MINIMAX LEARNING RATES FOR ESTIMATING BINARY CLASSIFIERS UNDER MARGIN CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study classification problems using binary estimators where the decision boundary is described by horizon functions and where the data distribution satisfies a geometric margin condition. We establish lower bounds for the minimax learning rate over broad function classes with bounded Kolmogorov entropy in Lebesgue norms. A key novelty of our work is the derivation of lower bounds on the worst-case learning rates under a geometric margin condition—a setting that is almost universally satisfied in practice but remains theoretically challenging. Moreover, our results deal with the noiseless setting, where lower bounds are particularly hard to establish. We apply our general results to classification problems with decision boundaries belonging to several function classes: for Barron-regular functions, Hölder-continuous functions, and convex functions with strong margins, we identify optimal rates close to the fast learning rates of $\mathcal{O}(n^{-1})$ for $n \in \mathbb{N}$ samples.

1 INTRODUCTION

How well can we solve classification problems with complex decision boundaries in deep learning? A lot of emphasis has been put on the noise in the problem. However, in practice, data sets may have very strong margins (see C3)) between the classes, which makes learning much simpler (see e.g. Figure 1). The presence of a margin seriously complicates the identification of lower bounds on learning success. This is intuitively clear, since in the extreme case, where certain regions between the classes almost surely do not contain any data points, many decision boundaries are valid.

In this manuscript, we overcome these issues and present lower bounds on learning under margin conditions.

For $n \in \mathbb{N}$ samples, an estimator can be defined as a measurable function $f : \Lambda^n \to \mathcal{F}$ and a binary classifier can be seen as an indicator function $\mathbb{1}_\Omega : X \to \{0, 1\}$, where $\Lambda := X \times \{0, 1\}$ for some set X containing a sequence $\{\boldsymbol{x}_i\}_{i=1}^n$ that belongs to a sample $((\boldsymbol{x}_i, \mathbb{1}_\Omega(\boldsymbol{x}_i)))_{i=1}^n \in \Lambda^n$, $\mathcal{F}$ is a measurable space, and $\Omega \subset X$ is the decision set.

Remark 1. *It is clear from the definition above, that the classifiers considered in this work are not corrupted by noise. This is an important assumption if we want to resolve the precise role of the regularity of the decision boundary and the margin conditions. Indeed, even the presence of low noise could yield vastly different lower bounds, because the learning problem then requires resolving the noise and this complication can mask the role of the decision boundary and the margin. An extended discussion of this is given in (Petersen & Voigtlaender, 2021, Section 1.1, Point 1). For a quick argument, we highlight, e.g., Stone (1982), where it was obtained that the optimal learning rate, to learn a function $f \in C^k([0, 1]^d)$ with $\|f\|_{C^k} \leq 1$ and noise defined as a parameter $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$ for $\sigma > 0$, is of the order of $O(n^{-k/(2k+d)})$, and decays slower than $n^{-1/2}$. On the other hand, in Krieg & Sonnleitner (2023), where the same problem is considered without noise, learning rates of the order of $\mathcal{O}(n^{-k/d})$ were obtained, in some cases faster than n^{-1}.*

1.1 CONDITIONS

In this paper, we study classification learning problems using binary estimators, when $\mathcal{F} := L^2([0, 1]^d)$; $X := [0, 1]^d$ and the following conditions are fulfilled:

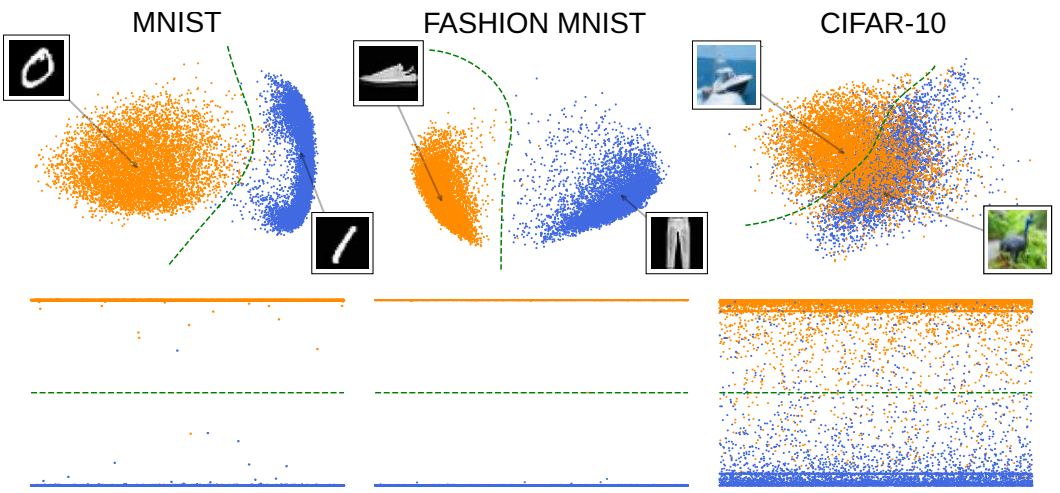

Figure 1: Geometric margin in common classification problems. The top row shows a two dimensional embedding on the first two principle components and a decision boundary identified by a support vector machine. Clearly MNIST Lecun et al. (1998) and Fashion MNIST Xiao et al. (2017) exhibit a strong margin between some classes. For the CIFAR-10 data Krizhevsky (2009) the margin is not visible in the two dimensional embedding. In the second row, we show the class probabilities predicted by a support vector classifier, which again shows extremely strong margin for MNIST and Fashon MNIST, but also reveals that the CIFAR-10 data exhibits a margin, albeit a weaker one. Which lower bounds on learning can be found in the presence of such various types of margins will be demonstrated in our main results Theorem 6, Corollary 8 and Corollary 9.

C1) Ω can be described by horizon functions: We define the general horizon function associated with $b \in \mathscr{C} \subset C([0,1]^{d-1}; [0,1])$ as

$$\begin{aligned} h_b : [0,1]^d &\to \{0,1\}, \\ \boldsymbol{x} = (x_1, \ldots, x_d) &\mapsto \mathbb{1}_{b(\boldsymbol{x}^{(d)}) \leq x_d}, \end{aligned} \tag{1}$$

where $\boldsymbol{x}^{(d)} := (x_1, \ldots, x_{d-1})$. Then it is fulfilled that

$$\Omega = \Omega(h) := \Omega_h = \left\{ \boldsymbol{x} \in [0,1]^d : h(\boldsymbol{x}) = 1 \right\} \quad \text{with} \quad h \in H_{\mathscr{C}} := \{h_b : b \in \mathscr{C}\}, \tag{2}$$

where $H_{\mathscr{C}}$ is the set of general horizon functions associated to $\mathscr{C}$ Petersen & Voigtlaender (2021). So, $\mathbb{1}_{\Omega_h} = h$.

Moreover, Λ is equipped with a probability measure $\boldsymbol{\mu}_h$, such that $\boldsymbol{\mu}_h \left([0,1]^d \times \{i\} \right) = \nu_i \in [0,1]$, for $i \in \{0,1\}$ and $\nu_0 + \nu_1 = 1$; $\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h$, where $h \in H_{\mathscr{C}}$, and μ_h is the marginal distribution of $[0,1]^d$ admitting a density function f_h with respect to the Lebesgue measure λ.

C2) Regular boundary: $\mathscr{C}$ is convex with $0 \in \mathscr{C}$. Furthermore, every $b \in \mathscr{C}$ satisfies

$$|b(\boldsymbol{z}) - b(\boldsymbol{z}')| \leq K \left\| \boldsymbol{z} - \boldsymbol{z}' \right\|^{\alpha} \quad \text{for all} \quad \boldsymbol{z}, \boldsymbol{z}' \in [0,1]^{d-1},$$

some $\alpha \in (0,1]$ and a constant $K > 0$.

C3) The margin condition is satisfied: There exist $C, \gamma > 0$, such that for all $\epsilon > 0$,

$$\mu_h \left(B_\epsilon^h \right) \leq C\epsilon^{\gamma} \quad \text{where} \quad B_\epsilon^h := \left\{ \boldsymbol{x} \in [0,1]^d : \operatorname{dist}(\boldsymbol{x}, \partial\Omega_h) \leq \epsilon \right\}$$

is the ball of radius $\epsilon > 0$ around $\partial\Omega_h$ with respect to the Euclidean distance

$$\operatorname{dist}(\boldsymbol{x}, \partial\Omega_h) := \inf_{\boldsymbol{x}' \in \partial\Omega_h} \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|,$$

and γ is called the margin exponent Christmann & Steinwart (2008); Kim et al. (2021).

Additionally, we define the measure $\boldsymbol{\mu}$ on Λ with marginal λ on $[0,1]^d$ and

$$\boldsymbol{\mu}\left([0,1]^d \times \{i\}\right) = \boldsymbol{\mu}_h\left([0,1]^d \times \{i\}\right) = \nu_i \quad \text{for all} \quad i \in \{0,1\}. \tag{3}$$

Under the above conditions, we establish lower bounds for the minimax error associated with the problem of estimating binary classifiers when the learning set satisfies C1) on a function class $\mathscr{C}$ with regularity C2), and the data distribution satisfies the margin condition C3); i.e., we lower bound the following inf-sup expression

$$\mathcal{I}_n(\mathscr{C}) := \inf_{A \in \mathcal{A}_n(L^2(\lambda))} \sup_{\substack{h \in H_{\mathscr{C}} \\ \mu_h \text{ satisfies C3)}}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h} \|A(\boldsymbol{S}_h) - h\|^2_{L^2(\mu_h)}, \tag{4}$$

where $\boldsymbol{S}_h := ((\boldsymbol{x}_i, h(\boldsymbol{x}_i)))_{i=1}^n$ is the sample of size $n \in \mathbb{N}$, for a function $h : [0,1]^d \to \{0,1\}$, and

$$\mathcal{A}_n(\mathcal{G}) := \{A : \Lambda^n \to \mathcal{G} : A \text{ is measurable}\} \quad \text{with} \quad \mathcal{G} \subseteq L^2(\lambda) := L^2([0,1]^d, \lambda).$$

Remark 2. *In particular, the margin condition C3) is satisfied when*

$$f_h(\boldsymbol{x}) \lesssim \begin{cases} \min\left\{\frac{\epsilon^\gamma}{\lambda(B_\epsilon^h)}, 1\right\} & \text{if } \boldsymbol{x} \in B_\epsilon^h \\ 1 & \text{otherwise,} \end{cases}$$

for almost every $\boldsymbol{x} \in [0,1]^d$ and for all $\epsilon > 0$. Moreover, C3) holds when

$$f_h(\boldsymbol{x}) \lesssim \operatorname{dist}^\gamma(\boldsymbol{x}, \partial\Omega_h) \quad \text{for almost every} \quad [0,1]^d.$$

1.2 Entropy

By Hurewicz & Wallman (1948); Kolmogorov & Tihomirov (1993), it is known that it is possible to characterize the "mass" of sets in metric spaces through the scaling law of the number of balls necessary to cover them with variable radii. Moreover, this characterization can be extended to spaces of densities with respect to distances that are not necessarily a metric, such as the distance associated with the Kullback-Leibler divergence or the Hellinger distance (see Yang & Barron (1999)). This leads to the definition of the concept of ε-covering entropy (or ε-entropy of Kolmogorov) which serves as a measure of complexity for sets, where a complex set is one that needs many elements to cover it and a simple set does not.

