$, *the following Lipschitz condition holds:*

$$\left| \frac{\partial^{\lfloor s_j \rfloor}}{(\partial x_j)^{\lfloor s_j \rfloor}} f(x_1, \dots, x_{j-1}, x'_j, x_{j+1}, \dots, x_d) - \frac{\partial^{\lfloor s_j \rfloor}}{(\partial x_j)^{\lfloor s_j \rfloor}} f(x) \right| \leq L |x'_j - x_j|^{s_j - \lfloor s_j \rfloor}.$$

If a function f belongs to the anisotropic Hölder space $\mathcal{H}(s, L)$, f has an Hölder regularity s_j in each direction $j = 1, \dots, d$. As a result, it can be well-approximated pointwise using a d-dimensional Taylor formula.

3. Main results

It is well-known that the behaviour of the rates of convergence $r_{n,f,\eta}(\mathcal{G})$ in (5) or $r^*_{n,f\eta}(\mathcal{G})$ in (6) is governed by the size of $\mathcal{G}$. In this paper, the size of the hypothesis space will be quantified in terms of ϵ-entropy with bracketing of the metric space $(\{\ell(g), g \in \mathcal{G}\}, L_2)$ as follows.

Definition 2 *Given a metric space $(\mathcal{F}, d)$ and a real number $\epsilon > 0$, the ϵ-entropy with bracketing of $(\mathcal{F}, d)$ is the quantity $\mathcal{H}_B(\mathcal{F}, \epsilon, d)$ defined as the logarithm of the minimal integer $N_B(\epsilon)$ such that there exist pairs $(f_j, g_j) \in \mathcal{F} \times \mathcal{F}$, $j = 1, \dots, N_B(\epsilon)$ such that $f_j \leq g_j$, $d(f_j, g_j) \leq \epsilon$, and such that for any $f \in \mathcal{F}$, there exists a pair (f_j, g_j) such that $f_j < f < g_j$.*

This notion of complexity allows to obtain local uniform concentration inequalities (see Van De Geer (2000) or van der Vaart and Weelner (1996)). Indeed, to reach fast rates of convergence (i.e. faster than $n^{-1/2}$), what really matters is not the total size of the hypothesis space but rather the size of a subclass of $\mathcal{G}$, made of functions with small errors. In this paper, we use an iterative localization principle originally introduced in Koltchinskii and Panchenko (2000) (see also Koltchinskii (2006) for such a generality). More precisely, to state exact oracle inequalities, we consider functions in $\mathcal{G}$ with small excess risk as follows:

$$\mathcal{G}(\delta) = \{g \in \mathcal{G} : R(g) - \inf_{g \in \mathcal{G}} R(g) \leq \delta\},$$

whereas to get non-exact oracle inequalities, we consider the following set:

$$\mathcal{G}'(\delta) = \{g \in \mathcal{G} : R(g) \leq \delta\}.$$

Originally, Mammen and Tsybakov (1999) (see also Tsybakov (2004b)) formulated an usefull condition to get fast rates of convergence in classification in the exact case. This assumption is known as the margin assumption and has been generalized by Bartlett and Mendelson (2006). coarsely speaking, a margin assumption guarantees a nice relationship between the variance and the expectation of any function of the excess loss class. In this contribution, it appears as follows:

Margin Assumption MA(κ) There exists some $\kappa \geq 1$ such that:

$$\forall g \in \mathcal{G}, \|\ell(g,\cdot) - \ell(g^*(g),\cdot)\|_{L_2}^2 \leq \kappa_0 \left[R(g) - \inf_{g\in\mathcal{G}} R(g)\right]^{1/\kappa},$$

for some $\kappa_0 > 0$ and where $g^*(g) \in \arg\min_{h\in\mathcal{G}} R(h)$ can depend on g when $|\mathcal{G}(0)| \geq 2$.

Gathering with a local concentration inequality (see Theorem 17 in Section 6) applied to the class $\mathcal{G}(\delta)$, this margin assumption is used in the exact-case to get fast rates. Note that provided that $\ell(g,\cdot)$ is bounded, **MA(κ)** implies **MA(κ')** for any $\kappa' \geq \kappa$. Interestingly, in the framework of finite dimensional clustering with k-means, Levrard (2012) proposes to give a sufficient condition to have **MA(κ)** with $\kappa = 1$. This condition is related with the geometry of f with respect to the optimal clusters and gives well-separated classes. It allows to interpret **MA(κ)** exactly as a margin assumption in clustering (see Section 4). In the sequel, we call the parameter κ in **MA(κ)** the margin parameter.

Recently, Lecué and Mendelson (2012) points out that one could wish non-exact oracle inequalities with fast rates under a weaker assumption. The idea is to relax significantly the margin assumption and use the loss class $\{\ell(g), g \in \mathcal{G}\}$ in **MA(κ)** instead of the excess loss class $\{\ell(g) - \ell(g^*), g \in \mathcal{G}\}$. This framework will be considered at the end of this section for completeness. It leads to non-exact oracle inequalities in the noisy case.

3.1 Exact Oracle inequalities

We are now on time to state the main exact oracle inequality.

Theorem 3 (Exact Oracle Inequality) *Suppose **(NA)**, **(K1)**, and **MA(κ)** holds for some margin parameter $\kappa \geq 1$. Suppose $f \in \mathcal{H}(s,L)$ and **K(m)** holds with $m = \lfloor \gamma \rfloor$. Suppose there exists $0 < \rho < 1$, $c > O$ such that for every $\epsilon > 0$:*

$$\mathcal{H}_B(\{\ell(g), g \in \mathcal{G}\}, \epsilon, L_2) \leq c\epsilon^{-2\rho}. \tag{9}$$

Then, for any $t > 0$, there exists some $n_0(t) \in \mathbb{N}^*$ such that for any $n \geq n_0(t)$, with probability greater than $1 - e^{-t}$, the deconvolution ERM $\hat{g}_n^\lambda$ is such that:

$$R(\hat{g}_n^\lambda) \leq \inf_{g\in\mathcal{G}} R(g) + Cn^{-\tau_d(\kappa,\rho,\beta,s)},$$

where $C > 0$ is independent of n and $\tau_d(\kappa,\rho,\beta,s)$ is given by:

$$\tau_d(\kappa,\rho,\beta,s) = \frac{\kappa}{2\kappa + \rho - 1 + (2\kappa - 1)\sum_{j=1}^d \beta_j/s_j},$$

and $\lambda = (\lambda_1, \ldots, \lambda_d)$ is chosen as:

$$\lambda_j \approx n^{-\frac{2\kappa-1}{2\kappa s_j}\tau_d(\kappa,\rho,\beta,s)}, \forall j = 1, \ldots d.$$

The proof of this result is postponed to Section 6. We list some remarks below.

Remark 4 (Comparison with Koltchinskii (2006) or Mammen and Tsybakov (1999)) *This result gives the order of the residual term in the exact oracle inequalities. The risk of the estimator* $\hat{g}_n^\lambda$ *mimics the risk of the oracle, up to a residual term detailled in Theorem 3. The price to pay for the error-in-variables model depends on the asymptotic behaviour of the characteristic function of the noise distribution. If* $\beta = 0 \in \mathbb{R}^d$ *in the noise assumption* ***(NA)****, the residual term in Theorem 3 satisfies:*

$$r_n(\mathcal{G}) = O\left(n^{-\frac{\kappa}{2\kappa+\rho-1}}\right).$$

It corresponds to the standard fast rates stated in Koltchinskii (2006) or Mammen and Tsybakov (1999) for the direct case.

Remark 5 (Comparison with Loustau (2012)) *In comparison with Loustau (2012), these rates deal with an anisotropic behaviour of the density* f*. If* $s_j = s$ *for any direction, we obtain the same asymptotics as in Loustau (2012) for supervised classification, namely:*

$$r_n(\mathcal{G}) = O\left(n^{-\frac{\kappa s}{s(2\kappa+\rho-1)+(2\kappa-1)\sum_{j=1}^d \beta_j}}\right).$$

The result of Theorem 3 gives a generalization of Loustau (2012) to the anisotropic case, in an unsupervised framework. It gives some intuition with respect to the optimality of this result.

Remark 6 (The anisotropic case is of practical interest) *The result of Theorem 3 gives some insights into the noisy quantization problem with an anisotropic density* f*. In this problem, due to the anisotropic behaviour of the density, the choice of the regularization parameters* λ_j*,* $j = 1, \ldots, d$ *depends on* j*. This result is of practical interest since it allows to consider different bandwidth coordinates for the deconvolution ERM. In finite dimensional noisy clustering with* $k \geq 2$*, this configuration arises when the optimal centers are not uniformly distributed over the support of the density. This case could not be treated at least from theoretical point of view using the previous isotropic approach stated in Loustau (2012) or Loustau and Marteau (2012).*

Remark 7 (Fast rates) *The most favorable cases arise when* $\rho \to 0$ *and* β *is small, whereas at the same time density* f *has sufficiently high Hölder exponents* s_j*. Indeed, fast rates occur when* $\tau_d(\kappa, \rho, \beta, s) \geq 1/2$*, or equivalently,* $(2\kappa - 1)\sum \beta_j/s_j < 1 - \rho$*. If* $\rho = 0$ *and* $\kappa = 1$ *(see the particular case of Section 4), we have the following condition to get fast rates:*

$$\sum_{j=1}^d \frac{\beta_j}{s_j} < 1.$$

Remark 8 (Choice of λ) *The optimal choice of* λ *in Theorem 3 optimizes a bias variance decomposition as in Loustau (2012). This choice depends on unknown parameters such as the margin parameter* κ*, the Hölder exponents* $(s_1, \ldots, s_d)$ *of the density* f *and the degree of illposedness* β*. A challenging open problem is to derive adaptive choice of* λ *to lead to the same fast rates of convergence. This could be the purpose of future works.*

Remark 9 (Comparison with Comte and Lacour (2012)) *It is also important to note that the optimal choice of the multivariate bandwidth λ does not coincide with the optimal choice of the bandwidth in standard nonparametric anisotropic density deconvolution. Indeed, it is stated in Comte and Lacour (2012) that under the same regularity and ill-posedness assumptions, the optimal choice of the bandwidth $\lambda = (\lambda_1, \ldots, \lambda_d)$ has the following asymptotics:*

$$\lambda_u \approx n^{-\frac{1}{s_u\left(2+\sum_{j=1}^d \frac{2\beta_j+1}{s_j}\right)}}.$$

The proposed asymptotic optimal calibration of Theorem 3 is rather different. It depends explicitely on parameter ρ, which measures the complexity of the decision set $\mathcal{G}$, and the margin parameter $\kappa \geq 1$. It shows rather well that our bandwidth selection problem is not equivalent to standard nonparametric estimation problems. It illustrates one more time that our procedure is not a plug-in procedure.

