Fast rates for plug-in estimators of density level sets

Philippe Rigollet, Régis Vert

Optimal rates for plug-in estimators of density level sets

Philippe Rigollet * and Régis Vert †

February 11, 2008

Abstract

In the context of density level set estimation, we study the convergence for general plug-in methods under two main assumptions on the density for a given level λ. More precisely, it is assumed that the density (i) is smooth in a neighborhood of λ and (ii) has γ-exponent at level λ. Condition (i) ensures that the density can be estimated at a standard nonparametric rate and condition (ii) is similar to Tsybakov's margin assumption that is stated for the classification framework. Under these assumptions, we derive fast rates of convergence for plug-in estimators, up to to order n^{-1}. Explicit convergence rates are given for plug-in estimators based on kernel density estimators when the underlying measure is the Lebesgue measure. Lower bounds proving optimality of the rates in a minimax sense when the density is Hölder smooth are also provided.

Mathematics Subject Classifications: Primary 62G05, Secondary 62C20, 62G05, 62G20.
Key Words: Density level sets, plug-in estimators, fast rates of convergence, kernel density estimators, minimax lower bounds.
Short title: Plug-in density level set estimation.

1 Introduction

Let Q be a positive σ-finite measure on $\mathcal{X} \subseteq \mathbb{R}^d$. Consider i.i.d random vectors $(X_1, \dots, X_n)$ with distribution P, having an unknown probability

*Georgia Institute of Technology. rigollet@math.gatech.edu
†Masagroup. regis.vert@masagroup.net

density p with respect to the measure Q. For a fixed $\lambda > 0$, we are interested in the estimation of the λ-*level set* of the density p:

$$\Gamma_p(\lambda) \triangleq \{x \in \mathcal{X} \ : \ p(x) \geq \lambda\} \ .$$

Throughout the paper we fix $\lambda > 0$ and when no confusion is possible we use the notation $\Gamma(\lambda)$ or simply Γ instead of $\Gamma_p(\lambda)$. When Q is the Lebesgue measure on $\mathbb{R}^d$, density level sets typically correspond to minimum volume sets of given P-probability mass, as shown in Polonik (1997).

Here are two possible applications of density level set estimation.

Anomaly detection: the goal is to detect an abnormal observation from a sample (see for example Steinwart et al., 2005, and references therein). One way to deal with that problem is to assume that abnormal observations do not belong to a group of concentrated observations. In this framework, observations are considered as abnormal when they do not belong to $\Gamma(\lambda)$ for some fixed $\lambda \geq 0$. The special case $\lambda = 0$, which corresponds to support estimation has been examined by Devroye and Wise (1980). In the general case, λ can be considered as a tolerance level for anomalies: the smaller λ, the fewer observations are considered as being abnormal.

Unsupervised or semi-supervised classification: these two problems amount to identify areas where the observations are concentrated with possible use of some available labels for the semi-supervised case. For instance, it can be assumed that the connected components of $\Gamma(\lambda)$, for a fixed λ, are clusters of homogeneous observations as described in Hartigan (1975). Note that this definition has been refined for example in Stuetzle (2003).

Remark 1.1 *In both applications, the choice of λ is critical and has to be addressed carefully. However, we do not treat this problem in this paper.*

There are essentially two approaches towards estimating density level sets from the sample $(X_1, \ldots, X_n)$. The most straightforward is to resort to *plug-in* methods where the density p in the expression for $\Gamma_p(\lambda)$ is replaced by its estimate computed from the sample. Another way to estimate density level sets is to resort to *direct* methods which are based on empirical excess-mass maximization. The *excess-mass* H is a functional that measures the quality of an estimator $\hat{G}$ and is defined as follows (Hartigan, 1987; Müller and Sawitzki, 1987):

$$H\big(\hat{G}\big) = P\big(\hat{G}\big) - \lambda Q\big(\hat{G}\big) \ .$$

Excess-mass measures how the P-probability mass concentrates in the region $\hat{G}$, and it is maximized by $\Gamma = \Gamma(\lambda)$. Hence, it acts as a risk functional in the density level set estimation (DLSE) framework and it is natural to measure the performance of an estimator $\hat{G}$ by its *excess-mass deficit* $H(\Gamma) - H(\hat{G}) \geq 0$. Further justifications for the well-foundedness of the excess mass criterion can be found in Polonik (1995). Recently, Gayraud and Rousseau (2005) proposed a Bayesian approach to DLSE and together with interesting comparative simulations.

While local versions of direct methods have been deeply analyzed and proved to be optimal in a minimax sense, over a certain family of well-behaved distributions (see Tsybakov, 1997), and although reasonable implementations have been recently proposed (see for instance Steinwart et al., 2005), they are still not very easy to use for practical purposes, compared to plug-in methods. Indeed, in practice, rather than specifying a value for λ, the user can specify a value for α, the P-probability mass of the level set. In this case, the value of λ is implied by that of α and efficient direct methods can be derived (Scott and Nowak, 2006). However, in general, using direct methods, one has to run an optimization procedure several times, one for different density level values, then choose a posteriori the most suited level according to the desired rejection rate. Plug-in methods do not involve such a complex process: the density estimation step is only performed once and the construction of a density level set estimate simply amounts to thresholding the density estimate at the desired level.

On the other hand, in the related context of binary classification where more theoretical advances have been developed, the different analysis proposed so far have mainly supported a belief in the superiority of direct methods. Yang (1999) shows that, under general assumptions, plug-in estimators cannot achieve a classification error risk convergence rate faster than $O\big(1/\sqrt{n}\big)$ (where n is the size of the data sample), and suffer from the curse of dimensionality. In contrast to that, under slightly different assumptions, direct methods achieve this rate $O\big(1/\sqrt{n}\big)$ whatever the dimensionality (see e.g. Vapnik, 1998; Devroye et al., 1996; Tsybakov, 2004b), and can even reach faster convergence rates- up to $O\big(1/n\big)$- under *Tsybakov's margin assumption* (see Mammen and Tsybakov, 1999; Tsybakov, 2004b; Tsybakov and van de Geer, 2005; Tarigan and van de Geer, 2006). This contributed to raising some pessimism concerning plug-in methods. Nevertheless such a comparison between plug-in methods and direct methods is far from being legitimate, since the aforementioned analyzes of both plug-in methods and direct ones have been carried out under the different sets of assumptions (those sets are not disjoint, but none of them is included in the other).

Recently, in the standard classification framework, Audibert and Tsybakov (2007) have combined a new type of assumption dealing with the smoothness of the *regression function* and the well known margin assumption. Under these assumptions, they derive fast convergence rates- even faster than $O\big(1/n\big)$ in some situations- for plug-in classification rules based on local polynomial estimators. This new result reveals that plug-in methods should not be considered as inferior to direct methods and, more importantly, that this new type of assumption on the regression function is a critical point in the general analysis of classification procedures.

In this paper we extend such positive results to the DLSE framework: we revisit the analysis of plug-in density level set estimators, and show that they can be also very efficient under smoothness assumptions on the underlying density function p. Unlike the global smoothness assumption used in Audibert and Tsybakov (2007), the local smoothness assumption introduced here emphasizes the predominant role of the smoothness close to the level λ as opposed to the smoothness for values of p far from the level under consideration. Related papers are Baíllo et al. (2001) and Baíllo (2003), who investigate plug-in estimators based on a certain type of kernel density estimates. Baíllo et al. (2001) also study the convergence for the symmetric difference under other assumptions and Baíllo (2003) derives almost sure rates of convergence for a quantity different from the one studied here. It is interesting to observe that she introduces a condition similar to the γ-exponent used here. Note however, that this definition of γ-exponent allows the density to have flat parts at level λ that cannot be estimated consistently by standard plug-in estimators when using the symmetric difference as a measure of error. To be able to achieve consistency in this setup, one has to resort to plug-in estimators with a positive offset.

The particular case $\lambda = 0$, corresponds to estimation of the support of density p and is often applied to anomaly detection. Following the pioneer paper of Devroye and Wise (1980), this problem has received more attention than the general case $\lambda \geq 0$ and has been treated using plug-in methods for example by Cuevas and Fraiman (1997). Unlike the previously cited papers, we derive fast rates of convergence and prove that these rates are optimal in a minimax sense. However, we do not treat the case $\lambda = 0$ for the which the rates are typically different than for $\lambda > 0$ as pointed out by Tsybakov (1997) for example. The techniques employed in the present analysis cannot be extended straightforwardly to that case.