Let $\mathcal{X} \neq \emptyset$ be a set and let $\rho : \mathcal{X} \times \mathcal{X} \to [0,\infty]$ be a distance function (without the need to be symmetric or satisfy some form of triangle inequality). For a set $\emptyset \neq K \subset \mathcal{X}$ and $\varepsilon > 0$, it is said that:

(N) $G(\varepsilon) \subseteq \mathcal{X}$ is an ε-net of K, when for all $x \in K$, there exists $y \in G(\varepsilon) \subset \mathcal{X}$, such that $\rho(x,y) \leq \varepsilon$.

We define

$$V_{K,\rho}(\varepsilon) := \log \min\{|G_\varepsilon| : G_\varepsilon \text{ satisfies (N)}\} \text{ as the } \varepsilon\text{-covering entropy of } K. \tag{5}$$

Moreover, a similar notion can be defined as follows.

(P) $G(\varepsilon) \subseteq K$ is an ε-packing (or ε-distinguishable) set in K, when for all $x_1, x_2 \in G(\varepsilon)$, it holds that $\rho(x_1, x_2) > \varepsilon$.

We define (the ε-capacity in Kolmogorov & Tihomirov (1993))

$$M_{K,\rho}(\varepsilon) := \log \max\{|G_\varepsilon| : G_\varepsilon \text{ satisfies (P)}\} \text{ as the } \varepsilon\text{-packing entropy of } K. \tag{6}$$

Indeed, when ρ is a metric, (5) and (6) are equivalent in the sense that (see Kolmogorov & Tihomirov (1993) and (Petersen & Voigtlaender, 2021, Remark 3.10))

$$M_{K,\rho}(2\varepsilon) \leq V_{K,\rho}(\varepsilon) \leq M_{K,\rho}(\varepsilon). \tag{7}$$

To conclude, we introduce the following distance functions between probability densities that will be applied to the notions of entropy. We define the Hellinger distance ρ_H and the associated distance of the Kullback-Leibler divergence ρ_{KL}, as follows. Let

$$D_{\boldsymbol{\mu}} := \left\{ f : \Lambda \to [0,\infty) : f \text{ measurable and } \int_\Lambda f \, d\boldsymbol{\mu} = 1 \right\}$$

be the set of all probability densities functions (with respect to $\boldsymbol{\mu}$) on Λ. The distances ρ_H and ρ_{KL}, between $p, q \in D_{\boldsymbol{\mu}}$ are defined as

$$\rho_H(p,q) := \left(\int_\Lambda (\sqrt{p} - \sqrt{q})^2 d\boldsymbol{\mu} \right)^{1/2} \quad \text{and} \quad \rho_{KL}(p,q) := \left(\int_\Lambda p \log(p/q) d\boldsymbol{\mu} \right)^{1/2}. \tag{8}$$

1.3 Entropy bounds for some function classes

We introduce some results on entropy bounds under the L^p norm with $p \geq 1$, for classes of functions previously studied in Guntuboyina & Sen (2013), Kolmogorov & Tihomirov (1993) and Petersen & Voigtlaender (2021). Our goal is to apply Theorem 6 to classes such as $\mathscr{C} \subset C([0,1]^{d-1}; [0,1])$ in C1).

1.3.1 Barron regular functions

A function $f : [0,1]^{d-1} \to \mathbb{R}$ is said to be of Barron class (see Barron (1993); Caragea et al. (2023); Petersen & Voigtlaender (2021)) with constant $C > 0$, if there are $c \in [-C, C]$ and a measurable function $F : \mathbb{R}^{d-1} \to \mathbb{C}$ satisfying

$$\int_{\mathbb{R}^{d-1}} |F(\boldsymbol{\xi})| \sup_{\boldsymbol{x} \in [0,1]^{d-1}} |\langle \boldsymbol{\xi}, \boldsymbol{x} \rangle| \, d\boldsymbol{\xi} \leq C \quad \text{and} \quad f(\boldsymbol{x}) = c + \int_{\mathbb{R}^{d-1}} (e^{i\langle \boldsymbol{x}, \boldsymbol{\xi} \rangle} - 1) \cdot F(\boldsymbol{\xi}) d\boldsymbol{\xi} \tag{9}$$

for all $\boldsymbol{x} \in [0,1]^{d-1}$. The set of all these Barron functions is denoted as $\mathcal{B}_C$. Propositions 4.4 and 4.6 in Petersen & Voigtlaender (2021) state the following lemma.

Lemma 3. *For all $0 < \varepsilon < 1$,*

$$\varepsilon^{-\frac{2(d-1)}{d+1}} \lesssim M_{\mathcal{B}_C, L^1([0,1]^{d-1})}(\varepsilon) \leq M_{\mathcal{B}_C, L^\infty([0,1]^{d-1})}(\varepsilon) \lesssim \varepsilon^{-\frac{2(d-1)}{d+1}} (1 + \log(1/\varepsilon)).$$

1.3.2 α-Hölder continuous functions

According to Clements (1963) and (Kolmogorov & Tihomirov, 1993, Sections §5, §9), let $\mathcal{H}_\alpha$ be the class of functions f from $[0,1]^{d-1}$ to $[0,1]$, that are bounded by a constant C and have all partial derivatives of the order $i \leq k := \beta - \alpha \in \mathbb{Z}_{\geq 0}$ with the kth order derivatives satisfying the α-Hölder condition with exponent $\alpha \in (0,1]$, i.e.

$$\left| f^{(i)}(\boldsymbol{x}_1) - f^{(i)}(\boldsymbol{x}_2) \right| \leq L \, \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_\infty^\alpha := L \left[\max_{j \in \{1,\ldots,d-1\}} |x_{1,j} - x_{2,j}| \right]^\alpha,$$

for some constant $L > 0$, where $\boldsymbol{x}_i := (x_{i,1}, \ldots, x_{i,d-1}) \in [0,1]^{d-1}$, for $i \in \{1,2\}$. Then, Clements (1963) and (Kolmogorov & Tihomirov, 1993, §9-1) and Theorem V) imply the next result.

Lemma 4. *For all $\alpha > 0$ and $p \in \{1, \infty\}$, $V_{\mathcal{H}_\alpha, L^p([0,1]^{d-1})}(\varepsilon) \approx \varepsilon^{-(d-1)/\alpha}$.*

1.3.3 Convex functions

We denote by $\mathcal{C}_B([a,b]^{d-1})$ the set of all convex functions on $[a,b]^{d-1}$ with $a < b$ and $d \geq 2$, that are uniformly bounded by B. For this class of functions, the ε-covering entropy has been studied when $d = 2$ in Dryanov (2009); by adding an uniformly Lipschitz condition in Bronshtein (1976); and without extra conditions for all $d \geq 2$, in Guntuboyina & Sen (2013). We present the following lemma, which is a consequence of Theorems 3.1 and 3.3 in Guntuboyina & Sen (2013), for the case that $a = 0$ and $b = 1$.

Lemma 5. *For all $p \in [1, \infty)$, $B > 0$ and $0 < \varepsilon \lesssim B$, it is fulfilled that*

$$V_{\mathcal{C}_B, L^p([0,1]^{d-1})}(\varepsilon) \approx \varepsilon^{-(d-1)/2},$$

where $\mathcal{C}_B := \mathcal{C}_B([0,1]^{d-1})$ and $d \geq 2$.

1.4 PREVIOUS WORKS AND OUR CONTRIBUTION

Some related work on binary classifiers under the margin condition C3) can be found in: (Christmann & Steinwart, 2008, Section 8), for SVMs where learning rates were sometimes as fast as n^{-1}, being n the number of data points; Kim et al. (2021), based on neural networks with hinge loss that achieves fast convergence rates when $d \lesssim \gamma$, particularly as fast as $n^{-(q+1)/(q+2)}$ when the margin exponent $\gamma \to \infty$, where q is a noise parameter; and García & Petersen (2025), where it is found that using ReLU neural networks, the strong margin conditions imply fast learning bounds that are close to $n^{-1}(1+\log n)$. Furthermore, in García & Petersen (2025); Kim et al. (2021), a regularity condition is assumed on the decision boundary, where $\partial\Omega$ can be described by classes of functions $\mathscr{C}$ satisfying condition C1) on the elements of some covering for the set Ω. In Kim et al. (2021), the class of functions $\mathscr{C} \subset \mathcal{H}_\alpha$ is considered, and in García & Petersen (2025), the class $\mathscr{C} = \mathcal{B}_C$. The last two works mentioned above, only provide upper bounds for the learning rate when binary classifiers are approximated by neural networks under the margin condition. Our contribution now, by finding minimax lower bounds for learning rates on binary estimators, is to confirm that these learning rates are indeed optimal. It is important to highlight that the margin plays a main role in the fast learning rates obtained in each function space, when γ is sufficiently large the curse of dimensionality is overcome.

In Yang & Barron (1999), it was shown that through the entropy in density spaces defined with distances as in (8), it is possible to determine lower bounds for the learning rate of binary estimators (see Lemma 17). Then, in Petersen & Voigtlaender (2021), a relation between the distances (8) in a specific set of densities and the norms $L^2(\boldsymbol{\lambda})$ (being $\boldsymbol{\lambda}$ the Lebesgue measure on $[0,1]^d$) in $H_{\mathscr{C}}$ and $L^1([0,1]^{d-1})$ in $\mathscr{C}$ was demonstrated, thus adapting the main results of Yang & Barron (1999) to function spaces $\mathscr{C}$ as in C1). However, the margin condition was not assumed in either Petersen & Voigtlaender (2021) or Yang & Barron (1999).

In this paper, we use an argument to lower bound the minimax expression $\mathcal{I}_n(\mathscr{C})$ in (4) based on carefully constructing a particular family of probability densities (see Lemma (12)) over a subset of the function class $\mathscr{C}$, where the margin condition C3) is satisfied (see Remark 7). Thus, assuming that conditions C1), C2), and C3) hold, our main results provide the following: In Theorem 6, under certain conditions (11), a general lower bound for $\mathcal{I}_n(\mathscr{C})$ is given when the class $\mathscr{C}$ has bounded entropies (5) and (6) in the Lebesgue norm L^1. In Corollary 8, we apply Theorem 6 when the class $\mathscr{C}$ has entropy bounded in a particular way; then we obtain specific lower bounds for $\mathcal{I}_n(\mathscr{C})$ that depend on the sample size n, the parameters controlling the entropies, the margin exponent γ in C3) and the α assumed in the regularity C2) of $\mathscr{C}$. Finally, in Corollary 9, we provide explicit lower bounds for $\mathscr{C}$ being:

- The Barron class, $\mathscr{C} = \mathcal{B}_C$, for which we obtain in (13) lower bounds that, for all $\gamma \geq 3$ (see Remark 10) take the form

$$\mathcal{I}_n(\mathcal{B}_C) \gtrsim \left[n \left(1 + 2\log n\right)^{\frac{(d+1)\gamma}{2(d-1)}} \right]^{-\frac{\gamma}{\gamma + \left(\frac{2(d-1)}{d+1}\right)}},$$

  where $2(d-1)/(d+1) \to 2$ when $d \to \infty$. While in García & Petersen (2025) were obtained upper bounds for the learning rate as

$$\mathcal{I}_n(\mathcal{B}_C) \lesssim n^{-\gamma/(2+\gamma)}(1+\log n), \quad \text{for all} \quad \gamma > 0.$$

  In conclusion, when $\gamma \geq 3$, our lower bound matches the upper bound in García & Petersen (2025) up to logarithmic factors. For large dimension d, the exponent of n in both bounds is $-\gamma/(2+\gamma)$, showing that the learning rate in García & Petersen (2025) is asymptotically optimal.

