# A general approximation lower bound in $L^p$ norm, with applications to feed-forward neural networks

El Mehdi Achour, Armand Foucault, Sébastien Gerchinovitz, François Malgouyres

# A general approximation lower bound in $L^p$ norm, with applications to feed-forward neural networks

**El Mehdi Achour**[1], **Armand Foucault**[1], **Sébastien Gerchinovitz**[2,1], **and François Malgouyres**[1]

[1]Institut de Mathématiques de Toulouse ; UMR5219 , Université de Toulouse ; CNRS , UPS IMT
F-31062 Toulouse Cedex 9, France
[2]IRT Saint Exupéry, 3 rue Tarfaya, 31400 Toulouse, France

{El_mehdi.achour,armand.foucault,francois.malgouyres} AT
math.univ-toulouse.fr
sebastien.gerchinovitz AT irt-saintexupery.com

## Abstract

We study the fundamental limits to the expressive power of neural networks. Given two sets $F$, $G$ of real-valued functions, we first prove a general lower bound on how well functions in $F$ can be approximated in $L^p(\mu)$ norm by functions in $G$, for any $p \geq 1$ and any probability measure $\mu$. The lower bound depends on the packing number of $F$, the range of $F$, and the fat-shattering dimension of $G$. We then instantiate this bound to the case where $G$ corresponds to a piecewise-polynomial feed-forward neural network, and describe in details the application to two sets $F$: Hölder balls and multivariate monotonic functions. Beside matching (known or new) upper bounds up to log factors, our lower bounds shed some light on the similarities or differences between approximation in $L^p$ norm or in sup norm, solving an open question by DeVore et al. [DHP21]. Our proof strategy differs from the sup norm case and uses a key probability result of Mendelson [Men02].

## 1 Introduction

Neural networks are known for their great expressive power: in classification, they can interpolate arbitrary labels [ZBH$^+$21], while in regression they have universal approximation properties [Cyb89, Hor91, LLPS93, KL20], with approximation rates that can outperform those of linear approximation methods [Yar18, DHP21]. Though the approximation problem is often only one part of the underlying learning problem (where generalization and optimization properties are also at stake), understanding the fundamental limits to the approximation properties of neural networks is key, both conceptually and for practical issues such as designing the right network architecture for the right problem.

**Setting and related works.** One way to quantify the expressive power of neural networks is through the following problem (some informal statements will be made more precise in the next sections). Let $G$ be the set of all functions $g_{\mathbf{w}} : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ that can be represented by tuning the weights $\mathbf{w} \in \mathbb{R}^W$ of a feed-forward neural network with a fixed architecture, and let $F$ be any set of real-valued functions on $\mathcal{X}$. A natural question is: how well functions $f \in F$ can be approximated by functions $g_{\mathbf{w}} \in G$? More precisely, given a norm $\|\cdot\|$ on functions, what is the order of magnitude of the (worst-case) *approximation error of $F$ by $G$* defined by

$$\sup_{f \in F} \inf_{g_{\mathbf{w}} \in G} \|f - g_{\mathbf{w}}\| \,, \tag{1}$$

Preprint. Under review.

and how small can it be given the numbers $W$, $L$ of weights and layers, and some properties of $F$?

The case when $\|\cdot\|$ is the sup norm (defined by $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$) is rather well understood at least in some special cases. For example, when $F$ is a Hölder ball of smoothness $s > 0$ and the network uses the ReLU activation function, Yarotsky [Yar17] derived a lower bound on (1) of the order of $W^{-2s/d}$, later refined to $(LW)^{-s/d}$ (up to log factors) by [Yar18, YZ20] when the depth of the network varies from $L = 1$ to $L \approx W$. Using the bit extraction technique, these authors showed that these lower bounds are achievable (up to log factors) with a carefully designed ReLU network architecture. Refined results in terms of width and depth were obtained by [SYZ22] when $s \leq 1$, while some other activation functions were also studied in [YZ20].

In this paper, we study (1) with the $L^p(\mu)$ norm, defined by $\|f\|_{L^p(\mu)} = \left(\int_X |f(x)|^p d\mu(x)\right)^{1/p}$, for $p \geq 1$ and some probability measure $\mu$ on $\mathcal{X}$. Since this corresponds to approximating functions in $F$ in a more "average" sense than in sup norm, a natural question is whether the same accuracy can be achieved with a smaller network or not. Unfortunately, however, the proof strategies behind the lower bounds of [Yar17, Yar18, YZ20, SYZ22] are specific to the sup norm (see Remark 1 in Section 3 for details). DeVore et al. [DHP21] indeed commented: "When we move to the case $p < \infty$, the situation is even less clear [...] we cannot use the VC dimension theory for $L^p(\Omega)$ approximation. [...] What is missing vis-à-vis Problem 8.13 is what the best bounds are and how we prove lower bounds for approximation rates in $L^p(\Omega)$, $p \neq \infty$."

**Existing lower bounds in $L^p(\mu)$ norm.** Several papers provided lower bounds in some special cases, under some restrictions on the set to approximate $F$, the neural network, the approximation metric, or the encoding map $f \in F \mapsto \mathbf{w}(f) \in \mathbb{R}^W$.

When $F$ is a space of smoothness $s$, a first result which is based on [DHM89] states that when imposing the weights to depend continuously on the function to be approximated, one can not achieve a better approximation rate than $W^{-\frac{s}{d}}$.

For the same $F$, another result for $p = 2$ and for activation functions which are continuous ([Mai99, MMR99]) prove a lower bound on the approximation of functions of smoothness $s$ on a compact of $\mathbb{R}^d$, by one hidden-layer neural networks, of order $W^{-\frac{s}{d-1}}$. A matching upper bound is proven for a particular activation function, which is sigmoidal but pathological ([MP99]). For this same activation function, they prove that contrary to the one-hidden-layer case, there is no lower bound in the case of two-hidden-layer networks. The result is based on the Kolmogorov-Arnold superposition theorem.

In [SX21], the authors study approximation by shallow neural networks with bounded weights and activations of the form $\mathrm{ReLU}^k$ for an integer $k$. They approximate the closure of the convex hull of shallow $\mathrm{ReLU}^k$-neural networks with constrained weights. They obtain optimal lower bounds of order $W^{-\frac{1}{2}-\frac{2k+1}{2d}}$ in any norm $\|\cdot\|_X$ where $X$ is a Banach space to which the approximation functions belong and such that these functions are uniformly bounded w.r.t. $\|\cdot\|_X$. Although we only consider approximation in $L^p(\mu)$ norm, our results complement the latter by addressing neural networks with unbounded weights and arbitrary depth, and general sets $F$.

Approximation lower bounds in $L^p(\mu)$ norm, $p \geq 1$, have also been studied in the quantized neural networks setting (networks with weights encoded with a fixed number of bits). In [PV18], under weak assumptions on the activation function, the authors prove a lower bound on the minimal number of nonzero weights $W$ that are required for a network to approximate a class of binary classifiers with $L^p$ error at most $\varepsilon$. They show that $W$ is at least of the order $\varepsilon^{-\frac{p(d-1)}{\beta}} \log_2^{-1}(1/\varepsilon)$, where $\beta$ is a smoothness parameter. Later works including [VP19, GR21] derive lower bounds for approximation by quantized networks in various norms.

**Main contributions and outline of the paper.** We prove lower bounds on the approximation error (1) in any $L^p(\mu)$ norm, for non-quantized networks of arbitrary depth, and general sets $F$. Our main contributions are the following.

In Section 2, we first prove a general lower bound for any two sets $F$, $G$ of real-valued functions on a set $\mathcal{X}$ (Theorem 1). The lower bound depends on the packing number of $F$, the range of $F$, and the fat-shattering dimension of $G$. We then derive a versatile corollary when $G$ corresponds to a piecewise-polynomial feed-forward neural network (Corollary 1), solving the question by DeVore et al. [DHP21]. Importantly, our proof strategy still relies on VC dimension theory, but differs from

the sup norm case in using a key probability result of Mendelson [Men02], to relate approximation in $L^p(\mu)$ norm with the fat-shattering dimension of $G$.

In Sections 3–4 we apply this corollary to the approximation of two sets: Hölder balls and multivariate monotonic functions. Beside matching (known or new) upper bounds up to log factors, our lower bounds shed some light on the similarities or differences between approximation in $L^p$ norm or in sup norm. In particular, with ReLU networks, Hölder balls are not easier to approximate in $L^p$ norm than in sup norm. On the contrary, the approximation rate for multivariate monotonic functions depends on $p$. In Section 5, we outline several other examples of function sets $F$ and $G$ for which the general lower bound (Theorem 1) can also be easily applied. Finally, some proofs are postponed to the supplement, while some details on other existing lower bound proof strategies are provided in the supplement, in Appendix C.

**Main definitions and notation.** We provide below some definitions and notation that will be used throughout the paper. We denote the set of positive integers $\{1, 2, \ldots\}$ by $\mathbb{N}^*$ and let $\mathbb{N} := \mathbb{N}^* \cup \{0\}$. All sets considered in this paper will be assumed to be nonempty, and *measurable set* will be used to denote a set $\mathcal{X}$ (implicitly) endowed with a $\sigma$-algebra.

Let $p \in [1, +\infty]$ and $\mathcal{X}$ be any measurable set endowed with a probability measure $\mu$. For any measurable function $f : \mathcal{X} \to \mathbb{R}$, the $L^p(\mu)$ norm of $f$ is defined by $\|f\|_{L^p(\mu)} = \left(\int_{\mathcal{X}} |f(x)|^p d\mu(x)\right)^{1/p}$ (possibly infinite) if $p < +\infty$, and $\|f\|_{L^\infty(\mu)} = \operatorname{ess\,sup}_{x\in\mathcal{X}} |f(x)|$. We will write $\lambda$ for the Lebesgue measure on $[0, 1]^d$.

For any $\varepsilon > 0$, two functions $f_1, f_2$ are said to be $\varepsilon$-*distant* in $\|\cdot\|$ if $\|f_1 - f_2\| > \varepsilon$. Let $F$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}$. A set $\{f_1, \ldots, f_N\} \subset F$ is said to be an $\varepsilon$-packing of $F$ in $\|\cdot\|$ (or just an $\varepsilon$-packing for short) if for any $i \neq j \in \{1, \ldots, N\}$, $f_i$ and $f_j$ are $\varepsilon$-distant in $\|\cdot\|$. The $\varepsilon$-packing number $M(\varepsilon, F, \|\cdot\|)$ is the largest cardinality of $\varepsilon$-packings (possibly infinite).

For $\gamma \geq 0$, we say that a set $S = \{x_1 \ldots, x_N\} \subset \mathcal{X}$ is $\gamma$-*shattered* by $F$ if there exists $r : S \to \mathbb{R}$ such that for any $E \subset S$, there exists $f \in F$ satisfying for all $i = 1, \ldots, N$, $f(x_i) > r(x_i) + \gamma$ if $x_i \in E$, and $f(x_i) < r(x_i) - \gamma$ if $x_i \notin E$. If this statement is true when we replace $\gamma$ by 0, we say that $F$ *pseudo shatters* $S$. The $\gamma$-*fat-shattering dimension*, denoted by $\mathrm{fat}_\gamma(F)$, is defined as the largest number $N \leq +\infty$ such that there exists $S \subset \mathcal{X}$ of cardinality $N$ which is $\gamma$-shattered by $F$. The *pseudo dimension* $\mathrm{Pdim}(G)$ is defined similarly but for sets that are pseudo shattered.

A formal definition of feed-forward neural networks is recalled in Appendix A. In short, in this paper, a *feed-forward neural network architecture* $\mathcal{A}$ of depth $L \geq 1$ is a directed acyclic graph with $d \geq 1$ input neurons, $L - 1$ hidden layers (if $L \geq 2$), and an output layer with only one neuron. Skip connections are allowed, i.e., there can be connections between non-consecutive layers. Given an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, a feed-forward neural network architecture $\mathcal{A}$, and a vector $\mathbf{w} \in \mathbb{R}^W$ of weights assigned to all edges and non-input neurons (linear coefficients and biases), the network computes a function $g_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ defined by recursively computing affine transformations for each hidden or output neuron, and then applying the activation function $\sigma$ for hidden neurons only (see Appendix A for more details). Finally, we define $H_{\mathcal{A}} := \{g_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^W\}$ to be the set of all functions that can be represented by tuning all the weights assigned to the network.

A function $\sigma : \mathbb{R} \to \mathbb{R}$ is *piecewise-polynomial* on $K \geq 2$ pieces, with maximal degree $\nu \in \mathbb{N}$, if there exists a partition $I_1, \ldots, I_K$ of $\mathbb{R}$ into $K$ nonempty intervals, such that $\sigma$ restricted on each $I_j$ is polynomial with degree at most $\nu$.

## 2 A general approximation lower bound in $L^p(\mu)$ norm

In this section, we provide our two main results: a general lower bound on the $L^p(\mu)$ approximation error of $F$ by $G$, i.e., $\sup_{f\in F} \inf_{g\in G} \|f - g\|_{L^p(\mu)}$, and a corollary when $G$ corresponds to a feed-forward neural network with a piecewise-polynomial activation function. The weak assumptions on $F$ make the last result applicable to a wide range of cases of interest, as shown in Sections 3–5.

### 2.1 Main results

Our generic lower bound reads as follows, and is proved in Section 2.2. We follow the conventions $0 \times \log^2(0) = 0$ and $P^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P) = +\infty$ when $P = 1$.

**Theorem 1.** *Let $1 \leq p < +\infty$ and $\mathcal{X}$ be a measurable set endowed with a probability measure $\mu$. Let $F$, $G$ be two sets of real-valued functions defined on $\mathcal{X}$, such that all functions in $F$ have the same finite range $[a,b]$, and $\mathrm{fat}_\gamma(G) < +\infty$ for all $\gamma > 0$. Then, there exists a constant $c > 0$ depending only on $p$ such that*

$$\sup_{f\in F}\inf_{g\in G}\|f-g\|_{L^p(\mu)} \geq \inf\left\{\varepsilon>0 : \log M\big(3\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) \leq c\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G)\log^2\left(\frac{2\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G)}{\varepsilon/(b-a)}\right)\right\}. \tag{2}$$

*In particular, if $\log M\big(\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) \geq c_0\varepsilon^{-\alpha}$ for some $c_0, \varepsilon_0, \alpha > 0$ and all $\varepsilon \leq \varepsilon_0$, and if $\mathrm{Pdim}(G) < +\infty$, then there exist constants $c_1, \varepsilon_1 > 0$ depending only on $b-a$, $p$, $c_0$, $\varepsilon_0$ and $\alpha$ such that*

$$\sup_{f\in F}\inf_{g\in G}\|f-g\|_{L^p(\mu)} \geq \min\left\{\varepsilon_1,\ c_1\,\mathrm{Pdim}(G)^{-\frac{1}{\alpha}}\log^{-\frac{2}{\alpha}}\big(\mathrm{Pdim}(G)\big)\right\}. \tag{3}$$

The first lower bound (2) is generic but requires solving an inequation.[1] In (3) we solve this inequation when $\log M\big(\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big)$ grows at least polynomially in $1/\varepsilon$ (which is typical of nonparametric sets) and when $G$ has finite pseudodimension $\mathrm{Pdim}(G)$. Though we will restrict our attention to such cases in all subsequent sections, we stress that the first bound should have broader applications. A first example is when $\mathrm{Pdim}(G) = +\infty$ but $\mathrm{fat}_\gamma(G) < +\infty$ is finite for all $\gamma > 0$ (e.g., for RKHS [Bel18]). The first bound should also be useful to prove (slightly) tighter lower bounds when $\log M\big(\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big)$ has a (slightly) different dependency on $1/\varepsilon$ (e.g., of the order of $\varepsilon^{-\alpha}\log^\beta(1/\varepsilon)$ as when $F$ is the set of all multivariate cumulative distribution functions [BGL07]).

In the rest of the paper, we focus on the important special case when the approximation set $G$ is the set of all real-valued functions that can be represented by tuning the weights of a feed-forward neural network with fixed architecture $\mathcal{A}$ and a piecewise-polynomial activation function. By combining Theorem 1 with known bounds on the pseudo-dimension [BHLM19], we obtain the following corollary, which bounds the approximation error in terms of the number $W$ of weights and the depth $L$ (i.e., the number of hidden and output layers). The proof is postponed to Appendix B.4.

**Corollary 1.** *Let $1 \leq p < +\infty$, $d \geq 1$ and $\mathcal{X}$ be a measurable subset of $\mathbb{R}^d$ endowed with a probability measure $\mu$. Let $F$ be a set of functions from $\mathcal{X}$ to $[a,b]$, $a,b \in \mathbb{R}$, such that $\log M\big(\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) \geq c_0\varepsilon^{-\alpha}$ for some $c_0, \varepsilon_0, \alpha > 0$ and all $\varepsilon \leq \varepsilon_0$.*

*Let $H_{\mathcal{A}}$ be the set of all real-valued functions on $\mathcal{X}$ that can be represented by a feed-forward neural network with a fixed architecture $\mathcal{A}$ of depth $L \geq 1$, $W \geq 1$ variable weights and a piecewise-polynomial activation function of maximal degree $\nu \in \mathbb{N}$ on $K \geq 2$ pieces. Then, for $W \geq W_{\min}$,*

$$\sup_{f\in F}\inf_{g\in H_{\mathcal{A}}}\|f-g\|_{L^p(\mu)} \geq \begin{cases} c_1 W^{-\frac{2}{\alpha}}\log^{-\frac{2}{\alpha}}(W) & \text{if } \nu \geq 2\,, \\ c_2(LW)^{-\frac{1}{\alpha}}\log^{-\frac{3}{\alpha}}(W) & \text{if } \nu = 1\,, \\ c_3 W^{-\frac{1}{\alpha}}\log^{-\frac{3}{\alpha}}(W) & \text{if } \nu = 0\,, \end{cases} \tag{4}$$

*where the constants $W_{\min}, c_1, c_2, c_3 > 0$ are independent from $W$.*

There are equivalent ways to write the above corollary. For example, given a target accuracy $\varepsilon > 0$ and a depth $L \geq 1$, (4) yields a lower bound on the minimum number $W$ of weights that are needed to get $\sup_{f\in F}\inf_{g\in H_{\mathcal{A}}}\|f-g\|_{L^p(\mu)} \leq \varepsilon$. Some earlier approximation results were written this way (e.g., [Yar17, PV18]).