A general plug-in approach has been studied previously by Molchanov (1998), where a result on the asymptotic distribution of the Hausdorff distance is given. In a recent paper, Cuevas et al. (2006) study general plug-in

estimators of the level sets. Under very general assumptions they derive consistency with respect to the Hausdorff metric and the measure of the symmetric difference. However, this very general framework does not allow them to derive rates of convergence.

This paper is organized as follows. Section 2 introduces the notation and definitions. Section 3 presents the main result, that is a new bound on the error of plug-in estimators based on general density estimators that satisfy a certain exponential inequality. We then apply, in Section 4, this result to the particular case of kernel density estimators, under the assumption that the underlying density belongs to some locally Hölder class of densities. Finally, minimax lower bounds are given in Section 5, as a way to assess the optimality of the upper bounds involved in the main result.

2 Notation and Setup

For any vector $x \in \mathbb{R}^d$, denote by $x^{(j)}$ its jth coordinate, $j = 1, \ldots, d$. Denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^d$ and by $\mathcal{B}(x, r)$ the closed Euclidean ball in $\mathcal{X}$ centered at $x \in \mathcal{X}$ and of radius $r > 0$.

The probability and expectation with respect to the joint distribution of $(X_1, \ldots, X_n)$ are denoted by $\mathbb{P}$ and $\mathbb{E}$ respectively. For any function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$ the sup-norm of f and by $\|f\| = \left(\int_{\mathbb{R}^d} f^2(x) \mathrm{d}x \right)^{1/2}$ its L_2-norm. Also, for any measurable function f on $\mathcal{X}$ and any set $A \subset f(\mathcal{X})$, we write for simplicity $\{x \in \mathcal{X} : f(x) \in A\} = \{f \in A\}$. Throughout the paper, we denote by C positive constants that can change from line to line and by c_j positive constants that have to be identified. Finally, A^c denotes the complement of the set A.

2.1 Plug-in density level set estimators with offset

For a fixed $\lambda > 0$, the plug-in estimator of $\Gamma(\lambda)$ is defined by

$$\hat{\Gamma}(\lambda) = \{x \in \mathcal{X} \, : \, \hat{p}_n(x) \geq \lambda\} \, ,$$

where $\hat{p}_n$ is a nonparametric estimator of p. For example, $\hat{p}_n$ can be a kernel density estimator of p,

$$\hat{p}_n(x) = \hat{p}_{n,h}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h} \right), \; x \in \mathcal{X} \, ,$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a suitably chosen kernel and $h > 0$ is the bandwidth parameter. For reasons that will be made clearer later, we consider in this

paper the more general family of *plug-in estimators with offset* ℓ_n, denoted by $\tilde{\Gamma}_{\ell_n}$ and defined as follows:

$$\tilde{\Gamma}_{\ell_n} = \tilde{\Gamma}_{\ell_n}(\lambda) = \hat{\Gamma}(\lambda + \ell_n) = \{x \in \mathcal{X} \,:\, \hat{p}_n(x) \geq \lambda + \ell_n\}\,,$$

where ℓ_n is a non-negative quantity that typically tends to 0 as n tends to infinity. This family includes in particular the estimator $\hat{\Gamma}$ when ℓ_n is taken equal to 0.

2.2 Measures of performance

Recall that Q is a positive σ-finite measure on $\mathcal{X}$ and define the measure $\tilde{Q}_\lambda$ that has density $|p(\cdot) - \lambda|$ with respect to Q. To assess the performance of a density level set estimator, we use the two pseudo-distances between closed sets G_1 and $G_2 \subseteq \mathcal{X}$:

(i) The Q-measure of the symmetric difference between G_1 and G_2:

$$d_{\triangle}(G_1, G_2) = Q(G_1 \triangle G_2)\,.$$

(ii) The $\tilde{Q}_\lambda$-measure of the symmetric difference between G_1 and G_2:

$$d_H(G_1, G_2) = \tilde{Q}_\lambda(G_1 \triangle G_2) = \int_{G_1 \triangle G_2} |p(x) - \lambda| \mathrm{d}Q(x)\,.$$

The quantity $d_{\triangle}(G_1, G_2)$ is a standard and natural way to measure the distance between two sets G_1 and G_2. Note that for any measurable set $G \subseteq \mathcal{X}$, the excess-mass $H(G)$ can be written

$$H(G) = \int_G (p(x) - \lambda)\, \mathrm{d}Q(x)\,.$$

Thus, we can rewrite,

$$\begin{aligned} H(\Gamma) - H(\hat{G}) &= \int_{\mathcal{X}} \left(\mathbb{1}_{\{p(\cdot) \geq \lambda\}}(x) - \mathbb{1}_{\hat{G}}(x)\right) (p(x) - \lambda)\, \mathrm{d}Q(x) \\ &= \int_{\Gamma \triangle \hat{G}} |p(x) - \lambda| \mathrm{d}Q(x) = d_H(\hat{G}, \Gamma)\,. \end{aligned}$$

This explains the notation d_H.

The following definition allows us to link d_H to $d_{\triangle}$.

Definition 2.1 *For any $\lambda, \gamma \geq 0$, a function $f : \mathcal{X} \to \mathbb{R}$ is said to have* **γ-exponent at level λ** *with respect to Q if there exist constants $c_0 > 0$ and $\varepsilon_0 > 0$ such that, for all $0 < \varepsilon \leq \varepsilon_0$,*

$$Q\{x \in \mathcal{X} : 0 < |f(x) - \lambda| \leq \varepsilon\} \leq c_0 \varepsilon^{\gamma} .$$

The assumption under which the underlying density has γ-exponent at level λ was first introduced by Polonik (1995). Its counterpart in the context of binary classification is commonly referred to as *margin assumption* (see Mammen and Tsybakov, 1999; Tsybakov, 2004b).

The exponent γ controls the slope of the function around level λ. When $\gamma = 0$, the condition is loose and when γ is positive, it constrains the rate at which the function approaches the level λ. A standard case corresponds to $\gamma = 1$, arising for instance in the case where the gradient of f has a coordinate bounded away from 0 in a neighborhood of $\{f = \lambda\}$.

We now show that the pseudo-distances $d_{\triangle}$ and d_H are linked when the density p has γ-exponent at level λ. The following proposition is a direct consequence of Proposition 6.1.

Proposition 2.1 *Fix $\lambda > 0$ and $\gamma \geq 0$. If the density p has γ-exponent at level λ w.r.t Q, then for any $L_Q > 0$, there exists $C > 0$ such that for any G_1, G_2 satisfying $Q(G_1 \triangle G_2) \leq L_Q$ we have*

$$d_{\triangle}(G_1, G_2) \leq Q(G_1 \triangle G_2 \cap \{p = \lambda\}) + C \left(d_H(G_1, G_2)\right)^{\frac{\gamma}{1+\gamma}} .$$

When controlling the expected distance $d_{\triangle}$ of an estimator to the density level set, the first term will be made negligible by using plug-in estimators with a sufficiently large offset. Note that when using plug-in estimators without offset, it is not possible to make the first term small enough and the use of the offset is justified.

3 Fast rates for plug-in density level sets estimators with offset

The first theorem states that rates of convergence for plug-in estimators with offset can be obtained using exponential inequalities for the corresponding nonparametric density estimator $\hat{p}_n$. In what follows, smoothness in the neighborhood of the level under consideration is particularly important and we define this neighborhood as follows:

$$\mathcal{D}(\eta) = \left\{p \in \ (\lambda - \eta, \lambda + \eta) \ \right\}, \quad \eta > 0$$

In the sequel, we will always use plug-in estimators with the same offset and we write for simplicity $\tilde{\Gamma}_{\ell_n} = \tilde{\Gamma}$.

