Architectural properties of neural networks for function approximation

by

Luca Venturi

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

New York University

September 2021

Joan Bruna

Afonso S. Bandeira

Dedication

Ai miei genitori, Franco e Lucia, e ai miei fratelli, Anna e Marco

Acknowledgements

This thesis would not be anywhere if it was not for the support and input from a numerous people. I would first like to thank those who I missed mentioning below, as there are likely some.

I would like to thank Joan Bruna for having guided me through this journey. Your enthusiasm for the subject and endless energy, mathematical creativity and ability to provide insights between different areas have been of great motivation for me. I am extremely indebted to your constant support.

I am profoundly grateful to my co-advisor Afonso Bandeira. Your mathematical talent and ability to distill simple ideas from seemingly complicated arguments have been a great source of inspiration for me. Thank you for your availability for listening to whatever I was working on and for the precious advice.

My sincere thanks goes to Benjamin Peherstorfer, for taking me on different projects, and keeping me interested in numerical analysis and reduced order models. I learned a lot from you on how to conduct research in a clear and organized way.

I would also like to thank Georg Stadler, for first introducing me to the program and helping me navigate through my first year at Courant.

Thanks to all my other collaborators during these years, Alberto Bietti, Dan Kushnir, Donsub Rim, Samy Jelassi, Terrence Alsup. Most of the projects mentioned in this thesis would not have been completed without you and I have learned a lot from you all.

Thanks to Eric Vanden-Ejden and Mahdi Soltanolkotabi, for being a part of the thesis committee.

Thanks to my advisors during the master and bachelor times, Gianluigi Rozza and Eugenio Regazzini, who taught me a lot, and without whom I would not have been here otherwise.

I wish to thank my cohort and all the other people I had the fortune to know during my mathematical life, both at and outside of NYU.

Thanks to all the people in the ‘Math of Deep Learning’ meetings, it has been a pleasure to interact with and learning from you all.
Thanks to IPAM, for giving me the opportunity to stay for three months in Los Angeles and participate to a great program. Also thanks to Lukas for sharing many surfing tentative.
Thanks to all the people who made my stay at Courant very enjoyable: Matteo, Jordan, Cem, Alec, and many others. Thanks to Francesco, for the many fun moments and conversations. A mention of honor goes to Juma, for being an incredible person and friend, whose love for math problems solving kept my primordial love for this subject alive during the last five years. Thanks to Rosario, who was always there to greet me when I was coming to or leaving from Courant every day.
Grazie a Giulia, per le chiamate su skype and per aver condiviso gli ‘up-and-downs’ di questo percorso fin dall’inizio.
Grazie ai Rapici: Dado, Lollo, Otto. Probabilmente non sarei finito a New York se prima non avessi vissuto con voi. Mi avete insegnato quanto matematica e ignoranza possano andare a bracetto, e l’arte accademica del procastinare.

There has many other people who I need to thank, whom I would not have gone anywhere without.
Grazie ai miei amici di sempre, Alessandro, Andrea, Antonio, Edoardo, Ludovico, per tutti i momenti felici vissuti assieme.
Un grazie di cuore a tutta la mia famiglia, per avermi sempre dato fiducia e supportato. Grazie ai miei genitori, per l’amore e l’educazione che mi avete donato. Grazie ai miei fratelli, per il legame che ci unisce. Grazie ai miei nonni, che non ce l’hanno fatta a vedermi arrivare fin qua.

At the end of my first year I had the fortune of meeting an incredible woman here at Courant, with whom I have been sharing my life since then. I am grateful to Sonica, for her love and constant support, for the smiles when I needed them the most, for teaching me so many things and for always providing me with a diverse perspective. Plus, my reputation would have been much worse

without her help in composing emails and messages.

Abstract

Deep neural networks have emerged, in recent years, as incredibly powerful models in machine learning. Despite this, theoretical understanding of neural networks are lacking, or unsatisfying. In this thesis, we consider two problems in the theory of neural networks, focusing on function approximation via feed-forward neural networks. The first part of this thesis deals with understanding the approximation capacities of neural networks in terms of their depth. We first discuss this problem in the low-dimensional case, focusing on a specific class of functions, namely solutions to transport problems. We then move to the high-dimensional setting, a regime more relevant to machine learning. We look at neural networks in the frequency domain, and we offer explanations on why and when depth is essential to computational efficiency. In the second part of the thesis we look at the problem of optimization, and we discuss the optimization landscape for shallow neural networks, focusing on distribution-free results on the existence of spurious minima regions. We finish with some concluding remarks and discussing a few open questions.

Contents

List of Figures

List of Appendices

List of Notations

w.l.o.g.	without loss of generality
w.r.t.	with respect to
i.i.d.	independent identically distributed
r.v.	random variable
PDE	partial differential equation
SVD	singular value decomposition
RKHS	representation kernel Hilbert space
UAT	universal approximation theorem
ERM	empirical risk minimization
$[\![n, m]\!]$	$\{n, n+1, \ldots, m\}$
$[n]$	$[\![1, n]\!]$
x	scalar variables are denoted as lowercase non-bold
$\mathbf{x}$	vector variables are denoted as lowercase bold
x_k	k-th entry of the vector $\mathbf{x}$

X	univariate random variables are denoted as uppercase non-bold
$\mathbf{X}$	tensor, matrix, and multivariate random variables are denoted as uppercase bold
$X_{k_1 \cdots k_m}$	entry $(k_1, \ldots, k_m)$ of the tensor $\mathbf{X}$
X_k	k-th entry of the random variable $\mathbf{X}$
$\mathbf{x}_k$	k-th column of the matrix $\mathbf{X}$
$\mathbf{X}_1 \circ \mathbf{X}_2$	outer product of tensor
$\mathbf{X}^{\circ k}$	outer product of the tensor $\mathbf{X}$ k-times with itself
$\mathbb{F}$	$\mathbb{R}$ or $\mathbb{C}$
$C(\Omega)$	the set of continuous functions $f : \Omega \to \mathbb{F}$
$C_c(\Omega)$	the set of continuous functions $f : \Omega \to \mathbb{F}$ with compact support
$C^k(\Omega)$	the set of C^k functions $f : \Omega \to \mathbb{F}$
$L^p(\Omega)$	the set of p-integrable functions $f : \Omega \to \mathbb{F}$ w.r.t. the Lebesgue measure
$L^p(\mu)$	the set of p-integrable functions $f : \Omega \to \mathbb{F}$ w.r.t. the measure μ
$L^p(\varphi)$	the set of p-integrable functions $f : \Omega \to \mathbb{F}$ w.r.t. the density φ (w.r.t. the Lebesgue measure)
$L^p(\mathbf{X})$	the set of p-integrable functions $f : \Omega \to \mathbb{F}$ w.r.t. the probability measure of the random variable $\mathbf{X}$
$\mathrm{TV}([a, b])$	the set of functions $f : [a, b] \to \mathbb{F}$ with finite total variation
$\|\mathbf{x}\|_p$	norm p of the vector $\mathbf{x}$

$\|f\|_p$	norm p of the function f, when the measure is clear from the context
$\|f\|_{\eta,p}$	norm p of the function $f \in L^p(\eta)$, for $\eta \in \{\Omega, \mu, \varphi, \mathbf{X}\}$
$\|f\|_V$	norm of the function f in the normed space V
$\langle f, g \rangle_\eta$	scalar product between the function $f, g \in L^2(\eta)$, for $\eta \in \{\Omega, \mu, \varphi, \mathbf{X}\}$
$\|\mu\|_1$	total variation of the finite signed Borel measure μ
$\|\mathbf{X}\|_{F,p}$	entrywise norm p of the tensor $\mathbf{X}$
$\|\mathbf{X}\|_{p,q}$	(p, q) operator norm of the matrix $\mathbf{X}$, that is $\max_{\mathbf{y} \,:\, \|\mathbf{y}\|_p=1} \|\mathbf{X}\mathbf{y}\|$
$\|\mathbf{X}\|_p$	p operator norm of the matrix $\mathbf{X}$, that is $\|\mathbf{X}\|_{p,p}$
$\mathbb{E}\mathbf{X}$	expectation of the r.v. $\mathbf{X}$
$\mathbf{X} \sim \mu$	denotes that the r.v. $\mathbf{X}$ has probability distribution μ
$\mathbb{E}_{\mathbf{X}\sim\mu}$	denotes that the expectation is taken w.r.t. the r.v. $\mathbf{X} \sim \mu$
$B^d_{r,p}$	the ℓ^p ball of radius r in $\mathbb{R}^d$, that is $\left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_p \leq r\right\}$
$\mathbb{S}^{d-1}$	the unit sphere in $\mathbb{R}^d$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	the normal distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$f(x) \lesssim g(x)$	$f(x) = O(g(x))$
$f(x) \gtrsim g(x)$	$f(x) = \Omega(g(x))$
$f(x) \simeq g(x)$	$f(x) = \Theta(x)$
$f(x) = \text{constant}(x)$	$\|f(x)\| = \Theta(1)$
$f(x) = \text{polylog}(x)$	$\|f(x)\| = \Theta(\left\|\log^k \|x\|\right\|)$ for some positive integer k

$f(x) = \mathrm{poly}(x)$	$\lvert f(x)\rvert = \Theta(\lvert x\rvert^k)$ for some positive integer k
$\hat{f}$, $\mathscr{F}(f)$	Fourier transform of f, that is $\hat{f}(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} f(\mathbf{x}) e^{-2\pi i \mathbf{x}^T \boldsymbol{\xi}}\, d\mathbf{x}$
$\check{f}$, $\mathscr{F}^*(f)$	inverse Fourier transform of f, that is $\check{f}(\mathbf{x}) = \int_{\mathbb{R}^d} f(\boldsymbol{\xi}) e^{2\pi i \mathbf{x}^T \boldsymbol{\xi}}\, d\boldsymbol{\xi}$
$f * g$	convolution of the functions f and g
$\mathrm{supp}(\eta)$	support of the function (or measure) η
$\mathrm{span}(A)$	linear space generated by the set A
$\mathrm{diam}(A)$	diameter of the set A, that is $\sup\{\lVert\mathbf{x} - \mathbf{y}\rVert_2 : \mathbf{x}, \mathbf{y} \in A\}$
$\dim(A), \dim_V(A)$	dimension of the subspace $A \subset V$
$\mathbb{1}\{A\}$, $\mathbb{1}_A$	indicator function of A

Chapter 1

Introduction

Neural networks (and more generally deep learning) have emerged in recent years as an incredibly powerful tool to perform machine learning tasks, most notably in computer vision and natural language processing. In general, deep learning models are defined as sums and composition of simple blocks, given by a linear transformation followed by the application of a non-linearity. These models, or *architectures*, are defined up to the choice of some parameters, which are determined by *training* the model over a dataset of interest. Roughly speaking, this account to perform some gradient-descent type algorithm to find the parameters which minimize a loss function defined over available data.

Despite an outstanding empirical success and a constant increase in model complexity, we are still far from a satisfying theoretical understanding of the performances of deep learning, even for fairly simple models. Classically, the theoretical questions arising in deep learning can be related to one of the following problems: *approximation*, that is to quantify how complex a model need to be to approximate a function of interest; *optimisation*, that is to understand how gradient base algorithms optimise the model parameters; *generalization*, that is to explain whether a given model trained on a number of samples can generalise over the whole data distribution. Notice that while we can define these questions separately, they are strictly related one to the other and they jointly

contribute the success of a model.

The work described in this thesis considers which role certain architectural features of feed-forward neural networks play in two of the three problems mentioned above, namely approximation and optimisation. In chapters 2 and 3 we focus on the role of depth of a neural network in terms of approximation. In chapter 4 we focus on the role of width of a neural network in terms of the optimisation landscape of square loss functions evaluated over the network. More in detail, the thesis is structured as follows.

1.1 Contributions and structure of the thesis

- In chapter 2 we consider a specific case study, namely the problem of approximating parametric transport partial differential equations (PDEs), a setting where classical reduced order modeling techniques are known to suffer of a slow Kolmogorov width decay, by neural networks. We show that shallow neural networks essentially suffer the same slow decay of the approximation rate, while this is not necessarily the case for their deep counterpart. We explain how this can inspire the definition of deep versions of reduced order models, which enjoy approximation capabilities similar to deep neural networks. This chapter is based on joint work with Donsub Rim, Benjamin Peherstorfer and Joan Bruna, partially presented in the work [RVBP20]. With respect to the paper, we focus the presentation on the neural networks point of view. For this reason, we report in detail only the proofs of results regarding feed-forward neural networks as defined in the classical sense; these results represent an original contribution not present in the paper. For the proofs of results regarding classical (and deep) reduced order models, we provide the main ideas and intuition, and we explain the conceptual connection to the neural networks world.

- In chapter 3 we establish results regarding limitations of shallow neural networks in approximating functions defined over an high-dimensional domain. These results mainly deal with

the Fourier representation of neural networks and target functions. We generalize existing results, which are limited to radial functions, regarding the existence of two-hidden-layer neural networks which need an exponential (in input dimension) number of parameters to be expressed as one-hidden-layer neural networks. We further establish that this is due to the fact that an essentially sufficient and necessary condition on a target function to be efficiently approximated by shallow models is the sparsity of its Fourier representation, a property which is not necessarily satisfied by deeper models. This chapter is based on joint work with Joan Bruna, Samy Jelassi and Tristan Ozuch [VJOB21]. With respect to the paper, the sections have been expanded to increase readability, with added intuitions and details behind the results and extended ideas of the proofs.

- In chapter 4, we consider the problem of describing the optimisation landscape of neural networks, in terms of properties amenable to descent methods. We look at global absence (or non-absence) of spurious valleys, which intuitively describe areas where descent methods could potentially get stuck far from global minima. We look at this problem for shallow feed-forward neural networks and square losses, and we show that distribution-independent absence of spurious valleys holds, for a fixed activation function, if and only if the network architecture 'fills' the functional space defined by shallow neural networks with the same activation. This chapter is based on joint work with Joan Bruna and Afonso Bandeira [VBB19]. Comparison to subsequent results have been included and section 4.4 has been extended to offer an extended intuition of the proof idea. The final section has been reworked to offer a better overview of the results, also with respect to recent advancements in the area.

- We conclude by presenting some related problems and open questions, in chapter 5.

For sake of simplicity of the exposition and to increase readability, the detailed proofs of the various results have been collected in the appendices (one for each of the chapters, minus this one). In the rest of this introduction, we introduce the main definitions and problems we will be dealing

with. We also discuss relevant literature and provide a more detailed description of the contents of the following chapters.

Finally, we mention that there are other projects that the author worked on during his PhD [KV20, AVP21, BVB21], but that are not included or fully presented in this thesis. This is due to little overlap with the material presented here, and to the desire of limiting the discussion of this thesis to theoretical understanding of neural networks.

1.2 Neural networks

For $L \geq 1$, an L-hidden-layer feed-forward neural network is a function

$$f : \mathbf{x} \in \mathbb{R}^d \to \mathbf{x}^{(L+1)}(\mathbf{x}) \in \mathbb{C}^{d_{L+1}} \tag{1.1}$$

where $\mathbf{x}^{(L)}$ is defined by recursion by $\mathbf{x}^{(0)}(\mathbf{x}) = \mathbf{x}$,

$$\mathbf{x}^{(k)}(\mathbf{x}) = \sigma^{(k)}(\mathbf{A}^{(k)}\mathbf{x}^{(k-1)}(\mathbf{x})) \text{ for } k \in [L] \quad \text{and} \quad \mathbf{x}^{(L+1)}(\mathbf{x}) = \mathbf{A}^{(L+1)}\mathbf{x}^{(L)}(\mathbf{x}) \,,$$

where

$$\begin{aligned} \mathbf{A}^{(k)} &= [\mathbf{a}_1^{(k)}| \cdots |\mathbf{a}_{d_k}^{(k)}]^T \in \mathbb{R}^{d_k \times d_{k-1}} \quad \text{for } k \in [L], \\ \mathbf{A}^{(L+1)} &= [\mathbf{a}_1^{(L+1)}| \cdots |\mathbf{a}_{d_{L+1}}^{(L+1)}]^T \in \mathbb{C}^{d_{L+1} \times d_L} \end{aligned}$$

(with $d_0 = d$) and $\sigma^{(k)} : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ are *activation* functions, that is $\left(\sigma^{(k)}(\mathbf{x})\right)_i = \sigma_i^{(k)}(x_i)$ for some function $\sigma_i^{(k)} : \mathbb{R} \to \mathbb{R}$. A neural network is therefore a sequence of sums and compositions of *ridge* functions, that is functions of the form $\mathbf{x} \mapsto \sigma(\mathbf{w}^T\mathbf{x})$. In the following, unless specified, we only consider neural networks (or, more simply, networks) as defined in (1.1). Most of the times we will deal with real-valued networks, that is $\mathbf{A}^{(L+1)} \in \mathbb{R}^{d_{L+1} \times d_L}$. We say that a network

has activation σ if $\sigma_i^{(k)}(x) = \sigma(x + b_i^k)$ for some bias term $b_i^k \in \mathbb{R}$ for all k, i. The function

$$\mathbf{x} \in \mathbb{R}^{d_{k-1}} \mapsto \sigma^{(k)}(\mathbf{A}^{(k)}\mathbf{x}) \in \mathbb{R}^{d_k}$$

is called k-th *hidden (or inner) layer* of *width* d_k, for $k \in [L]$, while we refer to the linear function defined by $\mathbf{A}^{(L+1)}$ as the last (or $L+1$-th) layer. We refer to the value $W(f) \doteq \max_{k\in[L]} d_k$ as width of the network f and to the vectors $\mathbf{a}_i^{(k)}$ as *weights* (of the k-th layer), for all k, i. A basic complexity measure for neural network (1.1) is given by the total number of units, or *size*:

$$N(f) \doteq \sum_{k=1}^{L} d_k \ .$$

Notice that this coincide with the network width if $L = 1$. The number of layers $L(f) = L$ is also a relevant measure of complexity, which we refer to as *depth*. Finally, in the following we sometimes require a control on the value of the weights; such controls are expressed in terms of norm p of the weights, that is

$$m_p(f) \doteq \max_{k,i} \|\mathbf{a}_{k,i}\|_p \ ,$$

for some $p \in [1, \infty]$.

1.2.1 Functional spaces of shallow neural networks

In the following, we will be looking at properties of certain classes of neural networks. In particular, we will be often dealing with spaces of one-hidden-layer networks, which we also refer to as *shallow*. This is contrast with neural networks with more than one hidden layer, which we refer to in the following as *deep*.

We denote the space of (scalar-valued) one-hidden-layer networks with at most N units by $\mathcal{F}_N$, and the space of one-hidden-layer networks with at most N units and given activation σ (resp., given activation σ and no bias terms) by $\mathcal{F}_N^\sigma$ (resp., $\mathcal{F}_N^{\sigma,0}$). Notice that every one-hidden-layer

network $f \in \mathcal{F}_N^\sigma$ can be equivalently written as

$$f : \mathbf{x} \in \mathbb{R}^d \mapsto \sum_{k=1}^{N} u_k \sigma \ \ \mathbf{w}_k^T \mathbf{x} + b_k \ \ = \int_{\mathbb{R}^{d+1}} \sigma(\mathbf{w}^T \mathbf{x} + b)\, d\pi_N(\mathbf{w}, b) \ ,$$

where π_N denotes the discrete signed measure $\pi_N = \sum_{k=1}^{N} u_k \delta_{(\mathbf{w}_k, b_k)}$. When the number of units grows to infinity, one can consider the following limit functional space

$$\mathcal{H}_\sigma^1 = \quad h = h_\pi^\sigma : \pi \text{ is a finite signed Radon measure on } \mathbb{R}^{d+1} \quad ,$$

where h_π^σ is defined as

$$h_\pi^\sigma : \mathbf{x} \in \mathbb{R}^d \mapsto \int_{\mathbb{R}^{d+1}} \sigma(\mathbf{w}^T \mathbf{x} + b)\, d\pi(\mathbf{w}, b) \ . \tag{1.2}$$

The space the space $\mathcal{H}_\sigma^1$ is a normed space, equipped with the norm

$$\gamma_1(h) = \inf_{\pi \,:\, h = h_\pi^\sigma} \|\pi\|_1 \ . \tag{1.3}$$

In particular, $\mathcal{F}_N^\sigma = \{h_\pi \in \mathcal{H}_\sigma^1 : |\mathrm{supp}(\pi)| \leq N\}$. Loosely speaking, the space $\mathcal{H}_\sigma^1$ consists of functions which are efficiently approximable by one-hidden-layer neural networks (of finite width). Consider $h_\pi^\sigma \in \mathcal{H}_\sigma^1$. By linearity, we can assume, w.l.o.g., that π is non-negative. Then, we can write

$$h_\pi^\sigma(\mathbf{x}) = v \cdot \mathbb{E}_{(\mathbf{w},b)\sim\hat{\pi}} \big[\sigma(\mathbf{w}^T \mathbf{x} + b)\big], \tag{1.4}$$

where $v = \|\pi\|_1$ and $\hat{\pi} = v^{-1}\pi$ is a probability measure. Sampling $\{(\mathbf{w}_k, b_k)\}_{k=1}^N$ from $\hat{\pi}$, one can approximate the expectation in (1.4) as

$$f_N(\mathbf{x}) \doteq \frac{v}{N} \sum_{k=1}^{N} \sigma(\mathbf{w}_k^T \mathbf{x} + b_k) \ . \tag{1.5}$$

There are many results in the literature that estabilish (typical Monte-Carlo-like) rates of convergence for this type of approximation, for example [BLM89, Bar93, YSW95, CB00, KB18]. In fact, many approximation results of an objective function by one-hidden-layer neural networks consist of approximating the objective function by a function in $\mathcal{H}^1_\sigma$ and then sampling as in eq. (1.5).

1.3 Approximation theory

A classic result about approximation by neural networks is the so-called Universal Approximation Theorem (UAT). It essentially states that any continuous function can be approximated by one-hidden-layer neural networks with an indefinite number of units. Several versions are available in the literature, see for example [Cyb89, HSW89, Hor91]; we report here the main results contained in [LLPS93].

Theorem 1.1 (UAT)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any function which is not a polynomial and with at most a finite number of discontinuity points; let $\mathcal{F}^\sigma \doteq \cup_{N=1}^\infty \mathcal{F}^\sigma_N$ be the space of shallow networks with activation σ. Then for any $K \subset \mathbb{R}^d$ compact, any finite measure μ on K which is absolutely continuous with respect to the Lebesgue measure, it holds that the space $\mathcal{F}^\sigma$ is dense in $C(K)$ (with respect to the L^∞ norm) and in L^p_μ.*

While this is a fundamental result, in the sense that it proves that neural networks are a reasonable class to consider to approximate generic functions, it has two main limitations. The first one is that it does not provide a rate of approximation, in terms of N, for any given objective function. The second one is that it only concerns shallow networks, and it is not clear (from this result) whether there is an advantage in considering deep networks.

There is extensive literature providing approximation rates for certain classes of functions, e.g. [BL91, Mha96, Pin99, MM00, Yar17]; a nice review of these results is given in [GRK20]. Works from the 90s /early 2000s deal with approximation by shallow networks with a smooth activation function σ. Such works essentially state that a function f of smoothness s can be ϵ-approximated

by a one-hidden-layer network with width $N \simeq \epsilon^{-d/s}$, where d is the input dimension. The following is a prototypical example of these type of results.

Theorem 1.2 (Informal, [Mha96, MM00]). *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ smooth. Then, for every function $f \in W_p^s([0,1]^d)$[1] and $\epsilon \in (0,1)$, there exists $f_N \in \mathcal{F}_N^\sigma$ such that*

$$\|f - f_N\|_p \leq \epsilon \|f\|_p \tag{1.6}$$

for some $N \lesssim \epsilon^{-d/s}$. Moreover, this rate is optimal, in the sense that there exists $f \in W_p^s([0,1]^d)$ such that any network $g \in \mathcal{F}^\sigma$ satisfying (1.6) must verify $N(g) \gtrsim \epsilon^{-d/s}$.

More recently, with the advent of piecewise-linear activation functions, such as the ReLU $\sigma(x) = x_+$, in practical application of neural networks, similar results have been shown for non-smooth activation as well.

Theorem 1.3 (Informal, [Yar17]). *Let σ be the ReLU activation function. Then, for every function $f \in W_p^s([0,1]^d)$ and $\epsilon \in (0,1)$, there exists a neural network $f_{L,N}$ with activation σ, depth $L \lesssim \log \frac{1}{\epsilon}$ and size $N \lesssim \epsilon^{-d/s} \mathrm{polylog} \frac{1}{\epsilon}$ such that*

$$\|f - f_{L,N}\|_p \leq \epsilon \|f\|_p \,. \tag{1.7}$$

Moreover, this rate is essentially optimal, in the sense that there exists $f \in W_p^s([0,1]^d)$ such that any network $f_{L,N}$ satisfying (1.7) must verify $N \gtrsim \epsilon^{-d/s}$.

Notice how these results are cursed by dimensionality: the number of units needed to obtain a certain approximation threshold grows exponentially with the input dimension d, unless the regularity of the objective function grows at least proportionally with d. Moreover, some of the cited

[1] For $\Omega \subset \mathbb{R}^d$, we denote by $W_p^s(\Omega)$ the Sobolev space of $L^p(\Omega)$ functions with derivatives up to degree s in $L^p(\Omega)$.

results do not specify the norm of the weights of the network $f_{L,N}$ which achieve such rate, that is $m_p(f_{L,N})$, which can be relevant in practical applications.

Finally, the approximation results cited so far prescribe a specific value to the depth of the network, needed to obtain a certain approximation rate, for different activation functions. Nevertheless, they do not provide insights on the trade-off between width and depth in approximation.

Remark 1. It must be noticed that approximation results (positive or negative) require an activation function to be fixed, or to belong to a properly defined class. Indeed, using Kolmogorov's superposition theorem [Kol56] it is easy to show the following (see [Pet20] for a proof).

Theorem 1.4. *There exists a continuous activation* $\sigma : \mathbb{R} \to \mathbb{R}$ *and constants* $C > 0$, $k \geq 1$ *integer such that, for every* $f \in C([-1,1]^d)$ *and* $\epsilon > 0$ *there exists a two-hidden-layer neural network* g *with at most* Cd^k *units such that* $\|f - g\|_\infty \leq \epsilon$.

In this thesis, we think about the activation function as a generic constant-Lipschitz function, and we require some more specific assumptions depending on the result.

1.3.1 Barron's theorem

The proofs of the approximation results cited above are essentially of two types. One type of proof consists in approximating the target function with some other basis function, such as polynomials or trigonometric polynomials, and then showing that each of these basis functions can be approximated by neural networks. In essence, proofs of this type show that neural networks perform as well as the underlying approximation procedure. The other type of proofs consists in showing that the target function admits an integral representation such as in (1.2); the approximation is then given by sampling as in (1.5).

The approximation rates mentioned in the previous section suffer from the curse of dimensionality, unless the objective function is highly regular. Following the latter proof technique, and moving away from the Sobolev spaces considered above, in the seminal work [Bar93], Barron

showed that, by terms of the Fourier transform, it is possible to describe a functional space for which approximation by shallow neural network holds at rate independent from the dimension. More specifically, Barron showed that if σ is some fixed sigmoidal[2] activation, μ is a probability measure supported on $[-1,1]^d$ and $f \in L^2(\mathbb{R}^d)$ satisfies

$$v = \int_{\mathbb{R}^d} \|\boldsymbol{\xi}\|_1 |\hat{f}(\boldsymbol{\xi})| \, d\boldsymbol{\xi} < \infty \,, \tag{1.8}$$

then it holds that

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f - f_N^\sigma\|_\mu^2 \leq \frac{4v^2}{N} \,.$$

Essentially, the proof of this result consists in controlling the norm γ_1 of f (introduced in (1.3)) by the quantity v. Various extensions of this result have since then been proved, with different assumptions on the activation function, on the error measure and showing the similar bounds for $m_\infty(f_N)$ as well; see for example the work [KB18] and references therein. Recently, a multi-layer version of this result has been proposed as well [BN20].

Condition (1.8) essentially requires the Fourier transform of ∇f to be integrable: this is because $\boldsymbol{\xi} \cdot \hat{f}(\boldsymbol{\xi}) = -i\mathscr{F}[\nabla f](\boldsymbol{\xi})$. In particular, it implies that $f \in C^1(\mathbb{R}^d)$. Intuitively, one can think about the constant v as a sort of L^1 norm: it is going to be large when $\hat{f}$ is 'spread' in the frequency domain. From a neural network perspective this makes sense, as one-hidden-layer networks have a sparse Fourier transform (see Section 1.5); it is thus reasonable to think about condition (1.8) as a *relaxed* sparsity condition in the frequency domain.

1.4 The depth-width tradeoff

A critical question for the use of neural networks is the choice of the architecture. For feed-forward neural networks this is equivalent to: should one should use wider or deeper networks? It is well

[2]That is, $\sigma : \mathbb{R} \to \mathbb{R}$ is a bounded measurable function such that $\lim_{x\to-\infty} \sigma(x) = 0$ and $\lim_{x\to\infty} \sigma(x) = 1$.

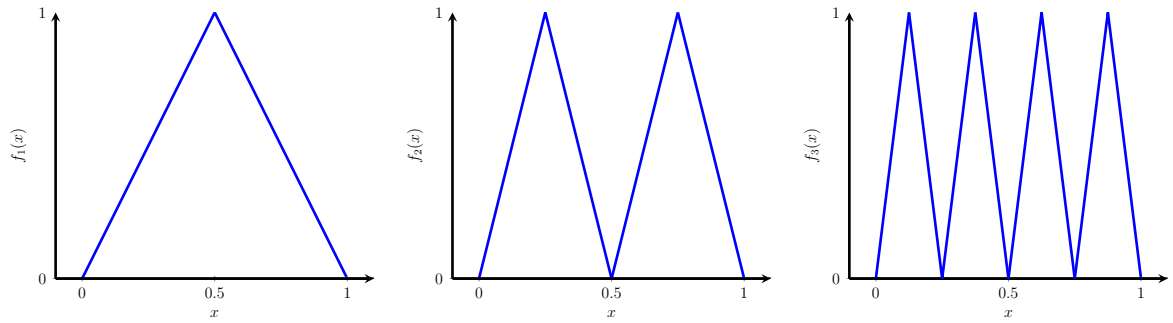

Figure 1.1: The function f_L is defined as the composition of h with itself L times. The plots, from left to right, show the graphs of, respectively, $f_1 = h$, $f_2 = h \circ h$ and $f_3 = h \circ h \circ h$.

known to practitioners that depth is essential to the performances of neural networks, but how can we quantify this fact?

This last question received particular interest in recent years. Consider, for example, the case of networks with the ReLU activation. Theorem 1.3 asks for the depth neural networks to increase as $\log \frac{1}{}$ in order to approximate a target function up to a certain accuracy ϵ. In this specific case, this is due to the difficulty of ReLU networks to approximate smooth function. Is this necessary? Or the same approximation rate can be obtained by a shallow model? In other terms, is it possible to showcase a function such that the approximation rate by shallow networks is substantially worse than the corresponding approximation rate by deep neural networks?

The answer to all these questions is positive. A simple example of such function is the *saw-tooth* function $f_L : [0,1] \to [0,1]$ defined as the linear interpolation of the points $\left\{ \left(\frac{k}{2^L}, \frac{1-(-1)^k}{2} \right) \right\}_{k=0}^{2^L}$. Consider the function $h : [0,1] \to \mathbb{R}$ defined by

$$h(x) = 2\ x_+ - (2x-1)_+ \ = \begin{cases} 2x & \text{if } x \in [0, 1/2] \\ 2 - 2x & \text{if } x \in [1/2, 1] \end{cases}.$$

Then $f_L = h^{\circ L}$, the composition of h with itself L times. This implies that f_L can be described exactly by a network with $O(L)$ units and depth $O(L)$. Now, say that we wish to describe the function f_L as a one-hidden-layer network; how many units are needed? Each one-hidden-layer

network, with the ReLU activation and N units, on $[0,1]$ is a piecewise linear function with at most $O(N)$ pieces. On the other hand, f_L is a piecewise linear function on $[0,1]$ with $O(2^L)$ pieces. Intuitively, this implies that the amounts of units needed to approximate the function f_L by shallow networks (with the ReLU activation) grows exponentially with L. This idea has been formalized in a seminal work by Telgarsky.

Theorem 1.5 ([Tel16]). *For any $L > 1$, f_{L^2+2} is a ReLU neural network of size $O(L^2)$ and depth $O(L^2)$, and any ReLU neural network g of depth at most L and size at most 2^L satisfies*

$$\int_0^1 |f_{L^2+2}(x) - g(x)|\, dx \geq \frac{1}{32}\,.$$

Roughly speaking, a similar reasoning shows that a depth of the order of $L \simeq \log \frac{1}{\epsilon}$ is needed to achieve exponentially efficient approximation by ReLU networks. Given a smooth function with positive curvature on a certain interval, the best approximation by piecewise linear function with M pieces achieves a uniform error of the order of M^{-2} [EEJ04]; a ReLU network with depth L and width N is a piecewise linear function with at most $O(N^L)$ pieces. Thus the following holds.

Theorem 1.6 (Informal, [LS16, Yar17, SS17a]). *Let $f : [0,1] \to \mathbb{R}$ be a non-linear sufficiently regular[3] function. Then any ReLU network g with depth at most L and width at most N satisfies $\|f - g\|_\infty \gtrsim N^{-2L}$. On the other hand, for every $\epsilon > 0$ there exists a ReLU network of depth $\mathrm{polylog}\frac{1}{\epsilon}$ with $\mathrm{polylog}\frac{1}{\epsilon}$ units such that $\|f - g\|_\infty \lesssim \epsilon$.*

Results on this line have been shown for different activations under different assumptions on the objective functions, such as polynomials [RT17], functions with a compositional structure [PMR+17] or piece-wise smooth function [PV18]. The result of [Tel16] has been further generalized using a notion of periodicity [CNPW19]. Moreover, this depth-width trade-off has been analyzed through different lens than approximation capabilities, such as classification capabili-

[3]For the lower bound, it is sufficient that $f \in C^2([0,1])$. For the upper bound, it is sufficient that f is $C^\infty([0,1])$ and it satisfies $(n!)^{-1}\|f^{(n)}\|_\infty \lesssim 1$ for any $n \geq 0$.

ties [MSS19], exact representability [ABMM16], Betti numbers [BS14], number of linear regions [PMB13, MPCB14], trajectory lengths [RPK+17], globale curvature [PLR+16] or topological entropy [BZL20]. In essence, all these results state that networks expressivity improve exponentially as we increase the depth.

1.4.1 Efficient modeling of transport partial differential equations: a study case

In chapter 2, we show a case of the aforementioned benefits of depth. We consider the problem of model order reduction of parametrized transport partial differential equations (PDEs). For this class of PDEs, standard model order reduction methods exhibits a slow convergence, making it difficult to use them in practice. We formally explain where such difficult stems from, generalizing existing lower bounds. Inspired by this fact, we look at the problem of approximating solutions via neural networks. We show that shallow models suffer from similar slow rates as (standard) reduced order models, while deep neural networks can potentially overcome this issue, thanks to the natural compositional structure of the solutions. Finally, we show how one can get inspiration from deep neural networks to define efficient deep reduced order models.

