A MODEL-BASED APPROACH TO DENSITY ESTIMATION IN SUP-NORM

Guillaume Maillard

A MODEL-BASED APPROACH TO DENSITY ESTIMATION IN SUP-NORM

GUILLAUME MAILLARD

ABSTRACT. Building on the $\ell-$estimators of Baraud [3], we define a general method for finding a quasi-best approximant in sup-norm to a target density $p^\star$ belonging to a given model m, based on independent samples drawn from distributions $p_i^\star$ which average to $p^\star$ (which does not necessarily belong to m). We also provide a general method for selecting among a countable family of such models. Both of these estimators satisfy oracle inequalities in the general setting. The quality of the bounds depends on the volume of sets C on which $|f|$ is close to its maximum, where $f = p - q$ for some $p, q \in m$ (or $p \in m$ and $q \in m'$, in the case of model selection). In particular, using piecewise polynomials on dyadic partitions of $\mathbb{R}^d$, we recover optimal rates of convergence for classes of functions with anisotropic smoothness, with optimal dependence on semi-norms measuring the smoothness of $p^\star$ in the coordinate directions. Moreover, our method adapts to the anisotropic smoothness, as long as it is smaller than 1 plus the degree of the polynomials.

1. INTRODUCTION

In regression, classification and density estimation, the model-based approach to estimation [14] consists in specifying a collection of *models*, together with a standard method for performing estimation within each model and a *penalty* or *model selection criterion* for selecting among the models. In density estimation, this approach can for-example be based on maximum likelihood or least-squares for estimating within a model [14, Example 1] and cross-validation for selecting among models.

This leads to a number of desirable practical and theoretical properties. First, the approach is very flexible and general since usually, a wide variety of different model collections are compatible with the basic method. Moreover, the analysis of the risk of model-based estimators naturally subdivides into an "approximation-theoretic" part dealing with the approximation

Date: June 15, 2022.

2020 *Mathematics Subject Classification.* Primary 62F35, 62G35, 62G07; Secondary 62C20, 62G10.

Key words and phrases. Density estimation, parametric estimation, robust estimation, Wasserstein loss, total variation loss, L_p-loss, minimax theory, robust testing.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017.

properties of the model m, and a "statistical part" dealing with the difficulty of estimating within m, which can be solved separately [18]. Appropriate penalties can be derived from concentration inequalities for (weighted) empirical processes [13, Chapter 1]. The resulting model selection estimators optimize the tradeoff between the approximation error and the penalty [4, Section 3]. This naturally leads to minimax-adaptive estimators, provided the model collection is well chosen [4, Section 1.4].

In density estimation, the model-based approach has mainly been used together with least-squares and maximum likelihood methods which target the minimizer of the squared L^2 and Küllback-Leibler distance to the underlying density. However, if we are interested instead in some other distance, the least-squares or maximum likelihood estimates may be arbitrarily far from optimal, as remarked by Devroye and Lugosi in the case of the L^1 loss [9, Chapter 6]. One is then left with the task of devising a data-driven method to minimize the given distance d over a model m. The difficulty here is that empirical risk minimization cannot be used in general, for lack of a suitable contrast function. This problem was first solved by Devroye and Lugosi [9, Chapter 6] in the case of the L^1 loss and by Baraud et al. [1] in the case of the Hellinger distance. More recently, Baraud [3] devised a general strategy called $\ell-$estimation, which applies to all L^p losses for $p \in [1, +\infty)$ (among others), but *not* however to the L^∞ distance in general. He also did not address the problem of model selection.

In this article, we treat the case of the sup-norm loss, establishing a general method for model-based density estimation in L^∞. Our method results from the application of a variant of Baraud's $\ell-$estimation to a certain parametrized family of semi-norms approximating the essential supremum. The estimation error of this method depends mainly on the measure of sets on which elements of the model m remain close to their extremal value. In addition, we develop an entirely data-driven method for model selection, using penalties derived from concentration inequalities.

As an application, we consider the class of piecewise polynomial functions on dyadic partitions of $\mathbb{R}^d$ and show that the resulting model-based estimator is minimax-adaptive over classes of functions with anisotropic smoothness. Our result improves on what was previously known in the literature for non model-based estimators: not only does our estimator converge at the optimal rate (a property already established by Lepski [11] for his adaptive kernel method), it also depends optimally on the underlying density, up to a constant depending only on the dimension and the degree of the polynomials.

This article is structured as follows. First, the setting is introduced and main notation defined in section 2 . The model-based estimator is defined in section 3 and a general oracle inequality is established. This general result is applied to models of piecewise polynomials in section 3.1. A minimax lower bound establishes the optimality of our estimator up to logarithmic factors. Section 4.1 addresses the model selection problem in the general setting, resulting in an estimator which satisfies an oracle inequality. In section 4.2, we consider the case of piecewise polynomials on *regular dyadic partition*, where one must select among such partitions, and show that our assumptions hold in that case.

In section 5, the resulting estimator is shown to be minimax-adaptive on classes of functions with anisotropic smoothness. A matching minimax lower bound is established, based on a result of Lepski [11] .

2. Setting and notation

Let $(E, \mathcal{E}, \mu)$ be a measure space, with $\sigma-$finite measure μ. Let $\mathscr{L}_\infty(E, \mu)$ be the set of measurable functions f on $(E, \mathcal{E}, \mu)$ such that

$$\|f\|_{\infty,\mu} = \sup \{r \geq 0 : \mu (\{x \in E : f(x) \geq r\}) > 0\} < +\infty$$

and let $L_\infty(E, \mu)$ denote the associated set of equivalent classes for the relation of equality $\mu-$almost everywhere. The topic of this article is density estimation on $L^\infty(E, \mu)$ with respect to the norm $\|\cdot\|_{\infty,\mu}$.

We assume that the observations $X_1, \ldots, X_n$ are independent but not necessarily i.i.d, which allows to consider possible outliers. Let $P_1^\star, \ldots, P_n^\star$ denote their marginals. We assume that the marginals have densities $p_1^\star, \ldots, p_n^\star$ belonging to $L_\infty(E, \mu)$ - otherwise, estimation in L_∞ norm is impossible.

Throughout this article, $\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star$ denotes the distribution of the observation $\boldsymbol{X} = (X_1, \ldots, X_n)$, and $\mathbf{p}^\star = (p_1^\star, \ldots, p_n^\star)$ denotes the corresponding $n-$uplet of probability densities. Moreover, $P^\star$ denotes the mixture distribution, $P^\star = \frac{1}{n} \sum_{i=1}^n P_i^\star$, and $p^\star$ denotes the corresponding probability density.

In case the data is not truly i.i.d, the estimators considered in this article estimate $p^\star$: in particular, they are robust to small departures from the i.i.d assumption (in the L^∞ sense).

2.1. Notations.

Bold capitals $\mathbf{P}$ will be used to denote either the product measure $\mathbf{P} = \bigotimes_{i=1}^n P_i$ or the n-uplet $(P_1, \ldots, P_n)$, depending on the context. The notation $\mathbb{E}[g(\boldsymbol{X})]$ is to be interpreted under the assumption that $\mathbf{X} \sim \mathbf{P}^\star$, while $\mathbb{E}_S[f(X)]$ denotes the expectation of $f(X)$ when $X \sim S$. The same conventions apply to $\mathrm{Var}\big(g(\boldsymbol{X})\big)$ and $\mathrm{Var}_S\big(f(X)\big)$. The same letter will always be used to denote a measurable function q and the corresponding (signed) measure $Q = qd\mu$: lowercase letters refer to functions and uppercase letters, to measures.

In addition, we shall use the following standard notation. For $x \in \mathbb{R}$, $x_- = \max\{0, -x\}$; for $x \in \mathbb{R}^d$, $B(x, r)$ denotes the closed Euclidean ball centered at x with radius $r \geqslant 0$. For a positive integer d, $L_\infty(\mathbb{R}^d)$ means $L_\infty(E, \mu)$ when $E = \mathbb{R}^d$, $\mathcal{E}$ is the Borel σ-algebra and $\mu = \lambda$ is the Lebesgue measure on $\mathbb{R}^d$.

2.2. **Models and losses.** Denote by $\mathcal{M}$ a collection of models m, each of which is a subset of $\mathscr{P} = L_\infty(E,\mu) \cap L_1(E,\mu)$. For reasons of technical convenience, we do not impose that the models m consist of densities.

In the following, we will always assume that $\mathcal{M}$, the model collection, as well as the models $m \in \mathcal{M}$, are *at most countable* in order to avoid measurability issues. Let $\mathscr{M}$ denote the union of all the models: $\mathscr{M} = \cup_{m\in\mathcal{M}} m$. In particular, $\mathscr{M}$ is countable. Since most of the models used by statisticians are separable, this assumption is not restrictive in practice: one can always replace an uncountable, separable model $\overline{m}$ by a dense countable subset m, without changing the approximation error.

Given the observation $\boldsymbol{X}$ and a model m, we want to design an estimator $\widehat{p}_m = \widehat{p}_m(\mathbf{X})$ of $p^\star$ with values in m which is as close as possible to $p^\star$ in norm $\|\cdot\|_{\infty,\mu}$. Since $\widehat{p}_m \in m$ by definition, $\|p^\star - \widehat{p}_m\|_{\infty,\mu}$ is lower bounded by

$$\inf_{q\in m} \|q - p^\star\|_{\infty,\mu} = d_{\infty,\mu}(p^\star, m),$$

the *approximation error* of the model m in $L^\infty(E,\mu)$. The best that can be expected of $\widehat{p}_m$ is that $\|p^\star - \widehat{p}_m\|_{\infty,\mu}$ be close to $d_{\infty,\mu}(p^\star, m)$. This term cancels when $p^\star \in \overline{m}$, where

$$\overline{m} = \{p \in \mathscr{P} \mid \inf_{q\in m} \|p - q\|_{\infty,\mu} = 0\}, \tag{1}$$

which generalizes the case of $\mathbf{p}^\star = (p,\dots,p)$ for some $p \in m$ (the "true model" case).

3. Estimator on a single model

To achieve model-based estimation in the norm $\|\cdot\|_{\infty,\mu}$, we adapt the general method of $\ell-$estimation introduced by Baraud [3]. For a given norm $\|\cdot\|$ and $p, q \in m \subset B$ (where B is a function space), this method relies on finding suitable measurable functions $g_{p,q}$ such that

- $\int (p-q) g_{p,q} d\mu = \|p - q\|$
- For all $f \in B$, $\int f g_{p,q} d\mu \le \|f\|$
- $\frac{1}{n}\sum_{i=1}^n g_{p,q}(X_i)$ is close to its expectation (over $p, q \in m$).

In the case of the norm $\|\cdot\|_{\infty,\mu}$ and the space $B = L^\infty(E,\mu)$, the first two requirements cannot be simultaneously satisfied in general, so we shall instead seek a suitable approximation of $\|p - q\|_{\infty,\mu}$ by $\int (p-q) g_{p,q} d\mu$, for some $g_{p,q}$ such that $\int |g_{p,q}| d\mu \le 1$. To that end, fix a VC class of measurable sets $\mathcal{C}$, with VC-dimension V. For any $f \in L_1(E,\mu)$ and any $h > 0$, let

$$|f|_h = \sup_{C\in\mathcal{C}} \frac{1}{\mu(C) + h} \left| \int_C f d\mu \right|.$$

This semi-norm is a norm whenever the sets of $\mathcal{C}$ have finite measure and generate the Borel sigma-algebra: this will be the case with all the examples which we will consider. Fix some $\varepsilon \in (0,1)$ and for any $(p,q) \in \mathscr{P}^2$ and any $h > 0$, choose some set $C_h(p,q)$ such that

$$\frac{\left|\int_{C_h(p,q)}(p-q)d\mu\right|}{\mu(C_h(p,q))+h} \geq (1-\varepsilon)|p-q|_h. \tag{2}$$

Let then $\varepsilon_h(p,q) \in \{-1,1\}$ be the sign of $\int_{C_h(p,q)}(p-q)d\mu$ and define

$$t_{p,q}^{(h)} = \varepsilon_h(p,q)\frac{P\left(C_h(p,q)\right) - \mathbb{1}_{C_h(p,q)}}{\mu\left(C_h(p,q)\right)+h},$$

as well as the associated *T-test*

$$T^{(h)}(\mathbf{X},p,q) = \frac{1}{n}\sum_{i=1}^{n} t_{p,q}^{(h)}(X_i).$$

Note that this construction can be carried out uniformly over all $p,q \in \mathscr{P}$. Let now

$$T_m^{(h)}(\mathbf{X},p) = \sup_{q\in m} T^{(h)}(\mathbf{X},p,q).$$

An $\ell-$estimator associated with the class $\mathcal{C}$, the model m, the parameter $h > 0$ and the tolerances ε,δ is, by definition, a random element $\widehat{p}_m^{(h)}$ such that

$$T_m^{(h)}\left(\mathbf{X},\widehat{p}_m^{(h)}\right) \leq \inf_{p\in m}\left\{T_m^{(h)}(\mathbf{X},p)\right\} + \delta.$$

Since m is countable and $\delta > 0$, such random elements exist. Note that $T_m^{(h)}$ and $\widehat{p}_m^{(h)}$ only depend on $C_h(p,q)$ for p,q belonging to the model m. The notation $\widehat{p}_m^{(h)}$ ignores the dependence on ε,δ which may be arbitrarily small.

For $\widehat{p}_m^{(h)}$ to be a valid estimator in L^∞ norm, it is necessary that $|\cdot|_h$ provide an adequate approximation to $\|\cdot\|_{\infty,\mu}$ on the model. Clearly, $|\cdot|_h$ does not uniformly approximate $\|\cdot\|_{\infty,\mu}$ on $\mathcal{P}$, so this property is model-dependent: this motivates the following definition.

Definition 1. *For any model $m \subset \mathscr{P}$ and any $h > 0$, let*

$$\kappa_m(h) = \inf_{p,q\in m}\frac{|p-q|_h}{\|p-q\|_{\infty,\mu}}.$$

This defines a function $\kappa_m : (0,+\infty) \to [0,1]$ which can be seen to have the following properties.

Lemma 1. *For any model $m \subset \mathscr{P}$,*

- *κ_m is non-increasing*
- *If $\kappa_m(h_0) > 0$ for some $h_0 > 0$, then $\kappa_m(h) > 0$ for all $h > 0$.*
- *κ_m is continuous, more precisely*

$$|\kappa_m(h_1) - \kappa_m(h_2)| \leq \left|1 - \frac{h_1 \wedge h_2}{h_1 \vee h_2}\right| \kappa_m(h_1 \wedge h_2).$$

Proof. The first property is obvious from the definition, the second is a consequence of the third. To prove the last inequality, note that

$$|\kappa_m(h_1) - \kappa_m(h_2)| \leq \sup_{p,q \in m} \left\{ \frac{1}{\|p-q\|_{\infty,\mu}} \big| |p-q|_{h_1} - |p-q|_{h_2} \big| \right\}.$$

Moreover, for any $f \in L^\infty$,

$$\begin{aligned} \big| |f|_{h_1} - |f|_{h_2} \big| &\leq \sup_{C \in \mathcal{C}} \left\{ \left| \int_C f \right| \left| \frac{1}{\mu(C)+h_1} - \frac{1}{\mu(C)+h_2} \right| \right\} \\ &\leq \sup_{C \in \mathcal{C}} \left\{ \frac{|h_1 - h_2| \, |\int_C f|}{(\mu(C)+h_1)(\mu(C)+h_2)} \right\} \\ &\leq |f|_{h_1 \wedge h_2} \frac{|h_1 - h_2|}{h_1 \vee h_2}. \end{aligned}$$

Together with the previous equation, this yields the result. □

Moreover, if $\mathcal{C}$ generates the Borel $\sigma-$algebra and m is a subset of a finite dimensional vector space, then by equivalence of norms, $\kappa_m(h) > 0$ for all $h > 0$.

To bound the stochastic error of the $\ell-$estimator, we introduce the following empirical process:

Definition 2. *For any $h > 0$, let*

$$Z(h) = \frac{1}{n} \sup_{C \in \mathcal{C}} \left\{ \frac{1}{\mu(C)+h} \left| \sum_{i=1}^n \mathbb{1}_C(X_i) - P_i^*(C) \right| \right\}.$$

Note that this definition does not depend on the model m. The risk of the $\ell-$estimator $\hat{p}_m^{(h)}$ may be related to the constant $\kappa_m(h)$ and the process $Z(h)$ as follows.

Proposition 1. *For any model $m \subset \mathscr{P}$ and any $h > 0$,*

$$(1-\varepsilon)\kappa_m(h) \times \left\| p^\star - \hat{p}_m^{(h)} \right\|_{\infty,\mu} \leq [2 + (1-\varepsilon)\kappa_m(h)] \inf_{p \in m} \|p^\star - p\|_{\infty,\mu} + 2Z(h) + \delta.$$

Proof. Let $p^\star = \frac{1}{n}\sum_{i=1}^n p_i^\star$ and $P^\star = \frac{1}{n}\sum_{i=1}^n P_i^\star$. For any $p, q \in \mathscr{P}$, let

$$\Delta_h(p,q) = \mathbb{E}\left[T^{(h)}(X,p,q)\right] = \varepsilon_h(p,q) \frac{(P - P^\star)(C_h(p,q))}{\mu(C_h(p,q)) + h}.$$

On the one hand,

$$\Delta_h(p,q) \leq |p - p^\star|_h \leq \|p - p^\star\|_{\infty,\mu}. \tag{3}$$

On the other hand,

$$\begin{aligned}\Delta_h(p,q) &= \varepsilon_h(p,q)\frac{(P-Q)(C_h(p,q))}{\mu(C_h(p,q))+h} + \varepsilon_h(p,q)\frac{(Q-P^\star)(C_h(p,q))}{\mu(C_h(p,q))+h}\\ &\geq (1-\varepsilon)|p-q|_h - \|q-p^\star\|_{\infty,\mu}\\ &\geq \kappa_m(h)(1-\varepsilon)\|p-q\|_{\infty,\mu} - \|q-p^\star\|_{\infty,\mu}\\ &\geq \kappa_m(h)(1-\varepsilon)\left(\|p-p^\star\|_{\infty,\mu} - \|p^\star-q\|_{\infty,\mu}\right) - \|q-p^\star\|_{\infty,\mu},\end{aligned}$$

from which it follows that

$$\Delta_h(p,q) \geq \kappa_m(h)(1-\varepsilon)\|p-p^\star\|_{\infty,\mu} - (1+\kappa_m(h)(1-\varepsilon))\|q-p^\star\|_{\infty,\mu}. \tag{4}$$

Moreover, by definition,

$$T^{(h)}(\mathbf{X},p,q) = \frac{\varepsilon_h(p,q)}{\mu(C_h(p,q))+h}\left[P(C_h(p,q)) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{C_h(p,q)}(X_i)\right],$$

which implies that for any $p,q\in\mathscr{P}$,

$$\left|T^{(h)}(\mathbf{X},p,q) - \Delta_h(p,q)\right| \leq Z(h). \tag{5}$$

Let now $p\in m$. On the one hand, by definition of $T_m^{(h)}(\mathbf{X},\cdot)$,

$$\begin{aligned}T_m^{(h)}\left(\mathbf{X},\hat{p}_m^{(h)}\right) &\geq T^{(h)}\left(\mathbf{X},\hat{p}_m^{(h)},p\right)\\ &= \Delta_h(\hat{p}_m^{(h)},p) + T^{(h)}\left(\mathbf{X},\hat{p}_m^{(h)},p\right) - \Delta_h(\hat{p}_m^{(h)},p)\\ &\geq \Delta_h(\hat{p}_m^{(h)},p) - Z(h) \text{ by equation (5)}\\ &\geq \kappa_m(h)(1-\varepsilon)\left\|\hat{p}_m^{(h)}-p^\star\right\|_{\infty,\mu} - (1+\kappa_m(h)(1-\varepsilon))\|p-p^\star\|_{\infty,\mu} - Z(h)\end{aligned} \tag{6}$$

by equation (4). On the other hand, for all $q\in m$, by equations (3) and (5),

$$\begin{aligned}T^{(h)}(\mathbf{X},p,q) &= \Delta_h(p,q) + T^{(h)}(\mathbf{X},p,q) - \Delta_h(p,q)\\ &\leq \|p-p^\star\|_{\infty,\mu} + Z(h),\end{aligned}$$

hence $T_m^{(h)}(\mathbf{X},p) \leq \|p-p^\star\|_{\infty,\mu} + Z(h)$. Finally, by equation (6) and definition of $\hat{p}_m^{(h)}$,

$$\begin{aligned}\delta + \|p-p^\star\|_{\infty,\mu} + Z(h) &\geq \delta + T_m^{(h)}(\mathbf{X},p)\\ &\geq T_m^{(h)}\left(\mathbf{X},\hat{p}_m^{(h)}\right)\\ &\geq \kappa_m(h)(1-\varepsilon)\left\|\hat{p}_m^{(h)}-p^\star\right\|_{\infty,\mu} - (1+\kappa_m(h)(1-\varepsilon))\|p-p^\star\|_{\infty,\mu} - Z(h),\end{aligned}$$

which yields

$$\kappa_m(h)(1-\varepsilon)\left\|\hat{p}_m^{(h)}-p^\star\right\|_{\infty,\mu} \leq (2+\kappa_m(h)(1-\varepsilon))\|p-p^\star\|_{\infty,\mu} + 2Z(h) + \delta.$$

As this is valid for any $p\in m$, the proposition is proved. □

To handle the stochastic process $Z(h)$, we state and prove a uniform Bernstein inequality. First, define the following family of events.