Theorem 3.1 *Fix $\lambda > 0$ and $\Delta > \lambda$. Let $\hat{p}_n$ be an estimator of the density p such that $Q(\hat{p}_n \geq \lambda) \leq M$, almost surely for some positive constant M and let $\mathcal{P}$ be class of densities on $\mathcal{X}$. Assume that there exists positive constants $\eta, c_1, c_2, c_3, c_4, c_\delta, c'_\delta, a$ and b, such that*

- *for Q-almost all $x \in \mathcal{D}(\eta)$ and for any δ such that $c_\delta n^{-a/2} < \delta < \Delta$, we have*

$$\sup_{p \in \mathcal{P}} \mathbb{P}\left(|\hat{p}_n(x) - p(x)| \geq \delta\right) \leq c_1 \mathrm{e}^{-c_2 n^a \delta^2}\,,\ n \geq 1\,, \tag{3.1}$$

- *for Q-almost all $x \in \mathcal{X} \setminus \mathcal{D}(\eta)$, for any δ such that $c'_\delta n^{-b/2} < \delta < \Delta$, we have*

$$\sup_{p \in \mathcal{P}} \mathbb{P}\left(|\hat{p}_n(x) - p(x)| \geq \delta\right) \leq c_3 \mathrm{e}^{-c_4 n^a \delta^2}\,,\ n \geq 1\,. \tag{3.2}$$

Then if p has γ-exponent at level λ for any $p \in \mathcal{P}$, the plug-in estimator $\tilde{\Gamma}$ with offset $\ell_n = n^{-\nu}$ for some $0 < \nu < a/2$ satisfies

$$\sup_{p \in \mathcal{P}} \mathbb{E}\left[d_H(\Gamma_p(\lambda), \tilde{\Gamma})\right] \leq c_5 n^{-\frac{(1+\gamma)a}{2}}\,, \tag{3.3}$$

$$\sup_{p \in \mathcal{P}} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda), \tilde{\Gamma})\right] \leq c_6 n^{-\frac{\gamma a}{2}}\,, \tag{3.4}$$

for $n \geq n_0 = n_0(\lambda, \eta, a, b, \varepsilon_0, c_\delta, c'_\delta)$ and where $c_5 > 0$ and $c_6 > 0$ depend only on c_1, c_2, c_3, c_4, M, a, b,γ and λ.

Before giving the proof of the theorem, we comment on its meaning. First note that the main consequence of (3.1) is that $|\hat{p}_n(x) - p(x)|$ is of order n^{-a} with polynomially high probability up to some logarithmic factors for any x in the neighborhood $\mathcal{D}(\eta)$. That is $\hat{p}_n$ is a good pointwise estimator of p in this neighborhood. Equation (3.2) is of the same flavor as (3.1) but in a weaker form. It entails that for x outside of $\mathcal{D}(\eta)$, $\hat{p}_n(x)$ is a consistent estimator of $p(x)$ with a polynomial rate of order $n^{-a \wedge b}$ up to a logarithmic factor that can be as slow as desired since b does not appear in the rates (3.3) or (3.4).

PROOF. Note first that the conditions of Proposition 2.1 are satisfied. Indeed, $Q(\{\hat{p}_n \geq \lambda + \ell_n\} \triangle \{p \geq \lambda\}) \leq Q(\hat{p}_n \geq \lambda) + Q(p \geq \lambda) \leq M + \lambda^{-1}$ and we choose $L_Q = M + \lambda^{-1}$. Therefore, if we prove that

$$\mathbb{E}Q\big(\Gamma_p(\lambda) \triangle \tilde{\Gamma} \cap \{p = \lambda\}\big) \leq Cn^{-\frac{\gamma a}{2}} \tag{3.5}$$

then (3.4) is a direct consequence of (3.3) and (3.5). We begin by proving (3.5). Remark that

$$\Gamma_p(\lambda) \triangle \tilde{\Gamma} \cap \{p = \lambda\} = \{\hat{p}_n + \ell_n < \lambda\} \cap \{p = \lambda\} \subset \{|\hat{p}_n - p| > \ell_n\} \cap \mathcal{D}(\eta)\,.$$

Therefore by the Fubini Theorem and assumption (3.1),

$$\mathbb{E}Q\big(\Gamma_p(\lambda) \triangle \tilde{\Gamma} \cap \{p = \lambda\}\big) \leq Q\{p = \lambda\}\mathrm{e}^{-c_2 n^a \ell_n^2} \leq Cn^{-\frac{\gamma a}{2}}\,,$$

which proves (3.5).

To prove (3.3), we use the same scheme as in the proof of Audibert and Tsybakov (2007, Theorem 3.1). Recall that $\tilde{\Gamma} \triangle \Gamma = (\tilde{\Gamma} \cap \Gamma^c) \cup (\tilde{\Gamma}^c \cap \Gamma)$. It yields

$$\mathbb{E}\left[d_H\big(\Gamma, \tilde{\Gamma}\big)\right] = \mathbb{E}\int_{\tilde{\Gamma}\cap\Gamma^c} |p(x) - \lambda| \mathrm{d}Q(x) + \mathbb{E}\int_{\tilde{\Gamma}^c\cap\Gamma} |p(x) - \lambda| \mathrm{d}Q(x)\,.$$

Define two sequences

$$\ell_n^a = n^{-a/2} \quad \text{and} \quad \ell_n^b = \left(\frac{c_4 n^{a\wedge b}}{2(1+\gamma)a\log n}\right)^{-1/2}\,.$$

Let n_0 be a positive integer such that $\ell_n^a < \ell_n^b < \eta \wedge \varepsilon_0 = r$ and $\ell_n^b > c'_\delta n^{-b/2}$ for all $n \geq n_0$. Consider the following disjoint decomposition:

$$\tilde{\Gamma}^c \cap \Gamma = \{\hat{p}_n < \lambda + \ell_n,\, p \geq \lambda\} \subset A_1 \cup A_2 \cup A_3\,, \tag{3.6}$$

where,

$$\begin{aligned}
A_1 &= \{\hat{p}_n < \lambda + \ell_n,\ \lambda \leq p \leq \lambda + \ell_n^a\}\,,\\
A_2 &= \{\hat{p}_n < \lambda + \ell_n,\ \lambda + \ell_n^a < p \leq \lambda + \ell_n^b\}\,,\\
A_3 &= \{\hat{p}_n < \lambda + \ell_n,\ p > \lambda + \ell_n^b\}\,.
\end{aligned}$$

Observe that $A_1 \subseteq \{|p - \lambda| \leq \ell_n^a\}$. It yields for $n \geq n_0$,

$$\mathbb{E}\int_{A_1} |p(x) - \lambda| \mathrm{d}Q(x) \leq \ell_n^a Q(A_1 \cap \{|p - \lambda| > 0\}) \leq c_0(\ell_n^a)^{1+\gamma} = c_0 n^{-\frac{(1+\gamma)a}{2}}\,, \tag{3.7}$$

where in the last inequality we used the γ-exponent of p. Then when $n \geq n_0$, we can decompose A_2 into the disjoint union:

$$A_2 = \bigcup_{j=1}^{J_n} \mathcal{X}_j, \quad \mathcal{X}_j = \left\{\hat{p}_n < \lambda + \ell_n,\, \lambda + 2^{j-1}\ell_n^a < p \leq \lambda + 2^j \ell_n^a\right\} \cap \mathcal{D}(r)\,,$$

where $J_n = \lfloor \log_2 \left(\frac{\ell_n^b}{\ell_n^a}\right) \rfloor + 2$ so that the $\mathcal{X}_j$ indeed form a partition of A_2. Hence,

$$\mathbb{E}\int_{A_2} |p(x) - \lambda| \mathrm{d}Q(x) = \sum_{j=1}^{J_n} \mathbb{E}\int_{\mathcal{X}_j} |p(x) - \lambda| \mathrm{d}Q(x)\,. \tag{3.8}$$

Remark that for sufficiently large n, we have $\ell_n \leq \ell_n^a/2$. It yields

$$\mathcal{X}_j \subset \{|\hat{p}_n - p| > 2^{j-2}\ell_n^a\} \cap \{|p(x) - \lambda| < 2^j \ell_n^a\}\,.$$

Using the Fubini Theorem and the previous inclusion, the general term of the sum in the right-hand side of (3.8) can be bounded from above by

$$2^j \ell_n^a \int_{\mathcal{D}(r)} \mathbb{P}\left[|\hat{p}_n(x) - p(x)| > 2^{j-2}\ell_n^a\right] \mathbb{I}_{\{0<|p(x)-\lambda|<2^j\ell_n^a\}} \mathrm{d}Q(x)\,.$$

