Minimax fast rates for discriminant analysis with errors in variables

Sébastien Loustau, Clément Marteau

Minimax fast rates for discriminant analysis with errors in variables

Sébastien Loustau and Clément Marteau

July 12, 2013

Abstract

The effect of measurement errors in discriminant analysis is investigated. Given observations $Z = X + \epsilon$, where ϵ denotes a random noise, the goal is to predict the density of X among two possible candidates f and g. We suppose that we have at our disposal two learning samples. The aim is to approach the best possible decision rule $G^\star$ defined as a minimizer of the Bayes risk.

In the free-noise case ($\epsilon = 0$), minimax fast rates of convergence are well-known under the margin assumption in discriminant analysis (see [27]) or in the more general classification framework (see [32, 2]). In this paper we intend to establish similar results in the noisy case, i.e. when dealing with errors in variables. We prove minimax lower bounds for this problem and explain how can these rates be attained, using in particular an Empirical Risk Minimizer (ERM) method based on deconvolution kernel estimators.

1 Introduction

In the problem of discriminant analysis, we usually observe two i.i.d. samples $X_1^{(1)}, .., X_n^{(1)}$ and $X_1^{(2)}, .., X_m^{(2)}$. Each observation $X_j^{(i)} \in \mathbb{R}^d$ is assumed to admit a density with respect to a σ-finite measure Q, dominated by the Lebesgue measure. This density will be denoted by f if the observation belongs to the first set (i.e. when $i = 1$) or g in the other case. Our aim is to infer the density of a new incoming observation X. This problem can be considered as a particular case of the more general and extensively studied binary classification problem (see [13] for a detailed introduction or [7] for a concise survey).

In this framework, a decision rule or classifier can be identified with a set $G \subset \mathbb{R}^d$, which attributes X to f if $X \in G$ and to g otherwise. Then, we can associate to each classifier G its corresponding Bayes risk $R_K(G)$ defined as:

$$R_K(G) = \frac{1}{2}\left[\int_{K/G} f(x)dQ(x) + \int_G g(x)dQ(x)\right], \tag{1.1}$$

where we restrict the problem to a compact set $K \subset \mathbb{R}^d$. The minimizer of the Bayes risk (the best possible classifier for this criterion) is given by:

$$G_K^\star = \{x \in K : f(x) \geq g(x)\}, \tag{1.2}$$

where the infimum is taken over all subsets of K. The Bayes classifier is obviously unknown since it explicitly depends on the couple (f, g). The goal is thus to estimate $G_K^\star$ thanks to a classifier $\hat{G}_{n,m}$ based on the two learning samples.

The risk minimizer (1.2) has attracted many attentions in the last two decades because it involves a quantity of applied motivating examples, including pattern recognition, spam filtering, or medical diagnostic. However, in many real-world problems, direct observations are not available and measurement errors occur. As a result, it could be interesting to take into account this

problem into the classification task. In this paper, we propose to estimate the Bayes classifier $G_K^\star$ defined in (1.2) thanks to noisy samples. For all $i \in \{1,2\}$, we assume that we observe:

$$Z_j^{(i)} = X_j^{(i)} + \epsilon_j^{(i)},\ j = 1, \dots n_i, \tag{1.3}$$

instead of the $X_j^{(i)}$, where in the sequel $n_1 = n$ and $n_2 = m$. The $\epsilon_j^{(i)}$ denotes i.i.d. random variables expressing measurement errors. We will see in this work that we are facing an inverse problem, and more precisely a deconvolution problem. Indeed, assume that for all $x \in \mathbb{R}^d$, $dQ(x) = \mu(x)dx$ for some bounded function μ. If ϵ admits a density η with respect to the Lebesgue measure, then the corresponding density of the $Z_j^{(i)}$ is the convolution product $(f.\mu) * \eta$ if $i = 1$ or $(g.\mu) * \eta$ if $i = 2$. This property gives rise to a deconvolution step in the estimation procedure. Deconvolution problems arise in many fields where data are obtained with measurement errors and are at the core of several nonparametric statistical studies. For a general review of the possible methodologies associated to these problems, we may mention for instance [29]. More specifically, we refer to [15] in density estimation, [9] for non-parametric prediction or [8] where goodness-of-fit tests are constructed in the presence of noise. The main key of all these studies is to construct a deconvolution kernel which may allow to annihilate the noise ϵ. More details on the construction of such objects are provided in Section 3. It is important to note that in this discriminant analysis setup, or more generally in classification, there is up to our knowledge no such a work. The aim of this article is to describe minimax rates of convergence in noisy discriminant analysis under the margin assumption.

In the free-noise case, i.e. when $\epsilon = 0$, [27] has attracted the attention on minimax fast rates of convergence (i.e. faster than $n^{-\frac{1}{2}}$). In particular, they propose an classifier $\hat{G}_{n,m}$ satisfying

$$\sup_{G_K^\star \in \mathcal{G}(\alpha,\rho)} \mathbb{E}\left[R_K(\hat{G}_{n,m}) - R_K(G_K^\star)\right] \leq C(n \wedge m)^{-\frac{\alpha+1}{2+\alpha+\rho\alpha}}, \tag{1.4}$$

for some positive constant C. Here, $\mathcal{G}(\alpha,\rho)$ denotes a nonparametric set of candidates $G_K^\star$ with complexity $\rho > 0$ and margin parameter $\alpha \geq 0$ (see Section 2.1 for a precise definition). In (1.4), the complexity parameter $\rho > 0$ is related to the notion of entropy with bracketing whereas the margin is used to relate the variance to the expectation. It allows [27] to get improved bounds using the so-called peeling technique of [17]. This result is at the origin of a recent and vast literature on fast rates of convergence in classification (see for instance [28, 2]) or in general statistical learning (see [21]). In these papers, the complexity assumption can be of two forms: a geometric assumption over the class of candidates $G_K^\star$ (such as finite VC dimension, or boundary fragments) or assumptions on the regularity of the regression function of classification (plug-in type assumptions). In [28], minimax fast rates are stated for finite VC classes of candidates whereas plug-in type assumptions have been studied in the binary classification model in [2] (see also [13, 31]). More generally, [21] proposes to consider $\rho > 0$ as a complexity parameter in local Rademacher complexities. It gives general upper bounds generalizing (1.4) and the results of [27] and [2]. In the present work, a plug-in type complexity assumption will be considered.

In all these results, empirical risk minimizers appear as good candidates to reach these fast rates of convergence. Indeed, given a class of candidates $\mathcal{G}$, a natural way to estimate $G_K^\star$ is to consider an Empirical Risk Minimization (ERM) approach. In standard discriminant analysis (e.g. in the free-noise case considered in [27]), the risk $R_K(G)$ in (1.2) can be estimated by:

$$R_{n,m}(G) = \frac{1}{2n}\sum_{j=1}^{n} \mathbf{1}_{\{X_j^{(1)} \in K/G\}} + \frac{1}{2m}\sum_{j=1}^{m} \mathbf{1}_{\{X_j^{(2)} \in G\}}. \tag{1.5}$$

It leads to an empirical risk minimizer $\hat{G}_{n,m}$, if it exists, defined as:

$$\hat{G}_{n,m} = \arg\min_{G \in \mathcal{G}} R_{n,m}(G). \tag{1.6}$$

Unfortunately, in the errors-in-variables model, since we observe noisy samples $Z = X + \epsilon$, the probability densities of the observed variables w.r.t. the Lebesgue measure are respectively convolution $(f.\mu) * \eta$ and $(g.\mu) * \eta$, where for instance $f.\mu(x) = f(x) \times \mu(x)$ for all $x \in \mathbb{R}^d$. As a result, classical ERM principle fails since:

$$\frac{1}{2n}\sum_{i=1}^{n} \mathbf{1}_{\{Z_i^{(1)} \in G^C\}} + \frac{1}{2m}\sum_{i=1}^{m} \mathbf{1}_{\{Z_i^{(2)} \in G\}} \xrightarrow[n,m\to\infty]{a.s.} \frac{1}{2}\left[\int_{K/G} (f.\mu) * \eta(x)dx + \int_{G} (g.\mu) * \eta(x)dx\right] \neq R_K(G).$$

As a consequence, we add a deconvolution step in the classical ERM procedure and study the solution of the minimization:

$$\min_{G \in \mathcal{G}} R_{n,m}^{\lambda}(G),$$

where $R_{n,m}^{\lambda}(G)$ is an asymptotically unbiased estimator of $R_K(G)$. This empirical risk uses kernel deconvolution estimators with smoothing parameter λ. It is called deconvolution empirical risk and will be of the form:

$$R_{n,m}^{\lambda}(G) = \frac{1}{2n}\sum_{j=1}^{n} h_{K/G,\lambda}(Z_j^{(1)}) + \frac{1}{2m}\sum_{j=1}^{m} h_{G,\lambda}(Z_j^{(2)}), \tag{1.7}$$

where the $h_{G,\lambda}(\cdot)$ are deconvoluted versions of indicator functions used in classical ERM for direct observations (see Section 3 for details).

In this contribution, we would like to describe as precisely as possible the influence of the error ϵ on the classification rates of convergence and the presence of fast rates. Our aim is to use the asymptotic theory of empirical processes in the spirit of [17] (see also [34]) when dealing with the deconvolution empirical risk (1.7). To this end, we study in details the complexity of the class of functions $\{h_{G,\lambda}, G \in \mathcal{G}\}$, giving the explicit form of functions $h_{G,\lambda}$. This complexity is related to the imposed complexity over $\mathcal{G}$.

We establish lower and upper bounds and discuss the performances of this deconvolution ERM estimator under a plug-in complexity assumption. As mentioned earlier, different complexity assumptions have been developed in the last decades. The boundary fragment regularity, considered by e.g. [22, 27] is the core of a future work.

We point out that the definition of the empirical risk (1.7) leads to a new and interesting theory of risk bounds detailed in Section 3 for discriminant analysis. In particular, parameter λ has to be calibrated to reach a bias/variance trade-off in the decomposition of the excess risk. Related ideas have been recently introduced in [20] in the Gaussian white noise model and density estimation setting for more general linear inverse problems using singular values decomposition. In our framework, up to our knowledge, the only minimax result is [18] which gives minimax rates in Hausdorff distance for manifold estimation in the presence of noisy variables. [11] gives also consistency and limiting distribution for estimators of boundaries in deconvolution problems, but no minimax results are proposed. In the free-error case, we can also apply this methodology. In this case, the empirical risk is given by the estimation of f and g using simple kernel density estimators. This idea has been already mentioned in [35] in the general learning context and called Vicinal Risk Minimization (see also [10]). However, even in pattern recognition and in the direct case, up to our knowledge, there is no asymptotic rates of convergence for this empirical minimization principle.

In this contribution, a classifier G is always identified with a subset of $\mathbb{R}^d$. Our aim is to mimic the set $G_K^{\star}$ from the noisy observations (1.3). In particular, we aim at understanding the relationship between the spatial position of an input $X \in \mathbb{R}^d$ and its affiliation to one of the candidate densities. For this purpose, we give a deconvolution strategy to minimize the excess

risk (1.1). This problematic falls into the general problem of prediction with measurement errors (see [9]). This is the classification counterpart of the more extensively studied model of regression with errors-in-variables (see [16] or more recently [29]). It is important to note that one could alternatively try to provide the best classifier for a noisy input Z. In this case, we are faced to a direct problem which is in some sense already treated in [27]. However, it could be interesting to compare the performances of the two different approaches.

At this step, remark that similar problems have been considered in the test theory. Indeed, if we deal with a new incoming (noise free) observation X having density f_X, our aim is exactly to test one of the following 'inverse' hypotheses:

$$H_0^{IP} : \ f_X = f, \text{ against } H_1^{IP} : \ f_X = g. \tag{1.8}$$

However, we do not set any kind of order (null and alternative) between H_0 and H_1. The risk $R_K(G)$ is then related to the sum of the first and second kind error. Alternatively, if we deal with a noisy input Z having density $(f_X.\mu) * \eta$, this would correspond to test:

$$H_0^{DP} : \ (f_X.\mu) * \eta = (f.\mu) * \eta, \text{ against } H_1^{DP} : \ (f_X.\mu) * \eta = (g.\mu) * \eta. \tag{1.9}$$

A natural question then arises: are the both problems (1.8) and (1.9) equivalent or comparable? This question as already been addressed in [23] or [24] in a slightly different setting. This could be the core of a future work, but it requires the preliminary study provided in these papers.

Finally, for practical motivation, we can refer to the monograph of Meister ([25]) for particular models with measurement errors, such as in medecine, econometry or astronomy. In the specific context of classification, we met two explicit examples. The first one is an example in oncology where we try to classify the evolution of cancer thanks to medical images (like IRM or X-ray). These images are noisy due to the data collection process or the interpretation of the practitioner. The second example comes from meteorology where the weather forecaster wants to predict the future raining day thanks to measures such as rain gauge or barometer (which have well-studied random errors).

The paper is organized as follows. In Section 2, the model assumptions are explicited and an associated lower bound is stated. This lower bound generalizes to the indirect case the well-known lower bound of [2] established in classification. Deconvolution ERM attaining these rates are presented in Section 3. We also consider in this section standard kernel estimators, which allow to construct a new minimax optimal procedure in the direct case. A brief discussion and some perspectives are gathered in Section 4 while Section 5 is dedicated to the proofs of the main results.

2 Lower bound

2.1 Model setting

In this section, we detail some common assumptions (complexity and margin) on the pair (f, g). We then propose a lower bound on the corresponding minimax rates.