3.2 Non-exact oracle inequalities

In this section, we also suggest a non-exact version of Theorem 3 without the margin assumption $\mathbf{MA}(\kappa)$. However, to get this result, we need an additional assumption about the compact K appearing in the empirical risk (8). The assumption has the following form:

Density assumption $\mathbf{DA}(c_0)$ There exists a constant $c_0 > 0$ such that the compact set K in (8) satisfies:

$$K \subset \{x : f(x) \geq c_0\}.$$

This assumption is trivially satisfied if $f > 0$ in $\mathbb{R}^d$ with a constant c_0 depending on the size of K. Assumption $\mathbf{DA}(c_0)$ is necessary to get fast rates in the context of non-exact oracle inequalities without the margin assumption $\mathbf{MA}(\kappa)$. We are now on time to state the following result.

Theorem 10 (Non-Exact Oracle Inequality) *Suppose **(NA)**, $\mathbf{DA}(c_0)$ and **(K1)** holds for some constant $c_0 > 0$. Suppose $f \in \mathcal{H}(s, L)$ and $\mathbf{K}(m)$ holds with $m = \lfloor s \rfloor$. Suppose there exists $0 < \rho < 1$, $c > O$ such that for every $\epsilon > 0$:*

$$\mathcal{H}_B(\{\ell(g), g \in \mathcal{G}\}, \epsilon, L_2) \leq c\epsilon^{-2\rho}.$$

Then, for any $t > 0$, there exists some $n_0(t) \in \mathbb{N}^*$ such that for any $\epsilon > 0$, for any $n \geq n_0(t)$, with probabilty higher than $1 - e^{-t}$, $\hat{g}_n^\lambda$ satisfies:

$$R(\hat{g}) \leq (1+\epsilon) \inf_{g \in \mathcal{G}} R(g) + Cn^{-\tau^*(\rho,\beta,s)},$$

where $C > 0$ is a constant which depends on $\epsilon, \beta, s, \rho, c_0$ and

$$\tau^*(\rho, \beta, s) = \frac{1}{1 + \rho + \sum_{j=1}^d \beta_j / s_j},$$

whereas $\lambda = (\lambda_1, \ldots, \lambda_d)$ is chosen as:

$$\lambda_j \sim n^{-\frac{\tau^*(\rho,\beta,s)}{2s_j}}, \forall j = 1, \ldots d.$$

Remark 11 (Same phenomenon as in Lecué and Mendelson (2012)) *The quantity $\tau^*(\rho, \beta, s)$ describes the order of the residual term in Theorem 10. We can see coarsely that $\tau^*(\rho, \beta, s) = \tau(1, \rho, \beta, s)$ where $\tau(1, \rho, \beta, s)$ appears in Theorem 3. As a result, this oracle inequality gives the same asymptotic as the previous result under **MA(κ)** with $\kappa = 1$, which corresponds to the strong margin assumption. Here, it holds without any margin assumption. The prize to pay is the constant in front of the infimum. This phenomenom has been already pointed out in Lecué and Mendelson (2012) in a supervised framework and in the direct case. Of course, constant $C > 0$ in front of the rate depends on $\epsilon > 0$ and exploses when ϵ tends to 0 (see condition (22) in the proof).*

Remark 12 (The density assumption) *Unfortunately, there is an additional assumption to get Theorem 10 in comparison to Theorem 3, namely the assumption **DA(c_0)**. This assumption is specific to the indirect framework where we need to control the variance of the convoluted loss $\ell_\lambda(g, Z)$ with respect to the variance of $\ell(g, X)$. More precisely, we need the following inequality (in dimension $d = 1$ for simplicity):*

$$\mathbb{E}_{\tilde{P}} \ell_\lambda(g, Z)^2 \leq \lambda^{-2\beta} \mathbb{E}_P \ell(g, X)^2, \forall g \in \mathcal{G}.$$

This can be done only if we restrict $\ell_\lambda(\cdot)$ to a region where $f > 0$. Otherwise, there is no reason to obtain such a control (see Lemma 23 and also the related discussion in Loustau (2012)).

4. Application to finite dimensional noisy clustering

The aim of this section is to use the general upper bound of Theorem 3 in the framework of noisy finite dimensional clustering. To frame the problem of finite dimensional clustering into the general study of this paper, we first introduce the following notation. Given some known integer $k \geq 2$, let us consider $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{C}$ the set of possible centers, where $\mathcal{C} \subseteq \mathbb{R}^{dk}$ is compact. The loss function $\gamma : \mathbb{R}^{dk} \times \mathbb{R}^d$ is defined as:

$$\gamma(\mathbf{c}, x) = \min_{j=1,\ldots k} \|x - c_j\|^2,$$

where $\|\cdot\|$ stands for the standard euclidean norm on $\mathbb{R}^d$. The corresponding true risk or clustering risk is given by $R(\mathbf{c}) = \mathbb{E}_P \gamma(\mathbf{c}, X)$. In the sequel, we introduce a constant $M \geq 0$ such that $\|X\|_\infty \leq M$. This boundedness assumption ensures $\gamma(\mathbf{c}, X)$ to be bounded. The performances of the empirical minimizer $\hat{\mathbf{c}}_n = \arg\min_{\mathcal{C}} P_n \gamma(\mathbf{c})$ (also called k-means clustering algorithm) have been widely studied in the literature. Consistency was shown by Pollard (1981) when $\mathbb{E}\|X\|^2 < \infty$ whereas Linder et al. (1994) or Biau et al. (2008) gives rates of convergence of the form $\mathcal{O}(1/\sqrt{n})$ for the excess clustering risk defined as $R(\hat{\mathbf{c}}_n) - R(c^*)$, where $c^* \in \mathcal{M}$ the set of all possible optimal clusters. More recently, Levrard (2012) proposes fast rates of the form $\mathcal{O}(1/n)$ under Pollard's regularity assumptions. It improves a previous result of Antos et al. (2005). The main ingredient of the proof is a localization argument in the spirit of Blanchard et al. (2008).

In this section, we study the problem of clustering where we have at our disposal a corrupted sample $Z_i = X_i + \epsilon_i$, $i = 1, \ldots, n$ where the ϵ_i's are i.i.d. with density η satisfying

(NA) of Section 2. For this purpose, we introduce the following deconvolution empirical risk minimization:

$$\arg\min_{\mathbf{c}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \gamma_\lambda(\mathbf{c}, Z_i), \tag{10}$$

where $\gamma_\lambda(\mathbf{c}, z)$ is a deconvolution k-means loss defined as:

$$\gamma_\lambda(\mathbf{c}, z) = \int_K \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right) \min_{j=1,\dots k} \|x - c_j\|^2 dx.$$

The kernel $\mathcal{K}_\eta$ is the deconvolution kernel introduced in Section 2 with $\lambda = (\lambda_1, \dots, \lambda_d) \in \mathbb{R}_+^d$ a set of positive bandwidths chosen later on. We investigate the generalization ability of the solution of (10) in the context of Pollard's regularity assumptions. For this purpose, we will use the following regularity assumptions on the source distribution P.

Pollard's Regularity Condition (PRC): The distribution P satisfies the following two conditions:

1. P has a continuous density f with respect to Lebesgue measure on $\mathbb{R}^d$,
2. The Hessian matrix of $\mathbf{c} \longmapsto P\gamma(\mathbf{c}, .)$ is positive definite for all optimal vector of clusters $\mathbf{c}^*$.

It is easy to see that using the compactness of $\mathcal{B}(0, M)$, $\|X\|_\infty \leq M$ and **(PRC)** ensures that there exists only a finite number of optimal clusters $\mathbf{c}^* \in \mathcal{M}$. This number is denoted as $|\mathcal{M}|$ in the rest of this section. Moreover, Pollard's conditions can be related to the margin assumption **MA(**κ**)** of Section 3 thanks to the following lemma due to Antos et al. (2005).

Lemma 13 (Antos et al. (2005)) *Suppose $\|X\|_\infty \leq M$ and **(PRC)** holds. Then, for any $\mathbf{c} \in \mathcal{B}(0, M)$:*

$$\|\gamma(\mathbf{c}, \cdot) - \gamma(\mathbf{c}^*(\mathbf{c}), \cdot)\|_{L_2} \leq C_1\|\mathbf{c} - \mathbf{c}^*(\mathbf{c})\|^2 \leq C_1 C_2 \left(R(\mathbf{c}) - R(\mathbf{c}^*(\mathbf{c}))\right),$$

where $c^*(\mathbf{c}) \in \arg\min_{\mathbf{c}^*} \|\mathbf{c} - \mathbf{c}^*\|$.

Lemma 13 ensures a margin assumption **MA(**κ**)** with $\kappa = 1$ (see Section 3). It is useful to derive fast rates of convergence. Recently, Levrard (2012) has pointed out sufficient conditions to have **(PRC)** as follows. Denote ∂V_i the boundary of the Voronoi cell V_i associated with c_i, for $i = 1, \dots, k$. Then, a sufficient condition to have **(PRC)** is to control the sup-norm of f on the union of all possible $|\mathcal{M}|$ boundaries $\partial V^{*,m} = \cup_{i=1}^k \partial V_i^{*,m}$, associated with $c_m^* \in \mathcal{M}$ as follows:

$$\|f_{|\cup_{m=1}^{\mathcal{M}} \partial V^{*,m}}\|_\infty \leq c(d)M^{d+1} \inf_{m=1,\dots,|\mathcal{M}|, i=1,\dots k} P(V_i^{*,m}),$$

where $c(d)$ is a constant depending on the dimension d. As a result, the margin assumption is guaranteed when the source distribution P is well concentrated around its optimal clusters, which is related to well-separated classes. From this point of view, the margin assumption **MA(**κ**)** can be related to the margin assumption in binary classification.
We are now ready to state the main result of this section.

