



**HAL**  
open science

## Statistical learning with indirect observations

Sébastien Loustau

► **To cite this version:**

| Sébastien Loustau. Statistical learning with indirect observations. 2012. hal-00664125v2

**HAL Id: hal-00664125**

**<https://hal.science/hal-00664125v2>**

Preprint submitted on 15 Feb 2012 (v2), last revised 10 Jul 2012 (v3)

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

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

# Statistical learning with indirect observations

Sébastien Loustau,

Université d'Angers, LAREMA  
 loustau@math.univ-angers.fr

**Abstract:** Given a random couple  $(X, Y)$  with unknown distribution  $P$ , the problem of statistical learning consists in the estimation of the Bayes  $g^* = \arg \min_{\mathcal{G}} \mathbb{E}_P l(g(X), Y)$ , where  $\mathcal{G}$  is a class of candidate functions and  $l$  is a loss function. In this paper we address this problem when we have at our disposal a corrupted sample  $\mathcal{D}_n = \{(Z_1, Y_1), \dots, (Z_n, Y_n)\}$  of i.i.d. indirect observations. It means that the inputs  $Z_i, i = 1, \dots, n$  are distributed from the density  $Af$ , where  $A$  is a known compact linear operator and  $f$  is the density of the direct input  $X$ .

## 1. Introduction

In this paper we consider the problem of learning from an indirect set of observations. The model can be described through 4 components:

- A generator **G** of random variables  $X \in \mathcal{X} \subseteq \mathbb{R}^d$  with unknown density  $f$  with respect to  $\nu$ , a  $\sigma$ -finite measure defined on  $\mathcal{X}$ ,
- A supervisor **S** who associates to  $X$  an output  $Y \in \mathcal{Y}$ , according to an unknown conditional probability,
- A known linear compact operator **A**:  $L_2(\nu, \mathcal{X}) \rightarrow L_2(\nu, \tilde{\mathcal{X}})$  which corrupts  $X$  given  $Z$  where  $Z$  has density  $Af$  with respect to  $\nu$ ,
- A learning Machine **LM** which given  $n$  i.i.d. observations  $(Z_i, Y_i), i = 1 \dots n$ , returns an estimator  $\hat{y}$  associated to any given  $x$  from the generator.

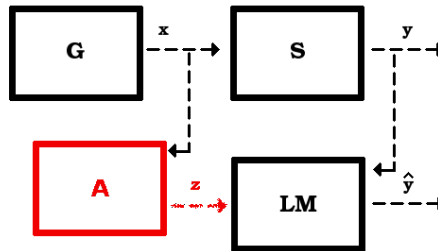


Figure 1. This representation has its origin in Vapnik [2000]. Here the presence of the nuisance operator  $A$  makes the problem an inverse problem.

The goal is to design a decision rule returning, for each new generator's value  $x$ , a value  $\hat{y}$  as close as possible to the supervisor's response  $y$ . Note that depending on the nature of the supervisor, this problem contains models of classification, density estimation, and regression. For example if operator  $A$  is a convolution product, we are faced to density deconvolution, classification with errors in variables or regression with errors in variables.

For this purpose, we introduce a loss function  $l : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}_+$  and a class  $\mathcal{G}$  of measurable functions  $g : \mathcal{X} \rightarrow \mathbb{R}$ . To define the best approximation, the problem is that of choosing from the given set of functions  $g \in \mathcal{G}$ , the one that minimizes the risk functional:

$$R_l(g) = \mathbb{E}l(g(X), Y), \quad (1.1)$$

where the expectation is taken over the joint distribution of  $(X, Y)$  denoted by  $P$ . The performances of a given  $g$  in measured through its non-negative excess risk, given by:

$$R_l(g) - R_l(g^*), \quad (1.2)$$

where  $g^*$  is the minimizer over  $\mathcal{G}$  of the risk (1.1). It is important to note that we do not address in this paper the problem of model selection of  $\mathcal{G}$  which consists in studying the difference  $R_l(g^*) - \inf_g R_l(g)$ , 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 ERM strategies based on a corrupted sample to minimize the excess risk (1.2).

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

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

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

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

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 the excess risk (1.2). To be concise, under complexity assumptions over  $\mathcal{G}$  (such as finite VC dimension ([Vapnik \[1982\]](#)), entropy conditions ([Van De Geer \[2000\]](#)), Rademacher complexity assumptions ([Koltchinskii \[2006\]](#)) and assumptions over the loss  $l$ , 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 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:

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

It is important to note that (1.4) 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 these important refinement in this introduction for the sack of concision whereas it is the main ingredient of the literature cited above. It allows to get fast rates of convergence in pattern recognition.

Here the framework is essentially different since 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.  $Af$  with  $A$  a known linear compact operator. 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 corrupted sample  $(Z_1, Y_1), \dots, (Z_n, Y_n)$  in standard ERM (1.3) fails since:

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

Due to the action of  $A$ , and provided that  $A \neq I$ , the empirical measure from the indirect sample, defined as  $\tilde{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{(Z_i, Y_i)}$  differs from  $P_n$  and we are faced to an ill-posed inverse problem. Note that this problem has been recently considered in [Klemela and Mammen \[2010\]](#) where  $L_2$ -ERM type estimators are proposed in the particular gaussian white noise model and in density estimation (see also [Butucea and Taupin \[2008\]](#) in a semi-parametric model of regression with errors).

In this paper, we propose to adopt a comparable strategy in statistical learning. Given a smoothing parameter  $\lambda = (\lambda_1, \dots, \lambda_d) \in \mathbb{R}_+^d$ , we propose to consider the following  $\lambda$ -Empirical Risk Minimization:

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

where  $R_n^\lambda(g)$  is called the  $\lambda$ -Empirical risk and is defined in a general way as:

$$R_n^\lambda(g) = \int_{\mathcal{X}} l(g(x), y) \hat{P}_\lambda(dx, dy). \quad (1.6)$$

Here  $\hat{P}_\lambda = \hat{P}_\lambda(Z_1, Y_1, \dots, Z_n, Y_n)$  is an estimator of the joint distribution  $P$  using the set of indirect inputs  $(Z_1, \dots, Z_n)$ . It will be related with standard regularization methods coming from the inverse problem literature (see [Engl et al. \[1996\]](#)) and as a consequence depends on a smoothing parameter  $\lambda \in \mathbb{R}_+^d$ . An explicit construction of  $\hat{P}_\lambda$  and the empirical risk (1.6) is detailed in Section 2 in pattern recognition with applications in Section 3.

To study the performances of the minimizer 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 Weelner [1996] or more recently Koltchinskii [2006]. Following the pioneering's work of Vapnik, we can write, in the presence of indirect observations, for  $\hat{g}_n^\lambda$  a solution of (1.5):

$$\begin{aligned} R_l(\hat{g}_n^\lambda) - R_l(g^*) &\leq R_l(\hat{g}_n^\lambda) - R_n^\lambda(\hat{g}_n^\lambda) + R_n^\lambda(g^*) - R_l(g^*) \\ &\leq R_l^\lambda(\hat{g}_n^\lambda) - R_n^\lambda(\hat{g}_n^\lambda) + R_n^\lambda(g^*) - R_l^\lambda(g^*) + (R_l - R_l^\lambda)(\hat{g}_n^\lambda - g^*) \\ &\leq \sup_{g \in \mathcal{G}} |(R_n^\lambda - R_l^\lambda)(g^* - g)| + \sup_{g \in \mathcal{G}} |(R_l^\lambda - R_l)(g - g^*)|, \end{aligned} \quad (1.7)$$

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

$$R_l^\lambda(g) = \mathbb{E}R_n^\lambda(g) = \int l(g(x), y) \mathbb{E}\hat{P}_\lambda(dx, dy). \quad (1.8)$$

Bounds (1.7) are called Inverse Vapnik's bounds. 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  $g^*$  using an empirical counterpart. This term can be controled thanks to uniform concentration inequalities such as Talagrand's type inequality, applied to a class of functions depending on a parameter.
- A bias term  $\sup_{g \in \mathcal{G}} |(R_l^\lambda - R_l)(g - g^*)|$ . It comes from the estimation of  $P$  into the expression of  $R_l(g)$  with estimator  $\hat{P}_\lambda$ . This term is specific to our method. However, it seems to be related to the usual bias term in nonparametric density estimation since we can see coarsely that:

$$R_l^\lambda(g) - R_l(g) = \int l(g(x), y) [\mathbb{E}\hat{P}_\lambda - P_\lambda](dx, dy).$$

The choice of  $\lambda$  is crucial in the decomposition (1.7). We will show below that the variance term explodes when  $\lambda$  tends to zero whereas the bias term vanishes. Parameter  $\lambda$  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.

In this work we restrict ourself to classification where  $\mathcal{Y} = \{0, 1, \dots, M\}$ . In other words, we consider the problem of pattern recognition with indirect observations, as illustrated in Figure 2 (see Devroye et al. [1996] for a survey in the direct case).