First of all, given a set $G \subset K$, simple algebra indicates that the excess risk $R_K(G) - R_K(G_K^\star)$ can be written as:

$$R_K(G) - R_K(G_K^\star) = \frac{1}{2} d_{f,g}(G, G_K^\star),$$

where the pseudo-distance $d_{f,g}$ over subsets of $K \subset \mathbb{R}^d$ is defined as:

$$d_{f,g}(G_1, G_2) = \int_{G_1 \Delta G_2} |f - g| dQ,$$

and $G_1 \Delta G_2 = [G_1^c \cap G_2] \cup [G_2^c \cap G_1]$ is the symmetric difference between two sets G_1 and G_2. In this context, there is another natural way of measuring the accuracy of a decision rule G through the quantity:

$$d_\Delta(G, G_K^\star) = \int_{G \Delta G_K^\star} dQ,$$

where d_Δ defines also a pseudo-distance on the subsets of $K \subset \mathbb{R}^d$.

In this paper, we are interested in the minimax rates associated to these pseudo-distances. In other words, given a class $\mathcal{F}$, one would like to quantify as precisely as possible the corresponding minimax risks defined as

$$\inf_{\hat{G}_{n,m}} \sup_{(f,g) \in \mathcal{F}} \mathbb{E}_{f,g} d_\square(\hat{G}_{n,m}, G_K^\star),$$

where the infimum is taken over all possible estimators of $G_K^\star$ and $d_\square$ stands for $d_{f,g}$ or d_Δ following the context. In particular, we will exhibit classification rules $\hat{G}_{n,m}$ attaining these rates. In order to obtain a satisfying study of the minimax rates mentioned above, one needs to detail the considered classes $\mathcal{F}$. Such a class expresses some conditions on the pair (f, g). They are often separated into two categories: margin and complexity assumptions.

A first condition is the well-known margin assumption. It has been introduced in discriminant analysis (see [27]) as follows.

Margin Assumption: *There exists positive constants $t_0, c_2, \alpha \geq 0$ such that for $0 < t < t_0$:*

$$Q\{x \in K : |f(x) - g(x)| \leq t\} \leq c_2 t^\alpha. \tag{2.1}$$

This assumption is related to the behavior of $|f - g|$ at the boundary of $G_K^\star$. It may give a variety of minimax fast rates of convergence which depends on the margin parameter α. A large margin corresponds to configurations where the slope of $|f - g|$ is high at the boundary of $G_K^\star$. The most favorable case arises when the margin $\alpha = +\infty$. In such a situation, $f - g$ has a discontinuity at the boundary of $G_K^\star$.

From a practical point of view, this assumption provides a precise description of the interaction between the pseudo distance $d_{f,g}$ and d_Δ. In particular, it allows a control of the variance of the empirical processes involved in the upper bounds, thanks to Lemma 2 in [27]. More general assumptions of this type can be formulated (see for instance [5] or [21]) in a more general statistical learning context.

For the sake of convenience, we will require in the following an additional assumption on the noise ϵ. We assume in the sequel that $\epsilon = (\epsilon_1, \ldots, \epsilon_d)'$ admits a bounded density η with respect to the Lebesgue measure satisfying:

$$\eta(x) = \prod_{i=1}^d \eta_i(x_i) \ \forall x \in \mathbb{R}^d. \tag{2.2}$$

In other words, the entries of the vector ϵ are independent. The assumption below describes the difficulty of the considered problems. It is often called the ordinary smooth case in the inverse problem literature.

Noise Assumption: *There exist $(\beta_1, \ldots, \beta_d)' \in \mathbb{R}_+^d$ and $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ positive constants such that for all $i \in \{1, \ldots, d\}$, $\beta_i > 1/2$,*

$$\mathcal{C}_1 |t|^{-\beta_i} \leq |\mathcal{F}[\eta_i](t)| \leq \mathcal{C}_2 |t|^{-\beta_i}, \text{ and } \left| \frac{d}{dt} \mathcal{F}[\eta_i](t) \right| \leq \mathcal{C}_3 |t|^{-\beta_i} \text{ as } |t| \to +\infty,$$

where $\mathcal{F}[\eta_i]$ 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 $i \in \{1, \ldots, d\}$.

Classical results in deconvolution (see e.g. [15], [16] or [8] among others) are stated for $d = 1$. Two different settings are then distinguished concerning the difficulty of the problem which is expressed through the shape of $\mathcal{F}[\eta]$. One can consider alternatively the case where $\mathcal{C}_1|t|^{-\beta} \leq |\mathcal{F}[\eta](t)| \leq \mathcal{C}_2|t|^{-\beta}$ as $|t| \to +\infty$, which yet corresponds to mildly ill-posed inverse problem or $\mathcal{C}_1 e^{-\gamma|t|^\beta} \leq |\mathcal{F}[\eta](t)| \leq \mathcal{C}_2 e^{-\gamma|t|^\beta}$ as $|t| \to +\infty$ which leads to a severely ill-posed inverse problem. This last setting corresponds to a particularly difficult problem and is often associated to low minimax rates of convergence.

In this contribution, we only deal with d-dimensional mildly ill-posed deconvolution problems. For the sake of brevity, we do not consider severely ill-posed inverse problems or possible intermediates (e.g. a combination of polynomial and exponential decreasing functions). Nevertheless, the rates in these cases could be obtained through the same steps.

The margin assumption is 'structural' in the sense that it describes the difficulty to distinguish an observation having density f from an other with density g. In order to provide a complete study, one also needs to set an assumption on the difficulty to find $G_K^\star$ in a possible set of candidates, namely a complexity assumption. In the classification framework, two different kinds of complexity assumptions are often introduced in the literature. The first kind concerns the regularity of the boundary of the Bayes classifier. Indeed, our aim is to estimate $G_K^\star$, which yet corresponds to a nonparametric set estimation problem. In this context, it seems natural to traduce the difficulty of the learning process by condition on the shape of $G_K^\star$. Another way to describe the complexity of the problem is to impose condition on the regularity of the underlying densities f and g. Such kind of condition is originally related to plug-in approaches and will be the investigated framework. Remark that these two assumptions are quite different and are convenient for distinct problems. In particular, a set $G_K^\star$ with a smooth boundary is not necessarily associated to smooth densities, and vice-versa.

In the rest of this section, lower bounds for the associated minimax rates of convergence are stated in the noisy setting. Corresponding upper bounds are presented and discussed in Section 3.

2.2 Lower bound for the plug-in assumption

The plug-in assumption considered in this paper is related to the regularity of the function $f - g$, expressed in terms of Hölder spaces. It corresponds to the same kind of assumption as in [2] for classification.
Given $\gamma, L > 0$, $\Sigma(\gamma, L)$ is the class of isotropic Hölder continuous functions ν having continuous partial derivatives up to order $\lfloor \gamma \rfloor$, the maximal integer strictly less than γ and such that:

$$|\nu(y) - p_{\nu,x}(y)| \leq L\|x - y\|^\gamma, \forall x, y \in \mathbb{R}^d,$$

where $p_{\nu,x}$ is the Taylor polynomial of ν at order $\lfloor \gamma \rfloor$ at point x and $\|.\|$ stands for the Euclidean norm on $\mathbb{R}^d$.

Plug-in Assumption. *There exist positive constants γ and L such that $f - g \in \Sigma(\gamma, L)$.*

We then call $\mathcal{F}_{\text{plug}}(Q)$ the set of all pairs (f, g) satisfying both the *margin* (with respect to Q) and the *plug-in* assumptions, since the previous assumption is often associated to plug-in rules in the statistical learning literature. The following theorem proposes a lower bound for the noisy discriminant analysis problem in such a setting.

Theorem 1 *Suppose that the noise assumption is satisfied. Then, there exists a measure Q_0 such that for all $\alpha \leq 1$,*

$$\liminf_{n,m\to+\infty} \inf_{\hat{G}_{n,m}} \sup_{(f,g)\in\mathcal{F}_{\mathrm{plug}}(Q_0)} (n\wedge m)^{\tau_d(\alpha,\beta,\gamma)} \mathbb{E}_{f,g} d_\square(\hat{G}_{n,m}, G_K^\star) > 0,$$

where the infinimum is taken over all possible estimators of the set $G_K^\star$ and

$$\tau_d(\alpha,\beta,\gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha)+d+2\displaystyle\sum_{i=1}^d \beta_i} & \text{for } d_\square = d_\Delta \\ \\ \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha)+d+2\displaystyle\sum_{i=1}^d \beta_i} & \text{for } d_\square = d_{f,g}. \end{cases}$$

Remark that we obtain exactly the same lower bounds as [2] in the direct case, which yet corresponds to the situation where $\beta_j = 0$ for all $j \in \{1,\dots,d\}$.

In the presence of noise in variables, the rates obtained in Theorem 1 are slower. The price to pay is an additional term of the form:

$$2\sum_{i=1}^d \beta_i.$$

This term clearly connects the difficulty of the problem to the tail behavior of the characteristic function of the noise distribution. This price to pay is already known in density estimation, regression with errors in variables or goodness-of-fit testing. Last step is to get a corresponding upper bound to validate this lower bound in the presence of noise in variables.

Remark that this lower bound is valid only for $\alpha \leq 1$. This restriction appears for some technical reasons in the proof (see Section 5). The main difficulty here is to use standard arguments from lower bounds in classification (see [1, 2]) in this deconvolution setting. More precisely, we have to take advantage of the noise assumption, related to the Fourier transform of the noise distribution η. To this end, we use in the proof of Theorem 1 an algebra based on standard Fourier analysis tools, and we have to consider sufficiently smooth objects. As a consequence in the lower bounds, we can check the margin assumption only for values of $\alpha \leq 1$. Nevertheless, we conjecture that this restriction is only due to technical reasons and that our result remains pertinent for all $\alpha \geq 0$. In particular, an interesting direction is to consider a wavelet basis which provides an isometric wavelet transform in L^2 in order to obtain the desired lower bound in the general case.

The measure Q_0 that we mention in Theorem 1 is explicitly constructed in the proof. For the sake of convenience, the construction of this measure is not reproduced here (we refer to Section 5.1 for an interested reader).

3 Upper bounds

3.1 Estimation of $G_K^\star$

In the free-noise case ($\epsilon_j^{(i)} = (0,\dots,0)$ for all $j \in \{1,\dots,n\}, i \in \{1,2\}$), we deal with two samples $(X_1^{(1)},\dots,X_n^{(1)})$, $(X_1^{(2)},\dots,X_m^{(2)})$ having respective densities f and g. A standard way to estimate $G_K^\star = \{x \in K : f(x) \geq g(x)\}$ is to estimate $R_K(\cdot)$ thanks to the data. For all $G \subset K$, the risk $R_K(G)$ can be estimated by the empirical risk defined in (1.5). Then the Bayes classifier $G_K^\star$ is estimated by $\hat{G}_{n,m}$ defined as a minimizer of the empirical risk (1.5) over a given family

of sets $\mathcal{G}$. We know for instance from [27] that the estimator $\hat{G}_{n,m}$ reaches the minimax rates of convergence in the direct case when $\mathcal{G} = \mathcal{G}(\gamma, L)$ corresponds to the set of boundary fragments with $\gamma > d-1$. For larger set $\mathcal{G}(\gamma, L)$, as proposed in [27], the minimization can be restricted to a $\delta-$net of $\mathcal{G}(\gamma, L)$. With an additional assumption over the approximation power of this $\delta-$net, the same minimax rates can be achieved in a subset of $\mathcal{G}(\gamma, L)$.

If we consider complexity assumptions related to the smoothness of $f-g$, we can show easily with [2] that an hybrid plug-in/ERM estimator reaches the minimax rates of convergence of [2] in the free-noise case. The principle of the method is to consider the empirical minimization (1.6) over a particular class $\mathcal{G}$ based on plug-in type decision sets. More precisely, following [2] for classification, we can minimize in the direct case the empirical risk over a class $\mathcal{G}$ of the form:

$$\mathcal{G} = \{\{f - g \geq 0\}, f - g \in \mathcal{N}_{n,m}\},$$

where $\mathcal{N}_{n,m}$ is a well-chosen $\delta-$net. With such a procedure, minimax rates can be obtained with no restriction over the parameter γ, α and d.

In noisy discriminant analysis, ERM estimator (1.6) is no longer available as mentioned earlier. Hence, we have to add a deconvolution step to the classical ERM estimator. In this context, we can construct a deconvolution kernel, provided that the noise has a non null Fourier transform, as expressed in the *Noise Assumption*. Such an assumption is rather classical in the inverse problem literature (see e.g. [15], [8] or [29]).