### 2.2 Proof of Theorem 1

In order to prove Theorem 1, we need two inequalities. The first one is straightforward (and appeared within proofs, e.g., in [YZ20]), but formalizes the key idea that if $G$ approximates $F$ with error $\varepsilon$, then $G$ has to be at least as large as $F$. We use the conventions $\log(+\infty) = +\infty$ and $+\infty \leq +\infty$.

**Lemma 1.** *Let $p \geq 1$ and $\mathcal{X}$ be a measurable space endowed with a probability measure $\mu$. Let $F$, $G$ be two sets of real-valued functions defined on $\mathcal{X}$. If $\sup_{f\in F}\inf_{g\in G}\|f-g\|_{L^p(\mu)} < \varepsilon$, then*

$$\log M\big(3\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) \leq \log M\big(\varepsilon, G, \|\cdot\|_{L^p(\mu)}\big).$$

[1] Note that any $\varepsilon \geq (b-a)/3$ is a solution to this inequation, since $\log M\big(3\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) = \log(1) = 0$ (because all functions in $F$ are $[a,b]$-valued). Therefore, the right-hand side of (2) is at most $(b-a)/3$.

*Proof.* Let $P_F = \{f_1, \dots, f_N\}$ be a $3\varepsilon$-packing of $F$, with $N \geq 1$. Let $P_G = \{g_1, \dots, g_N\}$ be a subset of $G$ such that $\|f_i - g_i\|_{L^p(\mu)} \leq \varepsilon$ for all $i$. Note that the existence of such a $P_G$ is guaranteed by the assumption $\sup_{f\in F} \inf_{g\in G} \|f-g\|_{L^p(\mu)} < \varepsilon$. Since the $f_i$'s are pairwise $3\varepsilon$-distant in $L^p(\mu)$, the triangle inequality entails that the $g_i$'s are also at least pairwise $\varepsilon$-distant in $L^p(\mu)$. Therefore, $P_G$ is an $\varepsilon$-packing of $G$, and the result follows. □

The second inequality is a fundamental probability result due to Mendelson [Men02]. It bounds from above the $\varepsilon$-packing number in $L^p(\mu)$ norm of any uniformly bounded function set in terms of its fat-shattering dimension. Crucially, the inequality holds for finite $p \geq 1$, as opposed to the lower bound strategy of Yarotsky [Yar17, Yar18] (see also [DHP21]), that relates the VC-dimension with the approximation error in sup norm. The next statement is a straightforward generalization of a result of [Men02] initially stated for $[a, b] = [0, 1]$ and for Glivenko-Cantelli classes $G$ (see Appendix B.1 for details).

**Proposition 1** ([Men02], Corollary 3.12)**.** *Let $G$ be a set of functions from a measurable set $\mathcal{X}$ to $[a, b]$, $a, b \in \mathbb{R}$, and such that $\mathrm{fat}_\gamma(G) < +\infty$ for all $\gamma > 0$. Then for any $1 \leq p < +\infty$, there exists $c > 0$ depending only on $p$ such that for every probability measure $\mu$ on $\mathcal{X}$ and every $\varepsilon > 0$,*

$$\log M\big(\varepsilon, G, \|\cdot\|_{L^p(\mu)}\big) \leq c\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G) \log^2\left(\frac{2(b-a)\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G)}{\varepsilon}\right) . \tag{5}$$

*Proof (of Theorem 1).* **Part 1.** We start by proving (2), using Proposition 1 as a key argument. Since functions in $G$ are not necessarily uniformly bounded, we will apply Proposition 1 to the "clipped version of $G$". More precisely, for any function $g \in G$, we define its clipping (truncature) to $[a, b]$ as the function $\tilde{g} : \mathcal{X} \to \mathbb{R}$ given by $\tilde{g}(x) = \min(\max(a, g(x)), b)$ for all $x \in \mathcal{X}$. We then set $G_{[a,b]} = \{\tilde{g} : g \in G\}$, which by construction consists of functions that are all $[a, b]$-valued.

Noting that clipping can only help since elements of $F$ are $[a, b]$-valued (see Lemma 4 in the supplement, Appendix B.2), we have

$$\sup_{f\in F} \inf_{g\in G} \|f - g\|_{L^p(\mu)} \geq \sup_{f\in F} \inf_{\tilde{g}\in G_{[a,b]}} \|f - \tilde{g}\|_{L^p(\mu)} . \tag{6}$$

Setting $\Delta := \sup_{f\in F} \inf_{\tilde{g}\in G_{[a,b]}} \|f - \tilde{g}\|_{L^p(\mu)}$, we now show that $\Delta$ is bounded from below by the right-hand side of (2). To that end, it suffices to show that every $\varepsilon > \Delta$ is a solution to the inequation

$$\log M\big(3\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) \leq c\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G) \log^2\left(\frac{2(b-a)\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G)}{\varepsilon}\right) . \tag{7}$$

The last inequality is true whenever $\varepsilon \geq (b-a)/3$ (see Footnote 1 and note $c\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G) \geq 0$). We only need to prove (7) when $\Delta < \varepsilon < (b-a)/3$. In this case, by definition of $\Delta$ and by Lemma 1 applied to $G_{[a,b]}$, we have

$$\begin{aligned}
\log M\big(3\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) &\leq \log M\big(\varepsilon, G_{[a,b]}, \|\cdot\|_{L^p(\mu)}\big) \\
&\leq c\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G_{[a,b]}) \log^2\left(\frac{2(b-a)\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G_{[a,b]})}{\varepsilon}\right) \\
&\leq c\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G) \log^2\left(\frac{2(b-a)\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G)}{\varepsilon}\right) ,
\end{aligned} \tag{8}$$

where the second inequality follows from Proposition 1 (note from Lemma 3 in the supplement, Appendix B.2 that $\mathrm{fat}_\gamma(G_{[a,b]}) \leq \mathrm{fat}_{\frac{\varepsilon}{32}}(G)$ for all $\gamma > 0$, which is finite by assumption), and where (8) follows from the next remark. Either $\mathrm{fat}_{\frac{\varepsilon}{32}}(G_{[a,b]}) = 0$, and (8) is true by the convention $0 \times \log^2(0) = 0$ and $c\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G) \geq 0$. Either $\mathrm{fat}_{\frac{\varepsilon}{32}}(G_{[a,b]}) \geq 1$, and (8) follows from $t \mapsto ct \log^2\big(\frac{2(b-a)t}{\varepsilon}\big)$ being non-decreasing on $[\varepsilon/(2(b-a)), +\infty)$ and $\varepsilon/(2(b-a)) \leq 1/6 \leq 1 \leq \mathrm{fat}_{\frac{\varepsilon}{32}}(G_{[a,b]}) \leq \mathrm{fat}_{\frac{\varepsilon}{32}}(G)$. To conclude, every $\varepsilon > \Delta$ satisfies (7), which implies that $\Delta$ is bounded from below by the right-hand side of (2). Combining with (6) concludes the proof of (2).

**Part 2.** Set $\varepsilon_1' = \min\big\{\frac{\varepsilon_0}{3}, 2(b-a)\big\}$. We now derive (3) from (2). To that end, setting $P = \mathrm{Pdim}(G)$, we show that every $\varepsilon > 0$ satisfying (7) is such that $\varepsilon \geq \min\big\{\varepsilon_1,\ c_1 P^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P)\big\}$, where

$\varepsilon_1 \in (0, \varepsilon_1']$ and $c_1 > 0$ will be defined later. Since the claimed lower bound on $\varepsilon$ is true when $\varepsilon \geq \varepsilon_1'$, in the sequel we consider any solution $\varepsilon$ to (7) such that $0 < \varepsilon < \varepsilon_1'$ (if such a solution exists).

By the assumption on $\log M\big(u, F, \|\cdot\|_{L^p(\mu)}\big)$ for $u = 3\varepsilon \leq \varepsilon_0$, and then using (7), we have, setting $r = 2(b-a)$,

$$c_0(3\varepsilon)^{-\alpha} \leq \log M\big(3\varepsilon, F, \|\cdot\|_{L^p(\mu)}\big) \leq c \operatorname{fat}_{\frac{\varepsilon}{32}}(G) \log^2\left(\frac{r \operatorname{fat}_{\frac{\varepsilon}{32}}(G)}{\varepsilon}\right) \leq cP \log^2\left(\frac{rP}{\varepsilon}\right) ,$$

where the last inequality is because $t \mapsto ct\log^2\big(\frac{rt}{\varepsilon}\big)$ is non-decreasing on $[\varepsilon/r, +\infty)$, with $\varepsilon/r \leq 1$, and $1 \leq \operatorname{fat}_{\frac{\varepsilon}{32}}(G) \leq \operatorname{Pdim}(G) = P$ (the lower bound of 1 follows from $c_0(3\varepsilon)^{-\alpha} > 0$).

Solving the inequation $c_0(3\varepsilon)^{-\alpha} \leq cP\log^2(rP/\varepsilon)$ for $\varepsilon$ (see Appendix B.3 for details), we get

$$\varepsilon \geq \min\big\{\varepsilon_1'',\ c_1 P^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}} P\big\} , \tag{9}$$

for some constants $\varepsilon_1'', c_1 > 0$ depending only on $p$, $c_0$, $b-a$ and $\alpha$. Setting $\varepsilon_1 = \min\{\varepsilon_1'', \varepsilon_1'\}$ and noting that $\varepsilon_1'$ only depends on $\varepsilon_0$ and $b-a$, we conclude the proof. □

## 3 Approximation of Hölder balls by feed-forward neural networks

In this section, we apply Corollary 1 to establish nearly-tight lower bounds for the approximation of unit Hölder balls by feed-forward neural networks. Our main result is Proposition 3, which solves an open question by [DHP21].

Throughout the section, for any $s > 0$, we denote by $n$ and $\alpha$ the unique members of the decomposition $s = n + \alpha$ such that $n \in \mathbb{N}$ and $0 < \alpha \leq 1$.

For a set $\mathcal{X} \subset \mathbb{R}^d$, we follow [YZ20] and define the Hölder space $\mathcal{C}^{n,\alpha}(\mathcal{X})$ as the space of $n$ times continuously differentiable functions with finite norm

$$\|f\|_{\mathcal{C}^{n,\alpha}} = \max\left\{\max_{\mathbf{n}:|\mathbf{n}|\leq n} \|D^{\mathbf{n}} f\|_\infty,\ \max_{\mathbf{n}:|\mathbf{n}|=n} \sup_{x\neq y} \frac{|D^{\mathbf{n}} f(x) - D^{\mathbf{n}} f(y)|}{\|x-y\|_2^\alpha}\right\},$$

where, for $\mathbf{n} = (n_1, \cdots, n_d) \in \mathbb{N}^d$, $D^{\mathbf{n}} f = \left(\frac{\partial}{\partial x_1}\right)^{n_1} \cdots \left(\frac{\partial}{\partial x_d}\right)^{n_d} f$ denotes the $|\mathbf{n}|$-order partial derivative of $f$. We denote

$$F_{s,d} = \{f \in \mathcal{C}^{n,\alpha}([0,1]^d) : \|f\|_{\mathcal{C}^{n,\alpha}} \leq 1\}.$$

Let $\lambda$ denote the Lebesgue measure over $[0,1]^d$. In this section, we provide nearly matching upper and lower bounds for the $L^p(\lambda)$ approximation error of elements of $F_{s,d}$ by feed-forward ReLU neural networks. The bounds are expressed in terms of the number of weights of the network.

### 3.1 Known bounds on the sup norm approximation error

[YZ20] gives matching (up to a certain constant) lower and upper bounds of the sup norm approximation error of the elements of $F_{s,d}$ by feed-forward ReLU neural networks.

**Proposition 2** ([YZ20])**.** *Let $d \in \mathbb{N}^*$, $s > 0$, $\gamma \in \big(\frac{s}{d}, \frac{2s}{d}\big]$. Consider $n \in \mathbb{N}$ and $\alpha \in (0,1]$ such that $s = n + \alpha$.*

*There exist positive constants $W_{\min}$ and $c_1$, depending only on $d$ and $n$, such that for any integer $W \geq W_{\min}$, there exists a feed-forward ReLU neural network architecture $\mathcal{A}$ with $L = O(W^{\gamma\frac{d}{s}-1})$ layers and $W$ weights such that*

$$\sup_{f\in F_{s,d}} \inf_{g\in H_{\mathcal{A}}} \|f-g\|_\infty \leq c_1 W^{-\gamma}. \tag{10}$$

*In the meantime, there exists a constant $c_2$ depending only on $d$ and $n$ such that for any feed-forward ReLU neural network of architecture $\mathcal{A}$ with $W$ weights and $L = o(W^{\gamma\frac{d}{s}-1}/\log W)$ layers*

$$\sup_{f\in F_{s,d}} \inf_{g\in H_{\mathcal{A}}} \|f-g\|_\infty \geq c_2 W^{-\gamma}. \tag{11}$$

It is worth stressing that, for any probability measure $\mu$ on $[0,1]^d$, the upper bound (10) is automatically generalized to any smaller $L^p(\mu)$ norm, when $1 \leq p < +\infty$. However, the lower bound (11) does not necessarily apply when $\|\cdot\|_\infty$ is replaced with $\|\cdot\|_{L^p(\mu)}$, $1 \leq p < +\infty$. The lower bound of the next subsection shows that, in this setting, approximation in $L^p(\lambda)$ norm is not easier than in sup norm, solving an open question of DeVore et al. [DHP21].

### 3.2 Nearly-matching lower bounds of the $L^p(\lambda)$ approximation error

We first state a lower bound on the packing number of $F_{s,d}$, which is rather classical though hard to find in this specific form (see [BS67] for the $L^\infty$ norm, or [ET96] for other Sobolev-type norms). For the sake of completeness, we give a proof of Lemma 2 in the supplement, Appendix D.1.

**Lemma 2.** *Let $s > 0$, $d \in \mathbb{N}^*$ and $1 \leq p < \infty$. There exist constants $\varepsilon_0, c_0 > 0$ such that for any $0 < \varepsilon \leq \varepsilon_0$,*

$$\log M\left(\varepsilon, F_{s,d}, \|\cdot\|_{L^p(\lambda)}\right) \geq c_0 \varepsilon^{-\frac{d}{s}}. \tag{12}$$

Given Lemma 2, we can use Corollary 1 to establish the next proposition and obtain the lower bound of the $L^p(\lambda)$ approximation error.

**Proposition 3.** *Let $d \in \mathbb{N}^*$, $s > 0$, $\gamma \in \left(\frac{s}{d}, \frac{2s}{d}\right]$ and $1 \leq p < +\infty$. Let $\varepsilon_0, c_0$ be defined as in Lemma 2. Consider $n \in \mathbb{N}$ and $\alpha \in (0,1]$ such that $s = n + \alpha$.*

*Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a piecewise-affine function. There exist positive constants $c_1, W_{\min}$ depending only on $d$, $p$, $c_0$, $\varepsilon_0$ and $\sigma$ such that for any architecture $\mathcal{A}$ of depth $1 \leq L \leq cW^{\gamma\frac{d}{s}-1}$ (where $c$ is a constant), $W \geq W_{\min}$ variable weights and activation $\sigma$, the set $H_{\mathcal{A}}$ satisfies*

$$\sup_{f \in F_{s,d}} \inf_{g \in H_{\mathcal{A}}} \|f - g\|_{L^p(\lambda)} \geq c_1 W^{-\gamma} \log^{-\frac{3s}{d}}(W). \tag{13}$$

Note that, since ReLU is a piecewise-affine function, we obtain a lower bound which matches, up to logarithmic factors, the upper bound presented in the previous subsection.

*Proof.* From Lemma 2, there exists a constant $c_0 > 0$ such that $\log M\left(\varepsilon, \mathcal{M}_d, \|\cdot\|_{L^p(\lambda)}\right) \geq c_0 \varepsilon^{-\frac{d}{s}}$. Therefore, using Corollary 1 and $L \leq c(W^{\gamma\frac{d}{s}-1})$, we obtain the result. □

**Remark 1** (Comparison with existing proof strategies in sup norm.)**.** *We would like to highlight a key difference between the proof of Proposition 3 and the lower bound proof strategies of [Yar17, Yar18, YZ20, SYZ22] that are specific to the sup norm. Their overall argument is roughly the following: if $G$ can approximate any $f \in F$ in sup norm at accuracy $\varepsilon > 0$, since $F$ contains many "oscillating" functions with oscillation amplitude roughly $\varepsilon$, then so must be the case for $G$ (the sup norm is key here: **all** oscillations of any $f \in F$ are well approximated). Therefore, a small $\varepsilon$ implies a large $\mathrm{VCdim}(G)$, which by contrapositive enables to lower bound the approximation error (1) with a decreasing function of $\mathrm{VCdim}(G)$, and therefore as a function of $L$ and $W$. In contrast, in the proof of Theorem 1, the key probability result of Mendelson (Proposition 1) enables us to show that, even if the oscillations of any $f \in F$ are only well approximated **on average** (in $L^p(\mu)$ norm) by $G$, then $\mathrm{Pdim}(G)$ must be large when $\varepsilon$ is small. The conclusion is then the same: the approximation error in $L^p(\mu)$ norm can be lower bounded as a function of $\mathrm{Pdim}(G)$, and therefore in terms of $L$, $W$. This solves the question of DeVore et al. [DHP21] mentioned in the introduction, showing in particular that VC dimension theory can (surprisingly) be useful to prove $L^p$ approximation lower bounds.*

## 4 Approximation of monotonic functions by feed-forward neural networks

In this section, we consider the problem of approximating the set $\mathcal{M}^d$ of all non-decreasing functions from $[0,1]^d$ to $[0,1]$. These are functions $f : [0,1]^d \to [0,1]$ that are non-decreasing along any line parallel to an axis, i.e., such that, for all $x, y \in [0,1]^d$,

$$x_i \leq y_i, \;\; i = 1, \ldots, d \implies f(x) \leq f(y)\,.$$

Next we focus on the approximation of $\mathcal{M}^d$ with Heaviside feed-forward neural networks. After proving an impossibility result for the sup norm, we show that the weaker goal of approximating $\mathcal{M}^d$ in $L^p$ norm is feasible, and derive nearly matching lower and upper bounds. Interestingly, the approximation rates depend on $p \geq 1$, which is in sharp contrast with the case of Hölder balls, that are not easier to approximate in $L^p$ norm than in sup norm (see Section 3).

