Fast rates for noisy clustering

Sébastien Loustau

Fast rates for Noisy Clustering

Sébastien Loustau LOUSTAU@MATH.UNIV-ANGERS.FR
LAREMA
Université d'Angers
2 Boulevard Lavoisier,
49045 Angers Cedex, France

Editor: Leslie Pack Kaelbling

Abstract

The effect of errors in variables in empirical minimization is investigated. Given a loss l and a set of decision rules $\mathcal{G}$, we prove a general upper bound for an empirical minimization based on a deconvolution kernel and a noisy sample $Z_i = X_i + \epsilon_i, i = 1, \dots, n$.

We apply this general upper bound to give the rate of convergence for the expected excess risk in noisy clustering. A recent bound from Levrard (2012) proves that this rate is $\mathcal{O}(1/n)$ in the direct case, under Pollard's regularity assumptions. Here the effect of noisy measurements gives a rate of the form $\mathcal{O}(1/n^{\frac{\gamma}{\gamma+2\beta}})$, where γ is the Hölder regularity of the density of X whereas β is the degree of illposedness.

Keywords: Empirical minimization, Inverse problem, Fast rates, k-means clustering

1. Introduction

Isolate meaningfull groups from the data is an interesting topic in data analysis with applications in many fields, such as biology or social sciences. This unsupervised learning task is known as clustering (see the early work of Hartigan (1975)). Let $X_1, \dots, X_n$ denote i.i.d. random variables with unknown law P on $\mathbb{R}^d$, with density f with respect to a σ-finite measure ν. The problem of clustering is to assign to each observation a cluster over a finite number of k possible items. From statistical viewpoint, this problem can be endowed into the general and extensively studied problem of empirical minimization (see Vapnik (2000), Koltchinskii (2006)) as follows. Let us consider a class of decision rules $\mathcal{G}$ and a loss function $l : \mathcal{G} \times \mathbb{R}^d$ where $l(g, x)$ measures the loss of g at point x. We aim at choosing from the data $X_1, \dots, X_n$ a candidate $g \in \mathcal{G}$ that minimizes the risk functionnal:

$$R_l(g) = \mathbb{E}l(g, X), \tag{1}$$

where the expectation is taken over the unknown distribution P. For instance the k-means algorithm proposes as criterion for partitioning the data the within cluster sum of squares $\mathbf{c} \mapsto \min_{\mathbf{c}} \|x - c_j\|^2$, where in the sequel $\|\cdot\|$ denotes the euclidian norm and $\mathbf{c} = (c_1, \dots, c_k)$ is the set of possible clusters, with corresponding decision rule $g_{\mathbf{c}}(x) = \arg\min_j \|x - c_j\|$. The performances of a given $g \in \mathcal{G}$ is measured through its non-negative excess risk, given by:

$$R_l(g) - R_l(g^*), \tag{2}$$

where g^* is a minimizer over $\mathcal{G}$ of the risk (1).

A classical way to tackle this issue in the direct case is to consider, if there exists, the Empirical Risk Minimizer (ERM) estimator defined as:

$$\hat{g}_n = \arg\min_{g\in\mathcal{G}} R_n(g), \tag{3}$$

where $R_n(g)$ denotes the empirical risk defined as:

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} l(g, X_i) := P_n l(g).$$

In the sequel the empirical measure of the direct sample $X_1, \dots, X_n$ will be denoted as P_n. A large litterature (see Vapnik (2000) for such a generality) deals with the statistical performances of 3) in terms of the excess risk (2). The central point of these papers is to control the complexity of the set $\mathcal{G}$ thanks to VC dimension (Vapnik (1982)), entropy conditions (Van De Geer (2000)), or Rademacher complexity assumptions in Bartlett et al. (2005); Koltchinskii (2006) (see also Massart and Nédélec (2006); Blanchard et al. (2008) in supervised classification). The main probabilistic tool for this problem is the statement of uniform concentration of the empirical measure to the true measure. This can be easily seen using the so-called Vapnik's bound:

$$\begin{aligned} R_l(\hat{g}_n) - R_l(g^*) &\leq R_l(\hat{g}_n) - R_n(\hat{g}_n) + R_n(g^*) - R_l(g^*) &(4)\\ &\leq 2\sup_{g\in\mathcal{G}} |(R_n - R_l)(g)| = 2\sup_{g\in\mathcal{G}} |(P_n - P)l(g)|. &(5)\end{aligned}$$

It is important to note that (4) can be improved using a local approach (see Massart (2000)) which consists in reducing the supremum to a neighborhood of g^*. We do not develop this important refinement in this introduction for the sake of concision whereas it is the main ingredient of the literature cited above. It allows to get fast rates of convergence.

In this paper the framework is essentially different since we observe a corrupted sample $Z_1, \dots, Z_n$ such that:

$$Z_i = X_i + \epsilon_i,\ i = 1, \dots n, \tag{6}$$

where the ϵ_i's are i.i.d. $\mathbb{R}^d$-random variables with density η with respect to the Lebesgue measure. As a result, from (6, the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is unobservable and standard ERM (3) is not available. Unfortunately, using the corrupted sample $Z_1, \dots, Z_n$ in standard ERM (3) seems problematic since:

$$\tilde{P}_n l(g) := \frac{1}{n}\sum_{i=1}^{n} l(g, Z_i) \longrightarrow \mathbb{E}l(g, Z) \neq R_l(g).$$

Due to the action of the convolution operator, the empirical measure of the indirect sample, defined as $\tilde{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ differs from P_n and we are faced to an ill-posed inverse problem. Note that this problem has been recently considered in Loustau and Marteau (2011) in discriminant analysis and in a more general supervised statistical learning context

in Loustau (2011). The main idea to get optimal upper bounds is to consider an empirical risk based on kernel deconvolution estimators.