- The α-Hölder continuous class, $\mathscr{C} = \mathcal{H}_\alpha$, where in (14) we found for all $\gamma \geq 3\alpha$ (see Remark 10), that

$$\mathcal{I}_n(\mathcal{H}_\alpha) \gtrsim n^{-\frac{1}{1+(d-1)/\gamma}}, \quad \text{for all} \quad \alpha \in (0,1].$$

  In contrast, under the noiseless assumption (see Remark 1), an upper bound was obtained in (Kim et al., 2021, Theorem 3.4), as follows

$$\left(\frac{\log^3 n}{n}\right)^{\frac{\alpha}{\alpha+(d-1)/\gamma}}, \quad \text{for all} \quad \gamma \geq 1 \quad \text{and all} \quad \alpha \in (0,1].$$

For values of α close to 1, the exponents of n in the lower and upper bounds become similar, indicating that the corresponding learning rates are nearly the same, up to logarithmic factors. In particular, for Lipschitz functions ($\alpha = 1$), the learning rate is asymptotically optimal. However, for small values of α, the lower and upper bounds themselves differ significantly.

- The Convex class, $\mathscr{C} = \mathcal{C}_B$, for which in (15), we obtain for all $\gamma \geq 3$ (see Remark 10),

$$\mathcal{I}_n(\mathcal{C}_B) \gtrsim n^{-\frac{2\gamma}{2\gamma+(d-1)}}.$$

When γ is large, the lower bound yields learning rates similar to those obtained for general Lipschitz functions in the previous item (case $\alpha = 1$).

2 Main results

In this section, we give the main results of our work. To begin with, we present the following theorem on general bounds for the minimax expression (4).

Theorem 6. *Let conditions C1), C2) and C3) hold. Let $V, M : (0,\infty) \to [0,\infty)$ be continuous from the right and non-increasing. If for all $\varepsilon \in (0,1)$, $\widetilde{\gamma} = \max\{2, \gamma/\alpha - 1\}$ and a constant $c > 1$, it is satisfied that*

$$V_{\mathscr{C},L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}} / \left(4\widetilde{\gamma}^2(2c)^{\frac{2}{\widetilde{\gamma}+1}}\right)\right) \leq V(\varepsilon) \quad \textit{and} \quad M_{\mathscr{C},L^1}\left((4\widetilde{\gamma})\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right) \geq M(\varepsilon). \tag{10}$$

Then, for every pair of sequences $\{\varepsilon_n\}_{n\in\mathbb{N}}, \{\widetilde{\varepsilon}_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ satisfying

$$n\varepsilon_n^2 = V(\varepsilon_n) = \left(M(\widetilde{\varepsilon}_n) - 2\log 2\right)/4, \tag{11}$$

we obtain $\widetilde{\varepsilon}_n^2 \leq (4\varepsilon_n)^2$ and

$$\mathcal{I}_n(\mathscr{C}) \geq \left[\frac{1}{8(4\widetilde{\gamma})^3}\right]\widetilde{\varepsilon}_n^2, \tag{12}$$

with $\mathcal{I}_n(\mathscr{C})$ as in (4)*.*

Remark 7. *For the proof of this theorem, we begin by defining a particular family of densities* (18) *that satisfy margin condition C3) over the subset $\mathscr{C}_\delta \subseteq \mathscr{C}$ defined in Lemma 12. Then, given that $\{\boldsymbol{x}_i\}_{i=1}^n$ is distributed according to the margin condition C3) with respect to a measure μ_h, in Lemma 13 we define a new class of densities $\mathcal{P}_{\mathscr{C}_\delta}$ over $\boldsymbol{\mu}$ for which the sample $\boldsymbol{S}_h = ((\boldsymbol{x}_i, h(\boldsymbol{x}_i)))_{i=1}^n$ is distributed. With this new set of densities $\mathcal{P}_{\mathscr{C}_\delta}$, we apply Lemma 17 ((Petersen & Voigtlaender, 2021, Theorem 3.11), which is a simplified version of (Yang & Barron, 1999, Theorems 1 and 2)), with a similar argument to the one used in Petersen & Voigtlaender (2021), to obtain our lower bound* (12)*. However, to apply Lemma 17 to the density class $\mathcal{P}_{\mathscr{C}_\delta}$, it was necessary to prove the results of Lemma 14 and Lemma 15.*

As a consequence of the previous theorem, when the entropy of the space $\mathscr{C}$ with respect to the Lebesgue L^1 norm is bounded in a particular way, we obtain the following corollary.

Corollary 8. *Let conditions C1), C2) and C3) hold. Let $\widetilde{\gamma} = \max\{2, \gamma/\alpha - 1\}$; $a \geq 1/2$; and $\mathcal{I}_n(\mathscr{C})$ as in* (4)*. Then, by cases:*

I) If $\varepsilon^{-a} \lesssim M_{\mathscr{C},L^1}(\varepsilon) \lesssim \varepsilon^{-a}(1 + \log(1/\varepsilon))$, for all $\varepsilon \in (0,1)$. Then

$$\mathcal{I}_n(\mathscr{C}) \gtrsim \left[n\left(1 + 2\log n\right)^{\frac{\widetilde{\gamma}+1}{a}}\right]^{-\frac{1}{1+\left(\frac{a}{\widetilde{\gamma}+1}\right)}}.$$

II) If $V_{\mathscr{C},L^1}(\varepsilon) \approx \varepsilon^{-a}$, for all $\varepsilon \in (0,1)$. Then

$$\mathcal{I}_n(\mathscr{C}) \gtrsim n^{-\frac{1}{1+\left(\frac{a}{\widetilde{\gamma}+1}\right)}}.$$

Finally, we apply the previous corollary to the spaces defined in Section 1.3 to obtain the next result.

Corollary 9. *Let $\gamma > 0$ be the margin exponent in C3). Then, with the notation* (4) *and the notation in Section 1.3, we obtain:*

I) *For Barron regular functions,*

$$\mathcal{I}_n(\mathcal{B}_C) \gtrsim \left[n \left(1 + 2 \log n\right)^{\frac{(d+1)\max\{3,\gamma\}}{2(d-1)}} \right]^{-\frac{\max\{3,\gamma\}}{\max\{3,\gamma\}+\left(\frac{2(d-1)}{d+1}\right)}}. \tag{13}$$

II) *For α-Hölder continuous functions,*

$$\mathcal{I}_n(\mathcal{H}_\alpha) \gtrsim n^{-\frac{\max\{3\alpha,\gamma\}}{\max\{3\alpha,\gamma\}+(d-1)}}, \tag{14}$$

for all $\alpha \in (0, 1]$.

III) *For convex functions,*

$$\mathcal{I}_n(\mathcal{C}_B) \gtrsim n^{-\frac{2\max\{3,\gamma\}}{2\max\{3,\gamma\}+(d-1)}}. \tag{15}$$

Remark 10. *In our results, the term $\max\{3\alpha, \gamma\}$ appears, which we had to assume since the proof of our main theorem is based on defining the particular family of densities given in* (18)*, for which one of the most important properties is that their square root is Lipschitz in b with respect to b on the $L^2(\lambda)$ (see* (20)*).*

In Kim et al. (2021) and García & Petersen (2025), upper bounds on the learning rate of approximating, by deep neural networks, binary classifiers with decision boundary satisfying the regularity condition C1) in function classes $\mathcal{H}_\alpha$ and $\mathcal{B}_\mathcal{C}$, respectively, were shown. See discussion of our results in comparison with those of Kim et al. (2021) and García & Petersen (2025) in Section 1.4.

Remark 11. *The main results Theorem 6, Corollary 8 and Corollary 9 identify precise optimal values for the learning rates of a binary classifier from noiseless samples, if a margin condition is satisfied and the decision boundary stems from a general class of functions. There are considerable limitations of these results that we would like to stress:*

1. *The established rates are understood in the inf-sup sense where all distributions that satisfy the margin condition are considered. In practice, many more favorable conditions on the data distribution may hold, which shows that the observed learning rates could be considerably faster.*

2. *As mentioned in Remark 1 we consider here noiseless data. This assumption is crucial to identify meaningful lower bounds, since it is otherwise unclear if the lower bounds stem from the noise or from the properties of the decision boundary (see (Petersen & Voigtlaender, 2021, Section 1.1, Point 1)). Still, noisy data appears extremely often in applications, and that situation is not covered by our main results.*

3. *Our approach requires precise bounds on the covering numbers of the spaces that model the decision boundary. We have described three such cases in this manuscript. However, for many, more exotic spaces, such estimates may not exist.*

REFERENCES

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500.

E. M. Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17: 393–398, 1976. doi: 10.1007/BF00967858.

Andrei Caragea, Philipp Petersen, and Felix Voigtlaender. Neural network approximation and estimation of classifiers with classification boundary in a Barron class. *The Annals of Applied Probability*, 33(4):3039 – 3079, 2023. doi: 10.1214/22-AAP1884.

Andreas Christmann and Ingo Steinwart. *Support Vector Machines*. Springer-Verlag New York, 2008. doi: 10.1007/978-0-387-77242-4.

G. F. Clements. Entropies of several sets of real valued functions. *Pacific Journal of Mathematics*, 13, 1963. doi: 10.2140/pjm.1963.13.1085.

D. Dryanov. Kolmogorov entropy for classes of convex functions. *Constructive Approximation*, 30: 137–153, 2009. doi: 10.1007/s00365-009-9053-3.

Jonathan García and Philipp Petersen. High-dimensional classification problems with barron regular boundaries under margin conditions. *arXiv:2412.07312*, 2025. doi: 10.48550/arXiv.2412.07312.

Adityanand Guntuboyina and Bodhisattva Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, 2013. doi: 10.1109/TIT.2012.2235172.

Witold Hurewicz and Henry Wallman. *Dimension theory*. Princeton University Press, 1948. URL https://mathscinet.ams.org/mathscinet-getitem?mr=6493.

Yongdai Kim, Ilsang Ohn, and Dongha Kim. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021. doi: 10.1016/j.neunet.2021.02.012.

A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional spaces. *Springer Nature*, pp. 86–170, 1993. doi: 10.1007/978-94-017-2973-4_7.

David Krieg and Mathias Sonnleitner. Random points are optimal for the approximation of sobolev functions. *IMA Journal of Numerical Analysis*, 44(3):1346–1371, 2023. doi: 10.1093/imanum/drad014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Philipp Petersen and Felix Voigtlaender. Optimal learning of high-dimensional classification problems using deep neural networks. *arXiv:2112.12555*, 2021. doi: 10.48550/arXiv.2112.12555.

Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982. doi: 10.1214/aos/1176345969.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564 – 1599, 1999. doi: 10.1214/aos/1017939142.

A Auxiliary Results

In this appendix, we present our auxiliary results used to prove Theorem 6.

