Electronic Journal of Statistics
Vol. 7 (2013) 2065–2097
ISSN: 1935-7524
DOI: 10.1214/13-EJS838

Inverse statistical learning

Sébastien Loustau

Laboratoire Angevin de Recherche en Maths
Université d'Angers
2 Boulevard Lavoisier
49045 Angers Cedex 01
e-mail: loustau@math.univ-angers.fr

Abstract: Let $(X,Y)\in\mathcal{X}\times\mathcal{Y}$ be a random couple with unknown distribution P. Let $\mathcal{G}$ be a class of measurable functions and ℓ a loss function. The problem of statistical learning deals with the estimation of the Bayes:

$$g^* = \arg\min_{g\in\mathcal{G}} \mathbb{E}_P \ell(g,(X,Y)).$$

In this paper, we study this problem when we deal with a contaminated sample $(Z_1,Y_1),\ldots,(Z_n,Y_n)$ of i.i.d. indirect observations. Each input Z_i, $i=1,\ldots,n$ is distributed from a density Af, where A is a known compact linear operator and f is the density of the direct input X.

We derive fast rates of convergence for the excess risk of empirical risk minimizers based on regularization methods, such as deconvolution kernel density estimators or spectral cut-off. These results are comparable to the existing fast rates in Koltchinskii (2006) for the direct case. It gives some insights into the effect of indirect measurements in the presence of fast rates of convergence.

AMS 2000 subject classifications: Primary 62G05; secondary 62H30.
Keywords and phrases: Statistical learning, inverse problem, classification, deconvolution, fast rates.

Received June 2012.

Contents

1. Introduction

In many real-life situations, direct data are not available and measurement errors occur. In many examples, such as medicine, astronomy, econometrics or meteorology, these measurement errors should not be neglected. Let us consider the following example from signal processing in oncology. Medical images (such as X-ray computed tomography, Magnetic Resonance Imaging) play an increasingly important role in diagnosing and treating cancer patients. In the clinical setting, imaging data allows to better evaluate whether a cancer patient is responding to therapy and to adjust the therapy accordingly. In such a setting, the response variable could be the total response to the treatment, a partial response or the absence of a response. However, image interpretation and management in clinical trials triggers a number of issues such as doubtful reliability of image analysis due to a high variability in image interpretation, censoring bias, and a number of operational issues due to complex image data workflow. Consequently, biomarkers, such as bidimensional measurements of lesions, suffer from measurement errors. For these reasons, the construction of decision rules from indirect observations may play a crucial role for this problem.

The problem of *inverse statistical learning* can be described as follows. Let us consider a generator of random inputs $X \in \mathcal{X}$, with unknown density distribution f with respect to some σ-finite measure ν, and (a possible) associated output $Y \in \mathcal{Y}$, from an unknown conditional probability. The joint law of (X, Y) is denoted as P. Given a class of functions $g \in \mathcal{G}$, the best possible decision rule, called an oracle, is defined as:

$$g^* \in \arg\min_{g\in\mathcal{G}} \mathbb{E}_P \ell(g, (X, Y)), \tag{1.1}$$

where $\ell(g, (x, y))$ measures the loss of g at point (x, y). For example, the set $\mathcal{G}$ can be made of functions $g : x \in \mathcal{X} \mapsto g(x) \in \mathcal{Y}$, whereas $\ell(g, (x, y)) = \Phi(y - g(x))$ can be a prediction loss function. The problem of inverse statistical learning consists in estimating the oracle g^* based on a set of indirect i.i.d. observations:

$$(Z_1, Y_1), (Z_2, Y_2), \dots, (Z_n, Y_n) \sim \widetilde{P}. \tag{1.2}$$

In (1.2), each input Z_i has density Af, where A is a known linear compact operator. The joint law of (Z, Y) is denoted as $\widetilde{P}$. We are facing an inverse problem.

The most extensively studied model with indirect observations is the additive measurement error model. In this case, we observe indirect inputs:

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

where $(\epsilon_i)_{i=1}^n$ are i.i.d. with known density η. It corresponds to a convolution operator $A_\eta : f \mapsto f * \eta$. Depending on the nature of the response $Y \in \mathcal{Y}$, we deal with classification with errors in variables, density deconvolution, or regression with errors in variables.

In this paper, we consider a bounded loss function ℓ such that for any $g \in \mathcal{G}$, $\ell(g,\cdot) : \mathcal{X} \times \mathcal{Y} \to [0,1]$ and a compact input space $\mathcal{X} \subset \mathbb{R}^d$. Given a class $\mathcal{G}$ of measurable functions $g : \mathcal{X} \to \mathbb{R}$, the performances of a given g is measured through its non-negative excess risk, given by:

$$R_\ell(g) - R_\ell(g^*),$$

where g^* is defined in (1.1) as a minimizer of the risk. It is important to point out that we do not adress in this paper the problem of model selection of $\mathcal{G}$. It consists in studying the difference $R_\ell(g^*) - \inf_g R_\ell(g)$, where the infimum is taken over all possible measurable functions g. Here, the target g^* corresponds to the oracle in the family $\mathcal{G}$. The purpose of this work is to use Empirical Risk Minimization (ERM) strategies based on a corrupted sample to mimic a minimizer g^* of the risk.

In the direct case, as we observe i.i.d. $(X_1, Y_1), \dots, (X_n, Y_n)$ with law P, a classical way is to consider ERM estimators defined as:

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} R_n(g), \tag{1.3}$$

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

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

In the sequel, the empirical measure of the direct sample $(X_1, Y_1), \dots, (X_n, Y_n)$ will be denoted as P_n. A large literature (see Vapnik (2000) for such a generality) deals with the statistical performances of (1.3) in terms of excess risk. To be concise, under complexity assumptions over $\mathcal{G}$ (such as finite VC dimension in Vapnik (1982), entropy conditions (van de Geer (2000)), or Rademacher complexity assumptions in Koltchinskii (2006)), it is possible to get both consistency and rates of convergence of ERM estimators (see also Massart and Nédélec (2006) in classification). The main probabilistic tool is the statement of uniform concentration of the empirical measure to the true measure. It comes from the so-called Vapnik's bound:

$$\begin{aligned} R_\ell(\hat{g}_n) - R_\ell(g^*) &\leq R_\ell(\hat{g}_n) - R_n(\hat{g}_n) + R_n(g^*) - R_\ell(g^*) \\ &\leq 2 \sup_{g \in \mathcal{G}} |(P_n - P)\ell(g)|. \end{aligned} \tag{1.4}$$

It is important to highlight that (1.4) can be improved using a local approach (see Massart (2000)). It consists in reducing the supremum to a neighborhood of g^*. We do not develop these important refinements 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 classification.

Here, the framework is essentially different. Given a linear compact operator A, we observe a corrupted sample $(Z_1, Y_1), \dots, (Z_n, Y_n)$ where Z_i, $i =$

$1,\dots,n$ are i.i.d. with density Af. As a result, the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$ is unobservable and standard ERM (1.3) is not available. Unfortunately, using the contaminated sample $(Z_1,Y_1),\dots,(Z_n,Y_n)$ in standard ERM (1.3) seems a wrong track:

$$\frac{1}{n}\sum_{i=1}^n \ell(g,(Z_i,Y_i)) \longrightarrow \mathbb{E}_{\widetilde{P}}\ell(g,(Z,Y)) \neq R_\ell(g).$$

Due to the action of A, the empirical measure from the indirect sample, denoted by $\widetilde{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{(Z_i,Y_i)}$, differs from P_n. We are facing an ill-posed inverse problem. This problem has been recently considered in Loustau and Marteau (2013) for discriminant analysis with errors in variables (see also Loustau and Marteau (2012) for completeness).

In this work, we suggest a comparable strategy in statistical learning. Given a smoothing parameter α, we consider the following α-Empirical Risk Minimization (α-ERM):

$$\arg\min_{g\in\mathcal{G}} R_n^\alpha(g), \tag{1.5}$$

where $R_n^\alpha(g)$ is defined in a general way as:

$$R_n^\alpha(g) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(g,(x,y))\hat{P}_\alpha(dx,dy). \tag{1.6}$$

The measure $\hat{P}_\alpha = \hat{P}_\alpha(Z_1,Y_1,\dots,Z_n,Y_n)$ is data-dependent to the set of indirect inputs $(Z_1,\dots,Z_n)$. It will be related to standard regularization methods coming from the inverse problem literature (see Engl et al. (1996)). Explicit constructions of $\hat{P}_\alpha$ and empirical risk (1.6) is detailled in Section 2 and Section 3. This construction depends on the inverse problem that we have at hand, and the regularization method used. Consequently, the smoothing parameter may be the bandwidth of some kernel estimator, or some threshold of a spectral cut-off. We denote it as α in its full generality.

To study the performances of the minimizer $\hat{g}_n^\alpha$ of the empirical risk (1.6), it is possible to use empirical processes theory in the spirit of van de Geer (2000); van der Vaart and Wellner (1996) or more recently Koltchinskii (2006). Following (1.4), in the presence of indirect observations, we can write[1]:

$$\begin{aligned}
&R_\ell(\hat{g}_n^\alpha) - R_\ell(g^*)\\
&\quad\le R_\ell(\hat{g}_n^\alpha) - R_n^\alpha(\hat{g}_n^\alpha) + R_n^\alpha(g^*) - R_\ell(g^*)\\
&\quad\le R_\ell^\alpha(\hat{g}_n^\alpha) - R_n^\alpha(\hat{g}_n^\alpha) + R_n^\alpha(g^*) - R_\ell^\alpha(g^*) + (R_\ell - R_\ell^\alpha)(\hat{g}_n^\alpha - g^*)\\
&\quad\le \sup_{g\in\mathcal{G}} |(R_n^\alpha - R_\ell^\alpha)(g-g^*)| + \sup_{g\in\mathcal{G}} |(R_\ell - R_\ell^\alpha)(g-g^*)|,
\end{aligned} \tag{1.7}$$

[1] where with a slight abuse of notations, we write:

$$(R_\ell - R_\ell^\alpha)(g-g') = R_\ell(g) - R_\ell(g') - R_\ell^\alpha(g) + R_\ell^\alpha(g').$$

where in the sequel, for any fixed $g \in \mathcal{G}$:

$$R_\ell^\alpha(g) = \int \ell(g,(x,y))\mathbb{E}_{\widetilde{P}^{\otimes n}} \hat{P}_\alpha(dx,dy) = \mathbb{E}_{\widetilde{P}^{\otimes n}} R_n^\alpha(g). \tag{1.8}$$

Bound (1.7) is an inverse counterpart of the classical Vapnik's bound (1.4). It consists in two terms:

- A variance term $\sup_{g\in\mathcal{G}} |(R_n^\alpha - R_\ell^\alpha)(g-g^*)|$ related to the ERM strategy. This term can be controlled thanks to uniform exponential inequalities such as Talagrand's concentration inequality, applied to a class of functions depending on the smoothing parameter α.
- A bias term $\sup_{g\in\mathcal{G}} |(R_\ell^\alpha - R_\ell)(g-g^*)|$: it comes from the estimation of P into the expression of $R_\ell(g)$ with estimator $\hat{P}_\alpha$. This additional term is specific to our problem. However, it seems to be related to the usual bias term in nonparametric statistics. Indeed, we can see easily that:

$$R_\ell^\alpha(g) - R_\ell(g) = \int \ell(g,(x,y))[\mathbb{E}\hat{P}_\alpha - P](dx,dy).$$

The choice of the smoothing parameter α is crucial in the decomposition (1.7). It has to be chosen as a trade-off between the bias term and the variance term. If we consider the classical errors-in-variables model with kernel deconvolution estimators (see Loustau and Marteau (2013)), the variance term explodes when the bandwidth of the kernel tends to zero whereas the bias term vanishes. As a consequence, the optimal value for α will depend on unknown parameters, such as the regularity, the margin, and the ill-posedness. The problem of adaptation is not adressed in this paper but it is an interesting future direction.

In this work, we consider $\mathcal{Y} = \{0,1,\dots,M\}$ for $M \geq 1$. In other words, we study the model of supervised classification with indirect observations (see Devroye et al. (1996) for a survey in the direct case). However, other issues could be adressed and fall into the general framework of this introduction, such as unsupervised classification with errors-in-variables (see Loustau (2012)). The problem of estimation of level sets, supports, or manifolds in the presence of indirect observations could also be treated similarly.