In this paper, we propose to adopt a comparable strategy in unsupervised statistical learning. To this end, we propose to construct a kernel deconvolution estimator of the density f of the form:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_i - x}{\lambda}\right), \tag{7}$$

where $\mathcal{K}_\eta$ is a deconvolution kernel and λ is a regularization parameter (see Section 2 for details). Given this estimator, we construct an empirical risk by plugging (7) into the true risk $R_l(g)$ to get a so-called deconvolution empirical risk minimization given by:

$$\arg\min_{g\in\mathcal{G}} R_n^\lambda(g) \text{ where } R_n^\lambda(g) := \frac{1}{n}\sum_{i=1}^{n} l_\lambda(g, Z_i), \tag{8}$$

whereas $l_\lambda(g,z)$ is a convolution of the loss function $l(g,\cdot)$ given by:

$$l_\lambda(g,z) = \int \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right) l(g,x)\nu(dx).$$

Note that in case no such minimum exists, we can consider δ-approximate minimizers as in Bartlett and Mendelson (2006).

In order to study the performances of a solution of (8, it is possible to use the empirical process machinery in the spirit of Koltchinskii (2006); Bartlett and Mendelson (2006); Blanchard et al. (2008). In the presence of indirect observations, for $\hat{g}_n^\lambda$ a solution of the minimization of (8, we have:

$$\begin{aligned}
R_l(\hat{g}_n^\lambda) - R_l(g^*) &\leq R_l(\hat{g}_n^\lambda) - R_n^\lambda(\hat{g}_n^\lambda) + R_n^\lambda(g^*) - R_l(g^*) &(9)\\
&\leq R_l^\lambda(\hat{g}_n^\lambda) - R_n^\lambda(\hat{g}_n^\lambda) + R_n^\lambda(g^*) - R_l^\lambda(g^*) + (R_l - R_l^\lambda)(\hat{g}_n^\lambda - g^*) &(10)\\
&\leq \sup_{g\in\mathcal{G}} |(R_n^\lambda - R_l^\lambda)(g^* - g)| + \sup_{g\in\mathcal{G}} |(R_l^\lambda - R_l(g - g^*)|, &(11)
\end{aligned}$$

where in the sequel, under integrability conditions and using Fubini:

$$R_l^\lambda(g) = \mathbb{E}R_n^\lambda(g) = \int l(g,x)\mathbb{E}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z-x}{\lambda}\right)\nu(dx). \tag{12}$$

Bounds (9) are comparable to (4) for the direct case. There consist in two terms:

- A variance term $\sup_{g\in\mathcal{G}} |(R_n^\lambda - R_l^\lambda)(g^* - g)|$ related to the estimation of $R_l(g)$ using an empirical couterpart. This term will be controled using standard tools from empirical process theory, namely a local approach in the spirit of Koltchinskii (2006). Here the empirical process is indexed by a class of functions depending on a smoothing parameter.

- A bias term $\sup_{g\in\mathcal{G}} |(R_l^\lambda - R_l)(g - g^*)|$ due to the estimation procedure using kernel deconvolution estimator. It seems to be related to the usual bias term in nonparametric density deconvolution since we can see coarselly that:

$$R_l^\lambda(g) - R_l(g) = \int l(g,x)\left[\mathbb{E}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z-x}{\lambda}\right) - f(x)\right]\nu(dx).$$

The choice of λ is crucial in the decomposition (9. We will show below that the variance term grows when λ tends to zero whereas the bias term vanishes. Parameter λ has to be chosen as a trade-off between these two terms, and as a consequence will depend on unknown parameters. The problem of adaptation is not adressed in this paper but is an interesting future direction.

The paper is organized as follows. In Section 2, we propose to give a general upper bound for (8, generalizing the results of Koltchinskii (2006) to indirect observations. Note that all the material of Section 2 is largely inspired from Loustau (2011) and gives an unsupervised counterpart of the previous results. Section 3 gives a direct application of the result of Section 2 in clustering by giving rates of convergence for a new deconvolution k-means algorithm. Fast rates of convergence are proposed which generalize the recent fast rates proposed in Levrard (2012) in the direct case.

2. General Upper bound

In this section we propose an upper bound for the expected excess risk of the estimator:

$$\hat{g}_n^\lambda := \arg\min \frac{1}{n}\sum_{i=1}^n l_\lambda(g, Z_i), \tag{13}$$

where $l_\lambda(g,z)$ is construct 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 $\mathcal{K}_j$. Then if we denote by $\lambda = (\lambda_1, \dots, \lambda_d)$ a set of (positive) bandwidths and by $\mathcal{F}[\cdot]$ the Fourier transform, we define $\mathcal{K}_\eta$ as:

$$\begin{aligned} \mathcal{K}_\eta \quad &: \quad \mathbb{R}^d \to \mathbb{R} \\ &\qquad t \mapsto \mathcal{K}_\eta(t) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)}\right](t). \end{aligned} \tag{14}$$

Moreover in the sequel we restrict the study to a compact set $K \subset \mathbb{R}^d$ and define $l_\lambda(g,z)$ as

$$l_\lambda(g,z) = \int_K \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right) l(g,x)\nu(dx),$$

where we write with a slight abuse of notation $\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)$ for $\frac{1}{\Pi_{i=1}^d \lambda_i}\mathcal{K}_\eta\left(\frac{z_1-x_1}{\lambda_1}, \dots, \frac{z_d-x_d}{\lambda_d}\right)$. The restriction to a compact K allows to control the variance in decomposition (9) thanks to Lemma 1 (we refer the reader to Loustau (2011) for a discussion).