Remark that for any $j \leq J_n$, we have $2^{j-2}\ell_n^a \leq \ell_n^b \leq \Delta$ for sufficiently large n. Using now (3.1) and the fact that p has γ-exponent at level λ, we get

$$\begin{aligned} \mathbb{E}\int_{A_2} |p(x) - \lambda| \mathrm{d}Q(x) &\leq c_0 c_1 \sum_{j\geq 1} \exp\left(-c_2 n^a (2^{j-2}\ell_n^a)^2\right) (2^j \ell_n^a)^{1+\gamma} \\ &\leq C(\ell_n^a)^{1+\gamma} = C n^{-\frac{(1+\gamma)a}{2}}\,. \end{aligned} \tag{3.9}$$

We now treat the integral over A_3 using the Fubini theorem and the fact that $\ell_n \leq \ell_n^b/2$ for sufficiently large n. We obtain

$$\begin{aligned} \mathbb{E}\int_{A_3} |p(x) - \lambda| \mathrm{d}Q(x) &\leq \sup_{\substack{G\subseteq\mathcal{X} \\ Q(G)\leq 1/\lambda}} \int_G |p(x) - \lambda| \mathbb{P}\left[|\hat{p}_n(x) - p(x)| > \ell_n^b/2\right] \mathrm{d}Q(x) \\ &\leq 2c_3 \exp\left(-c_4 n^a (\ell_n^b/2)^2\right) \leq 2c_3 n^{-\frac{(1+\gamma)a}{2}}\,, \end{aligned} \tag{3.10}$$

where in the last inequality, we used the fact that

$$\ell_n^b \geq \left(\frac{c_4 n^a}{2(1+\gamma)a \log n}\right)^{-1/2}\,.$$

In view of (3.6), if we combine (3.7), (3.9) and (3.10), we obtain

$$\mathbb{E}\int_{\tilde{\Gamma}^c\cap\Gamma} |p(x)-\lambda|\mathrm{d}Q(x) \leq Cn^{-\frac{(1+\gamma)a}{2}}\,.$$

In the same manner, it can be shown that for $n \geq n_0$,

$$\mathbb{E}\int_{\tilde{\Gamma}\cap\Gamma^c} |p(x)-\lambda|\mathrm{d}Q(x) \leq Cn^{-\frac{(1+\gamma)a}{2}}\,.$$

The only difference with the part of the proof detailed above is that in the step that corresponds to proving the equivalent of (3.10), we use the assumption that $Q(\hat{p}_n \geq \lambda) \leq M$, a.s. ∎

Remark 3.1 *It is sometimes the case, for some applications that $\tilde{\Gamma}$ is required to be included in Γ with high probability. When the offset ℓ_n is chosen sufficiently large, i.e. of order at least $n^{-a/2}$, it can be shown that the resulting performance of the density level set estimator is only altered by a logarithmic factor whereas it can be enforced that,*

$$\mathbb{E}Q\big(\tilde{\Gamma}\cap\Gamma^c\big) \leq Cn^{-\alpha}\,,$$

for any $\alpha > 0$ (Rigollet, 2007). In other words, $\tilde{\Gamma}$ is included in Γ with arbitrarily large probability.

4 Fast rates for plug-in estimators with offset based on kernel density estimators

In the rest of this paper, we fix the measure Q to be the Lebesgue measure on $\mathbb{R}^d$ denote by Leb_d.

In this section, we derive exponential inequalities of the type (3.1) when the estimator $\hat{p}_n$ is a kernel density estimator and the density p belongs to some Hölder class of densities. We begin by giving the definition of the Hölder classes of densities that we consider.

4.1 Hölder classes of densities

Fix $\beta > 0$ and $\lambda > 0$. For any d-tuples $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ and $x = (x_1, \ldots, x_d) \in \mathcal{X}$, we define $|s| = s_1 + \ldots + s_d$, $s! = s_1! \ldots s_d!$ and $x^s = x_1^{s_1} \ldots x_d^{s_d}$. Let D^s denote the differential operator

$$D^s = \frac{\partial^{s_1+\cdots+s_d}}{\partial x_1^{s_1}\ldots\partial x_d^{s_d}}\,.$$

Denote by $\lfloor \beta \rfloor$ the maximal integer that is strictly smaller than β and fix $x_0 \in \mathcal{X}$. For any real valued function g on $\mathcal{X}$ that is $\lfloor \beta \rfloor$-times continuously differentiable at point x_0, we denote by $g_{x_0}^{(\beta)}$ its Taylor polynomial of degree $\lfloor \beta \rfloor$ at point x_0:

$$g_{x_0}^{(\beta)}(x) = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(x - x_0)^s}{s!} D^s g(x_0) \,.$$

Let $L > 0$ and denote by $\Sigma(\beta, L, x_0)$ the set of functions $g : \mathcal{X} \to \mathbb{R}$ that are $\lfloor \beta \rfloor$-times continuously differentiable at point x_0 and satisfy

$$|g(x) - g_{x_0}^{(\beta)}(x)| \leq L \|x - x_0\|^\beta, \quad \forall\, x \in \mathcal{B}(x_0, r) \,,$$

for some $r > 0$. The set $\Sigma(\beta, L, x_0)$ is called (β, L, x_0)-*locally Hölder class* of functions.

We now define the class of densities that are considered in this paper.

Definition 4.1 *Fix $\beta > 0, L > 0, \lambda > 0$ and $\gamma \geq 0$. Recall the $\mathcal{D}(\eta)$ is the neighborhood defined by*

$$\mathcal{D}(\eta) = \big\{ p \in \ (\lambda - \eta, \lambda + \eta) \ \big\}, \quad \eta > 0$$

Let $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ denote the class of all probability densities p on $\mathcal{X}$ for which there exists $\eta > 0$ such that

(i) $p \in \Sigma(\beta, L, x_0)$ for all $x_0 \in \mathcal{D}(\eta)$, apart from a set of null Lebesgue measure Leb_d.

(ii) $\exists\, \beta' > 0$ such that $p \in \Sigma(\beta', L, x_0)$, for all $x_0 \notin \mathcal{D}(\eta)$, apart from a set of null measure Leb_d.

(iii) p has γ-exponent at level λ with respect to the Lebesgue measure.

(iv) p is uniformly bounded by a constant L^*.

The class $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ is the class of uniformly bounded (condition (iv)) densities that have γ-exponent at level λ with respect to Leb_d (condition (iii)) and that are smooth in the neighborhood of the level under consideration (condition (i)). Note that the parameters β' in condition (ii) and L^* in condition (iv) do not appear in the notation of the class. Indeed $\beta' > 0$ can be arbitrary close to 0 and this will not affect the rates of convergence. Actually, the role of condition (ii) is to ensure that any density from

the class can be consistently estimated at any point with an arbitrary slow polynomial rate. In the same manner, the constant L^* does not appear in the rates of convergence and only affects the constants. Conditions (i) and (ii) are smoothness conditions that ensure consistency of the nonparametric density estimator used in the plug-in estimator. The class of densities $\mathcal{P}_\Sigma = \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ is similar to the class of regression functions considered in Audibert and Tsybakov (2007). However, besides the additional assumption that functions in $\mathcal{P}_\Sigma$ are probability densities, the main improvement here is that the regularity of a density in $\mathcal{P}_\Sigma$ can be arbitrary low outside of a neighborhood of the level under consideration, yielding slower rates of pointwise estimation. We prove below (cf. Corollary 4.1) that fast rates of convergence for DLSE are possible for this larger class of densities, which corroborates the idea that the density need not be precisely estimated far from the level λ.

Intuitively, parameters γ and β are conflicting. Indeed, the parameter β ensures that the density p has a relatively small slope around level λ and the parameter γ requires p to have a slope that is not too small around level λ. The constraints on these parameters depend on whether the density p *crosses* level λ or simply *hits* it. We now recall the definition for these two terms that was introduced by Audibert and Tsybakov (2005).

Definition 4.2 *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to hit the level $\lambda > 0$ at $\mathcal{X}_0 \in \mathbb{R}^d$ if and only if $f(x_0) = \lambda$ and for any $\delta > 0$, there exists $x \in \mathcal{B}(x_0, \delta)$ such that $f(x) \neq \lambda$. Moreover, the function f is said to cross the level $\lambda > 0$ at $\mathcal{X}_0 \in \mathbb{R}^d$ if and only if $f(x_0) = \lambda$ and for any $\delta > 0$, there exists two points $x_+, x_- \in \mathcal{B}(x_0, \delta)$ such that $f(x_-) < \lambda$ and $f(x_+) > \lambda$.*

A straightforward consequence of Proposition 3.4 of Audibert and Tsybakov (2005) is the following proposition.