### 4.1 Warmup: an impossibility result in sup norm

We start this section by showing that approximating monotonic functions of $d \geq 2$ variables in sup norm is impossible with Heaviside neural networks.

**Proposition 4.** *For any neural network architecture $\mathcal{A}$ with the Heaviside activation, the set $H_{\mathcal{A}}$ satisfies*

$$\sup_{f \in \mathcal{M}^d} \inf_{g \in H_{\mathcal{A}}} \|f - g\|_\infty \geq \frac{1}{2}.$$

The proof of Proposition 4 is postponed to the supplement, Appendix E.2. We show a slightly stronger result, by exhibiting a single function $f \in \mathcal{M}^d$ such that the lower bound of $\frac{1}{2}$ holds simultaneously for all network architectures.

Next we study the approximation of $\mathcal{M}^d$ in $L^p(\lambda)$ norm.

### 4.2 Lower bound in $L^p(\lambda)$ norm

We start by proving a lower bound, as a direct consequence of Corollary 1 and a lower bound on the packing number due to [GW07].

**Proposition 5.** *Let $1 \leq p < +\infty$, $d \geq 1$, and let $\alpha = \max\{d, (d-1)p\}$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a piecewise-polynomial function having maximal degree $\nu \in \mathbb{N}$. There exist positive constants $c_1, c_2, c_3, W_{\min}$ depending only on $d$, $p$, and $\sigma$ such that for any architecture $\mathcal{A}$ of depth $L \geq 1$, $W \geq W_{\min}$ variable weights and activation $\sigma$, the set $H_{\mathcal{A}}$ satisfies*

$$\sup_{f \in \mathcal{M}^d} \inf_{g \in H_{\mathcal{A}}} \|f - g\|_{L^p(\lambda)} \geq \begin{cases} c_1 W^{-\frac{2}{\alpha}} \log^{-\frac{2}{\alpha}}(W) & \text{if } \nu \geq 2\,, \\ c_2 (LW)^{-\frac{1}{\alpha}} \log^{-\frac{3}{\alpha}}(W) & \text{if } \nu = 1\,, \\ c_3 W^{-\frac{1}{\alpha}} \log^{-\frac{3}{\alpha}}(W) & \text{if } \nu = 0\,. \end{cases} \tag{14}$$

*Proof.* From [GW07], there exist constants $\varepsilon_0, c_0 > 0$ such that for $\varepsilon \leq \varepsilon_0$, $\log M(\varepsilon, \mathcal{M}_d, \|\cdot\|_{L^p}) \geq c_0 \varepsilon^{-\alpha}$. Using Corollary 1, we obtain the result. □

### 4.3 Nearly-matching upper bound in $L^p(\lambda)$ norm

To the best of our knowledge, there does not exist any upper-bound of the $L^p(\lambda)$ approximation error of $\mathcal{M}^d$ with feed-forward neural networks. Checking that all the lower-bounds of Proposition 5 are tight is out of the scope of this paper and we leave it for future research[2]. However, we establish in the next proposition upper-bounds of the $L^p(\lambda)$ approximation error of $\mathcal{M}^d$ with feed-forward neural networks with the Heaviside activation function. This shows that, for the $L^p(\lambda)$ approximation error, the lower-bound obtained in (14), for $\nu = 0$, is tight up to logarithmic factors. The next proposition follows by reinterpreting a metric entropy upper bound of [GW07] in terms of Heaviside neural networks. The proof is postponed to Appendix E.1 in the supplement.

**Proposition 6.** *Let $1 \leq p < +\infty$, $d \in \mathbb{N} \setminus \{0, 1\}$ and let $\alpha = \max\{d, (d-1)p\}$. There exist positive constants $W_{\min}$ and $c$, depending only on $d$ and $p$, such that for any integer $W \geq W_{\min}$, there exists a feed-forward architecture $\mathcal{A}$ with two hidden layers, $W$ weights and the Heaviside activation function such that the set $H_{\mathcal{A}}$ satisfies*

$$\sup_{f \in \mathcal{M}^d} \inf_{g \in H_{\mathcal{A}}} \|f - g\|_{L^p(\lambda)} \leq \begin{cases} c W^{-\frac{1}{\alpha}} & \text{if } p(d-1) \neq d\,, \\ c W^{-\frac{1}{d}} \log(W) & \text{if } p(d-1) = d\,. \end{cases} \tag{15}$$

[2] Obtaining an upper-bound for ReLU networks seems challenging. For example, the bit extraction technique used in [Yar18] to find a sharp upper bound heavily relies on the local smoothness assumption of the function to approximate, which is not satisfied in general for monotonic functions.

## 5 Conclusion and other possible applications

We proved a general lower bound on the approximation error of $F$ by $G$ in $L^p(\mu)$ norm (Theorem 1), in terms of generic properties of $F$ and $G$ (packing number of $F$, range of $F$, fat-shattering dimension of $G$). The proof relies on VC dimension theory as in the sup norm case, but uses an additional key probabilistic argument due to Mendelson ([Men02], see Proposition 1), solving a question raised by DeVore et al. [DHP21].

In Sections 3 and 4 we detailed two applications, where Corollary 1 yields nearly optimal approximation lower bounds in $L^p$ norm, and which correspond to two examples where the approximation rate may depend or not depend on $p$.

Theorem 1 and Corollary 1 can be used to derive approximation lower bounds for many other cases. Corollary 1 only requires a (tight) lower bound on the packing number of $F$, for which approximation theory provides several examples. For instance, for the *Barron space* introduced in [Bar93], Petersen and Voigtlaender [PV21] showed a tight lower bound on the log packing number in $L^p(\lambda, [0,1]^d)$ norm, of order $\varepsilon^{-2d/(d+2)}$. Applying Corollary 1, this yields an approximation lower bound of $(LW)^{-\left(\frac{1}{2}+\frac{1}{d}\right)} \log^{-3\left(\frac{1}{2}+\frac{1}{d}\right)}(W)$ for ReLU networks (see Appendix F in the supplement for details). Other examples of sets $F$ for which tight lower bounds on the packing number (or metric entropy) are available include: multivariate cumulative distribution functions [BGL07], multivariate convex functions [GS13], and functions with other shape constraints [GJ14].

Theorem 1 can also be applied to other approximating sets $G$, beyond classical feed-forward neural networks, as soon as a (tight) upper bound on the fat-shattering dimension of $G$ is available. For example, upper bounds were derived by [WS22] on the VC dimension of partially quantized networks, while [Bel18] derived bounds on the fat-shattering dimension of some RKHS. Investigating such applications and whether the obtained approximation lower bounds are rate-optimal is a natural research direction for the future.


## Acknowledgements

The authors would like to thank Keridwen Codet for contributing to the results of Sections 4.1 and 4.3.

This work has benefited from the AI Interdisciplinary Institute ANITI, which is funded by the French "Investing for the Future – PIA3" program under the Grant agreement ANR-19-P3IA-0004. The authors gratefully acknowledge the support of IRT Saint Exupéry and the DEEL project.[3]


## References


[AB99] Martin Anthony and Peter L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge, 1999.

[ABDCBH97] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, jul 1997.

[Bar93] Andrew Barron. Universal approximation bounds for superpositions of a sigmoidal function. *Information Theory, IEEE Transactions on*, 39:930 – 945, 06 1993.

[Bel18] Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1348–1361. PMLR, 06–09 Jul 2018.

[BGL07] Ron Blei, Fuchang Gao, and Wenbo V. Li. Metric entropy of high dimensional distributions. *Proceedings of the American Mathematical Society*, 135(12):4009–4018, 2007.


[3] `https://www.deel.ai/`

[BHLM19] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[BS67] M Š Birman and M Z Solomjak. Piecewise-polynomial approximation of functions of the classes $w_p^\alpha$. 2(3):295–317, apr 1967.

[Cyb89] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems*, 2:303–314, 1989.

[DHM89] Ronald A DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *Manuscripta mathematica*, 63(4):469–478, 1989.

[DHP21] Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

[ET96] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, 1996.

[GJ04] Paul W. Goldberg and Mark Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 2004.

[GJ14] Piet Groeneboom and Geurt Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2014.

[GR21] Ingo Gühring and Mones Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 2021.

[GS13] Adityanand Guntuboyina and Bodhisattva Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, 2013.

[GW07] Fuchang Gao and Jon A. Wellner. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764, 2007.

[Hor91] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[KL20] Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2306–2327. PMLR, 09–12 Jul 2020.

[LLPS93] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

[Mai99] V.E Maiorov. On best approximation by ridge functions. *Journal of Approximation Theory*, 99(1):68–94, 1999.

[Men02] S. Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory*, 48, 2002.

[MMR99] V. Maiorov, R. Meir, and Joel Ratsaby. On the approximation of functional classes equipped with a uniform measure using ridge functions. *Journal of Approximation Theory*, 99:95–111, 1999.

[MP99] Vitaly Maiorov and Allan Pinkus. Lower bounds for approximation by mlp neural networks. *Neurocomputing*, 25(1):81–91, 1999.

[PV18] Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

[PV21] Philipp Petersen and Felix Voigtlaender. Optimal learning of high-dimensional classification problems using deep neural networks, 2021. arXiv:2112.12555.

[SX21] Jonathan W. Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and $n$-widths of shallow neural networks. 2021.

[SYZ22] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

[VP19] Felix Voigtländer and Philipp Petersen. Approximation in $L^p(\mu)$ with deep ReLU neural networks, 2019. arXiv:1904.04789.

[WS22] Yutong Wang and Clayton D. Scott. Vc dimension of partially quantized neural networks in the overparametrized regime. In *Proceedings of ICLR 2022*, 2022.

[Yar17] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

[Yar18] Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649, 2018.

[Yu97] Bin Yu. *Assouad, Fano, and Le Cam*, pages 423–435. Springer New York, New York, NY, 1997.

[YZ20] Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

[ZBH$^+$21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

# A general approximation lower bound in $L^p$ norm, with applications to feed-forward neural networks

Supplementary Material

This is the appendix for "A general approximation lower bound in $L^p$ norm, with applications to feed-forward neural networks".

## A Feed-forward neural networks: formal definition

In all this paper, we use the following classical graph-theoretic definitions for feed-forward neural networks given, e.g., in [BHLM19] (with slightly different terms and notation).

A *feed-forward neural network architecture* $\mathcal{A}$ of depth $L \geq 1$ is a directed acyclic graph $(V, E)$ with $d \geq 1$ nodes with in-degree 0 (also called the *input neurons*), a single node with out-degree 0 (also called the *output neuron*), and such that the longest path in the graph has length $L$.

We define layers $\ell = 0, 1, \ldots, L$ recursively as follows:

- layer 0 is the set $V_0$ of all input neurons; we assume that $V_0 = \{1, \ldots, d\}$ without loss of generality.
- for any $\ell = 1, \ldots, L$, layer $\ell$ is the set $V_\ell$ of all nodes that have one or several predecessors[4] in layer $\ell-1$, possibly other predecessors in layers $0, 1, \ldots, \ell-2$, but no other predecessors.

Layer $L$ consists of a single node: the output neuron. Layers $1, \ldots, L-1$ are called the *hidden layers*. Note that skip connections are allowed, i.e., there can be connections between non-consecutive layers.

Given a feed-forward neural network architecture $\mathcal{A}$ of depth $L \geq 1$, we associate real numbers $w_e \in \mathbb{R}$ to all edges $e \in E$ and $w_v \in \mathbb{R}$ to all nodes $v \in V_1 \cup \ldots \cup V_L$. These real numbers are called *weights* (they correspond to linear coefficients and biases) and are concatenated in a *weight vector* $\mathbf{w} \in \mathbb{R}^W$, where $W = \mathrm{Card}(E) + \sum_{\ell=1}^{L} \mathrm{Card}(V_\ell)$ is the total number of weights.

Given $\mathcal{A}$, an associated weight vector $\mathbf{w} \in \mathbb{R}^W$, and a function $\sigma : \mathbb{R} \to \mathbb{R}$ (called *activation function*), the network represents the function $g_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ defined recursively as follows. We write $P_v \subset V$ for the set of all predecessors of any node $v \in V$, and $w_{u \to v}$ for the weight on the edge from $u$ to $v$. The recursion from layer $\ell = 0$ to layer $\ell = L$ reads: given $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

- each input neuron $v \in \{1, \ldots, d\}$ outputs the value $y_v := x_v$;
- for any $\ell = 1, \ldots, L-1$, each neuron $v \in V_\ell$ outputs $y_v := \sigma\big(\sum_{u \in P_v} w_{u \to v} y_u + w_v\big)$;
- the unique output neuron $v \in V_L$ outputs $g_{\mathbf{w}}(x) := \sum_{u \in P_v} w_{u \to v} y_u + w_v$.

Finally, we define $H_{\mathcal{A}} := \{g_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^W\}$ to be the set of all functions that can be represented by tuning all the weights assigned to the network (the dependency on the activation function $\sigma$ is not written explicitly).

## B Main results: technical details

We provide technical details that were missing to establish Proposition 1, Theorem 1 and Corollary 1.

[4] A node $u \in V$ is a predecessor of another node $v \in V$ if there is a directed edge from $u$ to $v$.

### B.1 Proof of Proposition 1

Proposition 1 was originally stated by Mendelson [Men02, Corollary 3.12] when $[a,b] = [0,1]$ and when $G$ is a uniform Glivenko-Cantelli class (instead of the assumption on $\mathrm{fat}_\gamma(G)$). However, when $G$ only consists of $[0,1]$-valued functions, the fact that $G$ is a uniform Glivenko-Cantelli class is equivalent to $\mathrm{fat}_\gamma(G) < +\infty$ for all $\gamma > 0$ (see [ABDCBH97], Theorem 2.5, or [Men02], Theorem 2.4). Therefore, Corollary 3.12 in [Men02] can be rewritten as follows.

**Proposition 7** (Corollary 3.12 in [Men02], equivalent statement)**.** *Let $G$ be a set of functions from a measurable set $\mathcal{X}$ to $[0,1]$, such that $\mathrm{fat}_\gamma(G) < +\infty$ for all $\gamma > 0$. Then, for every $1 \le p < +\infty$, there is some constant $c_p > 0$ depending only on $p$ such that, for every probability measure $\mu$ on $\mathcal{X}$ and every $\varepsilon > 0$,*

$$\log M\left(\varepsilon, G, \|\cdot\|_{L^p(\mu)}\right) \le c_p \,\mathrm{fat}_{\frac{\varepsilon}{32}}(G) \log^2\left(\frac{2\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G)}{\varepsilon}\right).$$

We now explain how to derive Proposition 1 (with an arbitrary range $[a,b]$) as a straightforward consequence.

*Proof (of Proposition 1).* In order to apply Proposition 7, we reduce the problem from $[a,b]$ to $[0,1]$ by translating and rescaling every function in $G$. For $g \in G$, we define $\tilde{g} : \mathcal{X} \to [0,1]$ by $\tilde{g}(x) = \frac{g(x)-a}{b-a}$, and we set

$$\tilde{G} = \{\tilde{g} \;:\; g \in G\}\ .$$

Note that every $\tilde{g} \in \tilde{G}$ is indeed $[0,1]$-valued.

We now note that translation does not affect packing numbers nor the fat-shattering dimension, while rescaling only changes the scale $\varepsilon$ by a factor of $b-a$. More precisely, we have the following two properties:

**Property 1:** For all $u > 0$, $\mathrm{fat}_{\frac{u}{b-a}}(\tilde{G}) = \mathrm{fat}_u(G)$.

**Property 2:** For all $u > 0$, $M\Big(\frac{u}{b-a}, \tilde{G}, \|\cdot\|_{L^p(\mu)}\Big) = M\big(u, G, \|\cdot\|_{L^p(\mu)}\big)$.

Before proving the two properties (see below), we first conclude the proof of Proposition 1. By Property 1, $\mathrm{fat}_\gamma(\tilde{G}) = \mathrm{fat}_{\gamma(b-a)}(G)$, which by assumption is finite for all $\gamma > 0$. Since every $\tilde{g} \in \tilde{G}$ is $[0,1]$-valued, we can thus apply Proposition 7. Using it with $\tilde{\varepsilon} = \varepsilon/(b-a)$, we get

$$\log M\Big(\tilde{\varepsilon}, \tilde{G}, \|\cdot\|_{L^p(\mu)}\Big) \le c_p \,\mathrm{fat}_{\frac{\tilde{\varepsilon}}{32}}(\tilde{G}) \log^2\left(\frac{2\,\mathrm{fat}_{\frac{\tilde{\varepsilon}}{32}}(\tilde{G})}{\tilde{\varepsilon}}\right).$$

Combining with the two equalities in Properties 1 and 2, we obtain

$$\log M\big(\varepsilon, G, \|\cdot\|_{L^p(\mu)}\big) \le c_p \,\mathrm{fat}_{\frac{\varepsilon}{32}}(G) \log^2\left(\frac{2(b-a)\,\mathrm{fat}_{\frac{\varepsilon}{32}}(G)}{\varepsilon}\right),$$

which concludes the proof of Proposition 1.