Finally in the sequel for the sake of simplicity we restrict ourselves to moderately ill-posed inverse problem and introduce the following assumption:

Noise Assumption: There exist $(\beta_1, \dots, \beta_d)' \in \mathbb{R}_+^d$ such that for all $i \in \{1, \dots, d\}$,

$$|\mathcal{F}[\eta_i](t)| \sim |t|^{-\beta_i}, \text{as } t \to +\infty.$$

Moreover, we assume that $\mathcal{F}[\eta_i](t) \neq 0$ for all $t \in \mathbb{R}$ and $i \in \{1, \dots, d\}$.

Assumption **(NA)** deals with the asymptotic behavior of the characteristic function of the noise distribution. These kind of restrictions are standard in deconvolution problems (see Fan (1991); Meister (2009); Butucea (2007)). Note that straightforward modifications allow to consider severely ill-posed inverse problems, where the asymptotic behavior of the characteristic function of ϵ decreases exponentially to zero.

Under **(NA)**, the goal is to control the two terms of (9, namely the bias term and the variance term. The variance term is reduced to the study of the increments of the empirical process:

$$\nu_n^\lambda(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\lambda(g, Z_i) - \mathbb{E} l_\lambda(g, Z).$$

It will be controled thanks to a version of Talagrand's inequality due to Bousquet (2002). However it is important to note that here the empirical process is indexed by the class of functions $\{z \mapsto l_\lambda(g, z), g \in \mathcal{G}\}$, which depends on a regularization parameter $\lambda \in \mathbb{R}_+^d$. This parameter will be calibrated as a function of n so Talagrand's type inequality has to be used in a careful way. For this purpose, we need the following lemma.

Lemma 1 *Suppose $f \geq c_0 > 0$ on K. Then if **(NA)** holds, and $\mathcal{K}$ has compactly supported Fourier transform, we have:*

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

$$\exists C_1 > 0 : \forall g, g' \in \mathcal{G},\ \|l_\lambda(g) - l_\lambda(g')\|_{L_2(\tilde{P})} \leq C_1 \Pi_{i=1}^d \lambda_i^{-\beta_i} \|l(g) - l(g')\|_{L_2(P)}.$$

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

$$\exists C_2 > 0 : \sup_{g \in \mathcal{G}} \|l_\lambda(g, \cdot)\|_\infty \leq C_2 \Pi_{i=1}^d \lambda_i^{-\beta_i - 1/2}.$$

The proof of this result is presented in Loustau (2011) in a slightly different framework. Note that the assumption on f to be strictly positive on K appears for some technical reasons in the proof and could be avoided in some cases (see the discussion in Loustau (2011)).
The Lipschitz property **(i)** is a key ingredient to control the complexity of the class of functions $\{l_\lambda(g), g \in \mathcal{G}\}$ thanks to standard complexity arguments applied to the loss class $\{l(g), g \in \mathcal{G}\}$. Finally **(ii)** is necessary to apply Talagrand's type inequality to the empirical process $g \mapsto \nu_n^\lambda(g)$ above.

To control the excess risk of the procedure, we also need to control the bias term defined in (9) thanks to Lemma 2) below.

Lemma 2 *Suppose $f \in \Sigma(\gamma, L)$ the Hölder class of $\lfloor \gamma \rfloor$-fold continuously differentiable functions on R^d satisfying the Hölder condition. Let $\mathcal{K}$ a kernel of order $\lfloor \gamma \rfloor$ with respect to ν. Then if $l(g, \cdot) \in L_1(\nu, \mathbb{R}^d)$, we have:*

$$\forall g, g' \in \mathcal{G},\ \left| (R_l - R_l^\lambda)(g - g') \right| \leq C \sum_{i=1}^{d} \lambda_i^\gamma.$$

The proof is presented in Loustau (2011) and is omitted. Finally to state fast rates, we also require an additional assumption over the distribution P.

Definition 2.1 *We say that $\mathcal{F}$ is a Bernstein class with respect to P if there exists $\kappa_0 \geq 0$ such that for every $f \in \mathcal{F}$:*

$$\|f\|^2_{L_2(P)} \leq \kappa_0[\mathbb{E}_P f].$$

This notion of Bernstein class first appears in Bartlett and Mendelson (2006) in a more general form. Definition 2.1 corresponds to the ideal case where $\kappa = 1$. This assumption can be related to the well-known margin assumption in supervised classification, introduced in Mammen and Tsybakov (1999). Section 3 proposes an unsupervised version of this hypothesis.
From technical viewpoint, this requirement arises naturally in the proof when we want to apply functional Bernstein's inequality such as Talagrand's inequality. If we consider the loss class $\mathcal{F} = \{l(g) - l(g^*), g \in \mathcal{G}\}$, Definition 2.1 gives a perfect variance-risk correspondance.

We are now on time to state the main result of this section.