Proposition 4.1 *If $\gamma(\beta \wedge 1) > d$, there is no density $p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ that hits the level λ in the interior of $\mathcal{X}$. Conversely, if $\gamma(\beta \wedge 1) \leq d$, there exist densities in $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ that hit the level λ in the interior of $\mathcal{X}$.*

If $\gamma(\beta \wedge 1) > 1$, there is no density $p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ that crosses the level λ in the interior of $\mathcal{X}$. Conversely, if $\gamma(\beta \wedge 1) \leq 1$, there exist densities in $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$ that cross the level λ in the interior of $\mathcal{X}$.

4.2 Exponential inequalities for kernel density estimators

To estimate a density p from the class $\mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$, we can use a *kernel density estimator* defined by:

$$\hat{p}_n(x) = \hat{p}_{n,h}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{4.11}$$

where $h > 0$ is the bandwidth parameter and $K : \mathcal{X} \to \mathbb{R}$ is a kernel. This choice is not the only possible one and all we need is an estimator that satisfies exponential inequalities as in (3.1) and (3.2). The following lemma states that it is possible to derive such exponential inequalities for a kernel density estimator with a $\beta^\star$-valid kernel where $\beta^\star \geq \beta$. The definition of β-valid kernel is recalled in the appendix, Definition 6.1 (see also Tsybakov, 2004b, for example).

Lemma 4.1 *Let P be a distribution on $\mathbb{R}^d$ having a density p w.r.t. the Lebesgue measure and such that $\|p\|_\infty \leq L^*$ for some constant $L^* > 0$. Fix $\beta > 0$, $\beta^\star \geq \beta$, $L > 0$ and assume that $p \in \Sigma(\beta, L, x_0)$, where the neighborhood around x_0 is a ball of radius $r > 0$. Let $\hat{p}_n$ be a kernel density estimator with bandwidth $h > 0$ and $\beta^\star$-valid kernel K, given an i.i.d. sample $X_1, \ldots, X_n$ from P. Set*

$$\Delta = \frac{6L^*\|K\|^2}{\|K\|_\infty + L^* + L\int \|t\|^\beta K(t)\mathrm{d}t}.$$

Then, for all $\delta, h \leq r$ such that $\Delta > \delta > 2Lc_7h^\beta > 0$, we have,

$$\mathbb{P}\{|\hat{p}_n(x_0) - p(x_0)| \geq \delta\} \leq 2\exp\left(-c_8 nh^d\delta^2\right),$$

where $c_7 = \int \|t\|^\beta K(t)\mathrm{d}t$ and $c_8 = 1/(16L^*\|K\|^2)$.

Proof. For any $x_0 \in \mathbb{R}^d$,

$$|\hat{p}_n(x_0) - p(x_0)| = \frac{1}{n}\Big|\sum_{i=1}^{n} Z_i(x_0)\Big|,$$

with

$$Z_i(x) = \frac{1}{h^d} K\left(\frac{X_i - x}{h}\right) - p(x).$$

The expectation of $Z_i(x_0)$ is the pointwise bias of a kernel density estimator with bandwidth h. Under the assumptions of the theorem, it is controlled in the following way

$$|\mathbb{E}Z_i(x_0)| \leq Lc_7 h^\beta \,.$$

Indeed,

$$\begin{aligned}
&|\mathbb{E}Z_i(x_0)| = \\
&= \Big| \int \frac{1}{h^d} K\Big(\frac{t}{h}\Big) \big[p(x_0+t) - p(x_0)\big] \mathrm{d}t \Big| \\
&= \Big| \int K(t) \big[p(x_0+ht) - p(x_0)\big] \mathrm{d}t \Big| \\
&= \Big| \int K(t) \big[p(x_0+ht) - p_{x_0}^{(\beta)}(x_0+ht)\big] \mathrm{d}t \\
&\qquad\qquad + \int K(t) \big[p_{x_0}^{(\beta)}(x_0+ht) - p(x_0)\big] \mathrm{d}t \Big| \,.
\end{aligned} \tag{4.12}$$

To control the first term in the right hand side of (4.12), remark that since K has support $[-1,1]^d$, for any $h < r/\sqrt{d}$, we have $x_0 + ht \in \mathcal{B}(x_0, r)$ for any $t \in [-1,1]^d$. Thus, using the fact that p is in $\Sigma(\beta, L, x_0)$ we have

$$\Big| \int K(t) \big[p(x_0+ht) - p_{x_0}^{(\beta)}(x_0+ht)\big] \mathrm{d}t \Big| \leq L \int |K(t)| \|ht\|^\beta \mathrm{d}t \,.$$

Now, since K is a $\lfloor\beta\rfloor$-valid kernel (cf. Proposition 6.2) and $p_{x_0}^{(\beta)} - p(x_0)$ is a polynomial of degree at most $\lfloor\beta\rfloor$ with no constant term, the second term in the right hand side of (4.12) is zero. Therefore, it holds

$$|\mathbb{E}Z_i(x_0)| \leq Lh^\beta \int |K(t)| \|t\|^\beta \mathrm{d}t \,, \quad \text{for any } h \leq r \,.$$

Now denote for simplicity $Z_i = Z_i(x_0)$ and let $\overline{Z_i}$ be the centered version of Z_i. Then, when $Lc_7 h^\beta \leq \delta/2$,

$$\begin{aligned}
\mathbb{P}\{|\hat{p}_n(x_0) - p(x_0)| \geq \delta\} &\leq \mathbb{P}\Big\{ \frac{1}{n} \Big| \sum_{i=1}^n \overline{Z_i} \Big| \geq \delta - Lc_7 h^\beta \Big\} \\
&\leq \mathbb{P}\Big\{ \frac{1}{n} \Big| \sum_{i=1}^n \overline{Z_i} \Big| \geq \frac{\delta}{2} \Big\} \,.
\end{aligned}$$

The right-hand side of the last inequality can be bounded applying Bernstein's inequality to $\overline{Z_i}$ and $-\overline{Z_i}$ successively. For $h \leq 1$, one has

$$\big|\overline{Z_i}\big| \leq \|K\|_\infty h^{-d} + L^* + Lc_7 h^\beta \leq c_9 h^{-d} \,,$$

where $c_9 = \|K\|_\infty + L^* + Lc_7$ and

$$\mathrm{Var}\{Z_i\} \le h^{-d} \int K(u)^2 p(hu) \mathrm{d}u \le c_{10} h^{-d} \,,$$

where $c_{10} = L^* \|K\|^2$. Applying now Bernstein's inequality yields

$$\begin{aligned} \mathbb{P}\{|\hat{p}_n(x_0) - p(x_0)| \ge \delta\} &\le 2 \exp\left(-\frac{n(\delta/2)^2}{2\left(c_{10} h^{-d} + c_9 h^{-d} \delta/6\right)} \right) \\ &\le 2 \exp\left(-c_8 n h^d \delta^2 \right) \,, \end{aligned}$$

for any $\delta \le \Delta$ and where $\Delta = 6c_{10}/c_9$ and $c_8 = 1/(16c_{10})$. ∎

We can therefore apply Theorem 3.1. When the choice of h is optimal, i.e., $h = n^{-1/(2\beta+d)}$, it yields the following corollary.

Corollary 4.1 *Let the underlying measure Q be the Lebesgue measure on $\mathbb{R}^d$. Fix $\beta > 0$, $L > 0$, $\lambda > 0$, $\gamma > 0$ and consider the plug-in estimator $\tilde{\Gamma}$ with offset $\ell_n = n^{-\nu}$ for some $0 < \nu < \beta/(2\beta + d)$. The nonparametric estimator $\hat{p}_n$ is the kernel density estimator defined in* (4.11) *with bandwidth parameter $h = n^{-1/(2\beta+d)}$ and $\beta^\star$-valid kernel K, where $\beta^\star = \beta \vee \beta'$ and β' is the parameter from Definition 4.1. Then,*

$$\begin{aligned} \sup_{p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)} \mathbb{E}\left[d_H(\Gamma_p(\lambda), \tilde{\Gamma}) \right] &\le c_{11} n^{-\frac{(1+\gamma)\beta}{2\beta+d}} \,, \\ \sup_{p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda), \tilde{\Gamma}) \right] &\le c_{12} n^{-\frac{\gamma\beta}{2\beta+d}} \,, \end{aligned}$$

where $c_{11} > 0$ and $c_{12} > 0$ depend on the constants c_7 and c_8 that appear in Lemma 4.1, on $c_0, \beta, \beta', \gamma, d$ and on λ.