Theorem 14 *Assume **(NA)** holds, P satisfies **(PRC)** with density $f \in \mathcal{H}(s,L)$ and $\mathbb{E}\|\epsilon\|^2 < \infty$. Then, for any $t > 0$, for any $n \geq n_0(t)$, denoting by $\hat{\mathbf{c}}_n^\lambda$ a solution of (10), we have with probability higher than $1 - e^{-t}$:*

$$R(\hat{\mathbf{c}}_n^\lambda) \leq \inf_{\mathbf{c} \in \mathcal{C}} R(\mathbf{c}) + C\sqrt{\log\log(n)}\, n^{-\frac{1}{1+\sum_{j=1}^d \beta_j/s_j}},$$

where $C > 0$ is independent of n and $\lambda = (\lambda_1, \ldots, \lambda_d)$ is chosen as:

$$\lambda_j \approx n^{-\frac{1}{2s_j(1+\rho+\sum_{j=1}^d \beta_j/s_j)}}, \forall j = 1, \ldots d.$$

.

The proof is postponed to Section 6. Here follows some remarks.

Remark 15 (Fast rates of convergence) *Theorem 14 is a direct application of Theorem 3 in Section 3. The order of the residual term in Theorem 14 is comparable to Theorem 3. Due to the finite dimensional hypothesis space $\mathcal{C} \subset \mathbb{R}^{dk}$, we apply the previous study to the case $\rho = 0$. It leads to the fast rates $O\left(n^{-\frac{1}{1+\sum_{i=1}^d \beta_j/s_j}}\right)$, up to an extra $\sqrt{\log\log n}$ term. This term is due to the localization principle of the proof, which consists in applying iteratively the concentration inequality of Theorem 17. In the finite dimensional case, when $\rho = 0$, we pay an extra $\sqrt{\log\log n}$ term in the rate by solving the fixed point equation. Note that using for instance Levrard (2012), this term can be avoid. It is out of the scope of the present paper.*

Remark 16 (Optimality) *Lower bounds of the form $\mathcal{O}(1/\sqrt{n})$ have been stated in the direct case by Bartlett et al. (1998) for general distribution. An open problem is to derive lower bounds in the context of Theorem 14. For this purpose, we need to construct configurations where both Pollard's regularity assumption and noise assumption **(NA)** could be used in a careful way. In this direction, Loustau and Marteau (2012) suggests lower bounds in a supervised framework under both margin assumption and **(NA)**.*

5. Conclusion

This paper can be seen as a first attempt into the study of quantization with errors-in-variables. Many problems could be considered in future works, from theoretical or practical point of view.

In the problem of risk minimization with noisy data, we provide oracle inequalities for an empirical risk minimization based on a deconvolution kernel. The risk of the deconvolution ERM mimics the risk of the oracle, up to some residual term, called the rate of convergence. The order of these rates depends on the complexity of the hypothesis space in terms of entropy, the behaviour of the density f and the degree of ill-posedness. From the theoretical point of view, these results extend the previous study of Loustau (2012) to the unsupervised framework, the non-exact case and to an anisotropic behaviour of the density f. These significant extensions could be the core of many applications in unsupervised learning.

As an example, we turn into the problem of clustering with k-means. We consider the general approach and introduce a deconvolution kernel estimator of the density f in the distortion. It gives rise to a new stochastic minimization called deconvolution k-means. The method gives fast rates of convergence.

Another possible direct application of the result of this paper is to learn principal curves in the presence of noisy observations. In such a problem, the aim is to design a principal curve for an unknown distribution P when we have at our disposal a noisy dataset $Z_i = X_i + \epsilon_i$, $i = 1, \ldots, n$. To the best of our knowledge, this problem has not been considered in the literature. Following the ERM approach of this paper, it is possible to design a new procedure to state rates of convergence in the presence of noisy observations.

The general deconvolution ERM principle introduced in this paper can be used to design new algorithms to deal with unsupervised statistical learning with noisy observations. As a first step, the construction of a noisy version of the well-known k-means is a core of a future work. The construction of a noisy version of the Polygonal Line Algorithm (see Sandilya and Kulkarni (2002)) could also be investigated, to deal with learning principal curves from indirect observations.

6. Proofs

The main probabilistic tool for our needs is the following concentration inequality due to Bousquet.

Theorem 17 (Bousquet (2002)) *Let $\mathcal{G}$ a countable class of real-valued measurable functions defined on a measurable space $\mathcal{X}$. Let $X_1, \ldots, X_n$ be n i.i.d. random variables with values in $\mathcal{X}$. Let us consider the random variable:*

$$Z_n(\mathcal{G}) = \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \mathbb{E} g(X_1) \right|.$$

Then, for every $t > 0$:

$$\mathbb{P}(Z_n(\mathcal{G}) \geq U_n(\mathcal{G}, t)) \leq e^{-t},$$

where:

$$U_n(\mathcal{G}, t) = \mathbb{E} Z_n(\mathcal{G}) + \sqrt{\frac{2t}{n} [\sigma^2(\mathcal{G}) + (1 + b(\mathcal{G})) \mathbb{E} Z_n(\mathcal{G})]} + \frac{t}{3n},$$

and

$$\sigma^2(\mathcal{G}) = \sup_{g \in \mathcal{G}} \mathbb{E} g(X_1)^2 \text{ and } b(\mathcal{G}) = \sup_{g \in \mathcal{G}} \|g\|_\infty.$$

The proof of this result uses the so-called entropy method introduced by Ledoux (1996), and further refined by Massart (2000) or Rio (2000). The use of a ψ_1-version (see for instance Adamczak (2008)) has been considered in Lecué and Mendelson (2012), to alleviate the boundedness assumption.

This concentration inequality is at the core of the localization principle presented in Koltchinskii (2006), which consists in using Theorem 17 to functions in $\mathcal{G}$ with small error. In the following, we extend this localization approach to:

- the noisy set-up,

- the non-exact case.

For this purpose, we apply Theorem 17 to particular classes $\mathcal{G}$, namely excess loss classes for the exact case and loss classes for the non-exact case. These two extensions are proposed in Lemma 18 and 19 below. These results are at the core of the general exact and non-exact oracle inequalities of Theorem 3 and Theorem 10 in Section 3.

6.1 Intermediate lemmas

6.1.1 Notations

Let us first introduce the following notations. For any fixed $g \in \mathcal{G}$, we write:

$$R^\lambda(g) = \int_K \ell(g,x)\mathbb{E}_P \frac{1}{\lambda}\mathcal{K}\left(\frac{X-x}{\lambda}\right) dx \text{ and } R_n^\lambda(g) = \frac{1}{n}\sum_{i=1}^n \ell_\lambda(g, Z_i).$$

As a result, for any fixed $g \in \mathcal{G}$, we have the following equality:

$$R_n^\lambda(g) - R^\lambda(g) = \frac{1}{n}\sum_{i=1}^n \ell_\lambda(g, Z_i) - \mathbb{E}_{\tilde{P}}\ell_\lambda(g, Z).$$

With a slight abuse of notations, we also denote:

$$(R_n^\lambda - R^\lambda)(g-g') = R_n^\lambda(g) - R^\lambda(g) - R_n^\lambda(g') + R^\lambda(g').$$

The same notation is used for $R^\lambda(\cdot)$ and $R(\cdot)$ with the quantity $(R - R^\lambda)(g - g')$.

For a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, the following transformations will be considered:

$$\breve{\psi}(\delta) = \sup_{\sigma \geq \delta} \frac{\psi(\sigma)}{\sigma} \text{ and } \psi^\dagger(\epsilon) = \inf\{\delta > 0 : \breve{\psi}(\delta) \leq \epsilon\}.$$

Moreover, we need the following property (see Koltchinskii (2006)):

$$\forall \delta' \leq \delta,\ \psi(\delta) \leq \delta\breve{\psi}(\delta'). \tag{11}$$

We are also interested in the following discretization version of these transformations:

$$\breve{\psi}_q(\delta) = \sup_{\delta_j \geq \delta} \frac{\psi(\delta_j)}{\delta_j} \text{ and } \psi_q^\dagger(\epsilon) = \inf\{\delta > 0 : \breve{\psi}_q(\delta) \leq \epsilon\},$$

where for some $q > 1$, $\delta_j = q^{-j}$ for $j \in \mathbb{N}^*$.

Finally, in the sequel, constants $K, C > 0$ denote generic constants that may vary from line to line.

6.1.2 Exact case

The proof of Theorem 3 uses the following intermediate lemma.

Lemma 18 (Exact case) *Suppose there exists some function $a : \lambda \mapsto a(\lambda)$ and a constant $0 < r < 1$ such that:*

$$\forall g \in \mathcal{G}, \left|(R - R^\lambda)(g - g^*(g))\right| \leq a(\lambda) + r(R(g) - R(g^*(g))), \tag{12}$$

where $g^*(g) \in \arg\min_h R(h)$ can depend on g.
Then, for any $q > 1$, $\forall \delta \geq \bar{\delta}_\lambda(t)$, we have:

$$\mathbb{P}(R(\hat{g}_n^\lambda) \geq \inf_{g \in \mathcal{G}} R(g) + \delta) \leq \log_q \left(\frac{1}{\delta}\right) e^{-t},$$

where:

$$\bar{\delta}_\lambda(t) = \max\left(\delta_\lambda(t), \frac{8q}{1-r} a(\lambda)\right),$$

for $\delta_\lambda(t) = (U_\lambda(\cdot, t))^\dagger \left((1-r)/4q\right)$ and where we define, for some constant $K > 0$:

$$U_\lambda(\delta, t) := K\left[\mathbb{E}Z_\lambda(\delta) + \sqrt{\frac{t}{n}}\sigma_\lambda(\delta) + \sqrt{\frac{t}{n}\left(1 + 2b_\lambda(\delta)\right)\mathbb{E}Z_\lambda(\delta)} + \frac{t}{3n}\right],$$

where

$$Z_\lambda(\delta) := \sup_{g, g' \in \mathcal{G}(\delta)} \left|(R_n^\lambda - R^\lambda)(g - g')\right|,$$

$$\sigma_\lambda(\delta) := \sup_{g, g' \in \mathcal{G}(\delta)} \sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g, Z) - \ell_\lambda(g', Z))^2},$$

$$b_\lambda(\delta) := \sup_{g \in \mathcal{G}(\delta)} \|\ell_\lambda(g, \cdot)\|_\infty.$$

Proof The proof follows Koltchinskii (2006) extended to the noisy set-up.