Theorem 3 *Suppose **(NA)** holds and assumptions of Lemma 1-2 hold. Suppose $\{l(g) - l(g^*), g \in \mathcal{G}\}$ is Bernstein w.r.t. P where $g^* \in \arg\min_{\mathcal{G}} R_l(g)$ is unique and there exists $0 < \rho < 1$ such that for every $\delta > 0$:*

$$\mathbb{E} \sup_{g,g' \in \mathcal{G}(\delta)} \left| (\tilde{P} - \tilde{P}_n)(l_\lambda(g) - l_\lambda(g')) \right| \lesssim \frac{\Pi_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n}} \delta^{\frac{1-\rho}{2}}, \tag{15}$$

where $\mathcal{G}(\delta) = \{g \in \mathcal{G} : R_l(g) - R_l(g^*) \leq \delta\}$.
Then estimator $\hat{g} = \hat{g}^\lambda_n$ defined in (13) is such that:

$$\mathbb{E}R_l(\hat{g}) - R_l(g^*) \leq C n^{-\frac{\gamma}{\gamma(1+\rho)+2\beta}},$$

where $\bar{\beta} = \sum_{i=1}^d \beta_i$ and $\lambda = (\lambda_1, \ldots, \lambda_d)$ is chosen as:

$$\lambda_i \sim n^{-\frac{1}{\gamma(1+\rho)+2\beta}}, \forall i = 1, \ldots d.$$

.

The proof of this result iterates a version of Talagrand's inequality due to Bousquet (2002). It is presented in Section 5. coarsely speaking, Lemma 1, gathering with the complexity assumption (15, leads to a control of the variance term in decomposition (9. Then Lemma 2 gives the order of the bias term. The choice of λ explicited in Theorem 3 trades off these two terms, and gives the excess risk bound.

Note that the rates of convergence in Theorem 3 generalize previous results. When $\epsilon = 0$ Theorem 3 gives fast rates of convergence between $\mathcal{O}(1/\sqrt{n})$ to $\mathcal{O}(1/n)$ depending on the complexity parameter $\rho > 0$ in (15, which can be related with entropy or Rademacher complexities of the hypothesis set $\mathcal{G}$. The effect of the inverse problem depends on the asymptotic behavior of the characteristic function of the noise distribution ϵ. This is rather standard in the statistical inverse problem literature (Fan (1991) or Meister (2009)).

Moreover the control of the modulus of continuity in (15) is specific to the indirect framework and depends on the smoothing parameter λ. A comparable hypothesis arises in the direct case in Koltchinskii (2006), up to the constant depending on λ. It appears that it will be satisfied in our application using standard statistical learning argues, such as maximal inequalities and chaining.

Finally note that the complexity parameter involved in assumption (15) is smaller than the complexity proposed in Koltchinskii (2006) or Loustau (2011). Here the supremum is taken over the set $\{g, g' \in \mathcal{G}(\delta)\} \subset \{g, g' \in \mathcal{G} : P(l(g) - l(g'))^2 \leq c\delta\}$ provided that $\{l(g) - l(g^*), g \in \mathcal{G}\}$ is Bernstein with respect to P according to Definition 2.1. This indexing set is related to the localization's technique used in Theorem 3, namely a localization based on the excess risk instead of the $L_2(P)$-norm. This refinement is necessary to derive fast rates in Section 3 (see Bartlett and Mendelson (2006) for a related discussion).

3. Application to noisy clustering

Clustering is a basic problem in statistical learning where independent random variables $X_1, \dots, X_n$ are observed, with common source distribution P. The aim is to construct clusters to classify these data. However in many real-life situations, direct data $X_1, \dots, X_n$ are not available and measurement errors occur. Then we observe a corrupted sample $Z_i = X_i + \epsilon_i, i = 1, \dots n$ with unknown noisy distribution $\tilde{P}$. The problem of noisy clustering is to learn clusters for the direct dataset $X_1, \dots, X_n$ when only a contaminated version $Z_1, \dots, Z_n$ is observed.

To frame the noisy clustering problem as a statistical learning one, we first introduce the following notation. Let $\mathbf{c} = (c_1, \dots, c_k) \in \mathcal{C}$ the set of possible clusters, where $\mathcal{C} \subseteq \mathbb{R}^{dk}$ and $X \in \mathbb{R}^d$. The loss function $\gamma : \mathbb{R}^{dk} \times \mathbb{R}^d$ is defined as:

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

and the corresponding true risk or clustering risk is $R(\mathbf{c}) = \mathbb{E}\gamma(\mathbf{c}, X)$. 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 $Z_i = X_i + \epsilon_i$, $i = 1, \dots, n$ where the ϵ_i's are i.i.d. with density η satisfying **(NA)**. For this purpose, we introduce the following deconvolution empirical minimization:

$$\arg\min_{\mathbf{c}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \gamma_\lambda(\mathbf{c}, Z_i), \tag{16}$$

where $\gamma_\lambda(\mathbf{c}, z)$ is a deconvolution cluster sum of squares defined as:

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

for $\mathcal{K}_\eta$ the deconvolution kernel of Section 2 and $\lambda = (\lambda_1, \dots, \lambda_d) \in \mathbb{R}_+^d$ a set of positive bandwidths chosen later on. Note that here the existence of a minimizer in (16) could be managed as in Graf and Luschgy (2000) for the direct case. We investigate the generalization ability of the solution of (16) in the context of Pollard's regularity assumptions, thanks to the noisy empirical minimization results of Section 2. To this end, we will use the following assumptions on the source distribution P.

Assumption 1 (Boundedness assumption): The distribution P is such that:

$$P(\mathcal{B}(0, M)) = 1,$$

where $\mathcal{B}(0, M)$ denote the closed ball of radius M, with $M \geq 0$.

Note that **(A1)** imposes a boundedness condition on the random variable X. We will also need the following regularity requirement, first introduced in Pollard (1982).

Assumption 2 (Pollard's regularity condition): 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 $\mathcal{B}(0, M)$, **(A1)-(A2)** 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 the paper.
Moreover, Pollard's conditions can be interpreted as follows. Denote ∂V_i the boundary of the Voronoi cell V_i associated with c_i, for $i = 1, \dots, k$. Then Levrard (2012) has shown that a sufficient condition to have **(A2)** 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, **(A2)** 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, Pollard's regularity conditions can be related to the margin assumption in binary classification (see Tsybakov (2004)). We have in fact the following lemma due to Antos et al. (2005).