Definition 3. *Let* $P^\star = \frac{1}{n}\sum_{i=1}^n P_i^\star$ *and*

$$\Gamma = \log(\lceil \log_2 n \rceil) + \log\left(2\sum_{j=0}^{V\wedge n}\binom{n}{j}\right). \tag{7}$$

For any $x > 0$*, let* Ω_x *denote the event on which*

$$\frac{1}{n}\left|\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i^\star(C)\right| \leq \max\left(29\sqrt{P^\star(C)}\sqrt{\frac{\Gamma + x}{n}}, 20\frac{\Gamma + x}{n}\right), \tag{8}$$

for all $C \in \mathcal{C}$*.*

This class of events will govern the statistical behaviour of all procedures analyzed in this article. First, we prove the following proposition.

Proposition 2. *The event* Ω_x *has probability* $\mathbb{P}(\Omega_x) \geq 1 - 2e^{-x}$*.*

A result similar to proposition 2 was estalished by Baraud [2, Theorem 3] using similar methods. However, his result is stated for *suprema* of empirical processes over VC-classes and in particular, the variance term in the upper bound is the supremum of the variance of the empirical process over the class. What is novel about proposition 2, to the best of our knowledge, is that it provides a *pointwise* bound of the empirical process at each $C \in \mathcal{C}$ in terms of the variance of the process *at* C, for independent and not necessarily iid random variables.

Together with proposition 1, proposition 2 yields the following oracle inequality for the estimator $\hat{p}_m^{(h)}$.

Theorem 1. *Let* $p^\star = \frac{1}{n}\sum_{i=1}^n p_i^\star$ *and*

$$\Gamma = \log(\lceil \log_2 n \rceil) + \log\left(2\sum_{j=0}^{V\wedge n}\binom{n}{j}\right).$$

With probability greater than $1 - 2e^{-x}$*, for all countable models* $m \subset \mathscr{P}$ *and all* $h > 0$*,*

(9)
$$(1-\varepsilon)\kappa_m(h) \times \left\|\hat{p}_m^{(h)} - p^\star\right\|_{\infty,\mu} \leq [2 + (1-\varepsilon)\kappa_m(h)] \inf_{p\in m} \|p - p^\star\|_{\infty,\mu} + \delta$$
$$+ \max\left(58\sqrt{\frac{|p^\star|_h(\Gamma + x)}{hn}}, 40\frac{\Gamma + x}{hn}\right).$$

Proof. On Ω_x,

$$Z(h) \leq \max\left(29\sqrt{\frac{\Gamma+x}{n}}\sup_{C\in\mathcal{C}}\frac{\sqrt{P^*(C)}}{\mu(C)+h}, \frac{20(\Gamma+x)}{n}\sup_{C\in\mathcal{C}}\frac{1}{\mu(C)+h}\right)$$
$$\leq \max\left(29\sqrt{\frac{\Gamma+x}{n}}\sup_{C\in\mathcal{C}}\frac{\sqrt{P^*(C)}}{\mu(C)+h}, \frac{20(\Gamma+x)}{hn}\right).$$

For any $C \in \mathcal{C}$, by definition of $|p^\star|_h$,

$$\frac{\sqrt{P^\star(C)}}{\mu(C)+h} \leq \frac{\sqrt{|p^\star|_h(\mu(C)+h)}}{\mu(C)+h} \leq \sqrt{\frac{|p^\star|_h}{\mu(C)+h}} \leq \sqrt{\frac{|p^\star|_h}{h}}.$$

Hence, on Ω_x,

$$Z(h) \leq \max\left(29\sqrt{\frac{|p^\star|_h(\Gamma+x)}{hn}}, 20\frac{\Gamma+x}{hn}\right).$$

By proposition 1, equation (9) of the theorem holds on Ω_x. The conclusion follows from proposition 2. □

The quality of the bound provided by Theorem 1 depends on the constant $\kappa_m(h)$. If h_m is such that $\kappa_m(h_m) \geq \frac{1}{2}$ (say), then Theorem 1 yields a "true" oracle inequality, with remainder term of order $\sqrt{\frac{|p^\star|_{h_m}\Gamma}{nh_m}} \leq \sqrt{\frac{\|p^\star\|_{\infty,\mu}\Gamma}{nh_m}}$. Clearly, this value of h_m depends strongly on the class $\mathcal{C}$ and the model m. Later, we will show that $m, \mathcal{C}$ can be chosen such that equation (9) yields the minimax convergence rate over classes of smooth functions. More generally, one can ask when a constant h_m even exists. A sufficient condition for this is to have $\kappa_m(h) \to 1$ as $h \to 0$.

Existence of h_m provides an oracle inequality with a fixed constant (5, say) in front of the approximation error and a remainder term of order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ as $n \to +\infty$ for a fixed model m. This is the expected rate of convergence for finite-dimensional models.

We now show that, on $\mathbb{R}^d$ with Lebesgue measure μ, there are universal classes of sets $\mathcal{C}$ such that $\kappa_m(h) \to 1$ holds for all finite-dimensional m over which $\|\cdot\|_{\infty,\mu}$ is a norm.

Proposition 3. *Assume that $\mathcal{C}$ contains a sub-collection $\mathcal{C}_0$ satisfying the following conditions:*

- *For all $\delta > 0$, $\bigcup_{C\in\mathcal{C}_{0,\delta}} \overline{C} = \mathbb{R}^d$, where*
$$\mathcal{C}_{0,\delta} = \{C \in \mathcal{C}_0 : \operatorname{diam}(C) \leq \delta\}$$
- $\inf\left\{\frac{\mu(C)}{\operatorname{diam}(C)^d} : C \in \mathcal{C}_0\right\} > 0,$

where $\operatorname{diam}(C)$ *denotes the diameter of* C*. Then for any* $f \in \mathscr{P}$, $\lim_{h\to 0} |f|_h = \|f\|_{\infty,\mu}$*. As a consequence, if m is a subset of a finite-dimensional vector space,* $\lim_{h\to 0} \kappa_m(h) = 1$.

Proof. Fix some $\varepsilon > 0$. Assume without loss of generality that the set

$$A_\varepsilon = \{x \in \mathbb{R}^d : f(x) > (1-\varepsilon) \|f\|_{\infty,\mu}\}$$

has positive Lebesgue measure. Hence, by the Lebesgue differentiation Theorem, it contains a Lebesgue point x.

For each $k \in \mathbb{N}$, let $C_k \in \mathcal{C}_{0,\frac{1}{k}}$ such that $x \in \overline{C_k}$. In particular, $C_k \subset B(0, \mathrm{diam}(C_k))$. The assumptions of proposition 3 imply that $(C_k)_{k \geq 1}$ *shrinks to x nicely* in the sense of [16, section 7.9]. Hence, by [16, Theorem 7.10],

$$\lim_{k \to +\infty} \frac{1}{\mu(C_k)} \int_{C_k} f d\mu = f(x) \geq (1-\varepsilon) \|f\|_{\infty,\mu}.$$

Let $k \geq 1$ be such that $\frac{1}{\mu(C_k)} \int_{C_k} |f| d\mu \geq (1-2\varepsilon) \|f\|_{\infty,\mu}$. Then

$$\lim_{h \to 0} |f|_h \geq \lim_{h \to 0} \frac{1}{\mu(C_k) + h} \int_{C_k} f \geq (1-2\varepsilon) \|f\|_{\infty,\mu}.$$

Since this is true for any $\varepsilon > 0$, $\lim_{h \to 0} |f|_h = \|f\|_{\infty,\mu}$.

Let now $m \subset H$, where $H \subset \mathscr{P}$ is a finite dimensional vector space. Let K be the unit sphere of H in norm $\|\cdot\|_{\infty,\mu}$. The family of continuous functions

$$g_h : \begin{cases} K \to \mathbb{R} \\ f \mapsto |f|_h \end{cases}$$

is monotone with respect to the parameter h and converges pointwise at 0 to the constant function 1. Since K is compact, the convergence is uniform by Dini's theorem. In particular,

$$1 = \lim_{h \to 0} \inf_{x \in K} g_h(x) \leq \lim_{h \to 0} \kappa_m(h) \leq 1.$$

□

Classes of sets $\mathcal{C}$ which satisfy the assumptions of proposition 3 while having finite VC dimension include simplices, "box sets" (products of intervals), dyadic cubes, euclidean balls, ellipsoids, and many more.

3.1. **Piecewise polynomials.** To obtain more quantitative results about the constant $\kappa_m(h)$, it is necessary to look at specific classes of models. Here, we restrict attention to classes of piecewise polynomial functions on partitions of $\mathbb{R}^d$, because these classes are simple to define and have optimal approximation properties. However, we are confident that similar results could be proved for other classical function spaces, such as wavelet spaces or trigonometric polynomials. In the rest of this section, we shall assume that μ is the Lebesgue measure on $\mathbb{R}^d$.

First, it is necessary to introduce some definitions and notations. A (multivariate) *polynomial function* on $\mathbb{R}^d$ is a function of the form:

$$f : x \mapsto \sum_{a \in \mathcal{A}} c(a) \prod_{i=1}^{d} x_i^{a(i)},$$

where $\mathcal{A}$ is a finite set of functions $a : \{1, \ldots, d\} \to \mathbb{N}$ and $c : \mathcal{A} \to \mathbb{R}$ is a function. Its *degree* (in the usual sense) is defined to be

$$\deg(f) = \max_{a \in \mathcal{A}} \sum_{i=1}^{d} a(i).$$

It satisfies the usual relations, $\deg(fg) = \deg(f) + \deg(g)$ and $\deg(f + g) \leq \max(\deg(f), \deg(g))$. We define also the *directional degree* in direction $i \in \{1, \ldots, d\}$ to be

$$\deg_i(f) = \max_{a \in \mathcal{A}} a(i),$$

which satisfies the same relations. Let $\mathcal{P}_{\infty,d}$ be the space of all multivariate polynomial functions on $\mathbb{R}^d$. We define the following two families of spaces of polynomials with bounded degrees: first, given $r \in \mathbb{N}$, let

$$\mathcal{P}_{r,d} = \{f \in \mathcal{P}_{\infty,d} : \deg(f) \leq r\}.$$

Secondly, for all vectors $\mathbf{r} \in \mathbb{N}^d$, let

$$\mathcal{P}^{dir}_{\mathbf{r},d} = \{f \in \mathcal{P}_{\infty,d} : \forall i \in \{1, \ldots, d\}, \deg_i(f) \leq r_i\}.$$

The two families of spaces are related by the following inclusions:

$$\mathcal{P}^{dir}_{\mathbf{r},d} \subset \mathcal{P}_{\|\mathbf{r}\|_1,d} \subset \mathcal{P}^{dir}_{\|\mathbf{r}\|_1 \mathbf{1},d},$$

where $\mathbf{1}$ is the "all-one" vector, $\mathbf{1} = (1, \ldots, 1)$.

We can now define models of piecewise polynomial functions.

Definition 4. *Given a finite or countable and measurable partition $\mathcal{I}$ of $\mathbb{R}^d$ and $r \in \mathbb{N}$, let $m(r, \mathcal{I})$ denote the set of functions of the form*

$$f = \sum_{I \in \mathcal{I}} f_I \mathbb{1}_I,$$

where for each $I \in \mathcal{I}$, $f_I \in \mathcal{P}_{r,d}$ is a polynomial with rational coefficients and the set $\{I \in \mathcal{I} : f_I \neq 0\}$ is finite. Let $\bar{m}(r, \mathcal{I}) = \overline{m(r, \mathcal{I})}$, the closure of $m(r, \mathcal{I})$ in $L^\infty\left(\mathbb{R}^d\right)$.

Given $\mathbf{r} \in \mathbb{N}^d$, let $m_{dir}(\mathbf{r}, \mathcal{I}), \bar{m}_{dir}(\mathbf{r}, d)$ be defined similarly, with $\mathcal{P}^{dir}_{\mathbf{r},d}$ instead of $\mathcal{P}_{r,d}$.

Let the model $m = m(r, \mathcal{I})$ for some partition $\mathcal{I}$. If $\mathcal{I}$ is finite, then m is finite dimensional and the previous proposition applies. In general, to establish an explicit lower bound on $\kappa_m(h)$, we require the partition $\mathcal{I}$ to satisfy the following three conditions.

Assumption 1.

- *$\mathcal{C}$ contains translated and scaled copies of the interior $\mathring{I}$ of any $I \in \mathcal{I}$, i.e*

$$\left\{x + \lambda \mathring{I} : I \in \mathcal{I}, x \in \mathbb{R}^d, \lambda > 0\right\} \subset \mathcal{C}.$$

- *There is a lower bound on the volume of the elements of $\mathcal{I}$:*
$$h_0 := \min_{I\in\mathcal{I}} \mu(I) > 0.$$
- *The elements of $\mathcal{I}$ are bounded convex sets.*

Under assumption 1, for any $f = p - q \in m$, an appropriate set $C_{h,m}(f) \in \mathcal{C}$ can be constructed as follows. Since the collection $(f\mathbb{1}_I)_{I\in\mathcal{I}}$ has finite support, the supremum $\sup_{I\in\mathcal{I}} \|f\mathbb{1}_I\|_{\infty,\mu}$ is reached at some $I_*(f)$. Let $\mathring{I}_*(f)$ denote the topological interior of $I_*(f)$. f coincides on $I_*(f)$ with a polynomial f_*, which reaches its maximum on $\overline{I_*(f)}$ at some $x_*(f)$.

Let finally

$$C_{h,m}(f) = (1-\theta_m(h))x_*(f) + \theta_m(h)\mathring{I}_*(f), \tag{10}$$

where $\theta_m(h) \in (0,1)$ is a function given by equation (11) below.

By assumption 1, $C_{h,m}(f) \in \mathcal{C}$ and by convexity of $I_*(f)$, $C_{h,m}(f) \subset \mathring{I}_*(f)$. The following lower bound holds.

Proposition 4. *For all $u > 0$, let*

$$\gamma_{r,d}(u) = \max\left(\frac{1}{2(d+1)}\left[\frac{u^{-1}}{(2r^2)^d} \wedge 1\right], \left[1-(2r^2)^{\frac{d}{d+1}}u^{\frac{1}{d+1}}\right]_+^2\right).$$

Assume that hypothesis 1 holds. Let then

$$\theta_m(h) = \begin{cases} \frac{d}{d+1}\frac{1}{2r^2} \text{ if } \gamma_{r,d}\left(\frac{h}{h_0}\right) = \frac{1}{2(d+1)}\left[\frac{h_0}{(2r^2)^d h} \wedge 1\right] \\ \left(\frac{h}{2r^2 h_0}\right)^{\frac{1}{d+1}} \text{ otherwise }. \end{cases} \tag{11}$$

For all $f \in m = m(r,\mathcal{I})$,

$$\frac{\left|\int_{C_{h,m}(f)} f d\mu\right|}{\mu\left(C_{h,m}(f)\right) + h} \geq \gamma_{r,d}\left(\frac{h}{h_0}\right)\|f\|_{\infty,\mu}.$$

In particular, since m is a $\mathbb{Q}-$vector space, $\kappa_m(h) \geq \gamma_{r,d}\left(\frac{h}{h_0}\right)$.

Proof. The proof is carried out in appendix A.2. □

In particular, $\kappa_m(h)$ converges to 1 as $\frac{h}{h_0} \to 0$, and the rate of convergence depends only on the dimension d (and not on the partition $\mathcal{I}$).

Thus, the estimation error behaves essentially like $\frac{1}{\sqrt{h_0}}$, when h is well chosen. For concreteness, consider the case of the collection $\mathcal{C}$ of cartesian products of d intervals, with $\mathcal{I} \subset \mathcal{C}$ a partition of $\mathbb{R}^d$. Then Theorem 1 and Proposition 6 yield the following Corollary.

Corollary 1. *Let $m = m(r,\mathcal{I})$, $\mathcal{I} \subset \mathcal{C}$ satisfying 1, $\mathcal{C}$ the collection of cartesian products of d intervals. Let*

$$h_m = \frac{\left(1-\frac{1}{\sqrt{2}}\right)^{d+1}}{(2r^2)^d} h_0. \tag{12}$$

Let $\hat{p}_m^{(h_m)}$ be the $\ell-$estimator based on the sets $C_{h_m,m}(p-q)$ $(p,q \in m)$ defined above. Then

$$\mathbb{E}\left[\left\|\hat{p}_m^{(h_m)} - p^\star\right\|_{\infty,\mu}\right] \leq 5\min_{p\in m}\{\|p-p^\star\|_{\infty,\mu}\} + 274(2d+1)(3r)^{2d}\frac{\log(en)}{h_0 n} + 2\delta$$
$$+ 215\sqrt{2d+1}\min\left((3r)^d\sqrt{\|p^\star\|_{\infty,\mu}}\sqrt{\frac{\log(en)}{h_0 n}}, (3r)^{2d}\frac{\sqrt{\log(en)}}{h_0\sqrt{n}}\right).$$

Proof. The proof can be found in appendix A.3. □

The remainder term in the oracle inequality above is equivalent to

$$c_{r,d}\sqrt{\|p^\star\|_{\infty,\mu}}\sqrt{\frac{\log(n)}{h_0 n}}$$

for some constant $c_{r,d}$ (depending on r,d only), in the asymptotic regime where $h_0 \to 0$ and $h_0 n \to +\infty$. We show below that this is optimal for sufficiently "regular" partitions. Though we do not believe that the constant $c_{r,d}$ is optimal, exponential behaviour of the type r^{cd} is expected since

$$\dim(\mathcal{P}_{r,d}) = \binom{r+d}{d} \geq \left(\frac{r+d}{d}\right)^d.$$

To assess the optimality of the remainder term $\sqrt{\|p^\star\|_{\infty,\mu}}\sqrt{\frac{\log(n)}{h_0 n}}$ of Corollary 1 and more generally of Theorem 1, we prove a minimax lower bound on the class

$$m_L(0,\mathcal{I}) = \left\{\sum_{I\in\mathcal{I}} c_I \mathbb{1}_I : c \in [0,L]^{\mathcal{I}}, \sum_{I\in\mathcal{I}} c_I\mu(I) = 1\right\}$$

of pdfs which are piecewise constant on the blocks of the partition $\mathcal{I}$ and uniformly bounded by $L > 0$.

Note that the set $m_L(0,\mathcal{I})$ may be empty (if $\mathcal{I}$ does not contain blocks of finite measure), or a singleton (if $\mathcal{I}$ has exactly one block of finite measure). If $\mathcal{I}$ has a finite number of blocks of finite measure, then $m_L(0,\mathcal{I})$ will also be empty if L is too small. In such cases, estimation on $m_L(0,\mathcal{I})$ is trivial.

In general, the following minimax lower bound holds.

Theorem 2. *Let $(\mathcal{X},\mathcal{B},\mu)$ be a $\sigma-$finite measure space and $\mathcal{I}$ a countable, measurable partition of $\mathcal{X}$ into blocks of positive measure. Let*

$$\mathcal{X}_0 = \bigcup\{I \in \mathcal{I} : \mu(I) < +\infty\}.$$

For any $h > 0$, let

$$M(h) = |\{I \in \mathcal{I} : \mu(I) \leq h\}|.$$

For any $L > 0$ and $n \geq 1$, define $\psi_n(\mathcal{I},L) > 0$ by

$$\psi_n(\mathcal{I},L)^2 = \sup_{h>0}\left\{\frac{L}{hn}\log\left(1+\min\left(M(h), \left\lfloor\frac{1}{Lh}\right\rfloor\right)\right)\right\}. \quad (13)$$

Then, for any $\theta \in \left(\frac{1}{2}, 1\right)$ *and any* $L \geq \frac{1}{\theta\mu(\mathcal{X}_0)}$*,*

$$\inf_{\hat{p}} \sup_{p^\star \in m_L(0,\mathcal{I})} \mathbb{E}\left[\|\hat{p} - p^\star\|_{\infty,\mu}\right] \geq \frac{1}{40} \min\left((1-\theta)L, \sqrt{\theta(1-\theta)}\psi_n(\mathcal{I}, L)\right),$$

where the infimum runs over all estimators $\hat{p}$ *of* $p^\star$*, based on an iid sample of size* n *drawn from* $p^\star$*.*

Proof. The proof is based on standard multiple testing arguments. It can be found in appendix A.4. □

Though the class $m_L(0,\mathcal{I})$ is a simple one, and the proof of Theorem 2 uses standard "multiple testing" arguments, Theorem 2 is, to the best of our knowledge, the first minimax lower bound for classes of piecewise constant functions in density estimation in sup-norm.

The lower bound involves the parameters L, h, n and θ, as well as the function M which depends on the partition $\mathcal{I}$.

The parameter θ reflects the fact that if L is too small, then the model is empty, and if $L = \frac{1}{\mu(\mathcal{X}_0)}$, then the model contains precisely one element (the uniform distribution on $\mathcal{X}_0$). As soon as L is greater than this minimum value by constant factor $\frac{1}{\theta}$, the lower bound is of order

$$\min(L, \psi_n(\mathcal{I}, L)).$$

The minimum with L reflects the fact that we can always use any fixed $p_0 \in m_L(0,\mathcal{I})$ as an estimator, which has risk bounded by L. As soon as n is large enough, such that this trivial estimator is sub-optimal, the minimax risk becomes proportional to $\psi_n(\mathcal{I}, L)$.