Lemma 12. *Let $\mathscr{C}$ satisfy C2) with $\alpha \in (0,1]$ and $K > 0$. Then, the following statements hold.*

1) For all $\boldsymbol{x} \in [0,1]^d$,

$$\left(\frac{1}{\widetilde{K}}|b(\boldsymbol{x}^{(d)}) - x_d|\right)^{\frac{1}{\alpha}} \leq \operatorname{dist}(\boldsymbol{x}, \partial\Omega_{h_b}) \leq |b(\boldsymbol{x}^{(d)}) - x_d|, \tag{16}$$

where

$$1 < \widetilde{K} := \widetilde{K}(\alpha) = \begin{cases} \sqrt{1+K^2} & \text{if } \alpha = 1 \\ \max\{2^\alpha, 2K\} & \text{if } \alpha < 1. \end{cases} \tag{17}$$

2) Let $\delta \in (0, (4\widetilde{K})^{-\frac{1}{\alpha}}]$ and $\mathscr{C}_\delta := \{\delta b : b \in \mathscr{C}\}$. Let $\widetilde{f}_{h_b} : [0,1]^d \to [0,\infty)$ be defined by

$$\widetilde{f}_{h_b}(\boldsymbol{x}) := \begin{cases} |b(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} & \text{if } x_d \leq 1/2 \\ \widetilde{C}_b & \text{if } x_d > 1/2, \end{cases} \tag{18}$$

with $\widetilde{\gamma} := \max\left\{2, \frac{\gamma}{\alpha} - 1\right\}, \gamma > 0$ *and*

$$\widetilde{C}_b := 2\left(1 - \int_{[0,1]^{d-1}\times[0,1/2]} |b(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} d\boldsymbol{x}\right). \tag{19}$$

Then, $\mathscr{C}_\delta \subseteq \mathscr{C}$ *and for all* $b \in \mathscr{C}_\delta$*, we obtain that* $\widetilde{f}_{h_b}$ *is a Lebesgue density inducing a measure* $\widetilde{\mu}_{h_b}$ *that satisfies the margin condition C3), with margin exponent* γ *and for all* $0 < \epsilon \leq \delta$*. Moreover,* $\sqrt{\widetilde{f}_{h_b}}$ *is Lipschitz in* b *with respect to* b *on the* $L^2(\lambda)$ *norm, namely*

$$\left\|\sqrt{\widetilde{f}_{h_{b_1}}} - \sqrt{\widetilde{f}_{h_{b_2}}}\right\|_{L^2(\lambda)} \leq (3/8)\widetilde{\gamma}\left\|b_1 - b_2\right\|_{L^2([0,1]^{d-1})}, \quad \textit{for all} \quad b_1, b_2 \in \mathscr{C}_\delta. \tag{20}$$

Proof. By cases:

1) For all $\boldsymbol{x'} \in \partial\Omega_{h_b} = \left\{\boldsymbol{x} \in [0,1]^d : b(\boldsymbol{x}^{(d)}) = x_d\right\}$, and any $\boldsymbol{x} \in [0,1]^d$,

$$\|\boldsymbol{x} - \boldsymbol{x'}\|^2 = \left\|\boldsymbol{x}^{(d)} - \boldsymbol{x'}^{(d)}\right\|^2 + |b(\boldsymbol{x'}^{(d)}) - x_d|^2. \tag{21}$$

Let $t := \left\|\boldsymbol{x}^{(d)} - \boldsymbol{x'}^{(d)}\right\|$ and $w := |b(\boldsymbol{x}^{(d)}) - x_d|$. If $w - Kt^\alpha \leq 0$, we get $t^2 \geq (w/K)^{2/\alpha}$, and

$$\|\boldsymbol{x} - \boldsymbol{x'}\|^2 = t^2 + |b(\boldsymbol{x'}^{(d)}) - x_d|^2 \geq (w/K)^{2/\alpha}. \tag{22}$$

If $w - Kt^\alpha \geq 0$, condition C2) implies

$$\begin{aligned} w = |b(\boldsymbol{x}^{(d)}) - x_d| &\leq |b(\boldsymbol{x'}^{(d)}) - x_d| + |b(\boldsymbol{x}^{(d)}) - b(\boldsymbol{x'}^{(d)})| \\ &\leq |b(\boldsymbol{x'}^{(d)}) - x_d| + K\left\|\boldsymbol{x}^{(d)} - \boldsymbol{x'}^{(d)}\right\|^\alpha \\ &= |b(\boldsymbol{x'}^{(d)}) - x_d| + Kt^\alpha, \end{aligned}$$

therefore

$$\|\boldsymbol{x} - \boldsymbol{x'}\|^2 = t^2 + |b(\boldsymbol{x'}^{(d)}) - x_d|^2 \geq t^2 + (w - Kt^\alpha)^2. \tag{23}$$

So, when $\alpha = 1$, it follows that

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{x'}\|^2 &\geq (1 + K^2)t^2 - 2wKt + w^2 \\ &\geq \frac{4w^2(1 + K^2) - (2wK)^2}{4(1 + K^2)} \\ &= \frac{w^2}{1 + K^2}. \end{aligned} \tag{24}$$

Otherwise, $\alpha < 1$ and we set a threshold $t_0 := (w/(2K))^{1/\alpha}$ to minimize the right hand side of (23) on t. By cases:

- If $t \leq t_0$, we know that $w - Kt^\alpha \geq w - Kt_0^\alpha = w/2 > 0$, therefore
$$\|\boldsymbol{x} - \boldsymbol{x'}\|^2 \geq t^2 + (w - Kt_0^\alpha)^2 = t^2 + (w/2)^2 \geq (w/2)^2. \tag{25}$$
- If $t > t_0$, we obtain
$$\|\boldsymbol{x} - \boldsymbol{x'}\|^2 \geq t^2 + (w - Kt^\alpha)^2 \geq t^2 > (w/(2K))^{2/\alpha}. \tag{26}$$

Thus, (22), (24), (25) and (26), imply that

$$\|\boldsymbol{x} - \boldsymbol{x'}\| \geq \left(\frac{1}{\widetilde{K}}|b(\boldsymbol{x}^{(d)}) - x_d|\right)^{\frac{1}{\alpha}},$$

with $\widetilde{K}$ as in (17). Then, by the above inequality and identity (21), we arrive at

$$\left(\frac{1}{\widetilde{K}}|b(\boldsymbol{x}^{(d)}) - x_d|\right)^{\frac{1}{\alpha}} \leq \inf_{\boldsymbol{x'}\in\partial\Omega_{h_b}} \|\boldsymbol{x} - \boldsymbol{x'}\| \leq |b(\boldsymbol{x}^{(d)}) - x_d|,$$

where the upper bound was obtained given that $(x_1, \ldots, x_{d-1}, b(\boldsymbol{x}^{(d)})) \in \partial\Omega_{h_b}$.

2) It follows from C2) that $\delta b \in \mathscr{C}$ for all $b \in \mathscr{C}$, since $\delta < 1$. So, $\mathscr{C}_\delta \subseteq \mathscr{C}$.

For any $b \in \mathscr{C}_\delta$, to show that $\widetilde{f}_{h_b}$ is a density function with respect to λ, we get

$$\begin{aligned}\int_{[0,1]^d} \widetilde{f}_{h_b} d\lambda &= \int_{[0,1]^{d-1}\times[0,1/2]} \widetilde{f}_{h_b} d\lambda + \int_{[0,1]^{d-1}\times[1/2,1]} \widetilde{f}_{h_b} d\lambda \\ &= \int_{[0,1]^{d-1}\times[0,1/2]} |b(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} d\boldsymbol{x} + \int_{[0,1]^{d-1}\times[1/2,1]} \widetilde{C}_b d\boldsymbol{x} \\ &= \int_{[0,1]^{d-1}\times[0,1/2]} |b(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} d\boldsymbol{x} + (1/2)\widetilde{C}_b \\ &= 1.\end{aligned}$$

Now, for all $b \in \mathscr{C}_\delta$, we see that $\widetilde{f}_{h_b}$ satisfies the margin condition C3) for all $0 < \epsilon \leq \delta$, as follows. Using (16), we obtain that

$$x_d - b(\boldsymbol{x}^{(d)}) \leq |b(\boldsymbol{x}^{(d)}) - x_d| \leq \widetilde{K} \left(\operatorname{dist}(\boldsymbol{x}, \partial\Omega_{h_b})\right)^\alpha \leq \widetilde{K}\epsilon^\alpha \leq \widetilde{K}\delta^\alpha, \quad \text{for all} \quad \boldsymbol{x} \in B_\epsilon^{h_b}.$$

Therefore,

$$x_d \leq \widetilde{K}\delta^\alpha + b(\boldsymbol{x}^{(d)}) \leq \widetilde{K}\delta^\alpha + \delta \leq 2\widetilde{K}\delta^\alpha \leq 2\widetilde{K}/(4\widetilde{K}) = 1/2 \quad \text{for all} \quad \boldsymbol{x} \in B_\epsilon^{h_b}.$$

Then, $B_\epsilon^{h_b} \subseteq [0,1]^{d-1} \times [0,1/2]$, and

$$\begin{aligned}\widetilde{\mu}_{h_b}(B_\epsilon^{h_b}) &= \widetilde{\mu}_{h_b}(B_\epsilon^{h_b} \cap ([0,1]^{d-1} \times [0,1/2])) \\ &= \int_{B_\epsilon^{h_b}} |b(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} d\lambda \\ &\leq \int_{\left\{\boldsymbol{x}\in[0,1]^d : |b(\boldsymbol{x}^{(d)}) - x_d| \leq \widetilde{K}\epsilon^\alpha\right\}} |b(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} d\lambda \\ &\leq (\widetilde{K}\epsilon^\alpha)^{\widetilde{\gamma}} \int_{\left\{\boldsymbol{x}\in[0,1]^d : |b(\boldsymbol{x}^{(d)}) - x_d| \leq \widetilde{K}\epsilon^\alpha\right\}} 1 d\lambda \\ &\leq \widetilde{K}^{\widetilde{\gamma}} \epsilon^{\max\{2\alpha, \gamma-\alpha\}} (2\widetilde{K}\epsilon^\alpha) \\ &\leq 2\widetilde{K}^{\widetilde{\gamma}+1} \epsilon^{\max\{3\alpha,\gamma\}} \\ &\leq \left(2\widetilde{K}^{\widetilde{\gamma}+1}\right) \epsilon^\gamma.\end{aligned}$$

This proves that condition C3) is satisfied.

Moreover, as a consequence of the mean value theorem, it is well known that

$$\begin{aligned}||u|^\beta - |v|^\beta| &\leq \beta \max\{|u|, |v|\}^{\beta-1} ||u| - |v|| \\ &\leq \beta |u - v| \quad \text{for all} \quad u, v \in [-1,1] \quad \text{and} \quad \beta \geq 1. \end{aligned} \tag{27}$$

So, we apply (27) by cases:

- Let $g_b(\boldsymbol{x}) := |b(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}/2}$. Then, for all $b_1, b_2 \in \mathscr{C}_\delta$ and all $\boldsymbol{x} \in [0,1]^d$,

$$\begin{aligned}|g_{b_1}(\boldsymbol{x}) - g_{b_2}(\boldsymbol{x})| &= \left| |b_1(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}/2} - |b_2(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}/2} \right| \\ &\leq (\widetilde{\gamma}/2) |b_1(\boldsymbol{x}^{(d)}) - b_2(\boldsymbol{x}^{(d)})|.\end{aligned}$$

Therefore,

$$\begin{aligned}\int_{[0,1]^{d-1}\times[0,1/2]} |g_{b_1}(\boldsymbol{x}) - g_{b_2}(\boldsymbol{x})|^2 d\boldsymbol{x} &\leq (\widetilde{\gamma}/2)^2 \int_{[0,1]^{d-1}\times[0,1/2]} |b_1(\boldsymbol{x}^{(d)}) - b_2(\boldsymbol{x}^{(d)})|^2 d\boldsymbol{x} \\ &= (\widetilde{\gamma}^2/8) \|b_1 - b_2\|^2_{L^2([0,1]^{d-1})}.\end{aligned}$$

Thus, g_b is Lipschitz with respect to b on their respective L^2 norms.