Lemma 4 (Antos et al. (2005)) *Suppose **(A1)-(A2)** are satisfied. Then, for any* $\mathbf{c} \in \mathcal{B}(0, M)$*:*

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

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

Lemma 4 is useful to derive fast rates of convergence for two reasons.

Firstly, if we compil these two inequalities, we get a control of the variance of the excess loss $\gamma(\mathbf{c}) - \gamma(\mathbf{c}^*(\mathbf{c}))$ thanks to the excess clustering risk $R(\mathbf{c}) - R(\mathbf{c}^*(\mathbf{c}))$. Note that if $R(\mathbf{c}) - R(\mathbf{c}^*(\mathbf{c})) \leq 1$, it is clear that the loss class $\{\gamma(\mathbf{c}) - \gamma(\mathbf{c}^*(\mathbf{c})), c \in \mathcal{C}\}$ is Bernstein according to Definition 2.1 since we have coarsely:

$$P(\gamma(\mathbf{c},\cdot)-\gamma(\mathbf{c}^*(\mathbf{c}),\cdot))^2 \leq \mathrm{var}\left(\gamma(\mathbf{c},\cdot) - \gamma(\mathbf{c}^*(\mathbf{c}),\cdot)\right)+(R(\mathbf{c})-R(\mathbf{c}^*(\mathbf{c})))^2 \leq (C_1C_2+1)\left(R(\mathbf{c}) - R(\mathbf{c}^*(\mathbf{c}))\right).$$

Moreover the second inequality of Lemma 4 is necessary to control the complexity involved in Section 2 thanks to the following lemma:

Lemma 5 *Suppose **(A1)-(A2)** are satisfied. Suppose $\mathbb{E}\|\epsilon\|^2 < \infty$. Then:*

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

where C is a positive constant depending on $M, k, d, \tilde{P}, K, \eta$.

Note that $\mathbb{E}\|\epsilon\|^2 < \infty$ comes from Pollard (1982). Gathering with **(A1)**, it gives $\mathbb{E}\|Z\|^2 < \infty$ and allows to deal with indirect observations. The proof of Lemma 5 is presented in Section 5. It is based on Pollard (1982) extended to the noisy setting. Under **(A2)** and provided that $\mathbb{E}\|\epsilon\|^2 < \infty$, we use the following approximation of the convolution loss function $\gamma_\lambda(\cdot, x)$ at any point $\mathbf{c} \in \mathcal{C}$:

$$\gamma_\lambda(\mathbf{c}, z) = \gamma_\lambda(\mathbf{c}^*, z) + \langle \mathbf{c} - \mathbf{c}^*, \nabla_{\mathbf{c}}\gamma_\lambda(\mathbf{c}^*, z)\rangle + \|\mathbf{c} - \mathbf{c}^*\| R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z), \tag{17}$$

where $\nabla_{\mathbf{c}}\gamma_\lambda(\mathbf{c}^*, z)$ is the gradient of $\mathbf{c} \mapsto \gamma_\lambda(\mathbf{c}, z)$ at point $\mathbf{c}^*$ and $R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, z)$ is a residual term (see Pollard (1982) for details). With (17, the complexity term is controled with a maximal inequality due to Massart (2007), gathering with a chaining method.

We are now on time to state the main result of this section.

Theorem 6 *Assume **(NA)** holds, P satisfies **(A1)-(A2)** with density $f \in \Sigma(\gamma, L)$ and $\mathbb{E}\|\epsilon\|^2 < \infty$. Then, denoting by $\hat{\mathbf{c}}_n^\lambda$ a solution of (16, we have, for any $c^* \in \mathcal{M}$:*

$$\mathbb{E}R(\hat{\mathbf{c}}_n^\lambda) - R(\mathbf{c}^*) \leq Cn^{-\frac{\gamma}{\gamma+2\beta}},$$

where $\bar{\beta} = \sum_{i=1}^d \beta_i$, C is a positive constant whereas $\lambda = (\lambda_1, \ldots, \lambda_d)$ is chosen as:

$$\lambda_i \sim n^{-\frac{1}{\gamma+2\bar{\beta}}}, \forall i = 1, \ldots d.$$

The proof is a direct application of Section 2 when $|\mathcal{M}| = 1$ whereas when $|\mathcal{M}| \geq 2$, a more sophisticated geometry has to be considered (see Section 5 for details). Some remarks are in order.

Rates of convergence of Theorem 6 are fast rates when $2\bar{\beta} < \gamma$. 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 price to pay for the inverse problem is the quantity $2\sum_{i=1}^d \beta_i$, related to the tail behavior of the characteristic function of the noise distribution η in **(NA)**. This rate corresponds to the ideal case where $\rho = 0$ in Section 2,

due to the finite-dimensional structure of the set of clusters $\mathcal{C} = \{\mathbf{c} = (c_1, \dots, c_k), c_i \in \mathbb{R}^d\}$. An interesting extension is to consider richer classes such as kernel classes (see Mendelson (2003)) and to deal with kernel k-means.
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 optimality of Theorem 6, even in the direct case where $\epsilon = 0$. 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 (2011) proposes lower bounds in a supervised framework under both margin assumption and **(NA)**.

4. Conclusion

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

In the problem of empirical minimization with errors-in-variables, we provide the order of the expected excess risk, depending on the complexity of the hypothesis space, the regularity of the direct observations and the degree of ill-posedness. For the sake of concision, Theorem 3 only consider particular Bernstein classes and empirical minimization based on a deconvolution kernel estimator. A higher level of generality can be derived from Loustau (2011) but is out of the scope of the present paper.