This term, $\psi_n(\mathcal{I}, L)$, is somewhat complicated. For the purpose of this discussion, fix a partition $\mathcal{I}$ and let

$$h_0 = \inf_{I \in \mathcal{I}} \mu(I).$$

If $L < \frac{1}{h_0}$, then for any $h \in \left(h_0, \frac{1}{L}\right]$, $M(h) \geq 1$ and $\frac{1}{Lh} \geq 1$, which implies that $\psi_n(\mathcal{I}, L) \geq \sqrt{\frac{L \log 2}{hn}}$. On the other hand, if $L \geq \frac{1}{h_0}$, then since the models $(m_t(0,\mathcal{I}))_{t>0}$ are nested, the minimax risk on $m_L(0,\mathcal{I})$ is greater than the minimax risk on $m_{\frac{1}{h}}(0,\mathcal{I})$ for any $h > h_0$.

This yields the following corollary.

Corollary 2. *Let* $\mathcal{I}$ *be a countable partition of* $\mathcal{X}$ *into blocks of finite, positive measure. For any* $L \geq \frac{2}{\mu(\mathcal{X})}$*,*

$$\inf_{\hat{p}} \sup_{p \in m_L(0,\mathcal{I})} \mathbb{E}\left[\|\hat{p} - p\|_{\infty,\mu}\right] \geq \frac{1}{80} \min\left(L, \sqrt{\frac{L \log 2}{h_0 n}}, \frac{\sqrt{\log 2}}{h_0\sqrt{n}}\right),$$

where $h_0 = \inf_{I \in \mathcal{I}} \mu(I)$*.*

Comparing corollary 2 to the minimax upper bound resulting from Corollary 1, we see that Corollary 1 is optimal, possibly up to $\log n$ factors and the remainder term $\frac{1}{h_0 n}$, which is negligible relative to the minimax lower bound whenever $\sqrt{\frac{L}{h_0 n}} \ll L$, i.e whenever a non-trivial estimator is required.

If we assume additionally that

$$M(2h_0) = |\{I \in \mathcal{I} : h_0 \leq \mu(I) \leq 2h_0\}| \geq n^\alpha$$

and that $h_0 \leq \frac{1}{2Ln^\alpha}$ for some fixed $\alpha \in (0,1)$, then by equation (13),

$$\psi_n(\mathcal{I}, L) \geq \sqrt{\frac{L}{2h_0 n}} \sqrt{\log(1 + \lfloor n^\alpha \rfloor)} \geq \sqrt{\frac{\alpha L \log n}{2h_0 n}},$$

in which case the upper bound of Corollary 1 is optimal up to a constant depending only on α.

Moreover, if $\mathcal{I}_n$ are regular partitions of $\mathbb{R}^d$ into blocks of volume h_n, where $\limsup_{n \to +\infty} \{n^\alpha h_n\} < +\infty$, then

$$\liminf_{n \to +\infty} \left\{ \psi_n(\mathcal{I}_n, L) \times \sqrt{\frac{h_n n}{L \log n}} \right\} \geq \sqrt{\alpha},$$

which proves the asymptotic optimality of Corollary 1 in this non-parametric setting.

4. Model selection and adaptivity

4.1. **General approach.** Let $\mathcal{M}$ be a collection of *models* and let $\mathbf{M} = \cup_{m \in \mathcal{M}} m$. In principle, the tests $t_{p,q}^{(h)}$ for a fixed h could be used to select an element of $\mathbf{M}$. However, in order for this approach to work, it is necessary that $\inf_{m \in \mathcal{M}} \kappa_m(h) \geq \kappa_* > 0$: in particular, if the models are nested, the value of h chosen corresponds to that required for estimation on the largest model.

It would be desirable to instead use different values of h depending on the models to which p, q belong, so as to obtain an estimator which performs as well as the best single-model estimator in the collection $(\hat{p}_m^{h_m})_{m \in \mathcal{M}}$.

To achieve this goal of *model selection*, some means of estimating the statistical error $Z(h)$ is needed. Theorem 1 provides an upper bound on $Z(h)$ which is almost fully explicit: it only depends on $\mathbf{P}^\star$ through $|p^\star|_h$. We now show how this quantity can be estimated.

Definition 5. *For any $h > 0$, let*

$$|\hat{p}|_h = \sup_{C \in \mathcal{C}} \left\{ \frac{\sum_{i=1}^n \mathbb{1}_C(X_i)}{n(\mu(C) + h)} \right\}.$$

The following proposition shows that $|\hat{p}|_h$ is an adequate estimator of $|p^\star|_h$.

Proposition 5. *On Ω_x, for all $\theta \in (0,2)$,*

$$|p^\star|_h \leq \frac{1}{1-\frac{\theta}{2}}|\hat{p}|_h + \frac{29^2}{\theta(2-\theta)}\frac{\Gamma+x}{hn}$$
$$|\hat{p}|_h \leq \left(1+\frac{\theta}{2}\right)|p^\star|_h + \frac{29^2}{2\theta}\frac{\Gamma+x}{hn},$$

where Γ and Ω_x are given by Definition 3.

Proof. On Ω_x, by definition 3, for any $C \in \mathcal{C}$,

$$\begin{aligned}
\frac{\sum_{i=1}^n \mathbb{1}_C(X_i)}{n(\mu(C)+h)} &= \frac{1}{\mu(C)+h}\left[P^\star(C) + \frac{1}{n}\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i^\star(C)\right] \\
&\geq \frac{1}{\mu(C)+h}\left[P^\star(C) - \max\left(29\sqrt{P^\star(C)}\sqrt{\frac{\Gamma+x}{n}}, 20\frac{\Gamma+x}{n}\right)\right] \\
&\geq \frac{1}{\mu(C)+h}\min\left(\left(1-\frac{\theta}{2}\right)P^\star(C) - \frac{29^2}{2\theta}\frac{\Gamma+x}{n}, P^\star(C) - 20\frac{\Gamma+x}{n}\right) \\
&\geq \left(1-\frac{\theta}{2}\right)\frac{P^\star(C)}{\mu(C)+h} - \frac{29^2}{2\theta}\frac{\Gamma+x}{hn}
\end{aligned}$$

Taking a supremum on both sides with respect to $C \in \mathcal{C}$ yields

$$|\hat{p}|_h \geq \left(1-\frac{\theta}{2}\right)|p^\star|_h - \frac{29^2}{2\theta}\frac{\Gamma+x}{hn},$$

which yields the first equation. Similarly,

$$\begin{aligned}
\frac{\sum_{i=1}^n \mathbb{1}_C(X_i)}{n(\mu(C)+h)} &= \frac{1}{\mu(C)+h}\left[P^\star(C) + \frac{1}{n}\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i^\star(C)\right] \\
&\leq \frac{1}{\mu(C)+h}\left[P^\star(C) + \max\left(29\sqrt{P^\star(C)}\sqrt{\frac{\Gamma+x}{n}}, 20\frac{\Gamma+x}{n}\right)\right] \\
&\leq \frac{1}{\mu(C)+h}\max\left(\left(1+\frac{\theta}{2}\right)P^\star(C) + \frac{29^2}{2\theta}\frac{\Gamma+x}{n}, P^\star(C) + 20\frac{\Gamma+x}{n}\right) \\
&\leq \left(1+\frac{\theta}{2}\right)\frac{P^\star(C)}{\mu(C)+h} + \frac{29^2}{2\theta}\frac{\Gamma+x}{hn},
\end{aligned}$$

which yields the second equation. □

Based on Theorem 1 and the above proposition with $\theta = \frac{1}{2}$, let us define the following universal penalty. Let $\log_-$ denote the negative part of the log function, and let $a > 0$ be some parameter. Let then

$$\mathrm{pen}_a(h) = 29\sqrt{\frac{4}{3}}\sqrt{\frac{|\hat{p}|_h(\Gamma + a\log_-(h))}{hn}} + \sqrt{\frac{4}{3}}29^2\frac{\Gamma + a\log_-(h)}{hn}, \tag{14}$$

where Γ is defined by equation (7).

Assume that for any model $m \in \mathcal{M}$, there is an associated parameter $h_m > 0$, chosen such that $\hat{p}_m^{(h_m)}$ satisfies an oracle inequality on model m with fixed constant independent of m, i.e such that $\kappa_m(h_m) \geq \kappa_0 > 0$ for some constant κ_*.

In order to perform model selection, we need to control the behaviour of the tests $T^{(h)}(\mathbf{X}, p, q)$ when p, q belong to two different models. It may be that comparing two models is much harder (i.e, requires a much smaller value of h) than optimizing performance within a single model. For example, while it is feasible to optimize among piecewise constant functions on a given partition $\mathcal{I}$, selecting among partitions $\mathcal{I}$ is impossible in general since the set $(\mathbb{1}_{[a,b]})$ of indicator functions of intervals is non-separable in L^∞.

To avoid such cases, we make the following assumption.

Assumption 2. *There exists a constant*

$$\kappa_* = \inf_{m, m' \in \mathcal{M}} \{\kappa_{m \cup m'}(h_m \wedge h_{m'})\} > 0.$$

Qualitatively speaking, assumption 2 states that comparing p, q belonging to m, m' is not significantly harder than comparing p_1, q_1 belonging to the same model (m or m'). For example, this is always the case when models are nested.

Remark. *If $\mathcal{M}$ is totally ordered with respect to inclusion, then*

$$\kappa_* = \inf_{m \in \mathcal{M}} \{\kappa_m(h_m)\} \geq \kappa_0 > 0.$$

Proof. Let $m, m' \in \mathcal{M}$ and assume without loss of generality that $m' \subset m$. Since κ_m is a non-increasing function,

$$\kappa_{m \cup m'}(h_m \wedge h_{m'}) = \kappa_m(h_m \wedge h_{m'}) \geq \kappa_m(h_m).$$

This proves that $\kappa_* \geq \inf_{m \in \mathcal{M}}\{\kappa_m(h_m)\}$. On the other hand, taking $m = m'$ in assumption 2 yields $\kappa_* \leq \kappa_m(h_m)$. □

Assuming now that $\mathcal{M}, (h_m)_{m \in \mathcal{M}}$ satisfy hypothesis 2, we construct a model selection procedure as follows. For any $p \in \mathbf{M}$, let

$$h_p = \sup\{h_m : m \in \mathcal{M}, p \in m\}.$$

For any $p \in \mathbf{M}$, let then

$$T_{\mathcal{M}}(\mathbf{X}, p) = \sup_{q \in \mathbf{M}} \left\{ T^{(h_p \wedge h_q)}(\mathbf{X}, p, q) - \mathrm{pen}_a(h_q) \right\} + \mathrm{pen}_a(h_p).$$

A model selection $\ell-$estimator is defined to be any random element $\hat{p}_{\mathcal{M}}$ such that

$$T_{\mathcal{M}}(\mathbf{X}, \hat{p}_{\mathcal{M}}) \leq \inf_{p \in \mathbf{M}} \{T_{\mathcal{M}}(\mathbf{X}, p)\} + \delta.$$

The model selection $\ell-$estimator satisfies the following oracle inequality.

Theorem 3. *For all $y \geq e$, on an event $(\Omega_{a \log y})$ with probability greater than $1 - \frac{2}{y^a}$,*

(15)
$$(1-\varepsilon)\kappa_* \left\|\hat{p}_{\mathcal{M}} - p^\star\right\|_{\infty,\mu} \leq \inf_{m \in \mathcal{M}} \left\{(2 + (1-\varepsilon)\kappa_*) \inf_{p \in m}\{\|p - p^\star\|_{\infty,\mu}\} + 4\,\mathrm{pen}_a(h_m)\right\} + 29y\sqrt{\frac{2a}{3en}} + 29^2 \frac{4}{\sqrt{3}} \frac{ay}{en} + \delta.$$

Proof. The proof is postponed to appendix B.1. □

An interesting aspect of Theorem 3 is that the penalty only depends on h_m and on the fixed parameter a, but not on the number of models. In particular, the theorem also applies to countably infinite collections $\mathcal{M}$.

4.2. **Piecewise polynomials on regular dyadic partitions.** The key question concerning applications of Theorem 3 is for which collections of models assumption 2 holds. We have already seen that assumption 2 holds for nested model collections, however the assumption that models are nested is restrictive: it excludes classes of irregular partitions that one would like to use in order to adapt to potentially inhomogeneous or anisotropic smoothness of the target density.

Perhaps unexpectedly, assumption 2 turns out to be significantly weaker than nestedness. If h_m is chosen according to lemma 4 (for a value $h < \frac{h_0}{r^2}$), then assumption 2 holds over the class $m(r, \mathcal{I})$, where $r \in \mathbb{N}$ and $\mathcal{I}$ belongs to the set of *regular dyadic partitions*, i.e, partitions

$$\mathcal{I}(\mathbf{j}) = \left\{\prod_{i=1}^d [k_i 2^{-j_i}, (k_i+1)2^{-j_i}) : \mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{Z}^d\right\},$$

for some $\mathbf{j} \in \mathbb{Z}^d$. For completeness, we prove in appendix B.2 that the $\mathcal{I}(\mathbf{j})$ are indeed partitions of $\mathbb{R}^d$, with the property that $\mathcal{I}(\mathbf{j}')$ refines $\mathcal{I}(\mathbf{j})$ whenever $\mathbf{j}' \geq \mathbf{j}$.

Denote then

$$\mathfrak{I}_d = \left\{\mathcal{I}(\mathbf{j}) : \mathbf{j} \in \mathbb{Z}^d\right\}$$

and

$$\mathcal{M}_{\mathbf{r}} = \{m_{dir}(\mathbf{r}, \mathcal{I}) : \mathcal{I} \in \mathfrak{I}_d\}. \tag{16}$$

For any $m = m_{dir}(\mathbf{r}, \mathcal{I}) \in \mathcal{M}_{\mathbf{r}}$, let

$$h_m = \frac{\min_{I \in \mathcal{I}}\{\mu(I)\}}{(2\left\|\mathbf{r}\right\|_1^2)^d 4^{d+1}} = \frac{2^{-(j_1 + \ldots + j_d)}}{(2\left\|\mathbf{r}\right\|_1^2)^d 4^{d+1}}. \tag{17}$$

Consider the class $\mathcal{C}_{rec}$ of $d-$dimensional open rectangles with sides parallel to the axes, i.e

$$\mathcal{C}_{rec} = \{\prod_{i=1}^{d}(a_i, b_i) : a_i, b_i \in \mathbb{R}, a_i < b_i\}.$$

This class generates the Borel sigma-algebra, hence for any $h > 0$, $|\cdot|_h$ is a norm on $L_1 \cap L_\infty$.

The following Theorem shows that the model collection $\mathcal{M}_{\mathbf{r}}$ satisfies assumption 2 for a constant κ_* depending only on $\mathbf{r}$ and d.

Theorem 4. *For all $m, m' \in \mathcal{M}_{\mathbf{r}}$ and $h_m, h_{m'}$ defined by equation (17),*

$$\kappa_{m\cup m'}(h_m \wedge h_{m'}) \geq \left[4\left(1 + 4\sqrt{\prod_{i=1}^{d}(r_i+1)}\right)\right]^{-1}.$$

Proof. The proof can be found in appendix B.3. □

In light of Theorem 3, Theorem 4 implies that it is possible to perform *model selection* on the model collection $\mathcal{M}_{\mathbf{r}}$, in the sense that the model-selection estimator $\hat{p}_{\mathcal{M}_{\mathbf{r}}}$ defined in section 4.1 performs as well as the best estimator in the collection $\{\hat{p}_m^{(h_m)} : m \in \mathcal{M}_{\mathbf{r}}\}$, up to a constant depending only on $d, \mathbf{r}$.

5. Rates under anisotropic smoothness

The oracle inequality satisfied by the $\ell-$estimator (Theorem 1), together with the lower bound on κ_m for polynomial models (proposition 4) allow to recover minimax optimal rates on anisotropic Lipschitz spaces. Moreover, the model selection results in the previous section (Theorems 3 and 4) imply that this can be done in an adaptive manner, as we now show.

Let $\boldsymbol{\beta} \in \mathbb{R}^d$ be a multi-index, and let $C^{\boldsymbol{\beta}}$ denote the space of functions f which admit partial derivatives $\frac{\partial^{k_i} f}{\partial x_i^{k_i}}$ at all orders $k_i \leq \lfloor \beta_i \rfloor$, and are such that the semi-norms

$$|f|_{i,\beta_i} := \sup_{x\in\mathbb{R}^d} \sup_{t\in\mathbb{R}} \frac{1}{t^{\beta_i - \lfloor\beta_i\rfloor}} \left| \frac{\partial^{\lfloor\beta_i\rfloor} f}{\partial x_j^{\lfloor\beta_i\rfloor}}(x + te_i) - \frac{\partial^{\lfloor\beta_i\rfloor} f}{\partial x_i^{\lfloor\beta_i\rfloor}}(x) \right|$$

are finite for all $i \in \{1, \ldots, d\}$, where e_i denotes the standard basis of $\mathbb{R}^d$. Note that this only requires regularity along the coordinate directions, and in particular the cross-derivatives may fail to exist.

It is known that the minimax-optimal convergence rate on the class $C^{\boldsymbol{\beta}}$ is $\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}$, where β is the harmonic mean of the β_i. This follows from results of Lepski [11].

Let $\mathcal{C} = \mathcal{C}_{rec}$ be the class of products of d open intervals, and let $\hat{p}$ be the model-selection $\ell-$estimator defined in section 4.1, over the model collection $\mathcal{M}_{\mathbf{r}}$ defined in section 4.2 equation (16), with the values h_m specified by equation (17).

To prove that $\hat{p}$ attains the optimal rate when $p^\star \in C^{\boldsymbol{\beta}}$, an approximation result is needed in order to bound the term $\inf_{p \in m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}))} \|p - p^\star\|_{\infty,\mu}$. Such results have long been established in the approximation theory literature when $\boldsymbol{\beta} \in \mathbb{N}^d$, along with bounds on the approximation error expressed in terms of finite difference operators - the article [7] is particularly relevant. However, these results usually involve a non-explicit constant. Rather than adapt them to our setting, it is just as convenient to give a direct proof, which also provides an explicit constant.

Proposition 6. *Let $f \in C^{\boldsymbol{\beta}}(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$. Let $\mathbf{r} \geq \lfloor \boldsymbol{\beta} \rfloor$ be a vector of integers, let $\mathbf{h} = (h_j)_{1 \leq j \leq d}$ be non-negative real numbers and let $\mathcal{I}_{\mathbf{h}}$ denote the rectangular partition*

$$\left\{ \prod_{j=1}^{d} [k_j h_j, (k_j + 1) h_j) : k \in \mathbb{Z}^d \right\}.$$

There exists $f_d \in \overline{m_{dir}}(\mathbf{r}, \mathcal{I}_{\mathbf{h}})$ such that

$$\|f - f_d\|_{\infty,\mu} \leq 2 b_d(\mathbf{r}) \max_{1 \leq j \leq d} \left\{ \frac{h_j^{\beta_j}}{\lfloor \beta_j \rfloor !} |f|_{j,\beta_j} \right\},$$

where

$$b_d(\mathbf{r}) = 1 + \min_{\sigma \in \mathfrak{S}_d} \left\{ \sum_{j=1}^{d} \prod_{i=1}^{j} \left[\frac{2}{\pi} \log(1 + r_{\sigma(i)}) + 1 \right] \right\}. \tag{18}$$

Proof. The proof can be found in appendix C.1. □

Optimizing the bias-variance tradeoff between approximation error (given by proposition 6) and estimation error (given by $\mathrm{pen}_a(h_m)$ defined in equation (14)) yields the following Theorem.

Theorem 5. *Let $\boldsymbol{\beta} \in \mathbb{R}_+^d$ be such that $\lfloor \beta_j \rfloor \leq r_j$ for all j. Let*

$$\beta = \frac{1}{d} \sum_{j=1}^{d} \frac{1}{\beta_j}.$$

Let $p^\star = \frac{1}{n} \sum_{i=1}^{n} p_i^\star$. Assuming that $p^\star \in C^\beta \left(\mathbb{R}^d\right)$, let

$$L_\beta(p^\star) = \prod_{j=1}^{d} |p^\star|_{j,\beta_j}^{\frac{\beta}{d \beta_j}}.$$

For any $n \in \mathbb{N}$ such that $L_{\boldsymbol{\beta}}(p^\star)^{\frac{d}{\beta}} \leq \frac{n}{\log n}$,

$$\mathbb{E}\left[\|\hat{p} - p^\star\|_{\infty,\mu}\right] \leq C \|p^\star\|_{\infty,\mu}^{\frac{\beta}{2\beta+d}} L_{\boldsymbol{\beta}}(p^\star)^{\frac{d}{2\beta+d}} \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}$$

when $\|p^\star\|_{\infty,\mu} \geq L_{\boldsymbol{\beta}}(p^\star)^{\frac{d}{\beta+d}} \left(\frac{\log n}{n}\right)^{\frac{\beta}{\beta+d}}$, and

$$\mathbb{E}\left[\|\hat{p} - p^\star\|_{\infty,\mu}\right] \leq C L_{\boldsymbol{\beta}}(p^\star)^{\frac{d}{\beta+d}} \left(\frac{\log n}{n}\right)^{\frac{\beta}{\beta+d}}$$

else, where C is a constant which depends only on $\mathbf{r}, d$.