We now prove the two properties.

**Proof of Property 1.** We first show that $\mathrm{fat}_{\frac{u}{b-a}}(\tilde{G}) \ge \mathrm{fat}_u(G)$. To that end, let $S = \{x_1, \dots, x_m\}$ and $r : S \to \mathbb{R}$ be such that for any $E \subset S$, there exists $g \in G$ such that $g(x) > r(x) + u$ if $x \in E$ and $g(x) < r(x) - u$ otherwise. Setting $\tilde{r}(x) = \frac{r(x)-a}{b-a}$, we can see that $\tilde{g}(x) > \tilde{r}(x) + \frac{u}{b-a}$ if $x \in E$ and $\tilde{g}(x) < \tilde{r}(x) - \frac{u}{b-a}$ otherwise, which proves $\mathrm{fat}_{\frac{u}{b-a}}(\tilde{G}) \ge \mathrm{fat}_u(G)$. The reverse inequality is proved similarly.

**Proof of Property 2.** Let $\{g_1, \dots, g_m\}$ be a $u$-packing of $G$ in $L^p(\mu)$ norm. This means that $\|g_i - g_j\|_{L^p(\mu)} > u$ and therefore $\|\tilde{g}_i - \tilde{g}_j\|_{L^p(\mu)} > \frac{u}{b-a}$ for all $i \ne j \in \{1, \dots, m\}$, so that $\{\tilde{g}_1, \dots, \tilde{g}_m\} \subset \tilde{G}$ is a $\frac{u}{b-a}$-packing of $\tilde{G}$. This proves $M\big(\frac{u}{b-a}, \tilde{G}, \|\cdot\|_{L^p(\mu)}\big) \ge M\big(u, G, \|\cdot\|_{L^p(\mu)}\big)$. The reverse inequality is proved similarly. ∎

### B.2 Clipping can only help

The next two lemmas indicate that clipping (truncature) to a known range can only help. These are key to apply Proposition 1 in our setting. In the sequel, for a set $G$ of functions from a measurable set $\mathcal{X} \subset \mathbb{R}^d$ to $\mathbb{R}$, and for $a < b$, we denote by $G_{[a,b]}$ the set of all functions in $G$ whose values are truncated (clipped) to the segment $[a, b]$, that is, $G_{[a,b]} = \{\tilde{g} : g \in G\}$, where $\tilde{g} : \mathcal{X} \to \mathbb{R}$ is given by

$$\forall x \in \mathcal{X}, \quad \tilde{g}(x) = \min(\max(a, g(x)), b) \ .$$

**Lemma 3.** *Let $G$ be a set of functions defined on a set $\mathcal{X}$, and with values in $\mathbb{R}$. Let $G_{[a,b]}$ be defined as above. Then, for any $\gamma > 0$,*

$$\mathrm{fat}_\gamma(G) \geq \mathrm{fat}_\gamma(G_{[a,b]}) \ .$$

*Proof.* Let $\gamma > 0$. The case when $\mathrm{fat}_\gamma(G_{[a,b]}) = 0$ is straightforward. We thus assume that $\mathrm{fat}_\gamma(G_{[a,b]}) \geq 1$. To prove the result, we show that any subset $A$ of $X$ that is $\gamma$-*shattered* by $G_{[a,b]}$ is also $\gamma$-*shattered* by $G$. Let us consider such a subset $A = \{x^1, \ldots, x^N\} \subset X$, with cardinality $N \geq 1$. Hence, there exists $\{r_1, \ldots, r_N\} \subset \mathbb{R}$ such that for any $E \subset A$, there exists $\tilde{g} \in G_{[a,b]}$ such that $\tilde{g}(x_i) - r_i > \gamma$ if $x_i \in E$ and $\tilde{g}(x_i) - r_i < -\gamma$ otherwise. Note that this must imply that $r_i \in ]a, b[$ for all $i = 1, \ldots, N$ (indeed, by choosing $E$ such that $x_i \in E$ or not, we have either $r_i + \gamma < \tilde{g}(x_i) \leq b$ or $r_i - \gamma > \tilde{g}(x_i) \geq a$). Now fix $i \in \{1, \ldots, N\}$ and let us assume $\tilde{g}(x_i) - r_i > \gamma$ (by symmetry, the reversed case $\tilde{g}(x_i) - r_i < -\gamma$ is treated the same way). Because $r_i > a$, this implies that $\tilde{g}(x_i) > a$ and thus $g(x_i) \geq \tilde{g}(x_i)$ (by definition of $\tilde{g}$), which entails $g(x_i) - r_i > \gamma$. It follows that if $G_{[a,b]}$ $\gamma$-*shatters* $A$, then $G$ also $\gamma$-*shatters* $A$, and the result follows. □

The following lemma formalizes the well-known idea that it is easier to approach a function with values in a finite range by a function with values in the same range.

**Lemma 4.** *Let $G$ be a set of functions from a measurable set $\mathcal{X} \subset \mathbb{R}^d$ to $\mathbb{R}$ and let $G_{[a,b]}$ be defined as above. Assume $F$ is a set of functions from $\mathcal{X}$ to $[a, b]$. Then, for any probability measure $\mu$ on $\mathcal{X}$,*

$$\sup_{f \in F} \inf_{g \in G} \|f - g\|_{L^p(\mu)} \geq \sup_{f \in F} \inf_{\tilde{g} \in G_{[a,b]}} \|f - \tilde{g}\|_{L^p(\mu)} \ .$$

*Proof.* To prove the above result, it is enough to show that for any $f \in F$ and $g \in G$, the function $\tilde{g}$ is pointwise at least as close to $f$ as $g$ is, which for all $f \in F$ yields $\inf_{g \in G} \|f - g\|_{L^p(\mu)} \geq \inf_{\tilde{g} \in G_{[a,b]}} \|f - \tilde{g}\|_{L^p(\mu)}$. By definition of $G_{[a,b]}$, for any $x \in \mathcal{X}$, if $g(x) \in [a, b]$, then $|f(x) - g(x)| = |f(x) - \tilde{g}(x)|$. And if $g(x) \notin [a, b]$, then $|f(x) - \tilde{g}(x)| < |f(x) - g(x)|$ since $f(x) \in [a, b]$. It follows that the discrepancy $|f - \tilde{g}|$ is everywhere bounded by $|f - g|$, and the result follows. □

### B.3 Missing details in the proof of Theorem 1

We provide all details that were missing to derive (9), which is a direct consequence of Lemma 5 below. We follow the convention $aP^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P) = +\infty$ when $P = 1$.

**Lemma 5.** *Let $P \in \mathbb{N}^*$ and $c, \alpha, r > 0$. There exist constants $a, \varepsilon_1'' > 0$ depending only on $c$, $\alpha$ and $r$ such that, for all $\varepsilon \in (0, r)$ satisfying*

$$\varepsilon^{-\alpha} \leq cP \log^2\left(\frac{rP}{\varepsilon}\right) \ , \tag{16}$$

*we have*

$$\varepsilon \geq \min\left(\varepsilon_1'', aP^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P)\right) .$$

*Proof.* Assume $\varepsilon \in (0, r)$ is such that (16) holds. To show the result, we study the function $f : (1/r, +\infty) \to \mathbb{R}$ defined for all $x > 1/r$ by

$$f(x) = \frac{x^\alpha}{\log^2(rPx)} \ .$$

Note that (16) implies that $f(1/\varepsilon) \leq cP$. For all $P \geq 2$, we set

$$\varepsilon_P = P^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P) \ . \tag{17}$$

Let $P_1 \geq 2$ be such that $P_1^{\frac{1}{\alpha}} \log^{\frac{2}{\alpha}}(P_1) \geq \frac{\exp(\frac{2}{\alpha})}{r}$. For all $P \geq P_1$, we have $\frac{1}{\varepsilon_P} \geq \frac{\exp(\frac{2}{\alpha})}{r} > 1/r$ and

$$f\left(\frac{1}{\varepsilon_P}\right) = \frac{P \log^2(P)}{\log^2\left(rP^{1+\frac{1}{\alpha}} \log^{\frac{2}{\alpha}}(P)\right)} \ .$$

Since

$$\lim_{Q \to +\infty} \frac{\log^2(Q)}{\log^2\left(rQ^{1+\frac{1}{\alpha}} \log^{\frac{2}{\alpha}}(Q)\right)} = \frac{1}{(1+\frac{1}{\alpha})^2} =: c_1 \ ,$$

there exists $P_2$ such that for all $Q \geq P_2$, we have $\frac{\log^2(Q)}{\log^2\left(rQ^{1+\frac{1}{\alpha}} \log^{\frac{2}{\alpha}}(Q)\right)} \geq \frac{c_1}{2}$ .

Below we distinguish the cases $P \geq \max(P_1, P_2)$ and $P < \max(P_1, P_2)$.

**1st case:** $P \geq \max(P_1, P_2)$.
We have $f\left(\frac{1}{\varepsilon_P}\right) \geq \frac{c_1 P}{2}$ and $P \geq \frac{1}{c} f\left(\frac{1}{\varepsilon}\right)$ (by (16)), so that $f\left(\frac{1}{\varepsilon_P}\right) \geq \frac{c_1}{2c} f\left(\frac{1}{\varepsilon}\right)$. We now use Lemma 6 below with $b = \frac{c_1}{2c}$: setting $a := (b/2)^{1/\alpha} = (c_1/(4c))^{1/\alpha}$, there exists $x_1 > \max\left\{\frac{1}{r}, \frac{1}{ar}\right\}$ depending only on $r, b, \alpha$ such that $bf(x) \geq f(ax)$ for all $x \geq x_1$.

Therefore, if $\varepsilon < \frac{1}{x_1} =: \varepsilon_1$, then $\frac{c_1}{2c} f\left(\frac{1}{\varepsilon}\right) \geq f\left(\frac{a}{\varepsilon}\right)$. Therefore $f(\frac{1}{\varepsilon_P}) \geq f\left(\frac{a}{\varepsilon}\right)$.

Recall from (17) and $P \geq P_1$ that $\frac{1}{\varepsilon_P} \geq \frac{\exp(\frac{2}{\alpha})}{r}$. If $\varepsilon < \frac{ar}{\exp(\frac{2}{\alpha})} =: \varepsilon_2$, then we also have $\frac{a}{\varepsilon} \geq \frac{\exp(\frac{2}{\alpha})}{r}$. Therefore, using Lemma 6 again, $f(\frac{1}{\varepsilon_P}) \geq f(\frac{a}{\varepsilon})$ implies that $\frac{1}{\varepsilon_P} \geq \frac{a}{\varepsilon}$, that is,

$$\varepsilon \geq a\, \varepsilon_P \ .$$

Summarizing, when $\varepsilon \in (0, r)$ satisfies (16) and when $P \geq \max(P_1, P_2)$, either $\varepsilon \geq \varepsilon_1$ or $\varepsilon \geq \varepsilon_2$ or $\varepsilon \geq a\, \varepsilon_P$. Put differently,

$$\varepsilon \geq \min(\varepsilon_1, \varepsilon_2, a\, \varepsilon_P) \ . \tag{18}$$

**2nd case:** $P < \max(P_1, P_2) =: P_3$.
Using (16) and the fact that $t \mapsto ct \log^2\left(\frac{rt}{\varepsilon}\right)$ is non-decreasing on $[\varepsilon/r, +\infty)$, together with $\varepsilon/r \leq 1 \leq P \leq P_3$ yields $\varepsilon^{-\alpha} \leq cP_3 \log^2(rP_3/\varepsilon)$. This entails that, for some $\varepsilon_3 > 0$ depending only on $\alpha, c, P_3, r$,

$$\varepsilon \geq \varepsilon_3 \ . \tag{19}$$

**Conclusion:** combining the two cases, when $\varepsilon \in (0, r)$ satisfies (16), whatever $P \in \mathbb{N}^*$, we have (18) or (19). Setting $\varepsilon_1'' = \min(\varepsilon_1, \varepsilon_2, \varepsilon_3)$, we obtain

$$\varepsilon \geq \min\left(\varepsilon_1'', a\, P^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P)\right) .$$

(Note that this is also true in the case $P = 1$, by the convention $aP^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P) = +\infty$.) Since $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and $a$ only depend on $c, \alpha, r$, this concludes the proof. □

**Lemma 6.** *Let $\alpha, r > 0$ and $P \in \mathbb{N}^*$. We define $f(x) = \frac{x^\alpha}{\log^2(rPx)}$ for all $x > 1/r$. Then:*

*i) $f$ is increasing on $I := \left[\frac{\exp(\frac{2}{\alpha})}{r}, +\infty\right)$ and $\lim_{x \to +\infty} f(x) = +\infty$.*

*ii) for all $b > 0$, setting $a := (b/2)^{1/\alpha}$, there exists $x_1 > \max\left\{\frac{1}{r}, \frac{1}{ar}\right\}$ depending only on $r, b, \alpha$ such that,*

$$\forall x \geq x_1 \ , \qquad bf(x) \geq f(ax) \ .$$

*Proof.* **Proof of i):** The fact that $\lim_{x\to+\infty} f(x) = +\infty$ is because $\alpha > 0$. To see why $f$ is increasing on $I$, note that

$$f'(x) = \frac{\alpha x^{\alpha-1}\log^2(rPx) - x^\alpha 2\log(rPx)\frac{1}{x}}{\log^4(rPx)} = \frac{x^{\alpha-1}\log(rPx)(\alpha\log(rPx)-2)}{\log^4(rPx)} ,$$

so that $f'(x) > 0$ for all $x > \frac{\exp(\frac{2}{\alpha})}{rP}$, and in particular for all $x > \frac{\exp(\frac{2}{\alpha})}{r}$ (since $P \geq 1$). This proves that $f$ is increasing on $I$.

**Proof of ii):** Let $b > 0$ and set $a := (b/2)^{1/\alpha}$. Let $x_1 > \max\{\frac{1}{r}, \frac{1}{ar}\}$ depending only on $r, b, \alpha$ such that, for all $u \geq x_1$,

$$\frac{\log^2(ru)}{\log^2(rau)} \leq 2 .$$

(Such an $x_1$ exists since the ratio converges to 1 as $u \to +\infty$, and we can choose $x_1$ as a function of $r, a$ only.) Now, for all $x \geq x_1$, using the above inequality with $u = Px \geq x$ (since $P \geq 1$), we get

$$\frac{f(ax)}{f(x)} = a^\alpha \frac{\log^2(rPx)}{\log^2(rPax)} \leq 2a^\alpha = b ,$$

where the last equality is because $a := (b/2)^{1/\alpha}$. This proves that $bf(x) \geq f(ax)$ for all $x \geq x_1$. □

### B.4 Proof of Corollary 1

We first recall two key bounds on the VC-dimension of piecewise-polynomial feed-forward neural networks, proved by [GJ04] and [BHLM19].

Let $\mathcal{B}$ be any feed-forward neural network architecture of depth $L \geq 1$ with $W \geq 1$ weights, $d \geq 1$ input neurons, and $U \geq 1$ hidden or output neurons. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any piecewise-polynomial activation function on $K \geq 2$ pieces, with maximal degree $\nu \in \mathbb{N}$. Denote by $\mathrm{sgn}(H_\mathcal{B}) = \{\mathrm{sgn}(g_\mathbf{w}) : \mathbf{w} \in \mathbb{R}^W\}$ the set of all classifiers obtained by looking at the sign of the network's output, that is, the classifiers defined by $\mathrm{sgn}(g_\mathbf{w})(x) = \mathbb{1}_{\{g_\mathbf{w}(x)>0\}}$ for all $x \in \mathbb{R}^d$.

Goldberg and Jerrum [GJ04] showed that, for some constant $c_1' > 0$ depending only on $d$, $\nu$ and $K$, the VC-dimension of $\mathrm{sgn}(H_\mathcal{B})$ is bounded as follows (see also Theorem 8.7 in [AB99]):

$$\mathrm{VCdim}(\mathrm{sgn}(H_\mathcal{B})) \leq c_1' W^2 . \tag{20}$$

This bound was refined for piecewise-affine activation functions. Namely, Bartlett et al. [BHLM19, Theorem 7] proved that, if $U \geq 3$, then, for some $R \leq U + U(L-1)\nu^{L-1}$,

$$\mathrm{VCdim}(\mathrm{sgn}(H_\mathcal{B})) \leq L + \bar{L}W\log_2\Big(4e(K-1)R\log_2\big(2e(K-1)R\big)\Big) ,$$

where $\bar{L} = 1$ if $\nu = 0$, and $\bar{L} \leq L$ otherwise. Therefore, for some constants $W'_{\min} \geq 1$ and $c_2', c_3' > 0$ depending only on $d$ and $K$, we have, for all $W \geq W'_{\min}$ (which in particular implies $U \geq 3$),

$$\mathrm{VCdim}(\mathrm{sgn}(H_\mathcal{B})) \leq \begin{cases} c_2' LW\log(W) & \text{if } \nu = 1 , \\ c_3' W\log(W) & \text{if } \nu = 0 . \end{cases} \tag{21}$$

We are now ready to prove Corollary 1 from Theorem 1.