The performances of our deconvolution k-means algorithm is obtained thanks to a localization principle due to Koltchinskii (2006), where proofs iterate a Talagrand's inequality due to Bousquet (2002). With such a study, Koltchinskii (2006) provides the order of the excess risk in the direct case and allows to recover most of the recent results in the statistical learning context. There is nice hope that many statistical learning problem when dealing with indirect observations could be solved with similar argues.

In the problem of noisy k-means clustering, we propose fast rates of convergence to the order of $\mathcal{O}(1/n^{\frac{\gamma}{\gamma+2\beta}})$. Theorem 6 is a direct application of the result of Theorem 3 to the problem of clustering with Pollard's regularity assumptions and bounded source. It generalizes a recent result of Levrard (2012) where fast rates are stated for direct observations.

From practical viewpoint, this work proposes an empirical minimization to deal with the problem of noisy clustering. However the procedure (16) is not adaptive in many sense. Of course the dependency on the noise distribution η can be explored in a future work, and can be associated with the problem of unknown operator in the statistical inverse problem literature (see Marteau (2006) or Cavalier and Hengartner (2005)). Moreover the empirical minimization depends on the Hölder regularity of the density of the source distribution through the choice of the bandwidths $\lambda_i, i = 1, \dots d$. However for practical experiments, any data-dependent model selection procedure can be performed, such as cross-validation.

5. Proofs

In all the proofs constant $C > 0$ may vary from line to line.

5.1 Proof of Theorem 3

The proof uses the following intermediate lemma.

Lemma 7 *Suppose* $\{l_\lambda(g), g \in \mathcal{G}\}$ *is such that* $\sup \|l_\lambda(g)\|_\infty \leq K(\lambda)$. *Define:*

$$U_n^\lambda(\delta_j, t) := K\left[\phi_n^\lambda(\mathcal{G}, \delta_j) + \sqrt{\frac{t}{n}} D^\lambda(\mathcal{G}, \delta_j) + \sqrt{\frac{t}{n}(1 + K(\lambda))\phi_n^\lambda(\mathcal{G}, \delta_j)} + \frac{t}{n}\right],$$

$$\phi_n^\lambda(\mathcal{G}, \delta_j) := \mathbb{E} \sup_{g,g' \in \mathcal{G}(\delta_j)} |\tilde{P}_n - \tilde{P}|[l_\lambda(g) - l_\lambda(g')],$$

$$D^\lambda(\mathcal{G}, \delta_j) := \sup_{g,g' \in \mathcal{G}(\delta_j)} \sqrt{\tilde{P}(l_\lambda(g) - l_\lambda(g'))^2},$$

where $\delta_j = q^{-j}, j \in \mathbb{N}^*$, *for some* $q > 0$.
Then $\forall \delta \geq \delta_n^\lambda(t)$, *we have for* $\hat{g} = \hat{g}_n^\lambda$ *defined in (13:*

$$\mathbb{P}(R_l(\hat{g}) - R_l(g^*) \geq \delta) \leq c(\delta, q)e^{-t},$$

where:

$$\delta_n^\lambda(t) = \left(\inf\left\{\delta > 0 : \sup_{\delta_j \geq \delta} \frac{U_n(\delta_j, t)}{\delta_j} \leq \frac{1}{2q}\right\}\right) \vee \left(4q \sup_{g,g' \in \mathcal{G}} (R_l - R_l^\lambda)(g - g')\right).$$

The proof is a straightforward modification of the proof of Lemma 2 in Loustau (2011).
Proof [of Theorem 3] First note that, in dimension $d = 1$ for simplicity:

$$U_n^\lambda(\delta, t) \leq C\left(\phi_n^\lambda(\mathcal{G}, \delta) + \sqrt{\frac{t}{n}\phi_n^\lambda(\mathcal{G}, \delta)(1 + \lambda^{-\beta-1/2})} + \sqrt{\frac{t}{n}} D^\lambda(\mathcal{G}, \delta) + \frac{t}{n}\right).$$

Using Definition 2.1, gathering with the complexity assumption over $\tilde{\omega}_n(\mathcal{G}, \delta)$, we have:

$$\phi_n^\lambda(\mathcal{G}, \delta) \leq \mathbb{E} \sup_{g,g' \in \mathcal{G}(\delta)} |\tilde{P}_n - \tilde{P}|[l_\lambda(g) - l_\lambda(g')] \leq C\frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1-\rho}{2}}.$$

A control of $D^\lambda(\mathcal{G}, \delta)$ using Lemma 1, gathering with Definition 2.1 leads to:

$$U_n^\lambda(\delta, t) \leq C\left(\frac{\lambda^{-\beta}}{\sqrt{n}} \delta^{\frac{1-\rho}{2}} + \frac{\lambda^{-\beta/2}}{n^{3/4}} \delta^{\frac{1-\rho}{4}} \sqrt{\lambda^{-\beta-1/2} t} + \sqrt{\frac{t}{n}} \lambda^{-\beta} \delta^{\frac{1}{2}} + \frac{t}{n}\right).$$

We hence have from an easy calculation:

$$\delta_n^\lambda(t) \leq C\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2}{1+\rho}}.$$

To get the result we apply Lemma 7 with:

$$\delta = C(1+t)\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2}{1+\rho}},$$

noting that the choice of λ warrants that:

$$\lambda^{\gamma} \leq C(1+t)\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\right)^{\frac{2}{1+\rho}}.$$