Given $q > 1$, we introduce a sequence of positive numbers:

$$\delta_j = q^{-j}, \ \forall j \geq 1.$$

Given $n, j \geq 1$, $t > 0$ and $\lambda \in \mathbb{R}_+^d$, consider the event:

$$E_{\lambda,j}(t) = \left\{Z_\lambda(\delta_j) \leq U_\lambda(\delta_j, t)\right\}.$$

Then, we have, using Theorem 17, for some $K > 0$, $\mathbb{P}(E_{\lambda,j}(t)^C) \leq e^{-t}$, $\forall t > 0$.
We restrict ourselves to the event $E_{\lambda,j}(t)$.
Let $\epsilon < c\delta_{j+1}$ where $c > 0$ is chosen later on. Then, consider some $g \in \mathcal{G}(\epsilon)$, where:

$$\mathcal{G}(\epsilon) = \{g \in \mathcal{G} : R(g) - \inf_{g \in \mathcal{G}} R(g) \leq \epsilon\}.$$

Using assumption (12) and the definition of $\hat{g} := \hat{g}_n^\lambda$, one has:

$$\begin{aligned}
R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g) &\leq R(\hat{g}) - R(g) + \epsilon \\
&\leq (R - R^\lambda)(\hat{g} - g) + (R^\lambda - R_n^\lambda)(\hat{g} - g) + \epsilon \\
&\leq (R^\lambda - R_n^\lambda)(\hat{g} - g) + 2a(\lambda) + r(R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g)) + r(R(g) - \inf_{g\in\mathcal{G}} R(g)) + \epsilon
\end{aligned}$$

Hence, we have the following assertion:

$$\delta_{j+1} \leq R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g) \leq \delta_j \Rightarrow \delta_{j+1} \leq \frac{1}{1-r}\Big((R_n^\lambda - R^\lambda)(g - \hat{g}) + 2a(\lambda) + (1+r)\epsilon\Big).$$

On the event $E_{\lambda,j}(t)$, it follows that $\forall \delta \leq \delta_j$:

$$\begin{aligned}
\delta_{j+1} \leq R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g) \leq \delta_j \Rightarrow \delta_{j+1} &\leq \frac{1}{1-r}\left(U_\lambda(\delta_j, t) + 2a(\lambda) + (1+r)\epsilon\right) \\
&\leq \frac{1}{1-r}\left(\delta_j V_\lambda(\delta, t) + 2a(\lambda)(1+r)\epsilon\right),
\end{aligned}$$

where $V_\lambda(\delta, t) = \breve{U}_\lambda(\delta, t)$ satisfies property (11). We obtain, for any $\delta \leq \delta_j$:

$$\frac{1}{1-r} V_\lambda(\delta, t) \geq \frac{1}{q} - \frac{q^j(2a(\lambda) + (1+r)\epsilon)}{1-r}.$$

The assumption $a(\lambda) \leq (1-r)\delta/8q$ and the choice of $c = \frac{1-r}{4(1+r)}$ in the beginning of the proof gives the following lower bound:

$$V_\lambda(\delta, t) > \frac{1-r}{2q}.$$

It follows from the definition of the $\dagger$-transform that:

$$\delta < [U_\lambda(\cdot, t)]^\dagger\left(\frac{1-r}{2q}\right) = \delta_\lambda(t).$$

Hence, we have on the event $E_{\lambda,j}(t)$, for any $\delta \leq \delta_j$:

$$\delta_{j+1} \leq R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g) \leq \delta_j \Rightarrow \delta < \delta_n^\lambda(t),$$

or equivalently,

$$\delta_\lambda(t) \leq \delta \leq \delta_j \Rightarrow \hat{g} \notin \mathcal{G}(\delta_{j+1}, \delta_j),$$

where $\mathcal{G}(c, C) = \{g \in \mathcal{G} : c \leq R(g) - \inf_{g\in\mathcal{G}} R(g) \leq C\}$. We eventually obtain:

$$\bigcap_{\delta_j \geq \delta} E_{\lambda,j}(t) \text{ and } \delta \geq \delta_\lambda(t) \Rightarrow R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g) \leq \delta.$$

This formulation allows us to write by union's bound:

$$\mathbb{P}(R(\hat{g}) \geq \inf_{g\in\mathcal{G}} R(g) + \delta) \leq \sum_{\delta_j \geq \delta} \mathbb{P}(E_{\lambda,j}(t)^C) \leq \log_q\left(\frac{1}{\delta}\right) e^{-t},$$

since $\{j : \delta_j \geq \delta\} = \{j : j \leq -\frac{\log \delta}{\log q}\}$. ■

6.1.3 The non-exact case

The proof of Theorem 10 uses the following version of Lemma 18.

Lemma 19 (Non-exact case) *Suppose there exists* $a^*(\cdot,\cdot) : (r,\lambda) \in (0,1)\times\mathbb{R}^+ \mapsto a^*(r,\lambda)$ *such that for any* $(r,\lambda) \in (0,1) \times \mathbb{R}_+$*:*

$$\forall g \in \mathcal{G}, \left| R(g) - R^\lambda(g) \right| \leq a^*(r,\lambda) + rR(g). \tag{13}$$

Then, for any $q > 1$*,* $\alpha \in (0,1)$*,* $u \in (0, 1/q)$*,* $\delta \geq \bar{\delta}'_\lambda(t)$*:*

$$\mathbb{P}(R(\hat{g}_n^\lambda) \geq \delta) \leq \log \frac{1}{\delta} e^{-t},$$

where:

$$\bar{\delta}'_\lambda(t) = \max\left(\delta'_\lambda(t), \frac{2}{(1-r)\alpha u} a^*(r,\lambda), \frac{1+r}{(1-r)(1-\alpha)u} \inf_{g\in\mathcal{G}} R(g) \right)$$

for

$$\delta'_\lambda(t) = \left(U'_\lambda(\cdot,t)\right)^\dagger \left(\frac{(1-r)(1-qu)}{2q} \right),$$

and where we define, for some constant $K > 0$*:*

$$U'_\lambda(\delta,t) := K \left[Z'_\lambda(\delta) + \sqrt{\frac{t}{n}} \sigma'_\lambda(\delta) + \sqrt{\frac{t}{n} \left(1 + b'_\lambda(\delta)\right) \mathbb{E} Z'_\lambda(\delta)} + \frac{t}{3n} \right],$$

where here, we write for $\mathcal{G}'(\delta) = \{g \in \mathcal{G} : R(g) \leq \delta\}$*:*

$$Z'_\lambda(\delta) := \sup_{g\in\mathcal{G}'(\delta)} \left| (R_n^\lambda - R^\lambda)(g) \right|,$$

$$\sigma'_\lambda(\delta) := \sup_{g\in\mathcal{G}'(\delta)} \sqrt{\mathbb{E}_{\tilde{P}} (\ell_\lambda(g,Z))^2},$$

$$b'_\lambda(\delta) := \sup_{g\in\mathcal{G}'(\delta)} \|\ell_\lambda(g,\cdot)\|_\infty.$$

Proof The proof follows the proof of Lemma 18 applied to the non-exact case. Given $q > 1$, we introduce a sequence of positive numbers:

$$\delta_j = q^{-j}, \ \forall j \geq 1.$$

Given $n, j \geq 1$, $t > 0$ and $\lambda \in \mathbb{R}_+^d$, consider the event:

$$E'_{\lambda,j}(t) = \left\{ Z'_\lambda(\delta_j) \leq U'_\lambda(\delta_j, t) \right\}.$$

Then, we have that, using Theorem 17, $\mathbb{P}(E'_{\lambda,j}(t)^C) \leq e^{-t}$.
We restrict ourselves to the event $E'_{\lambda,j}(t)$.
Using assumption (13), we have, for any $g \in \mathcal{G}$ and any $r \in (0,1)$:

$$R(\hat{g}) \leq \frac{1}{1-r} \left((R^\lambda - R_n^\lambda)(\hat{g}) + a^*(r,\lambda) + R_n^\lambda(g) \right),$$

where we use the definition of $\hat{g} = \hat{g}_n^\lambda$. Moreover, note that, using again assumption (13):

$$\begin{aligned} R_n^\lambda(g) &= (R_n^\lambda - R^\lambda)(g) + (R^\lambda - R)(g) + R(g) \\ &\leq (R_n^\lambda - R^\lambda)(g) + a^*(r,\lambda) + (1+r)R(g) \end{aligned}$$

Then, we have, for $g = g^* \in \arg\min_{\mathcal{G}} R(g)$:

$$R(\hat{g}) \leq \frac{1}{1-r}\left((R_n^\lambda - R^\lambda)(g^* - \hat{g}) + 2a^*(r,\lambda) + (1+r)\inf_{g\in\mathcal{G}} R(g)\right).$$