The organization of the present contribution is as follows. In Section 2, we propose to give a general construction of the empirical risk (1.6) in classification thanks to the set of indirect observations. We state excess risk bounds for a solution of the α-ERM (1.5) under minimal assumptions over the loss function ℓ and the complexity of $\mathcal{G}$. It gives a generalization of the results of Koltchinskii (2006) when dealing with indirect observations. Section 3 gives applications of the result of Section 2 in two particular settings. In the errors-in-variables case, we use kernel deconvolution estimators in the empirical risk minimization. We deduce fast rates of convergence which coincide with a recent lower bound stated in Loustau and Marteau (2013). This illustrates rather well the asymptotic optimality of the method. We also consider the general case, where A is some known linear compact operator. In this case, we use a spectral cut-off by considering the Singular Value Decomposition (SVD) of the operator A. Section 4 concludes the paper with a discussion related to many open problems. Section 5 is dedicated to the proofs of the main results.

2. General upper bound

In this section, we detail the construction of the empirical risk (1.6) in classification. We give minimal assumptions to control the expected excess risk of the procedure. The construction of the empirical risk is based on the following decomposition of the true risk:

$$R_\ell(g) = \sum_{y\in\mathcal{Y}} p(y) \int_{\mathcal{X}} \ell(g,(x,y)) f_y(x)\nu(dx), \tag{2.1}$$

where $f_y(\cdot)$ is the conditional density of $X|Y = y$ and $p(y) = \mathbb{P}(Y = y)$, for any $y \in \mathcal{Y} = \{0,\dots,M\}$. With such a decomposition, we suggest to estimate each $f_y(\cdot)$ using a nonparametric density estimator. To state a general upper bound, given $n_y = \text{card}\{i : Y_i = y\}$, $k_\alpha : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and the set of inputs $(Z_i^y)_{i=1}^{n_y} = \{Z_i, i = 1,\dots,n : Y_i = y\}$, we consider a family of estimators such as:

$$\forall y \in \mathcal{Y}, \hat{f}_y(x) = \frac{1}{n_y}\sum_{i=1}^{n_y} k_\alpha(Z_i^y, x). \tag{2.2}$$

Assumption (2.2) provides a variety of nonparametric estimators of f_y. For instance, in Section 3, we construct deconvolution kernel estimators. This is a rather classical approach in deconvolution problems (see Fan (1991) or Meister (2009)). In this case, the smoothing parameter corresponds to the d-dimensional bandwidth of the deconvolution kernel. Another standard example where (2.2) holds is presented in Section 3. It corresponds to projection estimators (or spectral cut-off) of the conditional densities using the SVD of operator A. In this case, the smoothing parameter is the dimension of the projection method. Of course, many other regularization methods could be considered, such as Tikhonov regularization or Landweber (see Engl et al. (1996)). Moreover, it is important to note that we consider a constant smoothing level α for any class $y \in \mathcal{Y}$ in (2.2). Indeed, for the sake of simplicity, we restrict ourselves to similar regularity assumptions for the conditional densities f_y. Consequently, to get satisfying upper bounds, we will see that α does not necessary depend on the value $y \in \mathcal{Y}$.

Finally we plug estimators (2.2) in the true risk (2.1) to get an empirical risk defined as:

$$R_n^\alpha(g) = \sum_{y\in\mathcal{Y}} \int_{\mathcal{X}} \ell(g,(x,y)) \hat{f}_y(x)\nu(dx)\hat{p}(y),$$

where $\hat{p}(y) = \frac{n_y}{n}$ is an estimator of the quantity $p(y) = \mathbb{P}(Y = y)$. Thanks to (2.2), this empirical risk can be written as:

$$R_n^\alpha(g) = \frac{1}{n} \sum_{i=1}^n \ell_\alpha(g, (Z_i, Y_i)), \tag{2.3}$$

where $\ell_\alpha(g, (z, y))$ is a modified version of $\ell(g, (x, y))$ given by:

$$\ell_\alpha(g, (z, y)) = \int_{\mathcal{X}} \ell(g, (x, y)) k_\alpha(z, x) \nu(dx).$$

In this section, we study general upper bounds for the expected excess risk of the estimator:

$$\hat{g}_n^\alpha \in \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \ell_\alpha(g, (Z_i, Y_i)). \tag{2.4}$$

In case no such minimum exists, we can consider a δ-approximate minimizer as in Bartlett and Mendelson (2006) without significant change in the results.

The main idea is to use iteratively a concentration inequality for suprema of empirical processes due to Bousquet (2002). It allows to control the increments of the empirical process:

$$\nu_n^\alpha(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\ell_\alpha(g, (Z_i, Y_i)) - \mathbb{E}_{\widetilde{P}} \ell_\alpha(g, (Z, Y)) \right).$$

Here, it is important to note that Talagrand's type inequality has to be applied to the class of functions $\{(z, y) \mapsto \ell_\alpha(g, (z, y)), g \in \mathcal{G}\}$. This class depends on a regularization parameter α. This parameter will be calibrated as a function of n and that's why the concentration inequality has to be used carefully. For this purpose, we introduce in Definition 1 particular classes $\{\ell_\alpha(g), g \in \mathcal{G}\}$.

Definition 1. We say that the class $\{\ell_\alpha(g), g \in \mathcal{G}\}$ is a LB-class (Lipschitz bounded class) with respect to μ with parameters $(c(\alpha), K(\alpha))$ if these two properties hold:

($\mathbf{L}_\mu$) $\{\ell_\alpha(g), g \in \mathcal{G}\}$ is Lipschitz w.r.t. μ with constant $c(\alpha)$:

$$\forall g, g' \in \mathcal{G}, \, \|\ell_\alpha(g) - \ell_\alpha(g')\|_{L_2(\widetilde{P})} \leq c(\alpha) \|\ell(g) - \ell(g')\|_{L_2(\mu)}.$$

(B) $\{\ell_\alpha(g), g \in \mathcal{G}\}$ is uniformly bounded with constant $K(\alpha)$:

$$\sup_{g \in \mathcal{G}} \sup_{(z,y)} |\ell_\alpha(g, (z, y))| \leq K(\alpha).$$

A LB-class of loss function is Lipschitz and bounded with constants which depend on α. Examples of LB-classes are presented in Section 3. These properties are necessary to derive explicitly the upper bound of the variance in (1.7) as a function of α.

More precisely, the Lipschitz property $(\mathbf{L}_\mu)$ is a key ingredient to control the complexity of the class of functions $\{\ell_\alpha(g), g \in \mathcal{G}\}$. In the sequel, we use the following geometric complexity parameter:

$$\widetilde{\omega}_n(\mathcal{G}, \delta, \mu) = \mathbb{E} \sup_{g,g' \in \mathcal{G}: \|\ell(g) - \ell(g')\|_{L_2(\mu)} \leq \delta} \left| (\widetilde{P} - \widetilde{P}_n)(\ell_\alpha(g) - \ell_\alpha(g')) \right|. \tag{2.5}$$

This quantity corresponds to the indirect counterpart of more classical local complexities introduced in a variety of papers (see Bartlett et al. (2005), Koltchinskii (2006), Massart (2000)). Its control as a function of n,δ and α is a key point to get fast rates of convergence. This can be done thanks to the following lemma.

Lemma 1. *Consider a LB-class $\{\ell_\alpha(g), g \in \mathcal{G}\}$ with respect to μ with Lipschitz constant $c(\alpha)$. Then, given some $0 < \rho < 1$, we have for some $C_1 > 0$:*

$$\mathcal{H}_B(\{\ell(g),\, g \in \mathcal{G}\}, \epsilon, L_2(\mu)) \leq c\epsilon^{-2\rho} \Rightarrow \widetilde{\omega}_n(\mathcal{G}, \delta, \mu) \leq C_1 \frac{c(\alpha)}{\sqrt{n}} \delta^{1-\rho},$$

where $\mathcal{H}_B(\{\ell(g),\, g \in \mathcal{G}\}, \epsilon, L_2(\mu))$ denotes the ϵ-entropy with bracketing of the set $\{\ell(g),\, g \in \mathcal{G}\}$ with respect to $L_2(\mu)$ (see van der Vaart and Wellner (1996) for a definition).

With such a lemma, it is possible to control the complexity in the indirect setup thanks to standard entropy conditions related with the class $\mathcal{G}$. The proof is presented in Section 5. It is based on a maximal inequality due to van der Vaart and Wellner (1996) applied to the class:

$$\mathcal{F}_\alpha = \{\ell_\alpha(g) - \ell_\alpha(g') : \|\ell(g) - \ell(g')\|_{L_2(\mu)} \leq \delta\}.$$

Finally, in Definition 1, **(B)** is also necessary to apply Bousquet's inequality. This condition could be relaxed by dint of recent advances on empirical processes in an unbounded framework (see Lecué and Mendelson (2012) or Lederer and van de Geer (2012)).

Another standard assumption to get fast rates of convergence is the so-called Bernstein (or margin) assumption.

Definition 2. For $\kappa \geq 1$, we say that $\mathcal{F}$ is a Bernstein class with respect to μ with parameter κ if there exists $\kappa_0 \geq 0$ such that for every $f \in \mathcal{F}$:

$$\|f\|^2_{L_2(\mu)} \leq \kappa_0 [\mathbb{E}_P f]^{\frac{1}{\kappa}}.$$

This assumption first appears in Bartlett and Mendelson (2006) for $\mu = P$ when $\mathcal{F} = \{\ell(g) - \ell(g^*), g \in \mathcal{G}\}$ is the excess loss class. It allows to control the excess risk in statistical learning using functional's Bernstein inequality such as Talagrand's type inequality. In classification, it goes back to the standard margin assumption (see Mammen and Tsybakov (1999); Tsybakov (2004b)), where in this case $\kappa = \frac{p+1}{p}$ for a so-called margin parameter $p \geq 0$. The same kind of assumptions have been introduced originally in the related problem of excess mass by Polonik (1995).

Definition 2 has to be combined with the Lipschitz property of Definition 1. It allows us to have the following serie of inequalities:

$$\|\ell_\alpha(g) - \ell_\alpha(g^*)\|_{L_2(\widetilde{P})} \leq c(\alpha)\|f\|_{L_2(\mu)} \leq c(\alpha) \left(\mathbb{E}_P f\right)^{\frac{1}{2\kappa}}, \tag{2.6}$$

where $f \in \mathcal{F} = \{\ell(g) - \ell(g^*),\, g \in \mathcal{G}\}$.

Last definition provides a control of the bias term in (1.7) as follows:

Definition 3. The class $\{\ell_\alpha(g), g \in \mathcal{G}\}$ has approximation function $a(\alpha)$ and residual constant $0 < r < 1$ if the following holds:

$$\forall g \in \mathcal{G},\, (R_\ell - R_l^\alpha)(g - g^*) \leq a(\alpha) + r(R_\ell(g) - R_\ell(g^*)),$$

where with a slight abuse of notations, we write:

$$(R_\ell - R_\ell^\alpha)(g - g^*) = R_\ell(g) - R_\ell(g^*) - R_\ell^\alpha(g) + R_\ell^\alpha(g^*).$$

This definition warrants a control of the bias in the Inverse Vapnik's bound (1.7). It is straightforward that with Definition 3, we get a control of the excess risk as follows:

$$R_\ell(\hat{g}_n^\alpha) - R_\ell(g^*) \quad \leq \quad \frac{1}{1-r}\left(\sup_{g \in \mathcal{G}(1)} |(\widetilde{P}_n - \widetilde{P})(\ell_\alpha(g) - \ell_\alpha(g^*))| + a(\alpha)\right),$$

where in the sequel:

$$\mathcal{G}(\delta) = \{g \in \mathcal{G} : R_\ell(g) - R_\ell(g^*) \leq \delta\}.$$

Explicit functions $a(\alpha)$ and residual constant $r < 1$ are obtained in Section 3. There depend on the regularity conditions and allow to get fast rates of convergence.