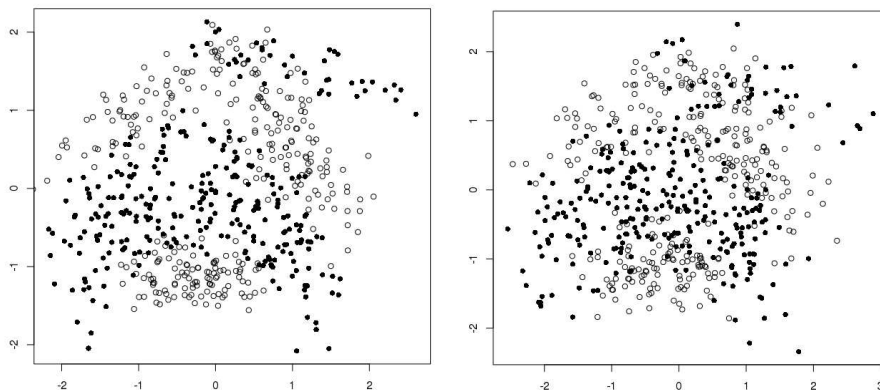


Figure 2. Representation of a binary classification sample "banana" (left) and a noisy version (right).

The paper is organized as follows. In Section 2, we propose to give an explicit construction of the empirical risk (1.6) in classification, thanks to the set of indirect observations. We state a general upper bound for the solution of the  $\lambda$ -Empirical Risk Minimization (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: (1) the errors-in-variables case where operator  $A$  is a convolution product, generalizing the results of Loustau and Marteau [2011]; (2) the general case using projections into the SVD of operator  $A$ . Rates of convergence are proposed which generalize the existing fast rates of convergence in classification stated in Koltchinskii [2006] and coincide with recent lower bounds proposed in discriminant analysis by Loustau and Marteau [2011]. Section 4 is devoted to a discussion related to the complexity assumption when we deal with indirect observations whereas Section 5 concludes the paper. Section 6 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 and give minimal assumptions to control the excess risk (1.2) of the procedure. The construction of the empirical risk is based on the following decomposition of the true risk:

$$R_l(g) = \sum_{y \in \mathcal{Y}} p(y) \int_{\mathcal{X}} l(g(x), y) f_y(x) \nu(dx), \quad (2.1)$$

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 propose to estimate each

conditional density  $f_y(\cdot)$  using the set of indirect observations  $Z_i$   $i = 1, \dots, n$ , thanks to a nonparametric estimator with smoothing parameter  $\lambda \in \mathbb{R}_+^d$ . To state a general upper bound, we consider a family of estimators such that:

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

where  $n_y = \text{card}\{i : Y_i = y\}$ ,  $k_\lambda : \tilde{\mathcal{X}} \times \mathcal{X} \rightarrow \mathbb{R}$  and the set of inputs  $(Z_i^y)_{i=1}^{n_y} = \{Z_i, i = 1, \dots, n : Y_i = y\}$ .

Here we consider a constant bandwidth  $\lambda$  for any  $y \in \mathcal{Y}$  in  $\hat{f}_y$ . It illustrates rather well the difference of our approach with plug-in type estimators (see [Audibert and Tsybakov \[2007\]](#) for instance). If we want to estimate  $f_y$ , for each  $y \in \mathcal{Y}$ , the bandwidth  $\lambda$  in (2.2) has to depend on  $n_y$  and the regularity of  $f_y$ . However here the aim is to estimate the true risk  $R_l(g)$  and to get satisfying upper bounds, we will see that  $\lambda$  does not necessary depend on the value  $y \in \mathcal{Y}$ . It is also important to note that assumption (2.2) provides a variety of nonparametric estimators of  $f_y$ . For instance if  $Af = f * \eta$  is a convolution operator, we can construct a deconvolution kernel provided that the noise has a nonnull Fourier transform. This is rather classical in deconvolution problems (see [Fan \[1991\]](#) or [Meister \[2009\]](#)). Another classical example of (2.2) is to consider projection estimators of the conditionnal densities using the SVD of operator  $A$  or any other regularizations methods (see [Engl et al. \[1996\]](#)). Section 3 gives some examples.

Finally given  $\hat{f}_y$  satisfying (2.2), we plug these estimators in the true risk (2.1) to get an empirical risk defined as:

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

where  $\hat{p}(y) = \frac{n_y}{n}$  is an estimator of the quantity  $p(y) = \mathbb{P}(Y = y)$ . Thanks to (2.2), the empirical risk in(1.6) can be written:

$$R_n^\lambda(g) = \frac{1}{n} \sum_{i=1}^n l_\lambda(g, (Z_i, Y_i)), \quad (2.3)$$

where  $l_\lambda(g, (z, y))$  is a modified version of  $l(g(x), y)$  given by:

$$l_\lambda(g, (z, y)) = \int_{\mathcal{X}} l(g(x), y) k_\lambda(z, x) \nu(dx).$$

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

$$\hat{g}_n^\lambda := \arg \min \frac{1}{n} \sum_{i=1}^n l_\lambda(g, (Z_i, Y_i)). \quad (2.4)$$

Note that in case no such minimum exists, we can consider a  $\delta$ -approximate minimizer as in [Bartlett and Mendelson \[2006\]](#) without significant change in

the results.

The main idea in the proof is to use iterate a Talagrand's type inequality due to Bousquet [2002]. It allows to control the increments of the empirical process:

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

Here it is important to note that Talagrand's inequality has to be applied to the class of functions  $\{(z, y) \mapsto l_\lambda(g, (z, y)), g \in \mathcal{G}\}$ . This class depends on a regularization parameter  $\lambda$ . This parameter will be calibrated as a function of  $n$  and that's why Talagrand's inequality has to be used in a careful way. For this purpose, we introduce in Definition 1 particular classes  $\{l_\lambda(g), g \in \mathcal{G}\}$  lipschitz with constant  $c(\lambda)$  and bounded by a constant  $K(\lambda)$ .

### 2.1. The result

In the sequel, with a slight abuse of notations, we write  $l_\lambda(g)$  for  $(z, y) \mapsto l_\lambda(g, (z, y))$ .

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

(**L** $_\mu$ )  $\{l_\lambda(g), g \in \mathcal{G}\}$  is lipschitz w.r.t.  $\mu$  with constant  $c(\lambda)$ :

$$\forall g, g' \in \mathcal{G}, \|l_\lambda(g) - l_\lambda(g')\|_{L_2(\tilde{P})} \leq c(\lambda) \|l(g) - l(g')\|_{L_2(\mu)},$$

for  $\mu = \nu \otimes P_Y$  or  $\mu = P$ .

(**B**)  $\{l_\lambda(g), g \in \mathcal{G}\}$  is uniformly bounded with constant  $K(\lambda)$ :

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

A LB-class of loss function is lipschitz and bounded with constants depending on  $\lambda$ . This properties are necessary to control the variance in (1.7) using Talagrand's type inequality to the class  $\{l_\lambda(g), g \in \mathcal{G}\}$ .

The lipschitz property (**L** $_\mu$ ) with  $\mu = P$  is stronger than (**L** $_\mu$ ) with  $\mu = \nu \otimes P_Y$  since we have coarselly for any measurable  $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ , if the  $f_y$ 's are bounded:

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

for some positive constant  $C$ , where in the sequel  $\nu_Y = \nu \otimes P_Y$ .

Note that to state a general upper bound, this lipschitz property has to be combined with Definition 2 below to have the following statement:

$$\|f\|_{L_2(\tilde{P})} \leq c(\lambda) \|f\|_{L_2(\mu)} \leq c(\lambda) (\mathbb{E}_P f)^{\frac{1}{2\kappa}}, \quad (2.5)$$



applied to the class of functions  $\{f = l(g) - l(g^*), g \in \mathcal{G}\}$ . Hence the dependence on  $\mu$  in  $(\mathbf{L}_\mu)$  is related to Definition 2 below and the Bernstein property of the loss class  $\mathcal{F} = \{l(g) - l(g^*), g \in \mathcal{G}\}$ .

Moreover the lipschitz property is a key ingredient to control the complexity of the class of functions  $\{l_\lambda(g), g \in \mathcal{G}\}$ . In the sequel, we use the following geometric complexity parameter:

$$\tilde{\omega}_n(\mathcal{G}, \delta, \mu) := \mathbb{E} \sup_{g, g' \in \mathcal{G}: \|l(g) - l(g')\|_{L_2(\mu)} \leq \delta} \left| (\tilde{P} - \tilde{P}_n)(l_\lambda(g) - l_\lambda(g')) \right|,$$

where  $\mu = \nu_Y$  or  $\mu = P$  following the context. The control of such a quantity is proposed in Section 4.

Finally **(B)** is necessary to apply Talagrand's inequality to a class of functions depending on a smoothing parameter  $\lambda$ .

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

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

This notion of Bernstein class first appears in [Bartlett and Mendelson \[2006\]](#) for  $\mu = P$ . This assumption arises naturally in statistical learning when we want to apply a functional Bernstein inequality such as Talagrand's type inequality. Here the dependence in measure  $\mu \in \{\nu_Y, P\}$  gives rise to two different assumptions. Since  $\|\cdot\|_{L_2(P)} \leq \|\cdot\|_{L_2(\nu_Y)}$ , a Bernstein class with respect to  $\nu_Y$  is also Bernstein with respect to  $P$ . The most favorable case where  $\mu = \nu_Y$  arises in binary classification (see for instance [Tsybakov \[2004b\]](#) or [Massart and Nédélec \[2006\]](#)). Section 3 states rates of convergence in these two different settings.