*Proof (of Corollary 1).* In order to apply Theorem 1, we first bound $P := \mathrm{Pdim}(H_\mathcal{A})$ from above. The bounds (20) and (21) were on the VC-dimension of $\mathrm{sgn}(H_\mathcal{B})$, for any feed-forward neural network architecture $\mathcal{B}$, while we need a bound on the pseudo-dimension. However, by a well-known trick (e.g., Theorem 14.1 in [AB99]), the pseudo-dimension of $H_\mathcal{A}$ is upper bounded by the VC-dimension of (the sign of) an augmented network architecture of depth $L$, with $d+1$ input neurons and $W+1$ weights.[5] Therefore, replacing $(d, W)$ with $(d+1, W+1)$ in (20) and (21),

[5] This is because $\mathrm{Pdim}(H_\mathcal{A}) = \mathrm{VCdim}\big(\{(x,r) \in \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{1}_{\{g(x)-r>0\}} : g \in H_\mathcal{A}\}\big)$, the output neuron of $\mathcal{A}$ is linear, and we allow skip connections.

we get that, for some constants $\tilde{W}_{\min} \geq 1$ and $\tilde{c_1}, \tilde{c_2}, \tilde{c_3} > 0$ depending only on $d$, $\nu$ and $K$, for all $W \geq \tilde{W}_{\min}$,

$$P \leq \begin{cases} \tilde{c_1} W^2 & \text{if } \nu \geq 2 \,, \\ \tilde{c_2} L W \log(W) & \text{if } \nu = 1 \,, \\ \tilde{c_3} W \log(W) & \text{if } \nu = 0 \,. \end{cases} \tag{22}$$

Now using Theorem 1, we have

$$\sup_{f \in F} \inf_{g \in H_{\mathcal{A}}} \|f - g\|_{L^p(\mu)} \geq \min \left\{ \varepsilon_1,\ c_1 P^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P) \right\} \,. \tag{23}$$

Noting that $P \mapsto \min \left\{ \varepsilon_1,\ c_1 P^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(P) \right\}$ is non-increasing and plugging (22) into (23), we get, for $W \geq W_{\min}$,

$$\sup_{f \in F} \inf_{g \in H_{\mathcal{A}}} \|f - g\|_{L^p(\mu)} \geq \min \left\{ \varepsilon_1, \begin{pmatrix} c_4 W^{-\frac{2}{\alpha}} \log^{-\frac{2}{\alpha}}(W^2) & \text{if } \nu \geq 2 \\ c_5 (L W \log(W))^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(L W \log(W)) & \text{if } \nu = 1 \\ c_6 (W \log(W))^{-\frac{1}{\alpha}} \log^{-\frac{2}{\alpha}}(W \log(W)) & \text{if } \nu = 0 \end{pmatrix} \right\}$$

for some constants $W_{\min} \geq 1$ and $c_4, c_5, c_6 > 0$ depending only on $d$, $\nu$ and $K$. Taking $W_{\min}$ large enough, the first term $\varepsilon_1$ is always larger than the second term in the above minimum, and the logarithmic terms $\log(W \log(W))$ and $\log(LW \log(W))$ can be upper bounded by a constant times $\log(W)$ (since $L \leq W$). Rearranging concludes the proof. □

# C Earlier works: two other lower bound proof strategies

Approximation lower bounds in a sense similar to ours have been obtained in other recent works. In the purpose of highlighting the differences between our approaches, we describe the lower bound proof strategies of Yarotsky [Yar17] and of Petersen and Voigtlaender [PV18].

## C.1 Approximation in sup norm of Sobolev unit balls with ReLU networks [Yar17]

Recall that the Sobolev space $\mathcal{W}^{n,\infty}([0,1]^d)$ is defined as the set of functions on $[0,1]^d$ lying in $L^\infty$ along with all their weak derivatives up to order $n$. We equip this space with the norm

$$\|f\|_{\mathcal{W}^{n,\infty}([0,1]^d)} = \max_{\mathbf{n} \in \mathbb{N}^d : |\mathbf{n}| \leq n} \operatorname*{ess\,sup}_{x \in [0,1]^d} |D^{\mathbf{n}} f(x)|,$$

and we let $F_{n,d}$ be the unit ball of this space.

We first state the sup norm lower bound and then we give a synthesized version of the proof.

**Proposition 8** ([Yar17])**.** *There exists positive constants $W_{\min}, c > 0$ such that for any feed-forward neural network with architecture $\mathcal{A}$, ReLU activation and $W \geq W_{\min}$ weights,*

$$\sup_{f \in F_{n,d}} \inf_{g \in H_{\mathcal{A}}} \|f - g\|_\infty \geq c W^{-\frac{2n}{d}}.$$

Details aside, the proof reads as follows. The author assumes that $H_{\mathcal{A}}$ approximates $F_{n,d}$ with error $\varepsilon$. Fixing $N = c(3\varepsilon)^{-1/n}$ for some constant $c > 0$ properly chosen, he constructs a set of functions in $F_{n,d}$ that can shatter a grid of $N^d$ points $x_1, \ldots, x_{N^d}$ evenly distributed over $[0,1]^d$. The assumption that $H_{\mathcal{A}}$ approximates $F_{n,d}$ in sup norm with error $\varepsilon$ allows to conclude that $H_{\mathcal{A}}$ also shatters $\{x_1, \ldots, x_{N^d}\}$, and hence, $\mathrm{VCdim}(H_{\mathcal{A}}) \geq N^d = c_{n,d} \varepsilon^{-\frac{d}{n}}$. The author concludes using the upper bound on $\mathrm{VCdim}(H_{\mathcal{A}})$ with respect to $W$ from [AB99] which yields $\mathrm{VCdim}(H_{\mathcal{A}}) \leq c' W^2$ for some constant $c'$.

It is worth stressing that in this proof, it is paramount to assume that $H_{\mathcal{A}}$ approximates $F_{n,d}$ in sup norm, rather than any $L^p$ norm with $p < +\infty$. The reason is that only this choice of norm allows to bound the discrepancy between $f \in F_{n,d}$ and $g_f \in H_{\mathcal{A}}$ chosen optimally with respect to $f$ at any chosen points. Our proof strategy relying on Proposition 1 allows to circumvent this issue by relating the pseudo-dimension to the metric entropy with respect to any $L^p$ norm, $1 \leq p < +\infty$.

### C.2 Approximation in $L^p$ norm of *Horizon functions* with quantized networks [PV18]

The authors study *quantized* neural networks, that is, networks with weights constrained to be representable with a fixed number of bits. They obtain a lower bound on the minimal number of weights in a quantized neural network that can approximate a set of *Horizon functions* in $L^p$ norm, $p > 0$, with error $\varepsilon > 0$. This lower bound is easily invertible to a bound on the approximation error and is thus comparable to the results we obtain in this paper.

Textually, the authors introduce the set of horizon functions as follows: "These are $\{0,1\}$-valued functions with a jump along a hypersurface and such that the jump surface is the graph of a smooth function" [PV18]. Denoting by $H$ the indicator function of the set $[0,+\infty) \times \mathbb{R}^{d-1}$, the set of horizon functions reads as

$$\mathcal{HF}_{\beta,d,B} = \left\{ f \circ T \in L^\infty\left(\left[-\frac{1}{2},\frac{1}{2}\right]^d\right) : \right.$$
$$\left. f(x) = H(x_1 + \gamma(x_2,\dots,x_d), x_2, \dots, x_d), \gamma \in \mathcal{F}_{\beta,d-1,B},\ T \in \Pi(d,\mathbb{R}) \right\},$$

where $\mathcal{F}_{\beta,d-1,B}$ denotes the set of Hölder functions over $[-1/2,1/2]^{d-1}$ whith smoothness parameter $\beta$ and with norm bounded by $B$ (see section 3), and $\Pi(d,\mathbb{R})$ denotes the group of $d$-dimensional permutation matrices.

In the following, for any nonzero integer $K$ and any neural network architecture $\mathcal{A}$, we denote by $H_{\mathcal{A}}^K \subset H_{\mathcal{A}}$ the set of $K$-quantized functions in $H_{\mathcal{A}}$; namely, the functions in $H_{\mathcal{A}}$ with weights representable over at most $K$ bits. The lower bound in [PV18] (Theorem 4.2) reads as follow:

**Proposition 9** ([PV18])**.** *Let $d \geq 2$. Let $p, \beta, B, c_0 > 0$ and let $\sigma : \mathbb{R} \to \mathbb{R}$ be such that $\sigma(0) = 0$. There exists positive constants $\varepsilon_0, c > 0$ depending only on $d, p, \beta, B$ and $c_0$ such that, for any $\varepsilon \leq \varepsilon_0$, setting $K = \lceil c_0 \log(1/\varepsilon) \rceil$, for any feed-forward neural network architecture $\mathcal{A}$ with $W$ weights and activation $\sigma$ such that $H_{\mathcal{A}}^K$ approximates $\mathcal{HF}_{\beta,d,B}$ in $L^p$ norm with error less than $\varepsilon$, we have*

$$W \geq c\varepsilon^{-\frac{p(d-1)}{\beta}} \log^{-1}(1/\varepsilon).$$

The proof of this result is based on a lemma giving a lower bound on the minimal number of bits $\ell$ necessary for a binary encoder-decoder pair to achieve an error less than $\varepsilon > 0$ in approximating $\mathcal{HF} := \mathcal{HF}_{\beta,d,B}$ in $L^p$ norm. Formally, given an integer $\ell > 0$, a binary encoder $E^\ell : \mathcal{HF} \to \{0,1\}^\ell$ and given a decoder $D^\ell : \{0,1\}^\ell \to \mathcal{HF}$, one can measure an approximation error

$$\sup_{f \in \mathcal{HF}} \|f - D^\ell(E^\ell(f))\|_{L^p},$$

which quantifies the loss of information due to the encoding $E^\ell$. Clearly, for an optimal choice of encoder, one can reduce this loss of information by increasing $\ell$. In particular, for $\varepsilon > 0$, it is possible to estimate

$$\ell_\varepsilon = \min\left\{ \ell > 0 \ : \inf_{E^\ell, D^\ell} \sup_{f \in \mathcal{HF}} \|f - D^\ell(E^\ell(f))\|_{L^p} \leq \varepsilon \right\},$$

with the convention that $\ell_\varepsilon = \infty$ if the above set is empty. The authors show that for $\varepsilon$ small enough (smaller than some $\varepsilon_0 > 0$), it holds that

$$\ell_\varepsilon \geq c\varepsilon^{-\frac{p(d-1)}{\beta}} \tag{24}$$

for some constant $c > 0$ depending only on $d, p, \beta$ and $B$. In other words, one can not achieve a loss of information smaller than $\varepsilon$ by encoding functions in $\mathcal{HF}$ over less than $c\varepsilon^{-\frac{p(d-1)}{\beta}}$ bits.

The rest of the proof consists in showing that for an integer $K > 0$, given a neural network architecture $\mathcal{A}$ with $W$ weight that can approximate $\mathcal{HF}$ in $L^p$ norm with error less than $\varepsilon > 0$, one can encode exactly (without loss of information, and for a given activation function) any function in $H_{\mathcal{A}}^K$ over a string of $\ell = c_1 W(K + \lceil \log_2 W \rceil)$ bits. This generates a natural encoder-decoder system where any function $f \in \mathcal{HF}$ is encoded as the bit string of length $\ell$ associated to $g_f \in H_{\mathcal{A}}^K$ chosen

to approximate $f$. It remains to observe that if we fix $K$, this automatically yields a lower bound on $\ell$ using inequality (24), and thus on $W$ by expressing $W$ through $\ell$ and $K$.

**Remark.** The authors in [PV18] study the neural network approximation in a setting slightly different from ours, since they focus on the approximation by quantized neural networks. This partly explains why their proof strategy differs from ours. However, it is worth pointing out that the proof of their lower bound on the minimal number of bits required to accurately encode a function in $\mathcal{HF}$ relies on a lower bound of the packing number of $\mathcal{HF}$, just like the lower bound of the packing number of the set to approximate is key in our proof strategy. An interesting question for the future would be to see whether our general lower bound (Theorem 1) yields lower bounds of the same order as those in [PV18] for quantized neural networks.

# D Hölder balls

## D.1 Proof of Lemma 2

Let $N \in \mathbb{N}^*$. For $\mathbf{m} = (m_1, \dots, m_d) \in \{0, \dots, N-1\}^d$, we let $x_{\mathbf{m}} := \frac{1}{N}(m_1 + 1/2, \dots, m_d + 1/2)$ and we denote by $C_{\mathbf{m}}$ the cube of side-length $\frac{1}{N}$ centered at $x_{\mathbf{m}}$, with sides parallel to the axes. We see that the $N^d$ cubes $C_{\mathbf{m}}$ decompose the cube $[0,1]^d$ in smaller parts which, up to negligible sets which will not be problematic, form a partition of $[0,1]^d$. We will use this decomposition to construct a packing of $F_{s,d}$. Denoting $\|\cdot\|$ the sup norm in $\mathbb{R}^d$, we define the $C^\infty$ test function $\phi : \mathbb{R}^d \to \mathbb{R}$ by:

$$\phi(x) = \exp\left(-\frac{\|x\|^2}{1 - \|x\|^2}\right),$$

for any $x \in \mathbb{R}^d$ such that $\|x\| < 1$, and $\phi(x) = 0$ otherwise. Recalling that $n \in \mathbb{N}$ and $\alpha \in (0,1]$ are such that $s = n + \alpha$, and since all the high-order partial derivatives of $\phi$ are uniformly bounded on $[0,1]^d$, $\|\phi\|_{\mathcal{C}^{n,\alpha}}$ is thus finite and is nonzero.

Let $c_s = \frac{1}{2}(2N)^{-s}\|\phi\|^{-1}_{\mathcal{C}^{n,\alpha}}$ and consider, for any tensor of signs $\sigma = (\sigma_{\mathbf{m}})_{\mathbf{m} \in \{0,\cdots,N-1\}^d} \in \{-1,1\}^{N^d}$, the function $f_\sigma$ defined as follows:

$$f_\sigma(x) = c_s \sum_{\mathbf{m} \in \{0,\dots,N-1\}^d} \sigma_{\mathbf{m}} \phi\left(2N(x - x_{\mathbf{m}})\right),$$

for all $x \in [0,1]^d$. There are $2^{N^d}$ different functions $f_\sigma$.

Let us prove that, for all $\sigma \in \{-1,1\}^{N^d}$, $f_\sigma \in F_{s,d}$. To do so, we study the constituents of $\|f_\sigma\|_{\mathcal{C}^{n,\alpha}}$ separately and show that they are all bounded by 1. For $\mathbf{m} \in \{0, \cdots, N-1\}^d$, we define the function $g_{\mathbf{m}}(x) = c_s \sigma_{\mathbf{m}} \phi\left(2N(x - x_{\mathbf{m}})\right)$. Note that because $\phi$ cancels outside $(-1,1)^d$, we have that $g_{\mathbf{m}}$ cancels everywhere outside the interior of $C_{\mathbf{m}}$, and the same holds for $D^{\mathbf{n}} g_{\mathbf{m}}$ for all $\mathbf{n} \in \mathbb{N}^d$ such that $|\mathbf{n}| \leq n$. For any such $\mathbf{n}$, we have

$$\|D^{\mathbf{n}} g_{\mathbf{m}}\|_\infty = c_s (2N)^{|\mathbf{n}|} \|D^{\mathbf{n}} \phi\|_\infty \leq c_s (2N)^s \|\phi\|_{\mathcal{C}^{n,\alpha}} \leq \frac{1}{2}.$$

Therefore,

$$\max_{\mathbf{n}:|\mathbf{n}| \leq n} \|D^{\mathbf{n}} f_\sigma\|_\infty \leq 1.$$

Now for any $\mathbf{n} \in \mathbb{N}^d$ such that $|\mathbf{n}| = n$, any $x, y \in [0,1]^d$, we have

$$\frac{|D^{\mathbf{n}} f_\sigma(x) - D^{\mathbf{n}} f_\sigma(y)|}{\|x - y\|_2^\alpha} = \frac{|D^{\mathbf{n}} g_{\mathbf{m}}(x) - D^{\mathbf{n}} g_{\mathbf{m}'}(y)|}{\|x - y\|_2^\alpha},$$

where $x \in C_{\mathbf{m}}$ and $y \in C_{\mathbf{m}'}$ for some multi-indexes $\mathbf{m}$ and $\mathbf{m}'$. We have to distinguish between the cases $\mathbf{m} = \mathbf{m}'$ and $\mathbf{m} \neq \mathbf{m}'$. In the former case, we have

$$\begin{aligned}
\frac{|D^{\mathbf{n}} f_\sigma(x) - D^{\mathbf{n}} f_\sigma(y)|}{\|x - y\|_2^\alpha} &= c_s (2N)^{n+\alpha} \frac{|D^{\mathbf{n}} \phi(2N(x - x_{\mathbf{m}})) - D^{\mathbf{n}} \phi(2N(y - x_{\mathbf{m}}))|}{\|2N(x - x_{\mathbf{m}}) - 2N(y - x_{\mathbf{m}})\|_2^\alpha} \\
&= c_s (2N)^s \frac{|D^{\mathbf{n}} \phi(x') - D^{\mathbf{n}} \phi(y')|}{\|x' - y'\|_2^\alpha} \\
&\leq c_s (2N)^s \|\phi\|_{\mathcal{C}^{n,\alpha}} = \frac{1}{2},
\end{aligned}$$

where at the second line, we used the changes of variables $x' = 2N(x - x_{\mathbf{m}})$ and $y' = 2N(y - x_{\mathbf{m}})$. In the case $\mathbf{m} = \mathbf{m}'$ ($x$ and $y$ belong to the same cube), we thus have

$$\frac{|D^{\mathbf{n}} f_\sigma(x) - D^{\mathbf{n}} f_\sigma(y)|}{\|x - y\|_2^\alpha} \leq 1.$$

In the case $\mathbf{m} \neq \mathbf{m}'$, observe that we have

$$|D^{\mathbf{n}} g_{\mathbf{m}}(x) - D^{\mathbf{n}} g_{\mathbf{m}'}(y)| \leq 2 \max\{|D^{\mathbf{n}} g_{\mathbf{m}}(x)|, |D^{\mathbf{n}} g_{\mathbf{m}'}(y)|\}. \tag{25}$$

Besides, recall that $D^{\mathbf{n}} g_{\mathbf{m}}$ and $D^{\mathbf{n}} g_{\mathbf{m}'}$ both cancel outside of the interiors of $C_{\mathbf{m}}$ and $C_{\mathbf{m}'}$ respectively. We can thus rewrite (25) as

$$\begin{aligned}
|D^{\mathbf{n}} g_{\mathbf{m}}(x) - D^{\mathbf{n}} g_{\mathbf{m}'}(y)| &\leq 2 \max\{|D^{\mathbf{n}} g_{\mathbf{m}}(x) - D^{\mathbf{n}} g_{\mathbf{m}}(y)|, |D^{\mathbf{n}} g_{\mathbf{m}'}(x) - D^{\mathbf{n}} g_{\mathbf{m}'}(y)|\} \\
&\leq 2c_s(2N)^n \max\{|D^{\mathbf{n}} \phi(2N(x - x_{\mathbf{m}})) - D^{\mathbf{n}} \phi(2N(y - x_{\mathbf{m}}))|, \\
&\qquad |D^{\mathbf{n}} \phi(2N(y - x_{\mathbf{m}'})) - D^{\mathbf{n}} \phi(2N(y - y_{\mathbf{m}'}))|\}.
\end{aligned}$$

This entails

$$\begin{aligned}
\frac{|D^{\mathbf{n}} f_\sigma(x) - D^{\mathbf{n}} f_\sigma(y)|}{\|x - y\|_2^\alpha} &\leq c_s 2(2N)^s \max\left\{ \frac{|D^{\mathbf{n}} \phi(x') - D^{\mathbf{n}} \phi(y')|}{\|x' - y'\|_2^\alpha}, \frac{|D^{\mathbf{n}} \phi(x'') - D^{\mathbf{n}} \phi(y'')|}{\|x'' - y''\|_2^\alpha} \right\} \\
&\leq c_s 2(2N)^s \|\phi\|_{\mathcal{C}^{n,\alpha}} = 1,
\end{aligned}$$

where $x' = 2N(x - x_{\mathbf{m}})$ and $y' = 2N(y - x_{\mathbf{m}})$, and $x'' = 2N(x - x_{\mathbf{m}'})$ and $y'' = 2N(y - x_{\mathbf{m}'})$.