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

Theorem 1. *Consider a LB-class $\{\ell_\alpha(g), g \in \mathcal{G}\}$ with respect to μ with parameters $(c(\alpha), K(\alpha))$ and approximation function $a(\alpha)$ such that:*

$$a(\alpha) \leq C_1 \left(\frac{c(\alpha)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}} \quad \text{and } K(\alpha) \leq \frac{c(\alpha)^{\frac{2\kappa}{2\kappa+\rho-1}} n^{\frac{\kappa+\rho-1}{2\kappa+\rho-1}}}{1 + \log\log_q n}, \tag{2.7}$$

for some $C_1 > 0$ and $q > 1$.

Suppose $\{\ell(g) - \ell(g^*), g \in \mathcal{G}\}$ is Bernstein with respect to μ with parameter $\kappa \geq 1$ where $g^* = \arg\min_{\mathcal{G}} R_\ell(g)$ is unique. Suppose there exists $0 < \rho < 1$ such that:

$$\mathcal{H}_B(\{\ell(g),\, g \in \mathcal{G}\}, \epsilon, L_2(\mu)) \leq C_2 \epsilon^{-2\rho}, \tag{2.8}$$

for some $C_2 > 0$.

Then estimator $\hat{g}_n^\alpha$ defined in (2.4) satisfies, for n great enough:

$$\mathbb{E}R_\ell(\hat{g}_n^\alpha) - R_\ell(g^*) \leq C \left(\frac{c(\alpha)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}},$$

where $C = C(C_1, C_2, \kappa, \kappa_0, \rho, q) > 0$.

The proof of this result is presented in Section 5. Here follows some remarks.

This upper bound generalizes the result presented in Koltchinskii (2006) to the indirect framework. Theorem 1 provides rates of convergence:

$$\left(c(\alpha)/\sqrt{n}\right)^{2\kappa/2\kappa+\rho-1}.$$

In the noise-free case, with standard ERM estimators, Tsybakov (2004b); V. Koltchinskii (2006) obtain fast rates $n^{-\kappa/2\kappa+\rho-1}$. In the presence of contaminated inputs, rates are slower since $c(\alpha)\to+\infty$ as $n\to+\infty$. Hence, Theorem 1 shows that the Lipschitz constant introduced in Definition 1 is central. It gives the price to pay for the inverse problem in the rates.

The behavior of the Lipschitz constant $c(\alpha)$ depend on the difficulty of the inverse problem through the degree of ill-posedness of operator A. Section 3 proposes to deal with mildly ill-posed inverse problems. In this case, $c(\alpha)$ depends polynomially on α.

Gathering with the complexity assumption (2.8), it leads to a control of the variance term in decomposition (1.7). The first statement of condition (2.7) gives the order of the bias term. It leads to the excess risk bound.

The second part of (2.7) is due to the use of a deviation's inequality from Bousquet (2002) to the class $\{\ell_\alpha(g), g\in\mathcal{G}\}$. In Section 3, we give explicit constants $c(\alpha)$ and $K(\alpha)$. It appears that this assumption is always guaranteed.

The control of the complexity in (2.8) is expressed in terms of bracketing entropy of the loss class. This assumption allows to control the complexity parameter $\widetilde{\omega}_n(\mathcal{G},\epsilon,\mu)$ defined in (2.5). As in Koltchinskii (2006), we can state Theorem 1 by using directly a control of the complexity (2.5). In the context of convex loss minimization, Koltchinskii (2006) considers many examples of hypothesis spaces, from finite VC classes to more complex functional classes such as kernel classes. This can be done in our context as well.

Finally, Theorem 1 requires the unicity of the Bayes g^*. Such a restriction can be avoided using a more sophisticated geometry as in (Koltchinskii, 2006, Section 4).

At this time, it is important to note that Theorem 1 depends on measure μ introduced in Definitions 1 and 2. In the rest of the paper, we will consider two particular cases: $\mu=\nu\otimes P_Y$ ($\mu=\nu_Y$ for short in the sequel) and $\mu=P$. The Lipschitz property $(\mathbf{L}_\mu)$ with $\mu=P$ is stronger than $(\mathbf{L}_\mu)$ with $\mu=\nu\otimes P_Y$. Indeed, for any measurable function $h:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$, if $\|f_y\|_\infty\leq C_y$, $\forall y\in\mathcal{Y}$:

$$\mathbb{E}_P f^2\leq\max_{y\in\mathcal{Y}}C_y\sum_{y\in\mathcal{Y}}p_y\int f(x,y)^2\nu(dx)=\max_{y\in\mathcal{Y}}C_y\,\|f\|^2_{L_2(\nu_Y)}.$$

Since $\|\cdot\|_{L_2(P)}\leq C\|\cdot\|_{L_2(\nu_Y)}$ for some $C>0$, a Bernstein class with respect to ν_Y is also Bernstein with respect to P (see Definition 2). The most favorable case ($\mu=\nu_Y$) arises in binary classification (see Tsybakov (2004b) or Massart and Nédélec (2006)). Section 3 states rates of convergence in these two different settings.

3. Applications

In this section, we propose to apply the general upper bound of Theorem 1 to give excess risk bounds for α-ERM strategies in two distinct frameworks. The first result deals with the errors-in-variables case where operator A is a convolution product. Using kernel deconvolution estimators, we obtain fast rates of convergence which coincide with recent minimax fast rates in discriminant analysis (see Loustau and Marteau (2013)). Then, we consider the general case where A is any linear compact operator. In this case, we introduce a family of projection estimators (or spectral cut-off) by diagonalizing A^*A. We also study two different settings in the sequel, namely $\mu = \nu_Y$ and $\mu = P$ (see the discussion at the end of Section 2).

3.1. Errors-in-variables case

The elementary model of indirect observations is the additive measurement error model with known error density. In this case, we suppose that we observe a corrupted training set (Z_i, Y_i), $i = 1, \dots, n$ where:

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

The sequence of random variables $\epsilon_1, \dots, \epsilon_n$ are i.i.d. $\mathbb{R}^d$-random variables with bounded density η with respect to the Lebesgue measure on $\mathbb{R}^d$. In this situation, operator A is exactly known as a convolution product with density η. Note that in practical applications, this knowledge cannot be guaranteed. However, in most examples, we could be able to estimate the error density η from replicated measurements. In the sequel, we do not address this problem and we focus on the deconvolution step itself.

In the errors-in-variables case, the difficulty of this inverse problem can be represented thanks to the asymptotic behavior of the Fourier transform of the noise density η. Assumption **(A1)** below concerns a polynomial asymptotic behavior of the characteristic function of the noise distribution. These kind of restrictions are standard in deconvolution problems (see Fan (1991); Butucea (2007); Meister (2009)).

(A1) *There exist* $(\beta_1, \dots, \beta_d)' \in \mathbb{R}_+^d$ *such that for all* $i \in \{1, \dots, d\}$, $\beta_i > \frac{1}{2}$ *and:*

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

where $\eta = \Pi_{i=1}^d \eta_i$ *and* $\mathcal{F}[\eta_i]$ *denotes the Fourier transform of* η_i. *Moreover, we assume that* $\mathcal{F}[\eta_i](t) \neq 0$ *for all* $t \in \mathbb{R}$ *and* $i \in \{1, \dots, d\}$.

Assumption **(A1)** focuses on moderately ill-posed inverse problems by considering polynomial decay of the Fourier transform. Notice that straightforward modifications in the proofs allow to consider severely ill-posed inverse problems.

In this framework, we construct kernel deconvolution estimators of the densities $f_y, y \in \mathcal{Y}$. For this purpose, let us introduce $\mathcal{K} = \prod_{j=1}^d \mathcal{K}_j : \mathbb{R}^d \to \mathbb{R}$ a

d-dimensional function defined as the product of d unidimensional functions $\mathcal{K}_j$. Then if we denote by $\lambda=(\lambda_1,\dots,\lambda_d)\in\mathbb{R}_+^d$ a set of (positive) bandwidths, we define $\mathcal{K}_\eta$ as

$$\begin{aligned}\mathcal{K}_\eta \quad &: \quad \mathbb{R}^d\to\mathbb{R}\\ &\qquad t\mapsto \mathcal{K}_\eta(t)=\mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)}\right](t).\end{aligned} \tag{3.1}$$

To apply Theorem 1, we require the following assumptions on the kernel $\mathcal{K}$.

(K1) *The kernel* $\mathcal{K}$ *satisfies*

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

where $\operatorname{supp} g=\{x:|g(x)|\geq 0\}$ *and* $[-S,S]=\bigotimes_{i=1}^d[-S_i,S_i]$.

We also need a Hölder regularity condition for the conditional densities:

(R1) *Given* $\gamma,L>0$*, for any* $y\in\mathcal{Y}$*,* $f_y\in\mathcal{H}(\gamma,L)$ *where:*

$\mathcal{H}(\gamma,L)=\{f\in\Sigma(\gamma,L): f$ *are bounded probability densities w.r.t. Lebesgue*$\}$,

and $\Sigma(\gamma,L)$ *is the class of isotropic Hölder continuous functions* f *having continuous partial derivatives up to order* $\lfloor\gamma\rfloor$*, the maximal integer strictly less than* γ *and such that:*

$$|f(y)-p_{f,x}(y)|\leq L|x-y|^\gamma,$$

where $p_{f,x}$ *is the Taylor polynomial of* f *at order* $\lfloor\gamma\rfloor$ *at point* x*.*

Moreover, we also need the associated classical assumption for the kernel $\mathcal{K}$.

(Korder) *The kernel* $\mathcal{K}$ *is of order* $m\in\mathbb{N}$ *if and only if:*

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

The Hölder regularity **(R1)**, gathering with **(Korder)** with $m=\lfloor\gamma\rfloor$ is standard to control the bias term of kernel estimators in density estimation or deconvolution (see for instance Tsybakov (2004a)).

In this context, we define the λ-ERM estimator based on (2.4) as:

$$\hat{g}_n^\lambda\in\arg\min_{g\in\mathcal{G}}R_n^\lambda(g), \tag{3.2}$$

where for any $g\in\mathcal{G}$:

$$R_n^\lambda(g)=\frac{1}{n}\sum_{i=1}^n\ell_\lambda(g,(Z_i,Y_i)),$$

and $\ell_\lambda(g,(z,y))$ is given by:

$$\ell_\lambda(g,(z,y)) = \int_{\mathcal{X}} \ell(g,(x,y)) \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) dx.$$

In the sequel, with a slight abuse of notations we write for any $z=(z_1,\dots,z_d)$, $x=(x_1,\dots,x_d)\in\mathbb{R}^d$, $\lambda=(\lambda_1,\dots,\lambda_d)\in\mathbb{R}_+^d$:

$$\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right) = \Pi_{i=1}^d \frac{1}{\lambda_i}\mathcal{K}_\eta\left(\frac{z_1-x_1}{\lambda_1},\dots,\frac{z_d-x_d}{\lambda_d}\right).$$

Theorem 2 below presents the rates of convergence of (3.2) with a kernel $\mathcal{K}$ satisfying **(K1)** and **(Korder)** with $m=\lfloor\gamma\rfloor$.

Theorem 2. *Suppose $\{\ell(g)-\ell(g^*), g\in\mathcal{G}\}$ is a Bernstein class with respect to ν_Y with parameter $\kappa\geq 1$. Suppose $0<\rho<1$ exists such that:*

$$\mathcal{H}_B(\{\ell(g),\, g\in\mathcal{G}\},\epsilon,L_2(\mu)) \leq C_2\epsilon^{-2\rho},$$

for some $C_2>0$.

*Under assumptions **(A1)** and **(R1)**, we have, for n great enough:*

$$\mathbb{E}R_\ell(\hat{g}_n^\lambda) - R_\ell(g^*) \leq Cn^{-\frac{\kappa\gamma}{\gamma(2\kappa+\rho-1)+(2\kappa-1)\sum_{i=1}^d \beta_i}},$$

where $C=C(K_1,S,\gamma,L,K_m,C_2,\rho,\kappa_0,\kappa)>0$ and $\lambda=(\lambda_1,\dots,\lambda_d)$ is given by:

$$\forall i\in\{1,\dots,d\},\ \lambda_i = n^{-\frac{2\kappa-1}{2\gamma(2\kappa+\rho-1)+2(2\kappa-1)\sum_{i=1}^d\beta_i}}. \tag{3.3}$$

The proof of this result is postponed to Section 5. Here follows some remarks.