- For all $b_1, b_2 \in \mathscr{C}_\delta$,

$$\begin{aligned}\frac{1}{4}\left|\widetilde{C}_{b_1}-\widetilde{C}_{b_2}\right|^2 &= \left(\int_{[0,1]^{d-1}\times[0,1/2]} \left(g_{b_2}^2(\boldsymbol{x})-g_{b_1}^2(\boldsymbol{x})\right) d\boldsymbol{x}\right)^2 \\ &\le \int_{[0,1]^{d-1}\times[0,1/2]} \left(g_{b_2}(\boldsymbol{x})-g_{b_1}(\boldsymbol{x})\right)^2 \left(g_{b_2}(\boldsymbol{x})+g_{b_1}(\boldsymbol{x})\right)^2 d\boldsymbol{x} \\ &\le 4\int_{[0,1]^{d-1}\times[0,1/2]} \left|g_{b_2}(\boldsymbol{x})-g_{b_1}(\boldsymbol{x})\right|^2 d\boldsymbol{x} \\ &\le (\widetilde{\gamma}^2/2)\left\|b_1-b_2\right\|^2_{L^2([0,1]^{d-1})}.\end{aligned}$$

  Also,

$$\begin{aligned}\left(\sqrt{\widetilde{C}_{b_1}}-\sqrt{\widetilde{C}_{b_2}}\right)^2 &\le \frac{1}{4}\left(\sqrt{\widetilde{C}_{b_1}}+\sqrt{\widetilde{C}_{b_2}}\right)^2\left(\sqrt{\widetilde{C}_{b_1}}-\sqrt{\widetilde{C}_{b_2}}\right)^2 \\ &= \frac{1}{4}\left|\widetilde{C}_{b_1}-\widetilde{C}_{b_2}\right|^2,\end{aligned}$$

  since for all $i \in \{1,2\}$,

$$\widetilde{C}_{b_i} \ge 2\left(1-\int_{[0,1]^{d-1}\times[0,1/2]} 1 d\boldsymbol{x}\right) = 1. \tag{28}$$

  Then,

$$\left(\sqrt{\widetilde{C}_{b_1}}-\sqrt{\widetilde{C}_{b_2}}\right)^2 \le (\widetilde{\gamma}^2/2)\left\|b_1-b_2\right\|^2_{L^2([0,1]^{d-1})}.$$

  Using the items above, we conclude that

$$\begin{aligned}&\left\|\sqrt{\widetilde{f}_{h_{b_1}}}-\sqrt{\widetilde{f}_{h_{b_2}}}\right\|^2_{L^2(\lambda)} \\ &= \int_{[0,1]^{d-1}\times[0,1/2]} \left|g_{b_1}(\boldsymbol{x})-g_{b_2}(\boldsymbol{x})\right|^2 d\boldsymbol{x} + \int_{[0,1]^{d-1}\times[1/2,1]} \left(\sqrt{\widetilde{C}_{b_1}}-\sqrt{\widetilde{C}_{b_2}}\right)^2 d\boldsymbol{x} \\ &\le (\widetilde{\gamma}^2/8)\left\|b_1-b_2\right\|^2_{L^2([0,1]^{d-1})} + (\widetilde{\gamma}^2/4)\left\|b_1-b_2\right\|^2_{L^2([0,1]^{d-1})} \\ &\le (3/8)\widetilde{\gamma}^2\left\|b_1-b_2\right\|^2_{L^2([0,1]^{d-1})}, \quad \text{for all} \quad b_1, b_2 \in \mathscr{C}_\delta.\end{aligned}$$

  Therefore (20) is fulfilled. □

Lemma 13. *Let $\mathcal{P}_{\mathscr{C}_\delta} := \{p_h : h \in H_{\mathscr{C}_\delta}\}$, where*

$$p_h(\boldsymbol{x}, y) := \begin{cases} h(\boldsymbol{x})\widetilde{f}_h(\boldsymbol{x})/\nu_1 & \text{if } y = 1 \\ (1-h(\boldsymbol{x}))\,\widetilde{f}_h(\boldsymbol{x})/\nu_0 & \text{if } y = 0, \end{cases} \tag{29}$$

and $\widetilde{f}_h$ is a Lebesgue density that defines a measure $\widetilde{\mu}_h$ as in (18)*. Then, p_h is a density function with respect to the measure $\boldsymbol{\mu}$ as in* (3)*. Furthermore,*

$$\boldsymbol{x} \sim \widetilde{\mu}_h \;\; \textit{if only if} \;\; (\boldsymbol{x}, h(\boldsymbol{x})) \sim p_h \quad \textit{(with respect to } \boldsymbol{\mu}\textit{)}. \tag{30}$$

Proof. For all $h \in H_{\mathscr{C}_\delta}$, we know that $p_h \in D_{\boldsymbol{\mu}}$, since

$$\begin{aligned}\int_\Lambda p_h(\boldsymbol{x}, y) d\boldsymbol{\mu} &= \nu_0 \int_{[0,1]^d} p_h(\boldsymbol{x}, 0) d\lambda + \nu_1 \int_{[0,1]^d} p_h(\boldsymbol{x}, 1) d\lambda \\ &= \int_{[0,1]^d} (1-h(\boldsymbol{x}))\, d\widetilde{\mu}_h + \int_{[0,1]^d} h(\boldsymbol{x}) d\widetilde{\mu}_h \\ &= \int_{[0,1]^d} 1 d\widetilde{\mu}_h = 1.\end{aligned}$$

In addition, for $(\boldsymbol{x}, h(\boldsymbol{x})) \in M \subset \Lambda$ and $\boldsymbol{x} \sim \widetilde{\mu}_h$, we have that

$$\begin{aligned}
\int_{[0,1]^d} \mathbb{1}_M(\boldsymbol{x}, h(\boldsymbol{x}))d\widetilde{\mu}_h &= \int_{[0,1]^d} \mathbb{1}_M(\boldsymbol{x}, \mathbb{1}_{\Omega_h}(\boldsymbol{x}))d\widetilde{\mu}_h \\
&= \int_{[0,1]^d \setminus \Omega_h} \mathbb{1}_M(\boldsymbol{x}, 0)d\widetilde{\mu}_h + \int_{\Omega_h} \mathbb{1}_M(\boldsymbol{x}, 1)d\widetilde{\mu}_h \\
&= \int_{[0,1]^d} \mathbb{1}_M(\boldsymbol{x}, 0)\left(1 - \mathbb{1}_{\Omega_h}(\boldsymbol{x})\right)d\widetilde{\mu}_h + \int_{[0,1]^d} \mathbb{1}_M(\boldsymbol{x}, 1)\mathbb{1}_{\Omega_h}(\boldsymbol{x})d\widetilde{\mu}_h \\
&= \int_{[0,1]^d} \mathbb{1}_M(\boldsymbol{x}, 0)\left(1 - h(\boldsymbol{x})\right)d\widetilde{\mu}_h + \int_{[0,1]^d} \mathbb{1}_M(\boldsymbol{x}, 1)h(\boldsymbol{x})d\widetilde{\mu}_h \\
&= \nu_0 \int_{[0,1]^d} \mathbb{1}_M(\boldsymbol{x}, 0)p_h(\boldsymbol{x}, 0)d\lambda + \nu_1 \int_{[0,1]^d} \mathbb{1}_M(\boldsymbol{x}, 1)p_h(\boldsymbol{x}, 1)d\lambda \\
&= \int_{\Lambda} \mathbb{1}_M(\boldsymbol{x}, y)p_h(\boldsymbol{x}, y)d\boldsymbol{\mu},
\end{aligned}$$

equivalent to $(\boldsymbol{x}, h(\boldsymbol{x})) \sim p_h$ with respect to $\boldsymbol{\mu}$ and (30) is satisfied. □

Lemma 14. *Let $h_1, h_2 \in H_{\mathscr{C}_\delta}$ be defined by $b_1, b_2 \in \mathscr{C}_\delta$, respectively. Then,*

$$\|h_1 - h_2\|^2_{L^2(\widetilde{\mu}_{h_i})} = \frac{1}{\widetilde{\gamma}+1}\|b_1 - b_2\|^{\widetilde{\gamma}+1}_{L^{\widetilde{\gamma}+1}([0,1]^{d-1})} \quad \textit{for all} \quad i \in \{1,2\}, \tag{31}$$

with $\widetilde{\mu}_{h_i}$ and $\widetilde{\gamma}$ as in Lemma 12. Moreover, for all $p_{h_1}, p_{h_2} \in \mathcal{P}_{\mathscr{C}_\delta}$, we have that

$$\|h_1 - h_2\|^2_{L^2(\widetilde{\mu}_{h_i})} \leq \rho^2_H(p_{h_1}, p_{h_2}) \leq 2\widetilde{\gamma}^2\|b_1 - b_2\|_{L^1([0,1]^{d-1})}, \tag{32}$$

for all $i \in \{1,2\}$.

Proof. For all $i \in \{1,2\}$, we know that $b_i \in \mathscr{C}_\delta$ and therefore $b_i(\boldsymbol{x}^{(d)}) \leq \delta$ for all $\boldsymbol{x}^{(d)} \in [0,1]^{d-1}$. Also, $\delta \in (0, (4\widetilde{K})^{-\frac{1}{\alpha}}]$, so $\delta \leq (4\widetilde{K})^{-1} < 1/2$. Then, by definition (1),

$$\begin{aligned}
G_{1,2} &:= \left\{\boldsymbol{x} \in [0,1]^d : h_1(\boldsymbol{x}) \neq h_2(\boldsymbol{x})\right\} \\
&= \left\{\boldsymbol{x} \in [0,1]^d : h_1(\boldsymbol{x}) = 0 \wedge h_2(\boldsymbol{x}) = 1\right\} \cup \left\{\boldsymbol{x} \in [0,1]^d : h_1(\boldsymbol{x}) = 1 \wedge h_2(\boldsymbol{x}) = 0\right\} \\
&= \left\{\boldsymbol{x} \in [0,1]^d : b_2(\boldsymbol{x}^{(d)}) \leq x_d < b_1(\boldsymbol{x}^{(d)})\right\} \cup \left\{\boldsymbol{x} \in [0,1]^d : b_1(\boldsymbol{x}^{(d)}) \leq x_d < b_2(\boldsymbol{x}^{(d)})\right\} \\
&= \left\{\boldsymbol{x} \in [0,1]^d : \min_{i\in\{1,2\}}\left\{b_i(\boldsymbol{x}^{(d)})\right\} \leq x_d < \max_{i\in\{1,2\}}\left\{b_i(\boldsymbol{x}^{(d)})\right\} < 1/2\right\} \\
&\subseteq [0,1]^{d-1} \times [0, 1/2].
\end{aligned} \tag{33}$$

Furthermore, h_1 and h_2 are indicator functions, thus $|h_1 - h_2| = \mathbb{1}_{G_{1,2}}$, and using (33), we get

$$\begin{aligned}
\|h_1 - h_2\|^2_{L^2(\widetilde{\mu}_{h_i})} &= \int_{[0,1]^d} |h_1 - h_2|^2 \widetilde{f}_{h_i} d\lambda \\
&= \int_{G_{1,2}} |b_i(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} d\boldsymbol{x} \\
&= \frac{1}{\widetilde{\gamma}+1}\int_{[0,1]^{d-1}} |b_1(\boldsymbol{x}^{(d)}) - b_2(\boldsymbol{x}^{(d)})|^{\widetilde{\gamma}+1} d\boldsymbol{x}^{(d)} \\
&= \frac{1}{\widetilde{\gamma}+1}\|b_1 - b_2\|^{\widetilde{\gamma}+1}_{L^{\widetilde{\gamma}+1}([0,1]^{d-1})} \quad \text{for all} \quad i \in \{1,2\}.
\end{aligned}$$