Let $\mathcal{K} = \prod_{j=1}^{d} \mathcal{K}_j : \mathbb{R}^d \to \mathbb{R}$ be a d-dimensional function defined as the product of d unidimensional functions $\mathcal{K}_j$. The properties of $\mathcal{K}$ leading to satisfying upper bounds will be precised later on. Then, if we denote by $\lambda = (\lambda_1, \ldots, \lambda_d)$ a set of (positive) bandwidths and by $\mathcal{F}[\cdot]$ the Fourier transform, we define $\mathcal{K}_\eta$ as:

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

In this context, for all $G \subset K$, the risk $R_K(G)$ can be estimated by

$$R^{\lambda}_{n,m}(G) = \frac{1}{2}\left[\frac{1}{n}\sum_{j=1}^{n} h_{K/G,\lambda}(Z_j^{(1)}) + \frac{1}{m}\sum_{j=1}^{m} h_{G,\lambda}(Z_j^{(2)})\right],$$

where for a given $z \in \mathbb{R}^d$:

$$h_{G,\lambda}(z) = \int_G \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)dx. \tag{3.2}$$

In the following, we study ERM estimators defined as:

$$\hat{G}^{\lambda}_{n,m} = \arg\min_{G\in\mathcal{G}} R^{\lambda}_{n,m}(G), \tag{3.3}$$

where parameter $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d_+$ has to be chosen explicitly. Functions $h_{G,\lambda}$ in equation (3.2) are at the core of the upper bounds. In particular, following the pioneering's works of Vapnik (see [35]), we have for $R^{\lambda}_K(\cdot) := \mathbb{E}R^{\lambda}_{n,m}(\cdot)$:

$$\begin{aligned} R_K(\hat{G}^{\lambda}_{n,m}) - R_K(G^\star_K) &\leq R_K(\hat{G}^{\lambda}_{n,m}) - R^{\lambda}_{n,m}(\hat{G}^{\lambda}_{n,m}) + R^{\lambda}_{n,m}(G^\star_K) - R_K(G^\star_K) \\ &\leq R^{\lambda}_K(\hat{G}^{\lambda}_{n,m}) - R^{\lambda}_{n,m}(\hat{G}^{\lambda}_{n,m}) + R^{\lambda}_{n,m}(G^\star_K) - R^{\lambda}_K(G^\star_K) \\ &\quad +(R_K - R^{\lambda}_K)(\hat{G}^{\lambda}_{n,m}) - (R_K - R^{\lambda}_K)(G^\star_K) \\ &\leq \sup_{G\in\mathcal{G}} |R^{\lambda}_K - R^{\lambda}_{n,m}|(G, G^\star_K) + \sup_{G\in\mathcal{G}} |R^{\lambda}_K - R_K|(G, G^\star_K), \end{aligned} \tag{3.4}$$

where we write for concision for any $G, G' \subset K$:

$$|R_K^\lambda - R_{n,m}^\lambda|(G, G') = |R_K^\lambda(G) - R_K^\lambda(G') - R_{n,m}^\lambda(G) + R_{n,m}^\lambda(G')|,$$

and similarly:

$$|R_K^\lambda - R_K|(G, G') = |R_K^\lambda(G) - R_K^\lambda(G') - R_K(G) + R_K(G')|.$$

As a result, to get risk bounds, we have to deal with two opposing terms, namely a so-called variability term:

$$\sup_{G \in \mathcal{G}} |R_K^\lambda - R_{n,m}^\lambda|(G - G_K^\star), \tag{3.5}$$

and a bias term (since $\mathbb{E}R_{n,m}^\lambda(G) \neq R_K(G)$) of the form:

$$\sup_{G \in \mathcal{G}} |R_K^\lambda - R_K|(G - G_K^\star). \tag{3.6}$$

The variability term (3.5) gives rise to the study of increments of empirical processes. In this work, this control is based on entropy conditions and uniform concentration inequalities. It is inspired by results presented for instance in [34] or [17]. The main novelty here is that in the noisy case, empirical processes are indexed by a class of functions which depends on the smoothing parameter λ. The bias term (3.6) is controlled by taking advantages of the properties of $\mathcal{G}$ and of the assumptions on the kernel $\mathcal{K}$. Indeed, it can be related to the standard bias term in nonparametric density estimation and can be controlled using smoothness assumptions of plug-in type. This bias term is inherent to the estimation procedure and its control is a cornerstone of the upper bounds.

The choice of λ will be a trade-off between the two opposing terms (3.5) and (3.6). Small λ leads to complex functions $h_{G,\lambda}$ and blasts the variance term whereas (3.6) vanishes when λ tends to zero. The kernel $\mathcal{K}$ has to be chosen in order to take advantage of the different conditions on $G_K^\star$. This choice will be operated according to the following definition.

Definition We say that $\mathcal{K}$ is a kernel of order $l \in \mathbb{N}^*$ if and only if:

- $\int \mathcal{K}(u)du = 1$.
- $\int u_j^k \mathcal{K}(u)du = 0 \ \forall \ k = 1, \dots l, \forall \ j = 1, \dots d$.
- $\int |u_j|^{l+1} |\mathcal{K}(u)|du < \infty, \forall \ j = 1, \dots d$.

In addition to this definition, we will require the following assumption on the kernel $\mathcal{K}$ which appears in (3.1).

Kernel Assumption. *The Kernel $\mathcal{K}$ is such that $\mathcal{F}[\mathcal{K}]$ is bounded and compactly supported.*

The construction of kernels of order l satisfying the kernel assumption could be managed using for instance the so-called Meyer wavelet (see [26]).

The following subsection intent to study deconvolution ERM estimator (3.3) and gives asymptotic fast rates of convergence. It validates the lower bounds of Theorem 1.

3.2 Upper bound

For all $\delta > 0$, using the notion of entropy (see for instance [34]) for Hölderian function on compact sets, we can find a δ-network $\mathcal{N}_\delta$ on $\Sigma(\gamma, L)$ such that:

- $\log(\mathrm{card}(\mathcal{N}_\delta)) \leq A\delta^{-d/\gamma}$,
- For all $h_0 \in \Sigma(\gamma, L)$, we can find $h \in \mathcal{N}_\delta$ such that $\|h - h_0\|_\infty \leq \delta$.

In the following, we associate to each $\nu := f - g \in \Sigma(\gamma, L)$, a set $G_\nu = \{x \in K : \nu(x) \geq 0\}$ and define the ERM estimator as:

$$\hat{G}_{n,m} = \arg\min_{\nu \in \mathcal{N}_\delta} R^\lambda_{n,m}(G_\nu), \tag{3.7}$$

where $\delta = \delta_{n,m}$ has to be chosen carefully. This procedure has been introduced in the direct case by [2] and referred to as an hybrid Plug-in/ERM procedure. The following theorem describes the performances of $\hat{G}_{n,m}$.

Theorem 2 *Let $\hat{G}_{n,m}$ the set introduced in (3.7) with*

$$\lambda_j = (n \wedge m)^{-\frac{1}{\gamma(2+\alpha)+2\sum_{i=1}^d \beta_i + d}},\ \forall j \in \{1, \ldots, d\},\ \text{and}\ \delta = \delta_{n,m} = \left(\frac{\prod_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n \wedge m}}\right)^{\frac{2}{d/\gamma+2+\alpha}}.$$

Given some $\sigma-$finite measure Q, suppose $(f, g) \in \mathcal{F}_{\mathrm{plug}}(Q)$ and the noise assumption is satisfied with $\beta_i > 1/2$, $\forall i = 1, \ldots d$. Consider a kernel $\mathcal{K}_\eta$ defined as in (3.1) where $\mathcal{K} = \Pi_{j=1}^d \mathcal{K}_j$ is a kernel of order $\lfloor \gamma \rfloor$, which satisfies the kernel assumption. Then, for all real $\alpha \geq 0$, if Q is the Lebesgue measure:

$$\lim_{n,m \to +\infty} \sup_{(f,g) \in \mathcal{F}_{\mathrm{plug}}(Q)} (n \wedge m)^{\tau_d(\alpha,\beta,\gamma)} \mathbb{E}_{f,g} d_\square(\hat{G}_{n,m}, G^\star_K) < +\infty,$$

where

$$\tau_d(\alpha, \beta, \gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha) + d + 2\sum_{i=1}^d \beta_i} & \text{for } d_\square = d_\Delta \\ \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha) + d + 2\sum_{i=1}^d \beta_i} & \text{for } d_\square = d_{f,g}. \end{cases}$$

Moreover, if $Q(x) = \mu(x)dx$, the same upper bounds hold provided that $\mu \in \Sigma(\gamma, L)$ and that $\max_{x \in \mathbb{R}^d} \mu(x) / \min_{x \in K} \mu(x) \leq c_0$ for some $c_0 > 0$.

Theorem 2 validates the lower bounds of Theorem 1. Deconvolution ERM are minimax optimal over the class $\mathcal{F}_{\mathrm{plug}}$. These optimal rates are characterized by the tail behavior of the characteristic function of the error distribution η. We only consider the ordinary smooth case whereas straightforward modifications lead to low rates of convergence in the super-smooth case.

Here, fast rates (i.e. faster than $1/\sqrt{n}$) are pointed out when $\alpha\gamma > d + 2\sum \beta_i$. This result is comparable to [2], where fast rates are proposed when $\alpha\gamma > d$. However, it is important to stress that large values of both α and γ correspond to restrictive situations. In this case, the margin parameter is high whereas the behavior of $f - g$ is smooth, which seems to be contradictory (see the related discussion in [2]). This situation arises for instance when Q is the uniform probability distribution in a ball centered at zero in $\mathbb{R}^d$, and the densities (f, g) satisfies $f(x) - g(x) = C\|x\|^2$ with an appropriate constant $C > 0$. In this case, $f - g \in \Sigma(\gamma, L)$ with arbitrarily large γ and the margin assumption holds with $\alpha = \frac{d}{2}$.

For the sake of concision, we do not study plug-in rules in this paper. Such algorithms are characterized by classifiers of the form

$$\tilde{G}_{n,m} = \left\{x \in K,\ \tilde{f}_n(x) - \tilde{g}_m(x) \geq 0\right\},$$

where $\tilde{f}_n - \tilde{g}_m$ is an (optimal) estimator of the function $f - g$. The performances of such kind of methods have been investigated by [2] in the binary classification model. We also mention for instance [19] or [6] for contributions in a more general framework.

Nevertheless, we point out that the choice of λ in Theorem 2 is the trade-off between the variability term (3.5) and the bias term (3.6). It is important to note that this asymptotic for λ is not the optimal choice in the problem of deconvolution estimation of $f - g \in \Sigma(\gamma, L)$ thanks to noisy data. Here the bandwidth depends on the margin parameter α and optimizes the classification excess risk bound. It highlights that the estimation procedure (3.7) is not a plug-in rule but an hybrid ERM/Plug-in estimator as in [2].

Finally, this deconvolution ERM appears to be minimax optimal when we deal with noisy data such that $\beta_i > \frac{1}{2}$, $\forall i = 1, \dots, d$. A natural question is to extend these results to the direct case where $\beta_i = 0$, $\forall i = 1, \dots, d$. Moreover, the minimax optimality of this procedure depends on the choice of λ in Theorem 2. In the following subsection, we deal with a similar approach in the direct case, using standard kernel estimators instead of deconvolution kernel estimators. Interestingly in this situation, the choice of λ is not crucial to derive optimal rates of convergence.

3.3 Upper bound in the free-noise case

In the free-noise setting, direct observations $X_j^{(1)}$, $j = 1, \dots, n$ and $X_j^{(2)}$, $j = 1, \dots, m$ are available. In this case, we can construct an estimation procedure based on (3.7) where a standard kernel estimator is used instead of a deconvolution kernel estimator. Following the noisy setting, we define in the direct case $\tilde{G}_{n,m}^{\lambda}$ as follows:

$$\tilde{G}_{n,m}^{\lambda} = \arg\min_{\nu \in \mathcal{N}_\delta} \tilde{R}_{n,m}^{\lambda}(G_\nu), \tag{3.8}$$

where here $\tilde{R}_{n,m}^{\lambda}(G)$ is an estimator of $R_K(G)$ defined as:

$$\tilde{R}_{n,m}^{\lambda}(G) = \frac{1}{2}\left[\frac{1}{n}\sum_{j=1}^{n} \tilde{h}_{K/G,\lambda}(X_j^{(1)}) + \frac{1}{m}\sum_{j=1}^{m} \tilde{h}_{G,\lambda}(X_j^{(2)})\right],$$

where for a given kernel $\mathcal{K}$:

$$\tilde{h}_{G,\lambda}(z) = \int_G \frac{1}{\lambda}\mathcal{K}\left(\frac{z-x}{\lambda}\right) dx.$$

The following theorem describes the performances of $\tilde{G}_{n,m}^{\lambda}$.

Corollary 1 *Let $\mathcal{F} = \mathcal{F}_{\text{plug}}(Q)$ and $\tilde{G}_{n,m}^{\lambda}$ the set introduced in (3.8) with*

$$\lambda_j \leq (n \wedge m)^{-\frac{1}{\gamma(2+\alpha)+d}},\ \forall j \in \{1, \dots, d\},\ \text{and}\ \delta = \delta_{n,m} = \left(\frac{1}{\sqrt{n \wedge m}}\right)^{\frac{2}{d/\gamma+2+\alpha}}.$$

Consider a kernel $\mathcal{K} = \Pi_{j=1}^{d}\mathcal{K}_j$ of order $\lfloor \gamma \rfloor$ satisfying the kernel assumption. Then, if Q is the Lebesgue measure, for any real $\alpha \geq 0$:

$$\lim_{n,m \to +\infty} \sup_{(f,g) \in \mathcal{F}_{\text{plug}}(Q)} (n \wedge m)^{\tau_d(\alpha,\gamma)} \mathbb{E} d_\square(\tilde{G}_{n,m}^{\lambda}, G_K^\star) < +\infty,$$

where

$$\tau_d(\alpha, \gamma) = \begin{cases} \dfrac{\gamma\alpha}{\gamma(2+\alpha)+d} \text{ for } d_\square = d_\Delta \\ \\ \dfrac{\gamma(\alpha+1)}{\gamma(2+\alpha)+d} \text{ for } d_\square = d_{f,g}. \end{cases}$$

Moreover, if $Q(x) = \mu(x)dx$, the same upper bounds holds provided that $\mu \in \Sigma(\gamma, L)$ and that $\max_{x \in \mathbb{R}^d} \mu(x) / \min_{x \in K} \mu(x) \leq c_0$ for some $c_0 > 0$.

These rates correspond to the lower bound of Theorem 1 for $\beta_j = 0$, $\forall j = 1, \dots, d$ (see also [2]). As a result, (3.8) provides a new procedure which reaches the minimax optimality in classification. Some remarks are in order.