Rates in Theorem 2 generalize the result of Koltchinskii (2006) (see also Tsybakov (2004b)) to the errors-in-variables case. Point out that if $\bar{\beta}=0$, we get the rates of the direct case. Here, the price to pay for the inverse problem of deconvolution can be quantified as:

$$\frac{(2\kappa-1)\sum_{i=1}^d\beta_i}{\gamma}.$$

Hence, the performances of the method depend on the behavior of the characteristic function of the noise distribution. In classification with fast rates, it is important to notice that the influence of the errors in variables is related to both parameters κ and γ. Same phenomenon also occurs in Loustau and Marteau (2013).

It is also interesting to study the minimax optimality of the result of Theorem 2 using the lower bound presented in Loustau and Marteau (2013). For this purpose, let us introduce a random couple (X,Y) with law P on $\mathcal{X}\times\{0,1\}$. Given some a priori $\mathcal{G}$ (chosen later on) and the class of associated candidates $\{g(x)=\mathbb{1}_G(x),\, G\in\mathcal{G}\}$, we consider the hard loss $\ell_H(g,(x,y))=|y-\mathbb{1}_G(x)|$.

In this case, the Bayes risk is defined as:

$$R_H(G) = \mathbb{E}|Y - \mathbb{1}_G(X)|.$$

It is easy to see that for $y \in \{0,1\}$ and $g(x) = \mathbb{1}_G(x)$, we have:

$$|\ell_H(g,(x,y)) - \ell_H(g',(x,y))| = ||y - \mathbb{1}_G(x)| - |y - \mathbb{1}_{G'}(x)|| = |\mathbb{1}_G(x) - \mathbb{1}_{G'}(x)|.$$

Gathering with the margin assumption, Lemma 2 in Mammen and Tsybakov (1999) allows us to write:

$$\begin{aligned} \|\ell_H(g) - \ell_H(g')\|^2_{L_2(\nu_Y)} = \|\mathbb{1}_G - \mathbb{1}_{G'}\|^2_{L_2(\nu)} &= d_\Delta(G,G') \\ &\leq \frac{c_0}{2}\left(R_H(g) - R_H(g')\right)^{\frac{p}{p+1}}, \end{aligned}$$

where $p \geq 0$ denoted the so-called margin parameter. As a result, provided that $G^* \in \mathcal{G}$ and under the margin assumption, the excess loss class $\{\ell_H(g) - \ell_H(g^*)\}$ is Bernstein with respect to ν_Y with parameter $\kappa = \frac{p+1}{p}$.

To apply Theorem 2, we need to check $(\mathbf{L}_\mu)$ and **(B)** from Definition 1. Remark that from Lemma 3 in Loustau and Marteau (2012), we have:

$$\|\ell_\lambda(g) - \ell_\lambda(g')\|^2_{L_2(\widetilde{P})} \leq C\Pi_{i=1}^d \lambda_i^{-\beta_i} d_\Delta(G,G'),$$

where for any $g = \mathbb{1}_G$:

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

Consequently, $\{\ell_\lambda(g), g = \mathbb{1}_G : G \in \mathcal{G}\}$ is a LB-class with respect to ν_Y with constants $c(\lambda)$ and $K(\lambda)$ given by:

$$c(\lambda) = \Pi_{i=1}^d \lambda_i^{-\beta_i} \text{ and } K(\lambda) = \Pi_{i=1}^d \lambda_i^{-\beta_i - 1/2}.$$

The last step is to control the complexity parameter $\widetilde{\omega}_n(\mathcal{G},\delta,\nu_Y)$ as a function of δ. For this purpose, we choose $\mathcal{G}(\gamma,L)$ the class of sets of the form $\{h(x) \geq 0\}$, where $h \in \Sigma(\gamma,L)$ has Hölder regularity $\gamma > 0$. With Lemma 5.1 in Audibert and Tsybakov (2007), a control of the d_Δ-entropy with bracketing of the class $\mathcal{G}(\gamma,L)$ can be found easily:

$$\log \mathcal{N}(\mathcal{G}(\gamma,L), d_\Delta, \epsilon) \leq c\epsilon^{-\frac{d}{\gamma p}}.$$

As a result, since $d_\Delta(G,G') = \|\mathbb{1}_G - \mathbb{1}_{G'}\|^2_{L_2(\nu)}$, we have:

$$\log \mathcal{N}(\{\mathbb{1}_G, G \in \mathcal{G}(\gamma,L)\}, L_2(\nu), \epsilon) \leq c\epsilon^{-\frac{2d}{\gamma p}}.$$

To get a control of the desired complexity term, we can apply Lemma 1 in Section 2 to get:

$$\widetilde{\omega}_n(\mathcal{G},\delta,\nu_Y) \leq C_1 \frac{c(\lambda)}{\sqrt{n}}\delta^{1-\frac{d}{\gamma p}},$$

for some $C_1 > 0$.

Finally, using Lemma 4 in Section 5, in the particular case of the hard loss, $\{\ell_\lambda(g), g \in \mathcal{G}\}$ has approximation power $a(\lambda)$ with constant $0 < r < 1$ given by:

$$a(\lambda) = \sum_{i=1}^{d} \lambda_i^{\gamma(p+1)} \text{ and } r = \frac{p}{p+1}.$$

In this case, Theorem 2 leads to:

$$\mathbb{E}R_H(\hat{g}_n^\lambda) - R_H(g^*) \leq Cn^{-\frac{(p+1)\gamma}{\gamma(p+2)+d+2\beta}}.$$

This rate corresponds to the minimax rates of classification with errors in variables stated in Loustau and Marteau (2013). It ensures the minimax optimality of the method in the errors-in-variables case for the hard loss. An open problem is to give a lower bound for more general losses.

3.2. General case with singular values decomposition

For the sake of completeness, we also propose another α-ERM strategy in the general case, where A is a known compact operator. In the recent statistical literature, several methods of regularization have been proposed: Tikhonov type regularizations, recurcive procedures in Hilbert space, or projection (or spectral cut-off) methods. A natural way of projection for ill-posed inverse problems is associated with the singular values decomposition (SVD) of A. Denote A^* the adjoint of A, and assume A^*A is a compact operator with eigenvalues $(b_k^2)_{k\in\mathbb{N}^*}$ with associated eigenfunctions $(\phi_k)_{k\in\mathbb{N}^*}$. Clearly, $\|A\phi_k\| = b_k$ and by setting $\varphi_k = \frac{A\phi_k}{\|A\phi_k\|}$, $(\varphi_k)_{k\in\mathbb{N}^*}$ is also orthonormal. Furthermore:

$$A\phi_k = b_k\varphi_k \text{ and } A^*\varphi_k = b_k\phi_k,\ k \in \mathbb{N}^*. \tag{3.4}$$

We may also write, for any $y \in \mathcal{Y}$, $f_y = \sum_{k\in\mathbb{N}^*} b_k^{-1}\langle Af_y, \varphi_k\rangle\phi_k$. Then, the family of projection estimators $\hat{f}_y$ of each f_y, $y \in \mathcal{Y}$ has the form:

$$\hat{f}_y = \sum_{k=1}^{N} \hat{\theta}_k\phi_k, \tag{3.5}$$

where $N \geq 1$ is the regularization parameter and $\hat{\theta}_k^y$ is an unbiased estimator of $\theta_k^y = \langle f_y, \phi_k\rangle$ given by:

$$\hat{\theta}_k^y = \frac{1}{n_y}\sum_{i=1}^{n_y} b_k^{-1}\varphi_k(Z_i^y).$$

In the sequel, as in **(A1)**, we restrict ourselves to moderately ill-posed inverse problem considering the rate of decrease of the singular values of A.

(A2) *There exists $\beta \in \mathbb{R}_+$ such that:*

$$b_k \sim k^{-\beta} \text{as } k \to +\infty.$$

In this case, the rate of decrease of the singular values is polynomial. As an example, we can consider the convolution operator above and from an easy calculation, the spectral domain is the Fourier domain and **(A2)** is comparable to **(A1)**. However, assumption **(A2)** can deal with any linear inverse problem and is rather standard in the statistical inverse problem literature (see Cavalier (2008)).

In this framework, we also need the following assumption on the regularity of the conditional densities into the basis of the operator A:

(R2) *For any $y \in \mathcal{Y}$, $f_y \in \mathcal{P}(\gamma, L)$ where:*

$\mathcal{P}(\gamma, L) = \{f \in \Theta(\gamma, L) : f$ *are bounded probability densities w.r.t. Lebesgue* $\}$,

and $\Theta(\gamma, L)$ is the ellipsoïd in the SVD basis defined as:

$$\Theta(\gamma, L) = \{f = \sum_{k\geq 1} \theta_k \phi_k \in L_2(\mathcal{X}) : \sum_{k\geq 1} \theta_k^2 k^{2\gamma+1} \leq L\}.$$

Considering the SVD (3.4), we propose to replace in the true risk the conditional densities f_y by the family of projection estimators (3.5). In this case, assumption (2.2) is satisfied with $k_N(z,x) = \sum_{k=1}^N b_k^{-1} \varphi_k(z) \phi_k(x)$. It gives the following expression of the empirical risk:

$$R_n^N(g) = \frac{1}{n} \sum_{i=1}^n \ell_N(g, Z_i, Y_i),$$

where:

$$\ell_N(g, z, y) = \sum_{k=1}^N b_k^{-1} \int_{\mathcal{X}} \phi_k(x) \ell(g, (x, y)) \nu(dx) \varphi_k(z).$$

Next theorem states the rates of convergence for the ERM estimator $\hat{g}_n^N$ defined as:

$$\hat{g}_n^N = \arg\min_{g\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^n \ell_N(g, Z_i, Y_i).$$

Theorem 3. *Suppose $\{\ell(g) - \ell(g^*), g \in \mathcal{G}\}$ is Bernstein class with respect to ν_Y with parameter $\kappa \geq 1$. Suppose $0 < \rho < 1$ exists such that:*

$$\mathcal{H}_B(\{\ell(g),\, g \in \mathcal{G}\}, \epsilon, L_2(\mu)) \leq C_2 \epsilon^{-2\rho},$$

*for some $C_2 > 0$. Then, under **(A2)** and **(R2)**, $\hat{g}_n^N$ satisfies, for n great enough:*

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

where $C = C(\gamma, L, C_2, \rho, \kappa, \kappa_0)$ and we choose N such that:

$$N = n^{\frac{2\kappa-1}{2\gamma(2\kappa+\rho-1)+2(2\kappa-1)\beta}}.$$

Theorem 3 shows that if A is a linear compact operator, we can derive rates of convergence under a regularity assumption related to the spectrum of A. From this point of view, the result of Theorem 3 is comparable with Theorem 2.

However, the regularity assumption **(R2)** is stronger than **(R1)** since $\Theta(\gamma, L) = \{(\theta_k)_k : \sum \theta_k^2 k^{2\gamma+1} \leq L\}$, whereas with **(R1)**, f_y has $\lfloor \gamma \rfloor$ derivatives. As a consequence, this result highlights a lack of optimality of projection methods for the problem of inverse statistical learning. This phenomenon can be explain by taking a closer look at the bias control in Lemma 4 and Lemma 6. Indeed, to get a satisfying approximation function $a(\alpha)$ (see Definition 3), we act in two steps: the first step is to control locally the bias $\mathbb{E}\hat{f}(x) - f(x)$. Then, we use Bernstein assumption to get:

$$(R_\ell^\alpha - R_\ell)(\hat{g}_n^\alpha - g^*) \leq (R_\ell(\hat{g}_n^\alpha) - R_\ell(g^*))^{1/2\kappa} [\mathbb{E}\hat{f}(x) - f(x)].$$

This allows us to derive with Young's inequality an approximation function of the form $a(\alpha) = [\mathbb{E}\hat{f}(x) - f(x)]^{2\kappa/2\kappa-1}$ with residual term $r = \frac{1}{2\kappa}$. Unfortunately, we known that projection estimators are minimax in statistical inverse problems when we consider the integrated L_2-risk. In this case, the bias term has the form $\|\mathbb{E}\hat{f}_N - f\|^2$, and can be controlled with a weaker assumption than **(R2)**. Note that in kernel deconvolution estimation, minimax results are stated for both pointwise and integrated risk. In this case, the control of $\mathbb{E}\hat{f}(x) - f(x)$ can be managed optimally in Lemma 4. This explains the optimality of Theorem 2 in comparison with Theorem 3.