We hence have on the event $E'_{\lambda,j}(t)$:

$$\delta_{j+1} \leq R(\hat{g}) \leq \delta_j \Rightarrow \delta_{j+1} \leq \frac{1}{1-r}\left(2U'_\lambda(\delta_j,t) + 2a^*(r,\lambda) + (1+r)\inf_{g\in\mathcal{G}} R(g)\right),$$

since in this case $R(g^*) \leq \delta_j$. On the event $E'_{\lambda,j}(t)$, it follows that $\forall \delta \leq \delta_j$:

$$\delta_{j+1} \leq R(\hat{g}) \leq \delta_j \Rightarrow \delta_{j+1} \leq \frac{1}{1-r}\left(2\delta_j V'_\lambda(\delta,t) + 2a^*(r,\lambda) + (1+r)\inf_{g\in\mathcal{G}} R(g)\right),$$

where $V'_\lambda(\delta,t) = \breve{U}'_\lambda(\cdot,t)$ is defined as above. We obtain, for any $u \in (0,1/q)$:

$$\frac{2}{1-r}V'_\lambda(\delta,t) \geq \frac{1}{q} - \frac{q^j}{1-r}(2a(\lambda) + (1+r)\inf_{g\in\mathcal{G}} R(g)) > \frac{1}{q} - u, \tag{14}$$

provided that for any $\alpha \in (0,1)$, since $\delta \leq \delta_j$:

$$a^*(r,\lambda) \leq \alpha\frac{u(1-r)\delta}{2} \text{ and } \inf_{g\in\mathcal{G}} R(g) \leq (1-\alpha)\frac{u(1-r)}{1+r}\delta.$$

From (14), on the event $E_{\lambda,j}(t)$, for any $\frac{1+r}{u(1-\alpha)(1-r)}\inf_{g\in\mathcal{G}} R(g) \vee \frac{2}{(1-r)\alpha u}a^*(r,\lambda) \leq \delta \leq \delta_j$:

$$\delta_{j+1} \leq R(\hat{g}) \leq \delta_j \Rightarrow \delta \leq \delta'_\lambda(t) := [U'_\lambda(\cdot,t)]^\dagger\left(\frac{1-r}{2q} - \frac{(1-r)u}{2}\right),$$

or equivalently, by definition of $\bar{\delta}'_\lambda(t)$:

$$\bar{\delta}'_\lambda(t) \leq \delta \leq \delta_j \Rightarrow \hat{g} \notin \mathcal{G}'(\delta_{j+1},\delta_j),$$

where here $\mathcal{G}'(c,C) = \{g \in \mathcal{G} : c \leq R(g) \leq C\}$. We eventually obtain:

$$\bigcap_{\delta_j \geq \delta} E_{\lambda,j}(t) \text{ and } \delta \geq \bar{\delta}'_\lambda(t) \Rightarrow R(\hat{g}) \leq \delta.$$

This formulation allows us to write by union's bound, exactly as in the proof of Lemma 18:

$$\mathbb{P}(R(\hat{g}) \geq \delta) \leq \sum_{\delta_j \geq \delta} \mathbb{P}(E_{\lambda,j}(t)^C) \leq \log_q\left(\frac{1}{\delta}\right)e^{-t}, \tag{15}$$

where $\delta \geq \bar{\delta}'_\lambda(t)$.

■

6.2 Proof of Theorem 3 and 10

6.2.1 Proof of Theorem 3

The proof of Theorem 3 is divided into two steps. Using Lemma 18, we obtain an exact oracle inequality when $|\mathcal{G}(0)| = 1$. For the general case, we will introduce a more sophisticated localization explain in (Koltchinskii, 2006, Section 4). Moreover, we begin the proof in dimension $d = 1$ for simplicity. A slightly different algebra is precised at the end of the proof to lead to the general case.

Case 1: $|\mathcal{G}(0)| = 1$.

When $|\mathcal{G}(0)| = 1$, it is important to note that $\mathbf{MA}(\kappa)$ holds with a minimizer $g^* \in \mathcal{G}$ which does not depend on g. Then, we can write, for any $g, g' \in \mathcal{G}(\delta)$:

$$\|\ell(g) - l(g')\|_{L_2} \le \|\ell(g) - l(g^*)\|_{L_2} + \|\ell(g') - l(g^*)\|_{L_2} \le 2\sqrt{\kappa_0}\delta^{1/2\kappa}.$$

Gathering with the entropy condition (9), we obtain:

$$\begin{aligned}
\mathbb{E} \sup_{g,g' \in \mathcal{G}(\delta)} \left|(R_n^\lambda - R^\lambda)(g - g')\right| &\le \mathbb{E} \sup_{\|\ell(g)-\ell(g')\|_{L_2} \le 2\sqrt{\kappa_0}\delta^{1/2\kappa}} \left|(R_n^\lambda - R^\lambda)(g - g')\right| \\
&\le C\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}},
\end{aligned}$$

where we use in last line Lemma 1 in Loustau (2012). Then, using the notations of Lemma 18:

$$\begin{aligned}
U_\lambda(\delta, t) &= K\left[\mathbb{E}Z_\lambda(\delta) + \sqrt{\frac{t}{n}}\sigma_\lambda(\delta) + \sqrt{\frac{t}{n}\left(1 + 2b_\lambda(\delta)\right)\mathbb{E}Z_\lambda(\delta)} + \frac{t}{3n}\right] \\
&\le K\left[\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}} + \sqrt{\frac{t}{n}}\sigma_\lambda(\delta) + \sqrt{\frac{t}{n}\left(1 + 2b_\lambda(\delta)\right)\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}}} + \frac{t}{3n}\right].
\end{aligned}$$

It remains to control the $L^2(\tilde{P})$-diameter $\sigma_\lambda(\delta)$ and the term $b_\lambda(\delta)$ thanks to Lemma 20. Using again assumption $\mathbf{MA}(\kappa)$, and the unicity of the minimizer g^*, gathering with the first assertion of Lemma 20, we can write:

$$\sigma_\lambda(\delta) = \sup_{g,g' \in \mathcal{G}(\delta)} \sqrt{\mathbb{E}_{\tilde{P}}(l_\lambda(g, Z) - l_\lambda(g', Z))^2} \le C\lambda^{-\beta}\sqrt{\kappa_0}\delta^{\frac{1}{2\kappa}}.$$

Now, by the second assertion of Lemma 20:

$$b_\lambda(\delta) = \sup_{g \in \mathcal{G}(\delta)} \|l_\lambda(g, \cdot)\|_\infty \le C\lambda^{-\beta-1/2}.$$

It follows that:

$$U_\lambda(\delta, t) \le K\left[\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}} + \sqrt{t}\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1}{2\kappa}} + \sqrt{\frac{t}{n}\left(1 + \lambda^{-\beta-1/2}\right)\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}}} + \frac{t}{3n}\right]. \tag{16}$$

We hence have the following assertion:

$$t \le \delta^{-\frac{\rho}{\kappa}} \wedge \sqrt{n}\lambda^{-\beta}\delta^{\frac{1-\rho}{2\kappa}} \Rightarrow U'_\lambda(\delta, t) \le K\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}}.$$

From an easy calculation, we hence get in this case:

$$\delta_\lambda(t) \leq K\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}},$$

where $K>0$ is a generic constant. We are now on time to apply Lemma 18 with:

$$\delta = K\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}} \text{ and } t' = t + \log\log_q n.$$

In this case, note that for any $t>0$ independent on n, the choice of λ in Theorem 3 warrants that, for any $n \geq n_0(t)$:

$$t + \log\log_q n \leq \delta^{-\frac{\rho}{\kappa}} \wedge \sqrt{n}\lambda^{-\beta}\delta^{\frac{1-\rho}{2\kappa}}.$$

Moreover, using Lemma 21, we have in dimension $d=1$:

$$\forall g \in \mathcal{G}, \left|(R-R^\lambda)(g-g^*)\right| \leq C\lambda^{2s} + \frac{1}{2}(R(g)-R(g^*)).$$

As a result condition (12) of Lemma 18 is satisfied with $r=1/2$ and $a(\lambda)=\lambda^{2s}$. We can also check that for n great enough, the choice of λ in Theorem 3 guarantees:

$$\lambda^{2s} \leq K\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}}.$$

Finally, we get the result since:

$$\log_q \frac{1}{\delta} e^{-t'} \leq \left(\frac{2\kappa}{2\kappa+\rho-1}\right)\log\left(\frac{\sqrt{n}}{\lambda^{-\beta}}\right)\frac{e^{-t}}{\log_q(n)} \leq e^{-t}.$$

For the d-dimensional case, we have the same algebra by replacing $\lambda^{-\beta}$ by $\Pi_{j=1}^d \lambda_j^{-\beta_j}$ in the previous calculus and λ^{2s} by $\sum_{j=1}^d \lambda_j^{2s_j}$ thanks to Lemma 21. The choice of λ_j, for $j=1,\ldots,d$ in Theorem 3 allows to conclude.

Case 2: $|\mathcal{G}(0)| \geq 2$.

When the infimum is not unique, the diameter $\sigma_\lambda^2(\delta)$ does not necessary tend to zero when $\delta \to 0$. We hence introduce the more sophisticated geometric parameter:

$$r(\sigma,\delta) = \sup_{g\in\mathcal{G}(\delta)} \inf_{g'\in\mathcal{G}(\sigma)} \sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z)-\ell_\lambda(g',Z))^2}, \text{ for } 0<\sigma\leq\delta.$$

It is clear that $r(\sigma,\delta) \leq \sqrt{\sigma_\lambda^2(\delta)}$ and for $\delta \to 0$, we have $r(\sigma,\delta) \to 0$. The idea of the proof is to use a modified version of Lemma 18 following (Koltchinskii, 2006, Theorem 4). More precisely, we have to apply the concentration inequality of Theorem 17 to the random variable:

$$W_\lambda(\delta) = \sup_{g\in\mathcal{G}(\sigma)} \sup_{g'\in\mathcal{G}(\delta):\sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z)-\ell_\lambda(g',Z))^2}\leq r(\sigma,\delta)+\epsilon} \left|(R_n^\lambda - R^\lambda)(g-g')\right|.$$