The choice of λ in Corollary 1 is not standard. It seems that if λ is small enough, the ERM procedure (3.8) is minimax. This result can be explain as follows. Here, λ is not a trade-off between two opposing terms. In the control of the variability term, it appears that with a good choice of $\mathcal{K}$, the variability term does not depend on the bandwidth λ of the kernel. As a result, we only need to control the bias term with a small bandwidth.

This property can also be interpreted heuristically as follows. It is clear that the estimation procedure (3.8) with kernel estimator $\mathcal{K}$ is not so far from the usual ERM estimator in the direct case. Indeed, if λ is small enough, we have coarsely:

$$\tilde{h}_{G,\lambda}(X_i) = \int_G \frac{1}{\lambda} \mathcal{K}\left(\frac{X_i - x}{\lambda}\right) dx \approx \mathbb{1}_G(X_i).$$

As a result, with a small enough bandwidth, the procedure (3.8) reaches the same asymptotic performances as standard ERM.

4 Conclusion

We have provided in this paper minimax rates of convergence in the framework of smooth discriminant analysis with errors in variables. In the presence of plug-in type assumptions, we replace the unknown densities f and g by deconvolution kernel estimators. It gives a new family of ERM estimators called deconvolution ERM. It reaches the minimax rates of convergence. These optimal rates are fast rates (faster than $n^{-\frac{1}{2}}$) when $\alpha\gamma > d + 2\sum_{i=1}^{d} \beta_i$ and generalize the result of [2]. As shown in Table 1, the influence of the noise ϵ can be compared with standard results in nonparametric statistics (see [15, 16] for regression and density estimation with errors in variables or [8] in Goodness-of-fit testing) using kernel deconvolution estimators. Note that this idea can be adapted to the direct case using kernel density estimators. It provides a new minimax optimal procedure in the direct case, under the plug-in assumption.

	Density estimation	Goodness-of-fit testing	Classification
Direct case ($\epsilon = 0$)	$n^{-\frac{2\gamma}{2\gamma+1}}$	$n^{-\frac{2\gamma}{2\gamma+1/2}}$	$n^{-\frac{\gamma(\alpha+1)}{\gamma(\alpha+2)+d}}$
Errors-in-variables	$n^{-\frac{2\gamma}{2\gamma+2\beta+1}}$	$n^{-\frac{2\gamma}{2\gamma+2\beta+1/2}}$	$n^{-\frac{\gamma(\alpha+1)}{\gamma(\alpha+2)+2\bar{\beta}+d}}$
Regularity assumptions	$f \in \Sigma(\gamma, L)$ $\lvert\mathcal{F}[\eta](t)\rvert \sim \lvert t\rvert^{-\beta}$	$f \in W(s, L)$ $\lvert\mathcal{F}[\eta](t)\rvert \sim \lvert t\rvert^{-\beta}$	$f - g \in \Sigma(\gamma, L)$ $\lvert\mathcal{F}[\eta_i](t)\rvert \sim \lvert t\rvert^{-\beta_i}$ $\forall i$

Table 1. Optimal rates of convergence in pointwise L^2-risk in density estimation (see [15]), optimal separation rates for goodness-of-fit testing on Sobolev spaces $W(s, L)$ (see e.g. [8]) and the result of this work in smooth discriminant analysis (where $\bar{\beta} := \sum_{i=1}^{d} \beta_i$).

It is important to note that considering the estimation procedure of this paper, we are facing two different problems of model selection or adaptation. First of all, the choice of the bandwidths clearly depends on parameters which may be unknown a priori (e.g. the margin α and the regularity γ of the densities). In this sense, adaptation algorithms should be investigated to choose automatically λ to balance the bias term and the variance term. The second step of adaptation would be to consider a family of nested $(\mathcal{G}_k) \subset \mathcal{G}$ and to choose the model which balances the approximation term and the estimation term. This could be done using for instance penalization techniques, such as [33] or [21].

This work can be considered as a first attempt into the study of risk bounds in classification with errors in variables. It can be extended in many directions. Naturally the first extension will be to state the same kind of result in classification. Another natural direction would be to consider more general complexity assumptions for the hypothesis space $\mathcal{G}$. In the free-noise case, [4] deal with local Rademacher complexities. It allows to consider many hypothesis spaces, such as VC classes of sets, kernel classes (see [30]) or even Besov spaces (see [25]). Another advantage of considering Rademacher complexities is to develop data-dependent complexities to deal with the problem of model selection (see [21, 3]). It also allows us to deal with the problem of non-unique solution of the empirical minimization.

Into the direction of statistical inverse problem, there are also many open problems. A natural direction for applications would be to consider unknown density η for the random noise ϵ. This is a well known issue in the errors-in-variables setting to deal with unknown operator of inversion. In this setting we can consider repeated measurements to estimate the density of the noise ϵ (see for instance [12] for both density estimation and regression with errors). Another natural extension will be to consider general linear compact operator $A : f \mapsto Af$ to generalize the case of deconvolution. In this case, ERM estimators based on standard regularization methods from the inverse problem literature (see [14]) appear as good candidates. This could be the material of future works.

Finally, the presence of fast rates in discriminant analysis goes back to [27]. In [27], the regularity assumption is related to the smoothness of the boundaries of the Bayes. If we consider a set of Hölder boundary fragments, [27] states minimax fast rates in noise-free discriminant analysis. These rates are attained by ERM estimators. A natural extension of the present contribution is to state minimax rates in the presence of Hölder boundary fragments, where the control of the bias term seems really more nasty. This is the purpose of a future work.

5 Proofs

In this section, with a slight abuse of notations, $C, c, c' > 0$ denotes generic constants that may vary from line to line, and even in the same line. Given two real sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, the notation $a \simeq b$ (resp. $a \lesssim b$) means that there exists generic constants $C, c > 0$ such that $ca_n \leq b_n \leq Ca_n$ (resp. $a_n \leq Cb_n$) for all $n \in \mathbb{N}$.

5.1 Proof of Theorem 1

The proof mixes standard lower bounds arguments from classification (see [1] and [2]) but then uses some techniques which are specific to the inverse problem literature (see for instance [8] or [29]).

Consider $\mathcal{F}_1 = \{f_{\overrightarrow{\sigma}}, \overrightarrow{\sigma} = (\sigma_1, \ldots, \sigma_k) \in \{0,1\}^k\}$ a finite class of densities with respect to a specific measure Q_0 and g_0 a fixed density (with respect to the same Q_0) such that $(f_{\overrightarrow{\sigma}}, g_0) \in \mathcal{F}_{\text{plug}}$ for all $\overrightarrow{\sigma} \in \{0,1\}^k$. The construction of $f_{\overrightarrow{\sigma}}$ as a function of $\overrightarrow{\sigma}$, the value of g_0 and the definition of Q_0 will be precised in Section 5.1.1. Then, for all estimator $\hat{G}_{n,m}$ of the set $G_K^\star$, we have:

$$\sup_{(f,g)\in\mathcal{F}_{plug}} \mathbb{E}_{f,g} d_\Delta(\hat{G}_{n,m}, G_K^\star) \geq \sup_{f\in\mathcal{F}_1} \mathbb{E}_{g_0}\left[\mathbb{E}_f\left\{d_\Delta(\hat{G}_{n,m}, G_K^\star)|Z_1^{(2)}, \ldots, Z_m^{(2)}\right\}\right]. \tag{5.1}$$

In a first time, we propose a triplet $(\mathcal{F}_1, g_0, Q_0)$. Then, we prove that each associated element satisfies our hypotheses. We finish the proof with a convenient lower bound for (5.1).

5.1.1 Construction of the triplet $(\mathcal{F}_1, g_0, Q_0)$

We only consider the case $d = 2$ for simplicity, whereas straightforward modifications lead to the general d-dimensional case. For g_0, we take the constant 1 over $\mathbb{R}^2$:

$$g_0(x) = 1, \forall x \in \mathbb{R}^2.$$

For any $z \in \mathbb{R}^2$ and positive δ, we write in the sequel $B(z, \delta) := \{x = (x_1, x_2) : |x_i - z_i| \leq \delta\}$.

For an integer $q \geq 1$, introduce the regular grid on $[0,1]^2$ defined as:

$$G_q = \left\{ \left(\frac{2p_1 + 1}{2q}, \ldots, \frac{2p_d + 1}{2q} \right), p_i \in \{0, \ldots q-1\}, i = 1, 2 \right\}.$$

Let $n_q(x) \in G_q$ the closest point to $x \in \mathbb{R}^2$ among points in G_q (by convention, we choose the closest point to 0 when it is non unique). Consider the partition $(\chi'_j)_{j=1,\ldots q^2}$ of $[0,1]^2$ defined as follows: x and y belongs to the same subset if and only if $n_q(x) = n_q(y)$. Fix an integer $k \leq q^2$. For any $i \in \{1, \ldots k\}$, we define $\chi_i = \chi'_i$ and $\chi_0 = \mathbb{R}^2 \backslash \cup_{i=1}^k \chi_i$ to get $(\chi_i)_{i=1,\ldots,k}$ a partition of $\mathbb{R}^2$. In the sequel, we note by $(z^j)_{j=1,\ldots k}$ the centers of the χ_j.

Then, we consider the measure Q_0 defined as $dQ_0(x) = \mu(x)dx$ where $\mu(x) = \mu_0(x) + \mu_1(x)$ for all $x \in \mathbb{R}^2$ with

$$\mu_0(x) = k\omega\rho(x_1 - 1/2)\rho(x_2 - 1/2) \text{ and } \mu_1(x) = (1 - k\omega)\rho(x_1 - a)\rho(x_2 - b)$$

where k, ω, a, b are constants which will be precised later on and where for all $x \in \mathbb{R}$, $\rho : \mathbb{R} \to [0,1]$ is the function defined as

$$\rho(x) = \frac{1 - \cos(x)}{\pi x^2}, \ \forall x \in \mathbb{R}.$$

Recall that ρ satisfies $\mathcal{F}[\rho](t) = (1 - |t|)_+$. It allows us to take advantage of the noise assumption. Moreover g defines a probability density w.r.t. to the measure Q_0 since $\int_{\mathbb{R}^2} \mu(x)dx = 1$.

Now, we have to define the class $\mathcal{F}_1 = \{f_{\overrightarrow{\sigma}}, \overrightarrow{\sigma}\}$. We first introduce φ as a $\mathcal{C}^\infty$ probability density function w.r.t. the measure Q_0 and such that

$$\varphi(x) = 1 - c^\star q^{-\gamma} \ \forall x \in [0,1]^2.$$

Now introduce a class of functions $\psi_j : \mathbb{R}^2 \to \mathbb{R}$, for $j = 1, \ldots, k$ defined for any $x \in \mathbb{R}^2$ as follows:

$$\psi_j(x) = q^{-\gamma} c_\psi \rho(2\pi q(x_1 - z_1^j))\rho(2\pi q(x_2 - z_2^j)) \cos(4\pi q(x_1 - z_1^j)) \cos(4\pi q(x_2 - z_2^j)),$$

where $(z^j)_{j=1,\ldots k}$ are the centers of the χ_j. The class $(\psi_j)_j$ is specific to the noisy case and the inverse problem literature (see [8] and [29]). With such notations, for any $\overrightarrow{\sigma} \in \{0,1\}^k$, we define:

$$f_{\overrightarrow{\sigma}}(x) = \varphi(x) + \sum_{l=1}^{k} \sigma_l \psi_l(x), \ \forall x \in \mathbb{R}^2.$$

Now we have to check that this choice of $\mathcal{F}_1$, g_0 and Q_0 provides the margin assumption and that the complexity assumption hold true.

5.1.2 Main assumptions check

In a first time, we prove that the $f_{\overrightarrow{\sigma}}$ define probability density functions w.r.t. the measure Q_0. Let $\overrightarrow{\sigma} \in \{0,1\}^k$. Remark that, considering the case $d = 1$ w.l.o.g:

$$\begin{aligned} \int_{\mathbb{R}} \psi_l(x)\mu_0(x)dx = \mathcal{F}[\psi_l\mu_0](0) &= c_\psi q^{-\gamma} \mathcal{F}[\rho(2\pi q.)\mu_0(.)](\pm 4\pi q) \\ &= c_\psi q^{-\gamma} k\omega \mathcal{F}[\rho] * \mathcal{F}[\rho(2\pi q.)](\pm 4\pi q). \end{aligned}$$

Then, since

$$\mathcal{F}[\rho(2\pi q.)](t) = \frac{1}{2\pi q}\mathcal{F}[\rho]\left(\frac{t}{2\pi q}\right) \ \forall t \in \mathbb{R},$$

and

$$\mathcal{F}[\rho(2\pi q.)](t) \neq 0 \Leftrightarrow -1 < \frac{t}{2\pi q} < 1 \Leftrightarrow -2\pi q < t < 2\pi q,$$

we get

$$\text{supp}\mathcal{F}[\rho] * \mathcal{F}[\rho(2\pi q.)] = [-2\pi q - 1; 2\pi q + 1] \text{ and } \int_{\mathbb{R}} \psi_l(x)\mu_0(x)dx = 0. \tag{5.2}$$

The same computations show that $\int_{\mathbb{R}} \psi_l(x)\mu_1(x)dx = 0$ and prove the desired result since φ is a probability density with respect to Q_0.

Concerning the regularity, $f_{\overrightarrow{\sigma}} \in \Sigma(\gamma, L)$ for q large enough since $f_{\overrightarrow{\sigma}}$ can be written as $q^{-\gamma}F_0(x)$ where F_0 is infinitely differentiable.