Same arguments conclude the proof for $d \geq 2$. ■

5.2 Proof of Lemma 5

The proof follows Levrard (2012) applied to the noisy setting. First note that, by smoothness assumptions over $\mathbf{c} \mapsto \min \|x - c_j\|$, we get, for any $\mathbf{c} \in (\mathbb{R}^d)^k$ and $\mathbf{c}^* \in \mathcal{M}$,

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

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

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

Splitting the expectation in two parts, we obtain:

$$\begin{aligned}\mathbb{E}\sup_{\mathbf{c}^* \in \mathcal{M}, \|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta} |\tilde{P}_n - \tilde{P}|(\gamma_\lambda(\mathbf{c}^*, .) - \gamma_\lambda(\mathbf{c}, .)) &\leq \mathbb{E}\sup_{\mathbf{c}^* \in \mathcal{M}, \|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta} |\tilde{P}_n - \tilde{P}|\left\langle \mathbf{c}^* - \mathbf{c}, \nabla_{\mathbf{c}} \gamma_\lambda(\mathbf{c}^*, .)\right\rangle \\ + \quad \sqrt{\delta}\mathbb{E}\sup_{\mathbf{c}^* \in \mathcal{M}, \|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta} |\tilde{P}_n - \tilde{P}|(-R_\lambda(\mathbf{c}^*, \mathbf{c} - \mathbf{c}^*, .)) &\qquad (18)\end{aligned}$$

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

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

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

$$\mathrm{var}Z_n \quad = \quad \frac{4}{n}\sum_{u=1}^{k}\sum_{j=1}^{d}(c_{u,j} - c_{u,j}^*)^2 \mathrm{var}\int_{V_u}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z - x}{\lambda}\right)(x_j - c_{u,j})dx$$

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

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 arguments originally stated in Loustau and Marteau (2011) for compactly supported $\mathcal{F}[\mathcal{K}]$. We hence have using for instance a maximal inequality due to Massart Massart (2007, Part 6.1):

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

We obtain for the first term in (18 the right order. To prove that the second term in (18) is smaller, note that from Pollard (1982), we have:

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

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

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

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

5.3 Proof of Theorem 6

Thr proof of Theorem 6 is divided into two steps. Using Theorem 3, we can bound the excess risk when $|\mathcal{M}|=1$. For the general case of a finite numbers of optimal clusters $|\mathcal{M}|\geq 2$, we need to introduce a more sophisticated localization explain in Koltchinskii (2006, Section 4).

First case: $|\mathcal{M}| = 1$**.**

The proof follows the proof of Theorem 3. Using the previous notations, we have:

$$U_n^\lambda(\delta, t) \leq C\left(\phi_n^\lambda(\mathcal{C}, \delta) + \sqrt{\frac{t}{n}\phi_n^\lambda(\mathcal{C}, \delta)(1 + K(\lambda))} + \sqrt{\frac{t}{n}}D^\lambda(\mathcal{C}, \delta) + \frac{t}{n}\right).$$

Using Lemma 4, gathering with Lemma 5, we have for $d = 1$ for simplicity:

$$\begin{aligned}
\phi_n^\lambda(\mathcal{C}, \delta) &\leq \mathbb{E}\sup_{\mathbf{c},\mathbf{c}' \in \mathcal{C}(\delta)} |\tilde{P}_n - \tilde{P}|[\gamma_\lambda(\mathbf{c}) - \gamma_\lambda(\mathbf{c}')] \\
&\leq \mathbb{E}\sup_{\|\mathbf{c}-\mathbf{c}^*\|^2 \leq \delta} |\tilde{P}_n - \tilde{P}|[\gamma_\lambda(\mathbf{c}) - \gamma_\lambda(\mathbf{c}^*)] + \mathbb{E}\sup_{\|\mathbf{c}'-\mathbf{c}^*\|^2 \leq \delta} |\tilde{P}_n - \tilde{P}|[\gamma_\lambda(\mathbf{c}') - \gamma_\lambda(\mathbf{c}^*)] \\
&\leq C\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1}{2}},
\end{aligned}$$

where $\mathbf{c}^*$ is the unique minimizer of the clustering risk. Moreover, by uniqueness of $\mathbf{c}^*$, we can write from Lemma 4:

$$\begin{aligned}
D^\lambda(\mathcal{G}, \delta) &:= \sup_{\mathbf{c},\mathbf{c}' \in \mathcal{C}(\delta)} \sqrt{\tilde{P}(\gamma_\lambda(\mathbf{c}) - \gamma_\lambda(\mathbf{c}'))^2} \\
&\leq C_1\lambda^{-\beta}\sup_{\mathbf{c},\mathbf{c}' \in \mathcal{C}(\delta)} \sqrt{P(\gamma(\mathbf{c}) - \gamma(\mathbf{c}'))^2} \\
&\leq C_1\lambda^{-\beta}\sup_{\mathbf{c} \in \mathcal{C}(\delta)} \sqrt{P(\gamma(\mathbf{c}) - \gamma(\mathbf{c}^*))^2} + \sup_{\mathbf{c}' \in \mathcal{C}(\delta)} \sqrt{P(\gamma(\mathbf{c}') - \gamma(\mathbf{c}^*))^2} \\
&\leq 2C_1\lambda^{-\beta}\sqrt{\delta}.
\end{aligned}$$

It follows:

$$U_n^\lambda(\delta, t) \leq C\left(\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1}{2}} + \frac{\lambda^{-\beta/2}}{n^{3/4}}\delta^{\frac{1}{4}}\sqrt{\lambda^{-\beta-1/2}t} + \sqrt{\frac{t}{n}}\lambda^{-\beta}\delta^{\frac{1}{2}} + \frac{t}{n}\right).$$

We hence have the result applying Lemma 7 with the choice of λ precised in Theorem 6.