We showed that, simultaneously, $\max_{\mathbf{n}:|\mathbf{n}|\leq n} \|D^{\mathbf{n}} f_\sigma\|_\infty \leq 1$, and $\max_{\mathbf{n}:|\mathbf{n}|=n} \sup_{x\neq y} \frac{|D^{\mathbf{n}} f_\sigma(x) - D^{\mathbf{n}} f_\sigma(y)|}{\|x-y\|_2^\alpha} \leq 1$. We conclude that for all $\sigma \in \{-1,1\}^{N^d}$

$$\|f_\sigma\|_{\mathcal{C}^{n,\alpha}} \leq 1,$$

and therefore $\{f_\sigma : \sigma \in \{-1,1\}^{N^d}\} \subset F_{s,d}$.

Let us now evaluate the distance between distinct elements of $\{f_\sigma : \sigma \in \{-1,1\}^{N^d}\}$. Let $\sigma^1$, $\sigma^2 \in \{-1,1\}^{N^d}$, with $\sigma^1 \neq \sigma^2$, and let $\mathbf{m} \in \{0, \ldots, N-1\}^d$ be such that $\sigma^1_{\mathbf{m}} = -\sigma^2_{\mathbf{m}}$. Let us estimate $\Delta_p$ the $L^p := L^p(\lambda)$ discrepancy between $f_{\sigma^1}$ and $f_{\sigma^2}$ on the cube $C_{\mathbf{m}}$, that is

$$\begin{aligned}
\Delta_p^p &= \int_{C_{\mathbf{m}}} |f_{\sigma^1}(x) - f_{\sigma^2}(x)|^p \mathrm{d}x \\
&= 2^p c_s^p \int_{C_{\mathbf{m}}} |\phi\left(2N(x - x_{\mathbf{m}})\right)|^p \mathrm{d}x \\
&= 2^p c_s^p (2N)^{-d} \|\phi\|_{L^p}^p.
\end{aligned}$$

It remains to find a subset among the functions $f_\sigma$ such that any two functions of this set differ on a significant number of cubes $C_{\mathbf{m}}$. According to the Varshamov-Gilbert Lemma [Yu97], there exists $\Gamma \subset \{-1,1\}^{N^d}$ with cardinal at least $\exp(N^d/8)$ such that for any $\sigma^1, \sigma^2 \in \Gamma$, such that $\sigma^1 \neq \sigma^2$, $\sigma^1$ and $\sigma^2$ differ on at least one fourth of their coordinates; i.e., $\sum_{k=1}^{N^d} \mathbb{1}_{\sigma^1_k \neq \sigma^2_k} \geq \frac{N^d}{4}$. We thus fix such a set $\Gamma \subset \{-1,1\}^{N^d}$. For any $\sigma^1, \sigma^2 \in \Gamma$, with $\sigma^1 \neq \sigma^2$,

$$\begin{aligned}
\|f_{\sigma^1} - f_{\sigma^2}\|_{L^p}^p &= \sum_{\mathbf{m}:\sigma^1_{\mathbf{m}} \neq \sigma^2_{\mathbf{m}}} \int_{C_{\mathbf{m}}} |f_{\sigma^1}(x) - f_{\sigma^2}(x)|^p \mathrm{d}x \\
&\geq \frac{N^d}{4} \Delta_p^p = \frac{2^{p-d} c_s^p}{4} \|\phi\|_{L^p}^p.
\end{aligned}$$

Finally, recalling the definition of $c_s$, we have for any $\sigma^1, \sigma^2 \in \Gamma$, with $\sigma^1 \neq \sigma^2$,

$$\|f_{\sigma^1} - f_{\sigma^2}\|_{L^p} \geq 2^{1 - \frac{d+2}{p}} \frac{1}{2} (2N)^{-s} \|\phi\|_{\mathcal{C}^{n,\alpha}}^{-1} \|\phi\|_{L^p} = cN^{-s},$$

where $c = 2^{-s - \frac{d+2}{p}} \frac{\|\phi\|_{L^p}}{\|\phi\|_{\mathcal{C}^{n,\alpha}}}$.

It follows that $\{f_\sigma \ : \ \sigma \in \Gamma\}$ is a $cN^{-s}$-packing of $F_{s,d}$. Given the lower bound on the size of $\Gamma$, this implies

$$M\left(cN^{-s}, F_{s,d}, \|\cdot\|_{L^p}\right) \geq \exp(N^d/8),$$

for all $n \in \mathbb{N}^*$. Consider now $\varepsilon > 0$, with $\varepsilon \leq c$, and let $N$ be the smallest integer such that $cN^{-s} \geq \varepsilon \geq c(2N)^{-s}$. Since we impose $N \geq 1$, this implies an upper bound $\varepsilon_0 := c$ on $\varepsilon$. On one side, we have

$$M\left(\varepsilon, F_{s,d}, \|\cdot\|_{L^p}\right) \geq M\left(cN^{-s}, F_{s,d}, \|\cdot\|_{L^p}\right),$$

and on the other side, since $2N \geq c^{\frac{1}{s}}\varepsilon^{-\frac{1}{s}}$,

$$\exp(N^d/8) \geq \exp(2^{-d}c^{\frac{d}{s}}\varepsilon^{-\frac{d}{s}}/8).$$

Combining the last three inequalities and setting $c_0 = 2^{-d}c^{\frac{d}{s}}/8$, we finally obtain

$$\log M\left(\varepsilon, F_{s,d}, \|\cdot\|_{L^p}\right) \geq c_0\varepsilon^{-d/s},$$

for all $0 < \varepsilon \leq \varepsilon_0$.

# E Monotonic functions

## E.1 Proof of Proposition 6

### E.1.1 Representing piecewise-constant functions with Heaviside neural networks

We first explain how to represent piecewise-constant functions on cubes with a feed-forward neural network, for some specific architecture and the Heaviside activation function.

**Proposition 10.** *Let $d \in \mathbb{N}^*$, $M \in \mathbb{N}^*$ and suppose that $(\mathcal{C}_n)_{1\leq n\leq M}$ is a partition of $[0,1]^d$ into $M$ cubes. There exists an architecture $\mathcal{A}$ with two-hidden layers, $2(d+1)^2M$ weights and the Heaviside activation function, such that for any $(\alpha_n)_{1\leq n\leq M} \in \mathbb{R}^M$, the function $\tilde{f}\colon [0,1]^d \to [0,1]$ defined by*

$$\forall x \in [0,1]^d, \quad \tilde{f}(x) = \sum_{1\leq i\leq M} \alpha_i \mathbb{1}_{\mathcal{C}_i}(x) \tag{26}$$

*satisfies $\tilde{f} \in H_{\mathcal{A}}$.*

*Proof.* Define $\sigma\colon \mathbb{R} \to \mathbb{R}$ by $\sigma(x) = \mathbb{1}_{x\geq 0}$ for all $x \in \mathbb{R}$.

Let $i \in \{1,\dots,M\}$. The cube $\mathcal{C}_i$ has $2d$ faces, that belong to the cube or not. These faces are supported by hyperplanes whose equations are of the form $\langle \mathbf{w}, x\rangle + b = 0$, with $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Let $J_i$ be the number of faces that belong to $\mathcal{C}_i$. We index the $J_i$ faces that belong to the cube from 1 to $J_i$, and the other faces from $J_i + 1$ to $2d$. Thus,

$$\mathcal{C}_i = \bigcap_{j=1}^{J_i} \{x \in \mathbb{R}^d\colon \left\langle \mathbf{w}_j^i, x\right\rangle + b_j^i \geq 0\} \cap \bigcap_{j=J_i+1}^{2d} \{x \in \mathbb{R}^d\colon \left\langle \mathbf{w}_j^i, x\right\rangle + b_j^i > 0\} \tag{27}$$

with $\mathbf{w}_j^i \in \mathbb{R}^d, b_j^i \in \mathbb{R}$ for all $j \in \{1,\dots,2d\}$. This set can be written as:

$$\left\{ x \in \mathbb{R}^d\colon \sum_{j=1}^{J_i} \mathbb{1}_{\{\left\langle \mathbf{w}_j^i, x\right\rangle + b_j^i \geq 0\}} + \sum_{j=J_i+1}^{2d} \mathbb{1}_{\{\left\langle \mathbf{w}_j^i, x\right\rangle + b_j^i > 0\}} \geq 2d \right\}. \tag{28}$$

Thus, to be on the "good" side of the $j$-th face can be coded by a perceptron (see Figure 1):

$$p_j\colon x \in \mathbb{R}^d \mapsto \begin{cases} \mathbb{1}_{\{\left\langle \mathbf{w}_j^i, x\right\rangle + b_j^i \geq 0\}} & \text{if the } j\text{-th face is included in } \mathcal{C}_i, \\ \mathbb{1}_{\{\left\langle \mathbf{w}_j^i, x\right\rangle + b_j^i > 0\}} & \text{otherwise,} \end{cases} \tag{29}$$

where $j \in \{1,\dots,2d\}$ and the $\mathbf{w}_j^i \in \mathbb{R}^d, b_j^i \in \mathbb{R}$ parametrize the equation of the hyperplane supporting the face. Let us remark that for all $x \in \mathbb{R}$:

$$\mathbb{1}_{x>0} = \mathbb{1}_{-x<0} = 1 - \mathbb{1}_{-x\geq 0} = 1 - \sigma(-x). \tag{30}$$

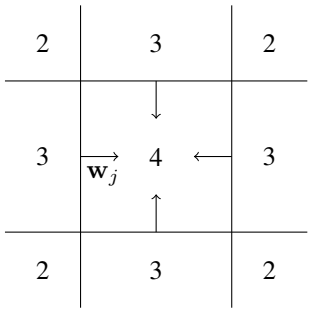


Figure 1: Values of the sum of the perceptrons around a face in dimension 2.

Thus the functions $p_j$ can be expressed with the activation function $\sigma$:

$$\forall j \in \{1, \dots, 2d\}, \forall x \in \mathbb{R}^d, p_j(x) = \begin{cases} \sigma\left(\left\langle \mathbf{w}_j^i, x \right\rangle + b_j^i\right) & \text{if the } j\text{-th face is included in } \mathcal{C}_i, \\ 1 - \sigma\left(-\left\langle \mathbf{w}_j^i, x \right\rangle - b_j^i\right) & \text{otherwise.} \end{cases} \tag{31}$$

Let us return to the expression of $\tilde{f}$. For all $x \in \mathbb{R}^d$,

$$\mathbb{1}_{\mathcal{C}_i}(x) = \begin{cases} 1 & \text{if } \sum_{j=1}^{2d} p_j(x) \geq 2d, \\ 0 & \text{otherwise,} \end{cases} \tag{32}$$

$$= \sigma\left(\sum_{j=1}^{2d} p_j(x) - 2d\right). \tag{33}$$

We deduce that $\tilde{f}$ is of the form, for all $x \in [0,1]^d$:

$$\tilde{f}(x) = \sum_{1 \leq i \leq M} \alpha_i \sigma\left(\sum_{j=1}^{J_i} \sigma\left(\left\langle \mathbf{w}_j^i, x \right\rangle + b_j^i\right) + \sum_{j=J_i+1}^{2d} \left(1 - \sigma\left(-\left\langle \mathbf{w}_j^i, x \right\rangle - b_j^i\right)\right) - 2d\right) \tag{34}$$

$$= \sum_{1 \leq i \leq M} \alpha_i \sigma\left(\sum_{j=1}^{2d} \varepsilon_{i,j} \sigma\left(\left\langle \tilde{\mathbf{w}}_j^i, x \right\rangle + \tilde{b}_j^i\right) - J_i\right), \tag{35}$$

where $\varepsilon_{i,j} = \pm 1$, depending on the $j$-th face is in $\mathcal{C}_i$ or not. This is the action of a Heaviside neural network with two hidden layers (see Figure 2).

It remains to count the weights and biases of $\tilde{f}$ :

- the architecture has $M$ edges going to the output layer, due to the $\alpha_i$;
- it has $M$ biases associated to the neurons of the second hidden layer (they correspond to the terms $-J_i$);
- between the second and the first hidden layer, the architecture has $M \times 2d$ edges (corresponding to the $\varepsilon_{i,j}$);
- it has $M \times 2d$ biases associated to the neurons of the first hidden layer (the $\tilde{b}_j^i$);
- it has $M \times 2d \times d$ edges between the first hidden layer and the entry (the $\tilde{\mathbf{w}}_j^i$).

Thus there are $2M + 2M \times 2d + M \times 2d \times d = 2(d^2 + 2d + 1)M = 2(d+1)^2 M$ weights and biases in total. □

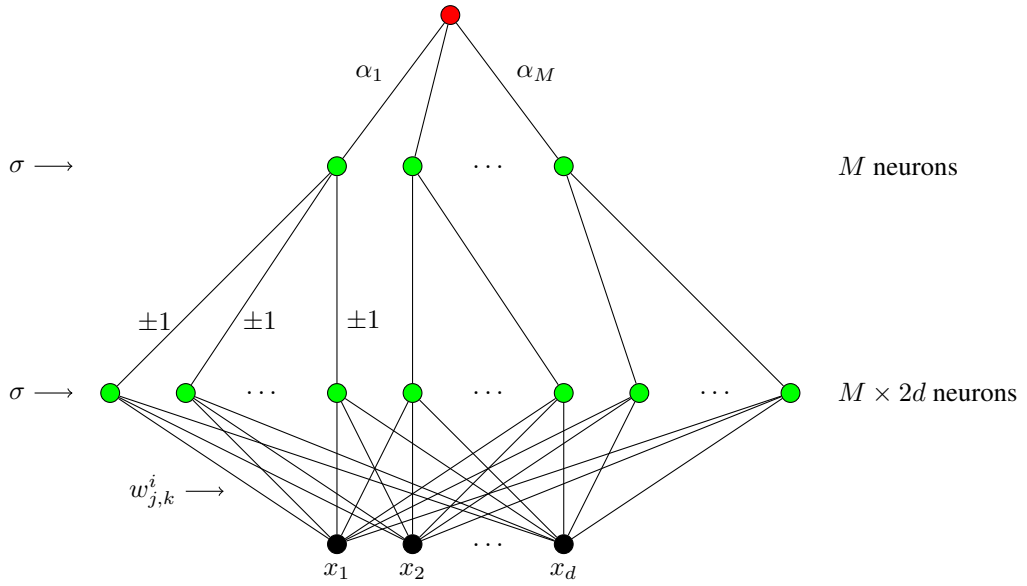


Figure 2: The function $\tilde{f}$ represented as a neural network.

### E.1.2 Main proof

Let $N \in \mathbb{N}$ and $f \in \mathcal{M}^d$. In this section, we partition $[0,1)^d$ into cubes whose sizes depend on the maximal variation of $f$. Then we use this partition to construct a piecewise constant approximation $\tilde{f}$ of $f$; we will bound from above the $L^p(\lambda)$ approximation error $\|f - \tilde{f}\|_{L^p(\lambda)}$ in function of $N$. This part is a direct reinterpretation of the proof of Proposition 3.1 in [GW07]. The result will then follow by an application of Proposition 10.

We first define some notation that will be used in the rest of the section, then we explain the algorithm used to divide $[0,1)^d$ into cubes. We fix the constant $K > 1$ the following way:

$$K := \begin{cases} 2^d & \text{if } p = 1, \\ 2^\beta & \text{otherwise, where } \beta = \frac{1}{2}(d - 1 + \frac{1}{p-1}). \end{cases}$$

We also define an integer $l$ that corresponds to the number of cube decompositions:

$$l := \left\lceil \frac{N \log 2}{\log K} \right\rceil = \begin{cases} \left\lceil \frac{N}{d} \right\rceil & \text{if } p = 1, \\ \left\lceil \frac{N}{\beta} \right\rceil & \text{otherwise .} \end{cases}$$

It is worth noting that this implies $K^{-l} \leq 2^{-N} < K^{-l+1}$.