1.5 Curse of dimensionality and the frequency domain

In the seminal work [ES16], Eldan and Shamir show that in high dimensions d, the trade-off between width and depth is even more striking, from the point of view of approximation. Eldan and Shamir provide an example of a function $f : \mathbb{R}^d \to \mathbb{R}$ such that f can be approximated by a two-hidden-layers neural network with $N \lesssim \mathrm{poly}(d)$ units, but which requires $N \gtrsim e^{\Omega(d)}$ units to be approximated by a one-hidden-layer neural network. This phenomena is often referred to as *depth-separation*, and holds under mild assumptions on the activation function.

The reason behind this is to attribute to the 'shape' of a one-hidden-layer neural network in the

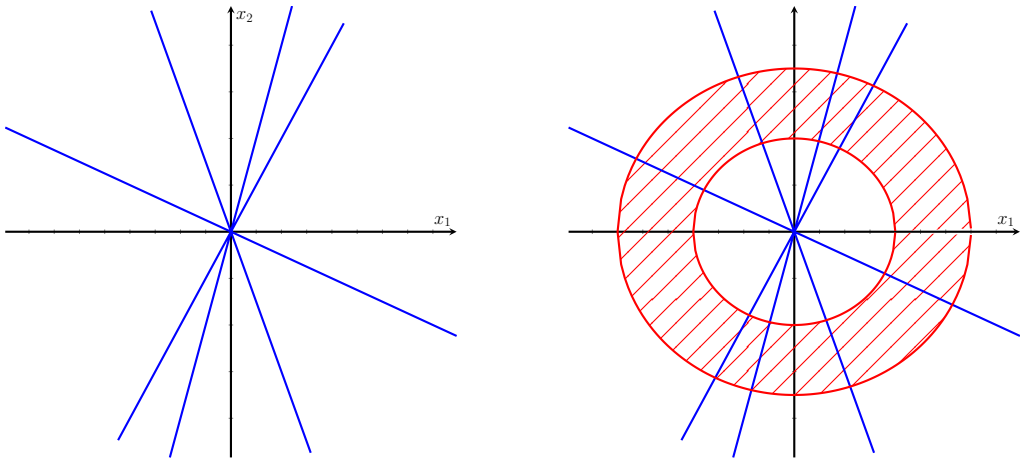

Figure 1.2: Left: the blue lines are the support of the Fourier transform of a one-hidden-layer network with $N = 4$ units. Right: One-hidden-layer networks with few units fail to approximate radial functions with high energy (represented by the red shaded area).

frequency domain. Consider a single ridge function $\psi_{\mathbf{w}} : \mathbf{x} \in \mathbb{R}^d \mapsto \sigma(\mathbf{w}^T\mathbf{x})$, for some continuous activation $\sigma : \mathbb{R} \to \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$. It is not difficult to show that the Fourier transform[4] of $\psi_{\mathbf{w}}$ satisfies

$$\operatorname{supp}(\mathscr{F}(\psi_{\mathbf{w}})) = \operatorname{span}(\{\mathbf{w}\}) \, .$$

By linearity, it follows that the Fourier transform of a one-hidden-layer neural network f_N with N units is supported on the union of N rays, as shown in Figure 1.2 (left). This implies that the Fourier transform of f_N is sparse at high frequencies, unless N grows exponentially with d; more formally, it holds that

$$\frac{r\mathbb{S}^{d-1} \cap \operatorname{supp}(\mathscr{F}(f_N)) + [-\epsilon, \epsilon]^d}{|r\mathbb{S}^{d-1}|} \lesssim N e^{-\Theta(d)}$$

for some given $\epsilon > 0$ and $r > 0$ large enough [ES16]. Intuitively, if the Fourier transform of a target function is uniformly distributed over a sphere of sufficiently large radius, then the number of units needed to well approximate (by shallow networks) such function grows exponentially with d. A prototypical example of such functions is given by radial functions $f(\mathbf{x}) = \varphi(\|\mathbf{x}\|)$, as their

[4]The Fourier transform of $\psi_{\mathbf{w}}$ is intended in the sense of distributions.

Fourier transform is also radial. Eldan and Shamir formalize this idea, by showcasing a radial function which can not be approximated up to a certain accuracy by one-hidden-layer networks, unless the number of units in the network increases exponentially with d. Here, the approximation error is measured in L^2, under an appropriately chosen measure with polynomial decay. Moreover, they show that such function can, on the other hand, be efficiently approximated by a two-hidden-layer network, where the first layer approximates the radial function $\mathbf{x} \mapsto \|\mathbf{x}\|$, and the second layer approximates the non-linearity φ.

This result has been further refined in [SS17a], and similar results have subsequently been shown in [Dan17a, JNS19]. All of these results deal with objective function which are (essentially) radial[5]. An open problem is to understand whether they can be extended to different classes of functions.

1.5.1 Depth-separation beyond radial functions

Roughly speaking, Barron's result [Bar93] establish a sparsity condition on the Fourier transform of the objective function which is *sufficient* for efficient approximation by shallow networks. On the other hand, the result by Eldan and Shamir [ES16] suggests that such sparsity is also *necessary*. In fact, looking again at Figure 1.2, this seems quite intuitive. In chapter 3, we focus on the high-dimensional regime and we further establish a formal understanding of approximation properties of neural networks by terms of Fourier representations.

As mentioned above, existing high-dimensional depth-separation results focus on functions with a radial structure. The first contribution of this chapter is to extend such results to a different class of functions, namely functions with piece-wise oscillatory structure, by building on the proof strategy in [ES16]. The piece-wise structure resembles the ones encountered in ReLU networks.

The oscillatory component of such functions needs to grow polynomially in d for the depth-

[5] In fact, Daniely considers an objective function of the form $f(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}^T\mathbf{y})$. Although, as noticed in [SES19], such functions can essentially be reduced to radial ones by a polarization identity.

separation result to hold. We complement the depth-separation result by showing that, if the rate of oscillation of the objective function is constant, then approximation by one-hidden-layer networks holds, uniformly over a set of constant radius, at a $\mathrm{poly}(d)$ rate for any fixed error threshold. The proof technique also sheds light on why the Fourier transform of a deep neural network is in general not sparse.

As mentioned, the common theme in the proof of such approximation lower bounds is the fact that one-hidden-layer fail to approximate high-energy functions whose Fourier representation is spread in the domain. The choice of the approximation domain plays a critical role in the proof depth-separation results, such as the one presented in section 3.2, and represents a source of gaps with the approximation upper bounds that we present in section 3.3. In section 3.4 we focus on approximation over an approximation domain of constant radius, namely the sphere $\mathbb{S}^{d-1}$ in dimension d. We provide a characterization of both functions which are efficiently approximable by one-hidden-layer networks and of functions which are provably not, in terms of their Fourier representation. We establish conditions in terms of sparsity or spreadness of such Fourier representation, marking a further step in formalizing the mentioned intuition.

1.6 Neural networks training

In supervised learning, once a certain network architecture is fixed, the weights of a network f as in (1.1) are found by optimising a loss function of the form

$$L(\boldsymbol{\theta}) = \mathbb{E}\, \ell(\Phi(\mathbf{X}; \boldsymbol{\theta}), \mathbf{Y})$$

where $\boldsymbol{\Phi}(\cdot; \boldsymbol{\theta})$ denotes the function f in (1.1) for a specific choice of parameters

$$\boldsymbol{\theta} = \left(\mathbf{A}^{(k)} \right)_{k \in [L+1]} \cup \left(\mathbf{b}^{k} \right)_{k \in [L]} .$$

The term ℓ denotes a convex function $\ell : \mathbb{R}^{d_{L+1}} \times \mathbb{R}^{d_y} \to [0, \infty)$ and the random variables $\mathbf{X}, \mathbf{Y}$ model the data distribution. In practice, the loss function is optimized following a gradient-based iterative algorithm, which, in its most basic form, is given by

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \cdot \mathrm{g}_k \,, \tag{1.9}$$

where g_k is an approximation of $\nabla L(\boldsymbol{\theta}_k)$ and $\eta_k > 0$ is a *step-size*. These type of algorithms are usually referred to as gradient descent algorithms, since they consist of taking repeated steps in the opposite direction of the (approximate) gradient of the function at the current point, this being the direction of steepest descent. The general idea is that, as $k \to \infty$, the iterate $\boldsymbol{\theta}_k$ should approach

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \,.$$

When the true gradient is used ($\mathrm{g}_k = \nabla L(\boldsymbol{\theta}_k)$), the method in (1.9) is called gradient descent (GD), and dates back to [C+47]. Under mild assumptions on the loss function L, GD is guaranteed to find an ϵ-approximate stationary point, that is a point $\boldsymbol{\theta}_k$ such that $\|\nabla L(\boldsymbol{\theta}_k)\|_2 \leq \epsilon$, for $\epsilon > 0$.

Theorem 1.7 ([Nes98])**.** *If the gradient ∇L is ν-Lipschitz and $\eta = \nu^{-1}$, then there exists*

$$k \leq \frac{\nu(L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*))}{\epsilon^2}$$

such that $\boldsymbol{\theta}_k$ is an ϵ-approximate stationary point.

When g_k is a random vector such that $\mathbb{E}\mathrm{g}_k = \nabla L(\boldsymbol{\theta}_k)$, the method in (1.9) is known as stochastic gradient descent (SGD), and dates back to [RM51]. Under suitable assumptions on the random vector g_k, results on the line of Theorem 1.7 are known for SGD as well; see e.g. [BCN18]. In the case that the function L is convex, stationary points correspond to global minima; in this setting, the results just cited can be tightened to yield provable convergence of the iterate (1.9) to a minima $\boldsymbol{\theta}^*$. In fact, a great amount of analysis has been carried out in the convex setting, leading to novel

algorithms and accompanying theory.

Most algorithms used nowadays to optimize a loss function L for the case neural networks consists of SGD or variant of it; we refer to [BGC17] for a review. This is due to two main factors: (i) the simplicity and versatility of gradient descent type algorithms, (ii) their unexpected efficacy in optimizing complex models, most often resulting in highly non-convex loss functions. Nevertheless, theoretical justifications are limited.

Many works have recently tried to explain the empirical success of gradient descent type algorithms at optimizing neural networks. A first line of work deals with understanding convergence properties of GD/SGD for generic non-convex objectives, such as convergence to approximate second-order stationary points (see e.g. [JNG+19] and references therein), under different assumption on the objective. A complementary line of works focus on understanding properties of the loss function, amenable to such convergence, such as characteristics of minima or saddles. More recently, a third line of results were published, describing at the behaviour of such algorithms in two different asymptotic (in the number of the parameters of the model) regimes: the mean field limit (see e.g. [RVE18b]) and the neural tangent kernel limit (see e.g. [JGH18]).

A critical factor that is known to be advantageuous in practice, for training neural networks, is what is called *over-parametrization*: increasing the number of parameters allows gradient descent methods to reach parameters for which the error (the value of the loss function) is zero. Thus, different works have been devoted to understand, theoretically, how increasing the number of parameters affects the *optimization landscape*.

From this point of view, the aforementioned asymptotic regimes represent limit cases. Roughly speaking, the mean field regime describes gradient descent dynamics in the space of infinite-width neural networks $\mathcal{H}^1_\sigma$, via gradient flow theory. On the other hand, the neural tangent kernel limit is amenable to random features methods.

1.6.1 The optimization landscape and over-parametrisation

The work presented in chapter 4 falls in the second line of works mentioned above – that is, the body of work devoted to characterizing properties of the landscape of loss functions evaluated on neural networks, which may (or may not) explain the success of gradients descent method to optimize them, such as absence of local minima or saddles. While the focus in this case is on finite-width regimes (non-asymptotic), a main question studied is how over-parametrization may affect such properties.

In chapter 4, we study a key topological property of the loss: the presence or absence of *spurious valleys*, defined as connected components of sub-level sets that do not include a global minimum. Focusing on a class of one-hidden-layer neural networks defined by smooth (but generally non-linear) activation functions and on the square loss $\ell(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$, we identify a notion of intrinsic dimension and show that it provides necessary and sufficient conditions for the absence of spurious valleys. More concretely, if the width of the network exceeds such intrinsic dimension, then spurious valleys are guaranteed not to exist, independently of the data distribution. Conversely, if the same condition does not hold, we show that spurious valleys do exist for certain data distributions. The condition on the network width N that we deem responsible for this phenomena essentially requires that the network architecture 'fills' the functional space defined by the same, that is $\mathcal{F}_N^\sigma = \mathcal{F}_{N+1}^\sigma$ (where σ is a fixed activation). We explain that this can only happen, essentially, for discrete data distributions or polynomial activations, where the network expressivity is limited. This implies that square losses evaluated on generic one-hidden-layer neural networks provably present local minima which are 'hard' to escape from.

We conclude by discussing certain sampling regimes which suggest that, although spurious valleys may exist in general, they are confined to low risk levels and avoided with high probability on over-parametrised models, as the number of parameters increases. As the work of this chapter was done previously to recent theoretical advancements on neural networks optimization, we review this last section in terms of some related works, including a brief review on the aforementioned

asymptotic regimes, and point out some limitations of this approach.

1.7 Other perspectives

There are a lot of other problems of interest in the theory of neural networks that we do not discuss in this thesis. While it is necessary to understand approximation and optimization properties of neural networks, another very important aspect is that of generalization, that we do not tackle here. The type of architectures considered in this thesis only include feed-forward networks; although, there is a huge variety of different type of neural networks being used in practice. The structure of the data they operate on is also another important aspect to consider. The depth separation problem discussed above is here tackled only from the approximation perspective; although, it remains important to understand whether such example offer separation in terms of learnability as well. Finally (but not exhaustively), in this thesis we consider networks of constant depth and we measure their complexity based on their size. But there is a growing use of *iterative* models, where the width is fixed and one achieves higher accuracy by increasing the depth.

In chapter 5 we briefly discuss on some of these problems and on some related ideas.

Chapter 2

The power of depth in model order reduction of certain transport problems

2.1 Introduction

Due to the outstanding success of neural networks in the machine learning field, there is recently been a spur in trying to use such methods in other areas of science. In particular, a number of authors have started to develop deep learning methods to numerically solve PDEs; see e.g. [HJW18, GHJVW18, BWJ19]. In this chapter we are interested in understanding the role of depth for approximation of solution to parametric PDEs by neural networks.

We are interested in PDEs of first order for which traditional model reduction fails. Model reduction derives reduced models to obtain computationally cheap approximations of PDE solutions in reduced (low-dimensional) subspaces of the typically high-dimensional solution spaces corresponding to numerical solution methods for PDEs such as finite-element and finite-volume methods [HRS$^+$16]. Model reduction methods achieve speedups compared to traditional numerical solution methods if the manifold induced by the solutions of the PDEs can be approximated well with low-dimensional subspaces. Note that the work [KPRS19] shows that deep networks

are at least as efficient as reduced models under certain assumptions. The Kolmogorov N-width of a solution manifold quantifies how well the solutions can be approximated in N-dimensional subspaces. Thus, if the Kolmogorov N-width of the solution manifold corresponding to a PDE decays slowly with the dimension N, then reduced models require potentially high-dimensional subspaces to provide approximate PDE solutions with acceptable accuracy. Advection-dominated PDEs represent a class of problems for which standard model order reduction is known to not be efficient, but theoretical results on the Kolmogorov N-width of such problems are limited to constant speed problems [OR15, GU19]. In section 2.2.1 we give a concise explanation of this phenomena, and show that existing theoretical results generalize to a larger class of equations.

Deep neural networks have recently emerged as a possible alternative solution. The works [Wel20, LC20] make use of deep neural networks. There also has been efforts to approximate the solution manifold of parametric PDEs directly with deep neural networks [KPRS19, LP21, GPR+20], by exploiting the expressive power of neural networks for approximating solutions of PDEs and nonlinear functions in general [RPK19, DDF+19, SZ19]. Deep neural networks also have been used to compute the reduced coefficients [WHR19]. The key challenge in these approaches is achieving the level of computational efficiency desired in model reduction, as these deep neural network constructions are more computationally expensive to evaluate or manipulate than the classical reduced models.

In this chapter, we exploit the limitations of shallow neural networks for approximating the solution manifold of transport problems. In section 2.3, we show polynomial lower bounds for the approximation of the solution manifold by shallow networks. We complement this result by showing that deep networks can potentially overcome this issue. Finally, we show that one can get inspiration from the positive results for efficient approximation of solutions to linear transport problems by deep neural networks, to define a *deep version* of reduced order models, which we show to be exponentially efficient in section 2.3.1.

2.2 Reduced order models

Model order reduction concerns the problem of providing efficiently computable and reliable solutions for parametrised partial differential equations, where the parameters describe certain characteristics of the problem. For the purpose of our exposition, we consider the case of a parametric PDE defined by a spatial variable $x \in \Omega \doteq (0,1)$, a time variable $t \in [0, t_F]$ (for a certain final time t_F) and a set of parameters $\boldsymbol{\mu} \in D \subseteq \mathbb{R}^P$, for some $P \geq 1$. The solutions to the PDE can be described as a map

$$u : (x, t, \boldsymbol{\mu}) \in \Omega \times [0,1] \times \mathcal{D} \mapsto u(x; t, \boldsymbol{\mu}) \in \mathbb{R} \, ,$$

where $u(\cdot; t, \boldsymbol{\mu}) \in \mathbb{V} \doteq L^2(\Omega)$ denotes the solution to the PDE defined by the parameter $\boldsymbol{\mu}$ at time t. We define the solution manifold as the set

$$\mathcal{M} \doteq \{u(\cdot; t, \boldsymbol{\mu}) : (t, \boldsymbol{\mu}) \in [0, t_F] \times \Omega\} \subset \mathbb{V} \, . \tag{2.1}$$

Computations of the solutions are in practice carried out in a reliable high-fidelity approximation space $\mathbb{V}_\delta \subseteq \mathbb{V}$; $\mathbb{V}_\delta$ is taken to be a linear space of finite dimension $N_\delta < \infty$. Assuming that the space $\mathbb{V}_\delta$ is spanned by some *basis* functions $\{\varphi_k\}_{k \in [N_\delta]} \subset \mathbb{V}$, a high-fidelity *full* solution can be found, of the form

$$u_\delta(x, t; \boldsymbol{\mu}) = \sum_{k=1}^{N_\delta} a_k(t, \boldsymbol{\mu}) \varphi_k(x) \, , \tag{2.2}$$

with coefficients $\{a_k(t, \boldsymbol{\mu})\}$ that depend on time and parameter. The size N_δ of the approximation space $\mathbb{V}_\delta$ is chosen so that, for each $(t, \boldsymbol{\mu})$, the full solution $u_\delta(\cdot, t; \boldsymbol{\mu})$ provides an approximation of the true solution $u(\cdot, t; \boldsymbol{\mu})$ up to a fidelity $\delta > 0$, that is

$$\| u_\delta(\cdot, t; \boldsymbol{\mu}) - u(\cdot, t; \boldsymbol{\mu}) \|_2 \leq \delta \, .$$

For a fixed N_δ, the *approximate solution manifold* is given by

$$\mathcal{M}_\delta \doteq \{u_\delta(\cdot; t, \boldsymbol{\mu}) : (t, \boldsymbol{\mu}) \in [0, t_F] \times \mathcal{D}\} .$$

Full solutions are typically computed with finite-difference, finite-element or finite-volume methods, which can be computationally expensive if a large N_δ is required to achieve the desired tolerance δ. Model reduction aims to construct reduced solutions in problem-dependent subspaces of much lower dimension $M \ll N_\delta$, to reduce computational costs [HRS+16]. Model reduction consists of an *offline stage* and an *online stage*. During the offline stage, the basis of the low-dimensional subspace, the reduced space $\mathbb{V}_M$, is constructed. A reduced basis is typically computed by collecting a finite subset $\mathcal{M}_S^\delta = \{u_\delta(\cdot, t_k; \boldsymbol{\mu}_k)\}_{k=1}^S$ of full solutions, where $\{(t_k, \boldsymbol{\mu}_k)\}_{k=1}^S \subset [0, t_F] \times \mathcal{D}$, and then computing a low-dimensional basis using, e.g., singular value decomposition (SVD). Let $\{\xi_k\}_{k=1}^M \subset \mathbb{V}$ be the set of the reduced-basis functions.

In the online phase, a *reduced solution* (or a *reduced-model solution*) is derived in the space spanned by the reduced basis,

$$u_M(x, t; \boldsymbol{\mu}) \doteq \sum_{k=1}^{M} \gamma_k(t, \boldsymbol{\mu}) \xi_k(x). \tag{2.3}$$

The coefficients $\{\gamma_k(t, \boldsymbol{\mu})\}_{k=1}^M$ of the reduced solutions are obtained by solving a system of equations for any given $(t, \boldsymbol{\mu}) \in [0, 1] \times \mathcal{D}$. The reduced system is derived using the PDE. In certain situations, the computational complexity of solving the reduced system scales with the dimension of the reduced space M only and is independent of the dimension of the full solutions N_δ. If the dimension M of the reduced space is small compared to the dimension N_δ of the full solutions, then solving for the reduced solution can be computationally cheaper than solving for the full solution. At the same time, the dimension M of the reduced space needs to be chosen sufficiently large so that the reduced solution are sufficiently accurate. For a fixed reduced basis $\{\xi_k\}_{k=1}^M$ with

M basis functions, we call the set of reduced solutions $\mathcal{M}_M$ the *reduced solution manifold*,

$$\mathcal{M}_M \doteq \{u_M(\cdot, t; \boldsymbol{\mu}) : (t, \boldsymbol{\mu}) \in [0, t_F] \times \mathcal{D}\} \,. \tag{2.4}$$

The *Kolmogorov N-width* [Pin12] of the reduced solution manifold provides a measure of optimal goodness of reduced order models for a given parametric PDE.

Definition 1. *The* Kolmogorov N-width *of a set of functions $\mathcal{M} \subset \mathcal{V}$ is defined as*

$$d_N(\mathcal{M}) = \inf_{\mathbb{V}_N} \sup_{u \in \mathcal{M}} \inf_{v \in \mathbb{V}_N} \|u - v\|_{\mathbb{V}} \,,$$

where the first infinimum is taken over all N-dimensional subspaces $\mathbb{V}_N$ of $\mathbb{V}$.

When the Kolmogorov N-width of a solution manifold $\mathcal{M}$ (2.1) is known, the smallest possible dimension M of its reduced manifold $\mathcal{M}$ (2.4) that satisfies the estimate

$$\|u(\cdot, t; \boldsymbol{\mu}) - u_M(\cdot, t; \boldsymbol{\mu})\|_2 \leq \epsilon \,, \quad \text{for all } (t, \boldsymbol{\mu}) \in [0, t_F] \times \mathcal{D} \,,$$

for given $\epsilon \in (0, 1)$, is also known. This implies that classical reduced models of the form (2.3) are not efficient for problems whose solution manifolds do not have a fast decaying Kolmogorov N-width [HRS+16].

2.2.1 The Kolmogorov N-width for advection problems

While it is known that the Kolmogorov N-width decays exponentially fast for many linear coercive parameterized partial differential equations [BMP+12, OR15], classical reduced models fail to be efficient not only for hyperbolic problems but for transport-dominated problems in general. This is well known fact in practice, but previous theoretical results are limited to constant-speed problems [OR15, Wel17, GU19] and Burger's equation [ELMV20]. In this section, we introduce the concept

of convective class, and we show that the Kolmogorov N-width of such a class decays at most polynomially. We use this concept to provide polynomial lower bounds for a large family of linear advections problems.

Definition 2. *We say that a set $\mathcal{M} \subset \mathbb{V}$ generates a $2N$-ball, for $N \geq 1$ integer, if there exists a set $B_{2N} := \{\phi_n\}_{n=1}^{2N} \subset \mathrm{span}(\mathcal{M})$ of linearly independent functions ϕ_n with the form*

$$\phi_n = \sum_{k=1}^{K} a_{n,k} u_{n,k} \quad \text{for some } K \geq 1,\ u_{n,k} \in \mathcal{M} \text{ and } \mathbf{a}_n \in \mathbb{R}^{k_n} \text{ with } \|\mathbf{a}_n\|_1 \leq 1. \tag{2.5}$$

The $2N$-ball B_{2N} is said orthogonal if $\phi_1, \dots, \phi_{2N}$ are orthogonal in $\mathbb{V}$. We say that the set $\mathcal{M}$ is α-convective[1]*, for some $\alpha > 0$, if for any $N \geq 1$ integer, $\mathcal{M}$ generates an $2N$-ball which generates an orthogonal $2N$-ball $B'_{2N} = \{\varphi_n\}_{n \in [2N]}$ with $\|\varphi_n\|_{\mathbb{V}} \gtrsim N^{-\alpha}$ for every n.*

Intuitively, if the solution manifold $\mathcal{M}$ generates $2N$-balls, approximating the manifold $\mathcal{M}$ by linear subspaces of a certain finite dimension is at least as difficult as approximating each of such $2N$-balls by finite linear subspaces of the same dimension. If such balls are orthogonal, such approximation rate can be controlled by the norm of the functions forming the balls; the α-convectivity notion essentially imposes that these norms decay at most polynomially. This intuition is formalized in the following; the proof reported below is a simplification of the one in [RVBP20], which holds for a more generic notion of α-convectivity.

Proposition 2.1. *Let $\mathcal{M} \subset \mathbb{V}$. If $\mathcal{M}$ generates an orthogonal $2N$-ball B_{2N}, then it holds that*

$$d_N(\mathcal{M}) \geq d_N(B_{2N}) \,. \tag{2.6}$$

If $\mathcal{M}$ is α-convective for some $\alpha > 0$, this implies that $d_N(\mathcal{M}) \gtrsim N^{-\alpha}$.

Proof. Let B_{2N} be a $2N$-orthogonal ball generated by $\mathcal{M}$ and let $\mathbb{V}_N$ be any linear subspace of $\mathbb{V}$

[1] This is a specific case of the full definition reported in [RVBP20], but we focus on this case here for sake of simplicity.

of dimension N. It holds that $B_{2N} = \{\phi_n\}_{n\in[2N]}$, where each function ϕ_n has the form (2.5). For any $\{w_k\}_{k\in[K]} \subset \mathbb{V}_N$, it holds that

$$\sup_{k\in[K]} \|u_{n,k} - w_k\|_{\mathbb{V}} \geq \sum_{k=1}^{K} |a_{n,k}| \|u_{n,k} - w_k\|_{\mathbb{V}} \geq \left\| \phi_n - \sum_{k=1}^{K} a_{n,k} w_k \right\|_{\mathbb{V}} \geq \inf_{v\in\mathbb{V}_N} \|\phi_n - v\|_{\mathbb{V}} .$$

Since the above holds for arbitrary $\{w_k\}_{k\in[K]} \subset \mathbb{V}_N$, it follows that

$$\sup_{u\in\mathcal{M}} \inf_{v\in\mathbb{V}_N} \|u - v\|_{\mathbb{V}} \geq \sup_{k\in[K]} \inf_{v\in\mathbb{V}_N} \|u_{n,k} - v\|_{\mathbb{V}} \geq \inf_{v\in\mathbb{V}_N} \|\phi_n - v\|_{\mathbb{V}} ,$$

which implies that

$$\sup_{u\in\mathcal{M}} \inf_{v\in\mathbb{V}_N} \|u - v\|_{\mathbb{V}} \geq \sup_{n\in[2N]} \inf_{v\in\mathbb{V}_N} \|\phi_n - v\|_{\mathbb{V}} .$$

Then, equation (2.6) follows by the definition of Kolmogorov N-width. Assume now that the functions $\{\phi_n\}_{n\in[2N]}$ are orthogonal and satisfy, for some costant $C > 0$, $\|\phi_n\|_{\mathbb{V}} \geq CN^{-\alpha}$ for all $n \in [2N]$. Let $\hat{\phi}_n = \phi_n/\|\phi_n\|_{\mathbb{V}}$. Then it follows

$$\sup_{n\in[2N]} \inf_{v\in\mathbb{V}_N} \|\phi_n - v\|_{\mathbb{V}} = \frac{C}{N^{\alpha}} \sup_{n\in[2N]} \inf_{v\in\mathbb{V}_N} \|\hat{\phi}_n - v\|_{\mathbb{V}} = \frac{C}{\sqrt{2}N^{\alpha}}$$

where the last equation follows by, e.g., Lemma 4.3 in [GU19]. □

2.2.1.1 Linear advection problems

In the rest of this chapter, we consider the following families of parametrized linear advection PDEs, defined by

$$\begin{cases} u_t + c(x,t,\boldsymbol{\mu})u_x = 0, & \text{for } (x,t) \in \mathbb{R} \times (0,t_F), \\ u(x,0;\boldsymbol{\mu}) = u_0(x), & \text{for } x \in \mathbb{R}, \end{cases} \tag{2.7}$$

for each $\boldsymbol{\mu} \in \mathcal{D}$, and we consider the (weak) solutions $u(x,t;\boldsymbol{\mu})$ for $x \in \Omega$ and $t \in [0,t_F]$. We make the following assumptions:

1. The function c is analytic in the space variable x and the time variable t (but not necessarily in $\boldsymbol{\mu}$). More specifically, we assume that for some $a,b>0$, $c(\cdot,\cdot;\boldsymbol{\mu})$ is an analytic function over a set $R \doteq \{(w,s) \in \mathbb{C} : |w-x| < a,\ x \in \Omega,\ |t| < b\}$ for every $\boldsymbol{\mu} \in \mathcal{D}$;

2. The function c is uniformly bounded away from zero, that is there exist $\nu > \iota > 0$ such that it holds $0 < \iota \leq c(x,t,\boldsymbol{\mu}) < \nu$ for any $(x,t,\boldsymbol{\mu}) \in R \times \mathcal{D}$;

3. The initial condition $u_0 \in L^\infty(\mathbb{R})$ is sufficiently regular on the interval of interest, that is, it holds $u_0 \in \mathrm{TV}([-\nu, 1+\nu])$ and $u_0 \in L^\infty([-\nu,1+\nu])$.

We will denote by $\mathcal{M}_c$ the solution manifold of such a parametrized PDE. One can solve for each solution in $\mathcal{M}_c$ by the method of characteristics by integrating along the characteristic curves [Eva98]. We will denote the characteristic curve for the initial condition x_0 by $X(t;x_0,\boldsymbol{\mu})$. Then the ODEs for the characteristic curves are

$$\begin{cases} X\ (t;x_0,\boldsymbol{\mu}) = c(X(t;x_0,\boldsymbol{\mu}),t;\boldsymbol{\mu}), \quad t \in (0,t_F)\ , \\ X(0;x_0,\boldsymbol{\mu}) = x_0\ . \end{cases}$$

By classical ODE theory [CL55, Chapter 1, Theorem 8.1], and thanks to the assumptions on c, $X(t;x_0,\boldsymbol{\mu})$ ($x_0 \in \Omega$, $\boldsymbol{\mu} \in \mathcal{D}$) is analytic with respect to the variable $t \in (0,t_F)$, for $t_F \leq \min\{a/\nu, b\}$. We will write X also as a function of its initial condition, $X(t,x;\boldsymbol{\mu}) := X(t;x,\boldsymbol{\mu})$. Since c is bounded away from zero, $\partial_x X > 0$ for $t \in (0,t_F)$, ensuring that the map is strictly increasing function of x. Furthermore, this implies that X is analytic with respect to x [RVBP20, Lemma 4.3]. If we express the transformation of the domain by

$$T_{(t,\boldsymbol{\mu})} : x \in \Omega \mapsto X(t,x;\boldsymbol{\mu})$$

it holds that $u_0(T^{-1}_{(t,\boldsymbol{\mu})}(x)) = u(x,t;\boldsymbol{\mu})$, and $u_0(x) = u(T_{(t,\boldsymbol{\mu})}(x),t;\boldsymbol{\mu})$ [DL89]. Since the results in the following sections do not depend on the values of the parameters in the assumptions on c above, we assume, for sake of simplicity, that $t_F = 1$ and $\nu = 1$.

2.2.1.2 Slow decay of the Kolmogorov N-width

In [OR15], it was proved that the solution manifold $\mathcal{M}_c$ is $\frac{1}{2}$-convective, in the case $c(\cdot;\mu) = \mu$ (where $\mathcal{D} \subset (0,1)$) and $u_0(x) = \mathbb{1}\{x \leq 0\}$. In fact, even for general c, it is not difficult to see why this is the case. The solution to the PDE is given by $u(x,t;\mu) = \mathbb{1}\{x \leq X(t;0,\mu)\}$. Since, for a fixed μ, the function $t \mapsto X(t;0,\mu)$ is continuous and increasing, there exists $0 < a < b < 1$ such that

$$\mathcal{B} = \{x \mapsto \mathbb{1}\{x \leq c\} : c \in [a,b]\} \subset \mathcal{M}_c.$$

The set $\mathcal{B}$ generates orthogonal $2N$-balls in the following way. For any $N \geq 1$, consider $a = x_0 < \cdots < x_{2N} = b$ a partition of $[a,b]$ in to $2N$ intervals of size $(b-a)/(2N)$. Then the functions

$$\phi_n(x) \doteq \frac{1}{2}\mathbb{1}\{x_{n-1} \leq x \leq x_n\} = \frac{1}{2}(\mathbb{1}\{x \leq x_n\} - \mathbb{1}\{x \leq x_{n-1}\})$$

satisfy (2.5) and are orthogonal in $\mathbb{V} = L^2([0,1])$. Moreover, it is easy to verify that $\|\phi_n\|_{\mathbb{V}} \geq N^{-1/2}$; this implies that $\mathcal{B}$, and thus $\mathcal{M}$, is $\frac{1}{2}$-convective. Thanks to Proposition 2.1, it follows that $d_N(\mathcal{M}) \geq N^{-1/2}$.

This idea can be extended to the case of $u_0 \in C^s(\Omega) \cap C^{s+1}(\Omega \setminus \{x_0\})$ for some $x_0 \in \Omega$ and $s \geq 0$. In this case, it follows that, at time t, the $(s+1)$-th derivative of the solution, $\partial_t^{s+1} u(\cdot,t;\boldsymbol{\mu})$, has a discontinuity at the point $X(t,x_0;\boldsymbol{\mu})$. The derivative $\partial_t^{s+1} u$ can be approximated by linear combination of the solution at different time increments, using a finite difference method. Such approximations, at different time steps, generate $2N$-balls for $N \geq 1$, which can be orthogonalised using Gram-Schmidt; in particular, the norm of the functions composing the orthogonal $2N$-balls can be lower bounded as $\omega(N^{s-1/2})$. This results is formalized in the following; for a detailed

proof of this result, we refer to section 4.2.1 of [RVBP20].

Proposition 2.2. *If u_0 satisfies the assumption above, then it holds that $d_N(\mathcal{M}_c) \gtrsim N^{-s-1/2}$.*

We remark that while here we focus on a linear advection problem (2.7), similar results have been also shown for other linear hyperbolic problems, such as the wave equation [GU19]. For non-linear problems, things are potentially even worse. For example, for the Burger's equation, it is possible to show that the collection of the characteristic curves themselves form a convective class. We refer to [RVBP20] for a proof of this fact.

2.3 PDE modeling via neural networks

In the work by Laakmann and Petersen [LP21] it has been shown that parametric solutions to (2.7) can be approximated by deep (ReLU) neural networks at a rate that is essentially the one provided in Theorem 1.3 (where the regularity refers to the regularity of the term c), by exploiting the solution formulation and the regularity of the characteristic curves. In this section we consider the same setup. While the work [LP21] shows upper bounds for approximation of parametric solutions by deep networks, we complement such results by showing lower bounds for approximation by shallow networks. We also discuss how, under certain assumptions, this implies a polynomial-versus-exponential separation from approximation with shallow-versus-deep networks.