In order to conclude this part, we only have to prove that the margin hypothesis is satisfied for all the couples $(f_{\overrightarrow{\sigma}}, g)$, namely for for some constant $c_2, t_0 > 0$, we have for $0 < t < t_0$:

$$Q_0\left(\left\{x \in [0,1]^d : |f_{\overrightarrow{\sigma}}(x) - g(x)| \leq t\right\}\right) \leq c_2 t^\alpha.$$

First, note that by construction of Q_0, we have $dQ_0(x) = (\mu_0(x) + \mu_1(x))dx$ and by choosing constant $a, b > 0$ large enough in μ_1, we can restrict ourselves to the study of the margin assumption with respect to $Q_0'(dx) = \mu_0(x)dx$.

Concerning the triplet (k, ω, q), we set

$$\begin{cases} k = q^2, \\ \omega = q^{-\alpha\gamma-2}. \end{cases}$$

In particular, we will have $k\omega = q^{-\alpha\gamma}$. Then, we will distinguish two different cases concerning the possible value of t. The first case concerns the situation where $C_1 q^{-\gamma} < t < t_0$ for some constant C_1. Then, we have for $Q_0'(dx) = \mu_0(x)dx$:

$$Q_0'\left(\left\{x \in [0,1]^2 : |f_{\overrightarrow{\sigma}}(x) - g(x)| \leq t\right\}\right) \leq \int_{[0,1]^2} \mu_0(x)dx \leq k\omega \leq Cq^{-\alpha\gamma} \leq Ct^\alpha.$$

Now, we consider the case where $t < C_1 q^{-\gamma}$. For all $\sigma \in \{0,1\}^k$:

$$\begin{aligned} Q_0'\left(\left\{x \in [0,1]^2 : |(f_\sigma - g)(x)| \leq t\right\}\right) &= \int_{[0,1]^2} k\omega \mathbf{1}_{|(f_\sigma - g)(x)| \leq t} dx \\ &\leq k\omega \sum_{j=1}^{k} \int_{\chi_j} \mathbf{1}_{|(f_\sigma - g)(x)| \leq t} dx \\ &\leq k^2 \omega \text{Leb}\left\{x \in \chi_1 : |(f_\sigma - g)(x)| \leq t\right\}, \end{aligned} \tag{5.3}$$

where without loss of generality, we suppose that $\sigma_1 = 1$ and we denote by $\text{Leb}(A)$ the Lebesgue measure of A.

Last step is to control the Lebesgue measure of the set $W_1 = \{x \in \chi_1 : |(f_\sigma - g)(x)| \leq t\}$. Since $f_\sigma - g = \sum_{j=1}^k \sigma_j \psi_j - c^\star q^{-\gamma}$, we have

$$W_1 = \left\{x \in \chi_1 : \left|\sum_{j=1}^k \sigma_j \psi_j(x) - c^\star q^{-\gamma}\right| \leq t\right\} = \left\{x \in \chi_1 : \left|\psi_1(x) - \left(c^\star q^{-\gamma} - \sum_{j=2}^k \sigma_j \psi_j(x)\right)\right| \leq t\right\}.$$

Moreover, note that on the square χ_j:

$$\begin{aligned}\sum_{l\neq j}\sigma_l\psi_l(x) &\leq q^{-\gamma}c_\psi\sum_{l\neq j}\frac{1}{2^4\pi^6q^4}\prod_{i=1}^{2}\frac{1}{|x_i-z_{l,i}|^2}\\ &\leq \frac{q^{-\gamma}c_\psi}{2^4\pi^6}\sum_{l\neq j}\frac{1}{|l-j|^4}\\ &\leq \frac{q^{-\gamma}c_\psi}{2^4\pi^6}\zeta(4)=\frac{q^{-\gamma}c_\psi\pi^4}{90\times 2^4\pi^6}:=c'q^{-\gamma},\end{aligned}\tag{5.4}$$

where $c'=\frac{c_\psi}{90\times 2^4\pi^2}$. Then, if we note by:

$$c_\infty=\sup_{x\in\chi_1}\rho(2\pi q(x_1-z_1^1))\rho(2\pi q(x_2-z_2^1))\cos(4\pi q(x_1-z_1^1))\cos(4\pi q(x_2-z_2^1)),$$

we have, for any $x\in\chi_1$:

$$\sum_{j=1}^{k}\sigma_j\psi_j(x)=\psi_1(x)+\sum_{j=2}^{k}\sigma_j\psi_j(x)\leq(c_\psi c_\infty+c')q^{-\gamma}.\tag{5.5}$$

Then, for all $x\in\chi_1$, we can define z^x as

$$z^x=\arg\min_{z:\psi_1(z)=c^\star q^{-\gamma}-\sum_{j=2}^{k}\sigma_j\psi_j(z)}\|x-z\|_2.$$

Indeed, inequality (5.5) ensures the existence of z^x provided that $c^\star<c_\psi c_\infty+c'$.

In order to evaluate the Lebesgue measure of W_1, the main idea is to approximate ψ_1 at each $x\in W_1$ by a Taylor polynomial of order 1 at z^x. We obtain

$$\begin{aligned}W_1 &= \{x\in\chi_1:|\psi_1(x)-\psi_1(z^x)|\leq t\},\\ &= \{x\in\chi_1:|\langle D\psi_1(z^x),x-z^x\rangle+\psi_1(x)-\psi_1(z^x)-\langle D\psi_1(z^x),x-z^x\rangle|\leq t\},\\ &\subset \{x\in\chi_1:|\,|\langle D\psi_1(z^x),x-z^x\rangle|-|\psi_1(x)-\psi_1(z^x)-\langle D\psi_1(z^x),x-z^x\rangle|\,|\leq t\}.\end{aligned}$$

Now, it is possible to see that there exists $c_0>0$ such that

$$|\langle D\psi_1(z^x),x-z^x\rangle|\geq c_0qq^{-\gamma}\|x-z^x\|_1,\ \forall x\in\chi_1.\tag{5.6}$$

Moreover, using again the inequality $\|x-z^x\|_1\leq C/q$, there exists a function $h:\mathbb{R}\to\mathbb{R}_+$ such that $qh(q)\to 0$ as $q\to\infty$ and which satisfies:

$$\frac{|\psi_1(x)-\psi_1(z^x)-\langle D\psi_1(z^x),x-z^x\rangle|}{\|x-z^x\|_1}\leq q^{-\gamma}h(q).\tag{5.7}$$

At this step, it is important to note that provided that $q:=q(n)\to\infty$ as $n\to\infty$, there exists some $n_0\in\mathbb{N}$ such that for any $n\geq n_0$, we have:

$$|\langle D\psi_1(z^x),x-z^x\rangle|>|\psi_1(x)-\psi_1(z^x)-\langle D\psi_1(z^x),x-z^x\rangle|.$$

Hence, we get the following inclusion

$$W_1\subset\left\{x\in\chi_1:c_0qq^{-\gamma}\|x-z^x\|_1\left(1-\frac{h(q)}{q}\right)\leq t\right\},\ \text{as } q\to+\infty.$$

With the property $qh(q)\to 0$ as $q\to\infty$ (or equivalently when $n\to\infty$), we can find n_0' large enough such that for any $n\geq n_0'$:

$$\mathrm{Leb}(W_1)\leq\mathrm{Leb}\left(\left\{x\in\chi_1:\|x-z^x\|_1\leq\frac{t}{2c_0}q^{\gamma-1}\right\}\right)\leq\frac{t}{2c_0qq^{1-\gamma}}.$$

Gathering with (5.3), we hence get, for $t < C_1 q^{-\gamma}$, provided that $\alpha \leq 1$:

$$\begin{aligned} Q_0' \left\{ x \in [0,1]^2 : |(f_\sigma - g)(x)| \leq t \right\} &\leq Ck^2 \omega \frac{t}{q^2 q^{-\gamma}} \\ &\leq Ck\omega \frac{t}{q^{-\gamma}} = Cq^{\gamma(1-\alpha)} t^\alpha t^{1-\alpha} \leq Ct^\alpha, \end{aligned}$$

where $C > 0$ is a generic constant.

5.1.3 Final minoration

Suppose without loss of generality that $n \leq m$. Now we argue as in [1] (Assouad Lemma for classification) and introduce ν, the distribution of a Bernoulli variable ($\nu(\sigma = 1) = \nu(\sigma = 0) = 1/2$). Then, denoting by $\mathbb{P}_{\overrightarrow{\sigma}}^{\otimes n}$ the law of $(Z_1^{(1)}, \ldots, Z_n^{(1)})$ when $f = f_{\overrightarrow{\sigma}}$, we get

$$\begin{aligned} &\sup_{\overrightarrow{\sigma} \in \{0,1\}} \mathbb{E}_f \left\{ d_\Delta(\hat{G}_{n,m}, G_K^*) | Z_1^{(2)}, \ldots, Z_m^{(2)} \right\} \\ \geq\ & \mathbb{E}_{\nu^{\otimes k}} \mathbb{E}_{f_{\overrightarrow{\sigma}}} d_\Delta(\hat{G}_{n,m}, G_K^*), \\ \geq\ & \mathbb{E}_{\nu^{\otimes k}} \mathbb{E}_{f_{\overrightarrow{\sigma}}} \sum_{j=1}^{k} \int_{\chi_j} \mathbf{1}(x \in \hat{G}_{n,m} \Delta G_K^\star) Q_0(dx), \\ =\ & \sum_{j=1}^{k} \mathbb{E}_{\nu^{\otimes(k-1)}} \int_\Omega \mathbb{E}_{\nu(d\sigma_j)} \int_{\chi_j} \mathbf{1}(x \in \hat{G}_{n,m}(\omega) \Delta G_K^\star) Q_0(dx) \mathbb{P}_{\overrightarrow{\sigma}}^{\otimes n}(d\omega) \\ \geq\ & \sum_{j=1}^{k} \mathbb{E}_{\nu^{\otimes(k-1)}} \int_\Omega \mathbb{E}_{\nu(d\sigma_j)} \int_{\chi_j} \mathbf{1}(x \in \hat{G}_{n,m}(\omega) \Delta G_K^\star) Q_0(dx) \left[\frac{\mathbb{P}_{\overrightarrow{\sigma}_{j,1}}^{\otimes n}}{\mathbb{P}_{\overrightarrow{\sigma}_j}^{\otimes n}} \wedge \frac{\mathbb{P}_{\overrightarrow{\sigma}_{j,0}}^{\otimes n}}{\mathbb{P}_{\overrightarrow{\sigma}_j}^{\otimes n}} \right] \mathbb{P}_{\overrightarrow{\sigma}}^{\otimes n}(d\omega) \quad (5.8) \end{aligned}$$

where $\overrightarrow{\sigma}_{j,r} = (\sigma_1, \ldots, \sigma_{j-1}, r, \sigma_{j+1}, \ldots, \sigma_k)$ for $r \in \{0,1\}$.

Moreover, note that from (5.4), we have on the square χ_j:

$$\sum_{l \neq j} \sigma_l \psi_l(x) \leq c' q^{-\gamma},$$

where $c' = \frac{c_\psi}{90 \times 2^4 \pi^2}$. Now it is easy to see that from the definition of the test functions ψ_j, for any integer $k_0, k_1 : k_1 > 2k_0$, on the square ring $B_j(k_0, k_1) = \{x \in \chi_j : \forall i \, |x_i - z_{j,i}| \leq \frac{1}{2k_0 q}$ and $|x_i - z_{j,i}| \geq \frac{1}{k_1 q}\}$:

$$\psi_j(x) - c^\star q^{-\gamma} \geq q^{-\gamma} \left[c_\psi k_0^4 \frac{\left(1 - \cos \frac{2\pi}{k_1}\right)^2}{\pi^6} (\cos 4\pi/k_0)^2 - c^\star \right] = q^{-\gamma},$$

provided that $c_\psi = \frac{\pi^6 (1 + c^\star)}{k_0^2 (1 - \cos 2\pi/k_1)^2 (\cos 4\pi/k_0)^2}$. Hence, since $c' = \frac{c_\psi}{90 \times 2^4 \pi^2}$, we can choose $k_0, k_1 \in \mathbb{N}$ such that $c' \leq 1$ to get on $B_j(k_0, k_1)$:

$$\sum_{l \neq j} \sigma_l \psi_l(x) \leq c' q^{-\gamma} \leq \psi_j(x) - c^\star q^{-\gamma}. \quad (5.9)$$

Now introduce binary valued functions:

$$\hat{f}(x) = \mathbf{1}(x \in \hat{G}_{n,m}) \text{ and } f_{\overrightarrow{\sigma}}^\star(x) = \mathbf{1}(x \in G_{K,\sigma}^\star),$$

where $G_{K,\sigma}^\star = \{f_{\overrightarrow{\sigma}} - g \geq 0\}$. From (5.9), we claim that for any $\overrightarrow{\sigma}$:

$$\forall x \in B_j(k_0, k_1), f_{\overrightarrow{\sigma}}^\star(x) = \sigma_j. \quad (5.10)$$

Indeed, since $f_{\overrightarrow{\sigma}} - g = \sum_{l=1}^{k} \sigma_l \psi_l - c^\star q^{-\gamma}$, gathering with (5.9), we have the following assertion:

$$f^\star_{\overrightarrow{\sigma}}(x) = 1 \Rightarrow (1+\sigma_j)\psi_j(x) \geq 2c^\star q^{-\gamma} \Rightarrow \sigma_j = 1,$$

provided that $c^\star \leq q^\gamma \min_{x \in B_j(k_0,k_1)} \psi_j(x)/2$. Moreover, this choice of $c^\star$ leads to the following assertion:

$$f^\star_{\overrightarrow{\sigma}}(x) = 0 \Rightarrow \sum_{l=1}^{k} \sigma_l \psi_l(x) \leq c^\star q^{-\gamma} \leq \min_{x \in B_j(k_0,k_1)} \psi_j(x)/2.$$