Proof. The proof is carried out in appendix C.2. □

Assume to simplify the discussion that $\mathbf{p}^\star = (p^\star, \ldots, p^\star))$ and that $p^\star \in C^{\boldsymbol{\beta}}(\mathbb{R}^d)$. Then, Theorem 5 yields the correct [11, Theorem 2] minimax convergence rate, $\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}$, with respect to the sample size n. Moreover, the dependence of the upper bound on $p^\star$ is explicit, through the term $\|p^\star\|_{\infty,\mu}^{\frac{\theta\beta}{2\beta+d}} L_{\boldsymbol{\beta}}(p^\star)^{\frac{d}{2\beta+d}}$. Note that the assumption $p^\star \in C^{\boldsymbol{\beta}}(\mathbb{R}^d)$, together with the fact that $p^\star$ is a density, imply a bound on $\|p^\star\|_{\infty,\mu}$ by a function of $\boldsymbol{\beta}, d$ and the semi-norms $|p^\star|_{j,\beta_j}$.

Consider now the class of functions

$$\mathcal{C}_{\mathbf{L},b}^{\boldsymbol{\beta}} = \left\{p \in C^{\boldsymbol{\beta}}(\mathbb{R}^d) : \|p\|_{\infty,\mu} \leq b, \forall j \in \{1, \ldots, d\}, |p|_{j,\beta_j} \leq L_j\right\},$$

as well as the class $\mathcal{P}_{\mathbf{L},b}^{\boldsymbol{\beta}}$ of probability density functions which belong to $\mathcal{C}_{\mathbf{L},b}^{\boldsymbol{\beta}}$. These function classes are non-decreasing as a function of b, moreover by the previous remark there is a function

$$b_{min}(\mathbf{L}, \boldsymbol{\beta}) = \sup\left\{\|p\|_{\infty,\mu} : p \in \mathcal{P}_{\mathbf{L},+\infty}^{\boldsymbol{\beta}}\right\}$$

such that $b \mapsto \mathcal{P}_{\mathbf{L},b}^{\boldsymbol{\beta}}$ is strictly increasing for $b < b_{min}(\mathbf{L}, \boldsymbol{\beta})$ and constant for all $b \geq b_{min}(\mathbf{L}, \boldsymbol{\beta})$.

Theorem 5 implies in particular the following minimax upper bound on the classes $\mathcal{P}_{\mathbf{L},b}^{\boldsymbol{\beta}}$:

Corollary 3. *Let $\boldsymbol{\beta} \in \mathbb{R}_+^d$ and $\frac{1}{\beta} = \frac{1}{d}\sum_{j=1}^d \frac{1}{\beta_j}$. Let $\mathbf{L} \in \mathbb{R}_+^d$ and*

$$L = \prod_{j=1}^d L_j^{\frac{\beta}{d\beta_j}}.$$

For all $b \leq b_{min}(\mathbf{L}, \boldsymbol{\beta})$ and all large enough n,

$$\inf_{\tilde{p}} \sup_{p^\star \in \mathcal{P}_{\mathbf{L},b}^{\boldsymbol{\beta}}} \mathbb{E}\left[\|\hat{p} - p^\star\|_{\infty,\mu}\right] \leq C b^{\frac{\beta}{2\beta+d}} L^{\frac{d}{2\beta+d}} \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}},$$

where the infimum runs over all estimators computed from an $n-$sample drawn from $p^\star$, and C is a constand depending only on $\boldsymbol{\beta}, d$.

In addition to the rate $\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}$, the optimality of which is known, a natural further question concerns the way in which the minimax risk depends on the parameters $b, \mathbf{L}$. To the best of our knowledge, this question has not yet been answered in the literature.

In fact, a careful reading of the proof of [11, Theorem 2] allows to strengthen Lepski's result into an asymptotic lower bound matching the upper bound of Corollary 3, proving the ℓ-estimator's optimality up to a constant depending only on $\boldsymbol{\beta}, d$.

Theorem 6. *Let $\boldsymbol{\beta} \in \mathbb{R}_+^d$ and $\frac{1}{\beta} = \frac{1}{d}\sum_{j=1}^d \frac{1}{\beta_j}$.*

Let p_b denote the isotropic, centered Gaussian pdf with norm $\|p_b\|_{\infty,\mu} = b$. For all $L \in \mathbb{R}_+^d$ and all $b > 0$ such that $p_b \in C^{\boldsymbol{\beta}}_{\frac{\mathbf{L}}{2},+\infty}$ and all large enough n,

$$\inf_{\tilde{p}} \sup_{p^\star \in \mathcal{P}^{\boldsymbol{\beta}}_{\mathbf{L},b}} \mathbb{E}\left[\|\tilde{p} - p^\star\|_{\infty,\mu}\right] \geq C b^{\frac{\beta}{2\beta+d}} \left(\prod_{j=1}^d L_j^{\frac{\beta}{\beta_j}}\right)^{\frac{1}{2\beta+d}} \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}},$$

where the infimum runs over all estimators computed from an $n-$sample drawn from $p^\star$, and C is a constand depending only on $\boldsymbol{\beta}, d$.

Proof. The proof is based on that of Lepski [11, Theorem 2]. It can be found in appendix C.3. □

Thus, the $\ell-$estimator $\hat{p}$ adapts not only to the smoothness $\boldsymbol{\beta}$ but also to the size of the semi-norms $\left(|p^\star|_{j,\beta_j}\right)_{1 \leq j \leq d}$ and of the norm $\|p^\star\|_{\infty,\mu}$. This property has not, to the best of our knowledge, been established for any estimator for density estimation in sup-norm, though such a result was known in the setting of white noise regression on $[0,1)^d$ under a Hölder regularity assumption [5].

Appendix A. Estimation on a single model: proofs

A.1. **Proof of proposition 2.** Let $u = \sqrt{\frac{\Gamma + x}{n}}$, $\mathcal{C}_u = \{C \in \mathcal{C} : P^\star(C) \geq u^2\}$. For any measurable t, let

$$\hat{R}_n(t) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i t(X_i),$$

where $\varepsilon_1, \dots, \varepsilon_n$ are iid Rademacher random variables independent from the sample. Let

$$\begin{aligned} Z_1(\mathcal{C}) &= \sup_{C\in\mathcal{C}_u} \frac{1}{n\sqrt{P^\star(C)}}\left|\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i^\star(C)\right| \\ \bar{Z}_1(\mathcal{C}) &= \sup_{C\in\mathcal{C}_u} \frac{|\hat{R}_n(C)|}{\sqrt{P^\star(C)}} \\ Z_2(\mathcal{C}) &= \frac{1}{n}\sup_{C\in\mathcal{C}\backslash\mathcal{C}_u}\left|\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i^\star(C)\right|. \end{aligned}$$

First consider $Z_2(\mathcal{C})$. For any $C \in \mathcal{C}\backslash\mathcal{C}_u$, $P^\star(C) \le u^2$ by definition. Hence, by [2, Theorem 3],

$$\mathbb{E}[Z_2(\mathcal{C})] \le 2u\sqrt{\frac{2\Gamma}{n}} + 8\frac{\Gamma}{n} \le \frac{2}{n}\left[\sqrt{2\Gamma(\Gamma+x)} + 4\Gamma\right].$$

By Bousquet's inequality 7, with probability greater than $1-e^{-x}$, for all $\theta > 0$,

$$\begin{aligned} Z_2(\mathcal{C}) &\le \frac{1+2\theta}{n}\left[2\sqrt{2\Gamma(\Gamma+x)} + 8\Gamma\right] + 2u\sqrt{\frac{2x}{n}} + \left(2+\frac{4}{\theta}\right)\frac{x}{n}, \\ &= \frac{1}{n}\left[(1+2\theta)\left(8\Gamma + 2\sqrt{2\Gamma(\Gamma+x)}\right) + 2\sqrt{2x(\Gamma+x)} + x\left(2+\frac{4}{\theta}\right)\right] \\ &\le \frac{1}{n}\left[(1+2\theta)\left(11\Gamma + x\right) + \Gamma + 3x + x\left(2+\frac{4}{\theta}\right)\right] \\ &= \frac{1}{n}\left[\Gamma\left(11(1+2\theta)+1\right) + x\left(6+2\theta+\frac{4}{\theta}\right)\right]. \end{aligned}$$

Solving the quadratic equation $11(1+2\theta)+1 = 6+2\theta+\frac{4}{\theta}$ yields $\theta = \frac{\sqrt{89}-3}{20} \approx 0.3217$ and

$$Z_2(\mathcal{C}) \le 20\frac{\Gamma+x}{n}. \tag{19}$$

Consider now $Z_1(\mathcal{C})$. For any $j \in \mathbb{N}$, let

$$\mathcal{C}_{u,j} = \{C \in \mathcal{C} : 2^j u^2 \le P^\star(C) \le 2^{j+1}u^2\}.$$

Note that $\mathcal{C}_{u,j}$ is empty for any $j \ge \lceil -2\log_2 u\rceil$, in particular for any $j \ge \lceil \log_2 n\rceil$. Let also

$$\xi_{u,j}(\mathbf{X}) = \left\{A \subset \{1,\dots,n\} : \exists C \in \mathcal{C}_{u,j}, A = \{i : X_i \in C\}\right\}.$$

Conditioning on the sample, we have that

$$\begin{aligned} \mathbb{E}_\varepsilon\left[\bar{Z}_1(\mathcal{C})\right] &= \mathbb{E}_\varepsilon\left[\max_{j=0,\dots,\lceil\log_2 n\rceil-1}\sup_{C\in\mathcal{C}_{u,j}}\frac{1}{n\sqrt{P^\star(C)}}\left|\sum_{i=1}^n \varepsilon_i \mathbb{1}_C(X_i)\right|\right] \\ &\le \mathbb{E}_\varepsilon\left[\max_{\sigma\in\{-1,1\}}\max_{j=0,\dots,\lceil\log_2 n\rceil-1}\max_{A\in\xi_{u,j}(\mathbf{X})}\frac{\sigma}{n2^{j/2}u}\sum_{i\in A}\varepsilon_i\right]. \end{aligned}$$

By Sauer's lemma, $|\xi_{u,j}(\mathbf{X})| \leq \sum_{k=0}^{V\wedge n}\binom{n}{k}$, hence

$$\log\left(2\sum_{j=1}^{\lceil \log_2 n\rceil}|\xi_{u,j-1}(\mathbf{X})|\right) \leq \log\left(\lceil \log_2 n\rceil\right) + \log\left(2\sum_{k=0}^{V\wedge n}\binom{n}{k}\right) \leq \Gamma. \tag{20}$$

Let

$$\hat{S} = \sup_{C\in\mathcal{C}_u}\left\{\frac{1}{n^2P^\star(C)}\sum_{i=1}^n \mathbb{1}_C(X_i)\right\}.$$

For any j and $A \in \xi_{u,j}(\mathbf{X})$, by definition, there is some $C \in \mathcal{C}_{u,j}$ such that $A = \{i : \mathbb{1}_C(X_i) = 1\}$. By Hoeffding's inequality [6, Section 2.6], the random variables $\frac{\sigma}{n2^{j/2}u}\sum_{i\in A}\varepsilon_i$ are sub-Gaussian with variance factor

$$\begin{aligned}\frac{|A|}{2^ju^2n^2} &\leq \frac{2}{n^2P^\star(C)}\sum_{i=1}^n \mathbb{1}_C(X_i)\\ &\leq 2\hat{S}.\end{aligned}$$

It follows by [6, Section 2.5] and equation (20) that

$$\mathbb{E}_\varepsilon\left[\bar{Z}_1(\mathcal{C})\right] \leq \sqrt{4\Gamma\hat{S}}, \tag{21}$$

hence $\mathbb{E}\left[\bar{Z}_1(\mathcal{C})\right] \leq 2\sqrt{\Gamma\mathbb{E}[\hat{S}]}$. On the other hand, since $P^\star(C) \geq u^2$ for any $C \in \mathcal{C}_u$,

$$\begin{aligned}\hat{S} &= \sup_{C\in\mathcal{C}_u}\left\{\frac{1}{n^2P^\star(C)}\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i^\star(C)\right\} + \frac{1}{n}\\ &\leq \frac{1}{nu}\sup_{C\in\mathcal{C}_u}\left\{\frac{1}{n\sqrt{P^\star(C)}}\left|\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i^\star(C)\right|\right\} + \frac{1}{n}.\end{aligned}$$

It follows by the symmetrization inequality that

$$\mathbb{E}\left[\hat{S}\right] \leq \frac{2}{nu}\mathbb{E}\left[\bar{Z}_1(\mathcal{C})\right] + \frac{1}{n}.$$

Thus, by equation (21),

$$\mathbb{E}\left[\bar{Z}_1(\mathcal{C})\right] \leq 2\sqrt{\Gamma\left(\frac{2}{nu}\mathbb{E}\left[\bar{Z}_1(\mathcal{C})\right] + \frac{1}{n}\right)}.$$

Together with symmetrization, solving this quadratic inequality yields

$$\mathbb{E}\left[Z_1(\mathcal{C})\right] \leq 2\mathbb{E}\left[\bar{Z}_1(\mathcal{C})\right] \leq \frac{16\Gamma}{nu} + 4\sqrt{\frac{\Gamma}{n}} \leq \frac{16\Gamma}{\sqrt{(\Gamma+x)n}} + 4\sqrt{\frac{\Gamma}{n}}.$$

By construction, for any $C \in \mathcal{C}_u$, $\frac{\mathbb{1}_C(X_i)}{n\sqrt{P^\star(C)}} \in \left(0, \frac{1}{nu}\right)$ for all i, moreover

$$\sum_{i=1}^n \mathrm{Var}\left(\frac{\mathbb{1}_C(X_i)}{n\sqrt{P^\star(C)}}\right) \leq \frac{\sum_{i=1}^n P_i^\star(C)}{n^2P^\star(C)} = \frac{1}{n}.$$

Hence, by Bousquet's inequality 7, for any $\theta > 0$, with probability greater than $1 - e^{-x}$,

$$\begin{aligned}
Z_1(\mathcal{C}) &\le (1+2\theta)\left(\frac{16\Gamma}{\sqrt{(\Gamma+x)n}} + 4\sqrt{\frac{\Gamma}{n}}\right) + 2\sqrt{\frac{2x}{n}} + \left(2+\frac{4}{\theta}\right)\frac{x}{\sqrt{(\Gamma+x)n}} \\
&= \frac{1}{\sqrt{(\Gamma+x)n}}\left[(1+2\theta)\left(16\Gamma + 4\sqrt{\Gamma(\Gamma+x)}\right) + 2\sqrt{2x(\Gamma+x)} + \left(2+\frac{4}{\theta}\right)x\right] \\
&\le \frac{1}{\sqrt{(\Gamma+x)n}}\left[(1+2\theta)\left(20\Gamma+2x\right) + 3x + \Gamma + \left(2+\frac{4}{\theta}\right)x\right] \\
&= \frac{\left(20(1+2\theta)+1\right)\Gamma + \left(7+4\theta+\frac{4}{\theta}\right)x}{\sqrt{(\Gamma+x)n}}
\end{aligned}$$

Solving the quadratic equation $20(1+2\theta)+1 = 7+4\theta+\frac{4}{\theta}$ yields $\theta = \frac{\sqrt{193}-7}{36} \approx 0.19146$ and $20(1+2\theta)+1 \le 29$. Hence, with probability greater than $1 - e^{-x}$,

$$Z_1(\mathcal{C}) \le 29\sqrt{\frac{\Gamma+x}{n}}. \tag{22}$$

To conclude, consider the event E_x on which equations (19) and (22) both hold. By the union bound, $\mathbb{P}(E_x) \ge 1 - 2e^{-x}$. On E_x, for any $C \in \mathcal{C}$,

- If $C \in \mathcal{C}_u$, then
$$\frac{1}{n}\left|\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i(C)\right| \le \sqrt{P^\star(C)}Z_1(\mathcal{C}) \le 29\sqrt{P^\star(C)}\sqrt{\frac{\Gamma+x}{n}}.$$
- If $C \notin \mathcal{C}_u$, then
$$\frac{1}{n}\left|\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i(C)\right| \le Z_2(\mathcal{C}) \le 20\frac{\Gamma+x}{n}.$$

Thus, in all cases, on E_x,

$$\frac{1}{n}\left|\sum_{i=1}^n \mathbb{1}_C(X_i) - P_i(C)\right| \le \max\left(29\sqrt{P^\star(C)}\sqrt{\frac{\Gamma+x}{n}}, 20\frac{\Gamma+x}{n}\right).$$

A.2. **Proof of Proposition 4.** Fix $f \in m$ and $h > 0$. To simplify notations, let $\theta = \theta_m(h)$, $I_* = I_*(f)$, $x_* = x_*(f)$ and $C = C_{h,m}(f) = (1-\theta)x_* + \theta\mathring{I}_*$.

By assumption, $C \in \mathcal{C}$, moreover by convexity of I_*, $C \subset \mathring{I}_*$ and $\mu(I_*/\mathring{I}_*) = 0$, hence

$$\mu(C) = \theta^d\mu(I_*) \ge \theta^d h_0.$$

Let $x \in C$. By definition, this means that there exists $y \in \mathring{I}_*$ such that $x = \theta y + (1-\theta)x_*$ or in other words, $y = x_* + \frac{x - x_*}{\theta} \in \mathring{I}_*$. f coincides on I_*

with a polynomial f_* with total degree $\deg(f_*) \le r$. For any $u \in (0;1]$, let

$$g(u) = f_*\left(x_* + \frac{u(x - x_*)}{\theta}\right)$$

For any $u \in (0;1)$, $(1-u)x_* + uy = x_* + \frac{u(x-x_*)}{\theta} \in \mathring{I}_*$, hence

$$|g(u)| = \left| f\left(x_* + \frac{u(x - x_*)}{\theta}\right)\right| \le \|f\|_{\infty,\mu}.$$

Assume without loss of generality that $\|f\|_{\infty,\mu} = f_*(x_*) \ge 0$. Hence, by Markov's inequality [8, Theorem 1.4],

$$\begin{aligned} f_*(x) - f_*(x_*) &= g(\theta) - g(0) \\ &\le \theta \sup_{u\in[0,1]} |g'(u)| \\ &\le 2r^2\theta \sup_{u\in[0,1]} |g(u)| \\ &\le 2r^2\theta \|f\|_{\infty,\mu}. \end{aligned}$$

By definition of x_*, this yields

$$f(x) = f_*(x) \ge (1 - 2r^2\theta) \|f\|_{\infty,\mu}$$

for all $x \in C$. Thus,

$$\begin{aligned} \frac{|\int_C f d\mu|}{\mu(C) + h} &\ge \frac{\mu(C)\left(1 - 2r^2\theta\right)}{\mu(C) + h} \|f\|_{\infty,\mu} \\ &\ge \frac{\theta^d h_0 \left(1 - 2r^2\theta\right)}{\theta^d h_0 + h} \|f\|_{\infty,\mu}. \end{aligned}$$

First, consider the case $\theta = \frac{d}{d+1}\frac{1}{2r^2}$, where

$$\frac{\theta^d h_0 \left(1 - 2r^2\theta\right)}{\theta^d h_0 + h} = \frac{1}{d+1}\frac{\theta^d h_0}{\theta^d h_0 + h}.$$

If $h \ge \frac{h_0}{2^d r^{2d}}$, then

$$\theta^d h_0 = \left(\frac{d}{d+1}\right)^d \frac{h_0}{2^d r^{2d}} \le h,$$

which implies that

$$\begin{aligned} \frac{\theta^d h_0 \left(1 - 2r^2\theta\right)}{\theta^d h_0 + h} &\ge \frac{1}{d+1}\frac{\theta^d h_0}{2h} \\ &\ge \frac{1}{d+1}\frac{1}{2^{d+1} r^{2d}}\frac{h_0}{h} \\ &\ge \gamma_{r,d}\left(\frac{h}{h_0}\right). \end{aligned}$$

If $h \le h_1 = \frac{h_0}{2^d r^{2d}}$, then

$$\frac{|\int_C f d\mu|}{\mu(C) + h} \ge \frac{|\int_C f d\mu|}{\mu(C) + h_1} \ge \gamma_{r,d}\left(\frac{h_1}{h_0}\right) = \frac{1}{2(d+1)} \ge \gamma_{r,d}\left(\frac{h}{h_0}\right).$$

Assume now that we are in the second case: $\theta = \left(\frac{h}{2r^2h_0}\right)^{\frac{1}{d+1}}$. Then

$$
\begin{aligned}
\frac{\theta^d h_0\left(1-2r^2\theta\right)}{\theta^d h_0 + h} &= \left(1-(2r^2)^{\frac{d}{d+1}}\left(\frac{h}{h_0}\right)^{\frac{1}{d+1}}\right)\left(1-\frac{h}{h+\left(\frac{h}{2r^2h_0}\right)^{\frac{d}{d+1}}h_0}\right)\\
&\geq \left(1-(2r^2)^{\frac{d}{d+1}}\left(\frac{h}{h_0}\right)^{\frac{1}{d+1}}\right)\left(1-\frac{1}{1+\left(\frac{1}{2r^2}\right)^{\frac{d}{d+1}}\left(\frac{h_0}{h}\right)^{\frac{1}{d+1}}}\right)\\
&\geq \left(1-(2r^2)^{\frac{d}{d+1}}\left(\frac{h}{h_0}\right)^{\frac{1}{d+1}}\right)^2\\
&\geq \gamma_{r,d}\left(\frac{h}{h_0}\right).
\end{aligned}
$$

A.3. **Proof of Corollary 1.** By proposition 6, the sets $C_{h_m,m}$ satisfy equation (2) with

$$1-\varepsilon = \frac{\gamma_{r,d}\left(\frac{h_m}{h_0}\right)}{\kappa_m(h_m)} = \frac{1}{2\kappa_m(h_m)}.$$

By Theorem 1, with probability greater than $1-e^{-x}$,

$$\left\|\hat{p}_m^{(h_m)}-p^\star\right\|_{\infty,\mu} \leq 5\min_{p\in m}\{\|p-p^\star\|_{\infty,\mu}\}+2\max\left(58\sqrt{\frac{|p^\star|_{h_m}(\Gamma_1+x)}{h_m n}}, 40\frac{\Gamma_1+x}{h_m n}\right)+2\delta,$$

where $\Gamma_1 = \Gamma + \log 2$. It follows that

$$\mathbb{E}\left[\left\|\hat{p}_m^{(h_m)}-p^\star\right\|_{\infty,\mu}\right] \leq 5\min_{p\in m}\{\|p-p^\star\|_{\infty,\mu}\}+116\sqrt{\frac{|p^\star|_{h_m}(\Gamma_1+1)}{h_m n}}+80\frac{\Gamma_1+1}{h_m n}+2\delta.$$

The collection $\mathcal{C}$ has VC-dimension at most $2d$, as can be easily proved by considering a subset of $2d$ points with extremal coordinates. Hence, for all $n \geq 2d$,

$$
\begin{aligned}
\Gamma_1 + 1 &\leq 1+\log 2+\log\left(\lceil \log_2 n\rceil\right)+\log\left(2\sum_{j=0}^{2d}\binom{n}{j}\right)\\
&\leq 1+2\log 2+\log\left(\lceil \log_2 n\rceil\right)+2d\log\left(\frac{en}{2d}\right)\\
&\leq (2d+1)\log(en).
\end{aligned}
$$

Remark also that $|p^\star|_{h_m} \leq \min\left(\|p^\star\|_{\infty,\mu}, \frac{1}{h_m}\right)$, which yields

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{p}_m^{(h_m)}-p^\star\right\|_{\infty,\mu}\right] \leq{}& 5\min_{p\in m}\{\|p-p^\star\|_{\infty,\mu}\} + 116\sqrt{2d+1}\min\left(\sqrt{\|p^\star\|_{\infty,\mu}}\sqrt{\frac{\log en}{h_m n}}, \frac{\sqrt{\log en}}{h_m\sqrt{n}}\right)\\
&+ 80\frac{(2d+1)\log en}{h_m n} + 2\delta.
\end{aligned}
$$

Using the definition of h_m (equation (12)) together with the inequalities $\sqrt{\frac{2}{1-\frac{1}{\sqrt{2}}}} \leq 3$, $\frac{116}{\sqrt{1-\frac{1}{\sqrt{2}}}} \leq 215$ and $\frac{80}{1-\frac{1}{\sqrt{2}}} \leq 274$ yields the result.

