Noisy classification with boundary assumptions

Sébastien Loustau, Clément Marteau

Noisy classification with boundary assumptions

Sébastien Loustau* and Clément Marteau†

Abstract

We address the problem of classification when data are collected from two samples with measurement errors. This problem turns to be an inverse problem and requires a specific treatment. In this context, we investigate the minimax rates of convergence using both a margin assumption, and a smoothness condition on the boundary of the set associated to the Bayes classifier. We establish lower and upper bounds (based on a deconvolution classifier) on these rates.

1 Introduction

Assume that we have at our disposal two *noisy* learning samples $\mathcal{S}_1 = (Z_1^{(1)}, \ldots, Z_n^{(1)})$ and $\mathcal{S}_2 = (Z_1^{(2)}, \ldots, Z_n^{(2)})$ satisfying:

$$Z_i^{(1)} = X_i^{(1)} + \epsilon_i^{(1)},\ \forall i \in \{1, \ldots, n\},\ \text{and}\ Z_j^{(2)} = X_j^{(2)} + \epsilon_j^{(2)},\ \forall j \in \{1, \ldots, m\}, \qquad (1.1)$$

where the $X_i^{(1)}$ (resp. $X_j^{(2)}$) denote independent identically distributed (i.i.d) random variables from unknown distribution F_1 (resp. F_2) and the $\epsilon_i^{(k)}$ denote random errors, independent of the $X_i^{(k)}$. For the sake of simplicity, F_1 (resp. F_2) admits a density f (resp. g) w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Moreover, we assume that the random errors are i.i.d. with *known* density η with respect to the Lebesgue measure.

The goal is to classify a new incoming observation X, assumed to have a density f or g (and independent of $\mathcal{S}_1$ and $\mathcal{S}_2$). In other words, one wants to determine whether the density of X is f or g. Remark that in this model, two independent random sources are involved. The first one corresponds to the fluctuations of the variable of interest which is governed by the distribution F_1 or F_2 following the corresponding label. The second one corresponds to measurement errors (or imprecisions) during the data collection process. This problem corresponds to a nonparametric measurement error model, or errors-in-variables model (see the monograph of [24] for an introduction). We are in fact faced to the so-called discrimination with errors in variables problem.

The free noise case has already been widely studied in the literature. We refer for instance to [11] for a complete survey. When the variables ϵ_i^j are equal to zero in (1.1), fast rates of convergence were obtain for the first time in [22], using both a complexity and a margin assumption. Similar results where obtained in [26, 23, 2] in slightly different settings. The previous papers were focused on empirical risk minimisation (ERM) algorithm and used margin assumptions (see Section for more details). Similar conditions were investigated in the last decades for instance in [3], [4] or [16] among other.

Concerning the error-in-variables model of classification, few results have been published. Up to our knowledge, the only minimax result is [21], where minimax fast rates were obtained

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

†Institut de Mathématiques de Toulouse, marteau@math.univ-toulouse.fr

using a regularity assumption on the class of densities f and g in model (1.1) (see also [19] in the general context of statistical learning). In a slightly different setting, we can however mention [9] where boundary estimation in a deconvolution framework was considered, or [14] for a study of empirical risk minimization (ERM) algorithm in inverse problem models. Finally, [15] proposes an empirical minimization based on a deconvolution approach for the problem of estimating geometric characters of a multivariate distribution in the presence of noisy measurements.

The aim of this paper is to provide a classifier in the error-in-variables model and to study the related minimax performances in terms of fast rates. For this purpose, we will use two different kind of assumptions on the model: a margin assumption and a complexity assumption. The margin assumption will traduce the difficulty to discriminate an observation from another. It has been introduced for the first time in [22]. The second assumption concerns the regularity of the boundary of the set

$$G_K^\star = \left\{ x \in K \subset \mathbb{R}^d, f(x) \geq g(x) \right\},$$

where we restrict the study to a compact subset $K \subset \mathbb{R}^d$. It is widely known in classification that the decision set G_K^* minimizes the so-called Bayes risk:

$$R_K(G) = \frac{1}{2} \left(\int_{K \setminus G} f(x)dx + \int_G g(x)dx \right).$$

Hence, the construction of a good classifier is more or less related to provide a good estimation (in a sense which will be precised later on) of $G_K^\star$. Remark that, contrary to [2] or [21], minimax results are investigated with no restriction over the regularity of f and g.

The structure of this paper is as follows. In Section 2, we present in detail our model and assumptions. Then, we construct a deconvolution classifier. Lower bounds are provided in Section 3. The performances of our classifier are studied in Section 4. Section 5 concludes the paper whereas Section 6 proposes to highlight the main ideas used in the proofs. Section 7 is dedicated to the proofs of the main results, whereas Section 8 adds some useful materials about noisy empirical processes.

2 A deconvolution classifier

2.1 Model

In this paper, a classifier is related to a subset $G \subset \mathbb{R}^d$ which traduces some hint on the places where there may be a greater probability to find an observation having distribution F_1. In the sequel, we restrict our investigations to a compact set $K \subset \mathbb{R}^d$. Using a slight abuse of notation, a classifier will be denoted by a measurable subset of the observations $\hat{G} = \hat{G}(\mathcal{S}_1, \mathcal{S}_2)$. In other words, the new incoming observation X will be associated to the first (resp. second) label if it belongs to the set $\hat{G}$ (resp. $K \setminus \hat{G}$).

In order to measure the performances of a given classifier $\hat{G} \subset K$, we will use the Bayes risk defined as:

$$R(\hat{G}) = \frac{1}{2} \left(\int_{K \setminus \hat{G}} f(x)dx + \int_{\hat{G}} g(x)dx \right).$$

The best possible classifier $G_K^\star$ then satisfies

$$G_K^\star = \arg\min_{G \subset K} R_K(G) = \{x \in K, f(x) \geq g(x)\},$$

where the infimum is taken over all possible subset of $K \subset \mathbb{R}^d$. In some sense, a good classifier should mimic (at least asymptotically) the behavior of $G_K^\star$. Hence, the excess risk

$$R_K(\hat{G}) - R_K(G_K^*),$$

will be of first interest all along the paper.

For all $G \subset K$, using simple algebra, we get,

$$\begin{aligned} R_K(G) - R_K(G_K^\star) &= \frac{1}{2}\int_K (f-g)(x)(\mathbf{1}_{G_K^\star}(x) - \mathbf{1}_G(x))dx, \\ &= \frac{1}{2}\int_K |f(x)-g(x)||\mathbf{1}_{G_K^\star}(x) - \mathbf{1}_G(x)|dx, \end{aligned}$$

since the product $(f(x)-g(x)).(\mathbf{1}_{G_K^\star}(x) - \mathbf{1}_G(x))$ is nonnegative for all $x \in K$. Then,

$$R_K(\hat{G}) - R_K(G_K^\star) = \frac{1}{2}\int_K |f(x)-g(x)|\mathbf{1}_{G\Delta G_K^\star}(x)dx := \frac{1}{2}d_{f,g}(G, G_K^\star),$$

where for all $G_1, G_2 \subset K$, $G_1 \Delta G_2 = \{K \setminus G_1 \cap G_2\} \cup \{G_1 \cap K \setminus G_2\}$. The term $d_{f,g}$ is a pseudo distance on the subsets of K. The excess risk corresponds to a measure of the difference between $\hat{G}$ and the Bayes risk $G_K^\star$, where the symmetric difference is balanced by the value of $f-g$. This term is avoided when using for instance the pseudo-distance d_Δ, defined as

$$d_\Delta(G_1, G_2) = Q(G_1 \Delta G_2)\ \forall G_1, G_2 \subset K,$$

where Q denotes the Lebesgue measure on $\mathbb{R}^d$.

Remark 1. The pseudo-distance $d_{f,g}$ is related to the densities f and g whereas d_Δ is entirely determined by the symmetric difference between G_1 and G_2. In some sense, $d_{f,g}$ is more related with the prediction task whereas d_Δ is a more related with a set estimation problem. In some favorable cases, i.e. when $Q(|f-g| \leq t_0) = 0$ for some $t_0 > 0$, it is clear that these two pseudo-distance are equivalent (see the margin assumption below and Lemma 2 in [22]). This particular case is known as the strong margin assumption case.

Remark 2. Our goal is to provide the best possible estimation of the set $G_K^\star$ from two noisy learning samples. From the prediction point of view, we are in fact interested in the estimation of the class of a new incoming observation X. We could also address the following problem: given a new noisy incoming observation, try to guess the corresponding label. These two problems are rather close but a precise comparison is beyond the scope of the present paper. We refer for instance to [18] where a similar problem was addressed in a goodness-of-fit testing framework, or to [21] for a related discussion.

In this paper, our aim is to establish minimax rates of convergence for both $d_{f,g}$ and d_Δ. In order to get these rates, we will need some assumptions on the model.

2.2 Assumptions

Following for instance [22], we will use two different conditions in order to obtain minimax rates of convergence. The first one is related to the behavior of the function $f-g$ at the boundary of $G_K^\star$. Recall that Q denotes the Lebesgue measure on $\mathbb{R}^d$.

Margin Assumption: There exist constant $t_0, c_2 \in \mathbb{R}_+$ and $\alpha \in \bar{\mathbb{R}}_+$ such that $\forall 0 < t < t_0$,

$$Q\{x \in K : |f(x)-g(x)| \leq t\} \begin{cases} \leq c_2 t^\alpha \text{ if } \alpha \in \mathbb{R}_+, \\ = 0 \text{ if } \alpha = +\infty. \end{cases} \tag{2.1}$$

This condition expresses the difficulty of distinguishing a distribution from another at the boundary of $G_K^\star$. It has been explicitly introduced for the first time in [22]. The case $\alpha = +\infty$ corresponds to the best situation when $f - g$ does not hit or cross the frontier of the Bayes set. It is the so-called strong margin assumption. In this case, it is well-known from [22] that d_Δ and $d_{f,g}$ are equivalent. If $\alpha \in \mathbb{R}$, the most favorable cases corresponds to large values for α: the distributions F_1 and F_2 are rather different of each side of $G_K^\star$. Small values for α correspond to more difficult situations where fast rates can not be expected.

Regularity Assumption. The second condition is related to the complexity of the problem. Since we are dealing with a nonparametric set estimation problem, it seems natural to use an assumption on the regularity of the boundary of $G_K^\star$. More precisely, we will deal with the family of boundary fragments on K. All along the paper, we state $K = [0,1]^d$ without loss of generality. A set $G \subset [0,1]^d$ belongs to a class of boundary fragments (see [17]) if there exists $b : [0,1]^{d-1} \to [0,1]$ such that:

$$G = \{x = (x_1, \dots x_d) \in [0,1]^d \ : x_d \leq b(x_1, \dots, x_{d-1})\} := G_b.$$

For given $\gamma, L > 0$ the class of Hölder boundary fragments is then defined as:

$$\mathcal{G}(\gamma, L) = \{G_b, b \in \Sigma(\gamma, L)\}, \tag{2.2}$$

where $\Sigma(\gamma, L)$ is the class of isotropic Hölder continuous functions $b(x_1, \dots, x_{d-1})$ having continuous partial derivatives up to order $\lfloor \gamma \rfloor$, the maximal integer strictly less than γ and such that:

$$|b(y) - p_{b,x}(y)| \leq L|x-y|^\gamma, \forall x, y \in \mathbb{R}^{d-1},$$

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

In the sequel, we restrict the class $\mathcal{G}$ of possible candidate sets $G \subset K$ for which both the margin and Hölder boundary fragment assumptions are satisfied. It requires, in turn, restrictions on the class $\mathcal{F}$ of possible density couple (f, g). Our result are given in a minimax framework over the following class $\mathcal{F}$. For positive constants $\gamma, L, c_2, t_0, \alpha$ and c_1, the class $\mathcal{F}$ is defined as:

$$\mathcal{F}(\alpha, \gamma) = \{(f,g) : f \text{ and } g \text{ are densities w.r.t. the Lebesgue measure}, \|f\|_\infty \vee \|g\|_\infty \leq c_1,$$
$$\{x \in K : f(x) \geq g(x)\} \in \mathcal{G}(\gamma, L) \text{ and (2.1) holds for } \alpha \in \bar{\mathbb{R}}\}. \tag{2.3}$$

In the free-noise case, [22] has proved that the rates in d_Δ and $d_{f,g}$ can be completely characterized by both margin and boundary fragment assumptions. Remark that alternative hypotheses can be set on the model. For instance, [2] or [21] deal with a plug-in type assumption on the regularity of $f - g$.