We first show that shallow neural networks suffer of similar limitations of reduced order models for approximation of parametric solutions. In the case of a smooth non-linear initial condition and ReLU networks, this follows for example from existing bounds on approximation of smooth functions by neural networks. Nevertheless, such results do not apply, in general, to the case of non-smooth initial conditions and piece-wise polynomial activation. We show that in this case, one can leverage the fact that the solution is a wave moving at a non-linear speed to obtain a lower bound.

In the following, we consider neural networks with (p, r, s) semi-algebraic[2] activation functions, as defined in [Tel16]. This includes activation functions such as the ReLU $\sigma(x) = x_+$ or the step-function $\sigma(x) = \mathbb{1}\{x \leq 0\}$. Consider the PDE (2.7) with the initial condition $u_0(x) = \mathbb{1}\{x \leq 0\}$. Notice that such initial condition can be approximated at any accuracy by shallow networks with semi-algebraic activation and constant width. The solution u is given by

$$u(x, t; \boldsymbol{\mu}) = u_0(X^{-1}(x; t, \boldsymbol{\mu})) = \mathbb{1}\{x \leq X(0; t, \boldsymbol{\mu})\} .$$

Unless the term c is constant (in all variables), it holds that the characteristic curve map $(t, \boldsymbol{\mu}) \mapsto X(t; 0, \boldsymbol{\mu})$ is C^∞ and non-linear. On the other hand any shallow network approximation of u is only allowed to depend on linear combinations of $\boldsymbol{\mu}$ and t. Let $f_N \in \mathcal{F}_N^\sigma$ be a neural network $f_N : \mathbb{R}^{P+2} \to \mathbb{R}$, with σ a (p, r, s) semi-algebraic activation. Then, at fixed t and $\boldsymbol{\mu}$, the function $x \in \Omega \mapsto f_N(x, t, \boldsymbol{\mu})$ is a piece-wise polynomial of degree s with (at most) prN breakpoints of the form

$$\alpha_{n,j}(t, \boldsymbol{\mu}) = \mathbf{w}_n^T(t, \boldsymbol{\mu}) + b_{n,j}$$

for some $\mathbf{w}_n \in \mathbb{R}^{P+1}$ and $b_{n,j} \in \mathbb{R}$, $n \in [N_u]$, $N_u \leq N$, and $j \in [pr]$. Let $\mathcal{A}(t, \boldsymbol{\mu}) = \{\alpha_{n,j}(t, \boldsymbol{\mu})\}_{n,j}$ be the set of breakpoints for a given pair $(t, \boldsymbol{\mu})$ and let

$$\epsilon(t, \boldsymbol{\mu}) = \min_{\alpha \in \mathcal{A}(t,\boldsymbol{\mu}) \cup \{0,1\}} |X(0, t; \boldsymbol{\mu}) - \alpha| .$$

The term $\epsilon(t, \boldsymbol{\mu})$ denotes the distance between the solution breakpoint $X(0, t; \boldsymbol{\mu})$ and the closest point in $\mathcal{A}(t, \boldsymbol{\mu}) \cup \{0, 1\}$, at given $t, \boldsymbol{\mu}$. By definition, the network $f_N(\cdot, t, \boldsymbol{\mu})$ is a polynomial of

[2] A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called (p, r, s)-semi-algebraic if there exist p polynomials $\{q_k\}_{k\in[p]}$ of degree at most r, and m triples $\{(p_k, G_k, L_k)\}_{k\in[m]}$ where p_k is a polynomial of degree at most s and $G_k, L_k \subset [p]$, such that $\sigma(x) = \sum_{k=1}^{m} p_k(x) \prod_{j\in L_k} \mathbb{1}\{q_j(x) < 0\} \prod_{j\in G_k} \mathbb{1}\{q_j(x) \geq 0\}$.

degree s in the interval $[X(0,t;\boldsymbol{\mu}) - \epsilon(t,\boldsymbol{\mu}), X(0,t;\boldsymbol{\mu}) + \epsilon(t,\boldsymbol{\mu})] \subseteq [0,1]$. Then, it holds that

$$\begin{aligned}
\|u(\cdot,t,\boldsymbol{\mu}) - f_N(\cdot;t,\boldsymbol{\mu})\|_2^2 = & \int_0^1 (u(x,t,\boldsymbol{\mu}) - f_N(x;t,\boldsymbol{\mu}))^2\, dx \\
= & \int_0^{X(0,t;\boldsymbol{\mu})} (1 - f_N(x,t,\boldsymbol{\mu}))^2\, dx + \int_{X(0,t;\boldsymbol{\mu})}^1 (f_N(x,t,\boldsymbol{\mu}))^2\, dx \\
\geq & \int_{X(0,t;\boldsymbol{\mu}) - \ (t,\boldsymbol{\mu})}^{X(0,t;\boldsymbol{\mu})} (1 - f_N(x,t,\boldsymbol{\mu}))^2\, dx + \int_{X(0,t;\boldsymbol{\mu})}^{X(0,t;\boldsymbol{\mu}) + \ (t,\boldsymbol{\mu})} (f_N(x,t,\boldsymbol{\mu}))^2\, dx \\
\geq & \inf_{p \,:\, \deg(p) \leq s} \left[\int_{X(0,t;\boldsymbol{\mu}) - \ (t,\boldsymbol{\mu})}^{X(0,t;\boldsymbol{\mu})} (1 - p(x))^2\, dx + \int_{X(0,t;\boldsymbol{\mu})}^{X(0,t;\boldsymbol{\mu}) + \ (t,\boldsymbol{\mu})} p^2(x)\, dx \right] \\
= & \, \epsilon(t,\boldsymbol{\mu}) \inf_{p \,:\, \deg(p) \leq s} \int_{-1}^1 (u_0(x) - p(x))^2\, dx \gtrsim \epsilon(t,\boldsymbol{\mu}) \,,
\end{aligned}$$

where the infimum is taken over all polynomials $p : \mathbb{R} \mapsto \mathbb{R}$ of degree at most s. Therefore, we get that, for any $f_N \in \mathcal{F}_N^\sigma$, it holds

$$\begin{aligned}
\sup_{(t,\boldsymbol{\mu}) \in [0,1] \times \mathcal{D}} & \|f_N(\cdot,t,\boldsymbol{\mu}) - u(\cdot,t;\boldsymbol{\mu})\|_2 \geq \\
& \geq \inf_{\substack{\mathbf{w}_1,\ldots,\mathbf{w}_{N+1} \in \mathbb{R}^{P+1} \\ \mathbf{b}_1,\ldots,\mathbf{b}_{N+1} \in \mathbb{R}^{pr}}} \sup_{(t,\boldsymbol{\mu}) \in [0,1] \times \mathcal{D}} \min_{\substack{k \in [N+1] \\ j \in [pr]}} X(t;0,\boldsymbol{\mu}) - \mathbf{w}_k^T(t,\boldsymbol{\mu}) - b_{k,j} \ ^{1/2} \,.
\end{aligned}$$

The sup-inf problem in the value in the lower bound above is similar to the problem of fitting the function $(t,\boldsymbol{\mu}) \mapsto X(t;0,\boldsymbol{\mu})$ with a piece-linear function with $O(N)$ pieces, but slightly different; instead it consists of approximating such function in each point as the closest value to an ensemble of $O(N)$ linear functions. If the function is smooth and non-linear, then one can apply the following.

Lemma 2.3. *Let $f : [0,1]^d \to \mathbb{R}$ be a C^2 function which is non linear. Then it holds that*

$$\inf_{\substack{\mathbf{w}_1,\ldots,\mathbf{w}_N \in \mathbb{R}^d \\ \mathbf{b}_1,\ldots,\mathbf{b}_N \in \mathbb{R}^M}} \sup_{\mathbf{x} \in [0,1]^d} \inf_{\substack{k \in [N] \\ j \in [M]}} f(\mathbf{x}) - \mathbf{w}_k^T \mathbf{x} - b_{k,j} \ \geq \frac{C}{MN^3}$$

where $C > 0$ is a constant only depending on f.

The proof of this result follows a similar idea to the proof of the lower bound in Theorem 1.3, and it is reported in section A.1.1. The application of Lemma 2.3 then gives us the following result.

Proposition 2.4. *Assume that* $u_0(x;\boldsymbol{\mu}) = \mathbb{1}\{x \leq 0\}$. *If the function* c *in* (2.7) *is not constant and* σ *is a* (p, r, s)*-semi-algebraic activation, then it holds that*

$$\inf_{f_N \in \mathcal{F}_N^\sigma} \sup_{(t,\mu)\in[0,1]\times\mathcal{D}} \|u(\cdot,t;\boldsymbol{\mu}) - f_N(\cdot,t,\boldsymbol{\mu})\|_2 \gtrsim \frac{1}{(pr)^{1/2} N^{3/2}} \,.$$

Notice that the rate obtained for the lower bound is faster than the one on the Kolmogorov N-width given in section 2.2.1.2 for the same PDE. A similar result can be shown for some more generic initial conditions u_0 such that $u_0 \in C^s \setminus C^{s+1}$, for $s \geq 0$. We provide further details on extensions of Proposition to more general initial conditions 2.4 in section A.1.2. The remarkable fact about the above lower bound is that it holds for any semi-algebraic activation: the proof highlights the fact that, despite of the degree of the activation, the transport map $(t, \boldsymbol{\mu}) \mapsto X(0, t; \boldsymbol{\mu})$ can only be captured by (N) linear functions. On the other hand, deep neural networks do not suffer from this limitation: the following proposition shows that approximation by deep networks can potentially yield exponential rates. The proof is reported in in section A.1.2.

Proposition 2.5. *Consider* $u_0(x) = \mathbb{1}\{x \leq 0\}$. *Assume that the map* $T_0 : (t, \boldsymbol{\mu}) \in [0, 1] \times \mathcal{D} \mapsto X(0, t, \boldsymbol{\mu})$ *can be uniformly approximated by polynomials at an exponential rates, that is*

$$\inf_{p \,\in\, \mathbb{P}_{\leq r}^{P+1}} \sup_{(t,\boldsymbol{\mu})\in[0,1]\times\mathcal{D}} |T_0(t,\boldsymbol{\mu}) - p(t,\boldsymbol{\mu})| \leq e^{-\omega(r)} \,,$$

where $\mathbb{P}_{\leq r}^{P+1}$ *denotes the space of polynomials (with real coefficients) of degree at most* r *in* $P+1$ *(real) variables. Then the solution* u *can be* ϵ*-approximated by a neural network of depth* $\mathrm{polylog}^{\frac{1}{\epsilon}}$ *with* $\mathrm{polylog}^{\frac{1}{\epsilon}}$ *units, that is there exists a network* f_N *(with ReLU and step-function activations) of*

size $N = O\ \log^{P+3}\frac{1}{}$ *and depth* $O\ \log\frac{1}{}$ *which verifies*

$$\sup_{(t,\boldsymbol{\mu})\in[0,1]\times\mathcal{D}} \|f_N(\cdot,t,\boldsymbol{\mu}) - u(\cdot,t;\boldsymbol{\mu})\|_{\mathbb{V}} \leq \epsilon\ .$$

2.3.1 Deep reduced order models

Given the benefits that we just discussed of deep neural networks versus their shallow counterpart for approximation, a possible strategy to overcome the limitations of classical reduced order models is to construct a deep version of the latter. Recall that, in standard reduced order modelling, one express the parametric solution in the form

$$u_M(x,t;\boldsymbol{\mu}) = \sum_{k=1}^{M} \gamma_k(t,\boldsymbol{\mu})\xi_k(x)\ ,$$

where the functions ξ_k are fixed elements of the space $\mathrm{span}(\mathcal{M}_\delta)$. Consider the case where full solution (2.2) to the PDE (2.7) are constructed as piecewise linear functions on an equidistant grid with N_δ grid points, which can be represented as a specific one-hidden-layer network whose weights and biases in the hidden layer are fixed. Namely, the full solutions have the form

$$u_\delta(x,t;\boldsymbol{\mu}) = \sum_{k=1}^{N_\delta} w_k(t,\boldsymbol{\mu})\sigma(h^{-1}x - k - 1)\ ,$$

where $\sigma(x) = x_+$ is the ReLU activation and $h = 1/(N_\delta - 1)$. If the second layer weights $\mathbf{w}_k(t,\boldsymbol{\mu})$ belong to a low-dimensional subspace of dimension $M \ll N_\delta$, then one can write

$$\mathbf{w}(t,\boldsymbol{\mu}) = \mathbf{V}\boldsymbol{\gamma}(t,\boldsymbol{\mu})$$

where $\mathbf{V} \in \mathbb{R}^{N_\delta \times M}$ has orthonormal columns and $\boldsymbol{\gamma}(t, \boldsymbol{\mu}) \in \mathbb{R}^M$. This leads to reduced order models of the form

$$u_M(x, t; \boldsymbol{\mu}) = \sum_{k=1}^{M} \gamma_k(t, \boldsymbol{\mu}) \xi_k(x)$$

where $\xi_k(x) = \sum_{j=1}^{N_\delta} v_{jk} \sigma(h^{-1}x - j - 1)$. Relating to neural networks, the full solution u_δ can be written as a one-hidden-layer ReLU network

$$u_\delta(x, t; \boldsymbol{\mu}) = \mathbf{w}(t, \boldsymbol{\mu})^T \boldsymbol{\sigma}(\mathbf{W}_0\, x + \mathbf{b}_0)\ ,$$

where $\mathbf{W}_0 = h^{-1}(1, \dots, 1) \in \mathbb{R}^{N_\delta \times 1}$ and $\mathbf{b}_0 = -(0, 1, \dots, N_\delta - 1) \in \mathbb{R}^{N_\delta}$. Here, the second layer weights depend on $t, \boldsymbol{\mu}$ while the first does not. The reduced order model has the form

$$u_M(x, t; \boldsymbol{\mu}) = \boldsymbol{\gamma}(t, \boldsymbol{\mu})^T \boldsymbol{\xi}(x)$$

where each ξ_k is a *reduced* activation. More generally, one could imagine to start with *full deep* solutions, that is high-fidelity approximations to the parametric solution, which have the form

$$u_\delta(x, t; \boldsymbol{\mu}) = \mathbf{A}_L(\boldsymbol{\sigma}_{L-1}(\mathbf{A}_{L-1}(\boldsymbol{\sigma}_{L-2}(\cdots \mathbf{A}_1(\boldsymbol{\sigma}_0(\mathbf{A}_0(x))\cdots) \tag{2.8}$$

where each $\boldsymbol{\sigma}_k$ is a (fixed) component-wise activation function and

$$\mathbf{A}_k(\mathbf{z}) = \mathbf{W}_k(t, \boldsymbol{\mu})\mathbf{z} + \mathbf{b}_k(t, \boldsymbol{\mu})\ .$$

Notice how the full deep solutions define an ensemble of deep neural networks, whose weights $\mathbf{W}_k \in \mathbb{R}^{N_{k+1} \times N_k}, \mathbf{b} \in \mathbb{R}^{N_{k+1}}$ depend on the solution parameters $t, \boldsymbol{\mu}$. For sake of simplicity, we consider the case where each activation is either a step-function $\sigma(x) = \mathbb{1}\{x \leq 0\}$ or a ReLU $\sigma(x) = x_+$. Notice that in the model (2.8), both the size $N_\delta = \sum_{k=1}^{L} N_k$ and the depth $L_\delta = L$ depend on the fidelity δ, and are potentially very large. Assume now that the full deep solution

weights belong to low-dimensional spaces, that is

$$\mathbf{W}_k(t,\boldsymbol{\mu}) = \mathbf{U}_k\boldsymbol{\Gamma}_k(t,\boldsymbol{\mu})\mathbf{V}_k^T\,, \qquad \mathbf{b}_k(t,\boldsymbol{\mu}) = \mathbf{U}_k\mathbf{c}_k(t,\boldsymbol{\mu})\,,$$

for some $\mathbf{U}_k \in \mathbb{R}^{N_{k+1}\times M_{k+1}}$, $\mathbf{V}_k \in \mathbb{R}^{N_k\times M_k}$ with orthonormal columns, where

$$M \doteq \sum_{k=1}^{L} M_k \ll \sum_{k=1}^{L} N_k = N_\delta\,.$$

Then, one can write

$$u_\delta(x,t;\boldsymbol{\mu}) = u_M(x,t;\boldsymbol{\mu}) \doteq \mathbf{B}_L(\boldsymbol{\xi}_{L-1}(\mathbf{B}_{L-1}(\boldsymbol{\xi}_{L-2}(\cdots\mathbf{B}_1(\boldsymbol{\xi}_0(\mathbf{B}_0(x))\cdots)$$

where

$$\mathbf{B}_k(\mathbf{z}) = \boldsymbol{\Gamma}_k(t,\boldsymbol{\mu})\mathbf{z} + \mathbf{c}_k(t,\boldsymbol{\mu}) \quad \text{and} \quad \boldsymbol{\xi}_k(\mathbf{z}) = \mathbf{V}_{k+1}^T\boldsymbol{\sigma}(\mathbf{U}_k\mathbf{z})\,.$$

Notice that the functions $\boldsymbol{\xi}_k : \mathbb{R}^{M_k} \to \mathbb{R}^{M_{k+1}}$ do not depend on the parameters $t, \boldsymbol{\mu}$: we refer to them as *reduced activations* (notice that they do not necessarily operate component-wise). The model u_δ thus define a deep-equivalent of the reduced order model previously introduced. Making a parallelism with the results for approximation by (standard) neural networks, one would expect deep reduced models to be more efficient to represent solutions. In fact, this is the case, and the reason lies in the compositional structure of solution $u(x,t;\boldsymbol{\mu}) = u_0(X^{-1}(x,t;\boldsymbol{\mu}))$ to the PDE (2.7). Thanks to the analyticity in the spatial variable x, the transport map $x \mapsto X(x,t;\boldsymbol{\mu})$ can be represented by a (ordinary) reduced model at an exponential (in the number of basis functions) rate. The inverse transport map $x \mapsto X^{-1}(x,t;\boldsymbol{\mu})$ can then be represented by a deep reduced model by implementing the bisection method with a deep network of constant width. Finally, composing with the initial condition u_0 gives a deep reduced model approximation to the solution.

Proposition 2.6. *Assume that the transport map $T_{(t,\boldsymbol{\mu})}$ is analytic and uniformly bounded on the*

closed ρ-Bernstein ellipse [Tre19] for some $\rho > 1$. Then, for any $\epsilon > 0$, there exists a deep reduced order solution u_M of depth $O\ \log \frac{1}{\epsilon}$ and size $M = O\ \log \frac{1}{\epsilon}$ such that

$$\sup_{(t,\boldsymbol{\mu})\in[0,1]\times\mathcal{D}} \|u(\cdot,t;\boldsymbol{\mu}) - u_M(\cdot,t;\boldsymbol{\mu})\|_{\mathbb{V}} \leq \epsilon\ .$$

We refer to [RVBP20] for a fully detailed proof of this result; the proof idea can be extended to other types of transport PDEs, such as Burger's equation. The reduced deep models introduced here are a generalization of Manifold Approximations via Transported Subspaces (MATS) [RPM19], with additional hidden layers, where each layer has a low-rank representation. Reduced deep models are reminiscent of the compression framework for deep networks that is being studied theoretically for improving generalization bounds [NBS17, AGNZ18], or being utilized in practice to accelerate the performance of large networks in practical applications [CWT+15, NPOV15, CWZZ18]. However, the fact that a reduced deep model is a set of networks with a specifically designed degree of freedom, rather than a single network exhibiting low-rank structure in its weights, distinguishes it from the compression frameworks.

Chapter 3

High-dimensional depth-separation for neural networks

3.1 Introduction

The seminal work [Bar93] provides dimension-free quadratic approximation rates by shallow networks under a condition of sparsity of the Fourier transform. Recent works [ES16, Dan17a] suggest that this property is essentially necessary in order to recover polynomial approximation rates, by constructing examples of deep networks which are spread in direction and away from zero in the frequency regime, and by showing that these function can not be efficiently approximated by a shallow counterpart. These depth-separation phenomena occur in the high-dimensional regime, where approximation by neural networks of standard Sobolev spaces is cursed (section 1.3). On the other hand, proofs of such high-dimensional depth-separation phenomena are currently limited to radial functions, that is of the form $f(\mathbf{x}) = \varphi(\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_2)$.

In this chapter we extend the results just cited, further cementing Barron's intuition. We describe rates of approximation by one-hidden-layer networks in terms of the number of units N of the network, by looking at the Fourier representation of the function to be approximated. We con-

sider two types of approximation rate, inspired by the work [SES19]: (i) the rate of approximation is polynomial in both the input dimension d and the error estimation ϵ, that is $N \simeq \mathrm{poly}(d, \epsilon^{-1})$ – we refer to this rate of approximation as *universal* approximation (ii) for any fixed error threshold ϵ, the number of units N needed for approximation of approximation depends at most polynomially on d, that is $N \simeq \mathrm{poly}(d)$ for any fixed error threshold ϵ – we refer to this rate of approximation as *fixed-threshold* approximation. We distinguish two fundamentally different regimes of approximation: relative to a heavy-tailed, unbounded data distribution, or relative to a concentrated distribution. Whereas the former captures the most general setup, the latter is motivated by practical machine learning applications.

First, we consider a class of two-hidden-layer networks exhibiting piece-wise oscillatory behavior, namely functions of the form

$$f_{r,\mathbf{w},\mathbf{v}} : \mathbf{x} \in \mathbb{R}^d \mapsto e^{2\pi i r\,(\mathbf{v}^T\mathbf{x} + \mathbf{w}^T\mathbf{x}_+)}\,.$$

In section 3.2, we show that, under appropriately heavy-tailed data distributions, approximation at a rate $N \simeq \mathrm{poly(d)}$ cannot hold (unconditionally on the weights of the approximant network), as long as the rate of oscillations r grows faster than d. On the other hand, $f_{r,\mathbf{w},\mathbf{v}}$ can be universally approximated (that is, at a rate $\mathrm{poly}(d, \epsilon^{-1})$) by a two-hidden-layer network with any practical activation of choice. The proof of this result (Theorem 3.2) extends the main idea introduced by the results of Eldan and Shamir [ES16] beyond the radial case.

In section 3.3, we show that the $\mathrm{poly}(d)$-oscillatory aspect and the heavy-tailed data distributions are necessary in the depth-separation result mentioned above. More specifically, we show that any deep network, with $O(1)$-bounded weights and $O(1)$-Lipschitz activation, can be fixed-threshold approximated by one-hidden-neural networks over a compact set of radius $O(1)$ (Theorem 3.6). This extends an equivalent result in [SES19], from the class of radial functions to the one of deep neural networks with Hölder activations.

Aforementioned depth separation results consider functions whose Fourier representation is spread in high frequencies. On the other hand, universal approximation results often require the function to be approximated to be, in some sense, sparse in the Fourier domain. Unfortunately, there are currently many gaps between these two types of results, one of them being the definition of approximation domain. In order to reduce the gap between the two results above, we consider approximation on a fixed compact domain, namely the unit sphere $\mathbb{S}^{d-1}$, where Fourier analysis can be done using spherical harmonics. We individuate two conditions on the spherical harmonics decomposition of a function $f \in C(\mathbb{S}^{d-1})$. The first is a sparsity condition on the decomposition, which we show to be sufficient to prove universal approximation (that is, at a rate $N \simeq \mathrm{poly}(d, \epsilon^{-1})$) of f by one-hidden-layer networks. The second is a high-energy spreadness condition on the spherical harmonics decomposition of f, which we show to imply that universal approximation of f by one-hidden-layer networks cannot hold. This is the content of section 3.4, of which the main results are summarized in section 3.4.2.

3.1.1 Neural network approximation rates

We measure the approximation error between two functions $f, g : \Omega \subseteq \mathbb{R}^d \to \mathbb{C}$ in terms of the $L^2(\mu)$ (with respect to a probability measure or density μ) or L^∞ norm. Notice that a L^2 lower bound implies a L^∞ one, and viceversa for an upper bound. The focus of this chapter is to establish upper and lower bounds for approximation of certain function classes by shallow neural networks, in high dimensions d. We distinguish two different approximation regimes of interest.

Definition 3. *We say that a sequence* $f^{(d)} : \Omega_d \subseteq \mathbb{R}^d \to \mathbb{C}\ _{d\geq 2}$ *is* universally approximable *by one-hidden-layer networks with activation σ if it is approximable at a* $\mathrm{poly}(d, \epsilon^{-1})$ *rate; that is if there exists some constants $\alpha > 0$ and $\beta > 0$ such that it holds*

$$\left\| f^{(d)} - f_N \right\|_{\Omega_d,\infty} \leq \epsilon$$

for some one-hidden-layer $f_N \in \mathcal{F}_N^\sigma$ *satisfying* $N + m_\infty(f_N) \leq \alpha(d\epsilon^{-1})^\beta$.

Definition 4. *We say that* $f^{(d)}{}_d$ *is* fixed-threshold approximable *if for any* $\epsilon \in (0,1)$ *it is* ϵ*-approximable at a* $\mathrm{poly}(d)$ *rate; that is if for any* $\epsilon > 0$ *there exists some constants* $\alpha > 0$ *and* $\beta > 0$ *such that for every* $\epsilon > 0$ *it holds*

$$f^{(d)} - f_N \;_{\Omega_d,\infty} \leq \epsilon$$

for some one-hidden-layer $f_N \in \mathcal{F}_N^\sigma$ *satisfying* $N + m_\infty(f_N) \leq \alpha d^\beta$.

These approximation schemes were introduced in [SES19]. To ensure significance of the approximation rates, in the following upper and lower bounds are stated for objective functions $f^{(d)}$ normalized such that $\|f^{(d)}\|_2 \leq 1$ or $\|f^{(d)}\|_\infty \leq 1$.

3.1.2 Activation assumptions

Finally, the results in the next sections generally hold for activations satisfying the following assumptions, which are satisfied by common activation such as the ReLU $\mathrm{ReLU}(x) = x_+$ or the sigmoid $\mathrm{sigmoid}(x) = (1 + e^{-x})^{-1}$ [ES16]. Most of the results can be easily generalized to hold under less strict conditions, but we take these assumptions for sake of simplicity.

Assumption 1. *Given an activation* $\sigma : \mathbb{R} \to \mathbb{R}$*, there exist constants* ι_σ *and* ν_σ *such that*

1. *it is* ι_σ*-Lipschitz and* $\sigma(0) \leq \iota_\sigma$*;*
2. *for any* L*-Lipschitz function* $f : \mathbb{R} \to \mathbb{R}$ *constant outside of an interval* $[-R, R]$ *and any* $\epsilon > 0$ *there exits* $f_N \in \mathcal{F}_N^\sigma$ *with* $\|f - f_N\|_\infty \leq \epsilon$ *such that* $N + w_\infty(f_N) \leq \nu_\sigma LR\epsilon^{-1}$.

Notice that this assumption implies that, given a (deep) neural network f with $\mathrm{poly}(d)$ weights and activations satisfying Assumption 1, then we are always able to replace the activations in f by any other activation satisfying Assumption 1, by paying an at most polynomial cost. This is formalized in the following lemma.

Lemma 3.1. *Let* $f^{(d)} : K_d \subset \mathbb{R}^d \to \mathbb{C}$ $_d$ *be neural networks with activations satisfying Assumption 1 and such that* $N(f^{(d)}) + w_\infty(f^{(d)}) + \mathrm{diam}(K^{(d)}) \leq \mathrm{poly}(d)$*; also let* σ *be any activation function satisfying Assumption 1. Then the sequence* $f^{(d)}$ $_d$ *is universally approximable by one-hidden-layer networks with activation* σ*.*

3.2 A depth separation example

Our starting point for the study of depth-separation is to consider a generic data distribution μ with adversarial properties against shallow approximations. In the seminal work [ES16], Eldan and Shamir establish an unconditional (with no restrictions on the norms of the weights of the network) depth-separation result by considering a density μ in $\mathbb{R}^d$ with tails $\mu(\|\mathbf{x}\|_2) \simeq \|\mathbf{x}\|_2^{-(d+1)/2}$ and a radial function $f^{(d)}(\mathbf{x}) = h_d(\|\mathbf{x}\|_2)$ with $h_d : \mathbb{R} \to \mathbb{R}$ a carefully chosen oscillating function with compact support. The proof in [ES16] reveals the limitations of shallow neural networks at approximating high-dimensional functions via a powerful harmonic analysis insight, that is particularly convenient in the setting of radial functions; see section 1.5. In this section, we show that their proof strategy can be extended to include more diverse function classes, namely those arising naturally from ReLU networks. Specifically, we consider networks of the form

$$f_{r,\mathbf{w},\mathbf{v}} : \mathbf{x} \in \mathbb{R}^d \mapsto \sigma_r\ \mathbf{v}^T\mathbf{x} + \mathbf{w}^T\mathbf{x}_+ \tag{3.1}$$

where $\mathbf{x}_+$ denotes the element-wise ReLU activation, $\mathbf{v},\ \mathbf{w} \in \mathbb{R}^d$ and $\sigma_r(t) = e^{2\pi i r t}$. We are thus considering a function which is piece-wise oscillatory, with constant envelope $|f_{r,\mathbf{w},\mathbf{v}}(\mathbf{x})| = 1$, and where the frequency of oscillations is controlled by r. The main result of this section can be summarized as follows.

Theorem 3.2 (Informal)**.** *Assume that* $\|\mathbf{w}\|_2, \|\mathbf{v}\|_2 = \Theta(1)$ *and that* $r = \Theta(d^k)$ *for some* $k \geq 2$*. Then there exists a (low-decay) product measure* μ *on* $\mathbb{R}^d$ *such that the function* $f_{r,\mathbf{w},\mathbf{v}}$ *is universally*

approximable by two-hidden-layer networks but it is not fixed-threshold approximable by one-hidden-layer networks.

3.2.1 The lower bound

Let $\psi \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ with $\|\psi\|_2 = 1$, and such that its Fourier transform $\hat{\psi}$ is compactly supported in $[-K, K]$, for some $K > 0$. Assume also that

$$\|\psi\|_1 < \sqrt{2/K}\,. \tag{3.2}$$

The condition ensure that the density ψ is sufficiently spread away from zero (see Remark 3). Our first objective is to establish depth separation for the approximation of $f_{r,\mathbf{w},\mathbf{v}}$ under the L^2 metric defined by the probability density φ^2, where $\varphi : \mathbf{x} \in \mathbb{R}^d \mapsto \prod_{j=1}^{d} \psi(x_j)$.

Theorem 3.3. *Let $f^{(d)} = f_{r_d,\mathbf{w}_d,\mathbf{v}_d}$, for some $r_d \in \mathbb{R}$, $\mathbf{w}_d, \mathbf{v}_d \in \mathbb{R}^d$. For a fixed $\gamma > 0$, define*

$$\tau_d \doteq \sup_{S \subseteq [d]} \|\mathbf{v}_d + \mathbf{w}_{d,S}\|_\infty\,, \quad \Omega_d \doteq \{j \in [d] : r_d |w_{d,j}| \geq \gamma d^2\} \quad \text{and} \quad \eta_d \doteq \frac{|\Omega_d|}{d}\,,$$

where $\mathbf{w}_{d,S} \in \mathbb{R}^d$ is defined by $w_{d,S,i} = w_i \mathbb{1}\{i \in S\}$. Assume that

(i) oscillations grow polynomially, that is $\tau_d \cdot r_d = \Theta(d^k)$ for some constant $k > 0$;

(ii) the vectors $\mathbf{w}_d$ are sufficiently spread, that is $\eta_d \geq \eta$ for some $\eta > 0$ independent of d;

(iii) the density φ^2 is sufficiently spread, i.e. $2K\|\psi\|_1^2 < 2^{2\eta}$.

Then there exists a constant $\alpha \in (0, 1)$ (independent of d) such that

$$\inf_{f_N \in \mathcal{F}_N} \|f^{(d)} - f_N\|_{\varphi^2,2}^2 \geq 1 - N \cdot \alpha^d \cdot O(d^{k+1})\,. \tag{3.3}$$

Notice that this lower bound is unconditional on the weights of the neurons $m_\infty(f_N)$.

The proof follows a similar strategy as in the work [ES16]. The approximation error can be expressed in the Fourier domain as

$$\|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} - f_N\|_{\varphi^2,2}^2 = \|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} \cdot \varphi - f_N \cdot \varphi\|_2^2 = \|\hat{f}_{r_d,\mathbf{w}_d,\mathbf{v}_d} * \hat{\varphi} - \hat{f_N} * \hat{\varphi}\|_2^2 \,.$$

Thanks to the assumptions, the target function $f_{r_d,\mathbf{w}_d,\mathbf{v}_d}$ satisfies a key property, namely that its Fourier transform has its energy sufficiently spread in the high-frequencies, after the convolution by $\hat{\varphi}$. Such frequency spread is caused by the shattering of the first ReLU layer, which effectively creates $\Theta(2^{\eta d})$ different frequencies. The piece-wise structure arising from the ReLU can be handled in the Fourier domain by the Hilbert transform of the function ψ, which has sufficient decay thanks to the assumptions. Noticing that $\|\hat{f}_{r_d,\mathbf{w}_d,\mathbf{v}_d} * \hat{\varphi}\|_2 = 1$, this is formalized in the following.

Lemma 3.4 (Informal)**.** *It holds that*

$$\hat{f}_{r_d,\mathbf{w}_d,\mathbf{v}_d} * \hat{\varphi} \;\; (\boldsymbol{\xi}) \;\lesssim 2^{-\eta d}\|\varphi\|_1\|\boldsymbol{\xi}\|_\infty^{-1} \qquad \textit{for } \boldsymbol{\xi} \gtrsim \mathrm{poly}(d) \,.$$

On the other hand, since $\hat{\varphi}$ is compactly supported and the Fourier transform of a single-unit network is localised in a frequency ray, the Fourier transform of $f_{r_d,\mathbf{w}_d,\mathbf{v}_d} \cdot \varphi$ is localised in a union of N tubes, of the form $T_{\boldsymbol{\alpha}} = \mathrm{span}(\{\boldsymbol{\alpha}\}) + [-K, K]^d$. This implies that

$$\inf_{f_N \in \mathcal{F}_N} \|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} - f_N\|_{\varphi^2,2}^2 \geq \inf_{f_N \in \mathcal{T}_{(N)}} \|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} - f_N\|_{\varphi^2,2}^2$$

where $\mathcal{T}_{(N)}$ denotes the set of L^2 functions such that their Fourier transform is supported on the union of N tubes $T_{\boldsymbol{\alpha}_1}, \ldots, T_{\boldsymbol{\alpha}_N}$ as above, for some arbitrary $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N \in \mathbb{R}^d$. Thanks to Plancherel's identity, and since $\|f_{r_d,\mathbf{w}_d,\mathbf{v}_d}\|_{\varphi^2,2} = 1$, it further holds that

$$\inf_{f_N \in \mathcal{T}_{(N)}} \|f_{r_d,\mathbf{w}_d,\mathbf{v}_d} - f_N\|_{\varphi^2,2}^2 \geq 1 - N \cdot \sup_{\boldsymbol{\alpha} \in \mathbb{S}^{d-1}} \left\| \mathbb{1}_{T_{\boldsymbol{\alpha}}} \cdot \hat{f}_{r_d,\mathbf{w}_d,\mathbf{v}_d} * \hat{\varphi} \right\|_2^2 ,$$

where $\mathbb{1}_{T_\alpha}$ denotes the indicator function of T_α. Lemma 3.4 can then be used to show that such projections are exponentially (in d) small, which implies equation (B.1). The detailed proof is deferred to section B.1.1.