To control the excess risk of the procedure, we also need to control the bias term defined in (1.7) as follows.

**Definition 3.** We said that the class  $\{l_\lambda(g), g \in \mathcal{G}\}$  has approximation function  $a(\lambda)$  and residual constant  $0 < r < 1$  if the following holds:

$$\forall g, g' \in \mathcal{G}, (R_l - R_l^\lambda)(g - g') \leq a(\lambda) + r(R_l(g) - R_l(g'))^2.$$

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

$$R_l(\hat{g}_n^\lambda) - R_l(g^*) \leq \frac{1}{1-r} \left( \sup_{g \in \mathcal{G}(1)} |(\tilde{P}_n - \tilde{P})(l_\lambda(g) - l_\lambda(g^*))| + a(\lambda) \right),$$

provided that  $\hat{g}_n^\lambda \in \mathcal{G}(1)$  where in the sequel  $\mathcal{G}(\delta) = \{g \in \mathcal{G} : R_l(g) - R_l(g^*) \leq \delta\}$ . Under regularity conditions over the conditional densities  $f_y$ , gathering with Definition 2 below, Section 3 proposes to give explicit function  $a(\lambda)$  and residual constant  $r < 1$  to get rates of convergence.

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

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

$$a(\lambda) \lesssim \left( \frac{c(\lambda)}{\sqrt{n}} \right)^{\frac{2\kappa}{2\kappa+\rho-1}} \vee \left( \frac{[c(\lambda)K(\lambda)]^{\frac{2\kappa}{4\kappa+\rho-1}}}{n^{\frac{3\kappa}{4\kappa+\rho-1}}} \right). \quad (2.6)$$

Suppose  $\{l(g) - l(g^*), g \in \mathcal{G}\}$  is Bernstein with respect to  $\mu$  with parameter  $\kappa > 1$  where  $g^* \in \arg \min_{\mathcal{G}} R_l(g)$  is unique. Suppose there exists  $0 < \rho < 1$  such that for every  $\delta > 0$ :

$$\tilde{\omega}_n(\mathcal{G}, \delta, \mu) := \mathbb{E} \sup_{g, g' \in \mathcal{G}: \|l(g) - l(g')\|_{L_2(\mu)} \leq \delta} |\tilde{P} - \tilde{P}_n|(l_\lambda(g) - l_\lambda(g')) \lesssim \frac{c(\lambda)}{\sqrt{n}} \delta^{1-\rho}. \quad (2.7)$$

Then estimator  $\hat{g} = \hat{g}_n^\lambda$  defined in (2.4) is such that:

$$\mathbb{E} R_l(\hat{g}) - R_l(g^*) \lesssim \left( \frac{c(\lambda)}{\sqrt{n}} \right)^{\frac{2\kappa}{2\kappa+\rho-1}} \vee \left( \frac{[c(\lambda)K(\lambda)]^{\frac{2\kappa}{4\kappa+\rho-1}}}{n^{\frac{3\kappa}{4\kappa+\rho-1}}} \right).$$

The proof of this result is presented in Section 5. Some remarks are in order. Assumption over  $\{l_\lambda(g), g \in \mathcal{G}\}$  introduced in Definition 1 is central. Gathering with the complexity assumption (2.7), it leads to a control of the variance term in decomposition (1.7). Then condition (2.6) gives the order of the bias term and leads to the excess risk bound.

This general upper bound generalizes the result presented in Koltchinskii [2006] to the indirect framework. Theorem 1 provides fast rates of convergence of the form:  $\mathcal{O}((c(\lambda)/\sqrt{n})^{2\kappa/2\kappa+\rho-1})$  provided that  $K(\lambda)$  is small enough. In the direct case with standard ERM estimators,  $l_\lambda(g) = l(g)$  and  $c(\lambda) = K(\lambda) = 1$  in Definition 1. In this case we obtain the fast rates  $\mathcal{O}(n^{-\kappa/2\kappa+\rho-1})$  of Koltchinskii [2006], Tsybakov [2004b]. In the presence of indirect observations, rates are slower since  $c(\lambda) \rightarrow +\infty$  as  $n \rightarrow +\infty$ . Hence the price to pay for the inverse problem is quantified by the Lipschitz constant  $c(\lambda)$  in Definition 1.

The behavior of constants  $c(\lambda)$  and  $K(\lambda)$  depend on the difficulty of the inverse problem through the degree of ill-posedness of operator  $A$ . Section 3 proposes to deal with the mildly ill-posed case where  $c(\lambda)$  and  $K(\lambda)$  grows polynomially when  $\lambda$  tends to zero.

Constant  $a(\lambda)$  depends essentially on the regularity of the conditional densities  $f_y, y \in \mathcal{Y}$ . It has to be upper-bounded by the level of variance in (2.6). Section 3 proposes to give explicit constants to calibrate  $\lambda$  and to get optimal rates of convergence.

Note that the control of the modulus of continuity in (2.7) is specific to the indirect framework and also depends on the constant  $c(\lambda)$ . A comparable hypothesis arises in the direct case in Koltchinskii [2006], up to the constant  $c(\lambda)$ . Section 4 is dedicated to the control (2.7). It appears that it will be satisfied under standard complexity conditions, such as  $L_2(\mu)$ -entropy of the loss class  $\{l(g), g \in \mathcal{G}\}$  (see Lemma 1 in Section 4 and the related discussion).

Finally Theorem 1 requires the unicity of the Bayes  $g^*$ . There is nice hope that such a restriction can be avoided using a more sophisticated geometry as in [Koltchinskii, 2006, Section 4].

### 3. Applications

In this section, we propose to apply the general upper bound of Theorem 1 to give rates of convergence of  $\lambda$ -ERM defined in (2.4). 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. Then we consider the general case using a family of projection estimators into the SVD base of the operator. These results hold when the loss class  $\{l(g) - l(g^*), g \in \mathcal{G}\}$  is Bernstein with respect to  $\mu = \nu \otimes P_Y$ , such as binary classification with hard loss for instance. For completeness, we also provide the same kind of results when  $\{l(g) - l(g^*), g \in \mathcal{G}\}$  is Bernstein with respect to  $P$ , where in this case we need to restrict the study to a compact set  $K \subseteq \mathcal{X}$ .

#### 3.1. Errors-in-variables case

We suppose that we observe a training set  $(Z_i, Y_i)$ ,  $i = 1 \dots, n$  where:

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

where the  $\epsilon_i$ 's are i.i.d.  $\mathbb{R}^d$ -random variables with density  $\eta$  with respect to  $\nu$ , the Lebesgue measure on  $\mathbb{R}^d$ . In this case operator  $A$  is a convolution product and the difficulty of this inverse problem can be represented thanks to the asymptotic behavior of the Fourier transform of the density  $\eta$ . Assumption **(A1)** below deals with the asymptotic behavior of the characteristic function of the noise distribution. These kind of restrictions are standard in deconvolution problems (see Butucea [2007], Fan [1991], Meister [2009]).

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

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

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

Hence we restrict ourselves to moderately ill-posed inverse problems by considering polynomial decay of the Fourier transform. Note that straightforward modifications in the proofs allow to consider severely ill-posed inverse problems. In this framework, we propose to construct a kernel deconvolution estimator of the densities  $f_y, y \in \mathcal{Y}$ . To this end, let us introduce  $\mathcal{K} = \prod_{j=1}^d \mathcal{K}_j : \mathbb{R}^d \rightarrow \mathbb{R}$  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

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

To apply Theorem 1, we also need the following assumption on the regularity of the conditional densities:

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

$$\mathcal{H}(\gamma, L) = \{f \in \Sigma(\gamma, L) : f \text{ are probability densities w.r.t. Lebesgue, } \forall x \in \mathbb{R}^d f(x) \leq M_0\},$$

and  $\Sigma(\gamma, L)$  is the Hölder class of  $\lfloor \gamma \rfloor$ -fold continuously differentiable functions on  $\mathbb{R}^d$  satisfying the Hölder condition.

This Hölder regularity is standard to control the bias term of kernel estimators in density estimation or density deconvolution (see for instance Tsybakov [2004a]).

In this context, for all  $g \in \mathcal{G}$ , we consider the empirical minimization (1.5) with empirical risk given by:

$$R_n^\lambda(g) = \frac{1}{n} \sum_{i=1}^n l_\lambda(g, (Z_i, Y_i)), \quad (3.2)$$

where  $l_\lambda(g, (z, y))$  is given by:

$$l_\lambda(g, (z, y)) = \int_{\mathbb{R}^d} l(g(x), y) \frac{1}{\lambda} K_\eta \left( \frac{z-x}{\lambda} \right) dx.$$

Theorem 2 below proposes to give the rates of convergence of  $\lambda$ -ERM under assumptions **(A1)**-**(R1)**.