The last hypothesis that we will introduce on the model concerns the measurement errors. Indeed, in the model (1.1), the density of the $Z_i^{(1)}$ (rep. $Z_i^{(2)}$) is nor f (resp. g) but rather $f * \eta$ (resp. $g * \eta$), where $*$ denotes the convolution product between two functions and η the density of the $\epsilon_i^{(j)}$ w.r.t. the Lebesgue measure. Contrary to the free-noise case, the $X_i^{(j)}$ are indirectly observed: we are faced to an inverse (deconvolution) problem.

Inverse problems have been widely investigated in the statistical literature. We mention for instance [24] or [6] for a general review of existing models and related results. In an estimation or testing framework, inverse problems are known for providing slower rates than in the direct cases. This can be explained by the loss of information related to the regularization of the operator. The behavior of the noise density η is hence of first importance if one want to evaluate this decay. In particular, we will see that we can take advantage of the shape of the Fourier transform of the noise in order to provide a precise description of the minimax rates in this setting. This is the purpose of the following assumption.

Noise Assumption: *There exists $\beta = (\beta_1, \dots, \beta_d)' \in \mathbb{R}_+^d$ such that for all $i \in \{1, \dots, d\}$, $\beta_i > 1/2$,*

$$|\mathcal{F}[\eta_i](t)| \sim |t|^{-\beta_i}, \text{ and } \left|\mathcal{F}'[\eta_i](t)\right| \leq C|t|^{-\beta_i} \text{ as } t \to +\infty,$$

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

Note that the hypothesis $\eta = \Pi_{i=1}^d \eta_i$ corresponds to non-degenerated random errors ϵ, whose coordinates are independent. This assumption could be relaxed as in [8] since we only need an assumption over the asymptotic behavior of $\mathcal{F}[\eta]$. In the literature, the noise assumption **(NA)** corresponds to *ordinary smooth* or *mildly ill-posed* inverse problem. The parameters β_i describe the difficulty of the related problem. Higher is β_i, smoother are $f * \eta$ and $g * \eta$ in the direction i. As a result, harder becomes the classification problem. In the sequel, we show that these coefficients play a crucial role in the expression of the minimax rates of convergence. For the sake of concision, we will not consider *severely ill-posed* problems, i.e. corresponding to exponentially decreasing Fourier transform. However, simple applications of the main result of this paper lead to the study of this particular case.

2.3 The classifier

We are now ready to propose a classifier in such a context. Our method is based on the empirical risk minimization (ERM) method. The main idea is to construct an estimator for the Bayes risk associated to each candidate G, and then to select the one associated to the lowest value. In the free-noise case, i.e. when data are observed without measurement errors, [22] have used the risk estimator $R_{n,m}(.)$ defined as

$$R_{n,m}(G) = \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i^{(1)} \in K \setminus G\}} + \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\{X_i^{(2)} \in G\}}\right], \; \forall G \subset K.$$

In particular, it is easy to see that for all $G \subset K$, $R_{n,m}(G)$ is an unbiased and consistent estimator of $R_K(G)$. When dealing with an error-in-variables model, the methodology is completely different. Indeed, for all $i \in \{1, \dots, n\}$, we get for instance

$$\mathbb{E}\left[\mathbf{1}_{\{Z_i^{(1)} \in K \setminus G\}}\right] = \int_{K \setminus G} f * \eta(x) dx.$$

In such a situation, $R_{n,m}(G)$ is nor an unbiased neither a consistent estimator of the risk $R_K(G)$. We are faced to an inverse (deconvolution) problem. In order to get round of this problem, we will propose a deconvolution ERM algorithm. This algorithm is heavily related to the properties of deconvolution kernel (see for instance [12] or [24]).

Let $\mathcal{K} = \prod_{j=1}^d \mathcal{K}_j : \mathbb{R}^d \to \mathbb{R}$ be a d-dimensional kernel defined as the product of d unidimensional kernels $\mathcal{K}_j$ (i.e. functions $\mathcal{K}_j : \mathbb{R} \to \mathbb{R}$ satisfying $\int \mathcal{K}_j = 1$). The properties of $\mathcal{K}$ leading to satisfying upper bounds will be made precise later on. 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 the deconvolution kernel $\mathcal{K}_\eta$ as

$$\begin{aligned} \mathcal{K}_\eta \; : \; & \mathbb{R}^d \to \mathbb{R} \\ & t \mapsto \mathcal{K}_\eta(t) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\mathcal{K}](\cdot)}{\mathcal{F}[\eta](\cdot/\lambda)}\right](t), \end{aligned} \tag{2.4}$$

provided that $\mathcal{K}$ (resp. η) belongs to $L_2(\mathbb{R}^d)$ and admits a Fourier transform. Note that in the sequel, for the sake of concision, we note, for any $\lambda \in \mathbb{R}_+^d$, $z, x \in \mathbb{R}^d$:

$$\frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right) := \frac{1}{\lambda_1\cdots\lambda_d}\mathcal{K}_\eta\left(\frac{z_1-x_1}{\lambda_1},\ldots,\frac{z_d-x_d}{\lambda_d}\right).$$

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

$$R_{n,m}^\lambda(G) = \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^n h_{K\backslash G,\lambda}(Z_i^{(1)}) + \frac{1}{m}\sum_{i=1}^m h_{G,\lambda}(Z_i^{(2)})\right],$$

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

$$h_{G,\lambda}(z) = \int_G \frac{1}{\lambda}\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)dx. \tag{2.5}$$

For all $G \subset K$, the function $h_{G,\lambda}$ more or less plays the role of an indicator function. In particular, for all $i \in \{1,\ldots,n\}$, we get for instance

$$\mathbb{E}[h_{G,\lambda}(Z_i^{(1)})/X_i^{(1)}] = \mathcal{K}_\lambda * \mathbf{1}_{\{.\in G\}}(X_i^{(1)}),\ \forall G \subset K, \tag{2.6}$$

where $\mathcal{K}_\lambda * \mathbf{1}_{\{.\in G\}}(x)$ denotes the convolution between $\mathcal{K}$ and the indicator function $\mathbf{1}_{\{.\in G\}}$ at a point $x \in \mathbb{R}$. This term can then be viewed as a smoothed indicator on the set G. Remark that due to (2.6), the estimator (2.5) will be biased. Indeed, for all $G \subset K$

$$\mathbb{E}R_{n,m}(G) = \int_{\mathbb{R}^d} f(x)\mathcal{K}_\lambda * \mathbf{1}_{\{.\in K/G\}}(x) + \int_{\mathbb{R}^d} g(x)\mathcal{K}_\lambda * \mathbf{1}_{\{.\in G\}}(x) := R^\lambda(G) \neq R(G).$$

The control of the related bias will be one of the main difficulty to establish minimax rates of convergence for this estimator.

In the following, we study ERM estimators defined as:

$$\hat{G}_{n,m}^\lambda = \arg\min_{G\in\mathcal{G}(\gamma,L)} R_{n,m}^\lambda(G), \tag{2.7}$$

where $\mathcal{G}(\alpha,\gamma)$ is defined in (2.2) and $\lambda = (\lambda_1,\ldots,\lambda_d) \in \mathbb{R}_+^d$ is a parameter that has to be chosen explicitly.

3 Lower bound

Theorem 1 states lower bounds for the minimax risks over the class $\mathcal{F}(\alpha,\gamma)$ defined in (2.3). The proof is postponed to Section 7.

Theorem 1 *Let $K = [0,1]^d$ and $\mathcal{F}(\alpha,\gamma)$ defined in* (2.3)*. Suppose that the noise assumption is satisfied for some β. Then we have:*

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

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

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

Remark 4. We obtain exactly the same lower bounds as [22] in the direct case, which yet corresponds to the situation where $\beta_j = 0$ for all $j \in \{1, \dots, d\}$. In this particular framework, the minimax rate of convergence mainly depends on γ and α. The coefficient γ corresponds to the regularity of the boundary of $G_K^\star$. Greater is γ, easier is the estimation. The term α is related to the margin assumption.

Remark 5. In the presence of noise in the variables, the rates obtained in Theorem 1 are slower. The price to pay is an additional term of the form

$$2\alpha \left[\sum_{i=1}^{d-1} \beta_i + \beta_d \gamma \right].$$

This term clearly connects the difficulty of the problem to the values of the coefficients $\beta_1, \dots, \beta_d$. Moreover, the above expression highlights a connection between the margin parameter and the ill-posedness. The role of the margin parameter over the inverse problem can be summarized as follows. Higher is the margin, higher is the price to pay for a given degree of ill-posedness. When the margin parameter is small, the problem is difficult at the boundary of $G_K^\star$ and we can only expect a non-sharp estimation of $G_K^\star$. In this case it is not significantly worst to add noise. On the contrary, for large margin parameter, there is nice hope to give a sharp estimation of $G_K^\star$ and then perturb the inputs variables have strong consequences in the performances.

Remark 6. In the above expression, the first $d-1$ components of ϵ do not have the same impact as the last (vertical) component. This is due to the fact that we consider boundary fragments with a given regularity γ. This regularity is expressed in a Hölder space of functions defined on the $d-1$ first directions.

Remark 7. Finally, we can compare the lower bound of Theorem 1 with the previous lower bound stated in [21] under plug-in type conditions. The main novelty here is that no restriction on $\alpha \in \bar{\mathbb{R}}$ is necessary to get the lower bound. It could be explain as follows. Coarsely, the case $\alpha = +\infty$ cannot be treated in [21] since the minimax approach is performed over a class of densities with Hölder regularity. If the strong margin assumption holds, $f - g$ is not continuous at the boundary. Moreover, Theorem 1 holds for arbitrary values for α. Since we do not suppose any assumption for the regularity of the densities f and g, the construction of the lower bound is easier. In particular, we can take advantage of the noise assumption and mix standard arguments from lower bounds in classification (see [1] and [22]) and inverse problems (see [5]).

4 Upper bounds

4.1 A preliminary result

For the sake of concision, in this section we propose to restrict the set $\mathcal{G}$ to $\mathcal{G}(\gamma, L)$, where all possible regularities γ satisfy $\gamma > d-1$. It allows us to control the bracketing entropy of $\mathcal{G}(\gamma, L)$ with a parameter $\rho = \frac{d-1}{\gamma} < 1$. We may also consider more general classes of candidates $\mathcal{G}$, with given entropy rates. This extension is presented in Section 6 using empirical processes theory in a more general framework.