Remark 2. Theorem 3.3 asks for two main conditions to hold. First, the magnitude of oscillations of the objective function (parametrised by r_d) must grow at least polynomially with d, similarly to the assumptions in the works [ES16] and [Dan17a]. Second, the data distribution μ with density φ^2 should be heavy-tailed, in order for its Fourier transform to be sufficiently localised. When r_d does not grow fast enough with d, the energy starts piling up at the low frequencies, creating an important roadblock to establish approximation lower-bounds, and leaving open the possibility of efficient shallow approximation. Similarly, when μ concentrates too quickly, the proof strategy also fails, due to the fact that in that case $\hat{\varphi}$ is too spread in the Fourier domain, creating full overlap of the energies.

Remark 3. The admissibility condition (3.2) is necessary since $\eta \leq 1$ by definition. Notice that

$$1 = \|\psi\|_2^2 = \|\hat{\psi}\|_2^2 \leq (2K)\|\hat{\psi}\|_\infty^2 \leq (2K)\|\psi\|_1^2$$

and therefore condition (3.2) can be considered as a requirement on the Fourier transform of ψ not being too concentrated in the origin. The choice $\psi(t) = \sqrt{3/2}\,\mathrm{sinc}^2(\pi t)$ corresponds to $K = 1$, $\|\psi\|_1 = \sqrt{3/2}$ and $\|\psi\|_2 = 1$, which verifies (3.2). In that case, from condition (ii) we need $\eta > \frac{\log_2 3}{2} \approx 0.79$. However, the choice $\psi(t) = C\mathrm{sinc}(\pi t)$ (the equivalent separable version of the of density considered in [ES16]) is not admissible, since ψ is not in L^1. The lower bound is optimized by finding compactly supported windows with an optimal L^1 to L^2 ratio of their Fourier transforms.

Remark 4. The theorem considers a separable ReLU transform $\mathbf{x} \mapsto \mathbf{x}_+$, combined with a separable data distribution μ with density φ^2. One could expect a similar lower bound to apply in the more general case of a layer of the form $\mathbf{x} \mapsto (\mathbf{U}\mathbf{x} + \mathbf{b})_+$, $\mathbf{U} \in \mathbb{R}^{d' \times d}, \mathbf{b} \in \mathbb{R}^{d'}$. Such general

case replaces the Hilbert transform of ψ with the Fourier transform of indicators of convex polytopes, which has been used in the context of ReLU networks to characterize spectral properties [RBA+19].

Example 1. We give an explicit example of a family of function $f^{(d)} : \mathbb{R}^d \to \mathbb{R}$ which satisfy the assumptions of Theorem 3.3. Consider the functions

$$f^{(d)}(\mathbf{x}) = \exp\left(2\pi i d^2 \sum_{k=1}^{d} \max\{0, x_k\}\right) .$$

Then, if μ_d is the product probability measure defined by the density in Remark 3, that is

$$\mu_d(d\mathbf{x}) = \prod_{k=1}^{d} \left[\frac{3}{2}\mathrm{sinc}^4(\pi x_k)\, dx_k\right] ,$$

then it holds that

$$\inf_{f_N \in \mathcal{F}_N} \left\| f_N(\mathbf{x}) - f^{(d)}(\mathbf{x}) \right\|^2_{\mu_d,2} \geq 1 - 1300N \cdot d^2 \cdot (0.75)^d .$$

For example, this implies that

$$\inf_{f_N \in \mathcal{F}_N} \left\| f_N(\mathbf{x}) - f^{(d)}(\mathbf{x}) \right\|_{\mu_d,2} \geq \frac{1}{2}$$

unless

$$N \geq \frac{1.3^d}{10^4 d^3} .$$

The numbers are obtained by explicitly tracking the constant in the proof of Theorem 3.3 (see section B.1.1 for more details).

3.2.2 The upper bound

According to the definition of neural networks we gave in section 1.2, the function $f_{r,\mathbf{w},\mathbf{v}}$ is naturally a two-hidden-layer neural network. Although, while there are cases of sinusoidal activations being used in practice, activations such as ReLU or sigmoid are more relevant to practical applications. The following theorem, proved in section B.1.2, shows that we can efficiently represent the function $f_{r,\mathbf{w},\mathbf{v}}$ in the hypothesis of the Theorem 3.3 as a two-hidden-layer neural network with fixed activation, such as the ReLU or the sigmoid. The main technical difference with Lemma 3.1 is that the result is proved for approximation w.r.t. the probability measure with density φ^2 introduced above.

Theorem 3.5. *Let σ be an activation satisfying Assumption 1. Assume that there exists a constant $k \geq 1$ such that $m_\infty(f_{r_d,\mathbf{v}_d,\mathbf{w}_d}) \leq O(d^k)$ and assume that ψ is such that $|\psi(x)| = O(|x|^{-1})$. Then, for every $\epsilon > 0$, there exists $f_N \in \mathcal{F}_N^\sigma$ with*

$$N + m_\infty(f_N) \leq O\ d^{2(1+k)}\epsilon^{-3/2} \qquad \textit{such that} \quad \|f_N - f_{r_d,\mathbf{w}_d,\mathbf{v}_d}\|^2_{\varphi^2,2} \leq \epsilon\ .$$

Theorems 3.3 and 3.5 therefore estabilish a depth separation result. If $f^{(d)} = f_{r_d,\mathbf{w}_d,\mathbf{v}_d}$ are defined with $r_d, \mathbf{w}_d, \mathbf{v}_d$ satisfying the assumptions of both theorems (that is, they satisfy assumptions (i)-(ii)-(iii) of Theorem 3.3 with $\tau_d \cdot r_d = \Theta(d^k)$), then Theorem 3.3 says that $f^{(d)}{}_d$ is not fixed-threshold approximable by one-hidden-layer networks, while Theorem 3.5 says that the sequence is universally approximable by two-hidden-layer networks with a fixed activation satisfying Assumption 1. For example, the family of functions considered in Example 1 satisfies such assumptions.

We thus identify two key aspects responsible for such depth separation: heavy-tailed data and oscillations growing with dimension. In the next sections we want to understand how necessary these two conditions are. The next section shows that if these two condition do not hold anymore, then a lower bound such as the one in Theorem 3.3 is not achievable; more specifically we show

that the objective function is fixed-threshold approximable by one-hidden-layer networks.

3.3 Approximation of deep networks by shallow ones

In this section, we show that any deep neural network f (which include the target functions considered in the previous section) can be approximated by shallow ones at a rate which is polynomial in d, as long as the rate of oscillation in the inner layers of f is constant in d and the metric is concentrated in a ball of constant radius. We start by reporting the result in a general form for two-hidden-layer networks and we discuss some consequences and extensions afterwards.

Consider a family of two-hidden-layers neural network $\{f^{(d)} : K_d \subset \mathbb{R}^d \to \mathbb{C}\}$ of the form

$$f^{(d)} : \mathbf{x} \in \mathbb{R}^d \mapsto \boldsymbol{\gamma}_d^T \mathbf{g}\ \mathbf{W}_d^T \mathbf{h}\ \mathbf{U}_d^T \mathbf{x} \quad \in \mathbb{C}\,, \tag{3.4}$$

where $\mathbf{h} = \mathbf{h}^{(d)} : \mathbb{R}^{p_d} \to \mathbb{R}^{p_d}$ and $\mathbf{g} = \mathbf{g}^{(d)} : \mathbb{R}^{o_d} \to \mathbb{R}^{o_d}$ are, respectively, component-wise 1-Lipschitz and $(1, \alpha)$-Holder[1] activation functions, and $\mathbf{U}_d \in \mathbb{R}^{d \times p_d}$, $\mathbf{W}_d \in \mathbb{R}^{p_d \times o_d}$, $\boldsymbol{\gamma}_d \in \mathbb{C}^{o_d}$. We wish to approximate $f^{(d)}$ by one-hidden-layer neural networks with a given activation.

Theorem 3.6. *Assume that* $\operatorname{diam}(K_d) = O(1)$ *and that the networks* $f^{(d)}$ *have* ℓ^1 *bounded weights, that is* $m_1(f^{(d)}) = O(1)$. *Then, for every activation* σ *satisfying Assumption 1.2 and every* $\epsilon \in (0, 1)$ *it holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f^{(d)} - f_N^\sigma\|_{K,\infty} \leq \epsilon \quad \textit{for some } N \leq \exp\ O\ \epsilon^{-1-2/\alpha} \log(p_d/\epsilon) \quad . \tag{3.5}$$

Moreover, it is possible to choose f_N^σ *attaining* (3.5) *with* $m_\infty(f_N^\sigma)$ *satisfying a bound similar to the one on* N, *for example* $m_\infty(f_N^\sigma) \leq (1 + N^2)$.

The proof is constructive and based on the following observation. Consider the case where $o_d = 1$, $\gamma_d = 1$, $p_d = p$ and $g(x) = x^r$ some positive integer r. If $h_k(x) = e^{ix}$ for all $k \in [p]$, then

[1] We say that a function $g : \mathbb{R} \to \mathbb{R}$ is $(1, \alpha)$-Holder if it holds that $|g(x) - g(y)| \leq |x - y|^\alpha$ for all $x, y \in \mathbb{R}$.

the function $f = f^{(d)}$ at (3.4) has form

$$f(\mathbf{x}) = \left(\sum_{k=1}^{N} w_k e^{i\mathbf{u}_k^T \mathbf{x}} \right)^r$$

for some $w \in \mathbb{R}^N$, $\mathbf{u}_k \in \mathbb{R}^d$, where $N = p$. By expanding the power we can write

$$f(\mathbf{x}) = \sum_{j_1 + \cdots + j_N = r} \binom{r}{j_1 \cdots j_N} w_1^{j_1} \cdots w_N^{j_N} \; e^{i\left(\sum_{h=1}^{N} j_h \mathbf{w}_h\right)^T \mathbf{x}},$$

that is a formulation of f as a one-hidden-layer network with activation $\sigma_1(t) = e^{2\pi i t}$ (in the following we refer to this type of networks as *shallow Fourier networks*) and a number of units that scales as N^r. Since both polynomials and trigonometric polynomials are universal approximators, with well known convergence rates, in the general case one can proceed as follows. Each of the non-linearities applied to the first hidden layer can be approximated by a trigonometric polynomial at a polynomial rate on the interval of interest. Similarly, every non-linearity applied to the second hidden layer can be approximated by a polynomial at a linear (in the degree of the polynomial) rate on the interval of interest. Assuming for simplicity that both rates behave as ϵ^{-1}, where $\epsilon > 0$ denotes the approximation error, the composition of the two approximation following the structure of the target network results in a shallow Fourier network (that is with activation $\sigma_1(t) = e^{2\pi i t}$) whose size N behaves, roughly speaking, as

$$N \simeq \Theta \left(p\epsilon^{-2} \right)^{-1}.$$

Moreover, it is also possible to control the value of the coefficients appearing in the final approximation. With this, we can approximate each summand in the shallow Fourier network by a one-hidden-layer network with activation σ with a controlled number of units, thanks to Assumption 1.2. A more detailed statement and a formal proof are reported in section 3.3.

In essence, in the Theorem 3.6, we show that it is possible to approximate a two-hidden-layer

neural network with $\mathrm{constant}(d)$ oscillations at a $\mathrm{poly}(d)$ rate over a compact set of $\mathrm{constant}(d)$ radius. On the other hand, it easy to show that it is also possible to obtain approximation at a $\mathrm{poly}(\epsilon^{-1})$ rate (see section B.3.2), for fixed d. Finally, existing results in the literature (see [SES19]) show that universal approximation is not possible, the counterexample being essentially a radial function.

Interestingly, the upper bound in Theorem 3.6 does not depend on the number of units in the second layer of the objective function. This parameter is *hidden* in the control we impose on the ℓ^1 norm of the objective weights. The proof technique of this upper bound highlights how the difficulty of approximating at $\mathrm{poly}(d, \epsilon^{-1})$ rate stems from the high-energy of the second layer, which requires the shallow network used for approximation to have a (potentially) exponential (in d) number of directions. Notice that the lower bound in Theorem 3.3 actually tells that the function is not fixed-threshold approximable. High oscillations in the lower bound (3.3) essentially ensure that an exponential (in d) number of neurons are necessary. An open question is then whether a low-decaying measure is, in general, necessary for such a result to hold.

Expanding on the proof technique above, it is possible to extend the result of Theorem 3.6 to approximation of L-hidden-layers networks by shallow ones, which gives a rate scaling as $\exp(O(\epsilon^{-L}\log(p/\epsilon)))$.

Theorem 3.7. *Let $f^{(d)}$ as in (1.1), with $O(1)$-Lipschitz activations, first hidden layer width $d_1 = p_d$, depth $L_d = L$ and bounded weights, that is $m_1(f^{(d)}) = O(1)$. Then for every $\epsilon > 0$ there exists a shallow Fourier network $f_N \in \mathcal{F}_N^\sigma$ with*

$$N \leq \quad p_d \cdot O\left(1 + \frac{1}{\epsilon^2}\right)^{O(L)\left(1+\frac{1}{\epsilon}\right)^{L-1}} \qquad \textit{such that} \qquad \left\| f^{(d)} - f_N \right\|_{B^d_{1,\infty},\infty} \leq \epsilon \,.$$

See section B.3.1 for a formal statement and its proof. While it has been shown that generic $O(1)$-Lipschitz function can not be (computably) represented by neural networks with $N \simeq \mathrm{poly}(d)$ units [VRPS21], an interesting related follow-up conjecture is whether our result can be general-

ized to any generic $O(1)$-Lipschitz function which is $\mathrm{poly}(d)$-computable. Notice that this is dependent on the choice of the uniform norm to measure the approximation error. For example, it has been shown that a rate $N \simeq \mathrm{poly}(d)$ is achievable for approximation in the L^2 norm with the uniform measure [HSSVG21].

Finally, notice that the approximation rate shown in Theorem 3.6 and Theorem 3.7 are actually polynomial in the size p_d of the first hidden layer of $f^{(d)}$ rather than in the input dimension d. Although, up to choosing a worse (yet constant) exponent in ϵ, we can replace p_d by d in the statement, by considering the function as a $(L+1)$-hidden-layer network, where the first layer is the identity.

3.3.1 Two cases of interest

Theorem 3.6 allows to recover, for any fixed threshold $\epsilon > 0$, a $\mathrm{poly}(d)$ rate for the approximation of $f_{r,\mathbf{w},\mathbf{v}}$ by one-hidden-layer networks and it can be seen as a generalization of Theorem 1 in [SES19]. This is the content of the following corollaries.

Corollary 3.8 (Radial functions)**.** *Let $f^{(d)}(\mathbf{x}) = \varphi_d(\|\mathbf{x}\|_2)$, where $\varphi_d : [-1,1] \to \mathbb{R}$ are 1-Lipschitz, and $K_d = B^d_{1,2}$. Then, for any $\epsilon \in (0,1)$ it holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \left\| f_N^\sigma - f^{(d)} \right\|_{K_d,\infty} \leq \epsilon \quad \text{for some } N \leq \exp\left(O\left(\epsilon^{-5}\log(d/\epsilon)\right)\right) .$$

Moreover, f_N^σ can be chosen so that $m_\infty(f_N^\sigma) \leq \exp(O(\epsilon^{-5}\log(d/\epsilon)))$.

Consider the functions $f^{(d)} : \mathbf{x} \in \mathbb{R}^d \mapsto e^{i\mathbf{w}_d^T(\mathbf{U}_d\mathbf{x})_+}$ for some $\mathbf{w}_d \in \mathbb{R}^{p_d}$, $\mathbf{U}_d \in \mathbb{R}^{p_d \times d}$. This is a more general version of the function $f_{r,\mathbf{w},\mathbf{v}}$ considered in section 3.2. If the weights are bounded, that is $m_1(f^{(d)}) = O(1)$, then Theorem 3.6 implies the following.

Corollary 3.9 (Shallow approximation of (3.1))**.** *If $r_d = O(1)$ and $K_d = B^d_{r_d,2}$, for any $\epsilon \in (0,1)$*

it holds that

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} D_\infty \|f_N^\sigma - f^{(d)}\|_{K_d,\infty} \leq \epsilon \quad \textit{for some } N \leq \exp\ O\ \epsilon^{-2} \log(p_d/\epsilon) \quad .$$

Moreover, f_N^σ can be chosen so that $m_\infty(f_N^\sigma) \leq \exp(O(\epsilon^{-2}\log(p_d/\epsilon)))$.

Although the result of Corollary 3.9 is established for approximation in the uniform norm over the unit ball, it is not difficult to extend it to a result in L^2 over a measure that concentrated over a compact set of constant (in d) radius, such as a normalized Gaussian. A formal statement of this fact, along with the proof, is reported in section B.3. Compared with the result of section 3.2, Corollary 3.9 implies the following. The function $f_{\mathbf{w},\mathbf{U}}$ can be approximated, at a $\mathrm{poly}(d)$ rate over a compact set of constant radius if its weights have constant norm. On the other hand, if the norm of the weights grows polynomially in d, then approximation at a $\mathrm{poly}(d)$ rate is not possible, under a polynomially slow decaying measure. An open question is whether approximation at a $\mathrm{poly}(d)$ rate is possible if only one of these two conditions hold.

3.4 Approximation by shallow networks: a spherical harmonics analysis

As already discussed, difficulties in approximating functions in high dimension by shallow networks appear when the function has a Fourier transform spread in a (exponential) number of directions in (polynomial) high energy. On the other hand, the presence of only one of these two conditions is not enough to prevent efficient approximability. While the previous results highlight this, the lower bound presented in Theorem 3.3 applies to a specific choice of error measure, with (polynomially) slowly decaying tails.

In this section, we aim to disentagle the role of the measure tail and understand how the Fourier representation can tell whether a function is efficiently approximable by a one-hidden-layer net-

work or not. In particular, we focus on approximation results for functions defined over the $(d-1)$-dimensional sphere $\mathbb{S}^{d-1}$, for which a rich literature of Fourier analysis is available.

First, we give a sufficient condition on the target function in terms of its spherical harmonics decomposition to be not efficiently approximable by shallow one-hidden-layer networks. This condition captures a slowly decaying and sufficiently spread spherical harmonic expansion. We also show that certain symmetry properties imply this condition. On the other hand, one may ask if a reverse statement holds. In this direction, building on existing theory, we provide a sufficient condition for approximation by one-hidden-layer networks.

3.4.1 Spherical harmonics decomposition

Let $d \geq 2$ and S^{d-1} (S when the dimension is clear from the context) be the uniform measure over $\mathbb{S}^{d-1}$. The spherical harmonics are a particular orthonormal basis for $L^2(S)$. They consists of

$$\bigcup_{k=0}^{\infty} \operatorname{span} \quad Y_{k,i}^d \quad {}_{i=1}^{N_k^d} \quad = \bigcup_{k=0}^{\infty} H_k^d$$

where $Y_{k,i}^d$ is a restriction to $\mathbb{S}^{d-1}$ of an homogeneous harmonic polynomial of degree k. The projection operator over H_k^d is given by

$$\mathcal{P}_k^d : f \in L^2(S) \mapsto f_k \doteq \sum_{i=1}^{N_k^d} \langle f, Y_{k,i}^d \rangle Y_{k,i}^d \, .$$

Similarly, $\mathcal{P}_I$ denotes the operator $\oplus_{i \in I} \mathcal{P}_i^d$, for any $I \subseteq \mathbb{N}$. The function f_k is referred to as the degree k spherical harmonic component of the function f. Since the spherical harmonic form an orthonormal basis of L_S^2, it holds that $f = \sum_{k=0}^{\infty} f_k$ and $\|f\|_2^2 = \sum_{k=0}^{\infty} \|f_k\|_2^2$ for every $f \in L^2(S)$, where $\|\cdot\|_2$ denotes the norm in $L^2(S)$. As spherical harmonics decomposition can be seen as a generalization of Fourier series to dimensions $d \geq 3$, in the following we refer to the spherical harmonics decomposition of a function as its Fourier representation, interchangeably. The operator

$\mathcal{P}_k$ can be associated with a kernel given by

$$\sum_{i=1}^{N_k^d} Y_{k,i}^d(\mathbf{x})\overline{Y_{k,i}^d(\mathbf{y})} = N_k^d P_k^d\ \mathbf{x}^T\mathbf{y}$$

where

$$N_k^d = \frac{(2k+d-2)(k+d-3)!}{k!(d-2)!} = \Theta\ \ \sqrt{\frac{k+d}{kd}}\frac{(k+d)^{k+d}}{k^k d^d}\frac{d^2}{(k+d)^2}$$

is the dimension of H_k^d and P_k^d is the $((d-2)/2)$-Gegenbauer polynomial defined as

$$P_k^d(x) = k!\,\Gamma\ \ \frac{d-1}{2}\ \ \sum_{j=0}^{\lfloor k/2\rfloor}(-1)^j\frac{(1-x^2)^j x^{k-2j}}{4^j j!(k-2j)!\Gamma\ \ j+\frac{d-1}{2}}\ .$$

Let ω_d be the Lebesgue area of the sphere:

$$\omega_d = \omega_{d-1}\frac{\sqrt{\pi}\,\Gamma\ \frac{d-1}{2}}{\Gamma\ \frac{d}{2}} = \frac{2\pi^{d/2}}{\Gamma\ \frac{d}{2}} = \Theta\ \ \frac{(2\pi e)^{d/2}}{d^{d/2-1/2}}\ \ = \Theta\ \ \sqrt{d}\ \ \frac{2\pi e}{d}\ \ ^{d/2}\ \ .$$

The polynomials $\{(N_k^d)^{1/2}P_k^d\}_{k\geq 0}$ form a basis of orthonormal polynomials for $L^2(\mu_d)$, where μ_d is the probability measure on $[-1,1]$ defined by

$$d\mu_d(t) = \alpha_d(1-t^2)^{(d-3)/2}\,dt\ ,$$

where $\alpha_d = \omega_{d-1}/\omega_d = \Theta(\sqrt{d})$. Notice that, given a function $f \in L^2(S)$, it holds

$$f_k(\mathbf{x}) = N_k^d\ \ _{\mathbb{S}^{d-1}} f(\mathbf{y})P_k^d\ \mathbf{x}^T\mathbf{y}\ \ dS(\mathbf{y})\ .$$

Moreover, if the function f only depends on a linear projection of the input, the Funk-Hecke formula holds.

Theorem 3.10 (Funk-Hecke formula)**.** *For every* $\sigma : [-1,1] \to \mathbb{C}$ *such that* $\mathbf{x} \in \mathbb{S}^{d-1} \mapsto \sigma(x_1)$ *is*

in $L^2(S)$*, and for every* $\mathbf{w} \in \mathbb{S}^{d-1}$*, it holds that*

$$\int_{\mathbb{S}^{d-1}} \sigma(\mathbf{w}^T\mathbf{x}) P_k^d(\boldsymbol{\xi}^T\mathbf{x})\, dS(\mathbf{x}) = \lambda_k P_k^d(\boldsymbol{\xi}^T\mathbf{w})$$

where $\lambda_k = \langle \sigma, P_k^d \rangle_{\mu_d}$.

Functions of the form

$$\mathbf{x} \in \mathbb{S}^{d-1} \mapsto \alpha P_k^d(\mathbf{w}^T\mathbf{x})$$

for some $\alpha \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{S}^{d-1}$, are called zonal harmonics. By the Funk-Hecke formula it follows that

$$\int_{\mathbb{S}^{d-1}} P_k^d(\mathbf{w}^T\mathbf{x}) P_k^d(\mathbf{v}^T\mathbf{x})\, dS(\mathbf{x}) = (N_k^d)^{-1} P_k^d(\mathbf{w}^T\mathbf{v})$$

for any $\mathbf{w}, \mathbf{v} \in \mathbb{S}^{d-1}$. This implies that H_k^d has an RKHS structure with kernel K given by

$$K(\mathbf{v}, \mathbf{w}) \doteq N_k^d P_k^d(\mathbf{v}^T\mathbf{w}) \,.$$

In particular, zonal harmonics actually span H_k^d. Moreover, it can be shown that there exists $\mathbf{w}_1, \ldots, \mathbf{w}_{N_k^d} \in \mathbb{S}^{d-1}$ such that $H_k^d = \operatorname{span}(\ P_k^d(\mathbf{w}_i^T \cdot) \ _{i=1}^{N_k^d})$ [EF14, Theorem 4.13]. For these facts and more details about spherical harmonics we refer to the books [AH12, DX13].

3.4.2 Concentration and spreadness in H_k^d and main results

Intuitively, one can say function $f \in C(\mathbb{S}^{d-1})$ is concentrated over $\mathbb{S}^{d-1}$ if there is an area $\Omega \subset \mathbb{S}^{d-1}$ such that the mass of f is concentrated over Ω. On the other hand one could say that f is spread if it assumes non-negligible values uniformly over the sphere. The spreadness/concentration of the function f can be quantified by looking at ratios of the type

$$\ell_{q,p}(f) \doteq \frac{\|f\|_q}{\|f\|_p}$$

for $1 \leq p < q \leq \infty$. Since the norms above are with respect to a probability measure, it holds that $\ell_{q,p} \geq 1$. Intuitively, the closest this ratio is to 1, the more spread is the function. On the other hand, the largest this ratio, the more concentrated the function is. Consider the case of a function $f_k \in H_k^d$. Then, it holds that

$$\ell_{\infty,2}(f_k) \leq \sqrt{N_k^d}$$

The equality is attained for functions of the type $f_k(\mathbf{x}) = \alpha P_k^d(\mathbf{w}^T\mathbf{x})$ for some $\alpha \in \mathbb{C}$ and $\mathbf{w} \in \mathbb{S}^{d-1}$, i.e. zonal harmonics. In this sense, zonal harmonics could be considered as the most concentrated functions in H_k^d. A similar inequality can be shown for the quantity $\ell_{2,1}$: it holds that

$$\ell_{2,1}(f_k) \leq \sqrt{N_k^d} \tag{3.6}$$

for $f_k \in H_k^d$. Nevertheless, in this case, zonal harmonics do not attain equality; the inequality is actually not tight; a more detail discussion on this quantity is reported in section 3.4.4.

Thanks to the Funk-Hecke formula, it holds that a one-hidden-layer $f_N \in \mathcal{F}_N$, with hidden layer weights given by $\mathbf{w}_1, \ldots, \mathbf{w}_N$, satisfies

$$\mathcal{P}_k^d f_N = \sum_{j=1}^{d} \alpha_j P_k^d(\mathbf{w}_j^T\mathbf{x})$$

for some $\boldsymbol{\alpha} \in \mathbb{C}^N$. In other words, its Fourier representation is concentrated along N directions. According to the remarks above, this implies that if the width N is relatively small, the Fourier components of the neural network f_N are relatively concentrated in space. One would then expect that such concentration can be used to determine whether a function can be approximated efficiently by a one-hidden-layer neural network or not. In the next sections, we show that this is indeed the case. Let $f \in C(\mathbb{S}^{d-1})$; assuming that $\|f_k^{(d)}\|_2 \simeq \mathrm{poly}(d, k^{-1})$, the results can be informally summarized as follows:

- If the spherical components of f are (exponentially) spread in $\ell_{\infty,2}$ sense, that is, for example,

$$\ell_{\infty,2}(f_k) \lesssim \epsilon^k \cdot \sqrt{N_k^d} = \epsilon^k \cdot \sup_{g \in H_k^d} \ell_{\infty,2}(g) \quad \text{for some } \epsilon \in (0,1)$$

  then f is provably not universally approximable by one-hidden-layer networks.

- If the spherical components of f are (polynomially) concentrated in $\ell_{2,1}$ sense, that is, for example,

$$\ell_{2,1}(f_k) \gtrsim \text{poly}(d^{-1}, k^{-1})\sqrt{N_k^d}$$

  then f is universally approximable by one-hidden-layer networks.

Notice that, on the other hand, if $\|f_k\|_2$ decreases exponentially fast then universal approximation follows, and similarly if $\|f_k\|_2$ decreases exponentially slowly then universal approximation can not hold. The first of the two conditions above expresses concentration of the Fourier decomposition, while the second expresses spreadness of the same. We notice at least two gaps between the two conditions. The first one is the expression of the concentration phenomena: one is with respect to $\ell_{\infty,2}$, while the other one is with respect to $\ell_{2,1}$. Second, the two regimes above do not include many other possible ones. For example, we suspect the existence of a regime which prevents universal approximability but allows for fixed-threshold one, a topic worth of future study. These results are properly formalized, stated and discussed in section 3.4.3 and section 3.4.4, respectively.

3.4.3 Inapproximability of functions with spread Fourier representation

As discussed above, one-hidden-layer functions have a *zonal* structure. In more detail, if $h(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b)$ for some $\mathbf{w} \in \mathbb{S}^{d-1}$ and $b \in \mathbb{R}$, then it is easy to see that

$$h_k(\mathbf{x}) = s_k \|h_k\|_2 \sqrt{N_k^d} P_k^d(\mathbf{w}^T\mathbf{x})$$

with $s_k \in \{\pm 1\}$. In particular, it follows that $\|h_k\|_\infty = |h_k(\pm \mathbf{w})| = (N_k^d)^{1/2}\|h_k\|_2$. This can be interpreted by saying that the Fourier components of single neurons are most concentrated (along the neuron direction) in space. Therefore, it is natural to expect that functions with spread Fourier decomposition are difficult to approximate by neural networks. The proposition below formalizes this fact. The proof follows a technique similar to the one used in [Dan17a] (see Remark 5 for a comparison) and essentially upper bounds the scalar product between the objective function and the network.

Proposition 3.11. *Let* $\left(f^{(d)}\right)_d$ *a sequence of functions such that* $f^{(d)} \in C(\mathbb{S}^{d-1})$. *Assume that for every* d *there exists* $I_d \subseteq \mathbb{N}$ *such that*

1. *It holds that* $\|f^{(d)}\|_2 \leq O(d^M) \cdot \|P_{I_d} f^{(d)}\|_2$ *for some* $M > 0$ *;*
2. *There exists a non-negative sequence* $\{c_{d,k}\}_{k \in I_d}$ *such that* $\|f_k^{(d)}\|_\infty \leq c_{d,k} \quad \overline{N_k^d}\|f^{(d)}\|_2$ *for all* $k \in I_d$ *and such that* $\left(\sum_{k \in I_d} c_{d,k}^2\right)^{1/2} \leq \epsilon^{d^\alpha} \cdot O(d^M)$ *for some* $\epsilon \in (0,1)$ *and* $\alpha > 0$.

Moreover, assume that $\|f^{(d)}\|_\infty = O(1)$ *and* $\|f^{(d)}\|_2 = \Omega(d^{-M})$ *. Then the sequence* $\left(f^{(d)}\right)_{d>2}$ *is not universally approximable by one-hidden-neural networks.*

Proof. Let $f_N : \mathbb{R}^d \to \mathbb{R}$ a one-hidden-layer network defined by

$$f_N(\mathbf{x}) = \sum_{i=1}^N u_i f^{\sigma_i, \mathbf{w}_i}(\mathbf{x}) \doteq \sum_{i=1}^N u_i \sigma_i \ \mathbf{w}_i^T \mathbf{x}$$

where $\mathbf{u} \in \mathbb{R}^N$, $\mathbf{w}_i \in \mathbb{S}^{d-1}$, and σ_i are linearly bounded activations. Thanks to Parseval's formula,

it holds that

$$
\begin{aligned}
\|f_N - f^{(d)}\|_2^2 &\geq \|\mathcal{P}_{I_d} f_N - \mathcal{P}_{I_d} f^{(d)}\|_2^2 \\
&\geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \sum_{j \in I_d} \sum_{i=1}^{N} u_i \langle f^{\sigma_i, \mathbf{w}_i}, f_j^{(d)} \rangle \\
&\geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \sum_{j \in I_d} \sum_{i=1}^{N} \frac{1}{\sqrt{N_j^d}} |u_i| \|f_j^{(d)}\|_\infty \|f_j^{\sigma_i, \mathbf{w}_i}\|_2 \\
&\geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \|f^{(d)}\|_2 \sum_{i=1}^{N} |u_i| \|f^{\sigma_i, \mathbf{w}_i}\|_2 \left[\sum_{j \in I_d} c_{d,j}^2 \right]^{1/2} \\
&\geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \cdot O(d^M) \cdot \epsilon^{d^\alpha} \|f^{(d)}\|_2 \sum_{i=1}^{N} |u_i| \|f^{\sigma_i, \mathbf{w}_i}\|_2 \, .
\end{aligned}
\tag{3.7}
$$

Finally, notice that it holds that

$$
\|f^{\sigma_i, \mathbf{w}_i}\|_2 \leq 2 \, m_\infty(f_N)
$$

and therefore

$$
\|f_N - f^{(d)}\|_2^2 \geq \Omega(d^{-2M}) - 4 \cdot O(d^M) \cdot \epsilon^{d^\alpha} \cdot m_\infty^2(f_N) \cdot N \, .
$$

This concludes the proof. □

We discuss two particular cases where the assumptions of Proposition 3.11 hold. Let $f^{(d)}_{d>2}$ be a sequence of functions $f^{(d)} \in C(\mathbb{S}^{d-1})$.

Example 2 (Constant control on $\ell_{\infty,2}$). Assume that assumption 1 in Proposition 3.11 holds with $I_d = \{k \in \mathbb{N} : k \geq d^2\}$ and that $\|f^{(d)}\|_2 = \Omega(d^{-M})$ for some constant $M > 0$. If it holds that

$$
\ell_{\infty,2}(f_k^{(d)}) \leq \bar{\ell}
$$

for all $k \geq d^2$ for some constant $\bar{\ell} \geq 1$, then it is easy to check that Proposition 3.11 holds.

This condition could be thought as the spherical harmonic components of the function $f^{(d)}$ being uniformly spread for high energy ($k \geq d^2$). Indeed assumption 2 holds with

$$c_{d,k} \doteq \frac{\bar{\ell}}{\overline{N_k^d}} \frac{\|f_k^{(d)}\|_2}{\|f^{(d)}\|_2}$$

since

$$\sum_{k=d^2}^{\infty} c_{d,k}^2 \leq \frac{\bar{\ell}^2}{N_{d^2}^d} = O(d^{3-d}) \,.$$

This is similar to the condition used in [Dan17a], discussed in the remark below.