**Theorem 2.** Suppose  $\{l(g) - l(g^*), g \in \mathcal{G}\}$  is a Bernstein class with respect to  $\nu_Y$  with parameter  $\kappa > 1$  and  $l(g(\cdot), y) \in L_2(\mathbb{R}^d)$ , for any  $y \in \mathcal{Y}$ . Suppose there exists  $0 < \rho < 1$  such that for every  $0 < \delta < 1$ :

$$\tilde{\omega}_n(\mathcal{G}, \delta, \nu_Y) \lesssim \frac{c(\lambda)}{\sqrt{n}} \delta^{1-\rho}.$$

If **(A1)** and **(R1)** hold, we have

$$\sup_{f_y \in \mathcal{H}(\gamma, l)} \mathbb{E} R_l(\hat{g}) - R_l(g^*) \lesssim n^{-\frac{\kappa\gamma}{\gamma(2\kappa+\rho-1)+2(\kappa-1)\bar{\beta}}},$$

where  $\bar{\beta} = \sum_{i=1}^d \beta_i$  and for a choice of  $\lambda = (\lambda_1, \dots, \lambda_d)$  given by:

$$\forall i \in \{1, \dots, d\}, \lambda_i \sim n^{-\frac{\kappa-1}{\gamma(2\kappa+\rho-1)+2(\kappa-1)\bar{\beta}}}. \quad (3.3)$$

The proof of this result is presented in Section 4. Few remarks are in order.

Rates in Theorem 2 generalizes the result of Koltchinskii [2006] (see also Tsybakov [2004b]) to the errors-in-variables case. Note 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  $\frac{2(\kappa-1)\bar{\beta}}{\gamma}$ , where  $\kappa > 1$ . Hence the performances of the method depends on the behavior of the characteristic function of the noise distribution. Moreover, in pattern recognition, we note that the influence of the errors in variables has to be related with the parameters  $\kappa$  of the Bernstein assumption and  $\gamma$  of the regularity of  $f_y$ . Same phenomenon arises in [Loustau and Marteau \[2011\]](#).

It is interesting to study the minimax optimality of the result of [Theorem 2](#) using the lower bounds presented in [Loustau and Marteau \[2011\]](#). To this end, we consider a random couple  $(X, Y)$  of law  $P$  where  $Y \in \{0, 1\}$ , a loss function  $l(g(x), y) = |Y - \mathbb{1}(X \in G)|$ , and a class of candidates  $\{g(x) = \mathbb{1}_G(x), G \in \mathcal{G}\}$ . In this case, the Bayes risk is defined as:

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

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

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

Gathering with the margin assumption, [[Mammen and Tsybakov, 1999](#), Lemma 2] gives:

$$\|l(g) - l(g')\|_{L_2(\nu_Y)}^2 = \|\mathbb{1}_G - \mathbb{1}_{G'}\|_{L_2(\mathbb{R}^d)}^2 = d_\Delta(G, G') \leq d_{f,g}(G, G')^{\frac{\alpha}{\alpha+1}}.$$

As a result the loss class  $\{l(g) - l(g'), g, g' \in \mathcal{G}\}$  is Bernstein with respect to  $\mu = \nu_Y$  with parameter  $\kappa = \frac{\alpha+1}{\alpha}$  since  $\frac{1}{2}d_{f,g}(G, G') = \mathbb{E}(l(g) - l(g'))$ .

To apply [Theorem 2](#), we are now on time to check [\(L \$\_\mu\$ \)](#) and [\(B\)](#) of [Definition 1](#). Note that [Lemma 3](#) in [Loustau and Marteau \[2011\]](#) states:

$$\|l_\lambda(g) - l_\lambda(g')\|_{L_2(\bar{P})}^2 \leq \prod_{i=1}^d \lambda_i^{-\beta_i} d_\Delta(G, G'),$$

where  $\{l_\lambda(G), G \in \mathcal{G}\}$  is defined in [\(3.2\)](#) with  $l(g(x), y) = |Y - \mathbb{1}(X \in G)|$ . It follows that  $\{l_\lambda(G), G \in \mathcal{G}\}$  is a LB-class with respect to  $\nu_Y$  with constants  $c(\lambda) = \prod \lambda_i^{-\beta_i}$  and  $K(\lambda) = \prod \lambda_i^{-\beta_i - 1/2}$ .

Last step is to control the complexity parameter  $\tilde{\omega}_n(\mathcal{G}, \delta, \nu_Y)$  as a function of  $\delta$ . Using [[Audibert and Tsybakov, 2007](#), Lemma 5.1], if  $\eta(x) := \mathbb{P}(Y = 1|X = x) \in \Sigma(\gamma, L)$ , we have a control of the  $L_2(\nu_Y)$ -entropy with bracketing of the class  $\{\mathbb{1}_G, G \in \mathcal{G}\}$  with exponent  $\rho = \frac{d}{\gamma\alpha}$ . As a result, we can apply [Lemma 1](#) in [Section 4](#) to get a control of the desired modulus of continuity as follows:

$$\tilde{\omega}_n(\mathcal{G}, \delta, \nu_Y) \lesssim \frac{c(\lambda)}{\sqrt{n}} \delta^{1 - \frac{d}{\gamma\alpha}}.$$

Hence we are on time to apply [Theorem 2](#) to get:

$$\mathbb{E}R_l(\hat{g}) - R_l(g^*) \lesssim n^{-\frac{(\alpha+1)\gamma}{\gamma(\alpha+2)+d+2\beta}},$$

which corresponds to the minimax rates of classification with errors in variables stated in [Loustau and Marteau \[2011\]](#). This result ensures the minimax optimality of the method in the errors-in-variables case.

### 3.2. General case with singular values decomposition

In this section we observe a training set  $(Z_i, Y_i)$ ,  $i = 1, \dots, n$  where  $Z_i$  are i.i.d. with law  $Af$ , where  $A : L_2(\mathcal{X}) \rightarrow L_2(\mathcal{Y})$  is a known linear compact operator. For simplicity, we also restrict ourselves to moderately ill-posed inverse problem considering the singular values decomposition of  $A$  as follows. Since  $A$  is compact,  $A^*A$  is auto-adjoint and compact and we can find an orthonormal base of eigenfunctions of  $A^*A$ , denoted by  $(\phi_k)_{k \in \mathbb{N}^*}$  such that  $A^*A\phi_k = b_k^2\phi_k$ , with  $(b_k)_{k \in \mathbb{N}^*}$  the decreasing sequence of singular values. Considering the image base  $\psi_k = A\phi_k/b_k$ , we have the following SVD (singular values decomposition):

$$A\phi_k = b_k\psi_k \text{ and } A^*\psi_k = b_k\phi_k, k \in \mathbb{N}^*. \quad (3.4)$$

In the sequel we make the following assumption:

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

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

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 \[2008\]](#)).

Considering the SVD (3.4), we propose to replace in the true risk the conditional densities  $f_y$  by a family of projection estimators given by:

$$\hat{f}_y(x) = \sum_{k=1}^N \hat{\theta}_k^y \phi_k(x), \quad (3.5)$$

where  $\hat{\theta}_k^y$  is an unbiased estimator of  $\theta_k^y = \int f_y \phi_k d\nu$  given by:

$$\hat{\theta}_k^y = \frac{1}{n_y} \sum_{i=1}^{n_y} b_k^{-1} \phi_k(Z_i). \quad (3.6)$$

In this case, assumption (2.2) is satisfied with  $k_N(z, x) = \sum_{k=1}^N b_k^{-1} \phi_k(z) \phi_k(x)$ . It gives the following expression of the empirical risk:

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

where here:

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

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

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

In this framework we also need the following assumption on the regularity of the conditional densities:

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

$$\mathcal{P}(\gamma, L) = \{f \in \Theta(\gamma, L) : f \text{ are probability densities w.r.t. Lebesgue, } \forall x \in \mathcal{X}, f(x) \leq M_0\},$$

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

$$\Theta(\gamma, L) = \{f(x) = \sum_{k \geq 1} \theta_k \phi_k(x) : \sum_{k \geq 1} \theta_k^2 k^{2\gamma} \leq L\}.$$

**Theorem 3.** Suppose  $\{l(g) - l(g^*), g \in \mathcal{G}\}$  is Bernstein class with respect to  $\nu_Y$  with parameter  $\kappa > 1$  such that  $l(g(\cdot), y) \in L_2(\nu)$ , for any  $y \in \mathcal{Y}$ . Suppose there exists  $0 < \rho < 1$  such that for every  $0 < \delta < 1$ :

$$\tilde{\omega}_n(\mathcal{G}, \delta, \nu_Y) \lesssim \frac{c(N)}{\sqrt{n}} \delta^{1-\rho}.$$

Then if **(A2)** and **(R2)** hold, estimators  $\hat{g}_n^N$  satisfies:

$$\sup_{f_y \in \mathcal{P}(\gamma, L)} \mathbb{E}R(\hat{g}_n^N) - R(g^*) \lesssim n^{-\frac{\kappa\gamma}{\gamma(2\kappa+\rho-1)+2(\kappa-1)\beta}},$$

where we choose  $N$  such that:

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

Theorem 3 shows that in pattern recognition with indirect observations, we can deal with any linear compact operator  $A$  using the SVD. From this point of view, this result could be compared with [Klemela and Mammen \[2010\]](#) where white noise model is considered.