From the above identity, it follows (31). Moreover, $h_i = \sqrt{h_i}$ and $1 - h_i = \sqrt{1-h_i}$, for all $i \in \{1, 2\}$. Hence,

$$\begin{aligned}
\rho_H^2(p_{h_1}, p_{h_2}) - 2 &= -2 \int_\Lambda \sqrt{p_{h_1} p_{h_2}}\, d\boldsymbol{\mu} \\
&= -2 \left[\int_{[0,1]^d \times \{1\}} \sqrt{p_{h_1} p_{h_2}}\, d\boldsymbol{\mu} + \int_{[0,1]^d \times \{0\}} \sqrt{p_{h_1} p_{h_2}}\, d\boldsymbol{\mu} \right] \\
&= -2 \left[\int_{[0,1]^d} \sqrt{\widetilde{f}_{h_1} \widetilde{f}_{h_2}} \left(\sqrt{h_1 h_2} + \sqrt{(1-h_1)(1-h_2)} \right) d\lambda \right] \\
&= 2 \left[\int_{[0,1]^d} \sqrt{\widetilde{f}_{h_1} \widetilde{f}_{h_2}}\, (h_1 - 2h_1 h_2 + h_2 - 1)\, d\lambda \right] \\
&= 2 \left[\int_{[0,1]^d} \sqrt{\widetilde{f}_{h_1} \widetilde{f}_{h_2}} \left((\sqrt{h_1} - \sqrt{h_2})^2 - 1 \right) d\lambda \right] \\
&= 2 \left[\int_{[0,1]^d} \sqrt{\widetilde{f}_{h_1} \widetilde{f}_{h_2}}\, (|h_1 - h_2| - 1)\, d\lambda \right].
\end{aligned}$$

We see that

$$\begin{aligned}
2 - 2 \int_{[0,1]^d} \sqrt{\widetilde{f}_{h_1} \widetilde{f}_{h_2}}\, d\lambda &= \int_{[0,1]^d} \widetilde{f}_{h_1} d\lambda - 2 \int_{[0,1]^d} \sqrt{\widetilde{f}_{h_1} \widetilde{f}_{h_2}}\, d\lambda + \int_{[0,1]^d} \widetilde{f}_{h_2} d\lambda \\
&= \int_{[0,1]^d} \left(\sqrt{\widetilde{f}_{h_1}} - \sqrt{\widetilde{f}_{h_2}} \right)^2 d\lambda.
\end{aligned}$$

Then the above two identities imply

$$\rho_H^2(p_{h_1}, p_{h_2}) = 2 \int_{[0,1]^d} \sqrt{\widetilde{f}_{h_1} \widetilde{f}_{h_2}} |h_1 - h_2|\, d\lambda + \int_{[0,1]^d} \left(\sqrt{\widetilde{f}_{h_1}} - \sqrt{\widetilde{f}_{h_2}} \right)^2 d\lambda. \tag{34}$$

Now we bound show (32) by cases:

- Upper bound: For all $i \in \{1, 2\}$, we know that $\widetilde{f}_{h_i} \leq 3$, since (19) and (28) implies

$$|b_i(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} \leq 1 \leq C_b \leq 2 \left(1 + \int_{[0,1]^{d-1} \times [0,1/2]} |b_i(\boldsymbol{x}^{(d)}) - x_d|^{\widetilde{\gamma}} d\boldsymbol{x} \right) \leq 3.$$

  We use (20), (33) and (34), to obtain

$$\begin{aligned}
\rho_H^2(p_{h_1}, p_{h_2}) &\leq 6 \int_{[0,1]^d} |h_1 - h_2|\, d\lambda + \left\| \sqrt{\widetilde{f}_{h_1}} - \sqrt{\widetilde{f}_{h_2}} \right\|_{L^2(\lambda)}^2 \\
&\leq 6 \int_{G_{1,2}} 1\, d\lambda + \widetilde{\gamma}^2 \left\| b_1 - b_2 \right\|_{L^2([0,1]^{d-1})}^2 \\
&= 6 \left\| b_1 - b_2 \right\|_{L^1([0,1]^{d-1})} + (3/8) \widetilde{\gamma}^2 \left\| b_1 - b_2 \right\|_{L^2([0,1]^{d-1})}^2,
\end{aligned}$$

  and we know that $|b_1(\boldsymbol{z}) - b_2(\boldsymbol{z})| \leq 1$ for all $\boldsymbol{z} \in [0,1]^{d-1}$, then

$$\rho_H^2(p_{h_1}, p_{h_2}) \leq (6 + (3/8)\widetilde{\gamma}^2) \left\| b_1 - b_2 \right\|_{L^1([0,1]^{d-1})} \leq 2\widetilde{\gamma}^2 \left\| b_1 - b_2 \right\|_{L^1([0,1]^{d-1})}.$$

- Lower bound: By (33), (34) and as $|h_1 - h_2| = \mathbb{1}_{G_{1,2}}$, we get

$$\begin{aligned}\rho_H^2(p_{h_1}, p_{h_2}) &= 2\int_{G_{1,2}} \sqrt{\widetilde{f}_{h_1}\widetilde{f}_{h_2}}\, d\lambda + \int_{([0,1]^d \setminus G_{1,2}) \cup G_{1,2}} \left(\sqrt{\widetilde{f}_{h_1}} - \sqrt{\widetilde{f}_{h_2}}\right)^2 d\lambda \\ &= \int_{G_{1,2}} (\widetilde{f}_{h_1} + \widetilde{f}_{h_2}) d\lambda + \int_{([0,1]^d \setminus G_{1,2})} \left(\sqrt{\widetilde{f}_{h_1}} - \sqrt{\widetilde{f}_{h_2}}\right)^2 d\lambda \\ &\geq \int_{[0,1]^d} |h_1 - h_2|^2 \widetilde{f}_{h_i} d\lambda \\ &= \|h_1 - h_2\|^2_{L^2(\widetilde{\mu}_{h_i})}, \quad \text{for all} \quad i \in \{1,2\}. \end{aligned}$$

□

Lemma 15. *Under the conditions C1), C2) and C3), let $\mathscr{C}(\xi) \subseteq \mathscr{C}_\delta \subseteq \mathscr{C}$ be a ξ-packing set as in (P), with respect to the L^1 norm, and $\xi \in (0,1)$. Let*

$$\mathscr{C}^\xi := \left\{\xi^{\widetilde{\gamma}} b : b \in \mathscr{C}(\xi)\right\} \subseteq \mathscr{C}(\xi) \quad \textit{and} \quad \mathcal{P}_{\mathscr{C}^\xi} := \left\{p_{h_b} \in \mathcal{P}_{\mathscr{C}_\delta} : b \in \mathscr{C}^\xi\right\},$$

with $\widetilde{\gamma} = \max\{2, \gamma/\alpha - 1\}$ as in Lemma 12. Then,

$$\mathcal{I}_n(\mathscr{C}) \geq \frac{1}{(4\widetilde{\gamma})^3}\left[\inf_{E \in \mathcal{A}_n(\mathcal{D}_{\boldsymbol{\mu}})} \sup_{p' \in \mathcal{P}_{\mathscr{C}^\xi}} \mathbb{E}_{\boldsymbol{S} \overset{iid}{\sim} p'} \rho_H^2(E(\boldsymbol{S}), p')\right], \tag{35}$$

where $\mathcal{I}_n(\mathscr{C})$ is defined in (4)*. Moreover,*

$$V_{\mathcal{P}_{\mathscr{C}^\xi}, \rho_H}(\varepsilon) \leq V_{\mathscr{C}, L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}/(4\widetilde{\gamma}^2)\right) \quad \textit{and} \quad M_{\mathcal{P}_{\mathscr{C}^\xi}, \rho_H}(\varepsilon) \geq M_{\mathscr{C}, L^1}\left((4\widetilde{\gamma})\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right), \tag{36}$$

for all $\varepsilon > 0$.

Proof. We begin by bounding $\mathcal{I}_n(\mathscr{C})$ from below. In order to show that

$$4\mathcal{I}_n(\mathscr{C}) \geq \widetilde{\mathcal{I}_n}(\mathscr{C}_\delta) := \inf_{A \in \mathcal{A}_n(H_{\mathscr{C}_\delta})} \sup_{\substack{h \in H_{\mathscr{C}_\delta} \\ \mu_h \text{ has density as in (18)}}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h} \|A(\boldsymbol{S}_h) - h\|^2_{L^2(\mu_h)} \tag{37}$$

is satisfied, we introduce (Petersen & Voigtlaender, 2021, Lemma 3.2).

Lemma 16. *Let (X, ρ) be a metric space, with distance function ρ, and let $\emptyset \neq M \subset X$ be separable. Then for each $\varepsilon > 0$ there exist a measurable map $\pi_\varepsilon : X \to M$ satisfying $\rho(x, \pi_\varepsilon(x)) \leq \varepsilon + \rho(x, M)$ for all $x \in M$.*

By condition C2), $\mathscr{C}_\delta$ is separable with respect to the L^2 norm, the continuous map $\mathscr{C}_\delta \to L^2(\lambda)$ defined by $b \to h = \mathbb{1}_{b(\boldsymbol{x}^{(d)}) \leq x_d}$ implies that $H_{\mathscr{C}_\delta}$ is separable. Then, for any $B \in \mathcal{A}_n\left(L^2(\lambda)\right)$ and $\varepsilon > 0$, Lemma 16 implies that there exists a map $\pi_\varepsilon : L^2(\lambda) \to H_{\mathscr{C}_\delta}$ such that

$$\|B(\boldsymbol{S}_h) - \pi_\varepsilon(B(\boldsymbol{S}_h))\|_{L^2(\mu_h)} \leq \varepsilon + \inf_{h \in H_{\mathscr{C}_\delta}} \|B(\boldsymbol{S}_h) - h\|_{L^2(\mu_h)} \leq \varepsilon + \|B(\boldsymbol{S}_h) - h\|_{L^2(\mu_h)},$$

for all $h \in H_{\mathscr{C}_\delta}$ and μ_h as in C3). Therefore,

$$\begin{aligned}\widetilde{\mathcal{I}_n}(\mathscr{C}_\delta) &\leq \sup_{\substack{h \in H_{\mathscr{C}_\delta} \\ \mu_h \text{ has density as in (18)}}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h} \|\pi_\varepsilon(B(\boldsymbol{S}_h)) - h\|^2_{L^2(\mu_h)} \\ &\leq \sup_{\substack{h \in H_{\mathscr{C}_\delta} \\ \mu_h \text{ satisfies C3)}}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h} \|\pi_\varepsilon(B(\boldsymbol{S}_h)) - h\|^2_{L^2(\mu_h)} \\ &\leq \sup_{\substack{h \in H_{\mathscr{C}_\delta} \\ \mu_h \text{ satisfies C3)}}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h} \left(\|\pi_\varepsilon(B(\boldsymbol{S}_h)) - B(\boldsymbol{S}_h)\|_{L^2(\mu_h)} + \|B(\boldsymbol{S}_h) - h\|_{L^2(\mu_h)}\right)^2 \\ &\leq \sup_{\substack{h \in H_{\mathscr{C}_\delta} \\ \mu_h \text{ satisfies C3)}}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h} \left(\varepsilon + 2\|B(\boldsymbol{S}_h) - h\|_{L^2(\mu_h)}\right)^2, \end{aligned}$$

where $\varepsilon \to 0$, implies

$$\widetilde{\mathcal{I}_n}(\mathscr{C}_\delta) \leq 4 \sup_{\substack{h \in H_{\mathscr{C}_\delta} \\ \mu_h \text{ satisfies C3)}}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h} \|B(\boldsymbol{S}_h) - h\|^2_{L^2(\mu_h)}.$$

Furthermore, B was arbitrary and $\mathscr{C}_\delta \subset \mathscr{C}$, so

$$\widetilde{\mathcal{I}_n}(\mathscr{C}_\delta) \leq 4 \inf_{B \in \mathcal{A}_n(L^2(\lambda))} \sup_{\substack{h \in H_{\mathscr{C}_\delta} \\ \mu_h \text{ satisfies C3)}}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \mu_h} \|B(\boldsymbol{S}_h) - h\|^2_{L^2(\mu_h)} \leq 4\mathcal{I}_n(\mathscr{C}),$$

and we obtain (37).