Proof. The results are direct consequences of Theorem 3.1 when $\hat{p}_n$ is chosen as in (4.11). We need to check that for such an estimator we have $\mathrm{Leb}_d(\hat{p}_n \ge \lambda) \le M$, almost surely for some $M > 0$. Note that since $K \in L_1(\mathbb{R}^d)$, we have

$$\infty > \int_{\mathbb{R}^d} |K(x)| \mathrm{dLeb}_d(x) \ge \int_{\{\hat{p}_n \ge \lambda\}} |\hat{p}_n(x)| \mathrm{dLeb}_d(x) \ge \lambda \mathrm{Leb}_d\{\hat{p}_n \ge \lambda\} \,.$$

Hence, the condition is satisfied with $M = \lambda^{-1} \int |K|$. All the other conditions of Theorem 3.1 are satisfied and we can apply it with $a = 2\beta/(2\beta + d)$ and $b = 2\beta'/(2\beta + d)$. ∎

5 Minimax lower bounds

The following theorem shows that the rates obtained in Corollary 4.1 are optimal in a minimax sense.

Theorem 5.1 *Let the underlying measure Q be the Lebesgue measure on $\mathbb{R}^d$. Fix $\lambda > 0$ and let L, β, γ be positive constants such that $\gamma\beta \leq d$. Then, for any $n \geq 1$ and any estimator $\hat{G}_n$ of $\Gamma_p(\lambda)$ constructed from the sample $X_1, \ldots, X_n$, we have*

$$\begin{aligned} \sup_{p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)} \mathbb{E}\left[d_H(\Gamma_p(\lambda), \hat{G}_n)\right] \geq & C n^{-\frac{(1+\gamma)\beta}{2\beta+d}}, \\ \sup_{p \in \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)} \mathbb{E}\left[d_\triangle(\Gamma_p(\lambda), \hat{G}_n)\right] \geq & C n^{-\frac{\gamma\beta}{2\beta+d}}. \end{aligned} \tag{5.1}$$

Proof. In view of Proposition 2.1 and since we have $\mathrm{Leb}_d[\Gamma_p(\lambda)] \leq \lambda^{-1} < \infty$, we only have to prove (5.1). To that end, we will use Lemma 6.2 with $d = d_\triangle$, $\varepsilon = \varepsilon_n \geq C n^{-\frac{\gamma\beta}{2\beta+d}}$ and $\mathcal{P} = \mathcal{P}_\Sigma(\beta, L, \lambda, \gamma)$. Thus our goal is to find a family $\mathcal{N}$ of densities that are in $\mathcal{P}$ such that the densities in $\mathcal{N}$ are close to each other for the Kullback-Leibler divergence and yield density level sets that are far for the symmetric distance. We now describe the construction of the family $\mathcal{N}$. It is similar to the construction used in Audibert and Tsybakov (2007), Section 6.2 and for the rest of this section any reference to this article will be about Section 6.2. However, the proof for densities involves different technical details. Indeed, here, the lower bound has to be on the Lebesgue measure of the symmetric difference -which is actually easier to handle than the excess-risk in classification- and most importantly, when construction less favorable distributions in the case of classification, one has two degrees of freedom: the marginal distribution and the regression function. Here we have only one degree of freedom: the density. The modification are substantial enough that we produce the entire proof, using results from Audibert and Tsybakov (2007) when possible.

We can assume without loss of generality that $\lambda = 1$. Consider the integer $q = \lfloor c_{13} n^{\frac{1}{2\beta+d}} \rfloor$ where c_{13} is a positive constant chosen large enough to ensure that $q \geq 1$, and the regular grid $\mathcal{G}$ on $[0,1]^d$ defined as

$$\mathcal{G} = \left\{ \left(\frac{2k_1+1}{2q}, \ldots, \frac{2k_d+1}{2q} \right), k_i \in \{0, \ldots, q-1\}, i = 1, \ldots, d \right\}.$$

Denote by $\{g_j\}_{1 \leq j \leq q^d}$ the elements of the grid, the choice of indexing being of no importance for what follows. Define the integer $m = \lfloor c_{14} q^{d-\gamma\beta} \rfloor$ for

some positive constant c_{14} and note that the condition $\gamma\beta \leq d$ ensures that $m \geq 2$ if c_{14} is chosen large enough. Let $\mathcal{J} = \{1, 3, \ldots, 2m-1\}$ be the set of odd integers between 1 and $2m-1$ and for any $j = 1, \ldots, 2m$, define the disjoint balls $B_j = \mathcal{B}(g_j, (4q)^{-1})$. Set $B_0 = [0,1]^d \setminus \bigcup_{j=1}^{2m} B_j$.

Let $\phi : \mathbb{R}^d \to \mathbb{R}_+$ and $\varphi_j : \mathbb{R}^d \to [0,1]$ be the functions defined in Audibert and Tsybakov (2007, Section 6.2) such that $\phi \in \Sigma(\beta, L, x_0)$ for any $x_0 \in \mathbb{R}^d$ and $\varphi_j(x) = q^{-\beta}\phi(q[x - g_j])\mathbb{1}_{\{x \in B_j\}}$.

Then, for any $\omega = (\omega_1, \ldots, \omega_m) \in \{0,1\}^m$, define the function on $[0,1]^d$

$$p_\omega(x) = 1 + \sum_{j \in \mathcal{J}} \omega_j [\varphi_j(x) - \varphi_{j+1}(x)] .$$

Consider a subset $\Omega \subset \{0,1\}^m$ of cardinality s and define the family $\mathcal{N}$ as

$$\mathcal{N} = \{p_\omega, \omega \in \Omega\} .$$

The set Ω, will be chosen in order to fulfill the conditions of Lemma 6.2.

First condition: $\mathcal{N} \subset \mathcal{P}_\Sigma(\beta, L, 1, \gamma)$.

First, as noted by Audibert and Tsybakov (2007), the function ϕ can always be adjusted so that $\|\varphi_j\|_\infty \leq 1$ for any j so that for any $\omega \in \Omega$, p_ω is a density that satisfies $\|p_\omega\|_\infty \leq 2$ and $p_\omega \in \Sigma(\beta, L, x_0)$ from the results of Audibert and Tsybakov (2007).

Therefore it remains to check that p_ω has γ-exponent at level 1 with respect to the Lebesgue measure. The following decomposition holds:

$$\begin{aligned}
&\mathrm{Leb}_d(x : 0 < |p_\omega(x) - 1| \leq \varepsilon) \\
&\quad = 2 \sum_{j \in \mathcal{J}} \mathrm{Leb}_d(x : 0 < |p_\omega(x) - 1| \leq \varepsilon, x \in B_j) \\
&\quad = 2m \int_{B_1} \mathbb{1}_{\{0 < \phi(q[x-g_1]) \leq \varepsilon q^\beta\}} \mathrm{d}x \qquad (5.2) \\
&\quad = 2mq^{-d} \int_{\mathcal{B}(0,1/4)} \mathbb{1}_{\{0 < \phi(x) \leq \varepsilon q^\beta\}} \mathrm{d}x \\
&\quad = 2mq^{-d} \mathbb{1}_{\{1 \leq \varepsilon q^\beta\}} \mathrm{d}x \leq 2c_{14}\varepsilon^\gamma ,
\end{aligned}$$

where the last but one inequality uses the fact that $\phi(x) = C_\phi \leq 1$ for any $x \in \mathcal{B}(0, 1/4)$ (see construction of ϕ in Audibert and Tsybakov, 2007).

Second condition (6.3): $d_\triangle(\Gamma_p, \Gamma_q) \geq \varepsilon_n, \forall\ p, q \in \mathcal{N}, p \neq q$.