In this section, we are interested in the performances of the estimator:

$$\hat{G}_{n,m} = \arg \min_{G \in \mathcal{G}(\gamma, L)} R_{n,m}^\lambda(G), \tag{4.1}$$

where $\mathcal{G}(\gamma, L)$ is defined in (2.2) for $\gamma > d-1$. Nevertheless, one may also define our ERM estimator for $\gamma \leq d-1$ by considering a network in a practical purpose, without significant change in the following results. We will also assume for clarity throughout this section that $n = m$. In order to get round of the assumption on both the shape of the noise and the boundary of $G_K^\star$, we will introduce some constraints on the kernel $\mathcal{K}$.

Kernel Assumption

(K1) The kernel $\mathcal{K}$ is such that the associated deconvolution kernel satisfies

$$\sup_{t\in\mathbb{R}^d} |\mathcal{F}[\mathcal{K}_\eta](t)| \leq C\prod_{i=1}^d \lambda_i^{-\beta_i}, \text{and } \|\mathcal{K}_\eta\|^2 \leq C\prod_{i=1}^d \lambda_i^{-2\beta_i}.$$

The assumption **(K1)** is necessary to control the variance of our classifier. It is satisfied for instance if the Fourier transform of $\mathcal{K}$ is bounded and compactly supported.

The following theorem provides a control of the expectation of the excess risk by a bias-variance decomposition, up to a residual term depending on the choice of the bandwidth λ.

Theorem 2 *Let $\hat{G}_n$ the set introduced in* (4.1) *where $\gamma > d-1$ and $n = m$. Suppose that the noise assumption is satisfied and consider a kernel $\mathcal{K}_\eta$ defined as in* (2.4) *satisfying* ***(K1)****. Then we have*

$$\mathbb{E}d_{f,g}(\hat{G}_{n,m}, G^*) \leq C \inf_{\lambda\in\mathbb{R}_+^d} \left[\left(\frac{\Pi_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n}} \right)^{\frac{2\gamma(\alpha+1)}{\gamma(\alpha+2)+(d-1)\alpha}} + \sup_{G\in\mathcal{G}} \left(R_K - R_K^\lambda \right)(G) + \sum_{i=1}^d (n\lambda_i)^{-\frac{\gamma}{\gamma+d-1}} \right],$$

where C is a positive constant and $(R_K - R_K^\lambda)(G) = R(G) - R^\lambda(G)$ for all $G \in \mathcal{G}$.

This result highlights a bias-variance decomposition of the excess risk. The proof is presented in Section 7. The main ingredient of the proof is a study of the increments of a noisy empirical process, indexed by a set of functions which depends on the regularization parameter $\lambda > 0$. At this step, some remarks are necessary:

Remark 9. The variance term is obtained thanks to extensions of the empirical process machinery and the peeling technique introduced by [13] in the direct case (see Section 8 for details). This term is related to the regularity of the distribution function η in the noise assumption. We can see coarsely that the price to pay for the inverse problem in the variance is summarized in the term $\Pi_{i=1}^d \lambda_i^{-\beta_i}$. Note that in the direct case, [22] has already stated fast rates of the form $n^{-\frac{\gamma(\alpha+1)}{\gamma(\alpha+2)+(d-1)\alpha}}$, which corresponds to $\beta = 0$ in Theorem 2.

Remark 10. The second term in Theorem 2 is a bias term due to the estimation of the true risk by a biased empirical risk. When dealing with a deconvolution ERM, the algebra is rather different. We have to provide a precise control of the bias of the ERM, namely the quantity

$$R(G) - \mathbb{E}R_n^\lambda(G) = \int (f-g)\left(\mathbf{1}_{G^C} - \mathcal{K}_\lambda * \mathbf{1}_{G^C}\right) dQ.$$

This term has to be controlled carefully to get minimax results. This is the focus of the next paragraph.

Remark 11. Finally, the last term in the upper bound is a residual term since we can see coarsely that

$$(n\lambda_i)^{-\frac{\gamma}{\gamma+d-1}} \leq \left(\frac{\Pi_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n}} \right)^{\frac{2\gamma(\alpha+1)}{\gamma(\alpha+2)+(d-1)\alpha}},$$

provided that $\lambda_i \to 0$ not too fast, for all $i \in \{1, \ldots, d\}$.

4.2 Control of the bias and related rates of convergence

The aim of this part is to investigate different available ways in order to control the bias. It is important to note that in a previous paper dedicated to plug-in type conditions, [21] provides a simple way to control the bias term. Indeed, under a Hölder regularity condition over the function $f-g$, we can bound the bias term as follows:

Lemma 1 (Loustau and Marteau [21]) *Suppose $f-g \in \Sigma(\gamma, L)$, the isotropic Hölder class of functions over $\mathbb{R}^d$. Suppose that the kernel $\mathcal{K}$ is of order γ. Then, we have*

$$\sup_{G\subset K}(R_K^\lambda - R_K)(G) \leq C\sum_{i=1}^d \lambda_i^\gamma.$$

The proof is straightforward since in this case, we can write:

$$R(G) - \mathbb{E}R_n^\lambda(G) = \int \mathbf{1}_{G^C}[(f-g) - \mathcal{K}_\lambda * (f-g)]dQ.$$

Then, the control of the bias term is reduced to the control of the bias term in standard non-parametric density estimation, which gives (see for instance [25]):

$$\sup_{x_0\in\mathbb{R}^d} |(f-g)(x_0) - \mathcal{K}_\lambda * (f-g)(x_0)| \leq \sum_{i=1}^d \lambda_i^\gamma.$$

Here, the problem is rather different since the regularity assumption deals with the boundary of G_K^*. It is well-known (see for instance [17]) that a regularity with respect to the boundary of a decision rule does not match with plug-in type conditions of Lemma 1. The following result proposes an upper bound under boundary assumptions.

Corollary 1 *Let $\hat{G}_n$ the set introduced in (4.1) where $\gamma > d-1$. Suppose the noise assumption is satisfied and consider a kernel $\mathcal{K}_\eta$ defined as in (2.4) satisfying **(K1)**. Suppose moreover that for any $j \in \{1,\dots,d\}$, $\int_{\mathbb{R}^d} |\mathcal{K}(z)||z_j|\,dz < \infty$ and $\Pi_{j=1}^{d-1}\mathcal{K}_j$ has compact support. Then, there exists a positive constant C such that*

$$\mathbb{E}d_\square(\hat{G}_{n,m}, G_K^*) \leq Cn^{-\kappa_d(\alpha\beta,\gamma)},$$

where

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

Following Corollary 1, lower and upper bounds do not match. The prize to pay for the errors-in-variables model is summarized in the term $2\gamma(\alpha+1)\sum_{i=1}^d \beta_i$ whereas the lower bound proposes a smaller term $2\alpha\sum_{i=1}^{d-1}\beta_i + 2\gamma\alpha\beta_d$. By the way, the corresponding error becomes negligible when γ is close to 1 and $\alpha \to \infty$. The proof is provided in Section 7 and uses Theorem 2 gathering with the following crude bound for the bias term:

$$\sup_{G\in\mathcal{G}}(R_K^\lambda - R_K)(G) \leq C\sum_{i=1}^d \lambda_i.$$

It is based on the following scheme. For all $G \subset K$, using Fubini, we have

$$
\begin{aligned}
&\int_{\mathbb{R}^d} (f-g)(x) \left(\mathcal{K}_\lambda * \mathbf{1}_G(x) - \mathbf{1}_G(x)\right) dx \\
&\quad = \int_{\mathbb{R}^d} (f-g)(x) \left(\int_{z\in\mathbb{R}^2} \mathcal{K}(z) \left[\mathbf{1}_G(x+\lambda z) - \mathbf{1}_G(x)\right] dz \right) dx \\
&\quad = \int_{\mathbb{R}^d} \mathcal{K}(z) \left(\int_{\mathbb{R}^d} (f-g)(x) \left[\mathbf{1}_G(x+\lambda z) - \mathbf{1}_G(x)\right] dx \right) dz.
\end{aligned}
$$

Since we do not have any conditions on the smoothness of $f-g$, the control of the bias reduces to the calculation of the Lebesgue measure between the sets G and $G+\lambda z$, which appears to be of order $\sum_i \lambda_i$. Hence, we can not take advantage on the smoothness of the boundary. In Section 5 below, we discuss several tracks to attack this problem. It appeals to different tools (such as convexification, or additional regularity assumptions) which do not fit with the machinery of the present paper.

5 Conclusion

Let us discuss the obtained results and highlight some open problems:

Comparison with [22] This paper can be seen as a generalization of the results of [22] to the error-in-variables case. We highlight, in the presence of noise, fast rates of convergence which depends on the Fourier transform of the noise distribution η. The price to pay depends on the triplet (γ, α, β) related with the regularity, margin and noise assumptions.

Choice of λ and model selection The main drawback of the deconvolution ERM of this paper is the calibration of the bandwidths λ. Under isotropic assumptions over the shape of f and g, we have shown that it is sufficient to choose only one bandwidth. An extension to the anisotropic case could be done easily (see for instance [20] in an unsupervised context). However, these calibrations are non-adaptive and depend on the smoothness assumptions. The data-driven choice of the bandwidth is a natural open problem. The bias variance trade-off to choose λ is not the usual one in non-parametric statistics and a careful study of this problem is necessary. In this direction, we can mention the recent work of [7]. Moreover, this problem of adaptation is compounded with the model selection of $\mathcal{G}$.

Adaptation to the operator The deconvolution classifier proposed in this paper depends on the known density of the noise ϵ. As a result, another issue would be to try to adapt to unknown error densities η. In this direction, it can be interesting to apply the same strategy, using for instance the estimator proposed in [10] for deconvolution with repeated measurements. In the presence of repeated measurements, the model (1.1) becomes, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$:

$$
Z_{i,k}^{(1)} = X_i^{(1)} + \epsilon_{ik},\ k = 1, \ldots, N_i, \text{ and } Z_{j,\ell}^{(2)} = X_j^{(2)} + \epsilon_{j\ell},\ \ell = 1, \ldots, M_j.
$$

In this case, the empirical risk associated to this problem can be written:

$$
R_{n,m}^\lambda(G) = \frac{1}{2}\left[\frac{1}{nN} \sum_{i=1}^n h_{K\backslash G,\lambda}(Z_{i,1}^{(1)}, \ldots, Z_{i,N_i}^{(1)}) + \frac{1}{mM} \sum_{j=1}^m h_{G,\lambda}(Z_{j,1}^{(2)}, \ldots, Z_{j,M_j}^{(2)}) \right],
$$

where $N = \sum_{i=1}^n N_i$, $M = \sum_{j=1}^m M_j$, and for some $j \in \{1, \ldots, m\}$:

$$
h_{G,\lambda}(Z_{j,2}^{(1)}, \ldots, Z_{j,M_j}^{(2)}) = \int_G w_j \sum_{\ell=1}^{M_j} \frac{1}{\lambda} \hat{L}\left(\frac{x - Z_{j,\ell}^{(2)}}{\lambda} \right) dQ(x), \tag{5.1}
$$

where w_j are weights satisfying $\sum_{j=1}^m w_j M_j = M$ and $\hat{L}$ is a ridged deconvolution estimator defined in [10]. An interesting open problem is to use the same methodology presented in this paper to study the performances of the minimizer of the deconvolution ERM using (5.1).