Rates of convergence in Theorem 3 are comparable with Theorem 2. If  $A$  is a convolution operator, the result above shows that  $\hat{g}_n^N$  using projection estimators in the SVD reaches the rate of Theorem 2 with  $d = 1$  using kernel deconvolution estimators. In this case the regularity assumption deals with ellipsoids in the SVD domain instead of Hölder classes. However we can conjecture that this result is also minimax, although a rigorous lower bound has to be managed.

Finally there is nice hope that this result can be extended to other linear regularization methods without significant change. Here we present the result for projections into the SVD domain for the sack of simplicity in the proofs but Tikhonov and Landweber regularization could be considered for instance.

### 3.3. Restriction to a compact $K$

In this subsection, we propose an alternative to Theorem 2-3 to deal with a weaker Bernstein assumption. For the sack of simplicity, we restrict in Theorem 2-3 to Bernstein class  $\{l(g) - l(g^*), g \in \mathcal{G}\}$  with respect to measure  $\nu_Y$  according to Definition 2. In this case, it is sufficient to deal with LB-class with respect to  $\nu_Y$  in Definition 1, thanks to (2.5). However Bernstein class with respect to  $\nu_Y$  appears only in particular case, such as classification with hard loss in the context of Mammen and Tsybakov [1999], Tsybakov [2004b] (see also Section 3.1). Here we propose to state Theorem 2-3 when we deal with a Bernstein class in the spirit of Bartlett and Mendelson [2006], namely such that:

$$\mathbb{E}_P f^2 \leq \kappa_0 (\mathbb{E}_P f)^{1/\kappa}, \forall f \in \mathcal{F} = \{l(g) - l(g^*), g \in \mathcal{G}\}.$$

The idea is to restrict the study to a set  $K \subseteq \mathbb{R}^d$  where  $f \geq c_0 > 0$  over  $K$ . To this end we can consider a set  $\mathcal{G}$  of classifiers  $g$  such that  $\{x \in \mathcal{X} : f(x) > 0\} \subset K$ . Another point of view is to consider a convolution loss defined as:

$$l_\lambda(g, z, y) = \int_K k_\lambda(z, x) l(g(x), y) \nu(dx). \quad (3.7)$$

It means that we consider the minimization of a true risk of the form:

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

With (3.7), it is straightforward to get  $(\mathbf{L}_\mu)$  with  $\mu = P$  since if  $f \geq c_0 > 0$  on  $K$ , one gets:

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

Roughly speaking, Assumption  $(\mathbf{L}_\mu)$  in Definition 1 which  $\mu = P$  provides a control of the variance of  $l_\lambda(g, (Z, Y))$  by the variance of  $l(g(X), Y)$ . To have a control of the  $L_2(\tilde{P})$ -norm with respect to the  $L_2(P)$ -norm, we need to restrict  $g$  to a set where  $f > 0$ . Otherwise, the variance of  $l_\lambda(g)$  according to the law of  $\tilde{P}$  cannot be compared with the variance of  $l(g)$  with respect to  $P$ . With this trick, we have the following result.

**Corollary 1.** *Suppose  $\{l(g) - l(g^*), g \in \mathcal{G}\}$  is a Bernstein class with respect to  $P$  with parameter  $\kappa > 1$  and  $l(g(\cdot), y) \in L_2(\nu)$ , for any  $y \in \mathcal{Y}$ . Suppose there exists  $0 < \rho < 1$  such that for every  $0 < \delta < 1$ :*

$$\tilde{\omega}_n(\mathcal{G}, \delta, P) \lesssim \frac{c(\lambda)}{\sqrt{n}} \delta^{1-\rho}.$$

1. If **(A1)** and **(R1)** hold,  $\hat{g}_n^\lambda$  with (3.7) satisfies:

$$\sup_{f_y \in \mathcal{H}(\gamma, l)} \mathbb{E} R_l(\hat{g}) - R_l(g^*) \lesssim n^{-\frac{\kappa\gamma}{\gamma(2\kappa+\rho-1)+2(\kappa-1)\beta}},$$



where  $\bar{\beta} = \sum_{i=1}^d \beta_i$  and for a choice of  $\lambda = (\lambda_1, \dots, \lambda_d)$  given by:

$$\forall i \in \{1, \dots, d\}, \lambda_i \sim n^{-\frac{\kappa-1}{\gamma(2\kappa+\rho-1)+2(\kappa-1)\bar{\beta}}}. \quad (3.8)$$

2. If **(A2)** and **(R2)** hold,  $\hat{g}_n^N$  with (3.7) satisfies:

$$\sup_{f_y \in \mathcal{P}(\gamma, l)} \mathbb{E}R(\hat{g}_n^N) - R(g^*) \lesssim n^{-\frac{\kappa\gamma}{\gamma(2\kappa+\rho-1)+2(\kappa-1)\bar{\beta}}},$$

where we choose  $N$  such that:

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

This corollary allows to get the same fast rates of convergence of Theorem 2-3 under a weaker Bernstein assumption. The price to pay for the  $\lambda$ -ERM with restricted loss (3.7) resides in the dependence on  $K$  of the estimation procedure.

#### 4. Complexity from indirect observations

In this section we propose to control the indirect modulus of continuity thanks to standard learning theory arguments. The first result relates the control of  $\tilde{\omega}_n(\mathcal{G}, \delta, \mu)$  to the bracketing entropy of the loss class, which generalizes the result of the direct case (see [van der Vaart and Weelner \[1996\]](#)) when  $A = Id$ .

**Lemma 1.** *Consider a LB-class  $\{l_\lambda(g), g \in \mathcal{G}\}$  with respect to  $\mu$  with lipschitz constant  $c(\lambda)$ . Then we have the following assertion:*

$$\mathcal{H}_B(\{l(g), g \in \mathcal{G}\}, \epsilon, L_2(\mu)) \leq c\epsilon^{-2\rho} \Rightarrow \tilde{\omega}_n(\mathcal{G}, \delta, \mu) \lesssim \frac{c(\lambda)}{\sqrt{n}} \delta^{1-\rho},$$

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

With such a Lemma, it is possible to control the complexity in the indirect setup thanks to standard entropy conditions. Note that here no boundedness assumption is required for the loss  $l$  since we deal with a class of lipschitz and bounded loss  $g \mapsto l_\lambda(g)$ . The proof is presented in Section 6 and follows [van der Vaart and Weelner \[1996\]](#). Lemma 1 allows to consider standard hypothesis sets  $\mathcal{G}$  from the machine learning theory as hypothesis space (we refer for instance to [Koltchinskii \[2006\]](#) for many examples). Theorem 2 or Theorem 3 give corresponding rates of convergence.

Another interesting direction is to get a control of the indirect modulus of continuity thanks to Rademacher complexities. This can be done using the symmetrization device as follows:

$$\tilde{\omega}_n(\mathcal{G}, \delta, \mu) \leq 2\mathbb{E} \sup_{f \in \mathcal{F}(\delta)} \frac{1}{n} \left| \sum_{i=1}^n \epsilon_i l_\lambda(g, (Z_i, Y_i)) \right|,$$

where  $\mathcal{F}(\delta) = \{g, g' \in \mathcal{G} : \|l(g) - l(g')\|_{L_2(\mu)} \leq \delta\}$ . In this case, we have for instance when the loss class  $\{l(g), g \in \mathcal{G}\} \subset L_2(\mu)$  is a  $D$ -dimensional subset with orthonormal base  $(\varphi_k)_{k=1}^D$ :

$$\begin{aligned}
& \mathbb{E} \sup_{f \in \mathcal{F}(\delta)} \frac{1}{n} \left| \sum_{i=1}^n \epsilon_i l_\lambda(g, (Z_i, Y_i)) \right| \\
& \leq \frac{1}{n} \mathbb{E} \sup_{\theta \in \mathbb{R}^D} \left| \sum_{k=1}^D \sum_{i=1}^n \epsilon_i \int_{\mathcal{X}} k_\lambda(Z_i, x) (\theta_k - \theta'_k) \varphi_k(x) \nu(dx) \right| \\
& \leq \frac{1}{n} \mathbb{E}_{P_Y^{\otimes n}} \sup_{\theta \in \mathbb{R}^D} \sqrt{\sum_{k=1}^D (\theta_k - \theta'_k)^2 \mathbb{E}_{\tilde{P}_Z^{\otimes n}} \mathbb{E}_{\epsilon^{\otimes n}} \sqrt{\sum_{k=1}^D \left( \sum_{i=1}^n \epsilon_i \int_{\mathcal{X}} k_\lambda(Z_i, x) \varphi_k(x) \nu(dx) \right)^2}} \\
& \leq \frac{\delta}{n} \mathbb{E}_{\tilde{P}_Z} \mathbb{E}_{\epsilon^{\otimes n}} \sqrt{\sum_{k=1}^D \sum_{i=1}^n \epsilon_i^2 \int_{\mathcal{X}} \mathbb{E}_{\tilde{P}_Z} k_\lambda^2(Z, x) \varphi_k^2(x) \nu(dx)} \\
& \leq \frac{\sup_{x \in \mathcal{X}} \|k_\lambda(Z, x)\|_{L_2(\tilde{P}_Z)} \delta \sqrt{D}}{\sqrt{n}},
\end{aligned}$$

provided that we restrict the study to a compact set  $\mathcal{X}$  (see Section 3.3 below). Note that this result corresponds to assumption (2.7) since it is easy to see from Definition 1 that  $\{l_\lambda(g), g \in \mathcal{G}\}$  is necessary lipschitz with constant  $c(\lambda) = \sup_{x \in \mathcal{X}} \|k_\lambda(Z, x)\|_{L_2(\tilde{P}_Z)}$ . Finally we obtain the control of the modulus of continuity stated in the direct case up to the term  $c(\lambda)$ .