Remark 5. Daniely [Dan17a] showed a depth-separation result using a result similar to Proposition 3.11. The difference in this case is that the author considers functions defined on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. Although, since $L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) = L^2(\mathbb{S}^{d-1}) \otimes L^2(\mathbb{S}^{d-1})$, the space $L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1})$ admits a decomposition in spherical harmonics

$$L^2(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) = \sum_{j,k=0}^{\infty} H_j^d \otimes H_k^d \,.$$

In particular, Daniely considers functions of the type

$$f^{(d)} : (\mathbf{x}, \mathbf{y}) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \mapsto h^{(d)}(\mathbf{x}^T \mathbf{y})$$

for some $h^{(d)} \in C([-1, 1])$. Such functions belong to $\overline{\sum_{k=0}^{\infty}} H_k^d \otimes H_k^d$ and satisfy

$$\ell_{\infty,2}(f_{k,k}^{(d)}) \leq \bar{\ell} \cdot \left(N_k^d\right)^{1/2} = \bar{\ell} \cdot \left(N_k^d\right)^{-1/2} \cdot \ell_{k,k}^*$$

where $\ell_{k,k}^* = \max_{f \in H_k^d \otimes H_k^d} \ell_{\infty,2}(f)$. The equation above resembles condition 2 in Proposition 3.11, since it implies that

$$\left\| f_{k,k}^{(d)} \right\|_{\infty} \leq \frac{\bar{\ell}}{\overline{N_k^d}} \frac{\|f_{k,k}^{(d)}\|_2}{\|f^{(d)}\|_2} \cdot \ell_{k,k}^* \cdot \|f^{(d)}\|_2$$

and since

$$c_d \doteq \left[\sum_{k \geq k_d} \left(\frac{\bar{\ell}}{\overline{N_k^d}} \frac{\|f_{k,k}^{(d)}\|_2}{\|f^{(d)}\|_2} \right)^2 \right]^{1/2} \leq \frac{\bar{\ell}}{\sqrt{N_{k_d}^d}}$$

which, for $k_d \geq d^2$ implies that $c_d \lesssim d^3\, 2^{-d}$. The proof is then concluded by choosing $I_d = \{(k,k) : k \geq k_d\}$, since (using the same notations as in the proof of Proposition 3.11), it holds

$$\|f_N - f^{(d)}\|_2^2 \geq \|\mathcal{P}_{I_d} f^{(d)}\|_2^2 - 2 \sum_{(j,j) \in I_d} \sum_{i=1}^{N} \left(\ell_{j,j}^*\right)^{-1} |u_i| \|f_{j,j}^{(d)}\|_\infty \|f_{j,j}^{\sigma_i, \mathbf{w}_i}\|_2$$

which is an equivalent of formula (3.7).

Example 3. Assume that assumption 1 in Proposition 3.11 holds with $I_d = \{k \in \mathbb{N} : k \geq \rho d^\beta\}$ for some $\rho > 0$, $\beta > 0$ and that $\|f^{(d)}\|_2 = \Omega(d^{-M})$ for some constant $M > 0$. If it holds that

$$\ell_{\infty,2}(f_k^{(d)}) \leq \epsilon^k \cdot O(d^M) \cdot \sqrt{N_k^d}$$

for all $k \geq \rho d^\beta$ for some constant $M > 0$, then Proposition 3.11 holds, since

$$\sum_{k=\rho d^\beta}^{\infty} \epsilon^k = \frac{\epsilon^{\rho d^\beta}}{1-\epsilon}\,.$$

This condition could also be thought as the spherical harmonic components of the function $f^{(d)}$ being uniformly spread for high energy ($k \geq d^2$), although in this case the spreadness is required to increase exponentially, as the degree increases, with respect to the maximum spreadness achievable (that is $(N_k^d)^{1/2}$).

Example 4 (Invariant functions). Finally, we show that certain symmetry assumptions can imply energy spreadness. Consider the case of a sign-invariant function $f \in C(\mathbb{S}^{d-1})$, that is such that $f(\boldsymbol{\epsilon} \circ \mathbf{x}) = f(\mathbf{x})$ for every $\boldsymbol{\epsilon} \in \{\pm 1\}^d$ and $\mathbf{x} \in \mathbb{S}^{d-1}$.

Lemma 3.12. *Let $f \in C(\mathbb{S}^{d-1})$ be a sign-invariant function. If*

$$\|f_k\|_\infty = \sup_{\boldsymbol{\epsilon}\in\{\pm 1\}^d} |f_k(\boldsymbol{\epsilon})| \tag{3.8}$$

for some $k \geq 16d^2$ and f is then it holds

$$\|f_k\|_\infty \leq 2 \cdot 2^{-d/2} \quad \overline{N_k^d} \|f_k\|_2 \, .$$

Proof. Notice that since f is Rademacher-symmetric, so is f_k. Consider the function

$$P : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto 2^{-d} N_k^d \sum_{\boldsymbol{\epsilon}\in\{\pm 1\}^d} P_k^d(\boldsymbol{\epsilon}^T \mathbf{x}) \, .$$

The function P satisfies $\|P\|_2 \leq 2 \cdot 2^{-d/2} \quad \overline{N_k^d}$ (see Lemma B.22). Let $\boldsymbol{\epsilon} \in \{\pm 1\}^d$. Then it holds

$$\|f_k\|_\infty = |f_k(\boldsymbol{\epsilon})| = |\langle f_k, P \rangle| \leq \|P\|_2 \|f_k\|_2 \leq 2 \cdot 2^{-d/2} \quad \overline{N_k^d} \|f_k\|_2 \, .$$

This concludes the proof. □

Under polynomial decay of $\|f_k\|_2$, the condition (3.8) can be relaxed to have the frequency $\mathbf{w}^{(k)} \in [0, \infty)^d$ such that $\|f_k\|_\infty = \ f_k(\mathbf{w}^{(k)})$ with

$$\inf_{j\in[d]} w_j^{(k)} \ \geq \text{poly}(d^{-1}) \, .$$

Further discussion on invariant functions are reported in section 5.3.

3.4.4 Efficient approximation under a sparsity condition of the spherical harmonics decomposition

Works by Barron [Bar93, KB18] essentially show that efficient approximation holds under a sparsity condition on the Fourier transform of the function to approximate; more specifically, for $f \in L^1(\mathbb{R}^d)$, the rate of (uniform) approximation is controlled by the quantity $\int_{\mathbb{R}^d} \|\mathbf{w}\|_1^2 |\hat{f}(\mathbf{w})| \, d\mathbf{w}$. In this section we show that an equivalent control can be determined for approximation on the sphere, in terms of spherical harmonics decomposition. For technical reason, the result is established for functions in $\hat{H}^d \doteq H_1^d \oplus \bigoplus_{k=1}^{\infty} H_{2k}^d$ (which correspond to the space of function in L_S^2 whose odd part is linear) and mainly for ReLu activation. We briefly discuss extensions to different activation functions in Remark 7. Consider the space of homogeneous one-hidden-layer neural networks with ReLU activations:

$$\mathcal{F}_N^{\mathrm{ReLU},0} = \left\{ f : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto \sum_{k=1}^{N} u_k \ \left(\mathbf{w}_k^T \mathbf{x} \right)_+ : \mathbf{u} \in \mathbb{R}^N, \mathbf{w}_k \in \mathbb{S}^{d-1} \right\}.$$

Since

$$\left(\mathbf{w}^T \mathbf{x} \right)_+ = \frac{1}{2} \ \mathbf{w}^T \mathbf{x} \ + \frac{1}{2} \ \left| \mathbf{w}^T \mathbf{x} \right| \ ,$$

every function in $\mathcal{F}_N^{\mathrm{ReLU},0}$ is the sum of a linear function with an even one. In other words, $\mathcal{F}_N^{\mathrm{ReLU},0} \subset \hat{H}^d$. Since any linear function belongs to $\mathcal{F}_2^{\mathrm{ReLU},0}$, it is equivalent to consider the problem of approximating even functions by homogeneous one-hidden-layer neural networks with activation $\mathrm{abs}(x) = |x|$, that is, elements of the space

$$\mathcal{F}_N^{\mathrm{abs},0} = \left\{ f : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto \sum_{k=1}^{N} u_k \ \left| \mathbf{w}_k^T \mathbf{x} \right| \ : \mathbf{u} \in \mathbb{R}^N, \mathbf{w}_k \in \mathbb{S}^{d-1} \right\}.$$

To study this, consider the corresponding functional space (as introduced in section 1.2.1)

$$\mathcal{H}^1 \doteq \{ h_\pi : \pi \text{ is a signed even Radon measure} \}$$

where h_π is defined to be the function

$$h_\pi : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto \int_{\mathbb{S}^{d-1}} \mathbf{w}^T \mathbf{x} \; d\pi(\mathbf{w}) \; .$$

The space $\mathcal{H}^1$ is a Banach space endowed with the norm $\gamma_1(h) = \inf_{h \,:\, h=h_\pi} \|\pi\|_1$. As discussed in the introduction, the space $\mathcal{H}^1$ consists of functions which are efficiently approximable by one-hidden-layer networks. More formally, the following holds.

Theorem 3.13 ([BLM89])**.** *Let $f \in \mathcal{H}^1$. Then it holds that*

$$\inf_{f_N \in \mathcal{F}_N^{\mathrm{abs},0}} \|f - f_N\|_\infty \leq c \frac{\gamma_1(f)}{N^{1/3}}$$

where $c > 0$ is a numerical constant. Moreover, f_N satisfying the bound can be chosen to satisfy $\gamma_1(f_N) \leq \gamma_1(f)$.

The question of interest can now be transposed to: which functions $f \in C(\mathbb{S}^{d-1})$ have a (polynomially) small norm $\gamma_1(f)$? One way to approach this problem is by the so-called Blaschke–Levy operator. Consider the transformation

$$T\varphi = \int_{\mathbb{S}^{d-1}} \mathbf{x}^T \mathbf{y} \; \varphi(\mathbf{y}) \, dS(\mathbf{y})$$

for functions $\varphi \in C(\mathbb{S}^{d-1})$. T can be described in terms of spherical harmonics [Rub98] as

$$T\varphi = \sum_{k \geq 0 \text{ even}} \sigma_k \varphi_k \quad \text{where} \quad \sigma_k = \frac{(-1)^{1+k/2}}{2\pi} \frac{\Gamma((k-1)/2)\Gamma(d/2)}{\Gamma((k+d+1)/2)} \; .$$

In particular, it holds that the functional T is an automorphism of $C^\infty_{even}(\mathbb{S}^{d-1})$ (the set of even function in $C^\infty(\mathbb{S}^{d-1})$) [Rub98] . Clearly, its inverse can be defined in terms of spherical harmonics

by

$$T^{-1} : \varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1}) \mapsto \sum_{k \geq 0 \text{ even}} \sigma_k^{-1} \varphi_k \, .$$

The following is immediate.

Proposition 3.14. *For any $\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1})$ it holds that $\varphi \in \mathcal{H}^1$ and*

$$\gamma_1(\varphi) = \left\| T^{-1}\varphi \right\|_1 .$$

Using these results, we can proceed similarly to the work [OWSS19] and obtain the following.

Proposition 3.15. *Let $f \in C(\mathbb{S}^{d-1})$ even. It holds that $f \in \mathcal{H}^1$ if and only if*

$$\sup_{\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1}) \, : \, \|\varphi\|_{\infty} \leq 1} \langle T^{-1}\varphi, f \rangle < \infty \, . \tag{3.9}$$

In this case,

$$\gamma_1(f) = \sup_{\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1}) \, : \, \|\varphi\|_{\infty} \leq 1} \langle T^{-1}\varphi, f \rangle \, .$$

Proof. Assume first that $f \in \mathcal{H}^1$. Then $f = h_\pi$ for some π even signed Radon measure. Thus

$$\begin{aligned}
\gamma_1(f) = \|\pi\|_1 &= \sup_{\varphi \in C(\mathbb{S}^{d-1}) \, : \, \|\varphi\|_{\infty} \leq 1} \int_{\mathbb{S}^{d-1}} \varphi(\mathbf{w}) \, d\pi(\mathbf{w}) \\
&= \sup_{\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1}) \, : \, \|\varphi\|_{\infty} \leq 1} \int_{\mathbb{S}^{d-1}} \varphi(\mathbf{w}) \, d\pi(\mathbf{w}) \\
&= \sup_{\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1}) \, : \, \|\varphi\|_{\infty} \leq 1} \int_{\mathbb{S}^{d-1}} T(T^{-1}\varphi)(\mathbf{w}) \, d\pi(\mathbf{w}) \\
&= \sup_{\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1}) \, : \, \|\varphi\|_{\infty} \leq 1} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} \mathbf{w}^T \mathbf{x} \; (T^{-1}\varphi)(\mathbf{x}) \, dS(\mathbf{x}) \, d\pi(\mathbf{w}) \\
&= \sup_{\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1}) \, : \, \|\varphi\|_{\infty} \leq 1} \langle T^{-1}\varphi, f \rangle \, .
\end{aligned}$$

This shows one side of the statement. On the other hand, assume that

$$\sup_{\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1}) \,:\, \|\varphi\|_{\infty} \leq 1} \langle T^{-1}\varphi, f \rangle < \infty \,.$$

Then, the transformation

$$S_f(\varphi) \doteq \langle T^{-1}\varphi, f \rangle$$

defines a bounded linear operator $S_f : C^{\infty}_{even} \to \mathbb{R}$. Since $C^{\infty}_{even}(\mathbb{S}^{d-1})$ is dense in $C_{even}(\mathbb{S}^{d-1})$ (the set of even function in $C(\mathbb{S}^{d-1})$), S_f can be extended to a bounded linear operator on $C_{even}(\mathbb{S}^{d-1})$. By setting

$$S_f(\varphi) = S_f(\varphi_{even})$$

we can extend it on $C(\mathbb{S}^{d-1})$. By the Riesz representation theorem, there exists a signed Radon measure π on $\mathbb{S}^{d-1}$ such that

$$S_f(\varphi) = \int_{\mathbb{S}^{d-1}} \varphi(\mathbf{w}) \, d\pi(\mathbf{w})$$

for every $\varphi \in C(\mathbb{S}^{d-1})$. Moreover, since $S_f(\varphi) = 0$ for every odd φ, we can assume that π is even. Let h_π be the function in $\mathcal{H}^1$ defined by π. Then it holds that

$$\langle T^{-1}\varphi, f \rangle = \|\pi\|_1 = \langle T^{-1}\varphi, h_\pi \rangle$$

for every $\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1})$. Since T is an automorphism over $C^{\infty}_{even}(\mathbb{S}^{d-1})$, then it holds

$$\langle \varphi, f \rangle = \langle \varphi, h_\pi \rangle$$

for every $\varphi \in C^{\infty}_{even}(\mathbb{S}^{d-1})$. Since f and h_π are even, this implies that $f = h_\pi$. This concludes the proof. □

Functions that satisfy equation (3.9) include all even functions in $C^{d+2}(\mathbb{S}^{d-1})$ if d is even and

all even functions in $C^{d+3}(\mathbb{S}^{d-1})$ if d is odd [Wei76]. This is inline with existing results that show approximability by neural networks for functions whose regularity is proportional to the dimension d (e.g. [MM00]).

Given $f \in C(\mathbb{S}^{d-1})$ even, the condition of Proposition 3.15 is implied by the (weak) convergence (as $N \to \infty$) of the series

$$S_N f = \sum_{k=0}^{N} \sigma_{2k}^{-1} f_{2k}$$

to a finite signed measure π. In this case $f = h_\pi$. In particular, a stronger condition is convergence in $L^1(S)$. This is implied if it holds that

$$\sum_{k \geq 0 \text{ even}} |\sigma_k|^{-1} \|f_k\|_1 < \infty \,. \tag{3.10}$$

Notice that, instead, the series converges in L^2_S if and only if

$$\sum_{k \geq 0 \text{ even}} \sigma_k^2 \|f_k\|_2^2 < \infty \,.$$

This is equivalent to asking that $f \in \mathcal{H}^2$, the RKHS given by the kernel function

$$k : (\mathbf{x}, \mathbf{y}) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \mapsto \int_{\mathbb{S}^{d-1}} \mathbf{x}^T \mathbf{w} \;\; \mathbf{w}^T \mathbf{y} \;\; dS(\mathbf{w}) \,.$$

Since in this case $\mathcal{H}^2$ can be described as

$$\mathcal{H}_2 \doteq \quad h_\pi \,:\, \pi \text{ is a signed even Radon measure with an } L^2_S \text{ density} \quad ,$$

it is clear that $\mathcal{H}^1 \subset \mathcal{H}^2$. We refer to [Bac17a] for more details about these statements. On the other hand, this also implies that the condition (3.10) is potentially much stronger than simply asking for $f \in \mathcal{H}^1$.

Example 5 (Highly concentrated function). Some computations show that

$$|\sigma_k|^{-1} \leq \Theta \quad d^{3/4}k^2 \quad \overline{N_k^d} \quad . \tag{3.11}$$

Using these observations it is then straightforward to prove the following.

Proposition 3.16. *Let* $f^{(d)}{}_d$ *a sequence of even functions in* $C(\mathbb{S}^{d-1})$. *Assume that there exist some constant* $M, N > 0$ *constant such that*

$$\overline{N_k^d}\|f_k^{(d)}\|_1 \leq O \ k^M d^N \ \cdot \|f_k^{(d)}\|_2 \quad \textit{and} \quad \sum_{k=0}^{\infty} k^{M+2}\|f_k^{(d)}\|_2 = O(d^N) \ .$$

Then the sequence $f^{(d)}{}_{d\geq 2}$ *is universally approximable by the space* $\mathcal{F}_N^{\mathrm{abs},0}$.

Proof. By Proposition 3.14 and equation (3.11) above we get that

$$\gamma_1(f^{(d)}) \leq \sum_{k\geq 0 \text{ even}} |\sigma_k|^{-1}\|f_k^{(d)}\|_1 \leq \Theta(d^{N+3/4}) \sum_{k\geq 0 \text{ even}} k^{2+M}\|f_k^{(d)}\|_2 \leq O(d^{3/4+2N}) \ .$$

The application of Theorem 3.13 concludes the proof. □

The proposition above requires essentially two conditions to hold. First, that the energy of the functions decreases fast enough (yet polynomially in k and d). The second condition is that the Fourier components of the function are concentrated enough, that is they are polynomially close to the bound (3.6). We remark that this condition is infact pretty strong; it requires the function f to be band-limited. According to [DFT16], it holds that

$$\|f_k^{(d)}\|_2 \leq C(d)k^{\frac{d-2}{4}}\|f_k^{(d)}\|_1 \ ,$$

for some function $C(d)$. Then $f^{(d)}$ would satisfy

$$\overline{N_k^d} \leq \text{poly}(k,d)\frac{\|f_k^{(d)}\|_2}{\|f_k^{(d)}\|_1} \leq \text{poly}(k,d)k^{\frac{d-2}{4}}$$

Since $\overline{N_k^d} \geq c(d)k^{\frac{d-2}{2}}$ for some $c(d)$, this implies that $k^{\frac{d-2}{4}}\text{poly}(k^{-1}) \leq H(d)$ for some function $H(d)$. It follows that k must satisfy $k \leq K(d)$ for some $K(d)$. Although, the rate of the function $K(d)$ does not follow from [DFT16]; we conjecture that $K(d)$ behaves as a power of d.

Example 6 (High energy zonal harmonics). The properties discussed in this section indicate that high-energy only does not yield not-universal-approximability. As an 'extreme' case, consider the case of a zonal harmonic $f(\mathbf{x}) \doteq P_k^d(\mathbf{w}^T\mathbf{x})$, for $\mathbf{x}, \mathbf{w} \in \mathbb{S}^{d-1}$ where $\mathbf{w}$ is fixed. Notice that $\|f\|_\infty = 1$. It holds that

$$\gamma_1(f) = \frac{\|f_k\|_1}{|\sigma_k|} \leq O(k^2d^{3/4}) \quad \overline{N_k^d}\|f_k\|_1 \leq O(k^2d^{3/4}) \quad \overline{N_k^d}f_k \;_2 = O(k^2d^{3/4}) \,,$$

which implies universal approximability by Theorem 3.13. Similarly, polynomial combinations of zonal harmonics can be well approximated, as expected.

Remark 6 (Ridge function). For a single neuron network $f(\mathbf{x}) = |\mathbf{w}^T\mathbf{x}|$, it holds $\|f\|_\infty = 1$ and $\|f\|_2 = d^{-1/2}$. The spherical components of f are given by

$$f_k(\mathbf{x}) = N_k^d \;\; T \;\; P_k^d(\mathbf{x}^T\cdot) \;\; (\mathbf{w}) \;\; = (\sigma_k N_k^d)P_k^d(\mathbf{w}^T\mathbf{x}) \,.$$

In particular, it holds

$$1 = \gamma_1(f) = \sum_{k\geq 0 \text{ even}} \sigma_k^{-1} f_k \;\; _1 = \sum_{k\geq 0 \text{ even}} N_k^d P_k^d(\mathbf{w}^T\cdot) \;\; _1 \,.$$

Therefore, understanding how tight (or strong) condition (3.10)å is highly correlated with understanding convergence of the series $\sum_{k\geq 0 \text{ even}} N_k^d \;\; P_k^d(\mathbf{w}^T\cdot) \;\; _1$, or equivalently, computing $P_k^d \;\; _{\mu_d,1}$.

Remark 7. While the result of this section mainly concern approximation by homogeneous one-hidden-layer networks with the ReLU (or absolute value) activation, they can easily be extended to any other activation satisfying Assumption 1, under the same assumptions. Moreover, notice that, thanks to Theorem 3.13, universal approximation by $\mathcal{F}_N^{\mathrm{ReLU},0}$ is equivalent to universal approximation by $\mathcal{H}_1 \oplus H_1^d$.

Chapter 4

On the optimization landscape of one-hidden-layer networks

4.1 Introduction

As discussed in section 1.6, loss functions evaluated on neural networks represent a rich class of objectives for which, despite their highly-non-convexity, simple local search heuristics, such as SGD, are able to efficiently recover zero-error minima. In particular, this is true in the over-parametrised regime, where the number of parameters exceeds the one needed to obtain a certain error threshold.

A considerable amount of literature has attempted to characterize the landscape of loss functions evaluated on neural networks by studying its critical points. Global optimality results have been obtained for architectures with linear activations [BH89, HM16, Kaw16, ZL17, LK17, YSJ18, LB18, Zha19, TKB19], quadratic activations [SJL17, DL18] and some more general non-linear activations, under appropriate regularity assumptions [SC16, NH17, FJZT17]. Negative examples have been shown as well, under different assumptions [SCP16, ZL17, SS17b, YSJ18]. Some other insights have been obtained by leveraging tools for complexity analysis of spin glasses [CHM+15]

and random matrix theory [PB17]. Other analysis involved studying goodness of the initialization of the parameter values [DFS16, SS16, DLT+17] or other topological properties of the loss function, such as connectivity of sub-level sets [DVSH18, FB17, Ngu21]. Optimization landscapes have also been studied in other contexts than neural networks, such as non-convex low rank problems [GJZ17], matrix completion [GLM16], problems arising in semidefinite programming [BVB16, BBV16] and implicit generative modeling [BALPO17].

The analysis in this chapter focuses mostly on the class of one-hidden-layer neural networks, with a hidden layer of size N, and covers both empirical and population risk landscapes. More specifically, we look at presence (or absence) of *spurious valleys*, defined as connected components of the sub-level sets that do not contain a global minima. We define two quantities depending on the functional space spanned by neural networks of different widths: the upper intrinsic dimension, defined as the dimension of this linear space, and the lower intrinsic dimension, defined as the minimum number of hidden units to describe any element of the functional space. Upper and lower intrinsic dimensions define only two scenarios: either (i) they are both finite, enabling positive results; or (ii) they are both infinite, implying the negative results. More specifically:

- In section 4.3.1.3 we show that, for empirical risk minimization (ERM) or polynomial activations, spurious valleys do not occur as long as the network is sufficiently over-parametrised. For the case of linear and quadratic activations, our results are (up to a constant factor) tight.

- For non-polynomial non-negative activations, for any hidden width, in section 4.4 we construct data distributions which yield spurious valleys with positive measure, whose value is arbitrarily far from the one of the global.

- Finally, drawing on connections with random features expansions, we show that, even if spurious valleys may appear in general, their measure decreases as the width increases. This holds up to a low energy threshold, which approaches the global minimum at a rate inversely proportional to the hidden layer size (up to log factors). We conclude by discussing limitation

of this approach to analyse the behaviour of optimization mechanism in neural networks and a variety of recent related results. This is the content of sections 4.5.1 and 4.5.2.

4.2 Spurios valleys and intrinsic dimensions of neural networks

Let $(\mathbf{X}, \mathbf{Y})$ be two random variables. These random variables take values in $\mathbb{R}^d$ and $\mathbb{R}^m$ and represent the input and output data, respectively. We consider oracle square loss functions $L : \Theta \to \mathbb{R}$ of the form

$$L(\boldsymbol{\theta}) \doteq \mathbb{E}[\ell(\boldsymbol{\Phi}(\mathbf{X}; \boldsymbol{\theta}), \mathbf{Y})] \tag{4.1}$$

where $\ell : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty)$ is convex in its first argument. For every $\boldsymbol{\theta} \in \Theta$, the function $\boldsymbol{\Phi}(\cdot; \boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}^m$ models the dependence of the output on the input as $\mathbf{Y} \simeq \boldsymbol{\Phi}(\mathbf{X}; \boldsymbol{\theta})$. We mainly focus on one-hidden-layer neural networks $\boldsymbol{\Phi}$, i.e. $\boldsymbol{\Phi}$ of the form

$$\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}^T\mathbf{x}) \tag{4.2}$$

where $\boldsymbol{\theta} = (\mathbf{U}, \mathbf{W}) \in \Theta \doteq \mathbb{R}^{m\times N} \times \mathbb{R}^{d\times N}$. Here N represents the width of the hidden layer and $\sigma : \mathbb{R}^N \to \mathbb{R}^N$ is a continuous identical element-wise activation function, that is $(\sigma(\mathbf{x}))_i = \sigma(x_i)$, for $i \in [N]$ and $\mathbf{x} \in \mathbb{R}^N$. Recall that we denote the space of shallow neural networks with activation σ, width N and output dimension $m = 1$ by $\mathcal{F}_N^\sigma$; notice that, for the matter of this chapter, the last row of $\mathbf{W}$ in (4.2) can be considered to be the bias term without loss of generality, up to re-define the distribution of the random variable $\mathbf{X}$.

The loss function $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta})$ is (in general) a non-convex object; it may present spurious (i.e. non global) local minima. In this chapter, we characterize L by determining absence or presence of spurious valleys, as defined below.

Definition 5. *For all $c \in \mathbb{R}$ we define the sub-level set of L as $\Omega_L(c) = \{\boldsymbol{\theta} \in \Theta : L(\boldsymbol{\theta}) \leq c\}$. We define a spurious valley as a path-connected component of a sub-level set $\Omega_L(c)$ which does not*

contain a global minimum of the loss $L(\boldsymbol{\theta})$.

Since, in practice, the loss (4.1) is minimized with a gradient descent based algorithm, then absence of spurious valleys is a desirable property, if we wish the algorithm to converge to an optimal parameter. It is easy to see that $L(\boldsymbol{\theta})$ not having spurious valleys is implied by the following property:

P.1 Given any *initial* parameter $\tilde{\boldsymbol{\theta}} \in \Theta$, there exists a continuous path $\boldsymbol{\theta} : t \in [0,1] \mapsto \boldsymbol{\theta}_t \in \Theta$ such that:

(a) $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}$;

(b) $\boldsymbol{\theta}_1 \in \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$;

(c) The function $t \in [0,1] \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing.

As pointed out in [FB17], this implies that L has no strict spurious (i.e. non global) local minima. The absence of generic (i.e. non-strict) spurious local minima is guaranteed if the path $\boldsymbol{\theta}_t$ is such that the function $L(\boldsymbol{\theta}_t)$ is strictly decreasing. For sake of clarity, we review these properties in the following lemma (the proof is reported in the section C.5).

Lemma 4.1. *Assume that* $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta})$ *is a continuous function. Then, property* ***P.1*** *implies absence of spurious valleys. In particular, this implies absence of strict spurious minima, and of (generally non-strict) spurious minima if property* ***P.1*** *holds with strictly decreasing paths* $t \mapsto L(\boldsymbol{\theta}_t)$. *Conversely, presence of spurious valleys implies existence of spurious minima.*

In the following, we prove absence of spurious valleys by proving that property **P.1** holds. Intuitively, one should think about spurious valleys as regions of the parameter space from which it is impossible to 'escape' without 'up-climbing' the loss value.

Notice that for many activation functions used in practice (such as the ReLU $\sigma(z) = z_+$), the parameter $\boldsymbol{\theta}$ determining the function $\boldsymbol{\Phi}(\cdot;\boldsymbol{\theta})$ is determined up to the action of a symmetry group

(e.g., in the case of the ReLU, σ is a positively homogeneous function). This already prevents strict minima: for any value of the parameter $\boldsymbol{\theta} \in \Theta$ there exists a (often large) *manifold* $\mathcal{U}_{\boldsymbol{\theta}} \subset \Theta$ intersecting $\boldsymbol{\theta}$ along which the loss function is constant.

ERM vs population loss In the following, we consider the loss (4.1) defined for a generic distribution $(\mathbf{X}, \mathbf{Y})$. In case of a distribution with a finite number of atoms, this corresponds to empirical risk minimization (ERM), which is (usually) the regime where machine learning algorithms perform optimization. On the other hand, for a generic data distribution, this loss is what is called *population* loss, and corresponds to the actual objective that machine learning algorithms aim to minimize. In our work we are interested in analyzing not only the ERM case, but more general population losses. While we in fact focus on highly over-parametrised neural networks, we aim to provide results which apply to the regime where number of data points goes to infinity before the number of parameters.

4.2.1 Intrinsic dimension of shallow networks

The main result of this chapter is to exploit that the property of absence of spurious valleys is related to the complexity of the functional space $\mathcal{F}^{\sigma} \doteq \cup_{N \geq 1} \mathcal{F}_N^{\sigma}$ defined by the network architecture. We therefore define two measures of such complexity which we will use to show, respectively, positive and negative results in this regard.

To simplify the discussion, we introduce some notation which we will use throughout the rest of the paper. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous activation function. For every $\mathbf{v} \in \mathbb{R}^d$ we denote $\psi_{\sigma,\mathbf{v}}$ to be the function $\psi_{\sigma,\mathbf{v}} : \mathbf{x} \in \mathbb{R}^d \mapsto \sigma(\mathbf{v}^T\mathbf{x}) \in \mathbb{R}$. Recall that we refer to each $\psi_{\sigma,\mathbf{v}}$ as a *ridge* function and that, if $\mathbf{X}$ is a random variable taking values in $\mathbb{R}^d$, we denote by $L^2(\mathbf{X})$ the space of L^2 function on $\mathbb{R}^d$ with respect to the probability measure induced by the random variable $\mathbf{X}$.

Finally, we define the space

$$\mathcal{R}_2(\sigma, d) = \quad \mathbf{X} \text{ random variable taking values in } \mathbb{R}^d : \psi_{\sigma,\mathbf{v}} \in L^2(\mathbf{X}) \text{ for every } \mathbf{v} \in \mathbb{R}^d$$

to be the space of (d-dimensional) input data distributions for which the filter functions have finite second moment.

Definition 6. *Let σ be a continuous activation function and $\mathbf{X} \in \mathcal{R}_2(\sigma, d)$. We define*[1]

$$\dim^*(\sigma, \mathbf{X}) = \dim_{L^2(\mathbf{X})}(\mathcal{F}^\sigma)$$

as the upper intrinsic dimension of the pair $(\sigma, \mathbf{X})$. We define the level d upper intrinsic dimension of σ as $\dim^*(\sigma, d) = \dim(V_\sigma) = \sup\{\dim^*(\sigma, \mathbf{X}) : \mathbf{X} \in \mathcal{R}_2(\sigma, d)\}$.

The upper intrinsic dimension $\dim^*(\sigma, \mathbf{X})$ defined above is therefore the dimension of the functional space spanned by the filter functions $\psi_{\sigma,\mathbf{v}} \in L^2(\mathbf{X})$ or, equivalently, of the image of the map $\Phi : \boldsymbol{\theta} \in \Theta \mapsto \Phi(\cdot; \boldsymbol{\theta}) \in L^2(\mathbf{X})$. Notice that $\dim^*(\sigma, \mathbf{X}) \leq \dim(L^2(\mathbf{X}))$. In particular, if the distribution of $\mathbf{X}$ is discrete, i.e. it is concentrated on a finite number of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$, then $\dim^*(\sigma, \mathbf{X}) \leq \dim(L^2(\mathbf{X})) \leq n$. Otherwise, if the distribution $\mathbf{X}$ is not discrete, then $\dim(L^2(\mathbf{X})) = \infty$.

The level d upper intrinsic dimension $\dim^*(\sigma, d)$ is defined as the dimension of the functional linear space $\mathcal{F}^\sigma$. We note that if $\mathbf{X} \in \mathcal{R}_2(\sigma, d)$ is a random variable with almost surely positive density with respect to the Lebesgue measure, then $\dim^*(\sigma, d) = \dim^*(\sigma, \mathbf{X})$.

The following lemma exhausts all the cases when the upper intrinsic dimension is not infinite.

Lemma 4.2. *Let σ be a continuous activation function and $\mathbf{X} \in \mathcal{R}_2(\sigma, d)$ such that $\dim(L^2(\mathbf{X})) =$*

[1]For any linear subspace $V \subseteq L^2(\mathbf{X})$, $\dim_{L^2(\mathbf{X})}(V)$ denotes the dimension of V as a subspace of $L^2(\mathbf{X})$.

∞. *If* $\sigma(z) = \sum_{i=0}^{k} a_i z^i$ *is a polynomial, then*

$$\dim^*(\sigma, \mathbf{X}) \leq \sum_{i=1}^{k} \binom{n+i-1}{i} \mathbf{1}_{\{a_i \neq 0\}} = O(d^k) \, .$$

Otherwise (i.e. if σ *is not a polynomial) it holds* $\dim^*(\sigma, \mathbf{X}) = \infty$.

The proof of the above lemma is based on the UAT (Theorem 1.1). We then define the lower intrinsic dimension, which corresponds to the concept of 'how many hidden neurons are needed to represent a generic function of $\mathcal{F}^\sigma$.

Definition 7. *Let* σ *be a continuous activation function and* $\mathbf{X} \in \mathcal{R}_2(\sigma, d)$. *We define*[2]

$$\dim_*(\sigma, \mathbf{X}) = \inf \ \ N \geq 1 \ : \ \mathcal{F}^\sigma \subseteq_{L^2(\mathbf{X})} \overline{\mathcal{F}_N^\sigma}^{L^2(\mathbf{X})}$$

as the lower dimension of the pair $(\sigma, \mathbf{X})$. *We define the level* d *lower dimension of* σ *as* $\dim_*(\sigma, d) = \sup\{\dim_*(\sigma, \mathbf{X}) : \mathbf{X} \in \mathcal{R}_2(\sigma, d)\}$.

If $\dim_*(\sigma, \mathbf{X})$ is finite, then it corresponds to the minimum number of hidden neurons which are needed to represent any function of $\mathcal{F}^\sigma$ with the neural network architecture (4.2). Clearly, this implies that

$$\dim_*(\sigma, \mathbf{X}) \leq \dim^*(\sigma, \mathbf{X})$$

for every continuous activation function σ and any $\mathbf{X} \in \mathcal{R}_2(\sigma, d)$. As with the upper intrinsic dimension, we note that if $\mathbf{X} \in \mathcal{R}_2(\sigma, d)$ is a random variable with almost surely positive density with respect to the Lebesgue measure, then $\dim_*(\sigma, d) = \dim_*(\sigma, \mathbf{X})$.

In the case of homogeneous polynomial activations $\sigma(z) = z^k$ with $k \geq 1$ integer, the level d lower dimension of σ is closely related to the notion of (maximal) symmetric tensor rank. Let

[2]For any subsets $V, W \subset L^2(\mathbf{X})$, we say that $V \subseteq_{L^2(\mathbf{X})} W$ if $V \subseteq W$ as subsets of $L^2(\mathbf{X})$ (and similar with other types of inclusion). We denote $\overline{V}^{L^2(\mathbf{X})}$ the closure of V in $L^2(\mathbf{X})$. We use the standard notation when it is clear from the context that these concepts are intended with respect to $L^2(\mathbf{X})$.