Direct VS inverse problem In this paper, we propose to study the rates of convergence to the Bayes of our deconvolution classifier in term of pseudo-distance $d_{f,g}$ and d_Δ. However, as discussed in Remark 2, a direct approach seems to be tractable in some particular cases. A systematic study of the rates of convergence of standard ERM using noisy measurements could be interesting. For this purpose, we have to control the difference between the Bayes risk with respect to X and the following Bayes risk:

$$R_K^\eta(G) = \frac{1}{2}\left[\int_{K\backslash G} f * \eta dQ + \int_G g * \eta dQ\right].$$

Another interesting comparison could be done in the problem of predicting a new incoming noisy observation. This paper shows that deconvolution ERM are optimal to predict a new X observation. Finding the more efficient rule to predict a new Z observation is also of practical interest. To this end, a relationship between the margin assumption and the noise assumption has to be done, which appears to be a challenging open problem.

Minimax optimality remains an open problem Lower and upper bounds of Theorem 1 and Corollary 1 do not match and the question of minimax rates in such a setting remains an open problem. However, our intuition is the following: the lower bound is valid and on the contrary, the estimation method of this paper suffers from a lack of optimality.

Firstly, an investigation of the upper bound above indicates that an *optimal* control of the bias would require an upper bound of order

$$\left[\left(\sum_{i=1}^{d-1}\lambda_i\right)^\gamma + \lambda_d\right]^{1+\frac{1}{\alpha}} \quad \text{instead of} \quad \sum_{i=1}^d \lambda_i.$$

Using simple algebra (in dimension 2 for the sake of convenience), we can re-write the bias as

$$\begin{aligned}&\int_{\mathbb{R}^d}(f-g)(x)\left(\mathcal{K}_\lambda * \mathbf{1}_G(x) - \mathbf{1}_G(x)\right)dx\\ &=\int_{\mathbb{R}^2}\mathcal{K}(z)\int_{\mathbb{R}}\left[\int_0^{(b(x_1+\lambda_1 z_1)-\lambda_2 z_2)_+}(f-g)(x)\mathbf{1}_{\{x_1+\lambda_1 z_1\in[0,1]\}}dx_2 - \int_0^{b(x_1)}(f-g)(x)dx_2\right]dx_1 dz.\end{aligned}$$

The exponent γ is related to the smoothness of the boundary of b. This smoothness properties could certainly be taken into account following some additional assumptions on the smoothness of $f-g$. Indeed, one may manage a Taylor expansion of b and then $f-g$ in the 2^{nd} direction, up to some additional technical constraints. Concerning the exponent $1+1/\alpha$, one might take advantage of the behavior of $f-g$ in the neighborhood of $G_K^\star$, but this may require an extended version of the margin assumption.

Another possibility is to use convex surrogate to deal with the bias term. Indeed, the difficulty to control the bias term seems to be related with the $0-1$ loss approach of this paper. Since we use the $0-1$ loss, the bias is upper bound by the Lebesgue measure between the sets G and $G+\lambda z$, which is of order $\sum\lambda_i$ (see Corollary 1 and the associated discussion). There is nice hope that a control of the bias term can be managed thanks to a smooth loss function, without any additional smoothness assumption. However, in this case, a precise study of the lower bound has to be performed and is out of the scope of the present paper.

Finally, the lack of optimality of the deconvolution ERM of this paper could be explain as follows. In the estimation procedure, the idea is to estimate the true risk by using a deconvolution kernel estimator of the densities f and g. As a result, the method seems to be strongly linked with plug-in type regularities for the Bayes decision rule G_K^*, which allows us to control easily the bias term in [21]. Here, the regularity assumption is rather different and deals with the boundaries of G_K^*. That's why an optimal control of the bias term seems problematic. Another way to obtain the fast rates of the lower bound should be to use another estimator of the true risk to deal with errors-in-variables. In classification with smooth boundaries, the estimation task could be summarize as an estimation of the boundary. As a result, in the presence of noisy observations, we have to study the effect of the inverse problem on the boundary of the Bayes G_K^* (that is with respect to the direct data). It could be a way to plug another estimator in the true risk, which allows to estimate optimally the boundary.

6 Preliminaries to the proofs

In this section, we provide a more general point of view about the problem of classification with noisy inputs. To be more precise, we state a generalization of Theorem 2 to study any possible candidate set $\mathcal{G}$ in the ERM minimization, given its entropy rates.

More formaly, we suggest to state upper bounds for deconvolution ERM of the form:

$$\hat{G}_n = \arg\min_{G \in \mathcal{G}} R_n^\lambda(G), \tag{6.1}$$

where $R_n^\lambda(\cdot)$ is the empirical risk defined in Section 1 and $\mathcal{G}$ is a set of possible candidates for $G_K^\star$. We want to study the rate of convergence of $\hat{G}_n$ to $G_K^\star$. This rate depends on the complexity of the class $\mathcal{G}$ in terms of $\delta-$entropy with bracketing. For $\delta > 0$, the bracketing entropy of $\mathcal{G}$ with respect to some distance d is denoted by $\mathcal{H}(\mathcal{G}, d, \delta)$ and corresponds to the minimal number such that $\mathcal{N}_B(\delta) = \exp(\mathcal{H}(\mathcal{G}, d, \delta))$ is an integer and such that there exists pairs (G_j, H_j), $j = 1, \ldots, N_B(\delta)$ of subsets of $\mathcal{G}$ satisfying

(1) $G_j \subset H_j$ for all $j \in \{1, \ldots, N_B(\delta)\}$,

(2) $d(G_j, H_j) \leq \delta$ for all $j \in \{1, \ldots, N_B(\delta)\}$,

(3) For any $G \in \mathcal{G}$, there exists $j \in \{1, \ldots, N_B(\delta)\}$ such that $G_j \subset G \subset H_j$.

We begin with a general upper bound when $\mathcal{G}$ has a given entropy rate. It allows to deduce easily Theorem 2 in Section 4.

Proposition 1 *Suppose $\mathcal{G}$ contains $G_K^\star$ and satisfies, for some $0 < \rho < 1$:*

$$\mathcal{H}(\mathcal{G}, d_\Delta, \delta) \leq c\delta^{-\rho}. \tag{6.2}$$

Let $\hat{G}_n$ the set introduced in (6.1) where $\mathcal{G}$ satisfies (6.2). Suppose the noise assumption is satisfied and consider a kernel $\mathcal{K}_\eta$ defined as in (2.4) satisfying the Kernel assumption. Then, we have:

$$\mathbb{E}d_{f,g}(\hat{G}_{n,m}, G_K^\star) \leq C \inf_{\lambda \in \mathbb{R}_+^d} \left[\left(\frac{\Pi_{i=1}^d \lambda_i^{-\beta_i}}{\sqrt{n}} \right)^{\frac{2(\alpha+1)}{\alpha+2+\rho\alpha}} + \sup_{G \in \mathcal{G}} \left(R_K - R_K^\lambda \right)(G) + \sum_{i=1}^d (n\lambda)^{-\frac{1}{1+\rho}} \right],$$

where $C > 0$ is a generic constant.

Such a result is an extension of Theorem 2 for general set $\mathcal{G}$ with given entropy conditions. The main ingredient of the proof is a generalization of Lemma 5.11 in [13]. The proof is postponed to Section 8.

Such a generality allows to deal with various constraints on the problem. In this paper, we deal with assumptions on the smoothness of the Bayes classifier boundary. Alternative constraints could be investigated. By the way, it may be possible to consider plug-in type assumption as in [2] or [21], convex sets or finite Vapnik Chervonenkis classes as in [23] or [16].

7 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 $(a_n)_{n\in\mathbb{N}}$, the notation $a_n \approx b_n$ (resp. $a_n \lesssim b_n$) 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 $n \in \mathbb{N}$.

7.1 Proof of Theorem 1

The proof starts as in [22] but then uses some arguments which are specific to the inverse problem literature (see for instance [5] or [24]).

Let $\mathcal{F}_1$ a finite class of densities and g_0 a fixed density such that $(f, g_0) \in \mathcal{F}_{\text{frag}}$ for all $f \in \mathcal{F}_1$. The contain of $\mathcal{F}_1$ and the value of g_0 will be precised later on. Then, for all estimator $\hat{G}_{n,m}$ of the set $G_K^\star$, we have

$$\begin{aligned}\sup_{(f,g)\in\mathcal{F}_{frag}} \mathbb{E}_{f,g}\, d_\Delta(\hat{G}_{n,m}, G_K^\star) &\geq \sup_{(f,g_0), f\in\mathcal{F}_1} \mathbb{E}_{f,g} d_\Delta(\hat{G}_{n,m}, G_K^\star),\\ &\geq \mathbb{E}_{g_0}\left[\frac{1}{\sharp\mathcal{F}_1}\sum_{f\in\mathcal{F}_1}\mathbb{E}_f\left\{d_\Delta(\hat{G}_{n,m}, G_K^\star)|X_1^{(2)},\dots,X_m^{(2)}\right\}\right]. \end{aligned} \tag{7.1}$$

7.1.1 Construction of $\mathcal{F}_1$

Concerning the density g_0, we deal with the uniform density on $[0,1]^2$, i.e.

$$g_0(x) = \mathbf{1}_{\{x\in[0,1]^2\}}, \forall x \in \mathbb{R}^2.$$

Now, we have to define the class $\mathcal{F}_1$. First, we consider a function φ infinitely differentiable defined on $\mathbb{R}$ such that $\text{supp}(\varphi) = [-1,1]$, $\varphi(t) \geq 0$ for all $t \in \mathbb{R}$ and $\|\varphi\|_\infty = \varphi(0) = 1$. Let $M \geq 2$ an integer which will be allowed to depend on n and $\tau > 0$ a positive constant. Then, for all $j \in \{1,\dots,M\}$, we set

$$\varphi_j(t) = \tau M^{-\gamma}\varphi\left(M\left[t - \frac{2j-1}{M}\right]\right),\ \forall t \in \mathbb{R}.$$

For all $\omega \in \{0,1\}^M$ and all $t \in \mathbb{R}$, we define

$$b(t,\omega) = \frac{1}{2} + \sum_{j=1}^M \omega_j\varphi_j(t).$$

In the specific case where $\omega_j = 1$ for all $j \in \{1,\dots,M\}$, we write $b(t, \mathbf{1})$. Then, let b_0 and $C^\star$ positive constants which will be precised later on. We define the function $f_0 : \mathbb{R}^2 \to \mathbb{R}$ as

$f_0(x) = 0$ for all $x \notin [0,1]^2$ and

$$f_0(x) = \begin{cases} 1 + 2\eta_0, \forall x_2 \in [0, 1/2], \\ \\ 1 - \eta_0 - b_0, \forall x_2 \in [b(x_1, \mathbf{1}), 1], \\ \\ 1 + \left(\frac{b(x,\mathbf{1}) - x_2}{c_2}\right)^{1/\alpha} - C^\star M^{-\gamma/\alpha}, \forall x_2 \in [1/2, b(x_1, \mathbf{1})], \end{cases}$$

where $C^\star = 3/2.(\tau/c_2)^{1/\alpha}$ and $b_0 > 0$ is such that $\int f_0(x)dx = 3/4$. The condition on $C^\star$ ensures that $f_0(x) < 1$ for all $x_2 \in [1/2, b(x_1, \mathbf{1})]$. We will also use the function f_1 defined as

$$f_1(x) = \begin{cases} 0, \forall x \in [0,1]^2, \\ \\ \frac{b_1}{(1+x_2)^2.(1+x_1)^2}, \forall x \notin [0,1]^2, \end{cases}$$

where $\mathcal{C}_1$ is such that $\int f_1(x)dx = 1/4$. Finally, the set $\mathcal{F}_1$ will be defined as

$$\mathcal{F}_1 = \left\{ f_\omega,\ \omega \in [0,1]^M \right\},$$

where for a given $\omega \in \{0,1\}^M$,

$$f_\omega(x) = f_0(x) + f_1(x) + \sum_{j=1}^{M} \omega_j \rho_j(x). \tag{7.2}$$

for some functions $(\rho_j)_{j=1\ldots M}$ which are explicited below. In order to complete the construction of the set $\mathcal{F}_1$, we have to provide a precise definition of the ρ_j and to prove that the f_ω define probability density functions for all $\omega \in \{0,1\}^M$.