Now we partition $[0,1)^d$ into dyadic cubes of the form $[a_1, b_1) \times \cdots \times [a_d, b_d)$. If $C$ is such a cube, we use the following convenient notation:

$$\underline{C} := (a_1, \ldots, a_d) \in \mathbb{R}^d, \qquad \overline{C} := (b_1, \ldots, b_d) \in \mathbb{R}^d,$$

to refer to the smallest and largest vertices of $C$. The cube decompositions process reads as follow:

- First we partition $[0,1)^d$ into $2^{Nd}$ cubes of side-length $2^{-N}$. We denote by $S_0$ the set of these cubes $C$ such that $f(\overline{C}) - f(\underline{C}) \leq K 2^{-N}$ and by $R_0$ the set of the remaining cubes.
- For $1 \leq i < l$, we partition each cube in the set $R_{i-1}$ (the remaining cubes at the step $i-1$) into $2^d$ cubes of equal size, and we denote by $S_i$ the set of obtained cubes $C$ of side-length $2^{-(i+N)}$ such that
$$f(\overline{C}) - f(\underline{C}) \leq K^{i+1} 2^{-N}. \tag{36}$$
Again, the set of remaining cubes is denoted by $R_i$.

- Lastly, we partition each cube in the set $R_{l-1}$ into $2^d$ cubes of equal size, and we denote by $S_l$ the set of obtained cubes of side-length $2^{-(l+N)}$.

Once the algorithm is done, each point in $[0,1)^d$ clearly belongs to one single cube of $\cup_{i=0}^{l} S_i$. For $i \in \{0,\dots,l\}$, we let $\tilde{S}_i = \cup_{C\in S_i} C$.

We now define the piecewise constant approximation of $f$ by

$$\forall x \in [0,1]^d, \quad \tilde{f}(x) = \sum_{C\in\bigcup_{0\le i\le l} S_i} f(\underline{C})\mathbb{1}_{x\in C},$$

where $\mathbb{1}_{x\in C}$ denotes the indicator function of the cube $C$. The number of cubes over which $\tilde{f}$ is constant is $\sum_{i=0}^{l}|S_i|$. This quantity is key in Proposition 10; in the next lemma, we bound from above $|S_i|$ for all $i = 0,\dots,l$. Then, we will estimate the error $\|f-\tilde{f}\|_{L^p(\lambda)}$.

**Lemma 7.** *With the above notation:*

$$\forall i \in \{0,\dots,l\}, \quad |S_i| \le dK^{-i}2^{i(d-1)+Nd+1}$$

*Moreover,*

$$\lambda(\tilde{S}_i) \le \begin{cases} 1 & \text{if } i = 0, \\ 2d(2K)^{-i} & \text{, otherwise.} \end{cases} \tag{37}$$

*Proof.* By construction, we have

$$\forall i \in \{1,\dots,l\}, \quad |S_i| + |R_i| = 2^d|R_{i-1}|,$$

since the set $S_i \cup R_i$ contains all the cubes of side-length $2^{-(i+N)}$, that have been constructed from the cubes of $R_{i-1}$. In particular,

$$\forall i \in \{1,\dots,l\}, \quad |S_i| \le 2^d|R_{i-1}|. \tag{38}$$

It remains to bound $|R_{i-1}|$ from above for $i \ge 1$. Define $V := \{\underline{C}\colon C \in R_{i-1}\}$ the set of the smallest vertices of the cubes in $R_{i-1}$. We consider the classes of these vertices under the "laying on the same extended diagonal" equivalence relation. Since the cubes have side-length $2^{-(i-1+N)}$, there are less than $d2^{(i-1+N)(d-1)}$ equivalence classes. According to the pigeonhole principle, the largest class has at least $\left\lceil \frac{|V|}{d2^{(i-1+N)(d-1)}} \right\rceil$ elements; let us refer to this class as $\mathcal{D}$. Let $(C_j)_{1\le j\le J}$ be the set of cubes in $R_{i-1}$ having a point in $\mathcal{D}$ as lowest vertex. Since $f$ is non-decreasing and according to (36), we have:

$$\begin{aligned} 1 \ge f(1,\dots,1) - f(0,\dots,0) &\ge \sum_{j=1}^{J} f(\overline{C_j}) - f(\underline{C_j}) \ge JK^i2^{-N} \\ &\ge \frac{|V|}{d2^{(i-1+N)(d-1)}}K^i2^{-N} = \frac{|R_{i-1}|}{d2^{(i-1+N)(d-1)}}K^i2^{-N}. \end{aligned}$$

Thus

$$|R_{i-1}| \le d2^{i(d-1)+Nd+1-d}K^{-i}.$$

The first statement of Lemma 7 follows from (38).

For $i = 0$, $\lambda(\tilde{S}_0) \le 1$. For $1 \le i \le l$, using the first statement of this lemma, we bound from above the measure of $\tilde{S}_i$:

$$\begin{aligned} \lambda(\tilde{S}_i) = \left(2^{-(i+N)}\right)^d |S_i| &\le dK^{-i}2^{i(d-1)+Nd+1}2^{-d(i+N)}, \\ &= 2d(2K)^{-i}. \end{aligned}$$

□

To show that $\tilde{f}$ is close to $f$ in $L^p(\lambda)$ norm, let us use the fact that $(S_i)_{0\le i\le l}$ is a partition of $[0,1)^d$ and decompose the error in three parts:

$$\|f-\tilde{f}\|^p_{L^p(\lambda)} = \int_{\tilde{S}_0} |f(x)-\tilde{f}(x)|^p\mathrm{d}x + \sum_{i=1}^{l-1}\int_{\tilde{S}_i} |f(x)-\tilde{f}(x)|^p\mathrm{d}x + \int_{\tilde{S}_l} |f(x)-\tilde{f}(x)|^p\mathrm{d}x.$$

In the next lemma, we control each term in the right-hand-side sum above to bound from above $\|f - \tilde{f}\|_{L^p(\lambda)}$ by a function of $N$ that is independent from $f$ and tends towards 0 when $N$ tends to $+\infty$.

**Lemma 8.** *For any $1 \leq p < +\infty$, there exists a constant $c_{d,p} > 0$ depending only on $d$ and $p$ such that*

$$\|f - \tilde{f}\|_{L^p(\lambda)} \leq c_{d,p} \begin{cases} 2^{-N} & \text{if } p(d-1) < d, \\ 2^{-N\frac{(1+1/\beta)}{p}} & \text{if } p(d-1) > d, \\ N^{\frac{1}{p}} 2^{-N} & \text{if } p(d-1) = d. \end{cases} \tag{39}$$

*Proof.* For $0 \leq i < l$, on any cube $C \in S_i$, we have

$$\forall x \in C, \quad |f(x) - \tilde{f}(x)| = |f(x) - f(\underline{C})| \leq f(\overline{C}) - f(\underline{C}) \leq K^{i+1} 2^{-N}, \tag{40}$$

since $f$ is non-decreasing, and by definition of $\tilde{f}$ and $S_i$.

- Using the fact that $\lambda(\tilde{S}_0) \leq 1$ and by (40):
$$\int_{\tilde{S}_0} |f(x) - \tilde{f}(x)|^p \mathrm{d}x \leq (2^{-N}K)^p. \tag{41}$$
- Using (37) and (40), we get for all $i \in \{1, \ldots, l-1\}$
$$\int_{\tilde{S}_i} |f(x) - \tilde{f}(x)|^p \mathrm{d}x \leq (K^{i+1}2^{-N})^p 2d(2K)^{-i}. \tag{42}$$
- On any $C \in S_l$, we have, for all $x \in C$, $|f(x) - \tilde{f}(x)| \leq |f(x) - f(\underline{C})| \leq 1$, and we get, using (37):
$$\int_{\tilde{S}_l} |f(x) - \tilde{f}(x)|^p \mathrm{d}x \leq 2d(2K)^{-l}. \tag{43}$$

Combining (41), (42) and (43) we get:

$$\begin{aligned} \|f - \tilde{f}\|^p_{L^p(\lambda)} &\leq (2^{-N}K)^p + \sum_{i=1}^{l-1} (K^{i+1}2^{-N})^p 2d(2K)^{-i} + 2d(2K)^{-l} \\ &\leq (2^{-N}K)^p + 2^{1-Np} K^p d \sum_{i=1}^{l-1} \left(\frac{K^{p-1}}{2}\right)^i + 2d(2K)^{-l}. \end{aligned} \tag{44}$$

It remains to bound the right-hand side of (44), depending on the value of $p$. Note that the behavior of this term depends on whether $\frac{K^{p-1}}{2}$ is bigger or smaller than 1.

- Suppose that $p(d-1) < d$. If $p = 1$, we have $\frac{K^{p-1}}{2} = \frac{1}{2} < 1$. If $p > 1$, we have:
$$(d-1)p < d \iff dp - p - d < 0 \iff d(p-1) < p \iff \frac{d}{p} < \frac{1}{p-1}.$$
Since we assumed that $(d-1)p < d$, we have $d - 1 < \frac{1}{p-1}$. Thus, $\beta$ being the arithmetic mean of $d-1$ and $\frac{1}{p-1}$, we have $d - 1 < \beta < \frac{1}{p-1}$. Then $K = 2^\beta < 2^{1/(p-1)}$ and hence $\frac{K^{p-1}}{2} < 1$ and $\frac{1}{2K} < K^{-p}$. Therefore
$$\sum_{i=1}^{l-1} \left(\frac{K^{p-1}}{2}\right)^i \leq \frac{K^{p-1}}{2 - K^{p-1}} \quad \text{and} \quad (2K)^{-l} \leq K^{-pl}.$$
When $p \geq 1$ and since $K^{-l} \leq 2^{-N}$, this leads to
$$\begin{aligned} \|f - \tilde{f}\|^p_{L^p(\lambda)} &\leq (2^{-N}K)^p + 2^{1-Np} K^p d \frac{K^{p-1}}{2 - K^{p-1}} + 2dK^{-pl} \\ &\leq \left(K^p + 2K^p d \frac{K^{p-1}}{2 - K^{p-1}} + 2d\right) 2^{-Np}. \end{aligned}$$

We thus have, setting $c_1 := \left(K^p + 2K^p d\frac{K^{p-1}}{2-K^{p-1}} + 2d\right)^{\frac{1}{p}}$,

$$\|f - \tilde{f}\|_{L^p(\lambda)} \leq c_1 2^{-N}.$$

- Suppose that $p(d-1) > d$. We have $d - 1 > \beta > \frac{1}{p-1}$. Then $K = 2^\beta > 2^{1/(p-1)}$ and hence $\frac{K^{p-1}}{2} > 1$, which entails

$$\begin{aligned}\|f - \tilde{f}\|^p_{L^p(\lambda)} &\leq (2^{-N}K)^p + 2^{1-Np}K^p d\frac{(K^{p-1}/2)^l}{K^{p-1}/2-1} + 2d(2K)^{-l} \\ &\leq 2^{-Np}K^p + 2^{-Np}K^{pl}\frac{2K^p d}{K^{p-1}/2-1}(2K)^{-l} + 2d(2K)^{-l}.\end{aligned}$$

  Since $p > 1 + \frac{1}{\beta}$, we have $2^{-Np} \leq 2^{-N(1+\frac{1}{\beta})}$. Also, since $K = 2^\beta$, $(2K)^{-l} = 2^{-l(\beta+1)}$, and since $l \geq \frac{N\log(2)}{\log(K)} = \frac{N}{\beta}$, we have $(2K)^{-l} \leq 2^{-\frac{N}{\beta}(\beta+1)} = 2^{-N(1+\frac{1}{\beta})}$. Finally, since $2^{-N}K^l < K$,

$$\|f - \tilde{f}\|^p_{L^p(\lambda)} \leq \left(K^p + K^p\frac{2K^p d}{K^{p-1}/2-1} + 2d\right)2^{-N(1+1/\beta)}$$

  We thus have, setting $c_2 := \left(K^p + \frac{2K^{2p}d}{K^{p-1}/2-1} + 2d\right)^{\frac{1}{p}}$,

$$\|f - \tilde{f}\|_{L^p(\lambda)} \leq c_2 2^{-\frac{N(1+1/\beta)}{p}}.$$

- Suppose that $p(d-1) = d$. It implies that $p - 1 = \frac{1}{d-1}$, then $\beta = d - 1$. We thus have $K^{p-1} = 2^{(d-1)(p-1)} = 2$. Therefore,

$$\|f - \tilde{f}\|^p_{L^p(\lambda)} \leq 2^{-Np}K^p + 2^{-Np}2K^p d(l-1) + 2d(K^p)^{-l}.$$

  On the one hand, we have $K^{-l} \leq 2^{-N}$. On the other, we have $2^{-N} < K^{-l+1}$, so $l - 1 < N\frac{\log 2}{\log K} = \frac{N}{d-1}$. Putting it all together, we get

$$\begin{aligned}\|f - \tilde{f}\|^p_{L^p(\lambda)} &\leq 2^{-Np}K^p + 2^{-Np}2K^p d(l-1) + 2d2^{-Np} \\ &\leq \left(K^p + 2K^p\frac{d}{d-1} + 2d\right)N2^{-Np}.\end{aligned}$$

  We thus have, setting $c_3 := \left(K^p + 2K^p\frac{d}{d-1} + 2d\right)^{\frac{1}{p}}$,

$$\|f - \tilde{f}\|_{L^p(\lambda)} \leq c_3 N^{\frac{1}{p}} 2^{-N}.$$

Finally, $c_1$, $c_2$ and $c_3$, only depend on $p$ and $d$. Hence, letting $c_{d,p} = \max\{c_1, c_2, c_3\}$ yield the result. □

According to Proposition 10, the function $\tilde{f}$ can be implemented by a Heaviside neural network with two hidden layers and $W = 2(d+1)^2\sum_{i=0}^{l}|S_i|$ weights. Using Lemma 7, we obtain

$$\begin{aligned}W = 2(d+1)^2\sum_{i=0}^{l}|S_i| &\leq 2(d+1)^2\sum_{i=0}^{l} dK^{-i}2^{i(d-1)+Nd+1} \\ &\leq 2^{Nd+2}d(d+1)^2\sum_{i=0}^{l}\left(\frac{2^{d-1}}{K}\right)^i.\end{aligned}$$

We let

$$W_N := 2^{Nd+2}d(d+1)^2\sum_{i=0}^{l}\left(\frac{2^{d-1}}{K}\right)^i.$$

Although we do not make the dependence explicit, $W_N$ also depends on $d$ and $p$. We have $\lim_{N\to+\infty} W_N = +\infty$.

**Lemma 9.** *With the above notation: For any $+\infty > p \geq 1$, there exist constants $W'_{\min}, c'_{d,p} > 0$ depending only on $d$ and $p \geq 1$ such that for all $N$ satisfying $W_N \geq W'_{\min}$*

$$\|f - \tilde{f}\|_{L^p(\lambda)} \leq c'_{d,p}\, g(W_{N+1})$$

*where, for all $W \geq 1$,*

$$g(W) = \begin{cases} W^{-1/d} & \text{if } (d-1)p < d, \\ W^{-\frac{1}{p(d-1)}} & \text{if } (d-1)p > d, \\ W^{-1/d} \log W & \text{if } (d-1)p = d. \end{cases}$$

*Proof.* Again, we separate the cases in function of $p$.

- Suppose that $p(d-1) < d$: if $p = 1$, $\frac{2^{d-1}}{K} = \frac{1}{2} < 1$; if $p > 1$, $\beta > d-1$ and $\frac{2^{d-1}}{K} = 2^{d-1-\beta} < 1$. Thus, for all $N \geq 1$,

  $$W_N \leq 2^{Nd} \left( \frac{4d(d+1)^2}{1 - 2^{d-1-\beta}} \right) =: 2^{Nd} c''_{d,p}.$$

  Writing the inequality for $N+1$, we obtain

  $$W_{N+1} \leq 2^{Nd} 2^d c''_{d,p}\,.$$

  That is: $2^{-N} \leq 2 \left( \frac{c''_{d,p}}{W_{N+1}} \right)^{1/d}$. Combined with (39), this provides

  $$\|f - \tilde{f}\|_{L^p(\lambda)} \leq 2c_{d,p} \left( \frac{c''_{d,p}}{W_{N+1}} \right)^{1/d} = c^1_{d,p} W_{N+1}^{-1/d},$$

  for $c^1_{d,p} = 2c_{d,p}(c''_{d,p})^{1/d}$.

- If $p(d-1) > d$, then $\beta < d-1$ and $\frac{2^{d-1}}{K} = 2^{d-1-\beta} > 1$. Thus,

  $$\begin{aligned} W_N &\leq 2^{Nd} 2^{(d-1-\beta)(l+1)} \left( \frac{4d(d+1)^2}{2^{d-1-\beta} - 1} \right) \leq 2^{Nd} 2^{(d-1-\beta)(N/\beta+2)} \left( \frac{4d(d+1)^2}{2^{d-1-\beta} - 1} \right) \\ &= 2^{N(d+(d-1)/\beta-1)} \left( \frac{4d(d+1)^2 2^{2(d-1-\beta)}}{2^{d-1-\beta} - 1} \right) =: 2^{N(1+\frac{1}{\beta})(d-1)} c''_{d,p}, \end{aligned}$$

  for a different constant $c''_{d,p}$. Writing again this inequality for $N+1$, we obtain

  $$W_{N+1} \leq c''_{d,p} 2^{(1+\frac{1}{\beta})(d-1)}\, 2^{N(1+\frac{1}{\beta})(d-1)},$$

  which we can write $2^{-N(1+\frac{1}{\beta})} \leq 2^{(1+\frac{1}{\beta})} \left( \frac{c''_{d,p}}{W_{N+1}} \right)^{\frac{1}{d-1}}$. And thus using (39),

  $$\|f - \tilde{f}\|_{L^p(\lambda)} \leq 2^{(1+\frac{1}{\beta})/p} c_{d,p} \left( \frac{c''_{d,p}}{W_{N+1}} \right)^{\frac{1}{p(d-1)}} = c^2_{d,p} W_{N+1}^{-\frac{1}{p(d-1)}},$$

  for $c^2_{d,p} = c_{d,p}\, 2^{(1+\frac{1}{\beta})/p}\, (c''_{d,p})^{\frac{1}{p(d-1)}}$.