Another possible powerful direction is to study directly the complexity of the class  $\{l_\lambda(g), g \in \mathcal{G}\}$  thanks to entropy numbers of compact operators. To this end, note that if  $\mathcal{X}$  is compact,  $l_\lambda(g, z, y) = \int_{\mathcal{X}} k_\lambda(z, x) l(g(x), y) \nu(dx)$  can be considered as the image of  $l(g)$  by the integral operator  $L_{k_\lambda}$  associated to the function  $k_\lambda$ . We hence have:

$$\{l_\lambda(g), g \in \mathcal{G}\} = L_{k_\lambda}(\{l(g), g \in \mathcal{G}\}).$$

Moreover it is clear that if  $k_\lambda$  is continuous,  $L_{k_\lambda}$  is well-defined and compact with operator norm verifying:

$$\|L_{k_\lambda}\| \leq \sqrt{\nu(\mathcal{X})} \sup_{(z,x)} |k_\lambda(z, x)|.$$

Using for instance Williamson et al. [2001], and provided that  $l$  is bounded and  $\mathcal{G}$  consists in bounded functions in  $L_2(\nu, \mathcal{X})$ , entropy of the class  $\{l_\lambda(g), g \in \mathcal{G}\}$  could be controled in terms of the eigenvalues of the integral operator. In this case, it is clear that the entropy of the class  $\{l_\lambda(g), g \in \mathcal{G}\}$  depends strongly on the spectrum of the operator  $A$ .

More precisely, if  $A$  is a convolution product, this paper proposes to use kernel deconvolution estimators where in this case  $k_\lambda(z, x) = \frac{1}{\lambda} \mathcal{K}_\eta(\frac{z-x}{\lambda})$ . As a result, operator  $L_{k_\lambda}$  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  $\frac{\mathcal{F}[\kappa]}{\mathcal{F}[\eta](\frac{1}{\lambda})}$ . At the end, the control of the entropy of the class of interest  $\{l_\lambda(g), g \in \mathcal{G}\}$  could be calculated thanks to an assumption over the behavior of the Fourier transform of the noise distribution  $\eta$  such as **(A1)**.

## 5. Conclusion

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

The main result is a general upper bound in the statistical learning context, when we observe indirect observations  $Z_i, i = 1, \dots, n$  with law  $Af$ . The proof is based on an uniform concentration inequality, which seems to fit the indirect case, provided that it is used carefully. For this purpose, we introduce lipschitz and bounded classes  $\{l_\lambda(g), g \in \mathcal{G}\}$ , depending on a smoothing parameter  $\lambda$ . 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  $c(\lambda)$  which explodes as  $\lambda \rightarrow 0$ . The behavior of this constant is related to the degree of ill-posedness. Here in the mildly ill-posed case,  $c(\lambda)$  grows polynomially as a function of  $\lambda$ .

The result of Section 2 proposes the same order of generality as the results of Koltchinskii [2006] in the direct case, which allows to recover most of the recent results in statistical learning theory and the area of fast rates. There is nice hope that many problems when dealing with indirect observations could be managed following the guiding thread of this paper.

The estimation procedure proposed in this paper can be discussed for several reasons. Firstly it is not adaptive in many sense. At the first glance, we can see three level of adaptation: (1) adaptation to the operator  $A$ ; (2) adaptation to the tunable parameter  $\lambda$ ; (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 minimization (see Blanchard et al. [2008], Koltchinskii [2006], Loustau [2009], Tsybakov and Geer [2005]). However the construction of adaptive versions of  $\lambda$ -ERM of the previous sections is a challenging open problem.

Moreover an alternative point of view would be to design the best possible decision rule for classify  $Z$ , thanks to the i.i.d. observations  $Z_i$ 's. In this case standard ERM using the indirect sample can be performed. However to compare this "direct" method with the results of this paper, a rigorous comparison of the two frameworks should be done, and precisely the respective excess risks. It could be the core of a future work, from both theoretical and practical point of view (see Laurent et al. [2011] for a related discussion in testing inverse problems).

## 6. 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 version due to Bousquet (see Bousquet [2002]) applied to a class of measurable functions  $f \in \mathcal{F}$  from  $\mathcal{X}$  into  $[0, K]$ . 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

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

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  $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ , the following transformations:

$$\bar{\psi}(\delta) := \sup_{\sigma \geq \delta} \frac{\psi(\sigma)}{\sigma} \quad \text{and} \quad \check{\psi}(\epsilon) := \inf\{\delta > 0 : \psi^b(\delta) \leq \epsilon\}.$$

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

$$\bar{\psi}^q(\delta) := \sup_{\delta_j \geq \delta} \frac{\psi(\delta_j)}{\delta_j} \quad \text{and} \quad \check{\psi}^q(\epsilon) := \inf\{\delta > 0 : \psi^{b,q}(\delta) \leq \epsilon\},$$

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

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

### 6.1. Proof of Theorem 1

**Lemma 2.** *Suppose  $\{l_\lambda(g), g \in \mathcal{G}\}$  is such that  $\sup \|l_\lambda(g)\|_\infty \leq K(\lambda)$  with approximation function  $a(\lambda)$  and residual constant  $0 < r < 1$  according to Definition 3. Define:*

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

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

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

Then  $\forall \delta \geq \delta_n^\lambda(t) := (\tilde{U}_{n,t}^\lambda)^q(\frac{1}{2q})$ , if  $a(\lambda) \leq \frac{1-r}{4q}\delta$  we have:

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

where  $c(\delta, q) = -\frac{\log(\delta)}{\log(q)}$ .

*Proof.* The proof follows [Koltchinskii \[2006\]](#) extended to the noisy set-up.

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

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

Consider the following event:

$$E_{n,j}^\lambda(t) = \left\{ \sup_{g, g' \in \mathcal{G}(\delta_j)} |\tilde{P}_n - \tilde{P}|[l_\lambda(g) - l_\lambda(g')] \leq U_n^\lambda(\delta_j, t) \right\}.$$

We hence have, using Bousquet's version of Talagrand's inequality (see [Bousquet \[2002\]](#)), for some  $K > 0$ ,  $\mathbb{P}(E_{n,j}^{\lambda, C}(t)) \leq e^{-t}$ ,  $\forall t \geq 0$ .

We restrict ourself to the event  $E_{n,j}^\lambda(t)$ . Let consider  $\delta_j > \delta$ ,  $0 < \epsilon < \frac{\delta_{j+1}}{4}$ , and  $g \in \mathcal{G}(\epsilon)$ . Then we have the following assertion:

$$\delta_{j+1} \leq R_l(\hat{g}) - R_l(g^*) \leq \delta_j \Rightarrow \delta_{j+1} \leq R_l(\hat{g}) - R_l(g) + \epsilon.$$

Using [Definition 3](#) we have from [\(1.7\)](#):

$$\begin{aligned} R_l(\hat{g}) - R_l(g) &\leq (\tilde{P}_n - \tilde{P})(l_\lambda(g) - l_\lambda(\hat{g})) + (R_l - R_l^\lambda)(\hat{g} - g) \\ &\leq (\tilde{P}_n - \tilde{P})(l_\lambda(g) - l_\lambda(\hat{g})) + a(\lambda) + r(R_l(\hat{g}) - R_l(g))^2. \end{aligned}$$

It gives coarsely that:

$$\delta_{j+1} \leq \frac{1}{1-r} \left( (\tilde{P}_n - \tilde{P})(l_\lambda(g) - l_\lambda(\hat{g})) + a(\lambda) \right) + \epsilon,$$

since  $|R_l(\hat{g}) - R_l(g)| \leq \delta_j + \epsilon \leq \delta_j + \frac{\delta_{j+1}}{4} \leq 1$ . On the event  $E_{n,j}^\lambda(t)$ , since  $\epsilon \leq \delta_j$  it follows coarsely that:

$$\delta_{j+1} \leq \frac{1}{1-r} U_n^\lambda(\delta_j, t) + \frac{1}{1-r} a(\lambda) + \epsilon.$$

Hence we obtain using  $V_n^\lambda(\delta, t) := \bar{U}_n^\lambda(\delta, t)$ :

$$V_n^\lambda(\delta, t) \geq \frac{1}{q} - q^j \left[ \frac{1}{1-r} a(\lambda) + \epsilon \right] \geq \frac{1}{2q},$$

since we have:

$$a(\lambda) \leq \frac{1-r}{4q}\delta \implies q^j \left( \frac{1}{1-r} a(\lambda) + \epsilon \right) \leq \frac{1}{2q}.$$

It follows from the definition of  $\tilde{\psi}$  that:

$$\delta \leq (\tilde{U}_n^\lambda(\delta_j, t))^q \left(\frac{1}{2q}\right) = \delta_n^\lambda(t).$$