Since $\mathscr{C}(\xi) \subseteq \mathscr{C}_\delta$, it follows from (18) and the same argument used to get (37) with Lemma 16, that

$$\begin{aligned} 4\widetilde{\mathcal{I}_n}(\mathscr{C}_\delta) &= 4 \inf_{A \in \mathcal{A}_n(H_{\mathscr{C}_\delta})} \sup_{h_b \in H_{\mathscr{C}_\delta}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \widetilde{\mu}_{h_b}} \|A(\boldsymbol{S}_{h_b}) - h_b\|^2_{L^2(\widetilde{\mu}_{h_b})} \\ &\geq \inf_{A \in \mathcal{A}_n(H_{\mathscr{C}(\xi)})} \sup_{h_b \in H_{\mathscr{C}(\xi)}} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \widetilde{\mu}_{h_b}} \|A(\boldsymbol{S}_{h_b}) - h_b\|^2_{L^2(\widetilde{\mu}_{h_b})}, \end{aligned}$$

where $A \in \mathcal{A}_n(H_{\mathscr{C}(\xi)})$ implies $A(\cdot) \in H_{\mathscr{C}(\xi)}$, and there exists a unique $b_A \in \mathscr{C}(\xi)$ such that $A(\cdot) = h_{b_{A(\cdot)}}$. Then, by Lemma 14, we get

$$\|A(\boldsymbol{S}_{h_b}) - h_b\|^2_{L^2(\widetilde{\mu}_{h_b})} = \frac{1}{\widetilde{\gamma}+1} \left\| b_{A(\boldsymbol{S}_{h_b})} - b \right\|^{\widetilde{\gamma}+1}_{L^{\widetilde{\gamma}+1}([0,1]^{d-1})} \geq \frac{1}{2^{\widetilde{\gamma}}} \left\| b_{A(\boldsymbol{S}_{h_b})} - b \right\|^{\widetilde{\gamma}+1}_{L^1([0,1]^{d-1})}.$$

Therefore,

$$(8\widetilde{\gamma})\widetilde{\mathcal{I}_n}(\mathscr{C}_\delta) \geq \inf_{b_A \in \mathscr{C}(\xi)} \sup_{b \in \mathscr{C}(\xi)} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \widetilde{\mu}_{h_b}} \left\| b_{A(\boldsymbol{S}_{h_b})} - b \right\|^{\widetilde{\gamma}+1}_{L^1([0,1]^{d-1})}. \tag{38}$$

By hypothesis, $\mathscr{C}(\xi)$ is a ξ-packing set as in (P), with respect to the L^1 norm, which means that

$$\left\| b_{A(\boldsymbol{S}_{h_b})} - b \right\|_{L^1([0,1]^{d-1})} > \xi, \quad \text{for all} \quad b_A, b \in \mathscr{C}(\xi). \tag{39}$$

Then, by (38) and (39), together with $\mathscr{C}^\xi = \left\{\xi^{\widetilde{\gamma}} b : b \in \mathscr{C}(\xi)\right\}$, we obtain

$$\begin{aligned} (8\widetilde{\gamma})\widetilde{\mathcal{I}_n}(\mathscr{C}_\delta) &\geq \inf_{b_A \in \mathscr{C}(\xi)} \sup_{b \in \mathscr{C}(\xi)} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \widetilde{\mu}_{h_b}} \left\| b_{A(\boldsymbol{S}_{h_b})} - b \right\|_{L^1([0,1]^{d-1})} \xi^{\widetilde{\gamma}} \\ &= \inf_{b_A \in \mathscr{C}(\xi)} \sup_{b \in \mathscr{C}(\xi)} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \widetilde{\mu}_{h_b}} \left\| \xi^{\widetilde{\gamma}} b_{A(\boldsymbol{S}_{h_b})} - \xi^{\widetilde{\gamma}} b \right\|_{L^1([0,1]^{d-1})} \\ &= \inf_{b_A \in \mathscr{C}^\xi} \sup_{b \in \mathscr{C}^\xi} \mathbb{E}_{\{\boldsymbol{x}_i\}_{i=1}^n \overset{iid}{\sim} \widetilde{\mu}_{h_b}} \left\| b_{A(\boldsymbol{S}_{h_b})} - b \right\|_{L^1([0,1]^{d-1})}. \end{aligned} \tag{40}$$

From (32),

$$2\widetilde{\gamma}^2 \left\| b_{A(\boldsymbol{S}_{h_b})} - b \right\|_{L^1([0,1]^{d-1})} \geq \rho^2_H(p_{h_{b_{A(\boldsymbol{S}_{h_b})}}}, p_{h_b}),$$

for all $b_A, b \in \mathscr{C}^\xi$, i.e. for all $p_{h_{b_{A(\boldsymbol{S}_{h_b})}}}, p_{h_b} \in \mathcal{P}_{\mathscr{C}^\xi}$. We define the map $E \in \mathcal{A}_n(\mathcal{P}_{\mathscr{C}^\xi})$ such that $E(\cdot) := p_{A(\cdot)} = p_{h_{b_{A(\cdot)}}}$, and (30), (40) imply

$$\begin{aligned} (16\widetilde{\gamma}^3)\widetilde{\mathcal{I}_n}(\mathscr{C}_\delta) &\geq \inf_{E \in \mathcal{A}_n(\mathcal{P}_{\mathscr{C}^\xi})} \sup_{p_{h_b} \in \mathcal{P}_{\mathscr{C}^\xi}} \mathbb{E}_{\boldsymbol{S}_{h_b} \overset{iid}{\sim} p_{h_b}} \rho^2_H(E(\boldsymbol{S}_{h_b}), p_{h_b}) \\ &\geq \inf_{E \in \mathcal{A}_n(\mathcal{D}_{\boldsymbol{\mu}})} \sup_{p' \in \mathcal{P}_{\mathscr{C}^\xi}} \mathbb{E}_{\boldsymbol{S} \overset{iid}{\sim} p'} \rho^2_H(E(\boldsymbol{S}), p'), \end{aligned} \tag{41}$$

since $\mathcal{A}_n(\mathcal{P}_{\mathscr{C}^\xi}) \subseteq \mathcal{A}_n(\mathcal{P}_{\mathscr{C}_\delta}) \subseteq \mathcal{A}_n(\mathcal{D}_{\boldsymbol{\mu}})$. In conclusion, (35) follows from (37) and (41).

Now we continue with the second part of the lemma, showing that (36) holds as follows. From Lemma 14 we have that

$$\frac{1}{2^{\widetilde{\gamma}}} \|b_1 - b_2\|^{\widetilde{\gamma}+1}_{L^1([0,1]^{d-1})} \leq \rho^2_H(p_{h_{b_1}}, p_{h_{b_2}}) \leq 2\widetilde{\gamma}^2 \|b_1 - b_2\|_{L^1([0,1]^{d-1})}, \tag{42}$$

for all $p_{h_{b_1}}, p_{h_{b_2}} \in \mathcal{P}_{\mathscr{C}_\delta}$. Using the same argument as in (39) and the right hand side of (42), we obtain

$$\begin{aligned}\rho_H^2(p_{h_{b_1}}, p_{h_{b_2}}) &\leq 2\widetilde{\gamma}^2 \|b_1 - b_2\|_{L^1([0,1]^{d-1})} \\ &= 2\widetilde{\gamma}^2 \xi^{\widetilde{\gamma}} \|b_1' - b_2'\|_{L^1([0,1]^{d-1})} \\ &\leq 2\widetilde{\gamma}^2 \|b_1' - b_2'\|^{\widetilde{\gamma}+1}_{L^1([0,1]^{d-1})},\end{aligned}$$

for all $p_{h_{b_1}}, p_{h_{b_2}} \in \mathcal{P}_{\mathscr{C}^\xi}$ and all $b_1', b_2' \in \mathscr{C}(\xi)$ associated with $b_1, b_2 \in \mathscr{C}^\xi$. Therefore,

$$\begin{aligned}M_{\mathcal{P}_{\mathscr{C}^\xi}, \rho_H}(\varepsilon) &\leq M_{\mathscr{C}(\xi), L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}/(2\widetilde{\gamma}^2)\right) \\ &\leq M_{\mathscr{C}, L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}/(2\widetilde{\gamma}^2)\right) \quad \text{for all} \quad \varepsilon > 0. \end{aligned} \tag{43}$$

Moreover, from the left hand side of (42),

$$\frac{1}{2\widetilde{\gamma}} \|b_1 - b_2\|^{\widetilde{\gamma}+1}_{L^1([0,1]^{d-1})} \leq \rho_H^2(p_{h_{b_1}}, p_{h_{b_2}}),$$

for all $p_{h_{b_1}}, p_{h_{b_2}} \in \mathcal{P}_{\mathscr{C}^\xi}$, and

$$V_{\mathcal{P}_{\mathscr{C}^\xi}, \rho_H}(\varepsilon) \geq V_{\mathscr{C}^\xi, L^1}\left((2\widetilde{\gamma})\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right) \geq V_{\mathscr{C}, L^1}\left((2\widetilde{\gamma})\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right) \quad \text{for all} \quad \varepsilon > 0. \tag{44}$$

Then, (7), (43) and (44) imply

$$V_{\mathscr{C}, L^1}\left((2\widetilde{\gamma})\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right) \leq V_{\mathcal{P}_{\mathscr{C}^\xi}, \rho_H}(\varepsilon) \leq M_{\mathcal{P}_{\mathscr{C}^\xi}, \rho_H}(\varepsilon) \leq M_{\mathscr{C}, L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}/(2\widetilde{\gamma}^2)\right),$$

for all $\varepsilon > 0$. Finally, from (7) and the above inequality, we conclude that

$$\begin{aligned}V_{\mathcal{P}_{\mathscr{C}^\xi}, \rho_H}(\varepsilon) \leq M_{\mathscr{C}, L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}/(2\widetilde{\gamma}^2)\right) \leq V_{\mathscr{C}, L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}/(4\widetilde{\gamma}^2)\right) \quad \text{and} \\ M_{\mathcal{P}_{\mathscr{C}^\xi}, \rho_H}(\varepsilon) \geq V_{\mathscr{C}, L^1}\left((2\widetilde{\gamma})\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right) \geq M_{\mathscr{C}, L^1}\left((4\widetilde{\gamma})\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right),\end{aligned}$$

for all $\varepsilon > 0$. So, (36) is fulfilled. □

B Proof of the main results

B.1 Proof of Theorem 6

To begin with, we present (Petersen & Voigtlaender, 2021, Theorem 3.11), which is a simplified version of (Yang & Barron, 1999, Theorems 1 and 2).