This localization guarantees the upper bounds of Theorem 3 when $|\mathcal{G}(0)| \geq 2$. However, to this end, we have to check (for $d=1$ for simplicity):

$$\lim_{\epsilon\to 0}\mathbb{E}\sup_{g\in\mathcal{G}(\sigma)}\sup_{g'\in\mathcal{G}(\delta):\sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z)-\ell_\lambda(g',Z))^2}\leq r(\sigma,\delta)+\epsilon}\left|(R_n^\lambda - R^\lambda)(g-g')\right| \leq C\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{1/2\kappa}, \quad (17)$$

and for $0<\sigma\leq\delta$:

$$r(\sigma,\delta)\leq C\lambda^{-\beta}\delta^{1/2\kappa}. \quad (18)$$

Using $\mathbf{MA}(\kappa)$ and Lemma 1 in Loustau (2012), it is clear that (17) holds since:

$$\begin{aligned}
&\mathbb{E}\sup_{g\in\mathcal{G}(\sigma)}\sup_{g'\in\mathcal{G}(\delta):\sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z)-\ell_\lambda(g',Z))^2}\leq r(\sigma,\delta)+\epsilon}\left|(R_n^\lambda - R^\lambda)(g-g')\right| \\
&\leq \mathbb{E}\sup_{g\in\mathcal{G}(\sigma),g^*\in\mathcal{G}(0)}\left|(R_n^\lambda - R^\lambda)(g-g^*)\right| + \mathbb{E}\sup_{g'\in\mathcal{G}(\delta)}\left|(R_n^\lambda - R^\lambda)(g'-g^*(g'))\right| \\
&\leq 2\mathbb{E}\sup_{(g,g^*)\in\mathcal{G}(\delta)\times\mathcal{G}(0)}\left|(R_n^\lambda - R^\lambda)(g^*-g)\right| \\
&\leq C\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{1/2\kappa}.
\end{aligned}$$

To check (18), note that with $\mathbf{MA}(\kappa)$ and the first assertion of Lemma 20, we have $\forall g\in\mathcal{G}(\delta), g'\in\mathcal{G}(\sigma)$:

$$\begin{aligned}
\sqrt{\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z)-\ell_\lambda(g',Z))^2} &\leq C\lambda^{-\beta}\|\ell(g)-\ell(g')\|_{L_2} \\
&\leq C\lambda^{-\beta}\delta^{1/2\kappa} + C\lambda^{-\beta}\|\ell(g^*(g))-\ell(g^*(g'))\|_{L_2},
\end{aligned}$$

for $0<\sigma\leq\delta$. Taking the infimum with respect to $g'\in\mathcal{G}(\sigma)$, we get:

$$\|\ell(g^*(g))-\ell(g^*(g'))\|_{L_2}=0.$$

6.2.2 Proof of Theorem 10

The main ingredient of the proof is Lemma 19. We want to find a convenient bound for the term (see the notations of Lemma 19):

$$U'_\lambda(\delta,t) = K\left[Z'_\lambda(\delta)+\sqrt{\frac{t}{n}}\sigma'_\lambda(\delta)+\sqrt{\frac{t}{n}\left(1+b'_\lambda(\delta)\right)\mathbb{E}Z'_\lambda(\delta)}+\frac{t}{3n}\right].$$

First note that since $\ell(g,\cdot)$ is bounded, we have the crude bound $\mathbb{E}_P\ell(g,X)^2\leq MR(g)$, where $M=\|\ell(g,\cdot)\|_\infty$. Hence, we have, using the entropy condition:

$$\begin{aligned}
\mathbb{E}Z'_\lambda(\delta) &= \mathbb{E}\sup_{g\in\mathcal{G}'(\delta)}\left|(R_n^\lambda - R^\lambda)(g)\right| \\
&\leq \mathbb{E}\sup_{\|\ell(g)\|_{L_2(P)}\leq\sqrt{M}\delta^{1/2}}\left|(R_n^\lambda - R^\lambda)(g)\right| \\
&\leq C\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2}},
\end{aligned}$$

where we use in last line Lemma 1 in Loustau (2012).
We obtain:

$$U'_\lambda(\delta,t) \quad \leq \quad K\left[\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2}}+\sqrt{\frac{t}{n}}\sigma'_\lambda(\delta)+\sqrt{\frac{t}{n}\left(1+b'_\lambda(\delta)\right)\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2}}}+\frac{t}{3n}\right].$$

Now, from Lemma 23, we have the following control of $\sigma'_\lambda(\delta)$:

$$\sigma'_\lambda(\delta)=\sup_{g\in\mathcal{G}'(\delta)}\sqrt{E_{\tilde{P}}\ell_\lambda(g)^2}\leq C\lambda^{-\beta}\sqrt{\mathbb{E}\ell(g,X)^2}\leq C\lambda^{-\beta}\sqrt{\delta},$$

where $C>0$ is a generic constant and where we use in the last inequality the boundedness assumption of $\ell(g,\cdot)$. Now by the second assertion of Lemma 20:

$$b'_\lambda(\delta)=\sup_{g\in\mathcal{G}(\delta)}\|l_\lambda(g,\cdot)\|_\infty\leq C\lambda^{-\beta-1/2}.$$

It follows that:

$$U'_\lambda(\delta,t)\leq K\left[\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2}}+\sqrt{\frac{t}{n}}\lambda^{-\beta}\delta^{\frac{1}{2}}+\sqrt{\frac{t}{n}}\frac{\lambda^{-\beta}}{\sqrt{n}}+\sqrt{\frac{t}{n}\left(1+\lambda^{-\beta-1/2}\right)\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2}}}+\frac{t}{3n}\right]. \tag{19}$$

We hence have in this case the following assertion:

$$t\leq\delta^{-2\rho}\wedge n\delta^{-\rho}\wedge\sqrt{n\lambda}\delta^{\frac{1-\rho}{2}}\Rightarrow U'_\lambda(\delta,t)\leq K\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2}}.$$

From an easy calculation, we hence get with the notations of Lemma 19:

$$\delta'_\lambda(t)\leq K\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2}{1+\rho}}, \tag{20}$$

where $K>0$ is a generic constant. Let us consider, for any $\epsilon>0$:

$$\delta=\frac{K\vee 2C}{\alpha_\epsilon u_\epsilon(1-r_\epsilon)r_\epsilon}\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2}{1+\rho}}+(1+\epsilon)\inf_{g\in\mathcal{G}}R(g),$$

where $(r_\epsilon,\alpha_\epsilon,u_\epsilon)\in(0,1)^2\times(0,1/q)$ are chosen later on as a function of $\epsilon>0$. Using Lemma 24, we have in dimension $d=1$, for any $r\in(0,1)$:

$$\forall g\in\mathcal{G},\ \left|(R-R^\lambda)(g)\right|\leq\frac{C}{r}\lambda^{2s}+rR(g).$$

As a result, condition (13) of Lemma 19 is satisfied with $a^*(r,\lambda)=C\lambda^{2s}/r$. The choice of λ in Theorem 10 warrants that:

$$\lambda^{2s}\leq\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2}{1+\rho}}. \tag{21}$$

Moreover, for any $\epsilon > 0$, we can find a triplet $(r_\epsilon, \alpha_\epsilon, u_\epsilon) \in (0,1)^2 \times (0, 1/q)$ such that:

$$1 + \epsilon \geq \frac{1 + r_\epsilon}{(1 - r_\epsilon) u_\epsilon (1 - \alpha_\epsilon)}. \tag{22}$$

Inequalities (20), (21) and (22) give us:

$$\delta \geq \max\left(\delta'_\lambda(t), \frac{1 + r_\epsilon}{(1 - r_\epsilon) u_\epsilon (1 - \alpha_\epsilon)} \inf_{g \in \mathcal{G}} R(g), \frac{2}{(1 - r_\epsilon)\alpha_\epsilon u_\epsilon} a^*(r_\epsilon, \lambda) \right).$$

Finally, we can apply Lemma 19 with the triplet $(r_\epsilon, \alpha_\epsilon, u_\epsilon)$, $t' = t + \log\log_q n$ and get the result since:

$$\log_q \frac{1}{\delta} e^{-t'} \leq \frac{2}{1+\rho} \log\left(\frac{\sqrt{n}}{\lambda^{-\beta}} \right) \frac{e^{-t}}{\log_q n} \leq e^{-t}.$$

6.3 Proof of Theorem 14

The proof of Theorem 14 uses a slightly different version of Theorem 3. First of all, an inspection of the proof of Theorem 3 shows that condition (9) in Theorem 3 can be replaced by the following control of the local complexity of the noisy empirical process:

$$\mathbb{E} \sup_{g, g' \in \mathcal{G}(\delta)} \left| (R_n^\lambda - R^\lambda)(g - g') \right| \leq C \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1-\rho}{2\kappa}}. \tag{23}$$

Hence, using Lemma 25 in the Appendix, gathering with condition **(PRC)**, we can have (23) with $\rho = 0$.
However, the case $\rho = 0$ is not treated in Theorem 3 where $\rho \in (0,1)$. From (23), and using the notations of Lemma 18, (16) in the proof of Theorem 3 becomes:

$$U_\lambda(\delta, t) \leq K \left[\frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}} + \sqrt{t} \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}} + \sqrt{\frac{t}{n} \left(1 + \lambda^{-\beta - 1/2}\right) \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}}} + \frac{t}{3n} \right].$$

We hence have the following assertion:

$$t \leq \sqrt{n} \lambda^{-\beta} \delta^{\frac{1}{2}} \Rightarrow U_\lambda(\delta, t) \leq K \left(1 + \sqrt{t}\right) \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}}.$$

Using the same algebra as above, we can use Lemma 18 with:

$$\delta = K \left(1 + \sqrt{t'}\right) \left(\frac{\lambda^{-\beta}}{\sqrt{n}} \right)^{\frac{2}{1+\rho}} \text{ and } t' = t + \log\log_q n.$$

In this case, note that the choice of $t' = t + \log\log_q n$ gives rise to the following asymptotic:

$$\delta \approx \sqrt{\log\log n} \frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1}{2}},$$

and leads to an extra $\sqrt{\log\log n}$ term in the rates of convergence.