3.3. Restriction to K to deal with a weaker Bernstein assumption

In this subsection, we develop an alternative to Theorems 2–3 to deal with a weaker Bernstein assumption. For the sake of simplicity, we restrict ourselves in Theorems 2–3 to Bernstein class satisfying (see Definition 2):

$$\|\ell(g) - \ell(g^*)\|_{L_2(\nu_Y)}^2 \leq \kappa_0 [\mathbb{E}_P(\ell(g) - \ell(g^*))]^{1/\kappa}.$$

However, Bernstein classes with respect to ν_Y appear only in particular cases, such as classification with hard loss in the context of Mammen and Tsybakov (1999); Tsybakov (2004b) (see the discussion after Theorem 2). Here, we present a corollary of Theorems 2–3. To deal with Bernstein classes in the spirit of Bartlett and Mendelson (2006). The excess loss class is Bernstein with respect to P if:

$$\|\ell(g) - \ell(g^*)\|_{L_2(P)}^2 \leq \kappa_0 [\mathbb{E}_P(\ell(g) - \ell(g^*))]^{1/\kappa}.$$

For this purpose, we restrict the study to a set $K \subseteq \mathbb{R}^d$ where $f \geq c_0 > 0$ over K. We introduce the following restricted loss:

$$\ell_{\alpha,K}(g, z, y) = \int_K k_\alpha(z, x) \ell(g, (x, y)) \nu(dx). \tag{3.6}$$

It means that, as in Mammen and Tsybakov (1999) (see also Loustau and Marteau (2013)), we deal with the minimization of a true risk of the form:

$$R_{\ell,K}(g)=\sum_{y\in\mathcal{Y}}p(y)\int_K \ell(g,(x,y))f_y(x)dx.$$

With (3.6), it is straightforward to get $(\mathbf{L}_\mu)$ with $\mu=P$ since if $f\geq c_0>0$ on K, one gets:

$$\sum_{y\in\mathcal{Y}}p_y\int_K(\ell(g,(x,y))-\ell(g',(x,y)))^2\nu(dx)\leq\frac{1}{c_0}\|\ell(g)-\ell(g')\|_{L_2(P)}.$$

Roughly speaking, Assumption $(\mathbf{L}_\mu)$ in Definition 1 whith $\mu=P$ provides a control of the variance of $\ell_\alpha(g,(Z,Y))$ by the variance of $\ell(g,(X,Y))$. To have such a control, we need to restrict the problem to $\{x:f(x)>0\}$. Otherwise, the variance of $\ell_\alpha(g,(Z,Y))$ could be significantly large compared with the variance of $\ell(g,(X,Y))$.

The following corollary points out the same performances for the α-ERM over K defined as:

$$\hat{g}_n^{\alpha,K}=\arg\min_{g\in\mathcal{G}}\sum_{i=1}^n\ell_{\alpha,K}(g,Z_i,Y_i).$$

Corollary 1. *Suppose $\{\ell(g)-\ell(g^*),g\in\mathcal{G}\}$ is a Bernstein class with respect to P with parameter $\kappa\geq1$. Suppose $0<\rho<1$ exists such that:*

$$\mathcal{H}_B(\{\ell(g),\,g\in\mathcal{G}\},\epsilon,L_2(P))\leq C_2\epsilon^{-2\rho},$$

for some $C_2>0$. Then:

1. *Under **(A1)** and **(R1)**, $\hat{g}_n^{\lambda,K}$ satisfies, for n great enough:*

$$\mathbb{E}R_{\ell,K}(\hat{g}_n^{\lambda,K})-R_{\ell,K}(g^*)\leq Cn^{-\frac{\kappa\gamma}{\gamma(2\kappa+\rho-1)+(2\kappa-1)\sum_{i=1}^d\beta_i}},$$

for a choice of $\lambda=(\lambda_1,\dots,\lambda_d)$ given by:

$$\forall i\in\{1,\dots,d\},\ \lambda_i=n^{-\frac{2\kappa-1}{2\gamma(2\kappa+\rho-1)+2(2\kappa-1)\sum_{i=1}^d\beta_i}}. \tag{3.7}$$

2. *Under **(A2)** and **(R2)**, $\hat{g}_n^{N,K}$ satisfies, for n great enough:*

$$\mathbb{E}R_{\ell,K}(\hat{g}_n^{N,K})-R_{\ell,K}(g^*)\leq Cn^{-\frac{\kappa\gamma}{\gamma(2\kappa+\rho-1)+(2\kappa-1)\beta}},$$

where we choose N such that:

$$N=n^{\frac{2\kappa-1}{2\gamma(2\kappa+\rho-1)+2(2\kappa-1)\beta}}.$$

This corollary allows to get the same fast rates of convergence of Theorems 2–3 under a weaker Bernstein assumption. The price to pay for the α-ERM with restricted loss (3.6) relies on the dependence on K of the estimation procedure.

4. Conclusion

This paper has tried to investigate the effect of indirect observations into the statement of fast rates of convergence in empirical risk minimization. Many issues could be considered in future works.

The main result is a general upper bound in classification whith indirect observations, when we observe indirect inputs Z_i, $i = 1, \dots, n$ with law Af. We state that under a standard complexity assumption over the hypothesis space, the proposed estimator $\hat{g}_n^\alpha$ reaches (fast) rates of convergence of the form:

$$\mathcal{O}\left((c(\alpha)/\sqrt{n})^{2\kappa/(2\kappa+\rho-1)} \right),$$

where $\alpha > 0$ is a smoothing parameter and $c(\alpha)$ depends on the operator A of inversion. The proof is based on a deviation inequality for suprema of empirical processes. It seems to fit the indirect case provided that it is used carefully. For this purpose, we introduce Lipschitz and bounded classes $\{\ell_\alpha(g), g \in \mathcal{G}\}$, depending on a smoothing parameter α. It allows us to quantify the effect of the inverse problem on the empirical process machinery. The price to pay is summarized in a constant $c(\alpha)$ which exploses as n tends to infinity. The behavior of this constant is related to the degree of ill-posedness. Here, in the mildly ill-posed case, $c(\alpha)$ grows polyniomally as a function of α.

The result of Section 2 suggests the same degree of generality as the results of Koltchinskii (2006) in the direct case. It is well-known that the work of Koltchinskii allows to recover most of the recent results in statistical learning theory and the area of fast rates. Consequently, there is a nice hope that many problems dealing with indirect observations could be managed following the guiding thread of this paper.

Section 3 gives some illustrations in different settings. In the errors-in-variables case, using deconvolution kernel estimators, we obtain fast rates. These rates coincide with recent minimax results stated in Loustau and Marteau (2013) for the hard loss. In the general case, it is also possible to construct another $\alpha-$ERM strategy based on projections using the SVD of the operator A.

The estimation procedure proposed in this paper leads to different challenging open problems of adaptation and model selection. The method is not adaptive in many senses. At the first glance, we can see three levels of adaptation: (1) adaptation to the operator A; (2) adaptation to choose the tunable parameter α; (3) adaptation or model selection of the hypothesis space $\mathcal{G}$. At this time, it is important to note that at least in the direct case, the same machinery used to analyzed the order of the excess risk can be applied to produce penalized empirical risk minimization (see Tsybakov and van de Geer (2005); Koltchinskii (2006); Blanchard et al. (2008); Loustau (2009)). Moreover, the problem of unknown operator of inversion A has been already considered in the literature (see for instance Delaigle et al. (2008)). A standard approach is to suppose that we have repeated measurements in order to estimate the Fourier tranform of η. It can be the purpose of a future work.

Another possible powerful direction is to study more precisely the complexity of the class $\{\ell_\alpha(g), g \in \mathcal{G}\}$. On the one hand, we can use particular properties of loss functions, such as convexity or Lipschitz properties, to control $\tilde{\omega}_n(\mathcal{G}, \delta, \mu)$ defined in (2.5). For instance, the same type of results as Theorems 2–3 could be derived using entropy conditions of the hypothesis set $\mathcal{G}$ itself (instead of the loss class $\{\ell(g), g \in \mathcal{G}\}$). It allows to consider many standard minimization problems over finite VC-set or reproducing kernel Hilbert spaces. On the other hand, another challenging direction can be the control of the complexity from indirect observations thanks to entropy numbers of compact operators. Note that since $\mathcal{X}$ is compact, $\ell_\alpha(g, z, y) = \int_{\mathcal{X}} k_\alpha(z, x)\ell(g, (x, y))\nu(dx)$ can be considered as the image of $\ell(g)$ by the integral operator L_{k_α} associated to the function k_α. Hence, we have:

$$\{\ell_\alpha(g), g \in \mathcal{G}\} = L_{k_\alpha}(\{\ell(g), g \in \mathcal{G}\}).$$

Furthermore, it is clear that if k_α is continuous, L_{k_α} is well-defined and compact. Using for instance Williamson et al. (2001), and provided that ℓ is bounded and $\mathcal{G}$ consists of bounded functions in $L_2(\nu, \mathcal{X})$, entropy of the class $\{\ell_\alpha(g), g \in \mathcal{G}\}$ could be controlled in terms of the eigenvalues of the integral operator. In this case, it is clear that the entropy of the class depends strongly on the spectrum of the operator A.

More precisely, if A is a convolution product, Section 3.1 deals with kernel deconvolution estimators with bandwidth λ. As a result, the integral operator L_{k_λ} is defined as the convolution product $L_{k_\lambda} f(z) = \frac{1}{\lambda}\mathcal{K}_\eta(\frac{\cdot}{\lambda}) * f(z)$. Its spectrum is related to the behavior of the Fourier transform of the deconvolution kernel estimator, which corresponds to the quantity $\mathcal{F}[\mathcal{K}]/\mathcal{F}[\eta](\cdot/\lambda)$. At the end, the control of the entropy of the class of interest $\{\ell_\lambda(g), g \in \mathcal{G}\}$ could be calculated thanks to an assumption over the behavior of the Fourier transform of the noise distribution η such as **(A1)**. It can produce promising results.

Finally, the aim of this contribution was to derive excess risk bounds under standard assumptions over the complexity and the geometry of the considered class $\mathcal{G}$. An alternative point of view would be to state oracle-type inequalities. Indeed, Theorems 1–3 could be written in terms of exact oracle inequalities of the form:

$$\mathbb{E}R_\ell(\hat{g}_n^\alpha) \leq \inf_{g \in \mathcal{G}} R_\ell(g) + r_n(\mathcal{G}),$$

where the residual term $r_n(\mathcal{G})$ corresponds to the rates of convergence in Theorems 1–3. In this setting, it is well-known that ERM estimators reach optimal fast rates under a Bernstein assumption. However, the Bernstein assumption presented in Definition 2 is a strong assumption related to the geometry of the class $\mathcal{G}$. Lecué and Mendelson (2012) proposes to relax significantly the Bernstein assumption and points out non-exact oracle inequalities of the form:

$$\mathbb{E}R_\ell(\hat{g}_n^\alpha) \leq (1 + \epsilon) \inf_{g \in \mathcal{G}} R_\ell(g) + r_n(\mathcal{G}),$$

for some $\epsilon > 0$. These results hold without Bernstein condition for any non-negative loss functions. There is a nice hope that such a study can be done in the presence of indirect observations, using some minor modifications in the proofs.