We hence have on the event  $E_{n,j}^\lambda(t)$ , for  $\delta_j \geq \max(\delta, \epsilon)$ :

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

or equivalently,

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

We eventually obtain:

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

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

$$\mathbb{P}(R(\hat{g}) - R(g^*) \geq \delta) \leq \sum_{\delta_j \geq \delta} \mathbb{P}(E_{n,j}^{\lambda C}(t)) \leq c(\delta, q)e^{-t},$$

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 1. We have:

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

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

$$\begin{aligned} \phi_n^\lambda(\mathcal{F}, \delta) &\leq \mathbb{E} \sup_{g, g' \in \mathcal{G}(\delta)} |\tilde{P}_n - \tilde{P}| [l_\lambda(g) - l_\lambda(g')] \\ &\leq \mathbb{E} \sup_{g, g' \in \mathcal{G}: P(l(g) - l(g'))^2 \leq K 2^{\frac{1}{\kappa}} \delta^{\frac{1}{\kappa}}} |\tilde{P}_n - \tilde{P}| [l_\lambda(g) - l_\lambda(g')] = \tilde{\omega}_n(\mathcal{G}, \delta^{\frac{1}{2\kappa}}) \\ &\leq \frac{c(\lambda)}{\sqrt{n}} \delta^{\frac{1-\rho}{2\kappa}}. \end{aligned}$$

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

$$U_n^\lambda(\delta, t) \leq c(K, \kappa) \frac{c(\lambda)}{\sqrt{n}} \delta^{\frac{1-\rho}{2\kappa}} + \frac{c(\lambda)^{1/2}}{n^{3/4}} \delta^{\frac{1-\rho}{4\kappa}} \sqrt{K(\lambda)t} + \sqrt{\frac{t}{n}} c(\lambda) \delta^{\frac{1}{2\kappa}} + \frac{t}{n}.$$

We hence have from an easy calculation:

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

To get the result we apply Lemma 1 with:

$$\delta = K'(1+t) \left( \left[ \frac{c(\lambda)}{\sqrt{n}} \right]^{\frac{2\kappa}{2\kappa+\rho-1}} \vee \frac{[c(\lambda)K(\lambda)]^{\frac{2\kappa}{4\kappa+\rho-1}} t^{\frac{2\kappa}{4\kappa+\rho-1}}}{n^{\frac{3\kappa}{4\kappa+\rho-1}}} \right),$$

provided that the approximation function obeys to the following inequality:

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

□

### 6.2. Proof of Theorem 2

Theorem 2 is straightforward application of Theorem 1 to the particular case of errors in variables using deconvolution kernel estimators and the general linear inverse problem using projection estimators.

First step is to check that the estimation procedure described in Section 3.1 gives rise to a LB-class with respect to  $\nu_Y$  where  $\nu$  is the Lebesgue measure on  $\mathbb{R}^d$ , thanks to the following lemma.

**Lemma 3.** *Suppose (A1) holds and suppose  $l(g(\cdot), y) \in L_2(\mathcal{X})$  for any  $y \in \mathcal{Y}$ . Consider a deconvolution kernel  $\mathcal{K}_\eta(t) = \mathcal{F}^{-1} \left[ \frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)} \right]$  where  $\mathcal{K}(t) = \prod_{i=1}^d K_i(t_i)$  with  $K_i$  having bounded compact supported Fourier transform. Then we have:*

$$\|l_\lambda(g) - l_\lambda(g')\|_{L_2(\tilde{P})} \lesssim \prod_{i=1}^d \lambda_i^{-\beta_i} \|l(g) - l(g')\|_{L_2(\nu_Y)},$$

and moreover:

$$\sup_{g \in \mathcal{G}} \|l_\lambda(g)\|_\infty \lesssim \prod_{i=1}^d \lambda_i^{-\beta_i - 1/2}.$$

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

$$\begin{aligned} \|l_\lambda(g) - l_\lambda(g')\|_{L_2(\tilde{P})}^2 &= \sum_{y \in \mathcal{Y}} p_y \int_{\tilde{\mathcal{X}}} \left[ \int_{\mathcal{X}} \frac{1}{\lambda} \mathcal{K}_\eta \left( \frac{z-x}{\lambda} \right) (l(g(x), y) - l(g'(x), y)) dx \right]^2 Af_y(z) dz \\ &= \sum_{y \in \mathcal{Y}} p_y \int_{\tilde{\mathcal{X}}} \left[ \frac{1}{\lambda} \mathcal{K}_\eta \left( \frac{\cdot}{\lambda} \right) * (l(g(\cdot), y) - l(g'(\cdot), y))(z) \right]^2 Af_y(z) dz \\ &\leq C \sum_{y \in \mathcal{Y}} p_y \int_{\tilde{\mathcal{X}}} \frac{1}{\lambda^2} |\mathcal{F}[\mathcal{K}_\eta \left( \frac{\cdot}{\lambda} \right)](t)|^2 |\mathcal{F}[l(g(\cdot), y) - l(g'(\cdot), y)](t)|^2 dt \\ &\leq C' \lambda^{-2\beta} \|l(g) - l(g')\|_{L_2(\nu_Y)}^2, \end{aligned}$$

where we use in last line the following inequalities:

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

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

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

$$\begin{aligned} \sup_{(z,y)} |l_\lambda(g, (z, y))| &\leq \sup_{(z,y)} \int_{\mathcal{X}} \left| \frac{1}{\lambda} \mathcal{K}_\eta \left( \frac{z-x}{\lambda} \right) l(g(x), y) \right| dx \\ &\leq C \sup_{z \in \mathcal{X}} \sqrt{\int_{\mathcal{X}} \left| \frac{1}{\lambda} \mathcal{K}_\eta \left( \frac{z-x}{\lambda} \right) \right|^2 dx} \\ &\leq \lambda^{-\beta-1/2}. \end{aligned}$$

A straightforward generalization leads to the  $d$ -dimensional case.  $\square$

Last step is to get an approximation function for the class  $\{l_\lambda(g), g \in \mathcal{G}\}$  with the following lemma:

**Lemma 4.** *Suppose (R1) holds and  $\mathcal{K}_\eta(t) = \mathcal{F}^{-1} \left[ \frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)} \right]$  such that  $K$  is a kernel of order  $\gamma$  with respect to the Lebesgue measure. Then if  $\{l(g) - l(g'), g, g' \in \mathcal{G}\}$  is Bernstein with parameter  $\kappa \geq 1$ , we have:*

$$\forall g, g' \in \mathcal{G}, (R_l^\lambda - R_l)(g - g') \leq a(\lambda) + r(R_l(g) - R_l(g'))^2,$$

where

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

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

$$(R_l^\lambda - R_l)(g - g') = \sum_{y \in \mathcal{Y}} p_y \int_{\mathcal{X}^2} K(u) (l(g(x), y) - l(g'(x), y)) (f_y(x + \lambda u) - f_y(x)) du dx.$$

Now since the  $f_y$ 's has  $l = \lceil \gamma \rceil$  derivatives, there exists  $\tau \in ]0, 1[$  such that:

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



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

Using the Bernstein assumption, one gets:

$$\begin{aligned}
(R_l^\lambda - R_l)(g - g') &\leq C\lambda^\gamma \sum_{y \in \mathcal{Y}} p_y \int_{\mathcal{X}} |l(g(x), y) - l(g'(x), y)| dx. \\
&\leq C\lambda^\gamma \sqrt{\sum_{y \in \mathcal{Y}} p_y \left( \int_{\mathcal{X}} |l(g(x), y) - l(g'(x), y)| dx \right)^2} \\
&\leq C \|l(g) - l(g')\|_{L_2(\nu)} \lambda^\gamma \\
&\leq C\lambda^\gamma (R_l(g) - R_l(g'))^{\frac{1}{2\kappa}} \\
&\leq C\lambda^{\frac{\kappa\gamma}{\kappa-1}} + \frac{1}{\kappa} (R_l(g) - R_l(g'))^2,
\end{aligned}$$

where we use in last line Young's inequality:

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

with  $r = \frac{1}{\kappa}$ . □

*Proof of Theorem 2.* The proof is a straightforward application of Theorem 1. From Lemma 3 and Lemma 4, condition (2.6) in Theorem 1 can be written:

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

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

### 6.3. 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  $\nu_Y$  with the following lemma.

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

$$\|l_\lambda(g) - l_\lambda(g')\|_{L_2(\bar{P})} \lesssim N^\beta \|l(g) - l(g')\|_{L_2(\nu_Y)},$$

and moreover:

$$\sup_{g \in \mathcal{G}} \|l_\lambda(g)\|_\infty \lesssim N^{\beta+1/2}.$$

*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 base and using the boundedness

assumptions over the  $f_y$ 's:

$$\begin{aligned}
\|l_N(g) - l_N(g')\|_{L_2(\bar{P})}^2 &= \sum_{y \in \mathcal{Y}} p_y \int_{\bar{\mathcal{X}}} \left( \sum_{k=1}^N b_k^{-1} \int_{\mathcal{X}} \phi_k(z) \phi_k(x) (l(g(x), y) - l(g'(x), y)) \nu(dx) \right)^2 A f_y(z) \nu(dz) \\
&\lesssim \sum_{y \in \mathcal{Y}} p_y \sum_{k=1}^N b_k^{-2} \int_{\bar{\mathcal{X}}} \phi_k(z)^2 \left( \int_{\mathcal{X}} (l(g(x), y) - l(g'(x), y)) \phi_k(x) \nu(dx) \right)^2 \nu(dz) \\
&\leq C N^{2\beta} \sum_{y \in \mathcal{Y}} p_y \sum_{k=1}^N \left( \int_{\mathcal{X}} (l(g(x), y) - l(g'(x), y)) \phi_k(x) \nu(dx) \right)^2 \\
&\leq C N^{2\beta} \|l(g) - l(g')\|_{L_2(\nu_Y)}^2.
\end{aligned}$$