- If $p(d-1) = d$, $\beta = d-1$ and $\frac{2^{d-1}}{K} = 1$. Thus,

  $$\begin{aligned} W_N &= 2^{Nd+2} d(d+1)^2 (l+1) \leq 2^{Nd+2} d(d+1)^2 \left( \frac{N}{\beta} + 2 \right) \\ &= 2^{Nd} \left( \frac{N}{\beta} + 2 \right) \left( 4d(d+1)^2 \right) =: 2^{Nd} \left( \frac{N}{d-1} + 2 \right) c''_{d,p} \\ &\leq 2^{d(d-1)\left(\frac{N}{d-1}+2\right)} \left( \frac{N}{d-1} + 2 \right) c''_{d,p} \\ &= \exp\left( d(d-1) \left( \frac{N}{d-1} + 2 \right) \log 2 \right) \left( \frac{N}{d-1} + 2 \right) c''_{d,p}. \end{aligned} \tag{45}$$

Setting:

$$\tilde{W}_N := \frac{d(d-1)W_N \log 2}{c''_{d,p}} \qquad \tilde{N} := d(d-1)\left(\frac{N}{d-1}+2\right)\log 2,$$

we can rewrite (45) as:

$$\tilde{W}_N \leq \tilde{N}\exp(\tilde{N})\,. \tag{46}$$

Since $d \geq 2$, there exists $W'_{min}$ such that, when such that $W_N \geq W'_{min}$ we have $\log(\tilde{W}_N) > 1$. We show by contradiction that (46) implies that

$$\tilde{N} \geq \log \tilde{W}_N - \log\log \tilde{W}_N,$$

which in turn gives

$$\begin{aligned} N &\geq \left(\frac{\log \tilde{W}_N - \log\log \tilde{W}_N}{d(d-1)\log 2} - 2\right)(d-1) \\ &\geq \frac{\log(\tilde{W}_N)}{d\log(2)} - \frac{\log\log \tilde{W}_N}{d\log(2)} + c, \end{aligned}$$

for an appropriate constant $c$. Since $(W_N)_{N\in\mathbb{N}}$ is non-decreasing, $W_{N+1} \geq W'_{min}$ and the inequality also holds for $N+1$. That is

$$N+1 \geq \frac{\log(\tilde{W}_{N+1})}{d\log(2)} - \frac{\log\log \tilde{W}_{N+1}}{d\log(2)} + c \tag{47}$$

Using (39), we obtain:

$$\|f-\tilde{f}\|_{L^p(\lambda)} \leq c_{d,p} N^{\frac{1}{p}} 2^{-N} \leq 2c_{d,p}(N+1)^{\frac{1}{p}} 2^{-(N+1)},$$

and since the function $t \longmapsto t^{\frac{1}{p}}2^{-t}$ is non-increasing for $t$ sufficiently large, up to a modification of $W'_{min}$, using (47), we obtain

$$\begin{aligned} \|f-\tilde{f}\|_{L^p(\lambda)} &\leq 2^{c+1} c_{d,p}\left(\frac{\log(W_{N+1})}{d\log(2)}\right)^{\frac{1}{p}} 2^{-\frac{\log(W_{N+1})}{d\log(2)}} 2^{\frac{\log\log W_{N+1}}{d\log(2)}}, \\ &\leq c^3_{d,p} W_{N+1}^{-\frac{1}{d}} (\log W_{N+1})^{\frac{1}{p}+\frac{1}{d}} = c^3_{d,p} W_{N+1}^{-\frac{1}{d}} \log W_{N+1}, \end{aligned}$$

for an appropriate constant $c^3_{d,p}$.

Taking $c'_{d,p} = \max(c^1_{d,p}, c^2_{d,p}, c^3_{d,p})$ provides the announced statement. □

**Proof of Proposition 6.** Take $W_{\min} = \max(W'_{\min}, W_0)$ and $c = c'_{d,p}$, where $W'_{\min}$ and $c'_{d,p}$ are from Lemma 9. Let $W \geq W_{\min}$, there exists $N$ such that

$$W_N \leq W < W_{N+1}.$$

Consider the architecture $\mathcal{A}$ with $W$ weights, as in Proposition 10, which allows to represent piecewise-constant functions with $\frac{W}{2(d+1)^2}$ cubic pieces, therefore it can also represent piecewise-constant functions with $\frac{W_N}{2(d+1)^2}$ pieces.

For any $f \in \mathcal{M}^d$, the function $\tilde{f}$ obtained for the parameter $N$ is a piecewise-constant function with at most $\frac{W_N}{2(d+1)^2}$ pieces, therefore we have $\tilde{f} \in H_{\mathcal{A}}$ and, according to Lemma 9, $\tilde{f}$ satisfies

$$\|f-\tilde{f}\|_{L^p(\lambda)} \leq c'_{d,p} g(W_{N+1}).$$

Moreover, since $g$ is non-increasing, we have

$$\|f-\tilde{f}\|_{L^p(\lambda)} \leq c'_{d,p} g(W).$$

Therefore, for any $f \in \mathcal{M}^d$,

$$\inf_{g\in H_{\mathcal{A}}} \|f-g\|_{L^p(\lambda)} \leq c'_{d,p} g(W)$$

and so does the supremum over $f$ in $\mathcal{M}^d$.

This concludes the proof of Proposition 6.

### E.2 Proof of Proposition 4

We first need some topological results.

#### E.2.1 Some useful topological results

In the following lemma, for any topological set $X$ and any subset $S \subset X$, we say that a point $x \in S$ is an *isolated point* of $S$ if there exists a neighborhood of $x$ that does not contain any other point of $S$.

**Lemma 10.** *Let $F$ be a closed set. Then an isolated point of $\partial F$ is an isolated point of $F$.*

*Proof.* Since $F$ is closed, $\partial F = F \setminus \mathring{F}$. Let $x$ be an isolated point of $\partial F$. Then there exists $r > 0$ such that $B(x, r) \cap \partial F = \{x\}$. Denote by $\tilde{B} := B(x, r) \setminus \{x\}$ the open ball without its center. Then $\tilde{B} \cap F = \tilde{B} \cap \mathring{F}$ since $\tilde{B} \cap \partial F = \emptyset$ and $\mathring{F} \subset F$. Observe that $\tilde{B} \cap F$ is a closed set of $\tilde{B}$ and $\tilde{B} \cap \mathring{F}$ is an open set of $\tilde{B}$. Since $\tilde{B}$ is connected, either $\tilde{B} \cap F = \emptyset$ or $\tilde{B} \cap F = \tilde{B}$. The latter cannot be true, for otherwise we would have $B(x, r) \subset F$, contradicting $x \in \partial F$. Hence $\tilde{B} \cap F = \emptyset$, i.e., $x$ is isolated in $F$. □

**Proposition 11.** *Let $\mathcal{C}$ be a compact, convex set which is not a singleton. Then $\partial\mathcal{C}$ does not contain any isolated point.*

*Proof.* We use a proof by contradiction. If $\partial\mathcal{C}$ contains an isolated point $x$, then $x$ is also an isolated point of $\mathcal{C}$ by Lemma 10 since $\mathcal{C}$ is closed. Let $y \in \mathcal{C}$. We have $[x, y] \subset \mathcal{C}$ by convexity, hence $y = x$ (if not, $x$ would not be isolated). Thus we have $\mathcal{C} = \{x\}$, which contradicts the fact that $\mathcal{C}$ is not a singleton. □

#### E.2.2 Proof of Proposition 4

**Step 1: we prove the result in dimension $d = 2$.** We define

$$\mathcal{C} = \left\{ x \in \mathbb{R}^2 : \sum_{i=1}^{2} (x_i - 1)^2 \leq 1 \right\} ,$$

whose intersection with $[0, 1]^2$ is displayed on Figure 3. We denote by $f : [0, 1]^2 \to \{0, 1\}$ the indicator function of the set $\mathcal{C} \cap [0, 1]^2$.

Note that $\mathcal{C}$ satisfies the hypothesis of Proposition 11. Since in addition no point in $\mathcal{C}^c \cap [0, 1]^2$ has all its coordinates strictly larger than those of a point in $\mathcal{C}$, we can see that $f$ lies in $\mathcal{M}^2$ (monotonic functions of 2 variables). Let $g \in H_{\mathcal{A}}$. The idea of the proof is to show that $\partial\mathcal{C}$ intersects the "discontinuity set" (or jump set) of $g$ in only a finite number of points. Since $f$ takes its values in $\{0, 1\}$, this implies that there are some points $x$ where either $f(x) = 0$ and $g(x) \geq \frac{1}{2}$ or $f(x) = 1$ and $g(x) \leq \frac{1}{2}$. We now give more details.

We show that $\|f - g\|_\infty \geq \frac{1}{2}$ by contradiction. Let $W \geq 1$ be the number of weights in the architecture $\mathcal{A}$. First note that $g = \sum_{j=1}^{K} \alpha_j \mathbb{1}_{A_j}$, where $K \leq 2^W$, and the $A_j$ are convex polytopes forming a partition of $[0, 1]^2$. Suppose that $\|f - g\|_\infty < \frac{1}{2}$ for a moment. This implies that $g > \frac{1}{2}$ on $\mathcal{C}$, and $g < \frac{1}{2}$ elsewhere.

Observe that under this assumption, $\partial\mathcal{C} \cap (0, 1)^2$ is included in the finite union $\bigcup_{1 \leq j \leq K} \partial A_j$. Indeed, if it were not true, then there would exist $x \in \partial\mathcal{C} \cap (0, 1)^2$ and $j \in \{1, \dots, K\}$ such that $x \in \mathring{A}_j$. Let $\epsilon > 0$ such that $B(x, \epsilon) \subset \mathring{A}_j$. We have $B(x, \epsilon) \not\subset \mathcal{C}$ (otherwise, the open ball would be included in the interior of $\mathcal{C}$ and $x$ would not lie on the boundary). Thus there exists $z \in B(x, \epsilon) \setminus \mathcal{C}$, which satisfies $g(z) < \frac{1}{2} < g(x)$ (recall that $g > \frac{1}{2}$ on $\mathcal{C}$, and $g < \frac{1}{2}$ elsewhere), which is not possible since $x, z \in \mathring{A}_j$ and $g$ is constant on $A_j$. This proves that $\partial\mathcal{C} \cap (0, 1)^2$ is included in the finite union $\bigcup_{1 \leq j \leq K} \partial A_j$.

Since the $A_j$ are polygons (recall that we work in dimension 2), their boundaries are finite unions of line segments. Then $\partial\mathcal{C} \cap (0, 1)^2$ is included in a finite union of line segments. Let us show that it is not possible, contradicting the assumption that $\|f - g\|_\infty < \frac{1}{2}$.

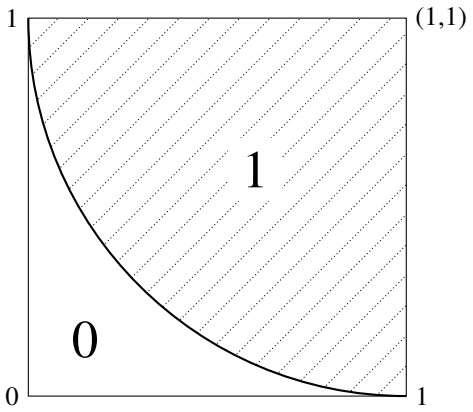


Figure 3: The set $\mathcal{C}$ and its indicator function $f$.

Let us prove that the intersection of a line segment with $\partial\mathcal{C} \cap (0,1)^2$ contains at most 2 points. Denote by $S$ a closed line segment: $\mathcal{C}$ and $S$ are convex and hence connected, thus $\mathcal{C} \cap S$ is either empty, a singleton or a line segment, as a connected compact set of the line supporting $S$. If it is empty, then *a fortiori*, $\partial\mathcal{C} \cap S = \emptyset$. If it is not, denote by $x$ and $y$ its extremities (assuming $x = y$ in the case of a singleton). By strict convexity, the open line segment $(x, y)$ is included in $\mathring{\mathcal{C}}$, hence $\partial\mathcal{C} \cap S \subset \{x, y\}$. In any case, we have $|\partial\mathcal{C} \cap S| \leq 2$.

Let $M$ be the finite number of line segments forming the boundaries of the $A_j$. If $\partial\mathcal{C} \cap (0,1)^2$ is included in the union of these line segments, it contains at most $2M$ points. Since $\partial\mathcal{C} \cap (0,1)^2$ is nonempty and finite, all its points are isolated points, thus $\partial\mathcal{C}$ contains at least one isolated point. This would contradict Proposition 11: since $\partial\mathcal{C}$ is the boundary of a compact, convex set which is not a single singleton, it can not have any isolated point. This leads to a contradiction, and proves the result in dimension 2.

**Step 2: we prove the result in any dimension $d \geq 2$,** by a reduction to dimension 2. We define

$$\mathcal{C} = \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{d} (x_i - 1)^2 \leq 1 \right\} ,$$

and the function $f : [0,1]^d \to \mathbb{R}$ by

$$f(x_1, \ldots, x_d) = \mathbb{1}_{(x_1,\ldots,x_d)\in\mathcal{C}} .$$

Let $g : [0,1]^d \to \mathbb{R}$ be any function in $H_{\mathcal{A}}$, that is, $g$ can be represented by a a Heaviside neural network with $d$ input neurons. Note that

$$\begin{aligned} &\sup_{x_1,x_2,x_3\ldots,x_d\in[0,1]} |f(x_1, x_2, x_3 \ldots, x_d) - g(x_1, x_2, x_3 \ldots, x_d)| \\ &\qquad \geq \sup_{x_1,x_2\in[0,1]} |f(x_1, x_2, 1 \ldots, 1) - g(x_1, x_2, 1 \ldots, 1)| \\ &\qquad \geq \frac{1}{2} , \end{aligned}$$

where the last inequality is by the result of Step 1, since $(x_1, x_2) \in [0,1]^2 \mapsto f(x_1, x_2, 1 \ldots, 1)$ is the indicator function of Step 1, and $(x_1, x_2) \in [0,1]^2 \mapsto g(x_1, x_2, 1 \ldots, 1)$ can be represented by a Heaviside neural network with 2 input neurons. This concludes the proof.

**Remark.** Note from the above proof that, though we only stated the impossibility result for piecewise-constant activation functions, it in fact holds more generally for piecewise-affine activation functions.

# F Barron space

In Section 5 we mentioned that the Barron space introduced in [Bar93] is one among several examples for which approximation theory provides ready-to-use lower bounds on the packing number.

This space has received renewed attention recently in the deep learning community, in particular because its "size" is sufficiently small to avoid approximation rates depending exponentially on the input dimension $d$. Next we detail how to apply Corollary 1 in this case.

**Definition of the Barron space.** We start by introducing the Barron space, as defined in [PV21]. Let $d \in \mathbb{N}^*$. For any constant $C > 0$, the Barron space $B_d(C)$ is the set of all functions $f$ for which there exists a measurable function $F : \mathbb{R}^d \to \mathbb{C}$ and some $c \in [-C, C]$ satisfying, for all $x \in [0,1]^d$,

$$f(x) = \int_{\mathbb{R}^d} (e^{ix\cdot\xi} - 1)F(\xi)\mathrm{d}\xi \qquad \text{and} \qquad \int_{\mathbb{R}^d} \|\xi\|_2 |F(\xi)|\mathrm{d}\xi \leq C,$$

where $x \cdot \xi$ denotes the standard scalar product in between $x$ and $\xi$.

**Known lower bound on the packing number.** Petersen and Voigtlaender [PV21] showed a tight lower bound on the log packing number in $L^p(\lambda, [0,1]^d)$ norm, which we recall below.

**Proposition 12** (Proposition 4.6 in [PV21])**.** *Let $1 \leq p \leq +\infty$. There exist constants $\varepsilon_0, c_0 > 0$ depending only on $d$ and $C$ such that for any $\varepsilon \leq \varepsilon_0$,*

$$\log M(\varepsilon, B_d(C), \|\cdot\|_{L^p}) \geq c_0 \varepsilon^{-1/(\frac{1}{2}+\frac{1}{d})}. \tag{48}$$

**Consequence on the approximation rate by piecewise-polynomial neural networks.** Plugging the lower bound of Proposition 12 in Corollary 1, we obtain the following lower bound on the approximation error of the Barron space by piecewise-polynomial neural networks.

**Proposition 13.** *Let $1 \leq p < +\infty$, $d \geq 1$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a piecewise-polynomial function on $K \geq 2$ pieces, with maximal degree $\nu \in \mathbb{N}$. Consider the Barron space $B_d(C)$ defined above, with $C > 0$. There exist positive constants $c_1, c_2, c_3, W_{\min}$ depending only on $d$, $p$, $C$, and $\sigma$ such that for any architecture $\mathcal{A}$ of depth $L \geq 1$, $W \geq W_{\min}$ variable weights, the set $H_{\mathcal{A}}$ satisfies*

$$\sup_{f \in B_d(C)} \inf_{g \in H_{\mathcal{A}}} \|f - g\|_{L^p(\lambda)} \geq \begin{cases} c_1 W^{-1-\frac{2}{d}} \log^{-1-\frac{2}{d}}(W) & \text{if } \nu \geq 2\,, \\ c_2 (LW)^{-\frac{1}{2}-\frac{1}{d}} \log^{-\frac{3}{2}-\frac{3}{d}}(W) & \text{if } \nu = 1\,, \\ c_3 W^{-\frac{1}{2}-\frac{1}{d}} \log^{-\frac{3}{2}-\frac{3}{d}}(W) & \text{if } \nu = 0\,. \end{cases} \tag{49}$$