5. Proofs

The main ingredient of the proofs is a concentration inequality for empirical processes in the spirit of Talagrand (Talagrand (1996)). We use precisely a Bennet deviation bound for suprema of empirical processes due to Bousquet (see Bousquet (2002)) applied to a class of measurable and bounded functions $f \in \mathcal{F}$. In this case it is stated in Bousquet (2002) that for all $t > 0$:

$$\mathbb{P}\left(Z \geq \mathbb{E}Z + \sqrt{2t(n\sigma^2 + (1+K)\mathbb{E}Z)} + \frac{t}{3}\right) \leq \exp(-t),$$

where

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} f(X_i) \right| \text{ and } \sup_{f \in \mathcal{F}} \operatorname{Var}(f(X_1)) \leq \sigma^2.$$

The proof of Lemma 2 below uses iteratively Bousquet's inequality and gives rise to solve the fixed point equation as in Koltchinskii (2006). For this purpose, we introduce, for a function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, the following transformations:

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

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 $j \in \mathbb{N}$.

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

5.1. Proof of Theorem 1

The following lemma is the key ingredient to get Theorem 1. For the sake of simplicity, we suppose that the oracle g^* is unique whereas a more sophisticated geometry can lead to the same kind of result without this assumption (see Koltchinskii (2006)).

Lemma 2. *Suppose $\{\ell_\alpha(g), g \in \mathcal{G}\}$ is such that:*

$$\sup_{g \in \mathcal{G}} \|\ell_\alpha(g)\|_\infty \leq K(\alpha).$$

Suppose $\{\ell_\alpha(g), g \in \mathcal{G}\}$ has approximation function $a(\alpha)$ and residual constant $0 < r < 1$ according to Definition 3. Define, for some constant $K > 0$:

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

$$\phi_n^\alpha(\mathcal{G}, \delta_j) = \mathbb{E} \sup_{g \in \mathcal{G}(\delta_j)} |\widetilde{P}_n - \widetilde{P}|[\ell_\alpha(g) - \ell_\alpha(g^*)],$$

$$D^\alpha(\mathcal{G}, \delta_j) = \sup_{g \in \mathcal{G}(\delta_j)} \sqrt{\widetilde{P}(\ell_\alpha(g) - \ell_\alpha(g^*))^2},$$

where $g^* = \arg\min_{\mathcal{G}} R_\ell(g)$ is unique.

Then $\forall \delta \geq \delta_n^\alpha(t) = [U_n^\alpha(\cdot, t)]_q^\dagger(\frac{1-r}{2q})$, if $a(\alpha) \leq \frac{1-r}{4q}\delta$ we have for $\hat{g} = \hat{g}_n^\alpha$:

$$\mathbb{P}(R_\ell(\hat{g}) - R_\ell(g^*) \geq \delta) \leq \log_q(\frac{1}{\delta})e^{-t}.$$

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 $\alpha \in \mathbb{R}_+^d$, consider the event:

$$E_{n,j}^\alpha(t) = \left\{ \sup_{g \in \mathcal{G}(\delta_j)} |\widetilde{P}_n - \widetilde{P}|[\ell_\alpha(g) - \ell_\alpha(g^*)] \leq U_n^\alpha(\delta_j, t) \right\},$$

where $\forall \delta > 0$, $\mathcal{G}(\delta) = \{g \in \mathcal{G} : R_\ell(g) - R_\ell(g^*) \leq \delta\}$. Then, we have, using Bousquet's version of Talagrand's concentration inequality (see Bousquet (2002)), for some $K > 0$, $\mathbb{P}(E_{n,j}^\alpha(t)^C) \leq e^{-t}$, $\forall t \geq 0$.

We restrict ourselves to the event $E_{n,j}^\alpha(t)$.

Using Definition 3, we have with a slight abuse of notations:

$$\begin{aligned} R_\ell(\hat{g}) - R_\ell(g^*) &\leq (\widetilde{P}_n - \widetilde{P})(\ell_\alpha(g^*) - \ell_\alpha(\hat{g})) + (R_\ell - R_\ell^\alpha)(\hat{g} - g^*) \\ &\leq (\widetilde{P}_n - \widetilde{P})(\ell_\alpha(g^*) - \ell_\alpha(\hat{g})) + a(\alpha) + r(R_\ell(\hat{g}) - R_\ell(g^*)). \end{aligned}$$

Hence, we have:

$$\delta_{j+1} \leq R_\ell(\hat{g}) - R_\ell(g^*) \leq \delta_j \Rightarrow \delta_{j+1} \leq \frac{1}{1-r}\left((\widetilde{P}_n - \widetilde{P})(\ell_\alpha(g^*) - \ell_\alpha(\hat{g})) + a(\alpha)\right).$$

On the event $E_{n,j}^\alpha(t)$, by definition of $\mathcal{G}(\delta_j)$, it follows that $\forall \delta \leq \delta_j$:

$$\begin{aligned} \delta_{j+1} \leq R_\ell(\hat{g}) - R_\ell(g^*) \leq \delta_j \Rightarrow \delta_{j+1} &\leq \frac{1}{1-r}U_n^\alpha(\delta_j, t) + \frac{1}{1-r}a(\alpha) \\ &\leq \frac{\delta_j}{1-r}V_n^\alpha(\delta, t) + \frac{1}{1-r}a(\alpha), \end{aligned}$$

where $V_n^\alpha(\delta,t)=\breve{U}_n^\alpha(\delta,t)$ satisfies (see Koltchinskii (2006)):

$$U_n^\alpha(\delta_j,t)\leq\delta_j V_n^\alpha(\delta,t),\forall\delta\leq\delta_j.$$

We obtain:

$$\frac{1}{1-r}V_n^\alpha(\delta,t)\geq\frac{1}{q}-\frac{q^j}{1-r}a(\alpha)>\frac{1}{2q},$$

since we have:

$$a(\alpha)\leq\frac{1-r}{4q}\delta\Longrightarrow\frac{q^j}{1-r}a(\alpha)<\frac{1}{2q}.$$

It follows from the definition of the †-transform that:

$$\delta<[U_n^\alpha(\cdot,t)]^\dagger(\frac{1-r}{2q})=\delta_n^\alpha(t).$$

Hence, we have on the event $E_{n,j}^\alpha(t)$, for $\delta_j\geq\delta$:

$$\delta_{j+1}\leq R_\ell(\hat{g})-R_\ell(g^*)\leq\delta_j\Rightarrow\delta<\delta_n^\alpha(t),$$

or equivalently,

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

where $\mathcal{G}(c,C)=\{g\in\mathcal{G}:c\leq R_\ell(g)-R_\ell(g^*)\leq C\}$. Finally, we obtain:

$$\bigcap_{j:\delta_j\geq\delta}E_{n,j}^\alpha(t)\text{ and }\delta\geq\delta_n^\alpha(t)\Rightarrow R_\ell(\hat{g})-R_\ell(g^*)\leq\delta.$$

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

$$\mathbb{P}(R_\ell(\hat{g})-R_\ell(g^*)\geq\delta)\leq\sum_{\delta_j\geq\delta}\mathbb{P}(E_{n,j}^\alpha(t)^C)\leq\log_q\left(\frac{1}{\delta}\right)e^{-t},$$

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

Proof of Theorem 1. The proof is a direct application of Lemma 2. We have, for some constant $K>0$:

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

First step is to control $\phi_n^\alpha(\mathcal{G},\delta)$. Assumption (2.8) and the Bernstein condition allow us to write:

$$\begin{aligned}
\phi_n^\alpha(\mathcal{G},\delta) &\leq \mathbb{E}\sup_{g\in\mathcal{G}(\delta)}|\widetilde{P}_n-\widetilde{P}|[\ell_\alpha(g)-\ell_\alpha(g^*)]\\
&\leq \mathbb{E}\sup_{g\in\mathcal{G}:\|\ell(g)-\ell(g^*)\|_{L^2(\mu)}\leq 2\sqrt{\kappa_0}\delta^{\frac{1}{2\kappa}}}|\widetilde{P}_n-\widetilde{P}|[\ell_\alpha(g)-\ell_\alpha(g^*)]\\
&\leq \widetilde{\omega}_n(\mathcal{G},\sqrt{\kappa_0}\delta^{\frac{1}{2\kappa}})\\
&\leq C\frac{c(\alpha)}{\sqrt{n}}\delta^{\frac{1-\rho}{2\kappa}},
\end{aligned}$$

where we use in last line Lemma 1.

A control of $D^\alpha(\mathcal{G},\delta)$ using the Lipschitz assumption leads to:

$$U_n^\alpha(\delta,t) \leq C\left[\frac{c(\alpha)}{\sqrt{n}}\delta^{\frac{(1-\rho)}{2\kappa}} + \frac{\sqrt{c(\alpha)}}{n^{3/4}}\delta^{\frac{1-\rho}{4\kappa}}\sqrt{K(\alpha)t} + \sqrt{\frac{t}{n}}c(\alpha)\delta^{\frac{1}{2\kappa}} + \frac{t}{n}\right].$$

Hence we have from an easy calculation:

$$\delta_n^\alpha(t) \leq C\max\left(\left(\frac{c(\alpha)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}}, \frac{[c(\alpha)K(\alpha)]^{\frac{2\kappa}{4\kappa+\rho-1}}}{n^{\frac{3\kappa}{4\kappa+\rho-1}}}t^{\frac{2\kappa}{4\kappa+\rho-1}}, \left(\frac{c(\alpha)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa-1}}t^{\frac{2\kappa}{2\kappa-1}}, \frac{t}{n}\right).$$

Consequently, for any $0 < t \leq 1$, for n large enough, we have:

$$\left(\frac{c(\alpha)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}} \geq \delta_n^\alpha(t+\log\log_q n),$$

provided that:

$$K(\alpha) \leq \frac{c(\alpha)^{\frac{2\kappa}{2\kappa+\rho-1}}n^{\frac{\kappa+\rho-1}{2\kappa+\rho-1}}}{1+\log\log_q n}.$$

It remains to use Lemma 2 with t replaced by $t+\log\log_q n$ to obtain:

$$\mathbb{P}\left(R_\ell(\hat{g}_n^\alpha) - R_\ell(g^*) \geq K(1+t)\left(\frac{c(\alpha)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}}\right) \leq e^{-t},$$

provided that the approximation function obeys to the following inequality:

$$a(\alpha) \leq K\frac{(1-r)}{4q}\left(\frac{c(\alpha)}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}}. \qquad \square$$

5.2. *Proof of Theorem 2*

Theorem 2 is a straightforward application of Theorem 1 to the particular case of errors in variables using deconvolution kernel estimators.

First step is to check that the estimation procedure described in Section 3.1 gives rise to a LB-class with respect to $\nu_Y = \nu \otimes P_Y$, where ν is the Lebesgue measure on $\mathbb{R}^d$ and P_Y is the law of Y.

Lemma 3. *Suppose **(A1)** holds and suppose $l(g(\cdot),y) \in L_2(\mathcal{X})$ for any $y \in \mathcal{Y}$. Suppose $\|f_y * \eta\|_\infty \leq c_{max}$ for any $y \in \mathcal{Y}$. Consider a deconvolution kernel $\mathcal{K}_\eta(t) = \mathcal{F}^{-1}[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)}]$ where $\mathcal{K}(t) = \Pi_{i=1}^d \mathcal{K}_i(t_i)$ where $\mathcal{K}_i$ have compactly supported and bounded Fourier transform. Then we have:*

$$\|\ell_\lambda(g) - \ell_\lambda(g')\|_{L_2(\widetilde{P})} \leq C_L \Pi_{i=1}^d \lambda_i^{-\beta_i}\|\ell(g) - \ell(g')\|_{L_2(\nu_Y)},$$

and moreover:

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

where $C_L, C_B > 0$ are constants depending on $\mathcal{X}$, c_{max}, ℓ, η and $\mathcal{K}$.