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

$$\begin{aligned}
\sup_{(z,y)} |l_\lambda(g, (z, y))| &\leq \sup_{(z,y)} \left| \sum_{k=1}^N b_k^{-1} \int_{\mathcal{X}} \phi_k(x) \phi_k(z) l(g, (x, y)) \nu(dx) \right| \\
&\leq \sup_{(z,y)} \sqrt{\sum_{k=1}^N b_k^{-2}} \sqrt{\sum_{k=1}^N \theta_k^{l(g)^2} \phi_k(z)^2} \\
&\leq C N^{\beta+1/2}.
\end{aligned}$$

□

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

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

$$\forall g, g' \in \mathcal{G}, (R_l^\lambda - R_l)(g - g') \leq a(\lambda) + r(R_l(g) - R_l(g'))^2,$$

where

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

*Proof.* We first write, since  $E_{Z^y} \hat{\theta}_k^y = \theta_k^y =: \int_{\mathcal{X}} f_y(x) \phi_k(x) \nu(dx)$ :

$$\begin{aligned}
R_l^N(g) = \mathbb{E} R_n^N(g) &= \mathbb{E} \int_{\mathcal{X}} l(g(x), y) \sum_{k=1}^N \hat{\theta}_k^y \phi_k(x) \nu(dx) \\
&= \sum_{y \in \mathcal{Y}} p_y \int_{\mathcal{X}} l(g(x), y) \sum_{k=1}^N \mathbb{E}_{Z^y} \hat{\theta}_k^y \phi_k(x) \nu(dx) \\
&= \sum_{y \in \mathcal{Y}} p_y \int_{\mathcal{X}} l(g(x), y) \sum_{k=1}^N \theta_k^y \phi_k(x) \nu(dx)
\end{aligned}$$

We hence write:

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

Using Cauchy-Schwarz twice, we have since  $(\phi_k)_{k \in \mathbb{N}}$  in an orthonormal base and provided that  $f_y \in \Theta(\gamma, L)$ :

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

We conclude the proof using Young's inequality exactly as in Lemma 4.  $\square$

*Proof of Theorem 3.* The proof is a straightforward application of Theorem 1. From Lemma 5 and Lemma 6, condition (2.6) in Theorem 1 can be written:

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

Applying Theorem 1 with a smoothing parameter  $N$  such that an equality holds above gives the rates of convergence.  $\square$

#### 6.4. Proof of Lemma 1

The proof uses the maximal inequality presented in [van der Vaart and Weelner \[1996\]](#) to the class:

$$\mathcal{F} = \{l_\lambda(g) - l_\lambda(g'), g, g' \in \mathcal{G} : P(l(g) - l(g'))^2 \leq \delta^2\}.$$

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

$$\begin{aligned} \tilde{\omega}_n(\mathcal{G}, \delta, \mu) &= \mathbb{E} \sup_{g, g' \in \mathcal{G}: \|l(g) - l(g')\|_{L_2(\mu)}^2 \leq \delta^2} \left| (\tilde{P}_n - \tilde{P})(l_\lambda(g) - l_\lambda(g')) \right| \\ &\leq \frac{\|F\|_{L_2(\tilde{P})}^2}{\sqrt{n}} \int_0^\eta \sqrt{1 + \mathcal{H}_B(\mathcal{F}, \epsilon \|F\|_{L_2(\tilde{P})}^2, L_2(\mu))} d\epsilon \\ &\quad + \frac{\sup_{f \in \mathcal{F}} \|f\|_{L_2(\tilde{P})}}{\sqrt{n}} \sqrt{1 + \mathcal{H}_B(\mathcal{F}, \eta \|F\|_{L_2(\tilde{P})}^2, L_2(\mu))} \end{aligned} \quad (6.1)$$

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

$$\begin{aligned} \|F\|_{L_2(\tilde{P})}^2 &= \int F^2(z) P(dz, dy) \\ &= \sum_{y \in \mathcal{Y}} p_y \int \left( \sup_{f \in \mathcal{F}} |l_\lambda(g, z, y) - l_\lambda(g', z, y)| \right)^2 Af_y(z) \nu(dz) \\ &\lesssim K(\lambda)^2. \end{aligned}$$

Moreover, we have since  $\{l_\lambda(g), g \in \mathcal{G}\}$  is a LB-class with respect to  $\mu$  with lipschitz constant  $c(\lambda)$ :

$$\mathcal{H}_B(\{l(g), g \in \mathcal{G}\}, \epsilon, L_2(\mu)) \leq c\epsilon^{-2\rho} \Rightarrow \mathcal{H}_B(\mathcal{F}, \epsilon, L_2(\tilde{P})) \lesssim c(\lambda)^{2\rho} \epsilon^{-2\rho}.$$

We hence have in (6.1), choosing  $\eta = \frac{c(\lambda)}{K(\lambda)^2} \delta$ :

$$\begin{aligned} \tilde{\omega}_n(\mathcal{G}, \delta) &\lesssim \frac{K(\lambda)^2}{\sqrt{n}} \int_0^\eta \sqrt{1 + \epsilon^{-2\rho} K(\lambda)^{-4\rho} c(\lambda)^{2\rho}} d\epsilon + \frac{c(\lambda)\delta}{\sqrt{n}} \sqrt{1 + \eta^{-2\rho} K(\lambda)^{-4\rho} c(\lambda)^{2\rho}} \\ &\lesssim \frac{\eta K(\lambda)^2}{\sqrt{n}} + \frac{\eta^{1-\rho} K(\lambda)^{2(1-\rho)} c(\lambda)^\rho}{\sqrt{n}} + \frac{c(\lambda)\delta}{\sqrt{n}} + \frac{c(\lambda)^{1+\rho} \eta^{-\rho} K(\lambda)^{-2\rho} \delta}{\sqrt{n}} \\ &\lesssim \frac{\eta^{1-\rho} K(\lambda)^{2(1-\rho)} c(\lambda)^\rho}{\sqrt{n}} + \frac{c(\lambda)^{1+\rho} \eta^{-\rho} K(\lambda)^{-2\rho} \delta}{\sqrt{n}} + \frac{c(\lambda)}{\sqrt{n}} \delta^{1-\rho}, \end{aligned}$$

provided that  $\delta \leq 1$ .

## References

- J-Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of statistics*, 35:608–633, 2007.
- P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135 (3):311–334, 2006.
- 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 applicatio 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 fonctionnal estimation from indirect observations. *Annals of Statistics*, 35:1907–1930, 2007.
- C. Butucea and M.L. Taupin. New m-estimators in semi-parametric regression with errors in variables. *Ann. Inst. H. Poincar Probab. Statist.*, 44 (3):393–421, 2008.
- L. Cavalier. New concentration inequalities in product spaces. *Inverse Problems*, 24:1–19, 2008.
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
- H.W. Engl, M. Hank, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers Group, Dordrecht, 1996.
- J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19:1257–1272, 1991.
- J Klemela and E. Mammen. Empirical risk minimization in inverse problems. *Annals of Statistics*, 38 (1):482–511, 2010.
- V. Koltchinskii. Local rademacher complexities and oracle inequalties in risk minimization. *The Annals of Statistics*, 34 (6):2593–2656, 2006.
- B. Laurent, J.M. Loubes, and C. Marteau. Testing inverse problems: a direct or an indirect problem? *Journal of Statistical Planning and Inference*, 141: 1849–1861, 2011.
- S. Loustau. Penalized erm over besov spaces. *Electronic journal of Statistics*, 3: 824–850, 2009.
- S. Loustau and C. Marteau. Discriminant analysis with errors in variables. *submitted*, 2011.
- E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27 (6):1808–1829, 1999.
- P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9 (2):245–303, 2000.
- 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.
- M. Talagrand. New concentration inequalities in product spaces. *Invent. Math*, 126:505–563, 1996.
- A.B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer-Verlag, 2004a.
- A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32 (1):135–166, 2004b.
- A.B. Tsybakov and S.A. Van De Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, 33 (3):1203–1224, 2005.
- S. Van De Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.
- A. W. van der Vaart and J. A. Weelner. *Weak convergence and Empirical*

- Processes. With Applications to Statistics.* Springer Verlag, 1996.
- V. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer Verlag, 1982.
- V. Vapnik. *The Nature of Statistical Learning Theory.* Statistics for Engineering and Information Science, Springer, 2000.
- R.C. Williamson, A.J. Smola, and B. Schölkopf. 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.