We first start with the construction of the ρ_j. For all $x \in \mathbb{R}$, let $\rho : \mathbb{R} \to [0,1]$ the function defined as

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

with associate Fourier transform $\mathcal{F}[\rho](t) = (1 - |t|)_+$. In particular, supp $\mathcal{F}[\rho] = [-1, 1]$. For all $j \in \{1, \ldots, M\}$ and $x_2 \in \mathbb{R}$, introduce

$$\rho_{(2)}(x_2) = \cos\left(\frac{x_2 - 1/2(1 + \tau M^{-\gamma})}{3/2\pi^{-1}\tau M^{-\gamma}}\right) \rho\left(\frac{x_2 - 1/2(1 + \tau M^{-\gamma})}{3\pi^{-1}\tau M^{-\gamma}}\right). \tag{7.3}$$

By the same way, for all $j \in \{1, \ldots, M\}$, we define

$$\rho_{j,(1)}(x_1) = \cos\left[\frac{\pi}{3}\left(\frac{x_1 - j/M}{M^{-1}}\right)\right] \rho\left[\frac{\pi}{6}\left(\frac{x_1 - j/M}{M^{-1}}\right)\right]. \tag{7.4}$$

Then, for all $j \in \{1, \ldots, M\}$ and $x = (x_1, x_2) \in [0,1]^2$, we set

$$\rho_j(x) = c^\star (\tau M^{-\gamma})^{1/\alpha}\ \rho_{(2)}(x_2) \rho_{j,(1)}(x_1), \tag{7.5}$$

for some constant $c^\star$ explicited below.

Now, we prove that the f_ω introduced in (7.2) define density functions. First, remark that

$$\sum_{j=1}^{M} |\rho_j(x)| \leq \begin{cases} CM^{-\gamma/\alpha}(1 + x_1)^{-2}(1 + x_2)^{-2},\ \forall x \notin [0,1]^2, \\ CM^{-\gamma/\alpha},\ \forall x \in [0,1]^2, \end{cases}$$

This ensures that $f_\omega \geq 0$ for all $\omega \in \{0,1\}^M$, at least for M large enough. Then recall that both f_0 and f_1 are designed in order to guarantee that $\int (f_0+f_1)(x)dx = 1$. Hence, we only have to show that $\int \rho_j(x)dx = 0$ for all $j \in \{1,\dots,M\}$. In fact, it is only necessary to prove that $\int \rho_{(2)}(x_2)dx_2 = 0$. First remark that $\int \rho_{(2)}(x_2)dx_2 = \int \tilde{\rho}_{(2)}(x_2)dx_2$ where $\tilde{\rho}_{(2)}(x_2) = \rho_{(2)}(x_2 + 1/2(1+\tau M^{-\gamma}))$ for all $x_2 \in \mathbb{R}$. Then, using simple algebra

$$\mathcal{F}[\rho_{(2)}](0) = \frac{1}{2}\mathcal{F}\left[\rho\left(\frac{\cdot}{3\pi^{-1}\tau M^{-\gamma}}\right)\right]\left(\pm\frac{1}{3/2\pi^{-1}\tau M^{-\gamma}}\right) = \frac{3}{2}\pi^{-1}\tau M^{-\gamma}\mathcal{F}[\rho]\,(\pm 2)$$
$$= 0,$$

since the support of the Fourier transform of ρ is $[-1;1]$. Hence, for all $\omega \in \{0,1\}^M$, f_ω is a density function.

In order to conclude the proof, we have to show that

$$(f_\omega, g_0) \in \mathcal{F}_{\text{frag}}\ \forall \omega \in \{0,1\}^M, \tag{7.6}$$

which allows to use the bound (7.1),

$$Q\left\{x \in K : |f_\omega(x) - g_0(x)| \leq \eta\right\} \leq c_2\eta^\alpha\ \forall \omega \in \{0,1\}^M \text{ and } \forall \eta \leq \eta_0, \tag{7.7}$$

which means that the *Margin assumption* is satisfied for our test functions and that

$$\mathbb{E}_{g_0}\mathbb{E}_{f_\omega}\left\{d_\Delta(\hat{G}_{n,m}, G_K^\star)|X_1^{(2)},\dots,X_m^{(2)}\right\} \geq Cn^{-\frac{\gamma}{\gamma\left(\frac{2}{\alpha}+1\right)+2\beta_1+2\beta_2\gamma+1}}, \tag{7.8}$$

for some positive constant C.

7.1.2 Main assumptions check

We first start with the proof of (7.6). First remark that for all $j \in \{1,\dots,M\}$, the function $\rho_j(.)$ is bounded from above by $CM^{-\gamma/\alpha}$ for some $C > 0$. Then, using simple algebra

$$\begin{aligned}
x_2 \in [1/2; b(x_1,\mathbf{1})] \quad &\Rightarrow \quad \frac{1}{2} \leq x_2 \leq \frac{1}{2} + \tau M^{-\gamma},\\
&\Rightarrow \quad -\frac{\tau M^{-\gamma}}{2} \leq x_2 - \frac{1}{2} - \frac{\tau M^{-\gamma}}{2} \leq \frac{\tau M^{-\gamma}}{2},\\
&\Rightarrow \quad -\frac{\pi}{6} \leq \frac{x_2 - 1/2(1+\tau M^{-\gamma})}{3\pi^{-1}\tau M^{-\gamma}} \leq \frac{\pi}{6},\\
&\Rightarrow \quad \rho_{(2)}(x_2) \geq \frac{9}{4\pi^3}.
\end{aligned}$$

The same kind on minoration holds for the function $\rho_{j,(1)}$. Hence the ρ_j are uniformly bounded from below on $[1/2; b(x_1,\mathbf{1})]$. For all $\omega \in \{0,1\}^M$ and for all $x \in [0,1]^2$, we then have

$$f_\omega(x) \geq 1 + \left(\frac{b(x,\mathbf{1}) - x_2}{c_2}\right)^{1/\alpha} \geq g_0(x),\ \forall x_2 \in [1/2, b(x_1,\omega)],$$

for $c^\star$ large enough. This ensures that

$$\{x \in [0,1]^2 : f_\omega(x) \geq g_0(x)\} = \{x \in [0,1]^2 : 0 \leq x_2 \leq b(x_1,\omega)\}.$$

In order to conclude the proof of (7.6), we only have to remark that the function $b(.,\omega)$ belongs to $\Sigma(\gamma, L)$ for all $\omega \in \{0,1\}^M$, at least for M small enough.

Now, we consider the margin assumption (7.7). First, we consider the case where $\eta < [\tau c_2^{-1}]^{1/\alpha} M^{-\gamma/\alpha} < \eta_0$. Clearly, following our choices of b_0 and $C^\star$, we have that

$$|f_\omega(x) - g_0(x)| \leq \eta \Rightarrow x_2 \in [1/2; b(x_1,\omega)] \Rightarrow x_2 \leq b(x_1,\omega).$$

Moreover, for all $x \in [0,1]^2$ such that $x_2 \leq b(x_1,\omega)$, we have

$$(f_\omega - g_0)(x) = \left(\frac{b(x,\mathbf{1}) - x_2}{c_2} \right)^{1/\alpha} + \sum_{j=1}^M \omega_j \rho_j(x) - C^\star M^{-\gamma/\alpha},$$

where

$$\sum_{j=1}^M \omega_j \rho_j(x) - C^\star M^{-\gamma/\alpha} > 0,\ \forall x_2 \in \left[\frac{1}{2}, b(x_1,\omega) \right].$$

Thus

$$|f_\omega(x) - g_0(x)| \leq \eta \Rightarrow \left(\frac{b(x,\omega) - x_2}{c_2} \right)^{1/\alpha} \leq \eta \Rightarrow x_2 \geq b(x_1,\omega) - c_2 \eta^\alpha,$$

which proves the margin assumption when $\eta < [\tau c_2^{-1}]^{1/\alpha} M^{-\gamma/\alpha}$. Now, in the case where $\eta_0 > \eta > [\tau c_2^{-1}]^{1/\alpha} M^{-\gamma/\alpha}$, we have

$$|f_\omega(x) - g_0(x)| \leq \eta \Rightarrow 1/2 < x_2 < b(x_1,\mathbf{1}),$$

which entails

$$Q\left\{ x \in K : |f_\omega(x) - g_0(x)| \leq \eta \right\} \leq \tau M^{-\gamma} \leq c_2 \eta^\alpha.$$

This concludes this part.

7.1.3 Final minoration

Now, we can deal with the lower bound (7.8). The proof is based on classical tools which can be found for instance in [25], [22], [5] or [24]. First remark that the shape of $G_K^\star$ depends on the value of ω. For the sake of convenience, we omit the dependency with respect to this quantity. For all $\omega \in \{0,1\}^M$, recall that

$$G_K^\star = \{x \in [0,1]^2 : f_\omega(x) \geq g_0(x)\} = \{x \in [0,1]^2 : 0 \leq x_2 \leq b(x_1,\omega)\}.$$

Using Assouad Lemma and classical tools designed for instance in [25], we get

$$\mathbb{E}\left[d_\Delta(\hat{G}_{n,m}, G_K^\star) | Y_1, \ldots, Y_m \right] \geq \frac{M}{2} \|\varphi_1\|_1 \int \min\left[dP_{11}, dP_{10}\right], \tag{7.9}$$

where P_{11} denotes the law of $(Z_i^{(1)})_{i=1\ldots n}$ when the density of the $X_i^{(1)}$ is $f_{\omega_{11}}$. In the following, we will choose M in order to guarantee that the term $\int \min\left[dP_{11}, dP_{10}\right]$ is bounded from below. Consequently, the lower bound will be determined by the corresponding value of $M\|\varphi_1\|_1$. Since the observations are independent

$$\int \min\left[dP_{11}, dP_{10}\right] \geq 1 - \sqrt{(1 + \chi^2(P_1, P_0))^n - 1},$$

where $\chi^2(P_a, P_b)$ denotes the chi-square divergence between two given probability measures P_a and P_b, and P_0, P_1 are the law of the variable $Z_1^{(1)} = X_1^{(1)} + \epsilon_1^{(1)}$ when the density of the $X_i^{(1)}$ is respectively $f_{\omega_{11}}$ or $f_{\omega_{10}}$. In the following, our aim is to find a satisfying upper bound for $\chi^2(P_1, P_0)$.