In this case, if $\sigma_j = 1$, we obtain:

$$\psi_j(x) + \sum_{l \neq j} \sigma_l \psi_l(x) \leq \min_{x \in B_j(k_0,k_1)} \psi_j(x)/2. \tag{5.11}$$

Last step is to show that (5.11) is a contradiction. For this purpose, note that:

$$\min_{x \in B_j(k_0,k_1)} \left(\psi_j(x) + \sum_{l \neq j} \sigma_l \psi_l(x) \right) \geq \min_{x \in B_j(k_0,k_1)} \psi_j(x) + \min_{x \in B_j(k_0,k_1)} \sum_{l \neq j} \sigma_l \psi_l(x) \geq \min_{x \in B_j(k_0,k_1)} \psi_j(x)/2,$$

where the last inequality is guaranteed when:

$$\min_{x \in B_j(k_0,k_1)} \psi_j(x)/2 \geq - \min_{x \in B_j(k_0,k_1)} \sum_{l \neq j} \sigma_l \psi_l(x).$$

Finally, the last inequality holds thanks to the positivity of $\psi_j(x)$ on the set $B_j(k_0, k_1)$ and the fact that $\forall j' \neq j$, $\text{sign}\psi_j = \text{sign}\psi_{j'}$. Indeed, $\forall j$, $\psi_j(x) = 0$ for $x \in \mathcal{Z}_{j,1} \cup \mathcal{Z}_{j,2}$ where:

$$\mathcal{Z}_{j,1} = \left\{ x \in \mathbb{R}^2 : |x_u - z_u^j| = \frac{k}{q},\, u \in \{1,2\},\, k \in \mathbb{N}^* \right\}$$

and

$$\mathcal{Z}_{j,2} = \left\{ x \in \mathbb{R}^2 : |x_u - z_u^j| = \frac{2k+1}{8q},\, u \in \{1,2\},\, k \in \mathbb{N} \right\}.$$

Note that by construction, $\forall j \neq j'$, $\mathcal{Z}_{j,2} = \mathcal{Z}_{j',2} = \mathcal{Z}_2$ does not depend on $j \in \{1, \ldots, k\}$. Moreover, for any $j \in \{1, \ldots, k\}$, ψ_j is alternatively positive and negative on the checkerboard associated with $\mathcal{Z}_2$. It leads to $\text{sign}\psi_j = \text{sign}\psi_{j'}$, $\forall j \neq j'$ since two centers z^j and $z^{j'}$ are separated by an odd number of squares (exactly 5) on both directions. We hence have by construction that (5.11) is a contradiction and then, (5.10) is shown.

Now we go back to the lower bound. We can write:

$$\begin{aligned}
& \mathbb{E}_{\nu(d\sigma_j)} \int_{\chi_j} \mathbf{1}(x \in \hat{G}_{n,m}(\omega) \Delta G_K^\star) Q_0(dx) = \mathbb{E}_{\nu(d\sigma_j)} \int_{\chi_j} \mathbf{1}(\hat{f} \neq f^\star_{\overrightarrow{\sigma}}) Q_0(dx) \\
\geq \;& E_{\nu(d\sigma_j)} \left[\int_{B_j} \mathbf{1}(\hat{f} \neq \sigma_j) Q_0(dx) \right] \\
= \;& \frac{1}{2} \left[\int_{B_j} [\mathbf{1}(\hat{f} \neq 1) + \mathbf{1}(\hat{f} \neq 0)] Q_0(dx) \right] \\
= \;& \frac{1}{2} \int_{B_j} Q_0(x) dx,
\end{aligned}$$

where we use (5.10) at the second line with $B_j := B_j(k_0, k_1)$. Then it follows from (5.8) that:

$$\begin{aligned}
&\sup_{\overrightarrow{\sigma}\in\{0,+1\}^k} \mathbb{E}_f\left\{d_\Delta(\hat{G}_{n,m}, G_K^\star)|Z_1^{(2)},\ldots,Z_m^{(2)}\right\} \\
&\geq \mathbb{E}_{\nu^{\otimes(k-1)}} \sum_{j=1}^k \int_\Omega \left[\frac{\mathbb{P}_{\overrightarrow{\sigma}_{j,0}}^{\otimes n}}{\mathbb{P}_{\overrightarrow{\sigma}_j}^{\otimes n}} \wedge \frac{\mathbb{P}_{\overrightarrow{\sigma}_{j,1}}^{\otimes n}}{\mathbb{P}_{\overrightarrow{\sigma}_j}^{\otimes n}}\right](d\omega)\frac{1}{2}\int_{\chi_j} Q_0(dx)\mathbb{P}_{\overrightarrow{\sigma}}^{\otimes n}(d\omega) \\
&= \sum_{j=1}^k \mathbb{E}_{\nu^{\otimes(k-1)}}[1-\mathbb{V}(\mathbb{P}_{\overrightarrow{\sigma},1}^{\otimes n}, \mathbb{P}_{\overrightarrow{\sigma},0}^{\otimes n})]\frac{1}{2}\int_{B_j} Q_0(dx) \\
&\geq \sum_{j=1}^k \mathbb{E}_{\nu^{\otimes(k-1)}}[1-\sqrt{\chi^2(\mathbb{P}_{\overrightarrow{\sigma},1}^{\otimes n}, \mathbb{P}_{\overrightarrow{\sigma},0}^{\otimes n})}]\frac{1}{2}\int_{B_j} Q_0(dx) \\
&= \sum_{j=1}^k [1-\sqrt{(1+\chi^2(\mathbb{P}_1,\mathbb{P}_0))^n-1})\frac{1}{2}]\int_{B_j} Q_0(dx), \qquad (5.12)
\end{aligned}$$

where $\mathbb{P}_i$, $i\in\{0,1\}$ is the law of $Z^{(1)}$ when $f = f_{\overrightarrow{\sigma}}$ with $\overrightarrow{\sigma} = (i,1,\ldots,1)$, $i\in\{0,1\}$, $\mathbb{V}(P,Q)$ is the total variation distance between distribution P and Q and $\chi^2(P,Q)$ is the χ^2 divergence between P and Q. Then we can write, if $\chi^2(\mathbb{P}_1,\mathbb{P}_0)\leq \frac{C}{n}$:

$$\sup_{\overrightarrow{\sigma}\in\{0,+1\}^k} \mathbb{E}_{f_{\overrightarrow{\sigma}},g_0} d_\Delta(\hat{G}_{n,m}, G_K^\star) \geq c'\sum_{j=1}^k \int_{B_j} Q_0(dx) = c'k\omega, \qquad (5.13)$$

where we use the definition of Q_0.

Next step is to find a satisfying upper bound for $\chi^2(\mathbb{P}_1,\mathbb{P}_0)$. We have, by construction of $f_{\overrightarrow{\sigma}}$:

$$\begin{aligned}
\chi^2(\mathbb{P}_1,\mathbb{P}_0) &= \int \frac{[(f_{\overrightarrow{\sigma},1}-f_{\overrightarrow{\sigma},0})\mu*\eta]^2}{f_{\overrightarrow{\sigma},0}*\eta}dx, \\
&\leq \int \frac{[(f_{\overrightarrow{\sigma},1}-f_{\overrightarrow{\sigma},0})\mu_0*\eta]^2}{f_{\overrightarrow{\sigma},0}\mu*\eta}dx + \int \frac{[(f_{\overrightarrow{\sigma},1}-f_{\overrightarrow{\sigma},0})\mu_1*\eta]^2}{f_{\overrightarrow{\sigma},0}\mu*\eta}dx.
\end{aligned}$$

The right hand side term can be considered as negligible with a good choice of the parameters a and b. Hence, we concentrate on the first one. First remark that for all $x\in\mathbb{R}^2$, for some $C>0$:

$$f_{\overrightarrow{\sigma},0}\mu*\eta \geq \frac{C}{(1+x_1^2)(1+x_2^2)},\ \forall x\in\mathbb{R}^2,\ \text{and}\ \left\{(f_{\overrightarrow{\sigma},+1}-f_{\overrightarrow{\sigma},0})\mu_0\right\}*\eta = q^{-\gamma}k\omega\{\psi_l\rho\}*\eta(x).$$

Then,

$$\begin{aligned}
\chi^2(\mathbb{P}_1,\mathbb{P}_0) &= \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{\{(f_{\omega_{11}}-f_{\omega_{10}})*\eta(x)\}^2}{f_{\omega_{11}}*\eta(x)}dx, \\
&\leq Cq^{-2\gamma}k\omega\int_{\mathbb{R}}\int_{\mathbb{R}}(1+x_1^2)(1+x_2^2)\{\psi_1\rho*\eta(x)\}^2dx.
\end{aligned}$$

Hence:

$$\begin{aligned}
\chi^2(\mathbb{P}_1,\mathbb{P}_0) &\leq Cq^{-2\gamma}k\omega\int_{\mathbb{R}}\int_{\mathbb{R}}\{\psi_1\rho*\eta(x)\}^2dx + Cq^{-2\gamma}k\omega\int_{\mathbb{R}}\int_{\mathbb{R}}x_2^2\{\psi_1\rho*\eta(x)\}^2dx \\
&\quad + Cq^{-2\gamma}k\omega\int_{\mathbb{R}}\int_{\mathbb{R}}x_1^2\{\psi_1\rho*\eta(x)\}^2dx + Cq^{-2\gamma}k\omega\int_{\mathbb{R}}\int_{\mathbb{R}}x_1^2x_2^2\{\psi_1\rho*\eta(x)\}^2dx, \\
&:= A_1+A_2+A_3+A_4.
\end{aligned}$$

In the following, we only consider the bound of $A_1 = Ck\omega q^{-2\gamma}\|(\psi_1\rho) * \eta\|^2$, the other terms being controlled in the same way. From the definition of ψ_1 and the conditions on η, we get

$$\begin{aligned}\|(\psi_1\rho) * \eta\|^2 &= \int (\psi_1\rho) * \eta(x)^2 dx = \prod_{i=1}^{2} \int |\mathcal{F}[\psi_1\rho](t_i)|^2 \, |\mathcal{F}[\eta_i](t_i)|^2 \, dt_i \\ &= \prod_{i=1}^{2} \int |\mathcal{F}[\rho(2\pi q\cdot)\rho](t_i - 4\pi q)|^2 \, |\mathcal{F}[\eta_i](t_i)|^2 \, dt_i.\end{aligned}$$

Using (5.2), the noise assumption, and the fact that $q \to +\infty$, we get

$$\begin{aligned}\|(\psi_1\rho) * \eta\|^2 &= Cq^{-2(\beta_1+\beta_2)} \prod_{i=1}^{2} \int |\mathcal{F}[\rho(2\pi q\cdot)\rho](t_i - 4\pi q)|^2 \, dt_i, \\ &= Cq^{-2(\beta_1+\beta_2)} \|\rho(2\pi q.)\rho\|^2, \\ &\leq Cq^{-2(\beta_1+\beta_2)} \|\rho(2\pi q.)\|^2 \leq Cq^{-2(\beta_1+\beta_2)-2}.\end{aligned}$$

Similar bounds are available for A_2, A_3 and A_4 as follows. First note that for all $t \in \mathbb{R}$:

$$\mathcal{F}[\psi_1\rho](t) = c_\psi q^{-\gamma} \mathcal{F}[\rho(2\pi q\cdot)\rho(\cdot)](t \pm 4\pi q),$$

and

$$\frac{d}{dt}\mathcal{F}[\psi_1\rho](t) = -(ic_\psi q^{-\gamma})^2 t.\mathcal{F}[\rho(2\pi q\cdot)\rho(\cdot)](t \pm 4\pi q),$$

for all t in a subset of $\mathbb{R}$ having a Lebesgue measure equal to 1. Then since $\mathcal{F}[\rho]$ and its weak derivative are bounded by 1 and supported on $[-1; 1]$, we have for instance for A_2:

$$\begin{aligned}A_2 &= Cq^{-2\gamma} k\omega \int_{\mathbb{R}} \int_{\mathbb{R}} x_2^2 \{\psi_1\rho * \eta(x)\}^2 dx \\ &\leq Cq^{-2\gamma} k\omega \int_{\mathbb{R}} \int_{\mathbb{R}} \left(\frac{d}{dx_2} \mathcal{F}[\psi_1\rho](x) \mathcal{F}[\eta](x) \right)^2 dx,\end{aligned}$$

which leads to the same asymptotics as in A_1. It leads to the following upper bound in the general d-dimensional case:

$$\chi^2(\mathbb{P}_1, \mathbb{P}_0) \leq Cq^{-2\gamma-\alpha\gamma-d-2(\beta_1+\beta_2)} \leq \frac{C}{n}, \text{ with } q = n^{\frac{1}{2\gamma+\alpha\gamma+d+2(\beta_1+\beta_2)}}. \tag{5.14}$$

Now using (5.13),

$$\sup_{\sigma\in\{0,1\}^k} \mathbb{E}_{f_{\vec{\sigma}}} d_\Delta(\hat{G}_{n,m}, G_K^\star) \geq c' k\omega = c' q^{-\alpha\gamma} = c' n^{\frac{-\alpha\gamma}{2\gamma+\alpha\gamma+d+2(\beta_1+\beta_2)}},$$

which concludes the proof of the lower bound.

5.2 Proof of Theorem 2

The proof is presented for $d = 2$ for simplicity whereas straightforward modifications lead to the $d-$dimensional case. In the sequel, we identify each $\nu \in \Sigma(\gamma, L)$ with a set $G_\nu = \{x : \nu(x) \geq 0\}$. By the same way, we identify $G_K^\star$ with $\nu^\star = f - g$. Moreover, we assume for simplicity that $n \leq m$.