7. Appendix

7.1 Technical lemmas for the exact case

Lemma 20 *Suppose* ***(NA)*** *holds, and $\mathcal{K}$ satisfies assumption* ***(K1)****. Suppose $\|f*\eta\|_\infty \leq \tilde{c}_\infty$ and $\sup_{g\in\mathcal{G}}\|\ell(g,\cdot)\|_{L_2(K)}<\infty$. Then, the two following assertions hold:*

(i) *$\ell(g)\mapsto\ell_\lambda(g)$ is Lipschitz with respect to λ:*

$$\forall g,g'\in\mathcal{G},\ \|\ell_\lambda(g,\cdot)-\ell_\lambda(g',\cdot)\|_{L_2(\tilde{P})}\leq C_1\Pi_{i=1}^d\lambda_i^{-\beta_i}\|\ell(g,\cdot)-\ell(g',\cdot)\|_{L_2},$$

where $C>0$ is a generic constant which depends on $\tilde{c}_\infty$ and constants in ***(K1)****.*

(ii) *$\{\ell_\lambda(g),g\in\mathcal{G}\}$ is uniformly bounded:*

$$\sup_{g\in\mathcal{G}}\|\ell_\lambda(g,\cdot)\|_\infty\leq C_2\Pi_{i=1}^d\lambda_i^{-(\beta_i+1/2)},$$

where $C_2>0$ is a generic constant which depends on constants in ***(K1)****.*

Proof Using Plancherel and the boundedness assumption over $f*\eta$, we have:

$$\begin{aligned}
\mathbb{E}_{\tilde{P}}(\ell_\lambda(g,Z)-\ell_\lambda(g',Z))^2 &= \int\left[\frac{1}{\lambda}\mathcal{K}_\eta(\frac{\cdot}{\lambda})*(\mathbb{1}_K\times(\ell(g,\cdot)-\ell(g',\cdot))(z)\right]^2 f*\eta(z)dz\\
&\leq C\int\frac{1}{\lambda^2}|\mathcal{F}[\mathcal{K}_\eta(\frac{\cdot}{\lambda})](t)|^2|\mathcal{F}[\mathbb{1}_K\times(\ell(g,\cdot)-\ell(g',\cdot))](t)|^2dt\\
&\leq C\lambda^{-2\beta}\|\ell(g)-\ell(g')\|_{L_2}^2,
\end{aligned}$$

where we use in last line the following inequalities:

$$\frac{1}{\lambda^2}|\mathcal{F}[\mathcal{K}_\eta(./\lambda)](s)|^2=|\mathcal{F}[\mathcal{K}_\eta](s\lambda)|^2\leq C\sup_{t\in\mathbb{R}}\left|\frac{\mathcal{F}[\mathcal{K}](t\lambda)}{\mathcal{F}[\eta](t)}\right|^2\leq C\sup_{t\in[-\frac{L}{\lambda},\frac{L}{\lambda}]}\left|\frac{1}{\mathcal{F}[\eta](t)}\right|^2\leq C\lambda^{-2\beta},$$

provided that **(K1)** holds.

By the same way, the second assertion holds since if $\ell(g,\cdot)\in L^2(K)$:

$$\begin{aligned}
|\ell_\lambda(g,z)| &\leq \int_K\left|\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)\ell(g,x)\right|dx\\
&\leq C\sqrt{\int_K\left|\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)\right|^2dx}\\
&\leq C\lambda^{-\beta-1/2}.
\end{aligned}$$

A straightforward generalization leads to the d-dimensional case. ∎

Lemma 21 *Suppose f belongs to the anisotropic Hölder spaces $\mathcal{H}(s,L)$ with $s=(s_1,\ldots,s_d)$. Let $\mathcal{K}$ a kernel satisfying assumption* ***K(m)*** *with $m=\lfloor s\rfloor \in \mathbb{N}^d$. Suppose* ***MA(κ)*** *holds with parameter $\kappa \geq 1$. Then, we have:*

$$\forall g \in \mathcal{G}, \left|(R-R^\lambda)(g-g^*(g))\right| \leq C\sum_{j=1}^d \lambda_j^{2\kappa s_j/(2\kappa-1)} + \frac{1}{2\kappa}(R(g)-\inf_{g\in\mathcal{G}} R(g)),$$

where $C>O$ is a generic constant.

Proof Note that we can write:

$$(R^\lambda - R)(g-g^*) \quad = \quad \int_K (\ell(g,x)-\ell(g^*,x))\left(\mathbb{E}\hat{f}_\lambda(x)-f(x)\right)dx,$$

where we omit the notation $g^* = g^*(g)$ for simplicity. The first part of the proof uses Proposition 1 stated in Comte and Lacour (2012).

Proposition 22 (Comte and Lacour (2012)) *Let $B_0(\lambda)=\sup_{x_0\in\mathbb{R}^d}|f(x_0)-\mathbb{E}\hat{f}_\lambda(x_0)|$. Then, if f belongs to the anisotropic Hölder space $\mathcal{H}(s,L)$, and $\mathcal{K}$ is a kernel of order $\lfloor s\rfloor$, we have:*

$$B_0(\lambda) \leq C\sum_{j=1}^d \lambda_j^{s_j},$$

where $C>0$ denotes some generic constant.

The rest of the proof uses the margin assumption **MA(κ)** as follows:

$$\begin{aligned}
\left|(R^\lambda - R)(g-g^*)\right| &\leq C\sum_{j=1}^d \lambda_j^{s_j}\int_K |\ell(g,x)-\ell(g^*,x)|dx. \\
&\leq C\sum_{j=1}^d \lambda_j^{s_j}\sqrt{\int_K |\ell(g,x)-\ell(g^*,x)|^2 dx} \\
&\leq C\sum_{j=1}^d \lambda_j^{s_j}\left(R(g)-R(g^*)\right)^{\frac{1}{2\kappa}} \\
&\leq C\sum_{j=1}^d \lambda_j^{2\kappa s_j/(2\kappa-1)} + \frac{1}{2\kappa}(R(g)-\inf_{g\in\mathcal{G}} R(g)),
\end{aligned}$$

where we use in last line Young's inequality:

$$xy^r \leq ry + x^{1/1-r}, \forall r<1,$$

with $r=\frac{1}{2\kappa}$. ■

7.2 Technical lemmas for the non-exact case

Lemma 23 *Suppose **(NA)** and $\mathbf{DA}(c_0)$ holds, and $\mathcal{K}$ satisfies assumption **(K1)**. Suppose $\|f*\eta\|_\infty \leq \tilde{c}_\infty$ and $\sup_{g\in\mathcal{G}}\|\ell(g,\cdot)\|_{L_2(K)}<\infty$. Then, we have:*

$$\forall g\in\mathcal{G},\ \sqrt{\mathbb{E}_{\tilde{P}}\ell_\lambda(g,Z)^2}\leq C_1'\Pi_{i=1}^d\lambda_i^{-\beta_i}\sqrt{\mathbb{E}_P\ell(g,X)^2},$$

*where $C_1'>0$ is a generic constant which depends on c_0, $\tilde{c}_\infty$ and constants in **(K1)**.*

Proof Using Plancherel and the boundedness assumption over $f*\eta$, we have as above:

$$\begin{aligned}
\mathbb{E}_{\tilde{P}}\ell_\lambda(g,Z)^2 &= \int\left[\frac{1}{\lambda}\mathcal{K}_\eta(\frac{\cdot}{\lambda})*\ \mathbb{1}_K\times\ell(g,\cdot)(z)\right]^2 f*\eta(z)dz\\
&\leq C\lambda^{-2\beta}\int_K|\ell(g,z)|^2dz\\
&\leq C\frac{\lambda^{-2\beta}}{c_0}\int_K|\ell(g,z)|^2f(z)dz\\
&\leq C\lambda^{-2\beta}P\ell(g,X)^2,
\end{aligned}$$

where we use in the third line assumption $\mathbf{DA}(c_0)$. ∎

Lemma 24 *Suppose f belongs to the anisotropic Hölder spaces $\mathcal{H}(s,L)$ with $s=(s_1,\ldots,s_d)$. Let $\mathcal{K}$ a kernel satisfying assumption **K(m)** with $m=\lfloor s\rfloor$. Then, we have, for any $r>0$:*

$$\forall g\in\mathcal{G},\ \left|R(g)-R^\lambda(g)\right|\leq\frac{C}{r}\sum_{j=1}^d\lambda_j^{2s_j}+rR(g),$$

where $C>O$ is a generic constant which does not depend on $r>0$.

Proof We follow the first part of the proof of Lemma 21 to get:

$$\left|R^\lambda(g)-R(g)\right| \leq C\sum_{j=1}^d\lambda_j^{s_j}\int_K|\ell(g,x)|dx.$$

Now using $\mathbf{DA}(c_0)$, we have, for any $r>0$:

$$\begin{aligned}
\left|R^\lambda(g)-R(g)\right| &\leq C\sum_{j=1}^d\lambda_j^{s_j}\sqrt{\int_K|\ell(g,x)|^2dx}\\
&\leq \frac{C\sum_{j=1}^d\lambda_j^{s_j}}{\sqrt{c_0}}\sqrt{\mathbb{E}_P\ell(g,X)^2}\\
&\leq C\sum_{j=1}^d\lambda_j^{s_j}\left(R(g)\right)^{\frac{1}{2}}
\end{aligned}$$

$$= \frac{C}{\sqrt{2r}} \sum_{j=1}^{d} \lambda_j^{s_j} \left(2rR(g)\right)^{\frac{1}{2}}$$
$$\leq \frac{C}{2r} \sum_{j=1}^{d} \lambda_j^{2s_j} + rR(g),$$

where we use in last line Young's inequality:

$$xy^a \leq ay + x^{1/1-a}, \forall a < 1,$$

with $a = \frac{1}{2}$. ■

7.3 Technical lemma for Theorem 14

Lemma 25 *Suppose* ***(PRC)****,* ***(NA)*** *and the kernel assumption* ***(K1)*** *are satisfied and* $\|X\|_\infty \leq M$*. Suppose* $\mathbb{E}\|\epsilon\|^2 < \infty$*. Then:*

$$\mathbb{E} \sup_{(\mathbf{c},\mathbf{c}^*)\in\mathcal{C}\times\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta} \left|(R_n^\lambda - R^\lambda)(\mathbf{c}^* - \mathbf{c})\right| \leq C\Pi_{i=1}^d \lambda_i^{-\beta_i} \frac{\sqrt{\delta}}{\sqrt{n}},$$

where $C > 0$ *is a positive constant.*

Proof The proof follows Levrard (2012) applied to the noisy setting. First note that in the sequel, we need to introduce the following notation:

$$(\tilde{P}_n - \tilde{P})(\gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}', Z) := \frac{1}{n}\sum_{i=1}^{n} \left[\gamma_\lambda(\mathbf{c}, Z_i) - \gamma_\lambda(\mathbf{c}', Z_i)\right] - \mathbb{E}_{\tilde{P}}\left[\gamma_\lambda(\mathbf{c}, Z) - \gamma_\lambda(\mathbf{c}', Z)\right].$$