Proof. We have in dimension $d = 1$ for simplicity, using the boundedness assumptions:

$$\begin{aligned}
&\|\ell_\lambda(g) - \ell_\lambda(g')\|^2_{L_2(\widetilde{P})} \\
= \quad & \sum_{y\in\mathcal{Y}} p_y \int_{\mathbb{R}^d} \left[\int_{\mathcal{X}} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) (\ell(g,(x,y))) - \ell(g',(x,y)))) dx \right]^2 f_y * \eta(z) dz \\
= \quad & \sum_{y\in\mathcal{Y}} p_y \int_{\mathbb{R}^d} \left[\frac{1}{\lambda} \mathcal{K}_\eta (\frac{\cdot}{\lambda}) * (\ell(g(\cdot),y) - \ell(g'(\cdot),y))(z) \right]^2 f_y * \eta(z) dz \\
\leq \quad & c_{\max} \sum_{y\in\mathcal{Y}} p_y \int_{\mathbb{R}^d} \frac{1}{\lambda^2} |\mathcal{F}[\mathcal{K}_\eta(\frac{\cdot}{\lambda})](t)|^2 |\mathcal{F}[\ell(g(\cdot),y) - \ell(g'(\cdot),y)](t)|^2 dt \\
\leq \quad & C_L \lambda^{-2\beta} \|\ell(g) - \ell(g')\|^2_{L_2(\nu_Y)},
\end{aligned}$$

where we use in last line the following inequalities:

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

provided that $\mathcal{F}[\mathcal{K}]$ is bounded and compactly supported.

By the same way, the second assertion holds since if $\ell(g(\cdot),y) \in L^2(\mathcal{X})$:

$$\begin{aligned}
|\ell_\lambda(g,(z,y))| \quad &\leq \quad \int_{\mathcal{X}} \left| \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) \ell(g,(x,y))) \right| dx \\
&\leq \quad C \sup_{z\in\mathcal{X}} \sqrt{ \int_{\mathcal{X}} \left| \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) \right|^2 dx } \\
&\leq \quad C \sup_{z\in\mathcal{X}} \sqrt{ \int_{\mathbb{R}^d} \left| \mathcal{F}[\mathcal{K}_\eta](\lambda t) \right|^2 dt } \\
&\leq \quad C_B \lambda^{-\beta-1/2},
\end{aligned}$$

provided that $\mathcal{F}[\mathcal{K}]$ is bounded and compactly supported.

A straightforward generalization leads to the d-dimensional case using assumption **(K1)**. □

The last step is to show that $\{\ell_\lambda(g), g \in \mathcal{G}\}$ satisfies Definition 3 with the following lemma.

Lemma 4. *Suppose **(R1)** holds for some $\gamma, L > 0$. Consider a deconvolution kernel $\mathcal{K}_\eta$ such that $\mathcal{K}$ is a kernel of order $\lfloor \gamma \rfloor$ with respect to the Lebesgue measure. Then if $\{\ell(g) - \ell(g^*), g \in \mathcal{G}\}$ is Bernstein with parameter $\kappa \geq 1$, we have:*

$$\forall g \in \mathcal{G}, (R^\lambda_\ell - R_\ell)(g - g^*) \leq a(\lambda) + r(R_\ell(g) - R_\ell(g^*)),$$

where

$$a(\lambda) = C\sum_{i=1}^{d} \lambda_i^{\frac{2\kappa\gamma}{2\kappa-1}} \text{ and } r = \frac{1}{2\kappa}.$$

Moreover, if $|\ell(g,(x,y)) - \ell(g',(x,y))| = |\ell(g,(x,y)) - \ell(g',(x,y))|^2$ *and* $\kappa > 1$, *we have:*

$$a(\lambda) = C\sum_{i=1}^{d} \lambda_i^{\frac{\kappa\gamma}{\kappa-1}} \text{ and } r = \frac{1}{\kappa}.$$

Proof. We consider the case $d = 1$ for simplicity. The first step is to compile the value of $R_\ell^\lambda(g)$ for a fixed $g \in \mathcal{G}$.Using the elementary property $\mathbb{E}K_\eta(\frac{Z-x}{\lambda}) = \mathbb{E}K(\frac{X-x}{\lambda})$, we can write:

$$\begin{aligned} R_\ell^\lambda(g) &= \mathbb{E}_{\widetilde{P}^{\otimes n}} R_n^\lambda(g) \\ &= \mathbb{E}_{P_Y^{\otimes n}} \sum_{y\in\mathcal{Y}} \hat{p}_y \int_{\mathcal{X}} \ell(g,(x,y))\mathbb{E}_{Z|Y=y}\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z^y - x}{\lambda}\right)dx \\ &= \sum p_y \int_{\mathcal{X}} \ell(g,(x,y))\mathbb{E}_{X|Y=y}\frac{1}{\lambda}\mathcal{K}\left(\frac{X^y - x}{\lambda}\right)dx \end{aligned}$$

Gathering with Fubini, we arrive at:

$$\begin{aligned} & (R_\ell^\lambda - R_\ell)(g - g^*) \\ = & \sum_{y\in\mathcal{Y}} p_y \int_{\mathcal{X}}\int_{\mathbb{R}} K(u)(\ell(g,(x,y)) - \ell(g^*,(x,y)))\left(f_y(x+\lambda u) - f_y(x)\right)dudx. \end{aligned}$$

Now since the f_y's has $\lfloor\gamma\rfloor$ derivatives, there exists $\tau \in]0,1[$ such that:

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

where we use in last line the Hölder regularity of the f_y's and that $\mathcal{K}$ is a kernel of order $\lfloor\gamma\rfloor$.

Using succesively Cauchy-Schwarz twice and the Bernstein assumption, one gets since $\mathcal{X}$ is compact:

$$\begin{aligned}
\left|(R_\ell^\lambda - R_\ell)(g-g^*)\right| &\leq C\lambda^\gamma \sum_{y\in\mathcal{Y}} p_y \int_{\mathcal{X}} |\ell(g,(x,y)) - \ell(g^*,(x,y))|dx \\
&\leq C\lambda^\gamma \sqrt{\sum_{y\in\mathcal{Y}} p_y \left(\int_{\mathcal{X}} |\ell(g,(x,y)) - \ell(g^*,(x,y))|dx\right)^2} \\
&\leq C\lambda^\gamma \sqrt{\sum_{y\in\mathcal{Y}} p_y \left(\int_{\mathcal{X}} |\ell(g,(x,y)) - \ell(g^*,(x,y))|^2 dx\right)} \\
&= C\|\ell(g) - \ell(g^*)\|_{L_2(\nu_Y)}\lambda^\gamma \\
&\leq C\lambda^\gamma \left(R_\ell(g) - R_\ell(g^*)\right)^{\frac{1}{2\kappa}} \\
&\leq C\lambda^{\frac{2\kappa\gamma}{2\kappa-1}} + \frac{1}{2\kappa}\left(R_\ell(g) - R_\ell(g^*)\right),
\end{aligned}$$

where we use in last line Young's inequality:

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

with $r = \frac{1}{2\kappa}$.

For the second statement, if $|\ell(g,(x,y)) - \ell(g',(x,y))| = |\ell(g,(x,y)) - \ell(g',(x,y))|^2$ and $\kappa > 1$, it is straightforward that 2κ can be replaced by κ to get the result. □

Proof of Theorem 2. The proof is a straightforward application of Theorem 1. From Lemma 3 and Lemma 4, condition (2.7) in Theorem 1 can be written:

$$\sum_{i=1}^d \lambda_i^{\frac{2\kappa\gamma}{2\kappa-1}} \lesssim \left(\frac{\Pi_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}} \Leftrightarrow \forall i = 1,\dots,d\ \lambda_i \lesssim n^{-\frac{2\kappa-1}{2\gamma(2\kappa+\rho-1)+2(2\kappa-1)\beta}}.$$

Applying Theorem 1 with a smoothing parameter λ such that equalities hold above gives the rates of convergence. □

5.3. *Proof of Theorem 3*

First step is to check that the estimation procedure described in Section 3.2 gives rise to a LB-class with respect to ν_Y with the following lemma.

Lemma 5. *Suppose **(A2)** holds and $l(g(\cdot),y) \in L_2(\nu)$ for any $y \in \mathcal{Y}$. Suppose $\|Af_y\|_\infty \leq c_{\max}$ for any $y \in \mathcal{Y}$. Then we have:*

$$\|\ell_\lambda(g) - \ell_\lambda(g')\|_{L_2(\widetilde{P})} \leq C_L N^\beta \|\ell(g) - \ell(g')\|_{L_2(\nu_Y)},$$

and moreover:

$$\sup_{g\in\mathcal{G}} \|\ell_\lambda(g)\|_\infty \leq C_B N^{\beta+1/2},$$

where $C_B, C_L > 0$ are constants depending on $c_{\max}$, ℓ and η.

Proof. The proof follows the proof of Lemma 3. We have in dimension $d = 1$ for simplicity since $(\phi_k)_{k\in\mathbb{N}^*}$ is an orthonormal basis and using the boundedness assumptions over the f_y's:

$$\begin{aligned}
&\|\ell_N(g) - \ell_N(g')\|^2_{L_2(\widetilde{P})} \\
= \quad & \sum_{y\in\mathcal{Y}} p_y \int_{\mathbb{R}^\lceil} \left(\sum_{k=1}^N b_k^{-1} \int_{\mathcal{X}} \phi_k(x)(\ell(g,(x,y))) \right. \\
& \qquad\qquad \left. - \ell(g',(x,y))))\nu(dx)\phi_k(z) \right)^2 Af_y(z)\nu(dz) \\
\leq \quad & c_{\max} \sum_{y\in\mathcal{Y}} p_y \sum_{k=1}^N b_k^{-2} \int_{\mathbb{R}^d} \left(\int_{\mathcal{X}} (\ell(g,(x,y))) \right. \\
& \qquad\qquad \left. - \ell(g',(x,y))))\phi_k(x)\nu(dx) \right)^2 \phi_k(z)^2\nu(dz) \\
\leq \quad & CN^{2\beta} \sum_{y\in\mathcal{Y}} p_y \sum_{k=1}^N \left(\int_{\mathcal{X}} (\ell(g,(x,y))) - \ell(g',(x,y))))\phi_k(x)\nu(dx) \right)^2 \\
\leq \quad & CN^{2\beta} \|\ell(g) - \ell(g')\|^2_{L_2(\nu_Y)}.
\end{aligned}$$

Moreover, the second assertion holds since if $l(g(\cdot), y) \in L_2(\nu)$:

$$\begin{aligned}
|\ell_N(g,(z,y))| \quad & \leq \quad \left| \sum_{k=1}^N b_k^{-1} \int_{\mathcal{X}} \phi_k(x)\phi_k(z)\ell(g,(x,y))\nu(dx) \right| \\
& \leq \quad \sqrt{\sum_{k=1}^N b_k^{-2}} \sqrt{\sum_{k=1}^N \left(\int \phi_k(x)\ell(g,(x,y))\nu(dx) \right)^2 \phi_k(z)^2} \\
& \leq \quad CN^{\beta+1/2}.
\end{aligned}$$

□

The last step is to control the bias term of the procedure with the following lemma:

Lemma 6. *Suppose* ***(R2)*** *holds and* $\{\ell(g) - \ell(g^*), g \in \mathcal{G}\}$ *is Bernstein with parameter* $\kappa \geq 1$. *Then we have:*

$$\forall g \in \mathcal{G}, (R_\ell^\lambda - R_\ell)(g - g^*) \leq a(\lambda) + r(R_\ell(g) - R_\ell(g^*)),$$

where

$$a(N) = C \sum_{i=1}^d N_i^{-\frac{2\kappa}{2\kappa-1}(\gamma-1/2)} \text{ and } r = \frac{1}{2\kappa}.$$

Moreover, if $|\ell(g,(x,y)) - \ell(g',(x,y))| = |\ell(g,(x,y)) - \ell(g',(x,y))|^2$ *and* $\kappa > 1$, *we have:*

$$a(N) = C \sum_{i=1}^d N_i^{-\frac{\kappa}{\kappa-1}(\gamma-1/2)} \text{ and } r = \frac{1}{\kappa}.$$