5.2.1 A first inequality

For all $G_\nu := \{\nu \geq 0\}$, we have, using the notations of Section 3:

$$\begin{aligned} & R^\lambda_{n,m}(G_\nu) - R^\lambda_{n,m}(G^\star_K) - R^\lambda_K(G_\nu) + R^\lambda_K(G^*_K) \\ = \quad & \frac{1}{2n}\sum_{i=1}^n U_i(G_\nu) + \frac{1}{2m}\sum_{i=1}^n V_i(G_\nu) := \frac{1}{2}T_{n,m}(G), \end{aligned}$$

where, for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,m\}$,

$$U_i(G_\nu) = \{h_{K/G^\star_K,\lambda}(Z^{(1)}_i) - h_{G^C_\nu,\lambda}(Z^{(1)}_i)\} - \mathbb{E}[h_{K/G^\star_K,\lambda}(Z^{(1)}_i) - h_{G^C_\nu,\lambda}(Z^{(1)}_i)],$$

and

$$V_j(G_\nu) = \{h_{G^\star_K,\lambda}(Z^{(2)}_j) - h_{G_\nu,\lambda}(Z^{(2)}_j)\} - \mathbb{E}[h_{G^\star_K,\lambda}(Z^{(2)}_j) - h_{G_\nu,\lambda}(Z^{(2)}_j)].$$

Then, for all $i \in \{1,\dots,n\}$, using successively Lemma 2 in Appendix and the margin assumption (Lemma 2 in [27]) we get:

$$\mathbb{E}[U_i(G_\nu)]^2 \quad \leq \quad c\lambda_1^{-2\beta_1}\lambda_2^{-2\beta_2} d_\Delta(G_\nu, G^\star_K) \leq c'\lambda_1^{-2\beta_1}\lambda_2^{-2\beta_2} d_{f,g}(G_\nu, G^\star_K)^{\frac{\alpha}{\alpha+1}},$$

and

$$|U_i(G_\nu)| \leq C\prod_{i=1}^2 \lambda_i^{-\beta_i - 1/2},$$

for some constant $C > 0$. The Bernstein's inequality leads to

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n U_i(G_\nu)\right| > a\right) \leq 2\exp\left[-\frac{Cna^2}{a\times\lambda_1^{-\beta_1-1/2}\lambda_2^{-\beta_1-1/2} + \lambda_1^{-2\beta_1}\lambda_2^{-2\beta_2} d_{f,g}(G_\nu, G^\star_K)^{\frac{\alpha}{\alpha+1}}}\right],$$

for all $a > 0$. Since $\beta_i > 1/2$ for all $i \in \{1,\dots,d\}$, the particular choice $a = d_{f,g}(G_\nu, G^\star_K)$ yields

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^n U_i(G_\nu)\right| > d_{f,g}(G_\nu, G^\star_K)\right) \quad & \leq \quad 2\exp\left[-Cn\lambda_1^{2\beta_1}\lambda_2^{2\beta_2} d_{f,g}(G_\nu, G^\star_K)^{2-\frac{\alpha}{\alpha+1}}\right] \\ & = \quad 2\exp\left[-Cn\lambda_1^{2\beta_1}\lambda_2^{2\beta_2} d_{f,g}(G_\nu, G^\star_K)^{\frac{2+\alpha}{\alpha+1}}\right]. \end{aligned}$$

In the upper bound above, we have implicitly use the fact that $d_{f,g}(G_\nu, G^\star_K)/2 \leq (d_{f,g}(G_\nu, G^\star_K)/2)^{\alpha/(\alpha+1)}$ since $d_{f,g}(G_1, G_2) \leq 2$ for all $G_1, G_2 \subset K$. Using the same algebra on the $V_j(G_\nu)$, we get

$$P\left(|T_{n,m}(G_\nu)| > d_{f,g}(G_\nu, G^\star_K)\right) \leq 2\exp\left[-Cn\lambda_1^{2\beta_1}\lambda_2^{2\beta_2} d_{f,g}(G_\nu, G^\star_K)^{\frac{2+\alpha}{\alpha+1}}\right].$$

This concludes the first part of the proof. Let t a positive parameter which will be chosen further and introduce the set $\mathcal{G}'$ defined as

$$\mathcal{G}' = \left\{G \in \mathcal{N}_{\delta_n}, d_{f,g}(G^\star_K, G) > t\delta_n^{1+\alpha}\right\},$$

where $\mathcal{N}_{\delta_n}$ is the δ_n network introduced in Section 3.2, with $\delta = \delta_n := \delta_{n,n}$. Using the upper bound above,

$$\begin{aligned} P\left(\exists G \in \mathcal{G}' : |T_{n,m}(G)| \geq \frac{1}{4}d_{f,g}(G, G^\star_K)\right) \quad & \leq \quad \sum_{G\in\mathcal{G}'} P\left(|T_{n,m}(G)| \geq \frac{1}{4}d_{f,g}(G, G^\star_K)\right) \\ & \leq \quad \sum_{G\in\mathcal{G}'} 2\exp\left[-Cn\lambda_1^{2\beta_1}\lambda_2^{2\beta_2} d_{f,g}(G, G^\star_K)^{\frac{2+\alpha}{\alpha+1}}\right] \\ & \leq \quad \sum_{G\in\mathcal{G}'} 2\exp\left[-Cn\lambda_1^{2\beta_1}\lambda_2^{2\beta_2}\left(t\delta_n^{1+\alpha}\right)^{\frac{2+\alpha}{\alpha+1}}\right] \\ & \leq \quad \sum_{G\in\mathcal{G}'} 2\exp\left[-Cn\lambda_1^{2\beta_1}\lambda_2^{2\beta_2} t^{\frac{2+\alpha}{\alpha+1}}\delta_n^{2+\alpha}\right]. \end{aligned}$$

Since $\log \mathrm{card}(\mathcal{N}_{\delta_n}) \leq A\delta_n^{-2/\gamma}$, we get

$$P\left(\exists G \in \mathcal{G}' : |T_{n,m}(G)| \geq \frac{1}{4} d_{f,g}(G, G_K^\star)\right) \leq \exp\left[A\delta_n^{-2/\gamma} - Cn\lambda_1^{2\beta_1}\lambda_2^{2\beta_2} t^{\frac{2+\alpha}{\alpha+1}} \delta_n^{2+\alpha}\right].$$

Thanks to the value of δ_n, we get $\delta_n^{-2/\gamma} \simeq n\lambda_1^{2\beta_1}\lambda_2^{2\beta_2}\delta_n^{2+\alpha}$. Hence, for t large enough,

$$\begin{aligned} P\left(\exists G \in \mathcal{G}' : |T_{n,m}(G)| \geq \frac{1}{4} d_{f,g}(G, G_K^\star)\right) &\leq \exp\left[-Ct\delta_n^{-2/\gamma}\right] \\ &= \exp\left[-Ct\left(\frac{\lambda_1^{-\beta_1}\lambda_2^{-\beta_2}}{\sqrt{n}}\right)^{-\frac{2}{\gamma}\frac{2}{2/\gamma+2+\alpha}}\right]. \quad (5.15) \end{aligned}$$

Now, using Lemma 1 in Appendix, we can find a set $G_n \in \mathcal{N}_{\delta_n}$ such that:

$$d_{f,g}(G_n, G_K^\star) \leq c_2 \|\nu^* - \nu_n\|_\infty^{\alpha+1} \leq c_2 \delta_n^{1+\alpha}.$$

Then, for all $G \in \mathcal{G}'$, we get

$$\frac{1}{8} d_{f,g}(G, G_K^\star) - \frac{3}{4} d_{f,g}(G_n, G_K^\star) \geq \frac{t}{8}\delta_n^{1+\alpha} - \frac{3c_2}{4}\delta_n^{1+\alpha} \geq \frac{c_2}{4}\delta_n^{1+\alpha},$$

provided that $t > 8c_2$. We eventually obtain:

$$\begin{aligned} & P\left(d_{f,g}(G_K^\star, \hat{G}_{n,m}) > t\delta_n^{1+\alpha}\right) \leq P\left(\exists G \in \mathcal{G}' : R_{n,m}^\lambda(G) \leq R_{n,m}^\lambda(G_n)\right) \\ = \; & P\left(\exists G \in \mathcal{G}' : \frac{1}{2} d_{f,g}^\lambda(G, G_K^\star) + T_{n,m}(G) - \frac{1}{2} d_{f,g}^\lambda(G_n, G_K^\star) - T_{n,m}(G_n) \leq 0\right) \quad (5.16) \end{aligned}$$

where for all $G_1, G_2 \subset K$,

$$\frac{1}{2} d_{f,g}^\lambda(G_1, G_2) := R_K^\lambda(G_1) - R_K^\lambda(G_2).$$

5.2.2 Control of the bias

Last step is to control the bias term. In particular, given $G_1, G_2 \subset K$, we want to measure the difference between $R_K(G_1) - R_K(G_2)$ and $R_K^\lambda(G_1) - R_K^\lambda(G_2)$. First of all, we have to explicit the term R_K^λ. Recall that for all $G_1 \subset K$,

$$\begin{aligned} 2R_K^\lambda(G_1) &:= 2\mathbb{E}R_{n,m}^\lambda(G_1), \\ &= \mathbb{E}[h_{K/G_1,\lambda}(Z_1^{(1)})] + \mathbb{E}[h_{G_1,\lambda}(Z_1^{(2)})], \\ &= \mathbb{E}\left[\int_{K/G_1} \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_1^{(1)} - x}{\lambda}\right) dx\right] + \mathbb{E}\left[\int_{G_1} \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{Z_1^{(2)} - x}{\lambda}\right) dx\right], \\ &= \int_{K/G_1} \mathbb{E}\left[\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{X_1^{(1)} + \epsilon_1^{(1)} - x}{\lambda}\right)\right] dx + \int_{G_1} \mathbb{E}\left[\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{X_1^{(2)} + \epsilon_1^{(2)} - x}{\lambda}\right)\right] dx. \end{aligned}$$

Using the properties of the deconvolution kernel, we can see that for all $x \in K$,

$$\mathbb{E}\left[\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{X_1^{(1)} + \epsilon_1^{(1)} - x}{\lambda}\right)\right] = \mathbb{E}\left[\frac{1}{\lambda}\mathcal{K}\left(\frac{X_1^{(1)} - x}{\lambda}\right)\right] = \int_{\mathbb{R}^d} \frac{1}{\lambda}\mathcal{K}\left(\frac{y-x}{\lambda}\right) f(y) dQ(y).$$

The same result holds true when replacing $X_1^{(1)}$ by $X_1^{(2)}$ and f by g. Hence, we obtain that

$$2R_K^\lambda(G_1) = \int_{K/G_1}\int_{\mathbb{R}^d} \frac{1}{\lambda}\mathcal{K}\left(\frac{y-x}{\lambda}\right) f(y) dQ(y) dx + \int_{G_1}\int_{\mathbb{R}^d} \frac{1}{\lambda}\mathcal{K}\left(\frac{y-x}{\lambda}\right) g(y) dQ(y) dx.$$

Moreoevr, if Q is not the Lebesgue measure, note that by assumption, there exists a constant $c_0 > 0$ such that:

$$\int_{G_1 \Delta G_2} dx \leq c_0^{-1} d_\Delta(G_1, G_2). \tag{5.17}$$

We then have:

$$\begin{aligned}
&\left|(R_K^\lambda - R_K)(G_1 - G_2)\right| \\
&\leq \frac{1}{2}\left|\int\left[\int \frac{1}{\lambda}\mathcal{K}\left(\frac{y-x}{\lambda}\right) f(y)\mu(y)dy - f(x)\mu(x)\right]\left[\mathbf{1}(x \in K/G_1) - \mathbf{1}(x \in K/G_2)\right] dx\right. \\
&\quad \left. + \int\left[\int \frac{1}{\lambda}\mathcal{K}\left(\frac{y-x}{\lambda}\right) g(y)\mu(y)dy - g(x)\mu(x)\right]\left[\mathbf{1}(x \in G_1) - \mathbf{1}(x \in G_2)\right] dx\right|, \\
&\leq \frac{1}{2}\int_{G_1 \Delta G_2} |\mathcal{K}_\lambda * (\nu^\star.\mu)(x) - \nu^\star.\mu(x)|\, dx, \\
&\leq \frac{1}{2c_0}\|\mathcal{K}_\lambda * (\nu^\star.\mu) - \nu^\star.\mu\|_\infty \int_{G_1 \Delta G_2} dx, \\
&\leq C d_\Delta(G_1, G_2)\left[\lambda_1^\gamma + \lambda_2^\gamma\right], \\
&\leq C\left[\lambda_1^\gamma + \lambda_2^\gamma\right] d_{f,g}(G_1, G_2)^{\frac{\alpha}{\alpha+1}},
\end{aligned}$$

for some $C > 0$, where $\mathcal{K}_\lambda(.) = \frac{1}{\lambda}\mathcal{K}(./\lambda)$. Indeed, provided that $\nu\mu \in \Sigma(\gamma, L)$ and $\mathcal{K}$ is a kernel of order $l = \lfloor\gamma\rfloor$, it is well-known that:

$$\|\mathcal{K}_\lambda * (\nu\mu) - \nu\mu\|_\infty \leq C\left[\lambda_1^\gamma + \lambda_2^\gamma\right]. \tag{5.18}$$