$\mathrm{S}^k\ \mathbb{R}^d$ be the space of order k symmetric tensors on $\mathbb{R}^d$. For any $\mathbf{T} \in \mathrm{S}^k\ \mathbb{R}^d$, the symmetric rank of $\mathbf{T}$ is defined as $\mathrm{rk}_{\mathrm{S}}(\mathbf{T}) = \min\ \ p \geq 1 : \mathbf{T} = \ \ \sum_{i=1}^{p} u_i \mathbf{w}_i^{\otimes k}$ for some $\mathbf{u} \in \mathbb{R}^p, \mathbf{w}_1, \ldots, \mathbf{w}_p \in \mathbb{R}^n$ [CGLM08]; let $\mathrm{S}_r^k(\mathbb{R}^d)$ denote the subspace of $\mathrm{S}^k(\mathbb{R}^d)$ of tensors of rank at most r. The symmetric border rank is defined as[3]

$$\mathrm{rk}_{\mathrm{S}}^*(\mathbf{T}) = \min\ \ r \geq 0 : \mathbf{T} \in \overline{\mathrm{S}_r^k(\mathbb{R}^d)} \quad .$$

A well-known fact is that, for $k > 2$, the border rank is in general only less or equal than the actual rank, as opposed to the matrix case $k = 2$. We define $\mathrm{rk}_{\mathrm{S}}(k, d) = \max\{\mathrm{rk}_{\mathrm{S}}(\mathbf{T}) : \mathbf{T} \in \mathrm{S}^k(\mathbb{R}^d)\}$ and $\mathrm{rk}_{\mathrm{S}}^*(k, d) = \max\{\mathrm{rk}_{\mathrm{S}}^*(\mathbf{T}) : \mathbf{T} \in \mathrm{S}^k(\mathbb{R}^d)\}$. As noticed in [KTB19], Proposition 5, these two values have the same asymptotic behaviour: it holds that $\frac{1}{2}\mathrm{rk}_{\mathrm{S}}(k, d) \leq \mathrm{rk}_{\mathrm{S}}^*(k, d) \leq \mathrm{rk}_{\mathrm{S}}(k, d)$.

Lemma 4.3. *Let $\sigma(z) = z^k$ for some integer $k > 0$. Then*

$$\frac{1}{2}\mathrm{rk}_{\mathrm{S}}(k, d) \leq \dim_*(\sigma, d) \leq \mathrm{rk}_{\mathrm{S}}(k, d) \ .$$

For the special cases $\sigma(z) = z$ and $\sigma(z) = z^2$, it follows, respectively, $\dim_*(\sigma, d) = 1$ and $\dim_*(\sigma, d) = d$.

Finally, the next lemma implies that for most non-polynomial activation functions of practical interest, the lower intrinsic dimension $\dim_*(\sigma, d)$ is infinite. Let φ denote the univariate standard Gaussian density.

Lemma 4.4. *Let σ be a continuous activation function such that $\sigma \in L^2(\varphi)$ and $d > 1$. Then $\dim_*(\sigma, d) = \infty$ if and only if σ is not a polynomial.*

The proof of the Lemma 4.4 is based on Hermite decomposition and on the correspondence between one-hidden-layer nets and symmetric tensors [MM18].

[3]Here the closure is intended with respect to the Euclidean topology. Notice that it is equivalent to consider the Zarisky topology, as noted in [CLQY20].

4.3 Finite intrinsic dimension and absence of spurious valleys

In this section we provide positive results regarding absence of spurious valleys. Essentially, the following results state that if the width of the network matches the dimension of the functional space $\mathcal{F}^\sigma$ spanned by its filter functions, then no spurious valleys exist. We first provide the main result (Theorem 4.5) in a general form, which allows a straight-forward derivation of two cases of interest: empirical risk minimization (Corollary 4.6) and polynomial activations (Corollary 4.7).

Theorem 4.5. *For any continuous activation function σ and random variable $\mathbf{X} \in \mathcal{R}_2(\sigma, d)$ with finite upper intrinsic dimension $\dim^*(\sigma, \mathbf{X}) < \infty$, the loss function*

$$L(\boldsymbol{\theta}) = \mathbb{E}[\ell(\boldsymbol{\Phi}(\mathbf{X}; \boldsymbol{\theta}), \mathbf{Y})]$$

for one-hidden-layer neural networks $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\boldsymbol{\sigma}(\mathbf{W}^T\mathbf{x})$ admits no spurious valleys in the over-parametrised regime $N \geq \dim^*(\sigma, \mathbf{X})$.

Sketch of the proof. The proof consists of showing that we can construct a descent path verifying property **P.1** starting from any parameters $\boldsymbol{\theta}$. The construction can be articulated in two main parts. First, we show that we can map the starting parameter $\boldsymbol{\theta}_0 = (\mathbf{U}_0, \mathbf{W}_0)$ to another parameter $\boldsymbol{\theta}_{1/2} = (\mathbf{U}_{1/2}, \mathbf{W}_{1/2})$ such that the functions $\mathbf{x} \mapsto \sigma(\mathbf{w}_{1/2,k}^T \mathbf{x})_{k \in [N]}$ form a basis of $\mathcal{F}^\sigma$. It follows that there exists a minimal function $\mathbf{f} \in (\mathcal{F}^\sigma)^m \doteq \{(f_1, \dots, f_m) : f_i \in \mathcal{F}^\sigma\}$, i.e.

$$\mathbf{f} \in \operatorname*{arg\,min}_{\mathbf{g} \in (\mathcal{F}^\sigma)^m} \mathbb{E}[\ell(\mathbf{g}(\mathbf{X}), \mathbf{Y})] \,,$$

which can be represented as $\mathbf{f} = \boldsymbol{\Phi}(\cdot; \boldsymbol{\theta}_1 = (\mathbf{U}_1, \mathbf{W}_{1/2}))$ for some $\mathbf{U}_1$. The second part of the path can be thus taken as $t \mapsto (1-t)\mathbf{U}_{1/2} + t\mathbf{U}_1$: as the loss function is convex, this is a descent path. □

The above result can be interpreted as follows: if the network is such that any of its output units Φ_i can be chosen from the whole linear space spanned by its filter functions $\mathcal{F}^\sigma$, then the associated optimization problem is such that there always exists a descent path to an optimal solution, for any initialization of the parameters.

Applying the observations in section 4.2.1 describing the cases of finite intrinsic dimension, we immediately get the following corollaries.

Corollary 4.6 (ERM)**.** *Consider n data points $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^m$. For one-hidden-layer neural networks $\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\boldsymbol{\sigma}(\mathbf{W}^T\mathbf{x})$, where σ is any continuous activation function, the empirical loss function*

$$L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{\Phi}(\mathbf{x}_i; \boldsymbol{\theta}), \mathbf{y}_i)$$

admits no spurious valleys in the over-parametrized regime $N \geq n$.

Comparison with existing results This results was essentially already shown in [LSSS14]. The only difference with our result is that we allow for rank degeneracy in the matrix $\sigma\ \mathbf{W}^T[\mathbf{x}_1|\cdots|\mathbf{x}_N]$. However, its proof illustrates the danger of studying empirical risk minimization landscapes in over-parametrised regimes, since it bypasses all the geometric and algebraic properties needed in the population risk setting - which may be more relevant to understand the generalization properties of the model.

Other works considered the landscape of empirical risk minimization for deep networks. For ReLU-like activations, multi-layer networks and square losses, [SC16] showed that (almost surely) there exists no differentiable spurious minima if one of the layer weights $\mathbf{W}_k \in \mathbb{R}^{d_{k-1} \times d_k}$ satisfy $d_k d_{k-1} \geq N$. [NH17] showed that no spurious minima occur for multilayer neural networks for a class of losses and activations, if one of the layers inner width exceeds the number of data points and the critical points verify certain non-degeneracy conditions.

Corollary 4.7 (Polynomial activations)**.** *For one-hidden-layer neural networks $\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\boldsymbol{\sigma}(\mathbf{W}^T\mathbf{x})$*

with polynomial activation function $\sigma(z) = a_0 + a_1 z + \cdots + a_k z^k$*, the loss function* $L(\boldsymbol{\theta}) = \mathbb{E}[\ell(\boldsymbol{\Phi}(\mathbf{X};\boldsymbol{\theta}), \mathbf{Y})]$ *admits no spurious valleys in the over-parametrized regime*

$$N \geq \sum_{i=1}^{k} \binom{d+i-1}{i} \mathbf{1}_{\{a_i \neq 0\}} = O(d^k) \, .$$

Under the hypothesis of Corollary 4.7 with $N = O(d^k)$, a generic function of $\mathcal{F}^\sigma$, $\boldsymbol{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{u}^T \sigma(\mathbf{W}^T \mathbf{x})$, can be also represented, for some $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\boldsymbol{\theta})$, in the generalized linear form

$$\boldsymbol{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\gamma}^T \boldsymbol{\varphi}(\mathbf{x})$$

with $\boldsymbol{\varphi}(\mathbf{x}) = (x_{k_1} \cdots x_{k_j})_{\{1 \leq k_1 \leq \cdots \leq k_j \leq d, j \in [k]\}}$. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ differ for their dimensions:

$$\dim(\boldsymbol{\gamma}) = O(d^k) < \dim(\boldsymbol{\theta}) = (d+1) \cdot O(d^k) = O(d^{k+1}) \, .$$

One would therefore like Corollary 4.7 to hold also (at least) for $p \geq O(d^{k-1})$. In the next section we address this problem for the linear activation $\sigma(z) = z$ and the quadratic activation $\sigma(z) = z^2$.

4.3.1 Improved over-parametrization bounds for homogeneous polynomial activations

The over-parametrization bounds obtained in Corollary 4.7 are quite non-desiderable in practical applications. We show that they can in fact be improved, for the case of linear and quadratic networks.

4.3.1.1 Linear networks case

Linear networks have been considered as a first order approximation of feed-forward multi-layers networks [Kaw16]. It was shown, in several works [Kaw16, FB17, YSJ18], that, for linear net-

works of any depth

$$\mathbf{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{W}_{L+1}\cdots\mathbf{W}_1\mathbf{x} \tag{4.3}$$

with $\boldsymbol{\theta} = (\mathbf{W}_{L+1}, \mathbf{W}_L, \dots, \mathbf{W}_2, \mathbf{W}_1) \in \mathbb{R}^{d_L\times m} \times \mathbb{R}^{d_{L-1}\times d_L} \times \cdots \mathbb{R}^{d_1\times d_2} \times \mathbb{R}^{d\times d_1}$, the loss function (4.1) has no spurious local minima, if $\min_{i\in[L]} d_i \geq \min\{d, m\}$. This corresponds exactly to the over-parametrization regime in Corollary 4.7, for the case of one-hidden-layer networks. The following theorem improves on Corollary 4.7 for the case of multi-layer linear networks, showing that no over-parametrisation is required in this case to avoid spurious valleys, for square loss functions.

Theorem 4.8 (Linear networks)**.** *For linear neural networks* (4.3) *of any depth* $L \geq 1$ *and of any hidden layer widths* $d_k \geq 1$, $k \in [L]$, *and any input-output dimensions* $d, m \geq 1$, *the square loss function* $L(\boldsymbol{\theta}) = \mathbb{E}\|\mathbf{\Phi}(\mathbf{X};\boldsymbol{\theta}) - \mathbf{Y}\|_2^2$ *admits no spurious valleys.*

4.3.1.2 Quadratic networks case

Quadratic activations $\sigma(z) = z^2$ have been considered in the literature [LSSS14, DL18, SJL17] as second order approximation of general non-linear activations. Corollary 4.7 says that, if $N \geq d(d+1)/2$, the loss function (4.1) admits no spurious valleys. In the following theorem we relax the over-parametrisation requirement and show that $N > 2d$ is sufficient for the statement to hold, in the case of square loss functions and one dimensional output ($m = 1$).

Theorem 4.9 (Quadratic networks)**.** *For one-hidden-layer neural networks* $\Phi(\mathbf{x};\boldsymbol{\theta}) = \mathbf{u}^T\boldsymbol{\sigma}(\mathbf{W}^T\mathbf{x})$ *with quadratic activation function* $\sigma(z) = z^2$ *and one-dimensional output* ($m = 1$)*, the square loss function* $L(\boldsymbol{\theta}) = \mathbb{E}|\Phi(\mathbf{X};\boldsymbol{\theta}) - Y|^2$ *admits no spurious valleys in the over-parametrised regime* $N \geq 2d + 1 = O(d)$.

Sketch of the proof. The proof (reported in section C.2) consists in constructing a path satisfying property **P.1** and improves upon the proof of Theorem 4.5 by leveraging the special linearized

structure of the network for quadratic activation. For every parameter $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W}) \in \mathbb{R}^N \times \mathbb{R}^{d \times N}$, we can write

$$\Phi(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^{N} u_k (\mathbf{w}_k^T \mathbf{x})^2 = \left\langle \sum_{k=1}^{N} u_k \mathbf{w}_k \mathbf{w}_k^T, \mathbf{x}\mathbf{x}^T \right\rangle_F .$$

We notice that $\Phi(\cdot; \boldsymbol{\theta})$ can also be represented by a neural network $\Phi(\cdot; \hat{\boldsymbol{\theta}})$ with d hidden units; indeed, if $\sum_{k=1}^{d} \sigma_k \mathbf{v}_k \mathbf{v}_k^T$ is the SVD of $\sum_{k=1}^{N} u_k \mathbf{w}_k \mathbf{w}_k^T$, then $\Phi(\mathbf{x}; \boldsymbol{\theta}) = \langle \sum_{k=1}^{d} \sigma_k \mathbf{v}_k \mathbf{v}_k^T, \mathbf{x}\mathbf{x}^T \rangle_F$. Therefore $N \geq d$ is sufficient to describe any element in $\mathcal{F}^\sigma$. A path to the symmetric matrix defining the optimal network is then constructed by mapping the above decomposition defined by the standard form of the network. □

The factor 2 in the statement is due to some technicalities in the proof, but a more involved proof might be able to extend the result to the regime $N \geq d$. The extension of such mechanism for higher order tensors (appearing as a result of multiple layers or high-order polynomial activations) using tensor decomposition also seems possible and is of interest for future work.

Comparison with other works The same optimization landscape has been considered in the works [SJL17] and [DL18]. In the first work, the authors show absence of spurious minima for the case of $N \geq 2d$ and of ERM (loss evaluated on n data points), but for fixed output layer weights; under some assumption on the output layer weights, the result is shown to still hold for $N \geq d$, if $d \leq n \leq O(d^2)$. This last condition can be removed by considering the regularized loss with non-zero weight decay, as shown in [DL18]; in the same work, the authors also proved absence of spurious minima in the case $N < d$ and $N(N+1) \geq 2n$ for a randomly regularized loss (with high probability).

By relaxing the statement to absence of spurious valleys, we showed that this holds for the square loss (both in population and ERM setting) and the optimisation problem over both layer weights if $N > 2d$.

4.3.1.3 Lower to upper intrinsic dimension gap

As observed in Lemma 4.3 $\dim_*(\sigma(z) = z, d) = 1$ and $\dim_*(\sigma(z) = z^2, d) = d$ for all integer $d \geq 1$. Therefore, Theorem 4.8 and Theorem 4.9 say that, for $\sigma(z) = z^k$, $k \in [2]$, and $m = 1$, the square loss function $L(\boldsymbol{\theta}) = \mathbb{E}|\Phi(\mathbf{X}; \boldsymbol{\theta}) - Y|^2$ admits no spurious valleys in the over-parametrized regime $N \geq O(\dim_*(\sigma, d))$. We conjecture that this holds for any (sufficiently regular) activation function with finite lower intrinsic dimension.

On the other hand, we point out that this might be due to the specific choice of the square loss function. In fact, it has been shown that, for generic convex losses ℓ and linear networks, a result equivalent to Theorem 4.8 holds if and only if ℓ is the square loss function [TKB19].

4.4 Infinite intrinsic dimension and presence of spurious valleys

This section is devoted to the construction of worst-case scenarios for non-over parametrised networks. The main result (Theorem 4.10) essentially states that, for networks with width smaller than the lower intrinsic dimension defined above, spurious valleys can be created by choosing adversarial data distributions. We then show how this implies negative results for under-parametrized polynomial architectures and a large variety of architectures used in practice.

A possible way to show existence of spurious valleys is to show that there exists an open set $U \subset \Theta$ such that

$$\min_{\boldsymbol{\theta} \in U} L(\boldsymbol{\theta}) > \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$$

and such that every path from a point in U_1 to any global minima of L must pass through a certain set $S \subset \Theta$ verifying

$$\min_{\boldsymbol{\theta} \in S} L(\boldsymbol{\theta}) > \max_{\boldsymbol{\theta} \in U} L(\boldsymbol{\theta}) \, .$$

Let's start by considering the *quadratic* case $\sigma(z) = z^2$, and let $N \in [2, d-1]$. As we already discussed in this case there is a surjective mapping from the space of one-hidden-layer networks $\mathcal{F}_N^\sigma$ to M_N^d, the set of symmetric $d \times d$ matrices of rank at most N: for every $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W}) \in \Theta = \mathbb{R}^N \times \mathbb{R}^{d \times N}$, it holds that

$$\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \langle \mathbf{x}\mathbf{x}^T, \mathbf{M}(\boldsymbol{\theta}) \rangle_F \ ,$$

where $\mathbf{M}(\boldsymbol{\theta}) \doteq \sum_{i=1}^N u_i \mathbf{w}_i \mathbf{w}_i^T$ is a continuous mapping of $\boldsymbol{\theta}$ to M_N^d. For any random vector $(\mathbf{X}, Y)$, with $\mathbf{X} \in \mathcal{R}(\sigma, d)$, the loss function

$$L : \boldsymbol{\theta} \in \Theta \mapsto \mathbb{E}|\Phi(\mathbf{X}; \boldsymbol{\theta}) - Y|^2$$

can be 'projected' to a corresponding loss function $\mathcal{L}$ over M_N^d, that is

$$\mathcal{L} : \mathbf{M} \in M_N^d \ \to \mathbb{E}|\langle \mathbf{M}, \mathbf{X}\mathbf{X}^T \rangle - Y|^2 \ .$$

Since the mapping $\boldsymbol{\theta} \mapsto \mathbf{M}(\boldsymbol{\theta})$ is continuous, it is equivalent to find two subset U and S of M_N^d as above, for the projected loss $\mathcal{L}$. Let $M^d_{(s_+, s_0, s_-)}$ be the set of symmetric $d \times d$ matrices with rank $d - s_0$ and signature (s_+, s_-), for $s_+ + s_- = d - s_0$. Clearly, a continuous path in M_N^d from a point in $M^d_{(k,0,N-k)}$ to a point $M^d_{(j,0,N-j)}$ for some $k \neq j$ must pass through M^d_{N-1}. Let $L_{(s_+,s_0,s_-)} \doteq \min_{\mathbf{M} \in M^d_{(s_+,s_0,s_-)}} \mathcal{L}(\mathbf{M})$ and $L_r \doteq \min_{\mathbf{M} \in M_r^d} \mathcal{L}(\mathbf{M})$. Then, the existence of $(\mathbf{X}, Y)$ such that

$$L_{N-1} > L_{(N-1,0,1)} > L_{(N,0,0)} \tag{4.4}$$

implies the statement, since any path in M_N^d from a local minima in $M^d_{(N-1,0,0)}$ to a global minima must pass through M^d_{N-1}. Let $\mathbf{X}$ be a d-dimensional Gaussian random variable, and $Y =$

$\langle \mathbf{A}, \mathbf{X}\mathbf{X}^T \rangle$, where $\mathbf{A} \in M^d_{N+1}$ has the form

$$\mathbf{A} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \sum_{k=2}^{N+1} \mathbf{v}_k \mathbf{v}_k^T \ ,$$

for some orthonormal vectors $\{\mathbf{v}_k\}_{k\in[N+1]} \subset \mathbb{R}^d$ and $\lambda_1 < 0$, $\lambda_2 > 0$ with $|\lambda_1| < |\lambda_2| < \sqrt{N}|\lambda_1|$. Then, it follows that

$$\mathcal{L}(\mathbf{M}) = C_1 \|\mathbf{A} - \mathbf{M}\|_F^2 + C_2 (\mathrm{tr}(\mathbf{A} - \mathbf{M}))^2$$

for $C_1 = 96$ and $C_2 = 9$. It holds that

$$L_{N-1} \geq C_1 \min_{\mathbf{M} \in M^d_{N-1}} \|\mathbf{A} - \mathbf{M}\|_F^2 = C_1(\lambda_1^2 + \lambda_2^2) \ ,$$

$$L_{(N-1,0,1)} \geq C_1 \min_{\mathbf{M} \in M^d_{(N-1,0,1)}} \|\mathbf{A} - \mathbf{M}\|_F^2 = C_1 \lambda_2^2 \ .$$

The lower bound of $L_{(N-1,0,1)}$ is found for $\mathbf{M} = \mathbf{M}^{(N-1,1)} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \sum_{k=2}^{N} \mathbf{v}_k \mathbf{v}_k^T$. Let

$$\mathbf{B} = \mathbf{M}^{(N-1,1)} + \frac{\mathrm{tr}(\mathbf{A} - \mathbf{M}^{(N-1,1)})}{N} \sum_{k=1}^{N} \mathbf{v}_k \mathbf{v}_k^T = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \sum_{k=2}^{N} \mathbf{v}_k \mathbf{v}_k^T + \frac{\lambda_2}{N} \sum_{k=1}^{N} \mathbf{v}_k \mathbf{v}_k^T \ .$$

Notice that $\mathbf{B} \in M^d_{(N-1,0,1)}$ since $|\lambda_2| < N|\lambda_1|$. Then it holds that

$$L_{(N-1,0,1)} \leq \mathcal{L}(\mathbf{B}) = C_1 \lambda_2^2 \ \left(1 + \frac{1}{N}\right) < L_{N-1}$$

since $|\lambda_2| < \sqrt{N}|\lambda_1|$. Similarly, it holds that

$$L_{(N,0,0)} \geq C_1 \min_{\mathbf{M} \in M^d_{(N,0,0)}} \|\mathbf{A} - \mathbf{M}\|_F^2 = C_1 \lambda_1^2 \ ,$$

and the minima is found for $\mathbf{M} = \mathbf{M}^{(N,0)} = \lambda_2 \sum_{k=2}^{N+1} \mathbf{v}_k \mathbf{v}_k^T$. Let

$$\mathbf{C} = \mathbf{M}^{(N,0)} + \frac{\mathrm{tr}(\mathbf{A} - \mathbf{M}^{(N,0)})}{N} \sum_{k=2}^{N+1} \mathbf{v}_k \mathbf{v}_k^T = \lambda_2 + \frac{\lambda_1}{N} \sum_{k=2}^{N+1} \mathbf{v}_k \mathbf{v}_k^T .$$

Notice that $\mathbf{C} \in M^d_{(N,0,0)}$ since $|\lambda_1| < |\lambda_2| < N|\lambda_2|$. Then it holds that

$$L_{(N,0,0)} \leq \mathcal{L}(\mathbf{C}) = C_1 \lambda_1^2 \left(1 + \frac{1}{N}\right) .$$

Equation (4.4) thus holds if

$$|\lambda_1| \sqrt{1 + \frac{1}{N}} < |\lambda_2| .$$

Therefore, choosing, for example, $\lambda_1 = -c$ and $\lambda_2 = 5c/4$ for some $c > 0$ proves existence of spurious valleys in the loss L for the choice of random variables $(\mathbf{X}, Y)$. Notice moreover, that the quantity that must be 'up-climbed' to escape a spurious valley located in a region corresponding to $M^d_{(N-1,0,1)}$ is quadratic in c, and therefore, arbitrarily large.

The proof for the general case (σ non quadratic) is based on a similar idea: an interpretation of lower intrinsic dimension as 'maximal rank'; see section 4.2.1. A formal statement of our result is given below.

Theorem 4.10. *Consider the square loss function $L(\boldsymbol{\theta}) = \mathbb{E}\|\boldsymbol{\Phi}(\mathbf{X}; \boldsymbol{\theta}) - \mathbf{Y}\|^2$ for one-hidden-layer neural networks $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$ with non-negative activation function $\sigma : \mathbb{R} \to [0, \infty)$ such that $\sigma \in L^2(\varphi)$ and $\sigma(0) = 0$. If $2 \leq N \leq \frac{1}{2}\dim_*(\sigma, d-1)$, then there exists a random vector $(\mathbf{X}, \mathbf{Y})$ such that the square loss function L admits spurious valleys. In particular, for any $M > 0$ large enough, the random variable $\mathbf{Y}$ can be chosen in such a way that there exists a (non-empty) open set $\Omega \subset \Theta$ such that*

$$M/2 + \min_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}) \geq \sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}) \geq \min_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}) \geq M + \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) \tag{4.5}$$

and any path $\boldsymbol{\theta} : [0,1] \to \Theta$ *such that* $\boldsymbol{\theta}_0 \in \Omega$ *and* $\boldsymbol{\theta}_1$ *is a global minima verifies*

$$\max_{t\in[0,1]} L(\boldsymbol{\theta}_t) \geq \min_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta}) + M \ . \tag{4.6}$$

Equation (4.5) in Theorem 4.10 says that any local descent algorithm, if initialized in $\boldsymbol{\theta}_0 \in \Omega$, at its best it will only be able to produce a final parameter value which is at least M far from optimality. Equation (4.6) implies that any path starting from parameter belonging to Ω must 'up-climb' at least $M/2$ in the loss value. In the following we refer to such property, as stated in Theorem 4.10, by saying that *the loss function has arbitrarily bad spurious valleys*. Note that this result ensures that spurious valleys have positive Lebesgue measure, so there is a positive probability that gradient descent methods initialized with a measure that is absolutely continuous with respect to Lebesgue will get stuck in a bad local minima. The random variables $(\mathbf{X}, Y)$ chosen in the proof of Theorem 4.10 correspond to a student-teacher scenario, where the planted solution has more neurons more than the network whose weights we want to optimize.

Applying the lemmas describing the values of the lower intrinsic dimension for different activation functions, we get the following corollaries.

Corollary 4.11 (Homogeneous even degree polynomial activations)**.** *Consider the case of activation* $\sigma(z) = z^{2k}$ *with* $k \geq 1$ *integer. For one-hidden-layer neural networks* $\boldsymbol{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$*, if* $d \geq 2$ *and the hidden layer width satisfies*

$$N \leq \begin{cases} d-1 & \text{if } k = 1 \\ \frac{1}{2}\,\mathrm{rk}_{\mathrm{S}}^*(2k, d-1) & \text{if } k > 1 \end{cases}$$

then there exists a random variable $(\mathbf{X}, \mathbf{Y})$ *such that the square loss function* $L(\boldsymbol{\theta}) = \mathbb{E}\|\boldsymbol{\Phi}(\mathbf{X};\boldsymbol{\theta}) - \mathbf{Y}\|^2$ *has arbitrarily bad spurious valleys.*

This follows by Theorem 4.10 and Lemma 4.3, since $\dim_*(\sigma(z) = z^{2k}, d) \geq \mathrm{rk}_{\mathrm{S}}^*(2k, d)$. For

the well known case $k = 1$ (symmetric matrices) it holds $\mathrm{rk}_{\mathrm{S}}(2, d) = d$; therefore Corollary 4.11 implies that the bound provided in Corollary 4.7 is almost (up to a factor 2) tight. Notice that our result is indeed in line with the results discussed in section 4.3.1.2.

Corollary 4.12 (Spurious valleys exist in generic architectures)**.** *If $d \geq 2$, for one-hidden-layer neural networks $\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\boldsymbol{\sigma}(\mathbf{W}\mathbf{x})$ with any hidden layer width $N \geq 1$ and continuous non-negative non-polynomial activation function $\sigma \in L^2(\varphi)$ wihr $\sigma(0) = 0$, then there exists a random variable $(\mathbf{X}, \mathbf{Y})$ such that the square loss function $L(\boldsymbol{\theta}) = \mathbb{E}\|\mathbf{\Phi}(\mathbf{X}; \boldsymbol{\theta}) - \mathbf{Y}\|_2^2$ has arbitrarily bad spurious valleys. This setting includes the ReLU activation functions $\sigma(z) = z_+$.*

This follows by Theorem 4.10 by observing that $\dim_*(\sigma, d) = \infty$ if σ is one of the above activation functions.

Discussion and comparison with previous works Several works showed existence of spurious minima: [SS17b] showed counterexamples under Gaussian input distributions, for $N = d - 1 \in \{8, \dots, 19\}$, using a computer-assisted proof; [SCP16] and [ZL17] provided a few numerical examples; [YSJ18] showed existence of spurious minima for ReLU-like activations under non-realizability, and provided counterexamples for smooth activations. For any number of hidden neurons N, we give a (constructive) proof of existence of a data distribution which creates spurious valleys, under the only assumption of non-negative continuous activation function. We also remark that while in the above works the authors proved existence of spurious local minima, we prove that, in fact, arbitrarily bad spurious valleys can exist, which is a stronger negative characterization.

The results of this section can be interpreted as worst-case scenarios for the problem of optimizing (4.1). We showed that, even for simple one-hidden-layer neural network architectures with non-linear activation functions used in practice (such as ReLU), global optimality results can not hold, unless we make some assumptions on the data distributions.

4.5 Increasing the width

In the previous section it was shown that whenever the number of hidden units N is below the lower intrinsic dimension, then one can show worst-case data distributions that yield a landscape with arbitrarily bad spurious valleys. A natural follow-up question is thus to consider the complexity of the energy landscape in a *typical* scenario, defined in terms of both parameter initialisation (how likely are descent algorithms to fall into a spurious valley?) and energy value (how deep are typical spurious valleys?).

4.5.1 A sampling regime

In a student-teacher setting, under sufficient regularity of the objective function and sufficient over-parametrization, one can leverage idea from random features methods [RR+07] to show that *as the network width increases, spurious valleys tend to be confined to decreasingly low loss value*. In the following, we formalize this argument and comment on its limitations afterwards.

Consider an oracle square loss of the form

$$L(\boldsymbol{\theta}) = \mathbb{E}|\Phi(\mathbf{X};\boldsymbol{\theta}) - Y|^2 \,,$$

where $\mathbf{X}$ is a d-dimensional square integrable random variable with distribution μ, $\Phi(\mathbf{x};\boldsymbol{\theta}) = \mathbf{u}^T\boldsymbol{\sigma}(\mathbf{W}^T\mathbf{x})$ and Y has the form $Y = f(\mathbf{X})$, where $f \in L^2(\mathbf{X})$. Assume here for simplicity that σ is a positively homogeneous activation function, such as the ReLU $\sigma(t) = t_+$. If f can be represented as

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}} g(\mathbf{w})\sigma(\mathbf{w}^T\mathbf{x})\, dS(\mathbf{w})$$

for some measurable $g : \mathbb{S}^{d-1} \to \mathbb{R}$, then one way to find a close-to-optimal parameter $\boldsymbol{\theta}$ is to

sample first layer weights $\mathbf{w}_1, \dots, \mathbf{w}_N \sim S$, and take $\boldsymbol{\theta} = (\mathbf{g}(\mathbf{W}), \mathbf{W})$, where

$$\mathbf{g}(\mathbf{W}) \doteq N^{-1}(g(\mathbf{w}_1), \dots, g(\mathbf{w}_N)).$$

If N is large enough, then this approach can work quite well; this is known as the *random features* method [RR+07]. Indeed, the average (over the sampling of $\mathbf{w}_1, \dots, \mathbf{w}_N$) error satisfies

$$\mathbb{E}_{\mathbf{w}_1,\dots,\mathbf{w}_N} L((\mathbf{g}(\mathbf{W}), \mathbf{W})) = \mathbb{E}_{\mathbf{w}_1,\dots,\mathbf{w}_N} \left\| \frac{1}{N} \sum_{k=1} \psi(\mathbf{w}_i) - \mathbb{E}_{\mathbf{w}} \psi(\mathbf{w}) \right\|_{\mu,2}^2 \leq \frac{1}{N} \mathbb{E}_{\mathbf{w}} \|\psi(\mathbf{w})\|_{\mu,2}^2 ,$$

where $\psi(\mathbf{w}) \in L^2(\mu)$ is defined by $\psi(\mathbf{w}) : \mathbf{x} \mapsto g(\mathbf{w})\sigma(\mathbf{w}^T\mathbf{x})$ and $\mathbb{E}_{\mathbf{w}}$ denotes the expectation for $\mathbf{w} \sim S$. Then, by applying a concentration argument, the bound on the average error can be transformed in a bound on the error with high probability. For example, assume that $g \in L^\infty(\mathbb{S}^{d-1})$, so that $C = \sup_{\mathbf{w}\in\mathbb{S}^{d-1}} \|\psi(\mathbf{w})\|_{\mu,2} < \infty$. Then, we can apply Example 6.3 from [BLM13] (Lemma C.12) to obtain that

$$\mathbb{P} \left[L((\mathbf{g}(\mathbf{W}), \mathbf{W})) \leq O(N^{-(1-\delta)}) \right] \geq 1 - e^{-N^\delta}$$

for any $\delta > 0$, for N sufficiently large. Notice that, given an initial parameter parameter value $\boldsymbol{\theta}_0 = (\mathbf{u}_0, \mathbf{W}_0)$, one can consider the path

$$\boldsymbol{\theta}_t = (t\mathbf{q}(\mathbf{W}_0) + (1-t)\mathbf{u}_0, \mathbf{W}_0) \quad \text{where} \quad \mathbf{q}(\mathbf{W}_0) = \operatorname*{arg\,min}_{\mathbf{u}\in\mathbb{R}^N} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=(\mathbf{u},\mathbf{W}_0)} .$$

By convexity of L, the function $t \in [0,1] \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing and it holds that

$$L(\boldsymbol{\theta}_1) \leq L((\mathbf{g}(\mathbf{W}_0), \mathbf{W}_0)) .$$

These two remarks then imply the following statement: for any $N > 0$ large enough, given any initial parameter $\boldsymbol{\theta}_0 = (\mathbf{u}_0, \mathbf{W}_0)$, where the columns of $\mathbf{W}_0$ have been sampled (independently) uniformly over the sphere, there exists a path $t \in [0,1] \mapsto \boldsymbol{\theta}_t$ such that the function $t \in [0,1] \mapsto$

$L(\boldsymbol{\theta}_t)$ is non-increasing and

$$L(\boldsymbol{\theta}_1) \leq O(N^{-1/2})$$

with probability greater or equal then $1 - e^{-N^{1/2}}$ (over the parameter initialization). From the point of view of spurious valleys, this statement can be interpreted by saying that, as the width increases, large loss spurious valleys (that is such that the minimum value in the valley is far from the global minima value of the loss function) are restricted to small regions of the parameter space. Notice that this argument relies on the regularity assumption of the label variable Y. By using more refined arguments, one can prove an equivalent result by assuming only $g \in L^2(\mathbb{S}^{d-1})$. This is the content of the following result, which is proved in section C.4.

Proposition 4.13. *Consider an initial parameter $\boldsymbol{\theta}_0 = (\mathbf{u}_0, \mathbf{W}_0) \in \mathbb{R}^N \times \mathbb{R}^{d \times N}$, where the columns of $\mathbf{W}_0$ are sampled independently uniformly over the sphere $\mathbb{S}^{d-1}$. Then there exists a path $t \in [0,1] \mapsto \boldsymbol{\theta}_t$ such that the function $t \in [0,1] \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing and*

$$L(\boldsymbol{\theta}_1) \leq \lambda \quad \textit{if} \quad N \geq O\ -\lambda^{-1} \log(\lambda \delta)$$

with probability greater or equal then $1 - \delta$, for every $\lambda, \delta \in (0,1)$.