By smoothness assumptions over $\mathbf{c} \mapsto \min\|x - c_j\|$, for any $\mathbf{c} \in \mathbb{R}^{dk}$ and $\mathbf{c}^* \in \mathcal{M}$, we have:

$$\gamma_\lambda(\mathbf{c}, z) - \gamma_\lambda(\mathbf{c}^*, z) = \langle \mathbf{c} - \mathbf{c}^*, \nabla_{\mathbf{c}}\gamma_\lambda(\mathbf{c}^*, z)\rangle + \|\mathbf{c} - \mathbf{c}^*\| R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z),$$

where, with Pollard (1982) we have:

$$\nabla_{\mathbf{c}}\gamma_\lambda(\mathbf{c}^*, z) = -2\left(\int \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)(x - c_1^*)\mathbf{1}_{V_1^*}(x)dx, ..., \int \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)(x - c_k^*)\mathbf{1}_{V_k^*}(x)dx\right)$$

and $R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z)$ satisfies:

$$|R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z)| \leq \|\mathbf{c} - \mathbf{c}^*\|^{-1}\left(|\langle \mathbf{c} - \mathbf{c}^*, \nabla_{\mathbf{c}}\gamma_\lambda(\mathbf{c}^*, z)\rangle| + \max_{j=1,\ldots k}(|\|z - \mathbf{c}_j\| - \|x - \mathbf{c}_j^*\|\right).$$

Splitting the expectation in two parts, we obtain:

$$\mathbb{E} \sup_{\mathbf{c}^*\in\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta} |\tilde{P}_n - \tilde{P}|(\gamma_\lambda(\mathbf{c}^*, .) - \gamma_\lambda(\mathbf{c}, .)) \leq \mathbb{E} \sup_{\mathbf{c}^*\in\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta} |\tilde{P}_n - \tilde{P}| \langle \mathbf{c}^* - \mathbf{c}, \nabla_{\mathbf{c}}\gamma_\lambda(\mathbf{c}^*, .)\rangle$$
$$+ \sqrt{\delta}\mathbb{E} \sup_{\mathbf{c}^*\in\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq\delta} |\tilde{P}_n - \tilde{P}|(-R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, .)) \quad (24)$$

To bound the first term in this decomposition, consider the random variable

$$Z_n = (\tilde{P}_n - \tilde{P}) \langle \mathbf{c}^* - \mathbf{c}, \nabla_{\mathbf{c}} \gamma_\lambda(\mathbf{c}^*, .) \rangle = \frac{2}{n} \sum_{u=1}^{k} \sum_{j=1}^{d} (c_{u,j} - c_{u,j}^*) \sum_{i=1}^{n} \int_{V_u} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{Z_i - x}{\lambda} \right) (x_j - c_{u,j}) dx.$$

By a simple Hoeffding's inequality, Z_n is a subgaussian random variable. Its variance can be bounded as follows:

$$\begin{aligned}
\mathrm{var} Z_n &= \frac{4}{n} \sum_{u=1}^{k} \sum_{j=1}^{d} (c_{u,j} - c_{u,j}^*)^2 \mathrm{var} \int_{V_u} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{Z - x}{\lambda} \right) (x_j - c_{u,j}) dx \\
&\leq \frac{4}{n} \delta \mathbb{E} \left(\int_{V_{u^+}} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{Z - x}{\lambda} \right) (x_j - c_{u^+,j}) dx \right)^2 \\
&\leq C \frac{4}{n} \delta \int \left| \mathcal{F} \left[\frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{\cdot}{\lambda} \right) \right] (t) \right|^2 \left| \mathcal{F}[(\pi_j - c_{u^+,j}) 1_{V_{u^+}}](t) \right|^2 dt \\
&\leq C \frac{4}{n} \delta \Pi_{i=1}^d \lambda_i^{-2\beta_i} \int_{V_{u^+}} (x_j - c_{u^+,j})^2 dx \\
&\leq C \Pi_{i=1}^d \lambda_i^{-2\beta_i} \frac{4}{n} \delta,
\end{aligned}$$

where $u^+ = \arg\max_u \int_{V_u} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{Z-x}{\lambda} \right) (x_j - c_{u,j}) dx$ and $\pi_j : x \mapsto x_j$, and where we use the same argument as in Lemma 20 under assumption **(K1)**. We hence have using for instance a maximal inequality due to Massart Massart (2007, Part 6.1):

$$\mathbb{E} \left(\sup_{\mathbf{c}^* \in \mathcal{M}, \|\mathbf{c} - \mathbf{c}^*\|^2 \leq \delta} (\tilde{P}_n - \tilde{P}) \langle \mathbf{c}^* - \mathbf{c}, \nabla_{\mathbf{c}} \gamma_\lambda(\mathbf{c}^*, .) \rangle \right) \leq C \frac{\Pi_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n}} \sqrt{\delta}.$$

We obtain for the first term in (24) the right order. To prove that the second term in (24) is smaller, note that from Pollard (1982), we have:

$$\begin{aligned}
|R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z)| &\leq \|\mathbf{c} - \mathbf{c}^*\|^{-1} \left(\langle \mathbf{c} - \mathbf{c}^*, \nabla_{\mathbf{c}} \gamma_\lambda(\mathbf{c}^*, z) \rangle + \max_{j=1,\ldots k} (| \|z - \mathbf{c}_j\|^2 - \|z - \mathbf{c}_j^*\|^2 | \right) \\
&\leq \|\nabla_{\mathbf{c}} \gamma_\lambda(\mathbf{c}^*, z)\| + \|\mathbf{c} - \mathbf{c}^*\|^{-1} \sum_{j=1,\ldots k} | \|z - \mathbf{c}_j\|^2 - \|z - \mathbf{c}_j^*\|^2 | \\
&\leq C(\Pi_{i=1}^d \lambda_i^{-\beta_i} + \|z\|)
\end{aligned}$$

we we use in last line:

$$\|\nabla_{\mathbf{c}} \gamma_\lambda(\mathbf{c}^*, z)\|^2 = 4 \sum_{j,k} \left(\int \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{z - x}{\lambda} \right) (x_j - c_{u,j}^*) \mathbf{1}_{V_u^*}(x) dx \right)^2 \leq C \Pi_{i=1}^d \lambda_i^{-2\beta_i}.$$

Hence it is possible to apply a chaining argument as in Levrard (2012) to the class

$$\mathcal{F} = \{ R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, \cdot), \mathbf{c}^* \in \mathcal{M}, \mathbf{c} \in \mathbb{R}^{kd} : \|\mathbf{c} - \mathbf{c}^*\| \leq \sqrt{\delta} \},$$

which has an enveloppe function $F(\cdot) \leq C(\Pi_{i=1}^d \lambda_i^{-\beta_i} + \|\cdot\|) \in L_2(\tilde{P})$ provided that $\mathbb{E}\|\epsilon\|^2 < \infty$. We arrive at the conclusion. ■

References

R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13 (34):1000–1034, 2008.

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

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

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

G. Biau and A. Fisher. Parameter selection for principal curves. *IEEE Transactions on Information Theory*, 58, 2012.

G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54 (2), 2008.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36 (2):489–531, 2008.

O. Bousquet. A bennet concentration inequality and its application to suprema of empirical processes. *C.R. Acad. SCI. Paris Ser. I Math*, 334:495–500, 2002.

C. Butucea. goodness-of-fit testing and quadratic functionnal estimation from indirect observations. *The Annals of Statistics*, 35:1907–1930, 2007.

F. Comte and C. Lacour. Anisotropic adaptive kernel deconvolution. to appear in Annales de l'Institut Henri Poincar, 2012.

J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19:1257–1272, 1991.

Siegfried Graf and Harald Luschgy. *Foundation of quantization for probability distributions*. Springer-Verlag, 2000. Lecture Notes in Mathematics, volume 1730.

J.A. Hartigan. *Clustering algorithms*. Wiley, 1975.

B. Kégl, A. Krzyzak, T. Linder, and Zeger K. Learning and design of principal curves. *IEEE Tansactions on Pattern Analysis and Machine Intelligence*, 22:282–297, 2000.

V. Koltchinskii. Local rademacher complexities and oracle inequalties in risk minimization. *The Annals of Statistics*, 34 (6):2593–2656, 2006.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–459. E. Gin, D. Mason and J. Wellner, eds., 2000.

G. Lecué and S. Mendelson. General non-exact oracle inequalities for classes with a subexponential envelope. *The Annals of Statistics*, 40 (2):832–860, 2012.

M. Ledoux. On talagrand's deviation inequalities for product measures. *ESAIM, Probability and Statistics*, 1:63–87, 1996.

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

T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40 (6), 1994.

S. Loustau. Inverse statistical learning. In revision to Electronic Journal of Statistics, 2012.

S. Loustau and C. Marteau. Minimax fast rates for discriminant analysis with errors in variables. In revision to Bernoulli, 2012.

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6):1808–1829, 1999.

P. Massart. About the constants in talagrand's inequality for empirical processes. *The Annals of Probability*, 29 (2):863–884, 2000.

P. Massart. Concentration inequalities and model selection. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics, Springer, 2007.

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

D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9 (1), 1981.

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

E. Rio. Inégalité de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119:163–175, 2000.

S. Sandilya and S.R. Kulkarni. Principal curves with bounded turn. *IEEE Tansactions on Information Theory*, 48:2789–2793, 2002.

A.B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer-Verlag, 2004a.

A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32 (1):135–166, 2004b.

S. Van De Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

A. W. van der Vaart and J. A. Weelner. *Weak convergence and Empirical Processes. With Applications to Statistics*. Springer Verlag, 1996.

V. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science, Springer, 2000.