A.4. **Proof of Theorem 2.** Any probability density $p \in m_L(0, \mathcal{I})$ is necessarily supported on $\mathcal{X}_0$. We can therefore assume that $\mathcal{X} = \mathcal{X}_0$, or equivalently that all blocks of $\mathcal{I}$ have finite measure.

Fix some $h \in \left(0, \frac{1}{L}\right]$. Let

$$\mathcal{I}_h = \{I \in \mathcal{I} : \mu(I) \leq h\}.$$

Let also

$$M = M(h) \wedge \left\lfloor \frac{1}{Lh} \right\rfloor.$$

If $M = 0$, the result is trivial. Assume now that $M \geq 1$, which implies that $h \leq \frac{1}{L}$.

Let $\mathcal{I}^0 \subset \mathcal{I}_h$, a subset with cardinality $M \geq 1$, which exists since $M(h) \geq M$. Let $J_0 = \cup \mathcal{I}^0$. By the assumptions on L, h,

$$\begin{aligned}
\sum_{I' \notin \mathcal{I}^0} \mu(I') &= \mu(\mathcal{X}) - \mu(J_0) \\
&\geq \frac{1}{\theta L} - \mu(J_0) \\
&\geq \frac{1}{\theta L} - Mh \\
&\geq \frac{1}{\theta L} - \frac{1}{L} > 0.
\end{aligned}$$

Let $\mathcal{I}^1 \subset \mathcal{I} \backslash \mathcal{I}^0$ and $J = \cup \mathcal{I}^1$ such that

$$0 < \frac{1}{\theta L} - \mu(J_0) \leq \mu(J) < +\infty.$$

Let $x \in \left(0, \frac{1}{\theta} - 1\right)$ to be specificed later, and define finally, for any $I \in \mathcal{I}^0$

$$p_I = (1+x)\theta L \mathbb{1}_I + \theta L \mathbb{1}_{J_0 \backslash I} + \frac{1 - \theta L \mu(J_0) - \theta L x \mu(I)}{\mu(J)} \mathbb{1}_J \tag{23}$$

$$p_0 = \theta L \mathbb{1}_{J_0} + \frac{1 - \theta L \mu(J_0)}{\mu(J)} \mathbb{1}_J. \tag{24}$$

Since $|\mathcal{I}^0| = M \leq \frac{1}{Lh}$ and $\mathcal{I}^0 \subset \mathcal{I}_h$,

$$1 - \theta L \mu(J_0) - \theta L x \mu(I) \geq 1 - (1+x)\theta L \mu(J_0) \geq 1 - LMh \geq 0.$$

This proves that the $(p_I)_{I\in\mathcal{I}^0}$ are probability densities, and p_0 a positive probability density. Moreover,

$$\begin{aligned}(1+x)\theta L &\le L\\ \frac{1-\theta L\mu(J_0)}{\mu(J)} &\le \frac{1-\theta L\mu(J_0)}{\frac{1}{\theta L}-\mu(J_0)}\\ &\le \theta L,\end{aligned}$$

which implies that p_0 and the p_I belong to $m_L(0,\mathcal{I})$.

The minimum distance between p_I and p_0 in sup-norm is

$$\min_{I\in\mathcal{I}^0}\|p_I-p_0\|_{\infty,\mu}\ge x\theta L. \tag{25}$$

The likelihood ratio between p_I and p_0 is

$$\frac{p_I}{p_0}=(1+x)\mathbb{1}_I+\mathbb{1}_{J_0\setminus I}+\left(1-\frac{x\theta L\mu(I)}{1-\theta L\mu(J_0)}\right)\mathbb{1}_J.$$

Hence, the chi-squared divergence is

$$\begin{aligned}\chi^2(P_I,P_0)&=(1+x)^2\theta L\mu(I)+\theta L[\mu(J_0)-\mu(I)]+\left(1-\frac{x\theta L\mu(I)}{1-\theta L\mu(J_0)}\right)^2[1-\theta L\mu(J_0)]-1\\ &=(1+2x+x^2)\theta L\mu(I)+\theta L[\mu(J_0)-\mu(I)]-1\\ &\quad+\left(1-\frac{2x\theta L\mu(I)}{1-\theta L\mu(J_0)}+\frac{(x\theta L\mu(I))^2}{(1-\theta L\mu(J_0))^2}\right)[1-\theta L\mu(J_0)]\\ &=x^2\theta L\mu(I)\left(1+\frac{\theta L\mu(I)}{1-\theta L\mu(J_0)}\right).\end{aligned}$$

Since $\mathcal{I}^0\subset\mathcal{I}_h$ and $|\mathcal{I}^0|=M$,

$$\frac{\theta L\mu(I)}{1-\theta L\mu(J_0)}\le\frac{\theta Lh}{1-\theta LMh}.$$

Since by assumption, $Lh\le 1$ and $M\le\frac{1}{Lh}$,

$$\frac{\theta L\mu(I)}{1-\theta L\mu(J_0)}\le\frac{\theta Lh}{1-\theta LMh}\le\frac{\theta}{1-\theta}.$$

It follows that, for any $I\in\mathcal{I}_0$,

$$\chi^2(P_I,P_0)\le\frac{x^2\theta L\mu(I)}{1-\theta}.$$

The KL-divergence between the distributions of two iid samples of size n, drawn respectively from P_I and P_0, is

$$\mathrm{KL}\left(P_I^{\otimes n},P_0^{\otimes n}\right)=n\mathrm{KL}\left(P_I,P_0\right)\le n\chi^2(P_I,P_0)\le\frac{nx^2\theta Lh}{1-\theta}. \tag{26}$$

Consider first the case $M=1$. Let I be the single element of $\mathcal{I}_0$, $\alpha>0$ and

$$x=\min\left(\frac{1}{\theta}-1,\sqrt{\frac{\alpha(1-\theta)}{\theta Lhn}}\right).$$

Then, by equation (26), $\mathrm{KL}\left(P_I^{\otimes n}, P_0^{\otimes n}\right) \leq \alpha$, moreover by equations (23), (24),

$$\|p_I - p_0\|_{\infty,\mu} \geq x\theta L \geq \min\left((1-\theta)L, \sqrt{\frac{\alpha\theta(1-\theta)L}{hn}}\right) := s_\alpha.$$

By [17, Theorem 2.2] and the following equation (2.9), for any density estimator $\hat{p}$ based on an $n-$sample,

$$\max_{P\in\{P_0,P_I\}} P^{\otimes n}\left(\|\hat{p}-p\|_{\infty,\mu} \geq \frac{s_\alpha}{2}\right) \geq \min\left(\frac{e^{-\alpha}}{4}, \frac{1-\sqrt{\frac{\alpha}{2}}}{2}\right).$$

Choosing $\alpha = \frac{1}{2}$, this yields

$$\begin{aligned}\inf_{\hat{p}} \sup_{p\in m_L(0,\mathcal{I})} \mathbb{E}\left[\|\hat{p}-p\|_{\infty,\mu}\right] &\geq 0.075 \times \min\left((1-\theta)L, 0.7\sqrt{\frac{\theta(1-\theta)L}{hn}}\right) \\ &\geq 0.075 \times \min\left((1-\theta)L, \sqrt{\frac{\theta(1-\theta)L}{hn}}\sqrt{\log(M+1)}\right).\end{aligned}$$

Consider now the case $M \geq 2$. Let $\alpha \in \left(0, \frac{1}{8}\right)$ and

$$x = \min\left(\frac{1}{\theta} - 1, \sqrt{\frac{\alpha(1-\theta)\log M}{\theta L h n}}\right).$$

By equation (26), for any $I \in \mathcal{I}_0$, $\mathrm{KL}\left(P_I^{\otimes n}, P_0^{\otimes n}\right) \leq \alpha \log M$. Moreover, for any distinct $I, I' \in \mathcal{I}_0$, by equation (23),

$$\|p_I - p_{I'}\|_{\infty,\mu} \geq x\theta L \geq \min\left((1-\theta)L, \sqrt{\frac{\alpha\theta(1-\theta)L}{hn}}\sqrt{\log M}\right) := t_\alpha.$$

Hence, by [17, Theorem 2.5], for all density estimators $\hat{p}$ based on an $n-$sample from P,

$$\sup_{p\in m_L(0,\mathcal{I})} P^{\otimes n}\left(\|\hat{p}-p\|_{\infty,\mu} \geq \frac{t_\alpha}{2}\right) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right).$$

Let $\alpha = \frac{37}{800}$. Since $M \geq 2$, it follows that

$$\sup_{p\in m_L(0,\mathcal{I})} P^{\otimes n}\left(\|\hat{p}-p\|_{\infty,\mu} \geq \frac{t_\alpha}{2}\right) \geq \frac{\sqrt{2}}{1+\sqrt{2}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log 2}}\right) \geq 0.317,$$

which implies that

$$\begin{aligned}\sup_{p\in m_L(0,\mathcal{I})} \mathbb{E}\left[\|\hat{p}-p\|_{\infty,\mu}\right] &\geq 0.158 \times \min\left((1-\theta)L, 0.21\sqrt{\frac{\theta(1-\theta)L}{hn}}\sqrt{\log M}\right) \\ &\geq 0.158 \times \min\left((1-\theta)L, 0.166\sqrt{\frac{\theta(1-\theta)L}{hn}}\sqrt{\log(M+1)}\right).\end{aligned}$$

This concludes the proof.

Appendix B. Model selection and adaptivity: Proofs

B.1. **Proof of Theorem 3.** Begin with the following proposition:

Proposition 7. *On Ω_x, for any $p, q \in \mathscr{P}$ and any $h > 0$,*

$$|T^{(h)}(\mathbf{X}, p, q) - \Delta_h(p, q)| \le Z(h) \le \mathrm{pen}_a(h) + 29\sqrt{\frac{4a}{3n}} g_1\left(\frac{x}{a}\right) + 29^2\sqrt{\frac{4}{3}}\frac{a}{n} g_2\left(\frac{x}{a}\right),$$

where

$$\begin{aligned} g_1(t) &= \sup_{u \ge 0}\left\{u\left(\sqrt{t} - \sqrt{\log_+ u}\right)\right\} \\ g_2(t) &= \sup_{u \ge 0}\left\{u\left(t - \log_+ u\right)\right\}. \end{aligned}$$

Proof. The first inequality is true by definition of $Z(h)$. On Ω_x, by Theorem 1 and proposition 5 with $\theta = \frac{1}{2}$,

$$\begin{aligned} Z(h) &\le \max\left(29\sqrt{\frac{|p^\star|_h(\Gamma + x)}{hn}}, 20\frac{\Gamma + x}{hn}\right) \\ &\le \max\left(29\sqrt{\frac{\Gamma + x}{hn}}\left[\frac{\sqrt{|\hat{p}|_h}}{\sqrt{1 - \frac{\theta}{2}}} + \frac{29}{\sqrt{\theta(2 - \theta)}}\sqrt{\frac{\Gamma + x}{hn}}\right], 20\frac{\Gamma + x}{hn}\right) \\ &\le \max\left(29\sqrt{\frac{4}{3}}\sqrt{\frac{|\hat{p}|_h(\Gamma + x)}{hn}} + \sqrt{\frac{4}{3}}29^2\frac{\Gamma + x}{hn}, 20\frac{\Gamma + x}{hn}\right) \\ &\le 29\sqrt{\frac{4}{3}}\sqrt{\frac{|\hat{p}|_h(\Gamma + x)}{hn}} + \sqrt{\frac{4}{3}}29^2\frac{\Gamma + x}{hn}. \end{aligned}$$

By equation (14) defining pen, it follows that on Ω_x,

$$\begin{aligned} Z(h) - \mathrm{pen}_a(h) &\le 29\sqrt{\frac{4}{3}}\sqrt{\frac{|\hat{p}|_h}{hn}}\left(\sqrt{\Gamma + x} - \sqrt{\Gamma + a\log_- h}\right) + \sqrt{\frac{4}{3}}\frac{29^2}{hn}(x - a\log_- h) \\ &\le 29\sqrt{\frac{4}{3}}\sqrt{\frac{|\hat{p}|_h}{hn}}\left(\sqrt{x} - \sqrt{a\log_- h}\right) + \sqrt{\frac{4}{3}}\frac{29^2}{hn}(x - a\log_- h). \end{aligned}$$

Note that for any $h > 0$,

$$|\hat{p}|_h = \sup_{C \in \mathcal{C}} \frac{\sum_{i=1}^n \mathbb{1}_C(X_i)}{n(\mu(C) + h)} \le \frac{1}{h}.$$

It follows that

$$\begin{aligned} Z(h) - \mathrm{pen}_a(h) &\le 29\sqrt{\frac{4a}{3n}}\frac{1}{h}\left(\sqrt{\frac{x}{a}} - \sqrt{\log_+\left(\frac{1}{h}\right)}\right) + \sqrt{\frac{4}{3}}29^2\frac{a}{n}\frac{1}{h}\left(\frac{x}{a} - \log_+\left(\frac{1}{h}\right)\right) \\ &\le 29\sqrt{\frac{4a}{3n}}g_1\left(\frac{x}{a}\right) + \sqrt{\frac{4}{3}}29^2\frac{a}{n}g_2\left(\frac{x}{a}\right). \end{aligned}$$

□

Now, let us control the expected value of the test $T^{(h_p\wedge h_q)}(\mathbf{X}, p, q)$.

Lemma 2. *Under assumption 2,*

$$(1-\varepsilon)\kappa_* \|p - p^\star\|_{\infty,\mu} - [(1-\varepsilon)\kappa_* + 1]\|q - p^\star\|_{\infty,\mu} \le \Delta_{h_p\wedge h_q}(p, q) \le \|p - p^\star\|_{\infty,\mu}.$$

Proof. Fix $p, q \in \mathbf{M}$ and let $h = h_p \wedge h_q$. Clearly,

$$\Delta_h(p, q) = \varepsilon_h(p, q)\frac{(P - P^\star)(C_h(p, q))}{\mu(C_h(p, q)) + h} \le |p^\star - p|_h \le \|p - p^\star\|_{\infty,\mu}.$$

Fix $\delta' > 0$. Let now $m, m' \in \mathcal{M}$ be such that $p \in m, q \in m'$ and $h_m \ge (1-\delta')h_p, h_{m'} \ge (1-\delta')h_q$. By lemma 1 and assumption 2,

$$\kappa_{m\cup m'}(h_p \wedge h_q) \ge \kappa_{m\cup m'}(h_m \wedge h'_m) - \left|1 - \frac{h_m \wedge h_{m'}}{h_p \wedge h_q}\right| \ge \kappa_* - \delta'.$$

It follows that

$$\begin{aligned} \Delta_h(p, q) &= \varepsilon_h(p, q)\frac{(P - P^\star)(C_h(p, q))}{\mu(C_h(p, q)) + h} \\ &= \varepsilon_h(p, q)\frac{(P - Q)(C_h(p, q))}{\mu(C_h(p, q)) + h} + \varepsilon_h(p, q)\frac{(Q - P^\star)(C_h(p, q))}{\mu(C_h(p, q)) + h} \\ &\ge (1-\varepsilon)|p - q|_h - \|p^\star - q\|_{\infty,\mu} \\ &\ge (1-\varepsilon)\kappa_{m\cup m'}(h)\|p - q\|_{\infty,\mu} - \|q - p^\star\|_{\infty,\mu} \\ &\ge (1-\varepsilon)(\kappa_* - \delta)(\|p - p^\star\|_{\infty,\mu} - \|q - p^\star\|_{\infty,\mu}) - \|q - p^\star\|_{\infty,\mu} \\ &\ge (1-\varepsilon)(\kappa_* - \delta')\|p - p^\star\|_{\infty,\mu} - \left[1 + (1-\varepsilon)(\kappa_* - \delta')\right]\|q - p^\star\|_{\infty,\mu}. \end{aligned}$$

Since $\delta' > 0$ is arbitrary, this proves the result. □

We can now carry out the proof of the Theorem. First, note that since $h \mapsto |\hat{p}|_h$ is a non-increasing function of h, pen_a (equation (14)) is also a non-increasing function of h for any $a > 0$. Hence, for any $p, q \in \mathbf{M}$,

$$\mathrm{pen}_a(h_p \wedge h_q) = \max\left(\mathrm{pen}_a(h_p), \mathrm{pen}_a(h_q)\right). \tag{27}$$

Let $\overline{p} \in \mathbf{M}$. By definition of T, pen and lemma 2, for any $p, q \in \mathbf{M}$,

$$\begin{aligned} &T^{(h_p\wedge h_q)}(\mathbf{X}, p, q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_p) \\ &\quad= \Delta_{h_p\wedge h_q}(p, q) + T^{(h_p\wedge h_q)}(\mathbf{X}, p, q) - \Delta_{h_p\wedge h_q}(p, q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_p) \\ &\quad\ge \kappa_*(1-\varepsilon)\|p - p^\star\|_{\infty,\mu} - (1 + \kappa_*(1-\varepsilon))\|q - p^\star\|_{\infty,\mu} - Z(h_p \wedge h_q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_p). \end{aligned}$$

To simplify notation, let $\kappa = (1-\varepsilon)\kappa_*$ and

$$R(a,x) = 29\sqrt{\frac{4a}{3n}}g_1\left(\frac{x}{a}\right) + 29^2\sqrt{\frac{4}{3}}\frac{a}{n}g_2\left(\frac{x}{a}\right). \tag{28}$$

By proposition 7 and equation (27), on Ω_x, for all $p, q \in \mathbf{M}$,

$$\begin{aligned}
&T^{(h_p\wedge h_q)}(\mathbf{X}, p, q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_p) + R(a,x) \\
&\quad\geq \kappa\left\|p - p^\star\right\|_{\infty,\mu} - (1+\kappa)\left\|q - p^\star\right\|_{\infty,\mu} - Z(h_p \wedge h_q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_p) + R(a,x) \\
&\quad\geq \kappa\left\|p - p^\star\right\|_{\infty,\mu} - (1+\kappa)\left\|q - p^\star\right\|_{\infty,\mu} - \mathrm{pen}_a(h_p \wedge h_q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_p) \\
&\quad\geq \kappa\left\|p - p^\star\right\|_{\infty,\mu} - (1+\kappa)\left\|q - p^\star\right\|_{\infty,\mu} - 2\,\mathrm{pen}_a(h_q)
\end{aligned}$$

in light of equation (27). In particular, taking $p = \hat{p}_{\mathcal{M}}$ and $q = \overline{p}$ yields

$$T_{\mathcal{M}}(\mathbf{X}, \hat{p}_{\mathcal{M}}) \geq \kappa\left\|\hat{p}_{\mathcal{M}} - p^\star\right\|_{\infty,\mu} - (1+\kappa)\left\|\overline{p} - p^\star\right\|_{\infty,\mu} - 2\,\mathrm{pen}_a(h_{\overline{p}}) - R(a,x). \tag{29}$$