Notice that assuming that f has the form for $g \in L^2(\mu)$ is equivalent to say that $f \in \mathcal{H}^2_+$, the RKHS defined by the kernel function

$$k : (\mathbf{x}, \mathbf{y}) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \mapsto \int_{\mathbb{S}^{d-1}} (\mathbf{w}^T \mathbf{x})_+ (\mathbf{w}^T \mathbf{y})_+ \, dS(\mathbf{w}) \, .$$

In the case of $\mu = S$, such RKHS is (up to linear terms) a subset of the functional space $\mathcal{H}^1$ introduced in section 3.4.4; in fact, it admits a similar description: any $f \in L^2(S)$ satisfies

$$f \in \mathcal{H}^2_+ \quad \text{if and only if} \quad \|f\|^2_{\mathcal{H}^2_+} \doteq \inf_{\rho \in L^2(S) \,:\, f = h^+_\rho} \|\rho\|^2_2 < \infty \, ,$$

where $h_\rho^+ : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto \int_{\mathbb{S}^{d-1}} (\mathbf{w}^T \mathbf{x})_+ \rho(\mathbf{w})\, d\mathbf{w}$. In particular in this case, an even function $f \in L^2(S)$ satisfies $f \in \mathcal{H}_+^2$ if and only if $\|f\|_{\mathcal{H}_+^2}^2 = \sum_{k=0}^{\infty} |\sigma_k|^{-2} \|f_k\|_2^2 < \infty$ (using the notation from 3.4.4); this also implies that $\mathcal{F}^\sigma$ (minus linear terms) is not contained in $\mathcal{H}^2$. A similar characterization can be shown for generic μ, but this should suffice to convince the reader that Proposition 4.13 holds under a strong assumption: that the function $Y = f(\mathbf{X})$ can be efficiently found by kernel methods. To conclude, it has been shown that it is essentially not possible to extend this type of analysis to the case of a 'planted solution' f. In [YS19], the authors show that, for any sufficiently large input dimension d and $\mathbf{X}$ standard d-dimensional Gaussian, there exists $b \in \mathbb{R}$ with $|b| \leq O(d^2)$ such that, for any $Y = (\mathbf{v}^T \mathbf{X} + b)_+$ with $\|\mathbf{v}\| = d^3$, any network $\Phi(x; \boldsymbol{\theta}) = \mathbf{u}^T \boldsymbol{\sigma}(\mathbf{W}^T \mathbf{x})$, where the columns have been sampled independently uniformly from the sphere, satisfies $L(\boldsymbol{\theta}) \geq 0.02$ with probability greater than $1 - e^{-\Theta(d)}$ unless

$$N \cdot m_\infty(\Phi(\cdot; \boldsymbol{\theta})) \geq e^{\Omega(d)} .$$

4.5.2 Related works

In the previous sections, we essentially showed that one-hidden-layer neural networks are amenable to being optimized by descent methods if and only if the architecture 'fills' the corresponding functional space, with respect to the data distribution. For generic data distributions, this only happens if the expressivity of such neural networks is bounded, that correspond to the activation being a polynomial. We also showed that under sufficiently regular teacher-students scenario, increasing the number of parameters has a benign effect on the optimization landscape. Although, in these same regimes, the same optimization can be performed, by e.g., kernel methods, while requiring less parameters. Given these observations, it thus seems unlikely that qualitative descriptions of the optimization landscape, while insightful, can alone explain the success of neural networks optimization. In other words, it seems evident that the algorithms being used play an important role in how such landscapes are *visited*. During the last years, a great amount of work has been dedicated

to understand such interplay. In the remaining of this section, we briefly review and comment on some recent relevant results.

4.5.2.1 Neural tangent kernel and lazy training

Neural networks optimization can be formulated as the search for a minima $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ of a loss function $\boldsymbol{\theta} \in \Theta \mapsto L(\boldsymbol{\theta})$ of the form

$$L(\boldsymbol{\theta}) = R(f(\boldsymbol{\theta})) \ ,$$

where R is a convex functional defined on some functional space $\mathcal{F}$, and $f(\boldsymbol{\theta}) \in \mathcal{F}$ represents the parametric model we are trying to learn. In the limit as the step-size goes to 0, gradient descent defines an ODE

$$\dot{\boldsymbol{\theta}}_t = -L(\boldsymbol{\theta}_t) = -\nabla f(\boldsymbol{\theta}_t)^T \cdot dR(f_t) \ , \tag{4.7}$$

where $f_t \doteq f(\boldsymbol{\theta}_t)$, $dR(f_t) \in \mathcal{F}$ denotes the differential of R in f_t and $\nabla f(\boldsymbol{\theta}_t)^T$ is the transpose of the Jacobian of f in $\boldsymbol{\theta}_t$, which defines a linear operator from $\mathcal{F}$ to Θ. For example, in the case of neural networks evaluated on a ERM loss, $f(\boldsymbol{\theta}) \in \mathbb{R}^{o \cdot n}$ represents the o-dimensional output values of the network evaluated on n data points, and all the differentials are in the usual euclidean sense. Then it follows that

$$\partial_t f_t = -\ \nabla f(\boldsymbol{\theta}_t) \cdot \nabla f(\boldsymbol{\theta}_t)^T \ \cdot dR(f_t) \doteq -\boldsymbol{\Sigma}_t \cdot dR(f_t) \ ,$$

where $\boldsymbol{\Sigma}_t = \nabla f(\boldsymbol{\theta}_t) \cdot \nabla f(\boldsymbol{\theta}_t)^T$ is an operator from $\mathcal{F}$ to itself (this is also referred to as the neural tangent kernel (NTK) for the case of neural networks [JGH18]). The loss $L_t = L(\boldsymbol{\theta}_t)$ thus verifies

$$\partial_t L_t = -dR(f_t)^T \cdot \boldsymbol{\Sigma}_t \cdot dR(f_t) \ .$$

The operator $\boldsymbol{\Sigma}_t$ is symmetric and admits an eigen-decomposition; if, for all t, all of its eigenvalues are bounded from below by a positive constant $\lambda > 0$, then it follows that

$$\partial_t L_t \leq -\lambda \|dR(f_t)\|^2 \,,$$

which implies convergence of $\boldsymbol{\theta}_t$ to a stationary points exponentially fast. Moreover, if the functional R is strongly convex, this stationary points is a global minima. Consider now the case of scalar-valued one-hidden-layer networks with N units, in the ERM setting. In this case we can write

$$f(\boldsymbol{\theta}) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \boldsymbol{\psi}(\boldsymbol{\theta}^k) \,, \tag{4.8}$$

where each $\boldsymbol{\psi}(\boldsymbol{\theta}_k) \in \mathbb{R}^n$ represents a unit of the network evaluated on the n data points and $\boldsymbol{\theta}^k$ denotes the component of the parameter corresponding to the same unit. If the components $\{\boldsymbol{\theta}_0^k\}_k$ of the initial parameters $\boldsymbol{\theta}_0$ are initialized i.i.d. at random according to some fixed distribution (e.g. a Gaussian), as it is often done in practice, then one gets that

$$\boldsymbol{\Sigma}_0 = \frac{1}{N} \sum_{k=1}^{N} \nabla \boldsymbol{\psi}_k(\boldsymbol{\theta}_0^k) \cdot \nabla \boldsymbol{\psi}_k(\boldsymbol{\theta}_0^k)^T \to \mathbb{E}\ \nabla \boldsymbol{\psi}_k(\boldsymbol{\theta}_0^1) \cdot \nabla \boldsymbol{\psi}_k(\boldsymbol{\theta}_0^1)^T \ \in \mathbb{R}^{n \times n}$$

as $N \to \infty$ by the law of large numbers. If the matrix in the RHS above is positive definite, then one can lower bound the eigenvalues of $\boldsymbol{\Sigma}_0$ at finite, large enough, N, by a concentration argument. Under suitable assumptions on the model, one can use this fact to show that, in the dynamics defined by GD (4.7), the inner weights of the network do not deviate much from initialization (the deviation scaling as $N^{-1/2}$ in N) and the matrix $\boldsymbol{\Sigma}_t$ remains positive semi-definite. Putting all of this together, one gets convergence to a global minimum for N large enough, with high probability over initialization. A series of works estabilished this fact in detail, under different assumption on the model, and extending the idea to actual GD and SGD iterations [Dan17b, LL18, AZLL18, AZLS19, DLL+19, ADH+19, OS20]. A major drawback of these results is that they

require the network size to increase polynomially in the number of data points n, which is usually quite undesirable; notice that this corresponds to the case described in Corollary 4.6. In [OS20], the authors notice that in fact, for two layer networks with N units, in the regime $N \gtrsim n\text{polylog}n$ zero error can be achieved by the random features method outlined in section 4.5.1.

As mentioned, this type of argument implies that the inner layer of the network do not deviate much from initialization. In fact, in the note [COB18], the authors point out how this is essentially an artifact of the $N^{-1/2}$ scaling in (4.8). The authors also point out how in this regime, the GD dynamics (4.7) do not deviate from those of learning a linearized version of the model

$$L(\boldsymbol{\theta}) = R(f(\boldsymbol{\theta}_0) + \nabla f(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) \, ,$$

For this reason, the authors refer to this regime as *lazy training*.

4.5.2.2 One-hidden-layer networks with infinite width: a mean-field limit

Consider again the case of one-hidden-layer neural networks as in (4.8), but scaled as

$$f(\boldsymbol{\theta}) = \frac{1}{N} \sum_{k=1}^{N} \psi(\boldsymbol{\theta}_k) \, .$$

In this case, one can write equivalently the network as

$$f(\boldsymbol{\theta}) = \int_{\Theta} \boldsymbol{\psi}(\boldsymbol{\theta}) \, d\pi_N(\boldsymbol{\theta}) \, ,$$

where $\pi_N = \frac{1}{N} \sum_{k=1}^{N} \delta_{\boldsymbol{\theta}^k}$ is a probability measure. The GD dynamics (4.7) can then equivalently be defined in terms of the evolution of π_N, via a gradient flow. A number of works have been devoted to studying the convergence of such dynamics, in the limit as $N \to \infty$ [CB18, MMN18, RVE18a, SS18], which, roughly speaking, corresponds to the case of model functions varying in the space $\mathcal{H}_\sigma^1$ introduced in section 1.2.1. Numerical experiments also seem to suggest that

this regime seems more appropriate to describe the advantages of over-parametrization for neural networks optimization [COB18]. Unfortunately, while the NTK approach could be carried out for the case of deeper networks as well, such mean field approach is not so easily extendable beyond the case of one-hidden-layer networks.

4.5.2.3 How do moderate changes in width affect the optimization landscape?

All the approaches described so far in this section describe how the optimization landscape of two layer neural networks *improves* in the asymptotic limit $N \to \infty$. In fact, by looking at the proof of Theorem 4.10, one can notice that the *adversarial* data distribution in section 4.4 is *width-dependent*: increasing the hidden layer width (in some cases of even a single unit), under the same data distribution, might potentially eliminate spurious valleys. A similar observation, along with some analysis, has been carried out in [SYS20], for the case of ReLU activations and Gaussian inputs. Understanding how mild over-parametrization affects the optimization landscape, given a data distribution, is an important future direction of research.

Chapter 5

Some related questions and open problems

In this chapter we discuss a potpourri of topics that are related to some of the problems considered in the previous chapters. We discuss a few related results and pose a few questions.

5.1 Approximation of convex bodies by zonoids

The problem of approximation even functions by $\mathcal{F}^{\mathrm{abs},0} \doteq \cup_N \mathcal{F}_N^{\mathrm{abs},0}$ has a nice geometric interpretation. Consider an element f of $\mathcal{F}^{\mathrm{abs},0}$; this has the form, for some $\mathbf{w}_k, \mathbf{v}_k \in \mathbb{R}^d$, $N_1, N_2 \geq 0$,

$$f(\mathbf{x}) = \sum_{k=1}^{N_1} \mathbf{w}_k^T \mathbf{x} \; - \sum_{k=1}^{N_2} \mathbf{v}_k^T \mathbf{x} \; = \sup_{\mathbf{t}\in[-1,1]^{N_1}} \sum_{k=1}^{N_1} t_k \mathbf{w}_k^T \mathbf{x} \; - \sup_{\mathbf{t}\in[-1,1]^{N_2}} \sum_{k=1}^{N_2} t_k \mathbf{v}_k^T \mathbf{x}$$
$$= \sup \; \mathbf{y}^T \mathbf{x} : \mathbf{y} = \sum_{k=1}^{N_1} t_k \mathbf{w}_k \, , \mathbf{t} \in [-1,1]^{N_1} \quad - \sup \; \mathbf{y}^T \mathbf{x} : \mathbf{y} = \sum_{k=1}^{N_2} t_k \mathbf{v}_k \, , \mathbf{t} \in [-1,1]^{N_2} \quad .$$

For a compact set convex body[1] $K \subset \mathbb{R}^d$, the function

$$s[K] : \mathbf{x} \in \mathbb{S}^{d-1} \mapsto \sup_{\mathbf{y}\in K} \mathbf{y}^T \mathbf{x}$$

[1] A set $K \subset \mathbb{R}^d$ is called a convex body if it is convex, compact and has non-empty interior.

is called support function of K. Thus, the space $\mathcal{F}^{\mathrm{abs},0}$ consists of all those function f which can be expressed as the difference of support functions of two set Z_+, Z_-, that is

$$f = s[Z_+] - s[Z_-] \,,$$

where the each of the sets Z_+, Z_- has the form

$$\mathcal{Z}(\mathbf{u}_1, \ldots, \mathbf{u}_N) \doteq \quad \sum_{k=1}^{N} t_k \mathbf{u}_k : \mathbf{t} \in [-1, 1]^N \tag{5.1}$$

for some $\mathbf{u}_1, \ldots, \mathbf{u}_N \in \mathbb{R}^d$, $N \geq 1$. Sets of the form (5.1) are known as (centrally symmetric) *zonotopes*. A zonotope is a polytope defined as sums of a finite number N of segments. Zonotopes represent a class of polytopes with certain regularity properties; in particular, a (centrally symmetric) polytope is a zonotope if and only if all of its faces are centrally symmetric [Sch14].

Elements of $\mathcal{H}^1$ are limit of functions in $\mathcal{F}^{\mathrm{abs},0}$, and this has an equivalent geometric interpretations. Let $h_\pi \in \mathcal{H}^1$ and assume, without loss of generality, that π is a finite non-negative Radon measure. Then h_π is the support function of a *zonoid*; the inverse is also true [Sch14]. A zonoid is defined as a limit of zonotopes in the Hausdorff metric. For two convex bodies K_1, K_2, the Haussdorff distance is defined as

$$d_H(K_1, K_2) = \max\left\{ \max_{\mathbf{x} \in K_1} d(\mathbf{x}, K_2), \max_{\mathbf{x} \in K_2} d(\mathbf{x}, K_1) \right\} .$$

The Haussdorf distance can also be formulated in terms of support functions, as it holds that $d_H(K_1, K_2) = \|s[K_1] - s[K_2]\|_\infty$ (where the infinity norm refers to the infinity norm over $\mathbb{S}^{d-1}$). Therefore the problem of approximating h_π by one-hidden-layer networks of finite width can be interpreted as approximating the respective zonoid by zonotopes.

Support functions of convex bodies also have a functional analysis interpretation. To each

convex body $K \subset \mathbb{R}^d$ one can associate a norm on $\mathbb{R}^d$:

$$\|\mathbf{x}\|_K \doteq \inf \ \ t > 0 : t^{-1}\mathbf{x} \in K \quad .$$

Reversely, to each norm $\|\cdot\|$ on $\mathbb{R}^d$, one can associate a convex body, the respective unit ball $B_{\|\cdot\|} \doteq \ \ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1 \quad$. This relation is bijective. For each norm $\|\cdot\|$, its dual (norm) is defined as the norm

$$\|\mathbf{x}\|_* \doteq \sup_{\mathbf{y} \in \mathbb{R}^d \,:\, \|\mathbf{y}\| \leq 1} \mathbf{x}^T\mathbf{y} = s[B_{\|\cdot\|}](\mathbf{x}) \, .$$

Similarly, for any convex body $K \subset \mathbb{R}^d$, its dual (or *polar*) is defined as the convex body

$$K^\circ \doteq \ \ \mathbf{x} \in \mathbb{R}^d : \sup_{\mathbf{y} \in K} \mathbf{x}^T\mathbf{y} \leq 1 \quad = \quad \mathbf{x} \in \mathbb{R}^d : s[K](\mathbf{x}) \leq 1 \quad .$$

It is easy to see that these definitions are strictly related; indeed, notice that it holds

$$\|\cdot\|_{K,*} = \|\cdot\|_{K^\circ} \qquad \text{and} \qquad B^\circ_{\|\cdot\|} = B_{\|\cdot\|_*} \, .$$

With these definition, one can interpret support functions as norms, and vice-versa:

$$s[K](\mathbf{x}) = \|\mathbf{x}\|_{K^\circ}$$

Consider now the following problem. Let $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^d$ such that $\|\mathbf{a}_k\|_2 = 1$ for all k, with $N > d$. Define the following norm:

$$\|\mathbf{x}\|_{\mathbf{A}} = \inf \ \ \|\mathbf{z}\|_1 : \mathbf{x} = \mathbf{A}\mathbf{z}, \ \mathbf{z} \in \mathbb{R}^N \quad ,$$

where $\mathbf{A} = [\mathbf{a}_1|\cdots|\mathbf{a}_N] \in \mathbb{R}^{d \times N}$. In a compressed sensing view, this correspond to the ℓ^1 norm of the sparsest signal $\mathbf{z}$ that recover the measurement $\mathbf{x}$ (though, notice that, in the compressed

sensing case, one usually assume $\mathbf{x}$ to admit an actual sparse $\mathbf{z}$ such that $\mathbf{x} = \mathbf{A}\mathbf{z} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is some small error; in the general case this is not the case, and it only holds that $\|\mathbf{z}\|_0 = d$). An interesting question is whether it is possible to efficiently approximate the norm $\|\cdot\|_{\mathbf{A}}$ by shallow models, that is, by $\mathcal{F}_N^{\text{abs},0}$. The norm $\|\cdot\|_{\mathbf{A}}$ can be expressed as the support function of a polytope:

$$\|\mathbf{x}\|_{\mathbf{A}} = s[\mathcal{C}_{\mathbf{A}}^\circ](\mathbf{x}) \,,$$

where $\mathcal{C}_{\mathbf{A}}$ is the convex hull of the points $\{\pm\mathbf{a}_1, \dots, \pm\mathbf{a}_N\}$, that is

$$\mathcal{C}_{\mathbf{A}} = \quad \mathbf{x} = \mathbf{A}\mathbf{z} : \mathbf{z} \in \mathbb{R}^N, \ \|\mathbf{z}\|_1 \leq 1 \quad .$$

In general, $\mathcal{C}_{\mathbf{A}}^\circ$ is not a zonoid; in fact, since it is a polytope, it is a zonoid if and only if it is a zonotope[2] [Sch14, Corollary 3.5.7]. As such, it is not possible to approximate $\mathcal{C}_{\mathbf{A}}^\circ$ (respectively, $\|\cdot\|_{\mathbf{A}}$) by zonotopes (respectively, elements of $\mathcal{F}_N^{\text{abs},0}$ with positive weights in the second layer). In some cases it is even possible to quantify the distance between the set $\mathcal{C}_{\mathbf{A}}^\circ$ and the class of zonoid. Consider the infinity norm $\|\cdot\|_\infty$ on $\mathbb{R}^d$. It can be expressed in the previous notation as $\|\cdot\|_\infty = \|\cdot\|_{\mathcal{E}_d}$, where $\mathcal{E}_d = \{\pm 1\}^d$ is the d-dimensional hypercube (with abuse of notation, we indentify the set $\mathcal{E}_d$ with the matrix whose columns are elements of $\mathcal{E}_d$). In particular, the infinity norm is the support function of the cross polytope $\Delta_d \doteq \quad \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 \leq 1 \quad$. The following holds.

Proposition 5.1. *Let $Z \subset \mathbb{R}^d$ be a zonoid. Then it holds that*

$$\left\| \|\cdot\|_\infty - s[Z] \right\|_\infty = d_H(\Delta_d, Z) \geq \frac{1}{e} - \frac{1}{\sqrt{d}} \,.$$

A proof of Proposition 5.1 is given in section D.1.1, based on a reduction from a result in [HLW10]. Although, going back to general case of a convex hull $\mathcal{C}_{\mathbf{A}}$, even if the polar $\mathcal{C}_{\mathbf{A}}^\circ$ is

[2] Any polytope is a a zonotope only if $d = 2$.

at positive distance from any zonoid, by the UAT (Theorem 1.1), we know that it is possible to approximate $\|\cdot\|_{\mathbf{A}}$ by difference of two zonotopes Z_+, Z_-. Notice that

$$\|s[\mathcal{C}^\circ_{\mathbf{A}}] - (s[Z_+] - s[Z_-])\|_\infty = d_H(\mathcal{C}^\circ_{\mathbf{A}} + Z_-, Z_+) \, .$$

Therefore, the problem can be geometrically interpreted as finding a 'regularization' of the polytope $\mathcal{C}^\circ_{\mathbf{A}}$ (obtained by summing it with a zonoid Z_-) such that the results is a zonoid. The efficiency question that we pose here is to understand whether this regulation can be made efficiently, that is by approximate steps using zonotopes with a $\mathrm{poly}(d)$ number of generating segments. We conjecture that this is not possible in the random case (that is for $\mathbf{a}_1, \dots, \mathbf{a}_N$ generated independently uniformly over the sphere $\mathbb{S}^{d-1}$) or the case considered above of the infinity norm $\|\cdot\|_\infty$. A possible way to proceed to prove this would be to show that the respective norms satisfy a condition such as in Lemma 3.12. Notice that in the case of the infinity norm $\|\cdot\|_\infty$, the function of interest can be approximated by a (ReLU) network of size $O(d)$ and depth $\lg d$. Proving the above conjecture would then offer another example of depth separation.

5.2 Not only approximation: learnability

In chapter 3, we discussed examples of families of functions which can be efficiently approximated by two-hidden-layers networks but not from one-hidden-layer ones. Although, from a practical point of view, approximability is a necessary property, but not it is not sufficient. In fact, while the family of functions considered in section 3.2 can be approximated by two-hidden-layers networks with polynomial complexity, it is not clear whether this family can be learned via an algorithm with polynomial complexity.

In some recent works [MSS19, MJSSS21], it has been shown that certain families of algorithms can efficiently learn deep networks only if such networks can be *weakly approximated* by shallow

models: this means that the hypothesis class (shallow models) can approximate the target (deep model) better than the trivial classifier, at least up to an inverse polynomial in some measure of complexity of the hypothesis class. Current results operate under the classification setup, and for different families of algorithms depending on the hypothesis and target classes; separations among different families of learning algorithms for deep neural networks was also shown in [AS20]. Inspired by such results and some questions posed in these works, we pose the following open questions, relating to our results in chapter 3.

- Consider a family of two-hidden-layer networks $f^{(d)} : \mathbb{R}^d \to \mathbb{C}$ $_{d\geq 1}$ and a family of probability measures $\mu^{(d)}$ $_{d\geq 1}$ over $\mathbb{R}^d$. Assume that such functions are not weakly approximable by one-hidden-layer networks with polynomial width, that is, there exists no polynomial functions $p, q : \mathbb{N} \to [1, \infty)$ such that

$$\inf_{h\in\mathcal{F}_{q(d)}} \|f^{(d)} - h\|_{\mu_d,2} \leq 1 - \frac{1}{p(d)} .$$

  Can such functions be learned via GD or SGD at a fixed threshold $\epsilon > 0$, that is, can GD or SGD with problem-agnostic $\mathrm{poly}(d)$ hyper-parameters and $\mathrm{poly}(d)$ iterations output a deep neural network $h^{(d)}$ satisfying

$$\|f^{(d)} - h^{(d)}\|_{\mu_d,2} \leq \epsilon \, ?$$

Notice that the families of functions satisfying the hypothesis of Theorem 3.2 are not weakly approximable. If the answer to the questions above is negative, this would mean that, while they provide an example of depth separation from the approximation point of view, such families are not actually learnable with standard algorithms used in deep learning.

5.3 Advantages of learning invariant functions

Many high-dimensional machine learning problems involve highly structured data such as images, text, or graphs, and may exhibit invariance to certain transformations of the input data, such as permutations, translations or rotations, and near-invariance to small deformations. Network architectures can be appropriately defined to exploit such invariances. For example, the success of deep convolutional architectures is often attributed in part to their ability to learn invariant representations of natural signals. On the other hand, if such invariances are not imposed a priori, simple architectures such as shallow neural networks, may fail to provide efficient estimators to invariant functions, as shown in Section 3.4.3.

In section 3.4 we looked at function approximation over the unit sphere $\mathbb{S}^{d-1}$ through the lens of spherical harmonics. In the recent work [MMM21], the authors studied, for groups G acting on $\mathbb{S}^{d-1}$ such as the cyclic group, the spherical harmonic decomposition for G-invariant functions, in the asymptotic limit of $d \to \infty$. Roughly speaking, they showed that, considering invariant functions only, the size of the set of fixed degree harmonic polynomial decreases of a factor $|G|^{-1}$ when we consider invariant polynomials only. This has implications in terms of generalisation bounds for kernel and random features methods, which are shown to benefit from such invariance. In the work [BVB21], we look at a similar set-up, but in the non asymptotic regime of bounded d. Let $\overline{N}_k^d$ be the dimension of the space of degree k harmonic invariant polynomials on $\mathbb{S}^{d-1}$. Roughly speaking, we show that

$$\frac{\overline{N}_k^d}{N_k^d} = \frac{1}{|G|} + o(k^{-\ell_d})$$

for a certain exponent ℓ_d. We find that this fact implies improvements in sample complexity by a factor of the order of the size of the group G when the sample size is large enough. Finally, we show how this analysis can be potentially extended beyond group invariance, estabilishing similar gains for geometrically stable functions. We refer to [BVB21] for more details.

While the discussed analysis are limited to generalization guarantees for kernel-based methods,

they could potentially provide insights for related questions, such as how approximation by neural networks can benefit from exploiting invariances, an important topic for future studies.

5.4 Very deep models

In the previous chapters, we mostly looked at neural networks of fixed depth and quantified their complexity in terms of their width. In contrast, one could instead consider networks whose width is fixed and ask whether approximation, up to a arbitrarily small accuracy, can be achieved by increasing the depth. For the case of ReLU networks, using the fact that such networks are piecewise linear functions, one can show that in fact arbitrarily deep networks of constant width form a class of universal approximators.

Theorem 5.2 ([HS17, Han19])**.** *Let $f : [0,1]^d \to \mathbb{R}$ be a continuous function. Then for every $\epsilon \in (0,1)$ there exists a ReLU network g of width at most $d+1$ such that*

$$\|f - g\|_\infty \leq \epsilon \, .$$

The depth of the network g can be upper bounded by $2d![\omega_f(\epsilon)]^{-d}$, where ω_f is the modulus of continuity of f. Moreover, this result is optimal in terms of minimal width.

These models are not of solely theoretical interest. The first reason is that models of increasing depth are being used in practice, an example being the one of residual networks [HZRS16]. Moreover, there is a variety of different iterative algorithms or models, that can be recast in this form. Chapter 2 provides one such example: the deep reduced models described in section 2.3.1 perform inversion of the characteristic maps by implementing the bisection method. This part of the model provides a network of fixed width whose accuracy to compute the inverse is increased by adding a set of (fixed) layers.

Another model that can be viewed in this way is LISTA [GL10]. Iterative Soft Thresholding

(ISTA, [DDDM04]) is an iterative algorithm to solve the ℓ^1 sparse coding problem

$$\mathbf{z}_{\mathbf{A}}^{\lambda}(\mathbf{x}) = \underset{\mathbf{z}\in\mathbb{R}^N}{\arg\min} \ \frac{1}{2}\|\mathbf{x} - \mathbf{A}\mathbf{z}\|^2 + \lambda\|\mathbf{z}\|_1$$

where $\mathbf{A} \in \mathbb{R}^{d\times N}$ is a given *dictionary* matrix. The relevant fact is that the k-th iteration of ISTA can be written as a deep (ReLU) network of depth k, where each layer performs one (same) iteration step. In its *learnable* version LISTA, the basic structure of the iteration is kept, but the weights can be learnt as in a standard neural network. This allows to obtain, under appropriate assumptions on the dictionary $\mathbf{A}$, deep ReLU networks of constant width which recover a sparse representation of the input $\mathbf{x}$, that is

$$\mathbf{z}_{\mathbf{A}}^{*}(\mathbf{x}) = \arg\min \ \|\mathbf{z}\|_1 : \mathbf{z}\in\mathbb{R}^N,\ \mathbf{x} = \mathbf{A}\mathbf{z} \quad , \tag{5.2}$$

up to an error ϵ, for a depth scaling as $O(\log\frac{1}{\epsilon})$ and input signals $\mathbf{x}$ admitting a sparse enough representation [CLWY18, Theorem 2]. A more explicit construction with the same property can be obtained by the homotopy method [JJL17], and it can actually be shown that the network weights defined by it are essentially optimal [CLWY18, Theorem 1]. This can be interpreted by saying that deep networks can efficiently recover sparse signals (5.2); on the other hand, it is not clear whether this is possible for their shallow counterpart. Notice that this is related to the question posed in section 5.1, although it fundamentally differs in terms of domain of approximation; in sparse coding, one is only interested in input measurements $\mathbf{x}$ with a sparse underlying signal $\mathbf{z}^*(\mathbf{x})$.

Normalizing flows [TVE+10, RM15] are a family of generative models which implement a mapping between a reference easy-to-sample-from distribution to a target distribution. In practical applications, they are implemented by learning a composition of parametric elementary blocks, via minimizing an empirical Kullback–Leibler divergence between the reference and the target distribution. The elementary blocks are defined to allow efficient computations of their gradients,

and their composition provide a neural network model of the sought mapping. Higher accuracy are found by increasing the number of elementary blocks, as at the core of the procedure lies a measure flow converging to transport between the reference and the target distributions [TVE+10]. In this sense, these type of models fall in the class of network with constant width but arbitrary depth.

This non-exhaustive list of examples shows that the regime of constant width and arbitrary depth, different from the view of neural networks of the classical UAT, represents an important regime to be studied and understood.

5.4.1 A multi-level study case

The type of models described above involve composing a (potentially parametric) basic operation, or layer, many times; enrolling these operations describes a *deep* model. Intuitively, the idea is that, at each step, a small portion of the target problem is learned. In some cases, one can take advantage of this to decompose the learning problem in a hierarchy of sub-problems in order to increase the overall algorithm efficiency. In this section, we briefly present an application of this idea to the case of Stein variational gradient descent (SVGD) [LW16], an iterative algorithm to compute a mapping between a reference distribution and a target one. In this sense, SVGD falls in the family of flows discussed above, although it operates by building layers in a RKHS, rather than 'standard' neural networks layers, via gradient based iterations.

Consider the problem of learning a mapping between a reference distribution η on $\mathbb{R}^d$ (such as a standard Gaussian distribution) and a target Gibbs distribution μ on $\mathbb{R}^d$; that is a function $\phi^* : \mathbb{R}^d \to \mathbb{R}^d$, such that $\mu = \phi^*_\# \eta$, where $\phi^*_\# \eta$ denotes the push-forward of η[3]. The SVGD methods constructs mappings $\phi^{(t)} : \mathbb{R}^d \to \mathbb{R}^d$, for $t = 1, 2, \ldots$, from η to a push-forward measure $\mu^t = \phi^{(t)}_\# \eta$ which increasingly approximates μ. The mapping $\phi^{(t)}$ has the form $\phi^{(t)} = \phi_t \circ \cdots \circ \phi_1$,

[3]Given a (measurable) function $f : \mathbb{R}^d \to \mathbb{R}^d$, the push-forward of a measure η on $\mathbb{R}^d$ is the measure $f_\# \eta$ defined as $f_\# \eta(A) = \eta(f^{-1}(A))$, for $A \subseteq \mathbb{R}^d$ measurable.

where each step is constructed as $\phi_t(\mathbf{x}) = \mathbf{x} - \epsilon g_t(\mathbf{x})$ where ϵ is a small step-size and g_t is chosen to maximally decrease the KL divergence of the reference distribution with the target distribution, by solving the following functional optimization,

$$g_t \in \underset{g \in \mathcal{H} \,:\, \|g\|_{\mathcal{H}} \leq 1}{\arg\max} \quad -\frac{d}{d\epsilon}\Big|_{=0} KL((I + \epsilon g)_{\#}\mu_{t-1} \mid \eta) \quad .$$

Above, KL denotes the KL divergence among the two distributions, I denotes the identity mapping over $\mathbb{R}^d$ and $\mathcal{H}$ denotes an RKHS over $\mathbb{R}^d$ to which the mapping g_t is set to belong. If we have access (up to constant) to the target distribution, then the map g_t can be computed explicitely, by approximating the KL divergence by sampling. While it is possible to show that, under certain conditions, the measure μ_t convergences to μ as $t \to \infty$ (see e.g. [Liu17]), a potential drawback of the method is that it can take a large time of iterations to converge. Each iterations involves a number of evaluations (equal to the number of samples used to approximate the KL) of the density of μ; if this density is computationally expensive to compute, the slow convergence represents a computational bottleneck.

Assume now that a sequence of densities $\{\mu^{(k)}\}_{k \geq 1}$ is available, such that it converges weakly to the target density μ as $k \to \infty$. Each density has an associated the computational cost, so that it is significantly cheaper to evaluate $\mu^{(k)}$ for low values of k rather than for large values of k. In this setting, a possible idea is to take advantage of the compositional structure of the mapping $f^{(t)}$ by sequentially applying SVGD on the sequence $\{\mu^{(k)}\}_k$. This defines a multilevel version of SVGD. The resulting approximate mapping takes the form $f_{\mathrm{ML}}^{(T_L)} = f^{(k_L),(t_L)} \circ \cdots \circ f^{(k_1),(t_1)}$, where $T_L \doteq \sum_{\ell=1}^{L} t_\ell$, and $f^{(k_\ell),(t_\ell)}$ is constructed by taking t_ℓ steps of SVGD from the (previous levels) reference measure to the target measure $\mu^{(k_\ell)}$. Calling c_k (respectively c_∞) the cost of evaluating $\mu^{(k)}$ (respectively μ), the cost to construct the SVGD mapping, in the standard and the multi-level setting, are proportional, respectively, to $c_\infty T$ and $\sum_{\ell=1}^{L} c_{k_\ell} t_\ell$.

In the case of certain Bayesian inverse problems described by parametric PDEs, a hierarchy

of measure $\{\mu^{(k)}\}$ can be constructed by considering increasingly accurate discretizations of the PDE domain. An heuristic criteria to run the multilevel version of SVGD is to run it for each level until the empirical gradient magnitude is below a certain threshold. This allows to recover the same accuracy (or more) of standard SVGD while notably reducing the computational effort. This is shown empirically in the work [AVP21], joint with Terrence Alsup and Benjamin Peherstorfer, which we refer to for further details. The work also presents theoretical justifications for this fact, although the author did not contribute to this part. Interestingly, a similar multilevel approach does not seem to yield, empirically, the same advantageous results when applied to a different method to construct deep mappings, such as the one discussed in [Par15].