By construction, for any $\omega, \omega' \in \{0,1\}^m$,

$$d_\triangle(\Gamma_{p_\omega}, \Gamma_{p_{\omega'}}) = 2\mathrm{Leb}_d(B_1) \sum_{j=1}^m \mathbb{1}_{\{\omega_j \neq \omega'_j\}} \geq Cq^{-d} \sum_{j=1}^m \mathbb{1}_{\{\omega_j \neq \omega'_j\}} .$$

We need to bound from below the Hamming distance between ω and ω', defined for any $\omega, \omega' \in \Omega$ by

$$\rho(\omega, \omega') = \sum_{j=1}^{m} \mathbb{1}_{\{\omega_j \neq \omega'_j\}} .$$

To do so we use the Varshamov-Gilbert bound (cf. Lemma 6.1) that guarantees the existence of Ω such that $\mathrm{card}(\Omega) \geq 2^{m/8}$ and $\rho(\omega, \omega') \geq m/8$ for any $\omega, \omega' \in \Omega$. We also choose Ω such that $\Omega \ni \omega_0 = (0, \dots, 0)$. For such Ω we have

$$d_{\triangle}(\Gamma_{p_\omega}, \Gamma_{p_{\omega'}}) \geq Cmq^{-d} \geq Cn^{-\frac{\gamma\beta}{2\beta+d}} .$$

THIRD CONDITION: $\max_{\omega \in \Omega} K(p_\omega, p_{\omega_0}) \leq C \log(s)$.

Note that for the above choice of Ω, we have $s = \mathrm{card}(\mathcal{N}) = \mathrm{card}(\Omega) \geq 2^{m/8}$. Therefore $\log(s) \geq Cm$ and we only have to prove that

$$\max_{\omega \in \Omega} K(p_\omega, p_{\omega_0}) \leq Cm .$$

Denote by $\xi_j(x) = \varphi_j(x) - \varphi_{j+1}(x)$. For any $p_\omega \in \mathcal{N}$, we have,

$$\begin{aligned} K(p_\omega, p_{\omega_0}) &= n \sum_{j \in \mathcal{J}} \int_{B_j \cup B_{j+1}} \log\left(1 + \omega_j \xi_j(x)\right) \big(1 + \omega_j \xi_j(x)\big) \mathrm{d}x \\ &\leq 2n \sum_{j \in \mathcal{J}} \omega_j \int_{B_j} \varphi_j^2(x) \mathrm{d}x , \\ &\leq 2nmq^{-(2\beta+d)} \int_{\mathcal{B}(0,1/4)} \phi^2(x) \mathrm{d}x \\ &\leq Cm . \end{aligned}$$

We can therefore apply Lemma 6.2 and Theorem 5.1 is proved. ∎

6 Appendix

Several results that can be omitted in a first reading are gathered in this appendix.

6.1 Equivalent formulation for the γ-exponent condition

The following proposition gives an equivalent formulation for the γ-exponent condition.

Proposition 6.1 *Fix $\lambda > 0, \gamma > 0$ and $L_Q > 0$.*

Define $\mathcal{L} = \mathcal{L}(\lambda) = \{p = \lambda\}$. The two following statements are equivalent.

(i) $\exists\, c > 0$ and $\varepsilon_0 > 0$, such that for any $0 < \varepsilon \leq \varepsilon_0$, we have

$$Q\left\{x \in \mathcal{X} : 0 < |p(x) - \lambda| \leq \varepsilon\right\} \leq c\varepsilon^{\gamma}\,.$$

(ii) $\exists\, c' > 0$ and $\varepsilon_1 > 0$, such that for any $0 < \varepsilon \leq \varepsilon_1$, we have

$$Q\left\{x \in \mathcal{X} : 0 < |p(x) - \lambda| \leq \varepsilon\right\} \leq L_Q$$

and for all $G \subseteq \mathcal{X} \setminus \mathcal{L}$ satisfying $Q(G) \leq L_Q$, we have

$$Q(G) \leq c' \left(\int_G |p(x) - \lambda| \mathrm{d}Q(x)\right)^{\frac{\gamma}{1+\gamma}}\,. \tag{6.1}$$

Proof. The proof of (i) $\Rightarrow$ (ii) essentially follows that of Tsybakov (2004b, Proposition 1). Define

$$\varepsilon_1 = \varepsilon_0 \wedge \Big(\frac{L_Q}{c(1+\gamma)}\Big)^{1/\gamma}\,.$$

Observe that for any $0 < \varepsilon \leq \varepsilon_1$, we have

$$Q\left\{x \in \mathcal{X} : 0 < |p(x) - \lambda| \leq \varepsilon\right\} \leq c\varepsilon^{\gamma} \leq c\varepsilon_1^{\gamma} = \frac{L_Q}{1+\gamma} \leq L_Q\,.$$

Define $\mathcal{A}_\varepsilon = \{x : |p(x) - \lambda| > \varepsilon\}$, for all $0 < \varepsilon \leq \varepsilon_0$. For any measurable set $G \subset \mathcal{X} \setminus \mathcal{L}$, we have

$$\begin{aligned}\int_G |p(x) - \lambda| \mathrm{d}Q(x) &\geq \varepsilon Q(G \cap \mathcal{A}_\varepsilon) \\ &\geq \varepsilon\left[Q(G) - Q(\mathcal{A}_\varepsilon^c \cap \mathcal{L}^c)\right] \\ &\geq \varepsilon\left[Q(G) - \underline{c}\varepsilon^{\gamma}\right]\,, \quad \forall\, \underline{c} > c\,,\end{aligned}$$

where the last inequality is obtained using (i). Maximizing the last term w.r.t $\varepsilon > 0$, we get

$$\Big(\int_G |p(x) - \lambda| \mathrm{d}Q(x)\Big)^{\frac{\gamma}{1+\gamma}} \geq Q(G)\Big(\frac{\gamma}{1+\gamma}\Big)^{\frac{\gamma}{1+\gamma}}\Big(\frac{1}{1+\gamma}\Big)^{\frac{1}{1+\gamma}}\underline{c}^{-1/(1+\gamma)}\,.$$

This yields (6.1) with $c' = \mathrm{e}^{-2/\mathrm{e}}\underline{\mathrm{c}}^{1/(1+\gamma)}$. Note that the maximum is obtained for $\varepsilon = \left(\frac{Q(G)}{\underline{\mathrm{c}}(1+\gamma)}\right)^{1/\gamma} \leq \varepsilon_0$ for sufficiently large $\underline{\mathrm{c}}$ and (i) is valid for this particular ε.

We now prove that (ii) $\Rightarrow$ (i). Consider $\varepsilon_1 > 0$ such that $Q(\mathcal{A}_\varepsilon^c \cap \mathcal{L}^c) \leq L_Q$ for any $0 < \varepsilon \leq \varepsilon_1$ and $c' > 0$ such that (6.1) is satisfied for any $G \subseteq \mathcal{X} \setminus \mathcal{L}$, $Q(G) \leq L_Q$. Taking $G = \mathcal{A}_\varepsilon^c \cap \mathcal{L}^c$ in (6.1) yields

$$\begin{aligned} Q\left\{x : 0 < |p(x) - \lambda| \leq \varepsilon\right\} &= Q\left(\mathcal{A}_\varepsilon^c \cap \mathcal{L}^c\right) \\ &\leq c' \left(\int_{\mathcal{A}_\varepsilon^c \cap \mathcal{L}^c} |p(x) - \lambda| \mathrm{d}Q(x)\right)^{\frac{\gamma}{1+\gamma}} \\ &\leq c' \left(\varepsilon Q\left(\mathcal{A}_\varepsilon^c \cap \mathcal{L}^c\right)\right)^{\frac{\gamma}{1+\gamma}} . \end{aligned}$$

Therefore,

$$Q\left\{x : 0 < |p(x) - \lambda| \leq \varepsilon\right\} \leq (c')^{1+\gamma}\varepsilon^\gamma .$$

This inequality yields (i) with $\varepsilon_0 = \varepsilon_1$ and $c = (c')^{1+\gamma}$. ∎

6.2 On β-valid kernels

We recall here the definition of β-valid kernels and give a property that is useful in the present study.