Proof. The first step is to compile the value of $R_\ell^N(g)$, for some fixed $g \in \mathcal{G}$. Noting that $\hat{\theta}_k^y$ is an unbiased estimator of θ_k^y, we have by a simple calculation:

$$\begin{aligned} R_\ell^N(g) = \mathbb{E}_{\widetilde{P}^{\otimes n}} R_n^N(g) &= \mathbb{E} \sum_{y\in\mathcal{Y}} \hat{p}_y \int_{\mathcal{X}} \ell(g,(x,y)) \sum_{k=1}^N \hat{\theta}_k^y \phi_k(x)\nu(dx) \\ &= \mathbb{E}_{P_Y^{\otimes n}} \sum_{y\in\mathcal{Y}} \hat{p}_y \int_{\mathcal{X}} \ell(g,(x,y)) \sum_{k=1}^N \mathbb{E}_{Z|Y=y} \hat{\theta}_k^y \phi_k(x)\nu(dx) \\ &= \sum_{y\in\mathcal{Y}} p_y \int_{\mathcal{X}} \ell(g,(x,y)) \sum_{k=1}^N \theta_k^y \phi_k(x)\nu(dx), \end{aligned}$$

where $\theta_k^y = \int f_y \phi_k d\nu$. We hence get:

$$\begin{aligned} &(R_\ell^\lambda - R_\ell)(g-g^*) \\ = \quad & \sum_{y\in\mathcal{Y}} p_y \int_{\mathcal{X}} (\ell(g,(x,y)) - \ell(g^*,(x,y))) \left(\sum_{k=1}^N \theta_k^y \phi_k(x) - \sum_{k\geq 1} \theta_k^y \phi_k(x) \right) \nu(dx) \\ = \quad & \sum_{y\in\mathcal{Y}} p_y \int_{\mathcal{X}} (\ell(g^*,(x,y)) - \ell(g,(x,y))) \sum_{k>N} \theta_k^y \phi_k(x)\nu(dx). \end{aligned}$$

Using Cauchy-Schwarz, we have since $(\phi_k)_{k\in\mathbb{N}}$ in an orthonormal basis:

$$\begin{aligned} &|(R_\ell^\lambda - R_\ell)(g-g^*)| \\ \leq \quad & \sum_{y\in\mathcal{Y}} p_y \sum_{k>N} \theta_k^y \sqrt{\int_{\mathcal{X}} (\ell(g^*,(x,y)) - \ell(g,(x,y)))^2 \nu(dx)} \sqrt{\int_{\mathcal{X}} \phi_k(x)^2 \nu(dx)} \\ = \quad & \sum_{y\in\mathcal{Y}} p_y \sum_{k>N} \theta_k^y \sqrt{\int_{\mathcal{X}} (\ell(g^*,(x,y)) - \ell(g,(x,y)))^2 \nu(dx)}. \end{aligned}$$

Using again Cauchy-Schwarz, gathering with **(R2)**, we arrive at:

$$\begin{aligned} &|(R_\ell^\lambda - R_\ell)(g-g^*)| \\ \leq \quad & \sqrt{\sum_{y\in\mathcal{Y}} p_y \int_{\mathcal{X}} (\ell(g,(x,y)) - \ell(g',(x,y)))^2 \nu(dx)} \sqrt{\sum_{y\in\mathcal{Y}} p_y \left(\sum_{k>N} \theta_k^y \right)^2} \\ \leq \quad & C \|\ell(g) - \ell(g')\|_{L_2(\nu_Y)} N^{-\gamma} \sqrt{\sum_{y\in\mathcal{Y}} p_y \sum_{k>N} (\theta_k^y)^2 k^{2\gamma+1}} \\ \leq \quad & C \left(R_\ell(g) - R_\ell(g') \right)^{\frac{1}{2\kappa}} N^{-\gamma}. \end{aligned}$$

We conclude the proof using Young's inequality exactly as in Lemma 4. □

Proof of Theorem 3. The proof is a straightforward application of Theorem 1. From Lemma 5 and Lemma 6, condition (2.7) in Theorem 1 can be written:

$$N^{\frac{-2\kappa\gamma}{2\kappa-1}} \lesssim \left(\frac{N^{\beta}}{\sqrt{n}}\right)^{\frac{2\kappa}{2\kappa+\rho-1}} \Leftrightarrow N \lesssim n^{\frac{2\kappa-1}{2\gamma(2\kappa+\rho-1)+2(2\kappa-1)\beta}}.$$

Applying Theorem 1 with a smoothing parameter N such thatan equality holds above gives the rates of convergence. □

5.4. Proof of Lemma 1

The proof uses the maximal inequality presented in van der Vaart and Wellner (1996) to the class:

$$\mathcal{F} = \{\ell_\alpha(g) - \ell_\alpha(g'),\, g, g' \in \mathcal{G} : P(\ell(g) - \ell(g'))^2 \leq \delta^2\}.$$

Indeed from Theorem 2.14.2 of van der Vaart and Wellner (1996), we can write, $\forall \eta > 0$:

$$\begin{aligned} \widetilde{\omega}_n(\mathcal{G}, \delta, \mu) &= \mathbb{E} \sup_{g,g' \in \mathcal{G}: \|\ell(g)-\ell(g')\|^2_{L_2(\mu)} \leq \delta^2} \left| (\widetilde{P}_n - \widetilde{P})(\ell_\alpha(g) - \ell_\alpha(g')) \right| \\ &\leq \frac{\|F\|^2_{L_2(\widetilde{P})}}{\sqrt{n}} \int_0^\eta \sqrt{1 + \mathcal{H}_B(\mathcal{F}, \epsilon\|F\|^2_{L_2(\widetilde{P})}, L_2(\mu))} d\epsilon \\ &+ \frac{\sup_{f\in\mathcal{F}} \|f\|_{L_2(\widetilde{P})}}{\sqrt{n}} \sqrt{1 + \mathcal{H}_B(\mathcal{F}, \eta\|F\|^2_{L_2(\widetilde{P})}, L_2(\mu))} \end{aligned} \tag{5.1}$$

where $F(z, y) = \sup_{f\in\mathcal{F}} |\ell_\alpha(g, z, y) - \ell_\alpha(g', z, y)|$ is the enveloppe function of the class $\mathcal{F}$. Since $\{\ell_\alpha(g), g \in \mathcal{G}\}$ is a LB-class with bounded constant $K(\alpha)$:

$$\begin{aligned} \|F\|^2_{L_2(\widetilde{P})} &= \int F^2(z, y)\widetilde{P}(dz, dy) \\ &= \sum_{y\in\mathcal{Y}} p_y \int \left(\sup_{f\in\mathcal{F}} |\ell_\alpha(g, z, y) - \ell_\alpha(g', z, y)| \right)^2 A f_y(z)\nu(dz) \\ &\leq C_B^2 K(\alpha)^2. \end{aligned}$$

Moreover, we have since $\{\ell_\alpha(g), g \in \mathcal{G}\}$ is a LB-class with respect to μ with Lipschitz constant $c(\alpha)$:

$$\mathcal{H}_B(\{\ell(g),\, g \in \mathcal{G}\}, \epsilon, L_2(\mu)) \leq c\epsilon^{-2\rho} \Rightarrow \mathcal{H}_B(\mathcal{F}, \epsilon, L_2(\widetilde{P})) \leq Cc(\alpha)^{2\rho}\epsilon^{-2\rho}.$$

Hence, we have in (5.1), choosing $\eta = \frac{c(\alpha)}{K(\alpha)^2}\delta$:

$$\begin{aligned} \widetilde{\omega}_n(\mathcal{G},\delta) &\le C\left[\frac{K(\alpha)^2}{\sqrt{n}}\int_0^\eta \sqrt{1+\epsilon^{-2\rho}K(\alpha)^{-4\rho}c(\alpha)^{2\rho}}\,d\epsilon \right.\\ &\qquad \left. + \frac{c(\alpha)\delta}{\sqrt{n}}\sqrt{1+\eta^{-2\rho}K(\alpha)^{-4\rho}c(\alpha)^{2\rho}}\right]\\ &\le C\left[\frac{\eta K(\alpha)^2}{\sqrt{n}} + \frac{\eta^{1-\rho}K(\alpha)^{2(1-\rho)}c(\alpha)^\rho}{\sqrt{n}} + \frac{c(\alpha)\delta}{\sqrt{n}}\right.\\ &\qquad \left. + \frac{c(\alpha)^{1+\rho}\eta^{-\rho}K(\alpha)^{-2\rho}\delta}{\sqrt{n}}\right]\\ &\le C\left[\frac{\eta^{1-\rho}K(\alpha)^{2(1-\rho)}c(\alpha)^\rho}{\sqrt{n}} + \frac{c(\alpha)^{1+\rho}\eta^{-\rho}K(\alpha)^{-2\rho}\delta}{\sqrt{n}}\right]\\ &\le C\frac{c(\alpha)}{\sqrt{n}}\delta^{1-\rho}, \end{aligned}$$

provided that $\delta \le 1$.

References

Audibert, J.-Y. and Tsybakov, A.B. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35: 608–633, 2007. MR2336861

Bartlett, P.L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005. MR2166554

Bartlett, P.L. and Mendelson, S. Empirical minimization. *Probability Theory and Related Fields*, 135(3): 311–334, 2006. MR2240689

Blanchard, G., Bousquet, O., and Massart, P. Statistical performance of support vector machines. *The Annals of Statistics*, 36(2): 489–531, 2008. MR2396805

Bousquet, O. A bennet concentration inequality and its application to suprema of empirical processes. *C. R. Acad. SCI. Paris Ser. I Math*, 334: 495–500, 2002. MR1890640

Butucea, C. Goodness-of-fit testing and quadratic functionnal estimation from indirect observations. *The Annals of Statistics*, 35: 1907–1930, 2007. MR2363957

Cavalier, L. Nonparametric statistical inverse problems. *Inverse Problems*, 24: 1–19, 2008. MR2421941

Delaigle, A., Hall, P., and Meister, A. On deconvolution with repeated measurements. *The Annals of Statistics*, 36(2): 665–685, 2008. MR2396811

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996. MR1383093

Engl, H.W., Hank, M., and Neubauer, A. *Regularization of Inverse Problems*. Kluwer Academic Publishers Group, Dordrecht, 1996. MR1408680

Fan, J. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19: 1257–1272, 1991. MR1126324

Koltchinskii, V. Local rademacher complexities and oracle inequalties in risk minimization. *The Annals of Statistics*, 34(6): 2593–2656, 2006. MR2329442

Lecué, G. and Mendelson, S. General non-exact oracle inequalities for classes with a subexponential envelope. *The Annals of Statistics*, 40(2): 832–860, 2012. MR2933668

Lederer, Y. and van de Geer, S. New concentration inequalities for suprema of empirical processes. Submitted, 2012.

Loustau, S. Penalized empirical risk minimization over Besov spaces. *Electronic Journal of Statistics*, 3: 824–850, 2009. MR2534203

Loustau, S. Fast rates for noisy clustering. `http://hal.archives-ouvertes.fr/hal-00695258`, 2012.

Loustau, S. and Marteau, C. Discriminant analysis with errors in variables. `http://hal.archives-ouvertes.fr/hal-00660383`, 2012.

Loustau, S. and Marteau, C. Minimax fast rates in discriminant analysis with errors in variables. *In revision to Bernoulli*, 2013.

Mammen, E. and Tsybakov, A.B. Smooth discrimination analysis. *The Annals of Statistics*, 27(6): 1808–1829, 1999. MR1765618

Massart, P. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9(2): 245–303, 2000. MR1813803

Massart, P. and Nédélec, E. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5): 2326–2366, 2006. MR2291502

Meister, A. *Deconvolution Problems in Nonparametric Statistics*. Springer-Verlag, 2009. MR2768576

Polonik, W. Measuring mass concentrations and estimating density contour clusters – An excess mass approach. *The Annals of Statistics*, 23(3): 855–881, 1995. MR1345204

Talagrand, M. New concentration inequalities in product spaces. *Invent. Math*, 126: 505–563, 1996. MR1419006

Tsybakov, A.B. *Introduction à l'estimation non-paramétrique*. Springer-Verlag, 2004a. MR2013911

Tsybakov, A.B. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1): 135–166, 2004b. MR2051002

Tsybakov, A.B. and van de Geer, S. Square root penalty: Adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, 33(3): 1203–1224, 2005. MR2195633

van de Geer, S. *Empirical Processes in M-estimation*. Cambridge University Press, 2000. MR1739079

van der Vaart, A.W. and Wellner, J.A. *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Verlag, 1996. MR1385671

Vapnik, V. *Estimation of Dependances Based on Empirical Data*. Springer Verlag, 1982. MR0672244

Vapnik, V. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science, Springer, 2000. MR1719582

Williamson, R.C., Smola, A.J., and Schölkopf, B. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6): 2516–2532, 2001. MR1873936