On the other hand, by lemma 2 for any $q \in \mathcal{M}$,

$$\begin{aligned}
&T^{(h_{\overline{p}}\wedge h_q)}(\mathbf{X}, \overline{p}, q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_{\overline{p}}) \\
&\quad= \Delta_{h_{\overline{p}}\wedge h_q}(\overline{p}, q) + T^{(h_{\overline{p}}\wedge h_q)}(\mathbf{X}, \overline{p}, q) - \Delta_{h_{\overline{p}}\wedge h_q}(p, q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_{\overline{p}}) \\
&\quad\leq \left\|\overline{p} - p^\star\right\|_{\infty,\mu} + Z(h_{\overline{p}} \wedge h_q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_{\overline{p}}).
\end{aligned}$$

It follows by proposition 7 and equation (27) that on Ω_x, for all $q \in \mathbf{M}$,

$$\begin{aligned}
&T^{(h_{\overline{p}}\wedge h_q)}(\mathbf{X}, \overline{p}, q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_{\overline{p}}) \\
&\quad\leq \left\|\overline{p} - p^\star\right\|_{\infty,\mu} + \mathrm{pen}_a(h_{\overline{p}} \wedge h_q) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_{\overline{p}}) + R(a,x) \\
&\quad= \left\|\overline{p} - p^\star\right\|_{\infty,\mu} + \max(\mathrm{pen}_a(h_{\overline{p}}), \mathrm{pen}_a(h_q)) - \mathrm{pen}_a(h_q) + \mathrm{pen}_a(h_{\overline{p}}) + R(a,x) \\
&\quad\leq \left\|\overline{p} - p^\star\right\|_{\infty,\mu} + 2\,\mathrm{pen}_a(h_{\overline{p}}) + R(a,x).
\end{aligned}$$

Hence, by definition of $T_{\mathcal{M}}$, on Ω_x,

$$T_{\mathcal{M}}(\mathbf{X}, \overline{p}) \leq \left\|\overline{p} - p^\star\right\|_{\infty,\mu} + 2\,\mathrm{pen}_a(h_{\overline{p}}) + R(a,x). \tag{30}$$

Thus, by equations (29), (30) and definition of $\hat{p}_{\mathcal{M}}$,

$$\begin{aligned}
&\left\|\overline{p} - p^\star\right\|_{\infty,\mu} + 2\,\mathrm{pen}_a(h_{\overline{p}}) + R(a,x) + \delta \\
&\quad\geq T_{\mathcal{M}}(\mathbf{X}, \overline{p}) + \delta \\
&\quad\geq T_{\mathcal{M}}(\mathbf{X}, \hat{p}_{\mathcal{M}}) \\
&\quad\geq \kappa\left\|\hat{p}_{\mathcal{M}} - p^\star\right\|_{\infty,\mu} - (1+\kappa)\left\|\overline{p} - p^\star\right\|_{\infty,\mu} - 2\,\mathrm{pen}_a(h_{\overline{p}}) - R(a,x).
\end{aligned}$$

This yields

$$\kappa\left\|\hat{p}_{\mathcal{M}} - p^\star\right\|_{\infty,\mu} \leq (2+\kappa)\left\|\overline{p} - p^\star\right\|_{\infty,\mu} + 4\,\mathrm{pen}_a(h_{\overline{p}}) + 2R(a,x) + \delta.$$

on Ω_x. Since $\overline{p}$ was arbitrary, it follows that on Ω_x

$$\begin{aligned}\kappa \left\|\hat{p}_{\mathcal{M}} - p^\star\right\|_{\infty,\mu} &\leq \inf_{p\in\mathbf{M}} \Big\{(2+\kappa)\left\|p-p^\star\right\|_{\infty,\mu} + 4\,\mathrm{pen}_a(h_p)\Big\} + 2R(a,x) + \delta \\ &= \inf_{p\in\mathbf{M}} \left\{(2+\kappa)\left\|p-p^\star\right\|_{\infty,\mu} + 4 \inf_{m\in\mathcal{M}:p\in m}\{\mathrm{pen}_a(h_m)\}\right\} + 2R(a,x) + \delta \\ &= \inf_{(p,m)\in\mathbf{M}\times\mathcal{M}:p\in m} \Big\{(2+\kappa)\left\|p-p^\star\right\|_{\infty,\mu} + 4\,\mathrm{pen}_a(h_m)\Big\} + 2R(a,x) + \delta \\ &= \inf_{m\in\mathcal{M}} \left\{(2+\kappa)\inf_{p\in m}\{\left\|p-p^\star\right\|_{\infty,\mu}\} + 4\,\mathrm{pen}_a(h_m)\right\} + 2R(a,x) + \delta.\end{aligned}$$

Setting $x = a\log y$, the event Ω_x occurs with probability greater than $1-\frac{2}{y^a}$ by proposition 2. Moreover, by equation (28),

$$R(a, a\log y) = 29\sqrt{\frac{4a}{3n}}g_1(\log y) + 29^2\sqrt{\frac{4}{3}\frac{a}{n}}g_2(\log y). \tag{31}$$

It remains to bound $g_1(t), g_2(t)$, where $t = \log y \geq 0$. First,

$$\begin{aligned}g_1(t) &= \sup_{u\geq 0}\left\{u\left(\sqrt{t}-\sqrt{\log_+ u}\right)\right\} \\ &= \sup_{1\leq u\leq e^t}\left\{u\left(\sqrt{t}-\sqrt{\log_+ u}\right)\right\} \\ &= \sup_{\theta\in[0,1]}\left\{e^{(1-\theta)t}\sqrt{t}\big(1-\sqrt{1-\theta}\big)\right\} \\ &\leq \frac{e^t}{2\sqrt{t}}\sup_{\theta\in[0,1]}\{\theta t e^{-\theta t}\} \\ &= \frac{e^{t-1}}{2\sqrt{t}}\mathbb{1}_{t\geq 1} + \frac{\sqrt{t}}{2}\mathbb{1}_{t<1} \\ &\leq \frac{e^{-1}y}{2\sqrt{\log y}}\mathbb{1}_{\log y\geq 1} + \frac{\sqrt{\log y}}{2}\mathbb{1}_{\log y<1}.\end{aligned}$$

Using the inequality $ve^{-v} \leq e^{-1}$ with $v = 2\log y$ yields $\sqrt{\log y} \leq \frac{y}{\sqrt{2e}}$, hence

$$g_1(\log y) \leq \frac{y}{2\sqrt{2e}}.$$

Similarly,

$$\begin{aligned}g_2(t) &= \sup_{u\geq 0}\big\{u\left(t-\log_+ u\right)\big\} \\ &= \sup_{1\leq u\leq e^t}\big\{u\left(t-\log_+ u\right)\big\} \\ &= \sup_{\theta\in[0,1]}\left\{\theta t e^{(1-\theta)t}\right\} \\ &= e^t\sup_{\theta\in[0,1]}\left\{\theta t e^{-\theta t}\right\} \\ &\leq ye^{-1}.\end{aligned}$$

Substituting these bounds into equation (31) yields the final result.

B.2. **Properties of dyadic partitions.**

Lemma 3. *If I, J are two dyadic intervals, then one of the following alternatives hold:*

- $I \subset J$
- $J \subset I$
- $I \cap J = \emptyset$.

Proof. Assume that I, J intersect at x. Without loss of generality, assume that $x \geq 0$ and that I is the longer of the two intervals. Then $I = [k_1 2^{-j_1}, (k_1+1)2^{-j_1+1})$ and $J = [k_2 2^{-j_2}, (k_2+1)2^{-j_2+1})$ where $j_1 \leq j_2$. Since $x \in I \cap J$,

$$k_1 \leq 2^{j_1} x < k_1 + 1$$
$$k_2 \leq 2^{j_2} x < k_2 + 1,$$

which proves that $k_i = \lfloor 2^{j_i} x \rfloor$. Let

$$x = \sum_{i=-\infty}^{+\infty} \varepsilon_i 2^{-i},$$

where $\varepsilon \in \{0,1\}^{\mathbb{Z}}$ has finite support. Then for any $j \in \mathbb{Z}$,

$$\begin{aligned} 2^{-j} \lfloor 2^j x \rfloor &= 2^{-j} \left\lfloor \sum_{i=-\infty}^{+\infty} \varepsilon_i 2^{j-i} \right\rfloor \\ &= 2^{-j} \sum_{i=-\infty}^{j} \varepsilon_i 2^{j-i} \\ &= \sum_{i=-\infty}^{j} \varepsilon_i 2^{-i}. \end{aligned}$$

In particular,

$$0 \leq 2^{-j_2} \left\lfloor 2^{j_2} x \right\rfloor - 2^{-j_1} \lfloor 2^{j_1} x \rfloor \leq \sum_{i=j_1+1}^{j_2} \varepsilon_i 2^{-i} \leq 2^{-j_1} - 2^{-j_2},$$

which proves that $J \subset I$. □

The following lemma is easily deduced from the previous one.

Lemma 4. *For any $\mathbf{j} \in \mathbb{Z}^d$, $\mathcal{I}(\mathbf{j})$ partitions $\mathbb{R}^d$. Moreover, if $\mathbf{j} \leq \mathbf{j}'$, then $\mathcal{I}(\mathbf{j}')$ refines $\mathcal{I}(\mathbf{j})$.*

Proof. Let $x \in \mathbb{R}^d$. For any $i \in \{1, \ldots, d\}$, x_i belongs to some dyadic interval I_i with length 2^{-j_i}. Then $x \in \prod_{i=1}^d I_i \in \mathcal{I}(\mathbf{j})$. Moreover, if $I, J \in \mathcal{I}(\mathbf{j})$ intersect at x, then for any $i \in \{1, \ldots, d\}$, $x_i \in I_i \cap J_i$, which implies by the previous lemma that $I_i \subset J_i$ or $J_i \subset I_i$. Since $|I_i| = |J_i| = 2^{-j_i}$, $I_i = J_i$. Hence $I = J$, which proves that $\mathcal{I}(\mathbf{j})$ is a partition.

Let now $I' \in \mathcal{J}(\mathbf{j}')$ and let $x \in I'$. Let $I \in \mathcal{I}(\mathbf{j})$ contain x. For any $i \in \{1, \dots, d\}$, $x_i \in I_i \cap I'_i$, hence by the previous lemma and since $2^{-j'_i} = |I'_i| \leq 2^{-j_i} = |I_i|$, $I'_i \subset I_i$. Hence, $I' \subset I$, which proves that $\mathcal{I}(\mathbf{j}')$ refines $\mathcal{I}(\mathbf{j})$. $\square$

B.3. **Proof of Theorem 4.** The class

$$\mathcal{C}_{rec,0} = \{C \cap [0,1)^d : C \in \mathcal{C}_{rec}\}$$

generates the Borel sigma-algebra on $[0,1)^d$, hence the semi-norms $|\cdot|_h$ defined with $\mathcal{C} = \mathcal{C}_{rec,0}$ are norms on $L_\infty([0,1)^d)$.

Let $\mathcal{H}_d$ denote the completion of $L_\infty([0,1)^d)$ with respect to a norm $|\cdot|_h$ with $\mathcal{C} = \mathcal{C}_{rec,0}$. Since the norms $|\cdot|_h$ are equivalent, this space does not depend on the choice of h.

For any $\tau > 0$, $d \in \mathbb{N}$ and $\mathbf{r} \in \mathbb{R}^d$, fix a linear projection $R^{(\tau)}_{d,\mathbf{r}} : (\mathcal{H}_d, |\cdot|_\tau) \to \left(\mathcal{P}^{dir}_{\mathbf{r},d}, |\cdot|_\tau\right)$ with operator norm less than

$$c_d(\mathbf{r}) := \sqrt{\dim(\mathcal{P}^{dir}_{\mathbf{r},d})} = \sqrt{\prod_{i=1}^{d}(r_i+1)}.$$

The existence of such a projection is guaranteed by [8, Theorem 7.6]

For any $x \in \mathbb{R}^d$ and any $S \subset \{1, \dots, d\}$, let $x_S = (x_i)_{i \in S}$. Given a function $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \subset \mathbb{R}^d$, define the function $f_S : \mathbb{R}^{d-|S|} \times \mathbb{R}^{|S|} \to \mathbb{R}$ by $f_S(x, y) = f(z)$, where

$$z_i = \begin{cases} y_i \text{ if } i \in S \\ x_i \text{ if } i \notin S. \end{cases}$$

Given $S \subset \{1, \dots, d\}$, we can then define the operator $R^{(\tau)}_S$ equal to $R^{(\tau)}_{|S|,\mathbf{r}_S}$ "applied to the variables $(x_i)_{i\in S}$", i.e the operator defined by

$$\begin{aligned} R^{(\tau)}_S &: L^\infty([0,1)^d) \to L^\infty([0,1)^d) \\ R^{(\tau)}_S f(x) &= R^{(\tau)}_{|S|,\mathbf{r}_S}\left(y \mapsto f_S(x_{S^c}, y)\right)(x_S) \text{ a.e.} \end{aligned}$$

Lemma 5. *Define the function*

$$\kappa_{\mathbf{r},d}(h) = \inf_{f \in \mathcal{P}^{dir}_{\mathbf{r},d}} \frac{\left|f\mathbb{1}_{[0,1)^d}\right|_h}{\left\|f\mathbb{1}_{[0,1)^d}\right\|_{\infty,\mu}}.$$

Let $J \subset [0,1)^d$,

$$J = \prod_{i=1}^{d} J_i \in \mathcal{C}_{rec},$$

$S \subset \{1, \ldots, d\}$ and $v_S = \prod_{i \in S} \mu(J_i)$. For any $\tau \geq v_S, h \geq \mu(J)$ and any $f \in L^\infty\left([0,1]^d\right)$,

$$\frac{1}{h}\left|\int_J R_S^{(\tau)} f\right| \leq \frac{c_{|S|}(\mathbf{r}_S)}{\kappa_{\mathbf{r}_S,|S|}(\tau)}|f|_h.$$

Proof. Let $J_S = \prod_{i \in S} J_i$, $J_{S^c} = \prod_{i \notin S} J_i$ and $h_{S^c} = \frac{h}{v_S}$. Note that $\mu(J_S) = v_S$ and $\mu(J_{S^c}) \leq h_{S^c}$. Since $R_{|S|,\mathbf{r}_S}^{(\tau)}$ is a bounded linear operator, by Fubini's theorem,

$$\begin{aligned}
\frac{1}{h}\int_J R_S^{(\tau)} f &= \frac{1}{v_S}\frac{1}{h_{S^c}}\int_{J_S}\int_{J_{S^c}} R_{|S|,\mathbf{r}_S}^{(\tau)}\left[y \mapsto f_S(x_{S^c}, y)\right](x_S)dx_{S^c}dx_S \\
&= \frac{1}{v_S}\int_{J_S} R_{|S|,\mathbf{r}_S}^{(\tau)}\left[y \mapsto \frac{1}{h_{S^c}}\int_{J_{S^c}} f_S(x_{S^c}, y)dx_{S^c}\right](x_S)dx_S.
\end{aligned}$$

Let

$$\overline{f}_S : y \mapsto \frac{1}{h_{S^c}}\int_{J_{S^c}} f_S(x_{S^c}, y)dx_{S^c}.$$

Since $R_{|S|,\mathbf{r}_S}^{(\tau)}\overline{f}_S \in \mathcal{P}_{\mathbf{r}_S,|S|}^{dir}$,

$$\begin{aligned}
\frac{1}{v_S}\left|\int_{J_S} R_{|S|,\mathbf{r}_S}^{(\tau)}\overline{f}_S\right| &\leq \left\|R_{|S|,\mathbf{r}_S}^{(\tau)}\overline{f}_S\right\|_{\infty,\mu} \\
&\leq \frac{\left|R_{|S|,\mathbf{r}_S}^{(\tau)}\overline{f}_S\right|_\tau}{\kappa_{\mathbf{r}_S,|S|}(\tau)} \\
&\leq \frac{c_{|S|}(\mathbf{r}_S)}{\kappa_{\mathbf{r}_S,|S|}(\tau)}\left|\overline{f}_S\right|_\tau \\
&\leq \frac{c_{|S|}(\mathbf{r}_S)}{\kappa_{\mathbf{r}_S,|S|}(\tau)}\left|\overline{f}_S\right|_{v_S}.
\end{aligned}$$

Thus,

$$\frac{1}{h}\left|\int_J R_S^{(v_S)} f\right| \leq \left|R_{|S|,\mathbf{r}_S}^{(v_S)}\overline{f}_S\right|_{v_S} \leq \frac{c_{|S|}(\mathbf{r}_S)}{\kappa_{\mathbf{r}_S,|S|}(\tau)}|\overline{f}_S|_{v_S}.$$

Moreover, for any rectangle $I_S \subset \mathbb{R}^S$,

$$\begin{aligned}
\frac{1}{\mu(I_S) + v_S}\left|\int_{I_S}\overline{f}_S(x_S)dx_S\right| &= \frac{1}{(\mu(I_S) + v_S)h_{S^c}}\left|\int_{I_S}\int_{J^{S^c}} f_S(x_{S^c}, x_S)dx_{S^c}dx_S\right| \\
&\leq \frac{1}{\mu(K) + h}\left|\int_K f\right|,
\end{aligned}$$

where $K \in \mathcal{C}$ is the unique rectangle such that $K_S = I_S$ and $K_{S^c} = J_{S^c}$. It follows that $|\overline{f}_S|_{v_S} \leq |f|_h$, which yields the result. □

For any $K = \prod_{i=1}^d [a_i, b_i)$, let

$$\begin{aligned}
l_K : [0,1)^d &\to K \\
u &\mapsto (a_i + u_i(b_i - a_i))_{1 \leq i \leq d}.
\end{aligned}$$

Define the corresponding composition operator,

$$\begin{aligned} A_K : L^\infty(K) &\to L^\infty([0,1)^d) \\ f &\mapsto f \circ l_K. \end{aligned}$$

Finally, for any $\mathbf{j}, \mathbf{j}' \in \mathbb{Z}^d$, let $\mathbf{j} \wedge \mathbf{j}' = (\min(j_i, j'_i))_i$, $S = \{i : j_i \leq j'_i\}$ and

$$\begin{aligned} R^{(\theta)}_{\mathbf{j},\mathbf{j}'} : L^\infty(\mathbb{R}^d) &\to L^\infty(\mathbb{R}^d) \\ f &\mapsto \sum_{K \in \mathcal{I}(\mathbf{j}\wedge\mathbf{j}')} \left[A_K^{-1} R_S^{\theta^{|S|}} A_K\right](f|_K)\mathbb{1}_K. \end{aligned}$$

For any $\theta \in (0,1)$, define the collection of sets

$$\mathcal{C}_{\mathbf{j},\mathbf{j}'}(\theta) = \left\{(1-\lambda)x + \lambda \mathring{I} : I \in \mathcal{I}(\mathbf{j}) \cup \mathcal{I}(\mathbf{j}'), x \in \overline{I}, 0 < \lambda \leq \theta\right\}$$

and the corresponding semi-norm

$$N_{\mathbf{j},\mathbf{j}'}(\theta, h, f) = \sup_{C \in \mathcal{C}_{\mathbf{j},\mathbf{j}'}(\theta)} \left\{\frac{1}{\mu(C) + h}\left|\int_C f\right|\right\}.$$

The operator $R^{(\theta)}_{\mathbf{j},\mathbf{j}'}$ satisfies the following properties.

Proposition 8. *Let $\mathbf{j}, \mathbf{j}' \in \mathbb{Z}^d$ and let $\theta > 0$.*

- *For any $p \in m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}))$, $R^{(\theta)}_{\mathbf{j},\mathbf{j}'}(p) = p$.*
- *For any $q \in m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}'))$, $R^{(\theta)}_{\mathbf{j},\mathbf{j}'}(q) \in m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}\wedge\mathbf{j}'))$*
- *For any $f \in L^\infty(\mathbb{R}^d)$,*

$$N_{\mathbf{j},\mathbf{j}'}\left(\theta, h, R^{(\theta)}_{\mathbf{j},\mathbf{j}'} f\right) \leq \frac{c_{|S|}(\mathbf{r}_S)}{\kappa_{\mathbf{r}_S,|S|}(\theta^{|S|})}|f|_h.$$

Proof.

- Let $K \in \mathcal{I}(\mathbf{j}\wedge\mathbf{j}')$. Let $x \in K$ and let $I \in \mathcal{I}(\mathbf{j})$ contain x. For any $i \in S$, $x_i \in I_i \cap K_i$, hence by lemma 3, $I_i \subset K_i$ or $K_i \subset I_i$. Moreover, for $i \in S, j_i = \min(j_i, j'_i)$, so $|K_i| = |I_i|$, which implies that $I_i = K_i$.