Definition 6.1 *Let K be a real-valued function on $\mathbb{R}^d$, with support $[-1,1]^d$. For fixed $\beta > 0$, the function $K(\cdot)$ is said to be a β-valid kernel if it satisfies $\int K = 1$, $\int |K|^p < \infty$ for any $p \geq 1$, $\int \|t\|^\beta |K(t)| \,\mathrm{d}t < \infty$, and, in case $\lfloor\beta\rfloor \geq 1$, it satisfies $\int t^s K(t)\mathrm{d}t = 0$ for any $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ such that $1 \leq s_1 + \ldots + s_d \leq \lfloor\beta\rfloor$.*

Example 6.1 *Let $\beta > 0$. For any β-valid kernel K defined on $\mathbb{R}^d$, consider the following product kernel*

$$\tilde{K}(x) = K(x_1)K(x_2)\ldots K(x_d)\mathbb{1}_{x\in[-1,1]^d} ,$$

for any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Then it can be easily shown that $\tilde{K}$ is a β-valid kernel on $\mathbb{R}^d$. Now, for any $\beta > 0$, an example of a 1-dimensional β-valid kernel is given in (Tsybakov, 2004a, section 1.2.2), the construction of which is based on Legendre polynomials. This eventually proves the existence of a multivariate β-valid kernel, for any given $\beta > 0$.

The following proposition holds.

Proposition 6.2 *Fix $\beta > 0$. If K is a β-valid kernel, then K is also a β'-valid kernel for any $0 < \beta' \leq \beta$.*

PROOF. Fix β and β' such that $0 < \beta' \leq \beta$. Observe that $\lfloor \beta' \rfloor \leq \lfloor \beta \rfloor$ yields that if $\lfloor \beta' \rfloor \geq 1$, for any β-valid kernel K, we have $\int t^s K(t) \mathrm{d}t = 0$ for any $s = (s_1, \ldots, s_d)$ such that $1 \leq s_1 + \ldots + s_d \leq \lfloor \beta' \rfloor$. It remains to check that

$$\int_{\mathbb{R}^d} \|t\|^{\beta'} |K(t)| \mathrm{d}t < \infty \,. \tag{6.2}$$

Consider the decomposition

$$\begin{aligned} \int_{\mathbb{R}^d} \|t\|^{\beta'} |K(t)| \mathrm{d}t &= \int_{\|t\| \leq 1} \|t\|^{\beta'} |K(t)| \mathrm{d}t + \int_{\|t\| \geq 1} \|t\|^{\beta'} |K(t)| \mathrm{d}t \\ &\leq \int_{\mathbb{R}^d} |K(t)| \mathrm{d}t + \int_{\|t\| \geq 1} \|t\|^{\beta} |K(t)| \mathrm{d}t \,. \end{aligned}$$

To prove (6.2), remark that since K is a β-valid kernel, we have $\int_{\mathbb{R}^d} |K(t)| \mathrm{d}t < \infty$ and

$$\int_{\|t\| \geq 1} \|t\|^{\beta} |K(t)| \mathrm{d}t \leq \int_{\mathbb{R}^d} \|t\|^{\beta} |K(t)| \mathrm{d}t < \infty \,.$$

∎

6.3 Technical lemmas for minimax lower bounds

We gather here technical results that are used in Section 5. For a recent survey on the construction of minimax lower bounds, see Tsybakov (2004a)[Chap. 2]. We first give a lemma related to subset extraction.

Fix an integer $m \geq 1$, and for any two $\omega = (\omega_1, \ldots, \omega_m)$ and $\omega' = (\omega'_1, \ldots, \omega'_m)$ in $\{0,1\}^m$ define the *Hamming distance* between ω and ω' by

$$\rho(\omega, \omega') = \sum_{i=1}^{m} \mathbb{1}_{\{\omega_i \neq \omega'_i\}} \,.$$

The following lemma holds.

Lemma 6.1 (Varshamov-Gilbert bound, 1962) *Fix $m \geq 8$. Then there exists a subset $\Omega = \{\omega^{(0)}, \ldots, \omega^{(M)}\}$ of $\{0,1\}$ such that $M \geq 2^{m/8}$ and*

$$\rho(\omega^{(j)}, \omega^{(k)}) \geq \frac{m}{8}, \quad \forall\ 0 \leq j < k \leq M \,.$$

Moreover, we can always take $\omega^{(0)} = (0, \ldots, 0)$.

For a proof of this lemma, see Tsybakov (2004a, Lemma 2.8, p. 89).

The next lemma can be found in Tsybakov (2004a, Theorem 2.5, p. 85) and is stated here in a form adapted to our purposes. It allows to derive minimax lower bounds in the context of DLSE. It involves the Kullback-Leibler divergence between two probability densities p and q on $\mathbb{R}^d$

$$K(p,q) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} \log\Big(\frac{p(x)}{q(x)}\Big) p(x)\mathrm{d}x & \text{if } P_p \ll P_q\,, \\ +\infty & \text{else.} \end{cases}$$

Lemma 6.2 *Let d be a pseudo-metric between subsets of $\mathcal{X} \subset \mathbb{R}^d$. Let $\mathcal{P}$ be a set of densities and assume that there exists a finite subset $\mathcal{N} \subset \mathcal{P}$ with $2 \leq \mathrm{card}(\mathcal{N}) = s < \infty$ and a constant $C > 0$, such that*

$$d\big(\Gamma_p(\lambda), \Gamma_q(\lambda)\big) \geq 2\varepsilon, \quad \forall\, p, q \in \mathcal{N}, p \neq q\,, \tag{6.3}$$

and there exists $p \in \mathcal{N}$ such that

$$\max_{q \in \mathcal{N}} K(q,p) \leq C \log(s)\,. \tag{6.4}$$

Then, there exists an absolute positive constant C' such that for any estimator $\hat{G}_n$ of $\Gamma_p(\lambda)$ constructed from the sample $X_1, \ldots, X_n$, we have

$$\sup_{p \in \mathcal{P}} \mathbb{E}\left[d(\Gamma_p(\lambda), \hat{G}_n)\right] \geq C'\varepsilon\,.$$

References

Audibert, J.-Y. and Tsybakov, A. (2005). Fast learning rates for plug-in classifiers under the margin condition. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires de Paris 6. Available at http://arxiv.org/abs/math/0507180.

Audibert, J.-Y. and Tsybakov, A. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.*, **35** 608–633.

Baíllo, A. (2003). Total error in a plug-in estimator of level sets. *Statist. Probab. Lett.*, **65** 411–417.

Baíllo, A., Cuesta-Albertos, J. A. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statist. Probab. Lett.*, **53** 27–35.

Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *Ann. Statist.*, **25** 2300–2312.

Cuevas, A., González-Manteiga, W. and Rodríguez-Casal, A. (2006). Plug-in estimation of general level sets. *Aust. N. Z. J. Stat.*, **48** 7–19.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, vol. 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, **38** 480–488.

Gayraud, G. and Rousseau, J. (2005). Rates of convergence for a Bayesian level set estimation. *Scand. J. Statist.*, **32** 639–660.

Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, **82** 267–270.

Hartigan, J. H. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.*, **27** 1808–1829.

Molchanov, I. S. (1998). A limit theorem for solutions of inequalities. *Scand. J. Statist.*, **25** 235–242.

Müller, D. W. and Sawitzki, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Tech. Rep. 398, SFB 123, Univ. Heidelberg.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.*, **23** 855–881.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, **69** 1–24.

Rigollet, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.*, **8** 1369–1392.

Scott, C. D. and Nowak, R. D. (2006). Learning minimum volume sets. *J. Mach. Learn. Res.*, **7** 665–704.

Steinwart, I., Hush, D. and Scovel, C. (2005). A classification framework for anomaly detection. *J. Mach. Learn. Res.*, **6** 211–232.

Stuetzle, W. (2003). Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, **20** 25–47.

Tarigan, B. and van de Geer, S. (2006). Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli*, **12**.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.*, **25** 948–969.

Tsybakov, A. B. (2004a). *Introduction à l'estimation non-paramétrique*, vol. 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin.

Tsybakov, A. B. (2004b). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32** 135–166.

Tsybakov, A. B. and van de Geer, S. A. (2005). Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, **33** 1203–1224.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New-York.

Yang, Y. (1999). Minimax nonparametric classification — part I: rates of convergence. *IEEE Trans. Inform. Theory*, **45** 2271–2284.

Philippe Rigollet
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30302-0160, USA
rigollet@math.gatech.edu

Régis Vert
Masagroup,
24 Bd de l'Hôpital,
75005 Paris, France
regis.vert@masagroup.net