Lemma 17. *Let $\emptyset \neq \mathcal{P} \subset D_{\boldsymbol{\mu}}$, and let $V, M : (0, \infty) \to [0, \infty)$ be continuous from the right, non-increasing, and with the following properties:*

1. *$V_{\mathcal{P}, \rho_{KL}}(\varepsilon) \leq V(\varepsilon)$ for all $\varepsilon > 0$;*
2. *$(\varepsilon_n)_{n \in \mathbb{N}} \subset (0, \infty)$ is chosen such that $\varepsilon_n^2 = V(\varepsilon_n)/n$ for all $n \in \mathbb{N}$;*
3. *$M(\varepsilon) \leq M_{\mathcal{P}, \rho_H}(\varepsilon)$ for all $\varepsilon > 0$;*
4. *$M(\varepsilon) > 2 \log 2$ for $\varepsilon > 0$ small enough;*
5. *$(\widetilde{\varepsilon}_n)_{n \in \mathbb{N}} \subset (0, \infty)$ is chosen such that $M(\widetilde{\varepsilon}_n) = 4n\varepsilon_n^2 + 2 \log 2$ for all $n \in N$.*

Then

$$\widetilde{\varepsilon}_n^2/8 \leq \inf_{E \in \mathcal{A}_n(\mathcal{D}_{\boldsymbol{\mu}})} \sup_{p' \in \mathcal{P}} \mathbb{E}_{\boldsymbol{S} \overset{iid}{\sim} p'} \rho_H^2(E(\boldsymbol{S}), p') \leq 2\varepsilon_n^2. \tag{45}$$

Remark 18. *Let $\emptyset \neq \mathcal{P} \subset D_{\boldsymbol{\mu}}$. In (Petersen & Voigtlaender, 2021, Page 15) it was shown that $V_{\mathcal{P}, \rho_{KL}}(\varepsilon) \leq V_{\mathcal{P}, \rho_H}(\varepsilon/(2c))$ for all $0 < \varepsilon < 1$ and a suitable absolute constant $c > 1$. Then, item 1 in Lemma 17 is fulfilled when $V_{\mathcal{P}, \rho_H}(\varepsilon/(2c)) \leq V(\varepsilon)$.*

Now, by hypothesis (10), Remark 18 and Lemma 15, it follows that

$$V_{\mathcal{P}_{\mathscr{C}^\xi},\rho_{KL}}(\varepsilon) \le V_{\mathcal{P}_{\mathscr{C}^\xi},\rho_H}(\varepsilon/(2c)) \le V_{\mathscr{C},L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}/\left(4\widetilde{\gamma}^2(2c)^{\frac{2}{\widetilde{\gamma}+1}}\right)\right) \le V(\varepsilon) \quad \text{and}$$
$$M_{\mathcal{P}_{\mathscr{C}^\xi},\rho_H}(\varepsilon) \ge M_{\mathscr{C},L^1}\left((4\widetilde{\gamma})\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right) \ge M(\varepsilon) \quad \text{for all} \quad \varepsilon > 0. \tag{46}$$

Therefore, Lemma 17 holds with $\mathcal{P} = \mathcal{P}_{\mathscr{C}^\xi}$, since items 1 and 3 are satisfied with (46), and items 2, 4, and 5 are satisfied with (11). Then, using (35) and (45), we conclude that

$$\mathcal{I}_n(\mathscr{C}) \ge \frac{1}{(4\widetilde{\gamma})^3}\left[\inf_{E\in\mathcal{A}_n(\mathcal{D}_{\boldsymbol{\mu}})}\sup_{p'\in\mathcal{P}_{\mathscr{C}^\xi}}\mathbb{E}_{\boldsymbol{S}\overset{iid}{\sim}p'}\,\rho_H^2(E(\boldsymbol{S}),p')\right] \ge \left[\frac{1}{8(4\widetilde{\gamma})^3}\right]\widetilde{\varepsilon}_n^2$$

and (12) is satisfied. □

B.2 Proof of Corollary 8

We consider each case as follows:

I) For all $\varepsilon \in (0,1)$, $\varepsilon^{-a} \lesssim M_{\mathscr{C},L^1}(\varepsilon) \lesssim \varepsilon^{-a}(1+\log(1/\varepsilon))$. So, (7) implies

$$V_{\mathscr{C},L^1}(\varepsilon) \lesssim \varepsilon^{-a}(1+\log(1/\varepsilon)) \quad \text{and} \quad \varepsilon^{-a} \lesssim M_{\mathscr{C},L^1}(\varepsilon).$$

To use Theorem 6, we set $\widetilde{\varepsilon}_n, \varepsilon_n \in (0,1)$,

$$V(\varepsilon) \approx \left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right)^{-a}\left(1+\log\left(\varepsilon^{-\frac{2}{\widetilde{\gamma}+1}}\right)\right) \gtrsim V_{\mathscr{C},L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right) \quad \text{and}$$
$$M(\varepsilon) \approx \left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right)^{-a} \lesssim M_{\mathscr{C},L^1}\left(\varepsilon^{\frac{2}{\widetilde{\gamma}+1}}\right),$$

such that (10) is satisfied. From (11),

$$n\varepsilon_n^2 \approx \left(\varepsilon_n^{\frac{2}{\widetilde{\gamma}+1}}\right)^{-a}\left(1+\log\left(\varepsilon_n^{-\frac{2}{\widetilde{\gamma}+1}}\right)\right) \approx \left(\left(\widetilde{\varepsilon}_n^{\frac{2}{\widetilde{\gamma}+1}}\right)^{-a} - 2\log 2\right)/4. \tag{47}$$

Also, $a \ge 1/2$ and $\widetilde{\varepsilon}_n \le 4\varepsilon_n$. Then,

$$n \gtrsim 4n\varepsilon_n^2 + 2\log 2 \approx \left(\widetilde{\varepsilon}_n^{\frac{2}{\widetilde{\gamma}+1}}\right)^{-a} \gtrsim \varepsilon_n^{-\frac{1}{\widetilde{\gamma}+1}}. \tag{48}$$

From (47) and (48),

$$n\varepsilon_n^2 \lesssim \left(\varepsilon_n^{\frac{2}{\widetilde{\gamma}+1}}\right)^{-a}(1+2\log n), \quad \text{i.e.} \quad (\varepsilon_n^2)^{1+\frac{a}{\widetilde{\gamma}+1}} \lesssim n^{-1}(1+2\log n).$$

Thus, $\varepsilon_n^2 \lesssim \left[n^{-1}(1+2\log n)\right]^{\frac{1}{1+\left(\frac{a}{\widetilde{\gamma}+1}\right)}}$. By (47),

$$\left(\widetilde{\varepsilon}_n^{\frac{2}{\widetilde{\gamma}+1}}\right)^{-a} \approx 4n\varepsilon_n^2 + 2\log 2 \lesssim n\left[n^{-1}(1+2\log n)\right]^{\frac{1}{1+\left(\frac{a}{\widetilde{\gamma}+1}\right)}} + 2\log 2$$
$$\lesssim n\left[n^{-1}(1+2\log n)\right]^{\frac{1}{1+\left(\frac{a}{\widetilde{\gamma}+1}\right)}}.$$

Therefore,

$$\widetilde{\varepsilon}_n^2 \gtrsim n^{-\frac{1}{\left(\frac{a}{\widetilde{\gamma}+1}\right)}}\left[n^{-1}(1+2\log n)\right]^{-\frac{1}{\left(1+\left(\frac{a}{\widetilde{\gamma}+1}\right)\right)\left(\frac{a}{\widetilde{\gamma}+1}\right)}}$$
$$= n^{-\frac{1}{1+\left(\frac{a}{\widetilde{\gamma}+1}\right)}}(1+2\log n)^{-\frac{1}{\left(1+\left(\frac{a}{\widetilde{\gamma}+1}\right)\right)\left(\frac{a}{\widetilde{\gamma}+1}\right)}},$$

and (12) implies

$$\mathcal{I}_n(\mathscr{C}) \gtrsim \left[n(1+2\log n)^{\frac{1}{\left(\frac{a}{\widetilde{\gamma}+1}\right)}}\right]^{-\frac{1}{1+\left(\frac{a}{\widetilde{\gamma}+1}\right)}}.$$

$II)$ We know that $V_{\mathscr{C},L^1}(\varepsilon) \approx \varepsilon^{-a}$, for all $\varepsilon \in (0,1)$. Then, to use Theorem 6, we choose

$$M(\varepsilon) \approx V(\varepsilon) \approx \left(\varepsilon^{\frac{2}{\gamma+1}}\right)^{-a},$$

such that (10) is satisfied. So, (11) implies

$$\left(\varepsilon_n^2\right)^{1+\left(\frac{a}{\gamma+1}\right)} \approx n^{-1}, \quad \text{i.e.} \quad \varepsilon_n^2 \approx n^{-\frac{1}{1+\left(\frac{a}{\gamma+1}\right)}}.$$

Thus, using the above inequality and (11),

$$\left(\widetilde{\varepsilon}_n^{\frac{2}{\gamma+1}}\right)^{-a} \approx 4n\varepsilon_n^2 + 2\log 2 \lesssim n^{1-\frac{1}{1+\left(\frac{a}{\gamma+1}\right)}} = n^{\frac{\left(\frac{a}{\gamma+1}\right)}{1+\left(\frac{a}{\gamma+1}\right)}}$$

and

$$\widetilde{\varepsilon}_n^2 \gtrsim n^{-\frac{1}{1+\left(\frac{a}{\gamma+1}\right)}}.$$

In conclusion, from (12), it follows

$$\mathcal{I}_n(\mathscr{C}) \gtrsim n^{-\frac{1}{1+\left(\frac{a}{\gamma+1}\right)}}.$$

□

B.3 Proof of Corollary 9

We split the proof into the following cases:

$I)$ It is well known that the space of Barron regular functions is Lipschitz, therefore $\mathcal{B}_C$ satisfies C2) with $\alpha = 1$. Furthermore, from Lemma 3, we have that item $I)$ of Corollary 8 holds with $a = 2(d-1)/(d+1)$. Then,

$$\mathcal{I}_n(\mathcal{B}_C) \gtrsim \left[n\left(1+2\log n\right)^{\frac{\max\{3,\gamma\}}{\left(\frac{2(d-1)}{d+1}\right)}}\right]^{-\frac{\max\{3,\gamma\}}{\max\{3,\gamma\}+\left(\frac{2(d-1)}{d+1}\right)}}.$$

$II)$ By definition, the α-Hölder continuous space satisfies C2). Then, with Lemma 4, we obtain that item $II)$ of Corollary 8 is satisfied with $a = (d-1)/\alpha$. Therefore,

$$\mathcal{I}_n(\mathcal{H}_\alpha) \gtrsim n^{-\frac{\max\{3,\gamma/\alpha\}}{\max\{3,\gamma/\alpha\}+(d-1)/\alpha}}.$$

$III)$ We know that $\mathcal{C}_B([0,1]^{d-1})$ is the set of all convex functions on $[0,1]^{d-1}$ that are uniformly bounded by B. Therefore, $\mathcal{C}_B$ is Lipschitz and condition C2) is fulfilled with $\alpha = 1$. Then, Lemma 5 implies that item $II)$ of Corollary 8 holds with $a = (d-1)/2$. Thus,

$$\mathcal{I}_n(\mathcal{C}_B) \gtrsim n^{-\frac{2\max\{3,\gamma\}}{2\max\{3,\gamma\}+(d-1)}}.$$

□