  Hence, $\{z \in K : z_{S^c} = x_{S^c}\} \subset I$, thus $p_S(x_{S^c}, \cdot)$ coincides on K_S with a polynomial from $\mathcal{P}^{dir}_{\mathbf{r}_S,d}$. Since l_K acts coordinatewise, the same is true of $A_K(p|_K)$, with K replaced by $[0,1)^d$. Since $R^{\theta^{|S|}}_{|S|,\mathbf{r}_S}$ is a projection, it follows that

  $$\left[A_K^{-1} R_S^{\theta^{|S|}} A_K\right](p|_K) = \left[A_K^{-1} A_K\right](p|_K) = p|_K.$$

  This proves the first point.
- Let $K \in \mathcal{I}(\mathbf{j}\wedge\mathbf{j}')$. Let $x \in K$ and let $I' \in \mathcal{I}(\mathbf{j}')$ contain x. For any $i \notin S$, $x_i \in I'_i \cap K_i$, hence by lemma 3, $I'_i \subset K_i$ or $K_i \subset I'_i$. Moreover, for $i \in S^c, j'_i = \min(j_i, j'_i)$, so $|K_i| = |I'_i|$, which implies that $I'_i = K_i$.

  Hence, $\{z \in K : z_S = x_S\} \subset I'$, thus $q_S(\cdot, x_S)$ coincides on K_{S^c} with a polynomial from $\mathcal{P}^{dir}_{\mathbf{r}_{S^c},d}$. Since l_K acts coordinatewise, the same is true of $A_K(q|_K) = q|_K \circ l_K$, with K replaced by $[0,1)^d$. Hence, $\tilde{q} = [A_K(q|_K)]_S$ can be written in the form

$$\tilde{q}(u_{S^c}, u_S) = \sum_{\alpha \in \mathcal{A}} c_\alpha(u_S) u_{S^c}^\alpha,$$

where $\mathcal{A} = \{\alpha \in \mathbb{N}^{d-|S|} : \alpha \le \mathbf{r}_{S^c}\}$. Hence, by definition of $R_S^{(\tau)}$, for any $\tau > 0$,

$$\begin{aligned}\left[R_S^{(\tau)} A_K\right](q|_K)(u) &= R^\tau_{|S|,\mathbf{r}_S}\left[y \mapsto \sum_{\alpha\in\mathcal{A}} c_\alpha(y) u_{S^c}^\alpha\right](u_S) \\ &= \sum_{\alpha\in\mathcal{A}} R^{(\tau)}_{|S|,\mathbf{r}_S}[c_\alpha](u_S) u_{S^c}^\alpha,\end{aligned}$$

which proves that $\left[R_S^{(\tau)} A_K\right](q|_K)$ coincides on $[0,1)^d$ with an element of $\mathcal{P}^{dir}_{\mathbf{r},d}$. It follows by definition of $R^{(\theta)}_{\mathbf{j},\mathbf{j}'}$ and linearity of l_K that $R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q$ belongs to $m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j} \wedge \mathbf{j}'))$.

- Let $C \in \mathcal{C}_{\mathbf{j},\mathbf{j}'}(\theta), C = (1-\lambda)x + \lambda \mathring{I}$ for some $I \in \mathcal{I}(\mathbf{j}) \cup \mathcal{I}(\mathbf{j}')$, some $x \in \overline{I}$ and some $\lambda \in (0,\theta]$. By convexity of I, $C \subset \mathring{I}$. By lemma 4, there exists one $K \in \mathcal{I}(\mathbf{j} \wedge \mathbf{j}')$ such that $C \subset I \subset K$. It follows that

$$\begin{aligned}\frac{1}{\mu(C)+h}\int_C R^{(\theta)}_{\mathbf{j},\mathbf{j}'} f &= \frac{1}{\mu(C)+h}\int_C A_K^{-1} R_S^{(\theta^{|S|})} A_K f \\ &= \frac{1}{\mu(C)+h}\int_C \left[R_S^{(\theta^{|S|})} A_K f\right](l_K^{-1}(x))dx \\ &= \frac{1}{\det(l_K^{-1})(\mu(C)+h)}\int_{l_K^{-1}(C)} R_S^{(\theta^{|S|})} A_K f \\ &= \frac{1}{\mu(J) + \frac{h}{\mu(K)}}\int_J R_S^{(\theta^{|S|})} A_K f,\end{aligned}$$

where $J = l_K^{-1}(C) = (1-\lambda)x + \lambda l_K^{-1}(\mathring{I})$. For all $i \in \{1,\ldots,d\}$, $|J_i| \le \lambda \le \theta$, which implies by lemma 3 that

$$\frac{1}{\mu(J) + \frac{h}{\mu(K)}}\left|\int_J R_S^{(\theta^{|S|})} A_K f\right| \le \frac{c_{|S|}(\mathbf{r}_S)}{\kappa_{\mathbf{r}_S,|S|}(\theta^{|S|})}\left|A_K f\right|_{\frac{h}{\mu(K)}}.$$

Now, for any $B \subset [0,1)^d$,

$$\begin{aligned}\frac{1}{\mu(B) + \frac{h}{\mu(K)}}\left|\int_B A_K f\right| &= \frac{1}{\mu(B) + \frac{h}{\mu(K)}}\left|\int_B f(l_K(y))dy\right| \\ &= \frac{1}{\mu(B) + \frac{h}{\mu(K)}}\frac{1}{\det(l_K)}\left|\int_{l_K(B)} f(x)dx\right| \\ &= \frac{1}{\mu(l_K(B)) + h}\left|\int_{l_K(B)} f(x)dx\right| \\ &\le |f|_h.\end{aligned}$$

This finally yields

$$\frac{1}{\mu(C)+h}\left|\int_C R^{(\theta)}_{\mathbf{j},\mathbf{j}'}f\right| \leq \frac{c_{|S|}(\mathbf{r}_S)}{\kappa_{\mathbf{r}_S,|S|}(\theta^{|S|})}|f|_h,$$

which proves the result, since $C \in \mathcal{C}_{\mathbf{j},\mathbf{j}'}(\theta)$ was arbitrary.

□

We can now complete the proof of Theorem 4. Let $p, q \in m \cup m'$, where $m = m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}))$ and $m' = m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}'))$. First, if $p, q \in m$ or $p, q \in m'$, then by proposition 4 and equation (17) defining h_m,

$$\begin{aligned}|p-q|_{h_m\wedge h_{m'}} &\geq \min\left(\kappa_m(h_m\wedge h_{m'}), \kappa_{m'}\left(h_m\wedge h_{m'}\right)\right)\|p-q\|_{\infty,\mu}\\ &\geq \min\left(\kappa_m(h_m), \kappa_{m'}\left(h_{m'}\right)\right)\|p-q\|_{\infty,\mu}\\ &\geq \frac{9}{16}.\end{aligned}$$

Let now $p \in m = m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}))$, $q \in m' = m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}'))$ and $S = \{i \in \{1, \dots, d\} : j_i \leq j'_i\}$. If $S = \emptyset$, then $\mathcal{I}(\mathbf{j})$ refines $\mathcal{I}(\mathbf{j}')$ by lemma 4, hence we are in the case described above.

Assume therefore that $k = |S| \geq 1$. Let $r = \|\mathbf{r}\|_1$, $\theta = \frac{1}{8r^2}$ and $h = h_m \wedge h_{m'}$. By proposition 4,

$$\begin{aligned}\kappa_{\mathbf{r}_S,k}(\theta^k) &\geq \gamma_{\|r_S\|_1,k}(\theta^k)\\ &\geq \left[1-\left(2\left\|\mathbf{r}_S\right\|_1^2\right)^{\frac{k}{k+1}}\left(\frac{1}{8\left\|\mathbf{r}\right\|_1^2}\right)^{\frac{k}{k+1}}\right]^2\\ &\geq \left(1-\frac{1}{4^{\frac{k}{k+1}}}\right)^2\\ &\geq \frac{1}{4},\end{aligned}$$

since $k \geq 1$. Remark also that $\theta = \theta_m(h_m) = \theta_{m'}(h_{m'})$, hence $C_{h_m,m}(f) \in \mathcal{C}_{\mathbf{j},\mathbf{j}'}(\theta)$ for all $f \in m$, and similarly for m'. It follows from proposition 4 that

$$N_{\mathbf{j},\mathbf{j}'}\left(\theta, h_m\wedge h_{m'}, f\right) \geq \left[1-r^{\frac{2d}{d+1}}\left(\frac{1}{r^{2d}4^{d+1}}\right)^{\frac{1}{d+1}}\right]^2\|f\|_{\infty,\mu} \geq \frac{1}{2}\|f\|_{\infty,\mu},$$

for any $f \in m \cup m' \subset m(r, \mathcal{I}(\mathbf{j})) \cup m(r, \mathcal{I}(\mathbf{j}'))$.

Now, by point 2 of proposition 8 above,

$$R^{(\theta)}_{\mathbf{j},\mathbf{j}'}q \in m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j}\wedge\mathbf{j}')) \subset m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j})) = m,$$

in particular $p - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q \in m$. Hence, by proposition 8,

$$\begin{aligned}\frac{1}{2}\left\|p - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q\right\|_{\infty,\mu} &\leq N_{\mathbf{j},\mathbf{j}'}(\theta, h, p - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q) \\ &= N_{\mathbf{j},\mathbf{j}'}\left(\theta, h, R^{(\theta)}_{\mathbf{j},\mathbf{j}'}(p-q)\right) \\ &\leq \frac{c_{|S|}(\mathbf{r}_S)}{\kappa_{\mathbf{r}_S,|S|}(\theta^{|S|})} |p-q|_h \\ &\leq 4c_d(\mathbf{r})|p-q|_h.\end{aligned}$$

For the same reasons, $q - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q \in m'$, hence by the triangle inequality,

$$\begin{aligned}\frac{1}{2}\left\|q - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q\right\|_{\infty,\mu} &\leq N_{\mathbf{j},\mathbf{j}'}(\theta, h, q - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q) \\ &\leq N_{\mathbf{j},\mathbf{j}'}(\theta, h, R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q - p) + N_{\mathbf{j},\mathbf{j}'}(\theta, h, p-q) \\ &\leq [1 + 4c_d(\mathbf{r})]\, |p-q|_h.\end{aligned}$$

It follows that

$$\begin{aligned}|p-q|_h &\geq \frac{1}{2(1+4c_d(\mathbf{r}))} \max\left(\left\|p - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q\right\|_{\infty,\mu}, \left\|q - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q\right\|_{\infty,\mu}\right) \\ &\geq \frac{1}{4(1+4c_d(\mathbf{r}))}\left[\left\|p - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q\right\|_{\infty,\mu} + \left\|q - R^{(\theta)}_{\mathbf{j},\mathbf{j}'} q\right\|_{\infty,\mu}\right] \\ &\geq \frac{\|p-q\|_{\infty,\mu}}{4(1+4c_d(\mathbf{r}))},\end{aligned}$$

which proves the theorem.

Appendix C. Rates under anisotropic smoothness: Proofs

C.1. Proof of Proposition 6.

We begin by a simple lemma.

Lemma 6. *Any $f \in C^{\boldsymbol{\beta}}(\mathbb{R}^d)$ is uniformly continuous, moreover*

$$\forall x, y, |f(y) - f(x)| \leq \sum_{i=1}^{d} |f|_{i,\beta_i \wedge 1} |y_i - x_i|^{\beta_i \wedge 1}.$$

Proof. Let $x, y \in \mathbb{R}^d$. Let $z_1 = x$ and for any $i \in \{2, \ldots, d+1\}$, $z_i = (x_1, \ldots, x_{i-1}, y_i, \ldots, y_d)$. By the triangle inequality,

$$|f(y) - f(x)| \leq \sum_{i=1}^{d} |f(z_{i+1}) - f(z_i)|.$$

Let e_i denote the $i-$th basis vector. By assumption, the function g_i defined on $[0,1]$ by

$$g_i : t \mapsto f\left(z + t(y_i - x_i)e_i\right)$$

belongs to $C^{\beta_i}(\mathbb{R})$, hence

$$|f(z_{i+1}) - f(z_i)| = g_i(1) - g_i(0) \leq |g_i|_{C^{\beta_i \wedge 1}} = |f|_{i,\beta_i \wedge 1} |y_i - x_i|^{\beta_i \wedge 1}.$$

This proves the lemma. □

For any $\delta > 0$, let

$$\mathcal{I}_{\mathbf{h},\delta}(f) = \{I \in \mathcal{I}_{\mathbf{h}} : \|f \mathbb{1}_I\|_{\infty,\mu} \geq \delta\}.$$

Then $\mathcal{I}_{\mathbf{h},\delta}(f)$ is finite. Indeed, let $(\delta_i)_{1\leq i\leq d}$ be such that

$$\sum_{i=1}^{d} |f|_{i,\beta_i\wedge 1}\delta_i^{\beta_i\wedge 1} \leq \frac{\delta}{2}.$$

For all $I = \prod_{i=1}^{d} I_i \in \mathcal{I}_{\mathbf{h},\delta}(f)$, let $x^I \in \overline{I}$ be such that $f(x^I) \geq \delta$ and define

$$J_i(I) = \begin{cases} \left(x_i^I, (\sup I_i) \wedge (x_i^I + \delta_i)\right) \text{ if } (\sup I_i) - x_i^I \geq \frac{h_i}{2} \\ \left(\inf I_i \vee (x_i^I - \delta_i), x_i^I\right) \text{ else.} \end{cases}$$

Let $J(I) = \prod_{i=1}^{d} J_i(I)$. Then, by the above lemma, $f(y) \geq \frac{\delta}{2}$ for any $y \in J(I)$, hence

$$\int_I |f(y)|dy \geq \int_{J(I)} |f(y)|dy \geq \frac{\delta}{2}\mu(J(I)) \geq \frac{\delta}{2}\prod_{i=1}^{d}\left(\delta_i \wedge \frac{h_i}{2}\right).$$

It follows that

$$\|f\|_{L^1} \geq \sum_{I\in\mathcal{I}_{\mathbf{h},\delta}(f)} \int_I |f(y)|dy \geq |\mathcal{I}_{\mathbf{h},\delta}(f)| \left[\frac{\delta}{2}\prod_{i=1}^{d}\left(\delta_i \wedge \frac{h_i}{2}\right)\right],$$

which implies the finiteness of $\mathcal{I}_{\mathbf{h},\delta}(f)$.

It follows that for any sequence $(g_I)_{I\in\mathcal{I}_{\mathbf{h},\delta}(f)}$ of elements of $\mathcal{P}_{\mathbf{r},d}^{dir}$,

$$\sum_{I\in\mathcal{I}_{\mathbf{h},\delta}(f)} g_I \mathbb{1}_I \in \overline{m_{dir}}(\mathbf{r}, \mathcal{I}_{\mathbf{h}}).$$

Moreover, since f is continuous,

$$\left\| f - \sum_{I\in\mathcal{I}_{\mathbf{h},\delta}(f)} g_I \mathbb{1}_I \right\|_{\infty,\mu} \leq \delta \vee \max_{I\in\mathcal{I}_{\mathbf{h},\delta}(f)} \sup_{x\in\overline{I}} |(f - g_I)(x)|.$$

Therefore, and since δ may be chosen arbitrarily small, it suffices to study uniform polynomial approximation of f on $\overline{I}$ for a given $I \in \mathcal{I}_{\mathbf{h}}$, say $\prod_{j=1}^{d}[0, h_j]$. Up to permuting $\boldsymbol{\beta}, \mathbf{r}$ and the arguments of f, we can also assume that the identity permutation achieves the minimum in equation (18) .

Let $\mathcal{P}_r(I)$ denote the space of univariate polynomial functions with degree at most r on the interval I. For $h > 0$, let S_h denote the scaling and translation operator:

$$S_h g : x \mapsto g\left(\frac{h}{2} + \frac{h}{2}x\right)$$

which maps $\mathcal{C}([0,h])$ isometrically to $\mathcal{C}([-1,1])$. For each $r \in \mathbb{N}$, polynomial interpolation at the Chebyshev nodes yields a continuous linear projection $Q_r : \mathcal{C}([-1,1]) \to \mathcal{P}_r([-1,1])$ with operator norm

$$\|Q_r\|_{\infty,\mu} \leq \frac{2}{\pi}\log(r+1) + 1 := a_1(r)$$

(for a reference, see [15, Theorem 1.2]).

For any $j \in \{1,\dots,d\}$, let $\mathbf{r}_j = (r_1,\dots,r_j)$ and $a_j(\mathbf{r}_j) = \prod_{i=1}^{j} a_1(r_i)$, with the convention $a_0(\mathbf{r}_0) = 1$. Define recursively functions $(f_j)_{0\leq j\leq d}$ by $f_0 = f$ and

$$f_j(x_1,\dots,x_d) = S_{h_j}^{-1} Q_{r_j} S_{h_j} \left[t \mapsto f_{j-1}(x_1,\dots,x_{j-1},t,x_{j+1},\dots,x_d)\right](x_j).$$

It follows by induction that f_j is polynomial as a function of $(x_i)_{1\leq i\leq j}$, with directional degree $\deg_i(f_j) \leq r_i$. In particular, $f_d \in \mathcal{P}^{dir}_{\mathbf{r},d}$.

Since S_h is an isometry and $\left\|Q_{r_j}\right\| \leq a_1(r_j)$, $\|f_j\|_{\infty,\mu} \leq a_1(r_j)\,\|f_{j-1}\|_{\infty,\mu}$, hence $\|f_j\|_{\infty,\mu} \leq a_j(\mathbf{r}_j)\,\|f\|_{\infty,\mu}$.

Moreover, by linearity and continuity of $S_h^{-1} Q_{r_j} S_h$, for all $i > j \geq 1$,

$$\partial_{x_i}^{\lfloor\beta_i\rfloor} f_j(x) = S_{h_j}^{-1} Q_{r_j} S_{h_j} \left[t \mapsto \partial_{x_i}^{\lfloor\beta_i\rfloor} f_{j-1}(x_1,\dots,x_{i-1},t,x_i,\dots,x_d)\right](x_j).$$

It follows that

$$|f_j|_{i,\beta_i} \leq \left\|S_{h_j}^{-1} Q_{r_j} S_{h_j}\right\| |f_{j-1}|_{i,\beta_i} \leq a_1(r_j)|f_{j-1}|_{i,\beta_i}.$$

By induction, this proves that $\|f_j\|_{i,\beta_i} \leq a_j(\mathbf{r}_j)\,\|f\|_{i,\beta_i}$. Fix now $j \in \{1,\dots,d\}$ and $x \in \prod_{i=1}^{d}[0,h_i]$. For any $P_j \in \mathcal{P}_{r_j}([0,h_j])$, $S_{h_j}^{-1} Q_{r_j} S_{h_j} P_j = P_j$, hence by the Lebesgue lemma,

$$\begin{aligned}
|f_j(x) - f_{j-1}(x)| &\leq \sup_{t\in[0,h_j]} |(f_j - f_{j-1})(x_1,\dots,x_{j-1},t,x_{j+1},\dots,x_d)| \\
&\leq [1 + \left\|S_{h_j}^{-1} Q_{r_j} S_{h_j}\right\|] \sup_{t\in[0,h_j]} |f_{j-1}(x_1,\dots,x_{j-1},t,x_{j+1},\dots,x_d) - P_j(t)|\,.
\end{aligned}$$

Choosing P_j to be the Taylor expansion of f_{j-1} at $(x_1,\dots,x_{j-1},0,x_{j+1},\dots,x_d)$ along the coordinate x_j yields

$$\begin{aligned}
|f_j(x) - f_{j-1}(x)| &\leq [1 + a_1(r_j)]\frac{h_j^{\beta_j}}{\lfloor\beta_j\rfloor!}|f_{j-1}|_{j,\beta_j} \\
&\leq [1 + a_1(r_j)]\frac{a_{j-1}(\mathbf{r}_{j-1})h_j^{\beta_j}}{\lfloor\beta_j\rfloor!}|f|_{j,\beta_j} \\
&\leq (a_j(\mathbf{r}_j) + a_{j-1}(\mathbf{r}_{j-1}))\,\frac{h_j^{\beta_j}}{\lfloor\beta_j\rfloor!}|f|_{j,\beta_j}.
\end{aligned}$$

It follows by the triangle inequality that

$$\begin{aligned}\|f - f_d\|_{\infty,\mu} &\leq \sum_{j=1}^{d} \|f_j - f_{j-1}\|_{\infty,\mu} \\ &\leq 2\left(\sum_{j=0}^{d} a_j(\mathbf{r}_j)\right) \max_{1\leq j\leq d}\left\{\frac{h_j^{\beta_j}}{\lfloor \beta_j \rfloor !}|f|_{j,\beta_j}\right\},\end{aligned}$$

which proves the result.

C.2. **Proof of Theorem 5.** Let C be a constant depending on $\mathbf{r}, d$ only, the value of which may change from line to line. Let c denote a numerical constant, which can also change from line to line.

By theorem 4, assumption 2 holds for some κ_* depending on $\mathbf{r}, d$ only. Hence, we can apply Theorem 3 with $a = 3$. Let

$$w_{n,d} = L_\beta(p^\star)^{\frac{d}{2\beta+d}}\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} \leq 1.$$

Let $\theta \in [0,1]$, to be chosen later, such that $\|p^\star\|_{\infty,\mu}^{\frac{\theta(\beta+d)}{2\beta+d}} \geq w_{n,d}$.