First, remark that we can find $\tilde{c} > 0$ such that for all $x \notin [0,1]^2$ and all $\omega \in \{0,1\}^M$, $f_\omega(x) \geq \tilde{c} f_1(x)$. Hence, using simple algebra, we get that

$$f_\omega * \eta(x) \geq \frac{C}{(1+x_1^2)(1+x_2^2)}, \ \forall x \in \mathbb{R}^2, \tag{7.10}$$

for some $C > 0$. In the following, given f, η_1 and η_2, we denote by $f * \eta$ the convolution product in dimension two, i.e.

$$f * \eta(x) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x_1 - y_1, x_2 - y_2)\eta_1(y_1)\eta_2(y_2) dy_1 dy_2, \ \forall x \in \mathbb{R}^2.$$

Then, using (7.2) and (7.10),

$$\begin{aligned} \chi^2(P_1, P_0) &= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{\{(f_{\omega_{11}} - f_{\omega_{10}}) * \eta(x)\}^2}{f_{\omega_{11}} * \eta(x)} dx, \\ &\leq C \int_{\mathbb{R}} \int_{\mathbb{R}} (1+x_1^2)(1+x_2^2)\{\rho_1 * \eta(x)\}^2 dx. \end{aligned}$$

Hence

$$\begin{aligned} \chi^2(P_1, P_0) &\leq C \int_{\mathbb{R}} \int_{\mathbb{R}} \{\rho_1 * \eta(x)\}^2 dx + C \int_{\mathbb{R}} \int_{\mathbb{R}} x_2^2 \{\rho_1 * \eta(x)\}^2 dx \\ &\quad + C \int_{\mathbb{R}} \int_{\mathbb{R}} x_1^2 \{\rho_1 * \eta(x)\}^2 dx + C \int_{\mathbb{R}} \int_{\mathbb{R}} x_1^2 x_2^2 \{\rho_1 * \eta(x)\}^2 dx, \\ &:= A_1 + A_2 + A_3 + A_4, \end{aligned}$$

where the ρ_j are defined in (7.5). In the following, we only consider the bound of A_1, the other terms being controlled in the same way. We get

$$\begin{aligned} A_1 &= C \int_{\mathbb{R}} \int_{\mathbb{R}} \{\rho_1 * \eta(x)\}^2 dx, \\ &= C M^{-2\gamma/\alpha} \int_{\mathbb{R}} \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} \int_{\mathbb{R}} \rho_{(2)}(x_2 - y_2)\rho_{j,(1)}(x_1 - y_1)\eta_1(y_1)\eta_2(y_2) dy_1 dy_2 \right\}^2 dx, \\ &= C M^{-2\gamma/\alpha} \int_{\mathbb{R}} \int_{\mathbb{R}} |\mathcal{F}[\rho_{(2)}](t_2)|^2 |\mathcal{F}[\rho_{1,(1)}](t_1)|^2 |\mathcal{F}[\eta_1](t_1)|^2 |\mathcal{F}[\eta_2](t_2)|^2 dt_1 dt_2, \\ &= C M^{-2\gamma/\alpha} A_{1,1} A_{1,2}, \end{aligned}$$

where

$$A_{1,1} = \int_{\mathbb{R}} |\mathcal{F}[\rho_{(1)}](t_1)|^2 |\mathcal{F}[\eta_1](t_1)|^2 dt_1, \ A_{1,2} = \int_{\mathbb{R}} \int_{\mathbb{R}} |\mathcal{F}[\rho_{1,(2)}](t_2)|^2 |\mathcal{F}[\eta_2](t_2)|^2 dt_2,$$

and $\rho_{(1)}$, $\rho_{1,(2)}$ are respectively defined in (7.3),(7.4). We first deal with the term $A_{1,2}$. Using simple algebra, we get

$$\begin{aligned} A_{1,2} &= \int_{\mathbb{R}} |\mathcal{F}[\rho_1^{(1)}](t_1)|^2 |\mathcal{F}[\eta_1](t_1)|^2 dt_1, \\ &= \int_{\mathbb{R}} \left| \mathcal{F}\left[\rho\left(\frac{\cdot}{3\pi^{-1}\tau M^{-\gamma}}\right)\right]\left(t_1 \pm \frac{1}{3/2\pi^{-1}\tau M^{-\gamma}}\right) \right|^2 |\mathcal{F}[\eta_1](t_1)|^2 dt_1, \\ &= (3\pi^{-1})^2 \tau^2 M^{-2\gamma} \int_{\mathbb{R}} \left| \mathcal{F}[\rho]\left(3\pi^{-1}\tau M^{-\gamma} t_1 \pm \frac{3}{3/2}\right) \right|^2 |\mathcal{F}[\eta_1](t_1)|^2 dt_1. \end{aligned}$$

Then, setting $s_1 = 3\pi^{-1}\tau M^{-\gamma}t_1$ and using the *Noise assumption*, we obtain

$$\begin{aligned}
A_{1,2} &= 3\pi^{-1}\tau M^{-\gamma}\int_{\mathbb{R}} |\mathcal{F}[\rho](s_1 \pm 2)|^2 \left|\mathcal{F}[\eta_1]\left(\frac{s_1}{3\pi^{-1}\tau M^{-\gamma}}\right)\right|^2 ds_1, \\
&= 3\pi^{-1}\tau M^{-\gamma}\int_1^3 |\mathcal{F}[\rho](s_1 \pm 2)|^2 \left|\mathcal{F}[\eta_1]\left(\frac{s_1}{3\pi^{-1}\tau M^{-\gamma}}\right)\right|^2 ds_1, \\
&\leq CM^{-\gamma-2\beta_2\gamma}\int_1^3 |\mathcal{F}[\rho](s_1 \pm 2)|^2 |s_1|^{-2\beta_1} ds_1, \\
&\leq CM^{-\gamma-2\beta_2\gamma}.
\end{aligned}$$

Using a similar algebra for the term $A_{1,1}$, we obtain

$$A_{1,2} \leq CM^{-1-2\beta_1}.$$

Similar bounds are available for A_2, A_3 and A_4 since $\mathcal{F}[\rho]$ and its weak derivative are bounded by 1 and supported on $[-1;1]$. In particular, we use the fact that for all $t \in \mathbb{R}$

$$\mathcal{F}[\rho_{1,(2)}](t) = 3\pi^{-1}\tau M^{-\gamma}\mathcal{F}[\rho](3\pi^{-1}\tau M^{-\gamma}t \pm 2),$$

and

$$\frac{d}{dt}\mathcal{F}[\rho_{1,(2)}](t) = -i(3\pi^{-1}\tau M^{-\gamma})^2 t.\mathcal{F}[\rho](3\pi^{-1}\tau M^{-\gamma}t \pm 2),$$

for all t in a subset of $\mathbb{R}$ having a Lebesgue measure equal to 1.

The above equations lead to the following upper bound:

$$\chi^2(P_1, P_0) \leq CM^{-\gamma(2/\alpha+1)-2\beta_1\gamma-2\beta_2-1}.$$

Then, $\chi^2(P_1, P_0) \leq C/n$ for some constant $C > 0$ as soon as

$$M = M_n \sim n^{\frac{1}{\gamma(2/\alpha+1)+2\beta_1+2\beta_2\gamma+1}}.$$

Finally, going back to equation (7.9), we obtain

$$\begin{aligned}
\mathbb{E}\left[d_\Delta(\hat{G}_{n,m}, G_K^\star)|Y_1,\ldots,Y_m\right] &\geq \frac{M_n}{2}\|\varphi_1\|_1 \int \min\left[dP_{11}, dP_{10}\right], \\
&\geq CM_n\|\varphi_1\|_1, \\
&= C\tau M_n^{-\gamma}\int_0^1 \varphi_1(t)dt, \\
&\sim M_n^{-\gamma} = n^{-\frac{\gamma}{\gamma(2/\alpha+1)+2\beta_1+2\beta_2\gamma+1}},
\end{aligned}$$

which concludes the proof.

□

7.2 Proof of Proposition 1

Consider the empirical processes $\nu_n^{(j)}$, for $j \in \{1,2\}$, defined as:

$$\nu_n^{(j)}(G) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[h_{G,\lambda}(Z_i^{(j)}) - \mathbb{E}h_{G,\lambda}(Z^{(j)})\right], \tag{7.11}$$

where the $h_{G,\lambda}(.)$ have been introduced in (2.5). In particular, remark that for all $i \in \{1, \dots, n\}$, $G \subset K$,

$$
\begin{aligned}
\mathbb{E}\left[h_{G,\lambda}(Z_i^{(1)})\right] &= \int_G \frac{1}{\lambda}\mathbb{E}\left[\mathcal{K}_\eta\left(\frac{X_i^{(1)} + \epsilon_i^{(1)} - x}{\lambda}\right)\right] dx, \\
&= \int_G \frac{1}{\lambda}\mathbb{E}\left[\mathcal{K}\left(\frac{X_i^{(1)} - x}{\lambda}\right)\right] dx = \int_{\mathbb{R}^d} f(x)\mathcal{K}_\lambda * \mathbf{1}_{\{.\in G\}}(x)dx.
\end{aligned}
$$

Hence, using (2.7), we can write

$$
\begin{aligned}
&\int (f-g)(\mathcal{K}_\lambda * \mathbf{1}_{\{.\in G_K^\star\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{.\in \hat{G}_n\}}) \\
&\leq \frac{1}{\sqrt{n}}(\nu_n^{(1)}(G^{\star C}) - \nu_n^{(1)}(\hat{G}_n^{\lambda C})) + \frac{1}{\sqrt{n}}(\nu_n^{(2)}(G^\star) - \nu_n^{(2)}(\hat{G}_n^\lambda). \qquad (7.12)
\end{aligned}
$$

Now denoting $\Lambda = \Pi_{i=1}^d \lambda_i^{-\beta_i - \frac{1}{2}}$, $c(\lambda) = \Pi_{i=1}^d \lambda_i^{-\beta_i}$ and $\rho = 2/\gamma$, consider the event

$$
\Omega := \{d_\Delta(\hat{G}_n, G_K^\star) \geq c(\lambda)^{-\frac{2}{1+\rho}} n^{-\frac{1}{1+\rho}} \Lambda^{\frac{2}{1+\rho}}\}.
$$

If the event Ω holds, using Lemma 2 of [22], we get

$$
\begin{aligned}
&\int (f-g)(\mathcal{K}_\lambda * \mathbf{1}_{\{.\in G_K^\star\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{.\in \hat{G}_n\}}) \\
&\leq \frac{d_\Delta^{\frac{1-\rho}{2}}(\hat{G}_{n,m}^\lambda, G^\star)c(\lambda)}{\sqrt{n}} \left[\frac{\nu_n^{(1)}(G^{\star C}) - \nu_n^{(1)}(\hat{G}_{n,m}^{\lambda C})}{c(\lambda)d_\Delta^{\frac{1-\rho}{2}}(\hat{G}_{n,m}^\lambda, G^\star) \vee c(\lambda)^{\frac{2\rho}{(1+\rho)}} n^{-\frac{1-\rho}{2+2\rho}} \Lambda^{\frac{1-\rho}{1+\rho}}}\right. \\
&\qquad \left. + \frac{\nu_n^{(2)}(G^\star) - \nu_n^{(2)}(\hat{G}_{n,m}^\lambda)}{c(\lambda)d_\Delta^{\frac{1-\rho}{2}}(\hat{G}_{n,m}^\lambda, G^\star) \vee c(\lambda)^{\frac{2\rho}{(1+\rho)}} n^{-\frac{1-\rho}{2+2\rho}} \Lambda^{\frac{1-\rho}{1+\rho}}}\right], \\
&\leq \frac{d_{f,g}^{\frac{1-\rho}{2}\frac{\alpha}{\alpha+1}}(\hat{G}_{n,m}^\lambda, G^\star)c(\lambda)}{\sqrt{n}}[V_n^{(1)} + V_n^{(2)}],
\end{aligned}
$$

where for $j \in \{1, 2\}$, $V_n^{(j)}$ is the random variable defined as

$$
V_n^{(j)} = \sup_{G \in \mathcal{G}} \frac{|\nu_n^{(j)}(G^\star) - \nu_n^{(j)}(G)|}{c(\lambda)\|\mathbf{1}_G - \mathbf{1}_{G^*}\|_{2,X^{(j)}}^{1-\rho} \vee c(\lambda)^{\frac{2\rho}{(1+\rho)}} n^{-\frac{1-\rho}{2+2\rho}} \Lambda^{\frac{1-\rho}{1+\rho}}}. \qquad (7.13)
$$