Second case: $|\mathcal{M}| \geq 2$

When the infimum is not unique, the diameter $D^\lambda(\mathcal{G}, \delta)$ does not necessary tend to zero when $\delta \to 0$. We hence introduce the more sophisticated geometric characteristic $r(\sigma, \delta)$ from Koltchinskii (2006) defined as:

$$r(\sigma, \delta) = \sup_{\mathbf{c} \in \mathcal{C}(\delta)} \inf_{\mathbf{c}' \in \mathcal{C}(\sigma)} \sqrt{\tilde{P}(\gamma_\lambda(\mathbf{c}) - \gamma_\lambda(\mathbf{c}'))^2},\ 0 < \sigma \leq \delta.$$

It is clear that $r(\sigma, \delta) \leq D^\lambda(\mathcal{G}, \delta)$ and for $\delta \to 0$, we have $r(\sigma, \delta) \to 0$. The idea of the proof of Theorem 6 is to use a modified version of Theorem 3 using $r(\sigma, \delta)$ instead of $D^\lambda(\mathcal{G}, \delta)$. Following Koltchinskii (2006, Theorem 4), we can use a modified version of Lemma 7 in order to guarantee the upper bounds of Theorem 3 when $|\mathcal{M}| \geq 2$. To this end, we have to check for $d = 1$ for simplicity:

$$\lim_{\epsilon \to 0} \mathbb{E}\sup_{g \in \mathcal{G}(\sigma)} \sup_{g' \in \mathcal{G}(\delta): P(l(g) - l(g'))^2 \leq r(\sigma,\delta) + \epsilon} \left|(\tilde{P} - \tilde{P}_n)(l_\lambda(g) - l_\lambda(g'))\right| \leq C\frac{\lambda^{-\beta}}{\sqrt{n}}\delta^{\frac{1-\rho}{2}}, \qquad (19)$$

and

$$r(\sigma,\delta)\sqrt{\frac{t}{n}} \leq C\lambda^{-\beta}\sqrt{\frac{t\delta}{n}}. \tag{20}$$

Note that from Lemma 4 and Lemma 5, it is clear that (19) holds since:

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}(\sigma)}\sup_{\mathbf{c}'\in\mathcal{C}(\delta):P(\gamma(\mathbf{c})-\gamma(\mathbf{c}'))^2\leq r(\sigma,\delta)+\epsilon}\left|(\tilde{P}-\tilde{P}_n)(\gamma_\lambda(g)-\gamma_\lambda(g'))\right| \\
&\leq \mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}(\sigma),c^*\in\mathcal{M}}\left|(\tilde{P}-\tilde{P}_n)(\gamma_\lambda(\mathbf{c})-\gamma_\lambda(\mathbf{c}^*))\right| + \mathbb{E}\sup_{\mathbf{c}'\in\mathcal{C}(\delta)}\left|(\tilde{P}-\tilde{P}_n)(\gamma_\lambda(\mathbf{c}')-\gamma_\lambda(\mathbf{c}^*(\mathbf{c}')))\right| \\
&\leq 2\mathbb{E}\sup_{(\mathbf{c},\mathbf{c}^*)\in\mathcal{C}\times\mathcal{M},\|\mathbf{c}-\mathbf{c}^*\|^2\leq c\delta}\left|(\tilde{P}_n-\tilde{P})(\gamma_\lambda(\mathbf{c}^*)-\gamma_\lambda(\mathbf{c}))\right| \\
&\leq C\lambda^{-\beta}\frac{\sqrt{\delta}}{\sqrt{n}}.
\end{aligned}$$

Finally (20) holds since we have with Lemma 4, $\forall \mathbf{c}\in\mathcal{C}(\delta), \mathbf{c}'\in\mathcal{C}(\sigma)$:

$$\begin{aligned}
\sqrt{\tilde{P}(\gamma_\lambda(\mathbf{c})-\gamma_\lambda(\mathbf{c}'))^2} &\leq C\lambda^{-\beta}\sqrt{P(\gamma(\mathbf{c})-\gamma(\mathbf{c}'))^2} \\
&\leq C\lambda^{-\beta}\left(\sqrt{P(\gamma(\mathbf{c})-\gamma(\mathbf{c}^*(\mathbf{c}))^2}+\sqrt{P(\gamma(\mathbf{c}')-\gamma(\mathbf{c}^*(\mathbf{c})))^2}\right) \\
&\leq C\lambda^{-\beta}\sqrt{\delta},
\end{aligned}$$

provided that $\sigma\leq\delta\leq 1$.

References

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.

P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005.

G. Biau, L. Devroye, and G. Lugosi. On the performances of clustering in hilbert spaces. *IEEE Trans. Inform. 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.

L. Cavalier and N.W. Hengartner. Adaptative estimation for inverse problems with noisy operators. *Inverse Problems*, 21 (4):1345–1361, 2005.

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.

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

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. Statistical learning with indirect observations. *http://hal.archives-ouvertes.fr/hal-00664125*, 2011.

S. Loustau and C. Marteau. Discriminant analysis with errors in variables. *http://hal.archives-ouvertes.fr/hal-00660383*, 2011.

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

C. Marteau. Regularization of inverse problems with unkown operator. *Mathematical Methods of Statistics*, 15 (4):415–443, 2006.

P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9 (2):245–303, 2000.

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

P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5):2326–2366, 2006.

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

S. Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

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.

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

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

V. Vapnik. *Estimation of Dependances Based on Empirical Data*. Springer Verlag, 1982.

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