For any $i \in \{1,\dots,d\}$, let

$$z_i = \left[w_{n,d}\,\|p^\star\|_{\infty,\mu}^{\frac{\theta\beta}{2\beta+d}}\right]^{\frac{1}{\beta_i}}\left(\frac{|p^\star|_{i,\beta_i}}{\lfloor \beta_i \rfloor !}\right)^{\frac{-1}{\beta_i}}.$$

Let $\mathbf{j} \in \mathbb{Z}^d$ be such that $2^{-j_i} \leq z_i < 2^{-j_i+1}$, for any $i \in \{1,\dots,d\}$. Let $m = m_{dir}(\mathbf{r}, \mathcal{I}(\mathbf{j})) \in \mathcal{M}_{\mathbf{r}}$. By Theorem 3, for all $y > 0$, on $\Omega_{3\log y}$

$$\|\hat{p} - p^\star\|_{\infty,\mu} \leq C \inf_{p\in m}\|p^\star - p\|_{\infty,\mu} + 4\,\mathrm{pen}_3(h_m) + \frac{cy}{\sqrt{n}}. \tag{32}$$

By equation (14) and proposition 5 with $\theta = \frac{2}{3}$,

$$\begin{aligned}\mathrm{pen}_3(h_m) &= 29\sqrt{\frac{4}{3}}\sqrt{\frac{|\hat{p}|_h(\Gamma + 3\log_-(h_m))}{h_m n}} + \sqrt{\frac{4}{3}}29^2\frac{\Gamma + 3\log_-(h_m)}{h_m n} \\ &\leq 29\frac{4}{3}\sqrt{\frac{|p^\star|_h(\Gamma + 3\log_-(h_m))}{h_m n}} + \frac{29^2}{h_m n}\sqrt{(\Gamma + 3\log y)(\Gamma + 3\log_- h_m)} \\ &\quad + \sqrt{\frac{4}{3}}29^2\frac{\Gamma + 3\log_-(h_m)}{h_m n}.\end{aligned} \tag{33}$$

By equation (17) and definition of $\mathbf{j}$,

$$\begin{aligned}
h_m &= \frac{\prod_{i=1}^d 2^{-j_i}}{(2\|\mathbf{r}\|_1^2)^d 4^{d+1}} \\
&\geq \frac{2^{-d}\prod_{i=1}^d z_i}{(2\|\mathbf{r}\|_1^2)^d 4^{d+1}} \\
&\geq \frac{1}{C}\left[w_{n,d}\|p^\star\|_{\infty,\mu}^{\frac{\theta\beta}{2\beta+d}}\right]^{\sum_{i=1}^d \frac{1}{\beta_i}} \prod_{i=1}^d \left(\frac{|p^\star|_{i,\beta_i}}{\lfloor\beta_i\rfloor!}\right)^{\frac{-1}{\beta_i}} \\
&= \frac{1}{C}\left[w_{n,d}\|p^\star\|_{\infty,\mu}^{\frac{\theta\beta}{2\beta+d}}\right]^{\frac{d}{\beta}} L_{\boldsymbol{\beta}}(p^\star)^{\frac{-d}{\beta}} \\
&\geq \frac{1}{C} w_{n,d}^{\frac{d}{\beta}} L_{\boldsymbol{\beta}}(p^\star)^{\frac{-d}{\beta}} \|p^\star\|_{\infty,\mu}^{\frac{\theta d}{2\beta+d}}.
\end{aligned}$$

Moreover, by the assumption on θ,

$$\begin{aligned}
h_m &\geq \frac{1}{C} w_{n,d}^{\frac{d}{\beta}} w_{n,d}^{-\frac{2\beta+d}{\beta}} \frac{\log n}{n} \|p^\star\|_{\infty,\mu}^{\frac{\theta d}{2\beta+d}} \\
&\geq \frac{1}{C} w_{n,d}^{-2} \frac{\log n}{n} w_{n,d}^{\frac{d}{\beta+d}} \\
&\geq \frac{\log n}{Cn}
\end{aligned}$$

since $w_{n,d} \leq 1$.

In particular, $\log_-(h_m) \leq \log C + \log n$. Moreover, the VC-dimension of $\mathcal{C}$ is $2d$, so $\Gamma \leq C \log n$. Since

$$\frac{\log n}{n h_m} \leq C \|p^\star\|_{\infty,\mu}^{\frac{-\theta d}{2\beta+d}} \frac{L_{\boldsymbol{\beta}}(p^\star)^{\frac{d}{\beta}} \log n}{n} w_{n,d}^{-\frac{d}{\beta}} = C \|p^\star\|_{\infty,\mu}^{\frac{-\theta d}{2\beta+d}} w_{n,d}^{\frac{2\beta+d}{\beta} - \frac{d}{\beta}},$$

it follows by equation (33) that

$$\text{pen}_3(h_m) \leq C \|p^\star\|_{\infty,\mu}^{\frac{1}{2}\left(1-\frac{\theta d}{2\beta+d}\right)} w_{n,d} + C \|p^\star\|_{\infty,\mu}^{\frac{-\theta d}{2\beta+d}} w_{n,d}^2 \left[1 + \sqrt{\frac{\log y}{\log n}}\right]. \tag{34}$$

Moreover, by proposition 6 and definition of z_i,

$$\begin{aligned}
\inf_{p \in m} \|p - p^\star\|_{\infty,\mu} &\leq C \max_{1 \leq i \leq d} \left\{\frac{2^{-\beta_i j_i}}{\lfloor\beta_i\rfloor!} |p^\star|_{i,\beta_i}\right\} \\
&\leq C \max_{1 \leq i \leq d} \left\{\frac{z_i^{\beta_i}}{\lfloor\beta_i\rfloor!} |p^\star|_{i,\beta_i}\right\} \\
&\leq C \|p^\star\|_{\infty,\mu}^{\frac{\theta\beta}{2\beta+d}} w_{n,d}.
\end{aligned}$$

Consider first the case $\|p^\star\|_{\infty,\mu} \geq w_{n,d}^{\frac{2\beta+d}{\beta+d}}$. Set then $\theta = 1$, which yields

$$\|p^\star\|_{\infty,\mu}^{\frac{-\theta d}{2\beta+d}} w_{n,d}^2 = \|p^\star\|_{\infty,\mu}^{\frac{\beta}{2\beta+d}} \|p^\star\|_{\infty,\mu}^{\frac{-(\beta+d)}{2\beta+d}} w_{n,d}^2 \leq \|p^\star\|_{\infty,\mu}^{\frac{\beta}{2\beta+d}} w_{n,d}^{-1} w_{n,d}^2 \leq \|p^\star\|_{\infty,\mu}^{\frac{\beta}{2\beta+d}} w_{n,d}.$$

It follows from equations (32), (34) that, on $\Omega_{3\log y}$,

$$\|\hat{p}-p^\star\|_{\infty,\mu} \le C\,\|p^\star\|_{\infty,\mu}^{\frac{\beta}{2\beta+d}}\,w_{n,d}\left[1+\sqrt{\frac{\log y}{\log n}}\right]+\frac{cy}{\sqrt{n}}.$$

Let $y = 2^{\frac{1}{3}}e^{\frac{x}{3}}$. With probability greater than $1-e^{-x}$,

$$\|\hat{p}-p^\star\|_{\infty,\mu} \le C\,\|p^\star\|_{\infty,\mu}^{\frac{\beta}{2\beta+d}}\,w_{n,d}\left[1+\sqrt{\frac{x+\log 4}{3\log n}}\right]+2^{\frac{1}{3}}\frac{ce^{\frac{x}{3}}}{\sqrt{n}}.$$

Since this holds for any $x>0$, it follows by [12, Lemma 21] that

$$\mathbb{E}\left[\|\hat{p}-p^\star\|_{\infty,\mu}\right] \le C\,\|p^\star\|_{\infty,\mu}^{\frac{\beta}{2\beta+d}}\,w_{n,d}+\frac{c}{\sqrt{n}},$$

which proves the result.

Consider now the case $\|p^\star\|_{\infty,\mu} < w_{n,d}^{\frac{2\beta+d}{\beta+d}}$. Let $\theta\in[0,1)$ solve the equation $\|p^\star\|_{\infty,\mu}^{\frac{\theta(\beta+d)}{2\beta+d}} = w_{n,d}$, which implies that

$$\begin{aligned}\|p^\star\|_{\infty,\mu}^{\frac{\theta\beta}{2\beta+d}}\,w_{n,d} &\le w_{n,d}^{\frac{\beta}{\beta+d}}w_{n,d} \le w_{n,d}^{\frac{2\beta+d}{\beta+d}}\\ \|p^\star\|_{\infty,\mu}^{\frac{-\theta d}{2\beta+d}}\,w_{n,d}^2 &\le w_{n,d}^{\frac{-d}{\beta+d}}w_{n,d}^2 \le w_{n,d}^{\frac{2\beta+d}{\beta+d}}.\end{aligned}$$

Moreover, since

$$\frac{1}{2}\left(1-\frac{\theta d}{2\beta+d}\right)\ge\frac{1}{2}\left(1-\frac{d}{2\beta+d}\right)\ge\frac{\beta}{2\beta+d}\ge\frac{\theta\beta}{2\beta+d}$$

and $\|p^\star\|_{\infty,\mu}<1$,

$$\|p^\star\|_{\infty,\mu}^{\frac{1}{2}\left(1-\frac{\theta d}{2\beta+d}\right)}\,w_{n,d}\le\|p^\star\|_{\infty,\mu}^{\frac{\theta\beta}{2\beta+d}}\,w_{n,d}\le w_{n,d}^{\frac{2\beta+d}{\beta+d}}.$$

It follows from equations (32), (34) that, on $\Omega_{3\log y}$,

$$\|\hat{p}-p^\star\|_{\infty,\mu}\le Cw_{n,d}^{\frac{2\beta+d}{\beta+d}}\left[1+\sqrt{\frac{\log y}{\log n}}\right]+\frac{cy}{\sqrt{n}}.$$

Let $y = 2^{\frac{1}{3}}e^{\frac{x}{3}}$. With probability greater than $1-e^{-x}$,

$$\|\hat{p}-p^\star\|_{\infty,\mu}\le Cw_{n,d}^{\frac{2\beta+d}{\beta+d}}\left[1+\sqrt{\frac{x+\log 4}{3\log n}}\right]+2^{\frac{1}{3}}\frac{ce^{\frac{x}{3}}}{\sqrt{n}}.$$

Since this holds for any $x>0$, it follows by [12, Lemma 21] that

$$\mathbb{E}\left[\|\hat{p}-p^\star\|_{\infty,\mu}\right]\le Cw_{n,d}^{\frac{2\beta+d}{\beta+d}}+\frac{c}{\sqrt{n}},$$

which proves the result.

C.3. **Proof of Theorem 6.** Denote by C a constant depending only on $\boldsymbol{\beta}, d$, the value of which may change from line to line.

The proof follows that of Lepski ([11, Theorem 2]) in the case $m = d$, $p_i = +\infty$ for $i \in \{1, \dots, d\}$, $\sigma = \frac{1}{\sqrt{2\pi}} \left(\frac{2}{b}\right)^{\frac{1}{d}}$, until equation (4.24). Thus, let $f_0 = p_{\frac{b}{2}}, (f^{(j)})_{j \in \mathbf{J}_n}$ belonging to $C^{\boldsymbol{\beta}}_{\mathbf{L},+\infty}$ be constructed as in [11], with $m = d, \mathbf{I}^* = \{1, \dots, d\}$. In particular,

$$\left\| f^{(j)} - f_0 \right\|_{\infty,\mu} = c_1^{\star} A_n = |g(0)|^d A_n$$

for all $j \in \mathbf{J}_n$, where A_n is a sequence converging to 0 . Moreover,

$$\begin{aligned} \mathcal{E}_n &:= \mathbb{E}^{(n)}_{f_0} \left[\frac{1}{\mathbf{J}_n} \sum_{j \in \mathbf{J}_n} \int_{\mathbb{R}^d} \frac{d\mathbb{P}^{(n)}_{f^{(j)}}}{d\mathbb{P}^{(n)}_{f_0}} (\mathbf{X}^{(n)}) - 1 \right]^2 \\ &= \frac{1}{|\mathbf{J}_n|} \sum_{j \in \mathbf{J}_n} \left\{ 1 + \int_{\mathbb{R}^d} \left[\frac{G_j^2(y)}{f_0(y)} \right] dy \right\}^n - \frac{1}{|\mathbf{J}_n|}. \end{aligned}$$

The G_j are supported in $\mathcal{Y}_n = \prod_{i=1}^d [0, \sqrt{\delta_{i,n}}]$, where the $\delta_{l,n}$ converge to 0, so that

$$\lim_{n \to +\infty} \inf_{y \in \mathcal{Y}_n} f_0(y) = f_0(0) = p_{b/2}(0) = \frac{b}{2}.$$

Hence, for all large enough n, by definition of the G_j,

$$\mathcal{E}_n \leq \frac{1}{|\mathbf{J}_n|} \left(1 + \frac{4}{b} A_n^2 \prod_{l=1}^d \delta_{l,n} \right)^n .$$

Let then $A_n, \delta_{l,n}$ satisfy the following equations for all large enough n:

$$\forall k \in \{1, \dots, d\}, A_n \delta_{k,n}^{-\beta_k} = \frac{1}{\|g\|_{\infty,\mu}^{d-1}} L_k \tag{35}$$

$$\frac{4}{b} n A_n^2 \prod_{l=1}^d \delta_{l,n} \leq \frac{1}{4} \log \left(\prod_{l=1}^d \delta_{l,n}^{-1} \right). \tag{36}$$

Note that by construction, $\frac{1}{4} \log \left(\prod_{l=1}^d \delta_{l,n}^{-1} \right) \leq \log(|\mathbf{J}_n|)$, so that $\mathcal{E}_n \leq 1$. Hence, by [10, Proposition 6] and the following [10, Corollary 2],

$$\lim \inf_{n \to +\infty} \inf_{\tilde{p}} \sup_{p \in \mathcal{P}^{\boldsymbol{\beta}}_{\mathbf{L},b}} |g(0)|^d A_n^{-1} \mathbb{E}_{\mathbf{X} \sim P^{\otimes n}} \left[\|\tilde{p}(\mathbf{X}) - p\|_{\infty,\mu} \right] \geq \frac{1}{2} \left(1 - \frac{1}{\sqrt{5}} \right).$$

Thus, the minimax convergence rate is at least A_n. It remains to solve equations (35), (36). To that end, let

$$\bar{L} = \prod_{k=1}^{d} L_k^{\frac{\beta}{d\beta_k}}$$

$$A_n = (\lambda b)^{\frac{\beta}{2\beta+d}} \bar{L}^{\frac{d}{2\beta+d}} \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}$$

for some $\lambda > 0$ to be chosen later. Let $C = \frac{1}{\|g\|_{\infty,\mu}^{d-1}}$ Solving (35) yields

$$\delta_{k,n} = \left(C\frac{A_n}{L_k}\right)^{\frac{1}{\beta_k}}$$

for all k, which implies that

$$\begin{aligned}
\prod_{l=1}^{d} \delta_{l,n} &= C^{\frac{d}{\beta}} A_n^{\frac{d}{\beta}} \prod_{l=1}^{d} \left(\frac{1}{L_l}\right)^{\frac{1}{\beta_l}} \\
&= \left(\frac{C}{\bar{L}} A_n\right)^{\frac{d}{\beta}} \\
\frac{1}{4}\log\left(\prod_{l=1}^{d} \delta_{l,n}^{-1}\right) &\sim_{n\to+\infty} \frac{-d}{4\beta}\log A_n \\
&\sim_{n\to+\infty} \frac{1}{4}\frac{d}{2\beta+d}\log n.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\frac{4}{b} n A_n^2 \prod_{l=1}^{d} \delta_{l,n} &= \frac{4}{b} n A_n^{2+\frac{d}{\beta}} C^{\frac{d}{\beta}} \bar{L}^{\frac{-d}{\beta}} \\
&= C^{\frac{d}{\beta}} \frac{4}{b} n A_n^{\frac{2\beta+d}{\beta}} \bar{L}^{\frac{-d}{\beta}} \\
&= 4\lambda C^{\frac{d}{\beta}}.
\end{aligned}$$

Thus, taking $\lambda < \frac{C^{\frac{-d}{\beta}}}{16}\frac{d}{2\beta+d}$ ensures that (36) holds for all n large enough.

Appendix D. Proofs

Lemma 7. *Let $X_1, \dots, X_n$ be independent random variables. Let $\mathcal{F}$ be a countable class of bounded measurable functions and*

$$Z = \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)]\right|.$$

Then with probability greater than $1 - e^{-x}$, for any $\theta > 0$,

$$Z \leq (1+2\theta)EZ + 2\sigma\sqrt{2xn} + \left(2 + \frac{4}{\theta}\right)cx,$$

where

$$\sigma^2 = \sup_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(f(X_i)) \right\}$$
$$c = \sup_{f\in\mathcal{F}} \{\|f\|_\infty\}.$$

Proof. Let

$$\tilde{Z} = \frac{Z}{c} = \sup_{\tilde{f}\in\frac{1}{c}\mathcal{F}} \left| \sum_{i=1}^{n} \tilde{f}(X_i) - \mathbb{E}[\tilde{f}(X_i)] \right|.$$

Let also $\tilde{\sigma} = \frac{\sigma}{c}$,

$$\tilde{\Sigma}^2 = \sup_{\tilde{f}\in\frac{1}{c}\mathcal{F}} \sum_{i=1}^{n} \left(\tilde{f}(X_i) - \mathbb{E}[\tilde{f}(X_i)]\right)^2.$$

By [6, Theorem 12.2] with $t = 2x + 2\sqrt{(n\tilde{\sigma}^2 + \tilde{\Sigma}^2)x}$,

$$\mathbb{P}\left(\tilde{Z} - E\tilde{Z} \geq 2x + 2\sqrt{(n\tilde{\sigma}^2 + \tilde{\Sigma}^2)x}\right) \leq e^{-x}.$$

Moreover, by [6, Theorem 11.8], $\tilde{\Sigma}^2 \leq 8E\tilde{Z} + 2\tilde{\sigma}^2$, so with probability greater than $1 - e^{-x}$, for any $\theta > 0$,

$$\begin{aligned}
\tilde{Z} &\leq E\tilde{Z} + 2\sqrt{x(8E\tilde{Z} + 2n\tilde{\sigma}^2)} + 2x \\
&\leq E\tilde{Z} + 2\sqrt{8xE\tilde{Z}} + 2\tilde{\sigma}\sqrt{2xn} + 2x \\
&\leq (1 + 2\theta)\, E\tilde{Z} + \frac{4x}{\theta} + 2\tilde{\sigma}\sqrt{2xn} + 2x.
\end{aligned}$$

In other words,

$$\frac{1}{c}Z \leq \frac{1+2\theta}{c} EZ + 2\frac{\sigma}{c}\sqrt{2xn} + \left(2 + \frac{4}{\theta}\right) x,$$

which proves the result. □

References

[1] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: rho - estimation. *Inventiones mathematicae*, 207(2):425–517, July 2016.

[2] Yannick Baraud. Bounding the expectation of the supremum of an empirical process over a (weak) VC-major class. *Electronic Journal of Statistics*, 10(2), January 2016.

[3] Yannick Baraud. Tests and estimation strategies associated to some loss functions. *Probability Theory and Related Fields*, 180, 08 2021.

[4] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 02 1999.

[5] Karine Bertin. Asymptotically exact minimax estimation in sup-norm for anisotropic Hölder classes. *Bernoulli*, 10(5):873 – 888, 2004.
[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, February 2013.
[7] W. Dahmen, R. de Vore, and K. Scherer. Multidimensional spline approximation. *SIAM Journal on Numerical Analysis*, 17(3):380–402, 1980.
[8] Ronald A DeVore and George G Lorentz. *Constructive Approximation*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Germany, 1993 edition, October 1993.
[9] Luc Devroye and Gabor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer, New York, NY, 2001 edition, January 2001.
[10] Gerard Kerkyacharian, Oleg Lepski, and Dominique Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability Theory and Related Fields*, 121:137–170, 01 2001.
[11] Oleg Lepski. Multivariate density estimation under sup-norm loss: Oracle approach, adaptation and independence structure. *The Annals of Statistics*, 41(2):1005 – 1034, 2013.
[12] Guillaume Maillard, Sylvain Arlot, and Matthieu Lerasle. Aggregated hold-out. *Journal of Machine Learning Research*, 22(20):1–55, 2021.
[13] Pascal Massart. *Concentration Inequalities and Model Selection*. Springer Berlin Heidelberg, 2007.
[14] Pascal Massart. A non asymptotic walk in probability and statistics. In Xihong Lin, Christian Genest, David L. Banks, Geert Molenberghs, David W. Scott, and Jane-Ling Wang, editors, *Past, Present, and Future of Statistical Science*. Chapman and Hall/CRC, 2014.
[15] T.J. Rivlin. *Chebyshev Polynomials*. Dover Books on Mathematics. Dover Publications, 2020.
[16] Walter Rudin. *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., USA, 1987.
[17] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, 2009.
[18] Ulrike von Luxburg and Bernhard Schölkopf. Statistical learning theory: Models, concepts, and results. In Dov M. Gabbay, Stephan Hartmann, and John Woods, editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 651–706. North-Holland, 2011.

Department of Mathematics,
University of Luxembourg
Maison du nombre
6 avenue de la Fonte
L-4364 Esch-sur-Alzette
Grand Duchy of Luxembourg
Email address: yannick.baraud@uni.lu

Email address: helene.halconruy@uni.lu

Email address: guillaume.maillard@uni.lu