Lemma 3 in Section 6 shows that the variable $V_n^{(1)} + V_n^{(2)}$ has controlled moments. Indeed, the bracketing entropy related to the set $\mathcal{G}$ is $\rho = (d-1)/\gamma = 1/\gamma$. Using the Young's inequality $xy^r \leq ry + (1-r)x^{1/(1-r)}$ with $r = \frac{1-\rho}{2}\frac{\alpha}{\alpha+1}$, we get

$$
\begin{aligned}
&\int (f-g)(\mathcal{K}_\lambda * \mathbf{1}_{\{.\in \hat{G}_n\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{.\in G_K^\star\}}) \\
&\leq c\left(\frac{c(\lambda)}{\tau^{-1}\sqrt{n}}[V_n^{(1)} + V_n^{(2)}]\right)^{\frac{2(\alpha+1)}{\alpha+2+\rho\alpha}} + \tau d_{f,g}(\hat{G}, G_K^\star). \qquad (7.14)
\end{aligned}
$$

Note that

$$
\begin{aligned}
d_{f,g}(\hat{G}, G^\star) &= \int (f-g)(\mathcal{K}_\lambda * \mathbf{1}_{\{.\in \hat{G}_n\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{.\in G_K^\star\}}) + \left(R_K - R_K^\lambda\right)(\hat{G}, G^\star) \\
&\leq \int (f-g)(\mathcal{K}_\lambda * \mathbf{1}_{\{.\in \hat{G}_n\}} - \mathcal{K}_\lambda * \mathbf{1}_{\{.\in G_K^\star\}}) + 2\sup_{G \in \mathcal{G}}\left(R_K - R_K^\lambda\right)(\hat{G}).
\end{aligned}
$$

From above, we have coarselly:

$$d_{f,g}(\hat{G}, G^\star)\mathbf{1}_\Omega \leq \left(\frac{1}{1-\tau}\right) c \left(\frac{c(\lambda)}{\tau^{-1}\sqrt{n}}[V_n^{(1)} + V_n^{(2)}]\right)^{\frac{2(\alpha+1)}{\alpha+2+\rho\alpha}} + 2\sup_{G\in\mathcal{G}}\left(R_K - R_K^\lambda\right)(G).$$

In order to end up the proof, let us consider the following decomposition:

$$d_{f,g}(\hat{G}, G^\star) \quad = \quad d_{f,g}(\hat{G}, G^\star)\mathbf{1}_\Omega + d_{f,g}(\hat{G}, G^\star)\mathbf{1}_{\Omega^C} \tag{7.15}$$

Moreover, note that on the event Ω^C, we have:

$$d_\Delta(\hat{G}_n, G_K^\star) \leq c(\lambda)^{-\frac{2}{1+\rho}} n^{-\frac{1}{1+\rho}} \Lambda^{\frac{2}{1+\rho}} = (n\lambda)^{-\frac{1}{1+\rho}}.$$

Hence, we can conclude that

$$d_{f,g}(\hat{G}, G^\star) \leq c_1\left(\frac{c(\lambda)}{\sqrt{n}}[V_n^{(1)} + V_n^{(2)}]\right)^{\frac{2(\alpha+1)}{\alpha+2+\rho\alpha}} + c_2\sup_{G\in\mathcal{G}}\left|R_K - R_K^\lambda\right|(G) + c_3(n\lambda)^{-\frac{1}{1+\rho}}. \tag{7.16}$$

Integrating the last inequality, we get the result of Theorem 2.

□

7.3 Proof of Theorem 2

The proof is a direct consequence of Proposition 1 as soon as we remark that

$$\mathcal{H}(\mathcal{G}(\gamma, L), d_\Delta, \delta) \leq c\delta^{-\frac{d-1}{\gamma}},\ \forall\ \delta > 0.$$

This result can be found in [27].

7.4 Proof of Corollary 1

Thanks to the previous proof, we only have to propose a bound for the term

$$\sup_{G\in\mathcal{G}}\left|R_K - R_K^\lambda\right|(G).$$

Let $G \in \mathcal{G}$ be fixed. For the sake of convenience, we restrict ourselves to the particular case where $d = 2$. The generalization to larger dimension is straightforward. Moreover, we restrict ourselves to the control of the bias term over the compact $K' = [\epsilon, 1-\epsilon]^{d-1} \times [0,1]$, where $\epsilon > 0$ is a small positive constant chosen later on to have:

$$\left|(R_K - R_K^\lambda)(G) - (R_{K'} - R_{K'}^\lambda)(G)\right| \leq C\psi_n(\alpha,\gamma,\beta), \tag{7.17}$$

where $\psi_n(\alpha,\gamma,\beta)$ is the expected rate of convergence. Using (7.17), we can conduct the proof of Corollary 1 over K'. Then, using Fubini, remark that, provided that $x_1 - \lambda_1 z_1 \in [0,1]$:

$$\begin{aligned}
(R_{K'} - R_{K'}^\lambda)(G) &= \int_{K'} (f-g)\left(\mathcal{K}_\lambda * \mathbf{1}_G - \mathbf{1}_G\right) d\lambda \\
&= \int_{K'} (f-g)(x)\left(\int_{\mathbb{R}^2} \frac{1}{\lambda}\mathcal{K}\left(\frac{x-z}{\lambda}\right)[\mathbf{1}_G(z) - \mathbf{1}_G(x)]\,dz\right)dx \\
&= \int_{K'} (f-g)(x)\left(\int_{\mathbb{R}^2} \mathcal{K}(z)\,[\mathbf{1}_G(x-\lambda z) - \mathbf{1}_G(x)]\,dz\right)dx, \\
&= \int_{\mathbb{R}^2} \mathcal{K}(z)\int_{K'} (f-g)(x)\,[\mathbf{1}_G(x-\lambda z) - \mathbf{1}_G(x)]\,dxdz.
\end{aligned}$$

Note that if $K' = [\epsilon, 1-\epsilon] \times [0,1]$ and $\text{supp}\mathcal{K}_1 = [-M, M]$, for any $\lambda_1 \leq \lambda_{\max}$, the choice of $\epsilon = M\lambda_{\max}$ ensures that:

$$x \in K' \Rightarrow 0 \leq x_1 - \lambda_1 z_1 \leq 1.$$

Moreover, since

$$x \in G \Leftrightarrow 0 \leq x_2 \leq b(x_1),$$

we get:

$$x - \lambda z \in G \Leftrightarrow 0 \leq x_2 \leq \min(1, b(x_1 - \lambda_1 z_1) + \lambda_2 z_2).$$

Finally, for all $z_1 \in \mathbb{R}$, since $b \in \Sigma(\gamma, L)$:

$$b(x_1 - \lambda_1 z_1) = p_{b,x_1}(x_1 - \lambda_1 z_1) + \mathcal{O}(|\lambda_1 z_1|^\gamma). \tag{7.18}$$

Hence, we obtain, using the crude bound $\|f-g\|_\infty \leq 2c_1$ and the assumptions over the kernel $\mathcal{K}$:

$$\begin{aligned}
&(R_{K'} - R_{K'}^\lambda)(G) \\
&= \int_{\mathbb{R}^2} \mathcal{K}(z) \int_\epsilon^{1-\epsilon} \left[\int_{\lambda_2 z_2}^{\min(1, b(x_1-\lambda_1 z_1)+\lambda_2 z_2)} (f-g)(x)dx_2 - \int_0^{b(x_1)} (f-g)(x) dx_2 \right] dx_1 dz \qquad (7.19)\\
&\leq 2c_1 \int_{\mathbb{R}^2} |\mathcal{K}(z)| \int_\epsilon^{1-\epsilon} |p_{b,x_1}(x_1 - \lambda_1 z_1) - b(x_1) + \mathcal{O}(|\lambda_1 z_1|^\gamma) + 2\lambda_2 z_2| \, dx_1 dz \\
&\leq C(\lambda_1 + \lambda_2), \qquad (7.20)
\end{aligned}$$

where $C > 0$ is a generic constant. Using (7.16) and (7.20), we obtain

$$d_{f,g}(\hat{G}, G^\star) \leq c_1 \left(\frac{c(\lambda)}{\sqrt{n}} [V_n^{(1)} + V_n^{(2)}] \right)^{\frac{2(\alpha+1)}{\alpha+2+\rho\alpha}} + C(\lambda_1 + \lambda_2) + c_3 (n\lambda)^{-\frac{1}{1+\rho}}.$$

We can conclude the proof with an appropriate choice for λ_1 and λ_2, noting that, in dimension $d = 2$ for simplicity, (7.17) holds since for any $\lambda_1 \leq \lambda_{\max}$:

$$\begin{aligned}
&\left| (R_K - R_K^\lambda)(G) - (R_{K'} - R_{K'}^\lambda)(G) \right| \leq \left| \int_{K \backslash K'} (f-g) \left(\mathbf{1}_G - \mathcal{K}_\lambda * \mathbf{1}_G \right) d\lambda \right| \\
&= \left| \int_{\mathbb{R}^2} \mathcal{K}(z) \left(\int_0^\epsilon + \int_{1-\epsilon}^1 \right) \left[\int_{\lambda_2 z_2}^{b(x_1 - \lambda_1 z_1)+\lambda_2 z_2} (f-g)(x) dx_2 - \int_0^{b(x_1)} (f-g)(x) dx_2 \right] dx_1 dz \right| \\
&\leq 2\epsilon \leq C(\lambda_1 + \lambda_2),
\end{aligned}$$

for $\epsilon = \lambda_{\max} M$.

□

8 Appendix

8.1 Technical Lemmas

Lemma 2 *Let Z be a random variable having density $f * \eta$ w.r.t. the Lebesgue measure. Assume that η satisfies the Noise assumption and that **(K1)** and **(K2)** hold. Then we have,*

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

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

where $C > 0$ is a generic constant.