Using the Young inequality:

$$xy^r \leq ry + (1-r)x^{1/(1-r)}, \ \forall x, y \in \mathbb{R}^+,$$

with $r = \alpha/(\alpha+1)$, $x = C\kappa^{-\alpha/\alpha+1}\left[\lambda_1^\gamma + \lambda_2^\gamma\right]$ and $y = \kappa d_{f,g}(G_1, G_2)$, where $\kappa > 0$ is chosen later on, we get for all $G_1, G_2 \subset K$:

$$\left|(R_K^\lambda - R_K)(G_1 - G_2)\right| \leq \left(1 - \frac{\alpha}{\alpha+1}\right)\left(\frac{C}{\kappa}\right)^\alpha \left[\lambda_1^\gamma + \lambda_2^\gamma\right]^{\alpha+1} + \frac{\alpha}{\alpha+1}\kappa d_{f,g}(G_1, G_2) \tag{5.19}$$

5.2.3 Conclusion of the proof

Hence, it follows from (5.16) and (5.19) that:

$$\begin{aligned}
&P\left(d_{f,g}(G_K^\star, \hat{G}_{n,m}) > t\delta_n^{1+\alpha}\right) \\
&\leq P\left(\exists G \in \mathcal{G}' : \left(\frac{1}{2} - \frac{\alpha}{\alpha+1}\kappa\right) d_{f,g}(G, G_K^\star) + T_{n,m}(G)\right. \\
&\qquad \left. - \left(\frac{1}{2} + \frac{\alpha}{\alpha+1}\kappa\right) d_{f,g}(G_n, G_K^\star) - T_{n,m}(G_n) + C\sum_{i=1}^2 \lambda_i^{\gamma(1+\alpha)} \leq 0\right), \\
&\leq P\left(\exists G \in \mathcal{G}' : T_{n,m}(G) \leq -\frac{1}{8} d_{f,g}(G, G_K^\star)\right) + P\left(T_{n,m}(G_n) \geq C\left(\delta_n^{1+\alpha} + \sum_{i=1}^2 \lambda_i^{\gamma(1+\alpha)}\right)\right),
\end{aligned}$$

where we have chosen $\kappa < \frac{\alpha+1}{4\alpha}$ in (5.19). In order to conclude, remark that the proposed choice of $(\lambda_j)_{j=1,\ldots,d}$ provides:

$$\delta_n^{1+\alpha} \simeq \sum_{i=1}^2 \lambda_i^{\gamma(1+\alpha)} \Leftrightarrow \forall i \in \{1,2\}, \left(\frac{\lambda_1^{-\beta_1}\lambda_2^{-\beta_2}}{\sqrt{n}}\right)^{\frac{2\gamma(\alpha+1)}{\gamma(2+\alpha)+2}} \simeq \lambda_i^{\gamma(\alpha+1)}.$$

Using (5.15), we eventually get

$$
\begin{aligned}
&P\left(d_{f,g}(G_K^\star,\hat{G}_{n,m}) > tn^{-\frac{\gamma(\alpha+1)}{\gamma(2+\alpha)+d+2\sum_{i=1}^d\beta_i}}\right)\\
&\quad\leq \quad \exp\left[-C_1 tn^{\frac{1}{\gamma(2+\alpha)+d+2\sum_{i=1}^d\beta_i}}\right]+\exp\left[-C_2 n^{\frac{1}{\gamma(2+\alpha)+d+2\sum_{i=1}^d\beta_i}}\right],
\end{aligned}
$$

where C_1, C_2 denote positive constants. In order to conclude, we can remark that

$$
\begin{aligned}
&n^{\tau_d(\alpha,\beta,\gamma)}\mathbb{E}_{f,g}d_{f,g}(G_K^\star,\hat{G}_{n,m})\\
&\quad\leq \quad t+\mathbb{E}_{f,g}d_{f,g}(G_K^\star,\hat{G}_{n,m})\mathbf{1}_{\left\{d_{f,g}(G_K^\star,\hat{G}_{n,m})>tn^{-\frac{\gamma(\alpha+1)}{\gamma(2+\alpha)+d+2\sum_{i=1}^d\beta_i}}\right\}},\\
&\quad\leq \quad t+2\exp\left[-C_1 tn^{\frac{1}{\gamma(2+\alpha)+d+2\sum_{i=1}^d\beta_i}}\right]+2\exp\left[-C_2 n^{\frac{1}{\gamma(2+\alpha)+d+2\sum_{i=1}^d\beta_i}}\right]\leq C
\end{aligned}
$$

for some positive constant C, where we have used the bound $d_{f,g}(G_1,G_2)\leq 2$ for all $G_1, G_2 \subset K$.

□

5.3 Proof of Corollary 1

The proof follows the same steps as the proof of Theorem 2. Note that in the direct case, using a kernel $\mathcal{K}$ with bounded Fourier transform, we have under the margin assumption:

$$
\mathbb{E}[U_i(G)^2]\leq Cd_\Delta(G,G_K^\star)\leq C' d_{f,g}(G,G_K^\star)^{\frac{\alpha}{\alpha+1}},\ \text{and}\ |U_i(G)|\leq C,
$$

for some constant $C>0$. Remark that the last inequality is more precise than in the error-in-variable case. Then using Bernstein's inequality, we have exactly as in the proof of Theorem 2:

$$
P\left(\left|\frac{1}{n}\sum_{i=1}^n U_i(G_\nu)\right|>a\right)\leq 2\exp\left[-\frac{Cna^2}{a+d_\Delta(G_\nu,G_K^*)}\right],
$$

for all $a>0$. Choosing $a=d_{f,g}(G_\nu,G_K^\star)$ and using the same algebra, we get a control of the upper bound provided that:

$$
\delta_n^{-\frac{2}{\gamma}}\simeq n\delta_n^{2+\alpha}\ \text{and}\ \delta_n^{1+\alpha}\geq\sum_{i=1}^2\lambda_i^{\gamma(1+\alpha)}.
$$

The choice of λ and δ_n in Corollary 1 concludes the proof.

□

6 Appendix

Lemma 1 *For any (f,g) satisfying the margin assumption with parameter $\alpha>0$, we have:*

$$
d_{f,g}(G_\nu,G_K^\star)\leq c_2\|\nu-\nu^\star\|_\infty^{\alpha+1},
$$

where $G_\nu=\{\nu\geq 0\}$ and $\nu^\star=f-g$.

PROOF. The proof is a straightforward modification of the proof of Lemma 5.1 in [2] which state a similar result in the binary classification framework. In the following, given $x \in \mathbb{R}$, we write $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = 0$ if $x = 0$, and $\text{sign}(x) = -1$ if $x < 0$. Then, we get

$$\begin{aligned} d_{f,g}(G_\nu, G_K^\star) &= \int_K |\nu^\star(x)| \mathbf{1}_{\left\{x \in G_K^\star \Delta G_\nu\right\}} dQ(x), \\ &= \int_K |\nu^\star(x)| \mathbf{1}_{\{\text{sign}(\nu^\star(x)) \neq \text{sign}(\nu(x))\}} dQ(x), \\ &\leq \int_K |\nu^\star(x)| \mathbf{1}_{\{0 < |\nu^\star(x)| \leq |\nu(x) - \nu^\star(x)|\}} dQ(x), \\ &\leq \|\nu - \nu^\star\|_\infty Q\left(\{x \in K : 0 < |\nu^\star(x)| \leq \|\nu(x) - \nu^\star(x)\|_\infty\}\right) \leq c_2 \|\nu - \nu^\star\|_\infty^{\alpha+1}, \end{aligned}$$

where we have used the Margin Assumption in order to get the last inequality.

□

Lemma 2 *Assume that η satisfies the Noise assumption . Let $\mathcal{K}_\eta$ a deconvolution kernel defined in (3.1) such that $\mathcal{F}[\mathcal{K}]$ is bounded and compactly supported. Iif $Q(x) = \mu(x)dx$, we assume that $\mu \in \Sigma(\gamma, L)$ and that $\max_{x \in K} \mu(x) / \min_{x \in K} \mu(x) \leq c_0$ for some $c_0 > 0$. Then, we have,*

$$(i) \qquad \mathbb{E}[h_{G,\lambda}(Z) - h_{G',\lambda}(Z)]^2 \leq C d_\Delta(G, G') \prod_{i=1}^d \lambda_i^{-2\beta_i},$$

$$(ii) \qquad \sup_{x \in K} |h_{G,\lambda}(x) - h_{G',\lambda}(x)| \leq C \prod_{i=1}^d \lambda_i^{-\beta_i - 1/2},$$

for some positive constant C.

PROOF. For the sake of convenience, we only consider the case where $d = 1$. We first prove (i). We have, using (5.17):

$$\begin{aligned} \mathbb{E}[h_{G,\lambda}(Z) - h_{G',\lambda}(Z)]^2 &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \frac{1}{\lambda} \mathcal{K}_\eta \left(\frac{z - x}{\lambda} \right) (\mathbf{1}_{\{x \in G\}} - \mathbf{1}_{\{x \in G'\}}) \mathbf{1}_{\{x \in K\}} dx \right]^2 (f\mu) * \eta(z) dz, \\ &\leq c \int_{\mathbb{R}} \frac{1}{\lambda^2} \left| \mathcal{F}[\mathcal{K}_\eta(./\lambda)](t) \right|^2 \left| \mathcal{F}[(\mathbf{1}_{\{.\in G\}} - \mathbf{1}_{\{.\in G'\}}) \mathbf{1}_{\{.\in K\}}](t) \right|^2 dt, \\ &\leq C \max_{x \in \mathbb{R}^d} \mu(x) \times \lambda^{-2\beta} \int_K \mathbf{1}_{\{t \in G \Delta G'\}} dt, \\ &\leq C \lambda^{-2\beta} d_\Delta(G, G'). \end{aligned}$$

Indeed, for all $s \in \mathbb{R}$, using assumptions on the kernel $\mathcal{K}_\eta$:

$$\frac{1}{\lambda^2} \left| \mathcal{F}[\mathcal{K}_\eta(./\lambda)](s) \right|^2 = \left| \mathcal{F}[\mathcal{K}_\eta](s\lambda) \right|^2 = \left| \frac{\mathcal{F}[\mathcal{K}](s\lambda)}{\mathcal{F}[\eta](s)} \right|^2 \leq C \sup_{s \in [-M/\lambda, M/\lambda]} \left| \frac{1}{\mathcal{F}[\mathcal{K}_\eta](s)} \right|^2 \leq C \lambda^{-2\beta}, \quad (6.1)$$

where $\mathcal{F}[\mathcal{K}] = 0$ on $\mathbb{R} \setminus [-M, M]$.

In order to prove (ii), we use the following algebra

$$\begin{aligned}
\sup_{z\in\mathbb{R}} |h_{G,\lambda}(z) - h_{G',\lambda}(z)| &\leq \sup_{z\in\mathbb{R}} \int_{G\Delta G'} \frac{1}{\lambda} \left| \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) \right| dx, \\
&\leq C \sup_{z\in\mathbb{R}} \int_K \frac{1}{\lambda} \left| \mathcal{K}_\eta \left(\frac{z-x}{\lambda} \right) \right| dx, \\
&\leq C \sup_{z\in\mathbb{R}} \sqrt{\int \frac{1}{\lambda^2} \mathcal{K}_\eta^2 \left(\frac{z-x}{\lambda} \right) dx} \\
&\leq C\lambda^{-1/2} \sqrt{\int_{[-M,M]} \left| \frac{\mathcal{F}[\mathcal{K}](t)}{\mathcal{F}[\eta](t/\lambda)} \right|^2 dt} \\
&\leq C\lambda^{-\beta-1/2},
\end{aligned}$$

where last line uses the noise assumption and assumptions on the kernel $\mathcal{K}_\eta$.

□

References

[1] J-Y. Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. Preprint, Laboratoire de Probabilités et Modéles Aléatoires, Univ. Paris VI and VII., 2004.

[2] J-Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of statistics*, 35:608–633, 2007.

[3] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

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

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

[6] Peter J. Bickel and Ya'acov Ritov. Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, 31(4):1033–1053, 2003.

[7] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

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

[9] P.J. Carroll, A. Delaigle, and P. Hall. Nonparametric prediction in measurement error models. *Journal of the American Statistical Asociation*, 104:993–1003, 2009.

[10] Olivier Chapelle, Jason Weston, Léon Bottou, L Eon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *Advances in Neural Information Processing Systems*, pages 416–422. MIT Press, 2001.

[11] A. Delaigle and I. Gijbels. Estimation of boundary and discontinuity points in deconvolution problems. *Statistica Sinica*, 16:773–788, 2006.

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

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

[14] W.H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Kluwer, 2000.

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

[16] J. Fan and Y.K Truong. Nonparametric regression with errors in variables. *Annals of Statistics*, 21 (4):1900–1925, 1993.

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

[18] Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry A. Wasserman. Minimax manifold estimation. *CoRR*, abs/1007.0549, 2010.

[19] Larry Goldstein and Karen Messer. Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.*, 20(3):1306–1328, 1992.

[20] J Klemela and E. Mammen. Empirical risk minimization in inverse problems. *Annals of Statistics*, 38 (1):482–511, 2010.

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

[22] A.P. Korostelev and A.B. Tsybakov. *Minimax theory of Image Reconstruction. Lecture Notes in Statistics.* Springer Verlag, 1993.

[23] 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.

[24] J.M. Loubes and C. Marteau. Goodness-of-fit strategies from indirect observations. *Working paper.*

[25] S. Loustau. Penalized empirical risk minimization over besov spaces. *Electronic journal of Statistics*, 3:824–850, 2009.

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

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

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

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

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

[31] Y. Tang. Minimax nonparametric classification - part i: Rates of convergence, part ii: Model selection for adaptation. *IEEE Trans. Inf. Theory*, 45:2271–2292, 1999.

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

[33] 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.

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

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