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

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

Indeed, for all $s\in\mathbb{R}$, using **(K3)**:

$$
\frac{1}{\lambda^2}\left|\mathcal{F}[\mathcal{K}_\eta(./\lambda)](s)\right|^2=\left|\mathcal{F}[\mathcal{K}_\eta](s\lambda)\right|^2\leq\sup_{t\in\mathbb{R}}\left|\mathcal{F}[\mathcal{K}_\eta](t)\right|^2\leq C\lambda^{-2\beta}, \tag{8.1}
$$

By the same way,

$$
\begin{aligned}
\sup_{x\in\mathbb{R}}|h_{G,\lambda}(x)-h_{G',\lambda}(x)| &= \sup_{x\in\mathbb{R}}\int_{G\Delta G'}\frac{1}{\lambda}\left|\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)\right|dx,\\
&\leq \sup_{x\in\mathbb{R}}\int_K\frac{1}{\lambda}\left|\mathcal{K}_\eta\left(\frac{z-x}{\lambda}\right)\right|dx,\\
&\leq C\sup_{x\in\mathbb{R}}\sqrt{\int\frac{1}{\lambda^2}\mathcal{K}_\eta^2\left(\frac{z-x}{\lambda}\right)dx}\leq\lambda^{-\beta-1/2},
\end{aligned}
$$

where the last line is inspired by (8.1).

□

8.2 Noisy Empirical process theory

In this paragraph, we present the main ingredient for the proof of Theorem 2 and Proposition 1. We intend to analyze the behaviour of the increments of a noisy empirical process related with error measurements. The framework is much more general than model (1.1) and deconvolution classifier of Section 2. Let us fix some notations.
Given a class of functions $\mathcal{G}$, we study the following risk minimization problem:

$$
g^\star=\arg\min_{g\in\mathcal{G}}R(g),
$$

where we have at our disposal a training sample $Z_1,\dots,Z_n$ of i.i.d. random variable with law P_Z. It differs from the law of X, denoted by P_X and since $R(g):=\mathbb{E}[g(X)]$, we are faced to an inverse problem. For this purpose, we consider a indirect ERM procedure that can be written:

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

where $g_\lambda:=\Phi_\lambda(g)$ is a smoothed version of $g\in\mathcal{G}$ and $\lambda\in\Lambda$ is a smoothing parameter (see the particular case $\Phi_\lambda(G)=h_G^\lambda$ in Section 2).

We are interested in the control of the variance of $\hat{g}_n^\lambda$, which is equivalent to the study of the increments of the empirical process ν_n defined as:

$$
\nu_n^\lambda(g)=\frac{1}{\sqrt{n}}\sum_{i=1}^n[g_\lambda(Z_i)-\mathbb{E}g_\lambda(Z)]. \tag{8.2}
$$

The empirical process (8.2) is indexed by the set $\mathcal{G}^\lambda = \{g_\lambda = \Phi_\lambda(g),\, g \in \mathcal{G}\}$. We are interested in the behaviour of this empirical process near some fixed function g_0, and we denote a neighbourhood of $g_0 \in \mathcal{G}$ by

$$\mathcal{G}(\delta) = \{g \in \mathcal{G} : \|g - g_0\|_{2,P_X} \leq \delta\}. \tag{8.3}$$

It is important to note that the localization is performed on the set $\mathcal{G}$ since we want to estimate $g^* \in \mathcal{G}$. The aim is to measure the influence of the parameter $\lambda \in \Lambda$ in the behaviour of (8.2) in the neighbourhood $\mathcal{G}(\delta)$. If we deal with a kernel deconvolution classifier, $\lambda \in \mathbb{R}^d$ is a set of bandwidths of a deconvolution kernel. However, in such a generality, λ could be any kind of regularization parameter (see [19]).

In order to apply concentration inequalities of Bernstein's type, we need the two following assumptions:

(A1) for any $\lambda \in \Lambda$, there exists $b(\lambda) : \sup_{g \in \mathcal{G}} \|g_\lambda\|_\infty \leq b(\lambda)$.

(A2) There exists a pseudo-distance d on $\mathcal{G}$ such that $\forall g, g' \in \mathcal{G}$, $\|g_\lambda - g'_\lambda\|_{2,Z} \leq c(\lambda) d(g, g')$.

These assumptions are satisfied for the particular case of the paper with Lemma 2 above where $b(\lambda) = \Pi_{i=1}^d \lambda_i^{-\beta_i - 1/2}$, $d = d_\Delta$ and $c(\lambda) = \Pi_{i=1}^d \lambda_i^{-\beta_i}$. The first assumption **(A1)** is necessary to use standard uniform concentration inequalities in the bounded case (such as Bernstein or Talagrand inequalities). Moreover, **(A2)** ensures a control of the entropy of $\mathcal{G}_\lambda = \{g_\lambda, g \in \mathcal{G}\}$ thanks to standard entropy condition over the pseudo-metric space $(\mathcal{G}, d)$. Indeed, using for instance [27], we have under the second asumption:

$$\mathcal{H}(\mathcal{G}^\lambda, \delta, L_2(P_Z)) \leq \mathcal{H}\left(\mathcal{G}, \frac{\delta}{c(\lambda)}, d\right).$$

Next lemma proposes a control of the increments of the noisy empirical process (8.2) when the class $\mathcal{G}^\lambda$ satisfies the two previous assumptions.

Lemma 3 *Consider a class of functions $\{g_\lambda, g \in \mathcal{G}\}$ satisfying **(A1)-(A2)**. Let $g_0 \in \mathcal{G}$ and $\mathcal{G}(\delta)$ the set introduced in (8.3). Suppose there exists some $0 < \alpha < 2$ such that:*

$$\mathcal{H}_B(\mathcal{G}, \delta, d) \leq c' \delta^{-\alpha}. \tag{8.4}$$

Let us consider $n_0 = \inf\{n \in \mathbb{N}^* : \delta_n(\lambda) < 1\}$ where

$$\delta_n(\lambda) := c(\lambda)^{\frac{\alpha}{2+\alpha}} b(\lambda)^{\frac{2}{2+\alpha}} n^{-\frac{1}{2+\alpha}}.$$

- *Then there exist constants c_1, c'_1 which depend on α, c' such that $\forall n \geq n_0$, $\forall \delta \in [\delta_n(\lambda), 1[$:*

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}(\delta)} |\nu_n^\lambda(g) - \nu_n^\lambda(g^0)| \geq c_1 c(\lambda) \delta^{1-\frac{\alpha}{2}}\right) \leq \exp\left(-c'_1 \delta^{-\alpha}\right).$$

- *There exists constants $c_2, c'_2 > 0$ which depends on α, c' such that for $T \geq c'_2$, for $n \geq n_0$:*

$$\mathbb{P}\left(\sup_{g \notin \mathcal{G}(\delta_n(\lambda))} \frac{|\nu_n^\lambda(g) - \nu_n^\lambda(g^0)|}{c(\lambda)\|g - g_0\|^{1-\frac{\alpha}{2}}} \geq T\right) \leq \exp\left(-\frac{T}{c_2}\right).$$

The proof is an application of [13] and consists in a noisy version of Lemma 5.13 in [13]. This result is of pratical interest to control the variance of our deconvolution estimator. In particular, we use in the proofs of Theorem 2 the fact that:

$$V_n = \sup_{g \in \mathcal{G}} \frac{|\nu_n^\lambda(g) - \nu_n^\lambda(g^0)|}{\|g - g_0\|^{1-\alpha/2} c(\lambda) \vee c(\lambda)^{\frac{2\alpha}{2+\alpha}} n^{-\frac{2-\alpha}{2(2+\alpha)}} b(\lambda)^{\frac{2-\alpha}{2+\alpha}}} = 0_{\mathbb{P}}(1),$$

as $n \to +\infty$.

Proof From Lemma 5.7, gathering with Lemma 5.8 of [13], for any $g \in \mathcal{G}(\delta)$, we have:

$$\mathbb{P}(\nu_n^\lambda(g) - \nu_n^\lambda(g_0) \geq a) \leq \exp\left(-\frac{a^2}{8c(\lambda)^2\delta^2}\right), \forall a \leq \sqrt{n}\frac{c(\lambda)^2}{b(\lambda)}\delta^2.$$

Next step is to use the following noisy version of Theorem 5.11. For any $a > 0$ satisfying:

$$C_0\left[\sqrt{2}\delta c(\lambda) \vee \int_{a/2^6\sqrt{n}}^{\sqrt{2}\delta c(\lambda)} \sqrt{\mathcal{H}_B(\mathcal{G}^\lambda, u, L_2(P_Z))}du\right] \leq a \leq \sqrt{n}\frac{c(\lambda)^2}{K(\lambda)}\delta^2 \wedge 8\sqrt{2n}\delta c(\lambda), \tag{8.5}$$

for some universal constant $C_0 > 0$, we have:

$$\mathbb{P}(\sup_{g\in\mathcal{G}(\delta)} |\nu_n^\lambda(g) - \nu_n^\lambda(g_0)| \geq a) \leq \exp\left(-\frac{a^2}{4C\delta^2c(\lambda)^2}\right),$$

where $C > 0$ depends on C_0.

Hence from assumption (8.4), for $n \geq n_0$, we have for any $\delta_n(\lambda) \leq \delta < 1$, by choosing $a = c_1c(\lambda)\delta^{1-\frac{\alpha}{2}}$ in (8.5):

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}(\delta)} |\nu_n^\lambda(g) - \nu_n^\lambda(g_0)| \geq c_1c(\lambda)\delta^{1-\frac{\alpha}{2}}\right) \leq \exp\left(-c_1'\delta^{-\alpha}\right).$$

To show the second statement, we apply the peeling device as in [13]. Introduce:

$$S = \inf\{s \geq 1 : 2^{-s} < \delta_n(\lambda)\}.$$

Then we have, for $T = 2^{1-\frac{\alpha}{2}}c_1$:

$$\begin{aligned}
\mathbb{P}\left(\sup_{g\notin\mathcal{G}(\delta_n(\lambda))} \frac{|\nu_n^\lambda(g) - \nu_n^\lambda(g_0)|}{c(\lambda)\|g - g_0\|^{1-\frac{\alpha}{2}}} \geq T\right) &\leq \sum_{s=1}^{S} \mathbb{P}\left(\sup_{2^{-s}\leq\|g-g^0\|\leq2^{-s+1}} \frac{|\nu_n^\lambda(g) - \nu_n^\lambda(g_0)|}{\|g - g_0\|^{1-\frac{\alpha}{2}}} \geq c(\lambda)T\right)\\
&\leq \sum_{s=1}^{S} \mathbb{P}\left(\sup_{g\in\mathcal{G}(2^{-s+1})} |\nu_n^\lambda(g) - \nu_n^\lambda(g_0)| \geq c(\lambda)c_1\left(2^{-s+1}\right)^{1-\frac{\alpha}{2}}\right)\\
&\leq \sum_{s=1}^{S} \exp\left(-c_1'\left(2^{-s+1}\right)^{-\alpha}\right) = \exp\left(-\frac{T}{c_2}\right),
\end{aligned}$$

where $c_2 > 0$ is a function of α, c_1 and c_1'.

□

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

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

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

[6] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24, 2008.

[7] M. Chichignoud and S. Loustau. Adaptive noisy clustering. Submitted, 2013.

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

[9] A. Delaigle and I. Gijbels. Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Statist. Math.*, 56:19–47, 2004.

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

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

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

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

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

[15] V.I. Koltchinskii. Empirical geometry of multivariate data: a deconvolution approach. *The Annals of Statistics*, 2000.

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

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

[18] 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.

[19] S. Loustau. Inverse statistical learning. In (minor) revision to Electronic Journal of Statistics, 2012.

[20] S. Loustau. Anisotropic oracle inequalities in noisy quantization. Submitted, 2013.

[21] S. Loustau and C. Marteau. Minimax fast rates for discriminant analysis with error in variables. In (minor) revision to Bernoulli, 2012.

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

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

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

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

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

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