Appendix A

Appendix to chapter 2

A.1 Proof of results on approximation by standard neural networks

A.1.1 Proof of Lemma 2.3

We prove Lemma 2.3 by showing a series of intermediate lemmas.

Lemma A.1. *Let $f : [0,1] \to \mathbb{R}$ be a C^1 function such that $c_1 \leq \| f \|_\infty \leq c_2$ for some $c_1, c_2 > 0$. Then, for any $a_1, \ldots, a_N \in \mathbb{R}$ there exists $[t_0, t_1] \subseteq [0,1]$ such that $t_1 - t_0 \geq c_1 (4 c_2 N)^{-1}$ and*

$$\inf_{t \in [t_0, t_1]} \inf_{k \in [N]} |f(t) - a_k| \geq \frac{c_1}{8N} .$$

Proof. Without loss of generality, we can assume that $c_1 \leq f(x) \leq c_2$ for all $x \in [0,1]$. This implies that f is increasing and takes values in $[f(0), f(1)]$, where

$$f(1) - f(0) \geq c_1 .$$

Define $f(0) = \alpha_0 < \cdots < \alpha_{M+1} = f(1)$ such that

$$\{\alpha_k\}_{k=0}^{M+1} = \quad \{a_k\}_{k\in[N]} \cap [f(0), f(1)] \quad \cup \{f(0), f(1)\} \, .$$

In particular, $M \leq N$. There exists $\bar{t} \in [0,1]$ such that $f(\bar{t}) = (\alpha_{k+1} + \alpha_k)/2$ for some $k \in [0, M]$ and such that

$$\inf_{k\in[0,M+1]} |f(\bar{t}) - \alpha_k| \geq \frac{c_1}{4N} \, .$$

Since f is c_2-Lipschitz, for any $t \in [\bar{t} - \frac{c_1}{8c_2N}, \bar{t} + \frac{c_1}{8c_2N}]$, it holds

$$|f(t) - \alpha_k| \geq |f(\bar{t}) - \alpha_k| - |f(t) - f(\bar{t})| \geq \frac{c_1}{4N} - c_2|\bar{t} - t| \, .$$

The result follows. □

Lemma A.2. *Let $f : [0,1] \to \mathbb{R}$ be a C^1 function and let $a_1, \ldots, a_N \in \mathbb{R}$ such that*

$$\inf_{t\in[0,1]} \inf_{k\in[N]} |f\,(t) - a_k| \geq \epsilon \, .$$

Then, for any $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^M$, it holds that

$$\sup_{t\in[0,1]} \inf_{\substack{k\in[N] \\ j\in[M]}} |f(t) - a_k t - b_{k,j}| \geq \frac{\epsilon}{4NM} \, .$$

Proof. Let $g_{k,j}(t) = f(t) - a_k t - b_{k,j}$ for $t \in [0,1]$ and $k \in [N]$, $j \in [M]$. By assumption, it holds that either $g_{k,j}(t) \geq \epsilon$ for all $t \in [0,1]$ or $g_{k,j}(t) \leq -\epsilon$ for all $t \in [0,1]$. This implies that $g_{k,j}$ is either strictly increasing or strictly decreasing and thus there exists only a point $t_{k,j} \in [0,1]$ such that

$$|g_{k,j}(t_{k,j})| = \min_{t\in[0,1]} |g_{k,j}(t)| \, .$$

In particular, it either holds that $t_{k,j} \in \{0,1\}$ and $|g_{k,j}(t_{k,j})| > 0$ or that $|g_{k,j}(t_{k,j})| = 0$. Now, let

$0 = x_0 < \cdots < x_{K+1} = 1$ such that

$$\{x_k\}_{k=0}^{K+1} = \{t_{k,j}\}_{k\in[N],j\in[M]} \cup \{0,1\} \,.$$

In particular, $K \leq NM$. Then there exists $\bar{t} = (x_k + x_{k+1})/2$ for some $k \in [0, K]$ such that

$$\inf_{k\in[0,K+1]} |x_k - \bar{t}| \geq \frac{1}{4NM} \,.$$

This implies that, for any $k \in [N]$ and $j \in [M]$, it holds

$$|g_{k,j}(\bar{t})| = \; g_{k,j}(t_{k,j}) + \int_{t_{k,j}}^{\bar{t}} g_{k,j}(t)\, dt \;\geq |g_{k,j}(t_{k,j})| + \epsilon|\bar{t} - t_{k,j}| \geq \frac{\epsilon}{4NM} \,.$$

This concludes the proof. □

Lemma A.3. *Let $f : [a, b] \to \mathbb{R}$ be a C^2 function such that $c_1 \leq \|f\ \|_\infty \leq c_2$ for some $c_2, c_1 > 0$. Then it holds that*

$$\inf_{\substack{a_1,\ldots,a_N\in\mathbb{R}\\ \mathbf{b}_1,\ldots,\mathbf{b}_N\in\mathbb{R}^M}} \sup_{t\in[a,b]} \inf_{\substack{k\in[N]\\ j\in[M]}} |f(t) - a_k t - b_{k,j}| \geq \frac{C}{MN^3}$$

where $C > 0$ is a constant that depends only on c_1,c_2 and $(b - a)$.

Proof. Fix any $a_1, \ldots, a_N \in \mathbb{R}$ and $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^M$, and let $\tilde{a}_k = \frac{4c_2N}{c_1(b-a)} a_k$ and $\tilde{\mathbf{b}}_k = \frac{4c_2N}{c_1(b-a)} \mathbf{b}_k$. Let $g(t) = f\,(a + (b - a)t)$ for $t \in [0, 1]$. Then $(b - a)c_1 \leq \|g\ \|_\infty \leq (b - a)c_2$. Thanks to Lemma A.1, there exists $[s_0, s_1] \subseteq [0, 1]$ such that

$$s_1 - s_0 = \frac{c_1}{4c_2N} \quad \text{and} \quad \inf_{s\in[s_0,s_1]} \inf_{k\in[N]} |g(s) - \tilde{a}_k| \geq \frac{c_1(b-a)}{8N} \,.$$

This is equivalent to say that there exists $[t_0, t_1] \subseteq [a, b]$ such that

$$t_1 - t_0 = \frac{c_1(b-a)}{4c_2N} \quad \text{and} \quad \inf_{t\in[t_0,t_1]} \inf_{k\in[N]} |f\,(t) - \tilde{a}_k| \geq \frac{c_1(b-a)}{8N} \,.$$

Let $h(t) = f(t_0 + t(t_1 - t_0))/(t_1 - t_0)$ for $t \in [0,1]$. It follows that

$$\inf_{t\in[0,1]} \inf_{k\in[N]} |h\,(t) - \tilde{a}_k| \geq \frac{c_1(b-a)}{8N}.$$

By Lemma A.2, it holds that

$$\sup_{t\in[0,1]} \inf_{\substack{k\in[N]\\ j\in[M]}} h(t) - \tilde{a}_k t - \tilde{b}_{k,j} \;\geq \frac{c_1(b-a)}{32MN^2}$$

which implies that

$$\sup_{t\in[t_0,t_1]} \inf_{\substack{k\in[N]\\ j\in[M]}} |f(t) - a_k t - b_j| \geq \frac{c_1^2(b-a)^2}{128c_2MN^3}\,.$$

This concludes the proof. □

We now conclude with the proof of Lemma 2.3. Let $f : [0,1]^d \to \mathbb{R}$ be a C^2 function which is non linear, and fix any $\mathbf{a}_1, \ldots, \mathbf{a}_N \in \mathbb{R}^d$ and $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^M$. Since f is non-linear, there exists $\mathbf{u} \in (0,1)^d$ and $\mathbf{v} \in \mathbb{R}^d$ such that $\mathbf{u} + t\mathbf{v} \in (0,1)^d$ for all $t \in [0,1]$ and such that, if

$$f_{\mathbf{v}} : t \in (-\epsilon, \epsilon) \mapsto f(\mathbf{u} + t\mathbf{v})$$

satisfies $c_1 \leq \|f_{\mathbf{v}}\|_\infty \leq c_2$ for some $c_1, c_2 > 0$. Then it holds that

$$\sup_{\mathbf{x}\in[0,1]^d} \inf_{k\in[N],j\in[M]} f(\mathbf{x}) - \mathbf{a}_k^T\mathbf{x} - b_{k,j} \;\geq \sup_{t\in[0,1]} \inf_{k\in[N],j\in[M]} |f_{\mathbf{v}}(t) - c_k t - d_{k,j}|\,,$$

where $c_k = \mathbf{v}^T\mathbf{a}_k$ and $d_{k,j} = b_{k,j} + \mathbf{u}^T\mathbf{a}_k$. Applying Lemma A.3 concludes the proof.

A.1.2 Generalizing Proposition 2.4 to different initial conditions

In this section we consider the problem of obtaining lower bounds as in section 2.3. We start by discussing how, under the same set-up, a similar proof technique can yield lower bounds for differ-

ent initial conditions, and discuss a few particular cases. We also discuss how different techniques can yield stronger lower bound, depending on the initial condition of the PDE.

Consider a PDE as in (2.7), and recall that there exists $\iota, \nu > 0$ such that the function $X(t, x; \boldsymbol{\mu})$ satisfies

$$0 < \iota \leq \partial_x X(t, x; \boldsymbol{\mu}) \leq \nu$$

for every $(x, t, \boldsymbol{\mu}) \in \Omega \times [0, 1] \times \mathcal{D}$. Consider a one-hidden-layer network $f_N \in \mathcal{F}_N^\sigma$; for sake of simplicity we consider here the case of σ being the ReLU but the following ideas can be generalized to semi-algebraic activations. Following the proof of Proposition 2.4, a way to proceed is to track the approximation of the solution over time in the transport of a point $x_0 \in \Omega$. Let $\alpha(t, \boldsymbol{\mu})$ be the closest breakpoint of f_N to $X(t, x_0; \boldsymbol{\mu})$ and

$$\epsilon(t, \boldsymbol{\mu}) = |X(t, x_0; \boldsymbol{\mu}) - \alpha(t, \boldsymbol{\mu})| \,.$$

Consider the component of the error in the interval

$$I(t, \boldsymbol{\mu}) = [X(t, x_0; \boldsymbol{\mu}) - \epsilon(t, \boldsymbol{\mu}), X(t, x_0; \boldsymbol{\mu}) + \epsilon(t, \boldsymbol{\mu})] \,;$$

we have

$$\begin{aligned}
\|u(\cdot, t; \boldsymbol{\mu}) - f_N(\cdot, t, \boldsymbol{\mu})\|_{\mathbb{V}} &\geq \inf_{a,b \in \mathbb{R}} \int_{X(t,x_0;\boldsymbol{\mu})- \ (t,\boldsymbol{\mu})}^{X(t,x_0;\boldsymbol{\mu})+ \ (t,\boldsymbol{\mu})} u_0(X^{-1}(t, x; \boldsymbol{\mu})) - ax - b \ ^2 \, dx \\
&\geq \inf_{a,b \in \mathbb{R}} \int_{X^{-1}(X(t,x_0;\boldsymbol{\mu})- \ (t,\boldsymbol{\mu}),t;\boldsymbol{\mu})}^{X^{-1}(X(t,x_0;\boldsymbol{\mu})+ \ (t,\boldsymbol{\mu}),t;\boldsymbol{\mu})} (u_0(y) - aX(t, y; \boldsymbol{\mu}) - b)^2 |\partial_y X(t, y; \boldsymbol{\mu})| \, dy \\
&\geq \iota \inf_{a,b \in \mathbb{R}} \int_{x_0 - \nu^{-1}}^{x_0 + \nu^{-1}} (u_0(y) - aX(t, y; \boldsymbol{\mu}) - b)^2 \, dy \\
&= \frac{\iota}{\nu} \inf_{a,b \in \mathbb{R}} \int_{-} u_0(x_0 + \nu^{-1} z) - aX(t, x_0 + \nu^{-1} z; \boldsymbol{\mu}) - b \ ^2 \, dz \doteq \Gamma(\epsilon; t, \boldsymbol{\mu}) \,.
\end{aligned}$$

Proceeding as in the proof of Proposition 2.4, we then obtain the following.

Lemma A.4. *Let* $\Gamma(\epsilon) \doteq \inf_{t\in[0,1],\boldsymbol{\mu}\in\mathcal{D}} \Gamma(\epsilon; t, \boldsymbol{\mu})$. *Then it holds that*

$$\sup_{(t,\boldsymbol{\mu})\in[0,1]\times\mathcal{D}} \|u(\cdot, t; \boldsymbol{\mu}) - f_N(\cdot, t, \boldsymbol{\mu})\|_{\mathbb{V}} \geq \ \Gamma(N^{-3})^{\ 1/2} .$$

In some cases, the leading behaviour of Γ around the origin can be computed explicitly. This is indeed what allows the proof of Proposition 2.4 to carry over. In this case, if $u_0(x) = \mathbb{1}\{x \leq x_0\}$, it holds that

$$u_0(X^{-1}(t, x; \boldsymbol{\mu})) = u_0(x + x_0 - X(t, x_0; \boldsymbol{\mu}))$$

Therefore, it is equivalent to consider $X(t, x; \boldsymbol{\mu}) = x + X(t, x_0; \boldsymbol{\mu}) - x_0$ (which gives $\iota = \nu = 1$). In this case one gets that

$$\begin{aligned}\Gamma(\epsilon; t, \boldsymbol{\mu}) &= \inf_{a,b} \ _- (\mathbb{1}\{z \leq 0\} - az - aX(t, x_0; \boldsymbol{\mu}) - b)^2 \, dz \\ &= \inf_{a,b} \ _- (\mathbb{1}\{z \leq 0\} - az - b)^2 \, dz \gtrsim \epsilon \, .\end{aligned}$$

For different initial conditions, one recovers the same results in the case that the transport map is linear in the spatial variable, that is $X(t, x; \boldsymbol{\mu}) = a(t, \boldsymbol{\mu})x + b(t, \boldsymbol{\mu})$. Notice that it equivalent to ask that the term c in the PDE is a linear function of x. In this case it holds

$$\begin{aligned}\Gamma(\epsilon; t, \boldsymbol{\mu}) &= \frac{\iota}{\nu} \inf_{a,b} \ _- u_0(x_0 + \nu^{-1}z) - a \cdot a(t, \boldsymbol{\mu}) \cdot \nu^{-1} \cdot z - a \cdot a(t, \boldsymbol{\mu})x_0 - a \cdot b(t, \boldsymbol{\mu}) - b \ ^2 \, dz \\ &= \inf_{a,b} \ _- u_0(x_0 + \nu^{-1}z) - az - b \ ^2 \, dz = \Gamma(\epsilon) \, .\end{aligned}$$

Depending on the choice of u_0, $\Gamma(\epsilon)$ can be shown to yield different rates as $\epsilon \to 0$. The following corollaries are two examples of this.

Corollary A.5. *If* $u_0 \in C^{s-1}(\Omega) \setminus C^s(\Omega)$ *is a piece-wise polynomial of degree* s*, with a break point in* $x_0 \in \Omega$*, then* $\Gamma(\epsilon) \gtrsim \epsilon^{2s+1}$.

Proof. By assumption, there exist $\alpha, \beta \in \mathbb{R}$, $\alpha \neq \beta$, and a polynomial p of degree $s-1$ such that

$$\begin{aligned}\Gamma(\epsilon; t, \boldsymbol{\mu}) &= \inf_{a,b} \int_{-}^{0} (\alpha z^s + p(z) - az - b)^2 \, dz + \int_{0} (\beta z^s + p(z) - az - b)^2 \, dz \\ &\geq \inf_{q:\deg(q)<s} \int_{-}^{0} (\alpha z^s - q(z))^2 \, dz + \int_{0} (\beta z^s + p(z) - q(z))^2 \, dz \quad ,\end{aligned}$$

where the infimum in the last equation is taken over all polynomials of degree at most $s-1$. Using the change of variables $z = \epsilon u$, we get

$$\begin{aligned}\Gamma(\epsilon; t, \boldsymbol{\mu}) &\geq \epsilon \inf_{q:\deg(q)<s} \int_{-}^{0} (\alpha \epsilon^s z^s - q(z))^2 \, dz + \int_{0} (\beta \epsilon^s z^s + p(z) - q(z))^2 \, dz \\ &= \epsilon^{2s+1} \inf_{q:\deg(q)<s} \int_{-}^{0} (\alpha z^s - q(z))^2 \, dz + \int_{0} (\beta z^s + p(z) - q(z))^2 \, dz \quad .\end{aligned}$$

This concludes the proof. □

Corollary A.6. *If the function u_0 is C^3 and not affine in a neighborhood of x_0, then $\Gamma(\epsilon) \gtrsim \epsilon^5$.*

Proof. This follows by applying Corollary C.3 in [PV18]. □

Notice that nevertheless, if $\Gamma(\epsilon) \lesssim \epsilon^s$ for some $s > 8/3$, a stronger lower bound holds (of the order of N^{-4}), if the resulting solution is C^3 non-linear in a region of the domain $\Omega \times [0,1] \times \mathcal{D}$. We further remark that the lower bound in [PV18] can be improved, for low dimensions ($d \leq 3$). This lower bound thus applies to all piece-wise C^3 smooth functions. We prove this in Section A.1.2.1. Moreover, similar lower bounds apply to solutions of other transport problems, such as the wave equation or certain linear transport equations with spatially dependent speed. We provide examples of this in Section A.1.2.2.

A.1.2.1 Enhanced lower bounds for smooth functions

We start by showing the following lemma.

Lemma A.7. *Consider K hyperplanes $a_i^T y = b_i$, $i \in [K]$, in $\mathbb{R}^d$, for $d \leq 3$. Then there exist at least one d-dimensional cube of the form $Q_d^{N,i} \doteq \bigtimes_{k=1}^{d} \left[\frac{i_k}{N}, \frac{i_k+1}{N}\right]$ for $i \in [\![0, N-1]\!]^d$ such that*

$$\mathring{Q}_d^{N,i} \cap \bigcup_{i=1}^{K} \{x : a_i^T x = b_i\} = \emptyset$$

as long as

$$K \leq c_d N - 1\,,$$

where $c_d \in (0, 1]$ is a constant only depending on d.

Proof. This is obvious for $d = 1$, with $c_1 = 1$. For the case $d \geq 2$, consider a cube a split of $[0, 1]^d$ in with a (N, d)-grid:

$$[0, 1]^d = \bigcup_{i \in [\![0, N-1]\!]^d} Q_d^{N,i}\,.$$

Given an hyperplane $\ell \doteq \{x : a^T x = b\}$, we wish to count the number of cubes $Q_d^{N,i}$ *intercepted* by the line:

$$n \doteq \#\{i \in [\![0, N-1]\!]^d : \ell \cap \mathring{Q}_d^{N,i} \neq \emptyset\}\,.$$

Let's consider the case $d = 2$ first; in this case hyperplanes are lines. We claim that in this case $n \leq 2N - 1$. Assume, w.l.o.g., that $a_1 a_2 < 0$ (the line is thus an increasing function in the first coordinate). Let t_k be the number of cubes that the line intercepts in the *stripe* $\left[\frac{k}{N}, \frac{k+1}{N}\right] \times [0, 1]$, for $k \in [0, N-1]$. Clearly, $n = \sum_{i=0}^{N-1} t_k$. In particular, the cubes intercepted in the k-th stripe are given by

$$\mathring{Q}_2^{N,(k,j)}$$

for $a_k \leq j \leq b_k$, with $t_k = c_k - b_k + 1$. Moreover, since the line is increasing, we know that $b_k \geq c_{k-1}$, for $k \in [N-1]$, which implies $t_k \leq c_k - c_{k-1} + 1$. Finally, notice that $c_k \leq N - 1$ and

$b_k \geq 0$. It follows that

$$n = \sum_{k=0}^{N-1} t_k = \sum_{k=0}^{N-1} (c_k - b_k + 1) \leq c_0 - b_0 + \sum_{k=1}^{N-1} (c_k - c_{k-1}) + N = c_{N-1} - b_0 + N \leq 2N - 1\,.$$

By an union bound, it follows that K lines intercept at most $K(2N-1)$ cubes $\mathring{Q}_2^{(N,i)}$. Therefore, the thesis holds as long as

$$K(2N-1) < N^2\,,$$

that implies the thesis with $c_2 = \frac{1}{2}$. Finally, let's consider the case $d = 3$; in this case hyperplanes are planes. Let $\ell_k \doteq \ell \cap \{x : x_k = 0\}$ and

$$n_k \doteq \quad i \in [\![0, N-1]\!]^d : \ell_k \cap \mathring{Q}_3^{N,i} \neq \emptyset$$

for $k \in [3]$. Assume, w.l.o.g., that $a_1 a_2 < 0$ and $a_3 a_2 < 0$ (the lines ℓ_3 and ℓ_1 are thus increasing functions in the first and third coordinate, respectively). By before, we know that $n_k \leq 2N - 1$. Similarly to before, we can write $n_3 = \sum_{k=1}^{N} t_k^3$, where

$$t_k^3 \doteq \quad i \in [\![0, N-1]\!] : \ell_3 \cap \mathring{Q}_3^{N,(k,i,0)} \neq \emptyset \quad ,$$

In particular, the cubes intercepted in the k-th 2-d stripe $\left]\frac{k}{N}, \frac{k+1}{N}\right[\times [0,1] \times \{0\}$ by the line ℓ_3 are given by

$$\mathring{Q}_3^{N,(k,j,0)}$$

for $b_k^3 \leq j \leq c_k^3$, with $t_k^3 = c_k^3 - b_k^3 + 1$. In the same way, we can write $n_1 = \sum_{k=1}^{N} t_k^1$, where

$$t_k^1 \doteq \quad i \in [\![0, N-1]\!] : \ell_1 \cap \mathring{Q}_3^{N,(0,i,k)} \neq \emptyset \quad .$$

In particular, the cubes intercepted in the k-th 2-d stripe $\{0\} \times [0,1] \times \left]\frac{k}{N}, \frac{k+1}{N}\right[$ by the line ℓ_1 are

given by

$$\mathring{Q}_3^{N,(0,j,k)}$$

for $b_k^1 \leq j \leq c_k^1$, with $t_k^1 = c_k^1 - b_k^1 + 1$. It follows that the cubes $\mathring{Q}_3^{N,(i,j,k)}$ intercepted by the plane ℓ are given by

$$\cup_{i=0}^{N-1} \cup_{k=0}^{N-1} \cup_{j=b_k^1+b_i^3}^{c_k^1+c_i^3} \mathring{Q}_3^{N,(i,j,k)} .$$

It follows that

$$\begin{aligned} n &= \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} c_k^1 + c_i^3 - b_k^1 - b_i^3 + 1 \;\; = \sum_{i=0}^{N-1} \sum_{k=0}^{N-1} t_k^1 + t_i^3 - 1 \\ &= N(n_1 + n_3 - N) \leq N(3N-2) . \end{aligned}$$

Therefore, the thesis holds as long as

$$KN(3N-2) < N^3 ,$$

that implies the thesis with $c_3 = \frac{1}{3}$. □

Remark 8. We believe that a similar proof should show the theorem for general $d \geq 1$, with $c_d = d^{-1}$.

Using the above lemma, the following is immediate. Let σ be the ReLU activation and $\mathcal{F}_N^\sigma$ be the space of one-hidden-layer neural networks with at most N units and activation σ.

Corollary A.8. *Let $d \leq 3$. For any $f_N \in \mathcal{F}_N^\sigma$ and ℓ^∞ ball $B \subset \mathbb{R}^d$ of radius $\epsilon > 0$, there exists an ℓ^∞ ball $Q_N \subset \mathbb{R}^d$ of radius $d^{-1}\epsilon(N+1)^{-1}$ such that $f_N|_{Q_N}$ is linear.*

Proof. Clearly, we only need to show this for $B_1 = [0,1]^d$. Given f_N as defined before, f_N is piece-wise linear with linear regions divided by the lines

$$\cup_{i=1}^N \ell_i \doteq \bigcup_{i=1}^N \;\; x \in [0,1]^d : \mathbf{w}_i^T \mathbf{x} = -b_i \quad ,$$

where $\{\mathbf{w}_i, b_i\}_{i=1}^N$ are the hidden layer weights of f_N. By the previous lemma, we deduce that we can find a cube

$$Q = \prod_{j=1}^{d} \frac{i_j}{K}, \frac{i_j + 1}{K} \quad ,$$

with $i_j \in [0, N-1]$ and $K = d^{-1}(N+1)$, such that $f|_Q$ is linear. □

We conclude by combining the above lemma with the proof of Proposition C.5 in [PV18].

Proposition A.9. *Let $f : [0,1]^d \to \mathbb{R}$ C^3 non-linear, for $d \leq 3$, and let $p \in (0, \infty)$. Then it holds that*

$$\inf_{f_N \in \mathcal{F}_N^\sigma} \|f - f_N\|_p \gtrsim N^{-\frac{d}{p}-2} .$$

Proof. Let $f_N \in \mathcal{F}_N^\sigma$. By Corollary A.8, it holds that there exists a ball B_N of radius $r_N \simeq N^{-1}$ such that

$$\|f - f_N\|_{[0,1]^d,p} \gtrsim \inf_{g:\mathbb{R}^d \to R \text{ affine}} \|f - h\|_{B_N,p} .$$

The proof of Proposition C.5 in [PV18] implies that

$$\inf_{g:\mathbb{R}^d \to R \text{ affine}} \|f - h\|_{B_N,p} \gtrsim r_N^{\frac{d}{p}+2},$$

which concludes the proof. □

The above proposition immediately implies, that for the solution u to the PDE 2.7, for C^3 initial conditions, depending on a scalar parameter $\mu \in \mathcal{D} = [0,1]$ ($P = 1$), one has

$$\inf_{f_N \in \mathcal{F}_N^\sigma} \sup_{(t,\boldsymbol{\mu}) \in [0,1] \times \mathcal{D}} \|u(\cdot, t; \mu) - f_N(\cdot, t, \mu)\|_{\mathbb{V}} \geq \inf_{f_N \in \mathcal{F}_N^\sigma} \|u - f_N\|_{[0,1]^3,2} \gtrsim N^{-7/2} .$$

We remark that the lower bound above may potentially be even improved, by showing a lower bound directly for the $L^\infty \otimes \mathbb{V}$ norm, rather then the overall L^2 norm. Clearly, the lower bound extends to the case of u_0 being piece-wise C^3, since we can restrict ourselves to the a domain

where u is C^3. Finally, the lower bound might be further improved, by using the following lemma, which we conjecture to hold, inspired by the work [Bra98].

Lemma A.10. *Consider a uniform grid in the cube $Q = [0,1]^3$ of size 2^{-M},*

$$Q = \cup_{i,j,k=1}^{2^M} Q_{ijk} \doteq \cup_{i,j,k=1}^{2^M} \frac{i-1}{2^M}, \frac{i}{2^M} \times \frac{j-1}{2^M}, \frac{j}{2^M} \times \frac{k-1}{2^M}, \frac{k}{2^M} .$$

Given N hyperplanes in $\mathbb{R}^3$, the maximum number of cubes Q_{ijk} that intersect at least one of them is upper bounded (up to multiplicative constants) by

$$N \cdot 2^M \lg M .$$

A.1.2.2 Lower bounds for other transport PDEs

It is easy to see that similar lower bound can be obtained for other types of transport PDEs. Here, we show two examples of this. In the following, let $\sigma : \mathbb{R} \to \mathbb{R}$ be a semi-algebraic activation.

Example 7 (linear advection equation with constant transport speed). Consider the PDE defined by

$$\begin{cases} u_t + \mu(x \cdot u)_x = 0, & \text{for } (x,t) \in \Omega \times (0,1), \\ u(x,0;\mu) = u_0(x;\mu), & \text{for } x \in \Omega, \\ u(0,t;\mu) = u_0(0;\mu), \end{cases}$$

where $\mu \in \mathcal{D} = [0,1]$ and $u_0(x) = \mathbb{1}\{x \leq 0.5\}$. The solution to this PDE is given by $u(x,t\ \mu) = e^{-\mu t} u_0(e^{-\mu t}x) = e^{-\mu t}\mathbb{1}\ \ x \leq \frac{1}{2}e^{\mu t}$. By a proof equivalent to the one of Proposition 2.4, it follows that

$$\inf_{f_N \in \mathcal{F}_N^\sigma} \sup_{(t,\mu)\in[0,1]\times\mathcal{D}} \|u(\cdot,t;\mu) - f_N(\cdot,t,\mu)\|_{\mathbb{V}} \gtrsim N^{-3/2} . \tag{A.1}$$

Example 8 (wave equation). Consider the wave equation considered in [GU19], that is the PDE

given by

$$\begin{cases} u_{tt} - \mu^2 \cdot u_{xx} = 0, & \text{for } (x,t) \in (-1,1) \times (0,1), \\ u(x,0;\mu) = \mathbb{1}\{x \leq 0\} - \mathbb{1}\{x > 0\}, & \text{for } x \in \Omega, \\ u_t(x,0;\mu) = 0, & \text{for } x \in \Omega, \\ u_t(-1,t;\mu) = 0, & \\ u_t(1,t;\mu) = -1. & \end{cases}$$

where $\mu \in \mathcal{D} = [0,1]$. The solution to this PDE is given by

$$u(x,t;\mu) = \mathbb{1}\{x < -\mu t\} - \mathbb{1}\{x \geq \mu t\}\,.$$

Once again, we can recast the proof of Proposition 2.4 to show the lower bound (A.1) for this solution, where in this case $\mathbb{V} = L^2_{[-1,1]}$.

A.1.3 Proof of Proposition 2.5

The proof is a straight-forward application Theorem 9 from [LS16]. Let $\epsilon \in (0, 0.1]$. By hypothesis, the function T_0 can be uniformly ϵ^2-approximated by a polynomial $T_{0,r}$ of degree $r = O(\log \frac{1}{\epsilon})$. Then, by [LS16, Theorem 9], $T_{0,r}$ can be uniformly ϵ^2-approximated by a shallow network (with ReLU and step function activations) $\hat{T}_0$ of size $O(r^{P+2} \log \frac{r}{\epsilon})$ and depth $O(r + \log \frac{1}{\epsilon})$. It follows that, for any $(t, \boldsymbol{\mu}) \in [0,1] \times \mathcal{D}$, it holds

$$\begin{aligned} \|u_0(\cdot - \hat{T}(t,\boldsymbol{\mu})) - u(\cdot,t;\boldsymbol{\mu})\|^2 &= \int_0^1 \left| u_0(x - \hat{T}(t,\boldsymbol{\mu})) - u_0(x - T_0(t,\boldsymbol{\mu})) \right|^2 dx \\ &\leq \hat{T}(t,\boldsymbol{\mu}) - T_0(t,\boldsymbol{\mu}) \\ &\leq \hat{T}(t,\boldsymbol{\mu}) - T_{0,r}(t,\boldsymbol{\mu}) \; + |T_{0,r}(t,\boldsymbol{\mu}) - T_0(t,\boldsymbol{\mu})| \leq 2\epsilon^2\,. \end{aligned}$$

By construction, $f_N(x, t, \boldsymbol{\mu}) \doteq u_0(x - \hat{T}(t, \boldsymbol{\mu}))$ is a network of size $O(\log^{P+3} \frac{1}{\epsilon})$ and depth $O(\log \frac{1}{\epsilon})$.

Appendix B

Appendix to chapter 3

B.1 Proofs of depth-separation results

B.1.1 Proof of Theorem 3.3

The proof of the lower bound follows the same strategy as [ES16]. For sake of simplicity in the following we remove the dimension d from the following notations: $\mathbf{w}_d = \mathbf{w}$ and $\mathbf{v}_d = \mathbf{v}$. In the following we always assume $d \geq 3$. Let $S \subseteq [d]$ a subset and let $\mathbf{I}_S$ be the truncated identity matrix defined as

$$\mathbf{I}_S := \sum_{s \in S} \mathbf{e}_s \mathbf{e}_s^\top .$$

Moreover, define the function $H_S(\mathbf{x})$ as

$$H_S(\mathbf{x}) \doteq \prod_{i : i \in S} \mathbf{1}_{x_i > 0} \prod_{j : j \in [d] \setminus S} \mathbf{1}_{x_j \leq 0} .$$

Lastly, for a subset $S \subseteq [d]$, let $\mathbf{v}_S := \mathbf{v} + \mathbf{I}_S\mathbf{w}$ and define the function $\sigma_{r,S}(\mathbf{x}) := \sigma_r(\mathbf{v}_S^T\mathbf{x})$. Therefore, the expression of $f_{r_d,\mathbf{w},\mathbf{v}}$ can be rewritten as:

$$f_{r_d,\mathbf{w},\mathbf{v}}(\mathbf{x}) = \sum_{S\subseteq[d]} g_S(\mathbf{x}) = \sum_{S\subseteq[d]} H_S(\mathbf{x})\sigma_{r_d,S}(\mathbf{x})$$

where $g_S(\mathbf{x}) := H_S(\mathbf{x})\sigma_{r_d,S}(\mathbf{x})$. Let the space of N-units one-hidden-layer networks be

$$\mathcal{F}_N = \quad f_N : \mathbf{x} \in \mathbb{R}^r \mapsto \sum_{k=1}^{N} \sigma_k(\mathbf{a}_k^T\mathbf{x}) : \mathbf{a}_k \in \mathbb{R}^d, \, \sigma_k \text{ are 1-Lipschitz activations} \quad .$$

Assume that

(A1) it holds that $\tau_d \cdot r_d \geq \beta d^k$ for some constant $k \geq 1$;

(A2) it holds that $\eta > \log_2 \ \|\psi\|_1 \quad \overline{K/2}$

Then, for large enough d, it holds

$$\inf_{f\in\mathcal{F}_N} \|f_{r_d,\mathbf{w},\mathbf{v}} - f\|_\varphi^2 \geq 1 - N \ \ 2^{1-2\eta}K\|\psi\|_1^2 \ \ ^d O(d \cdot \tau_d \cdot r_d) \, , \tag{B.1}$$

where we denote

$$\|g\|_\varphi^2 \doteq \int_{\mathbb{R}^d} |g(\mathbf{x})|^2\varphi^2(\mathbf{x}) \, d\mathbf{x}$$

for $g \in L^2_{\varphi^2}$. In particular, if $N \simeq \text{poly}(d)$, then the error (B.1) tends to 1 as $d \to \infty$.

To show equation (B.1), we proceed as follows. Let $\mathcal{F} = \{\widehat{f\varphi} : f \in \mathcal{F}_1\}$, and denote by $F := \widehat{\varphi \cdot f_{r_d,\mathbf{w},\mathbf{v}}} = \hat{f}_{r_d,\mathbf{w},\mathbf{v}} * \hat{\varphi}$. Since $\hat{\varphi}$ has compact support in $[-K, K]^d$ and the Fourier transform of a one-unit shallow network $f(\mathbf{x}) = \sigma(\mathbf{x}^T\mathbf{a})$ has support in the line $\{\boldsymbol{\xi} : \boldsymbol{\xi} = \alpha\mathbf{a}, \, \alpha \in \mathbb{R}\}$, it follows that any function in $\mathcal{F}$ is supported in a tube $T = \{\boldsymbol{\xi} : \boldsymbol{\xi} = \alpha\mathbf{a} + [-K, K]^d, \, \alpha \in \mathbb{R}\}$ of radius K. For each tube T of radius K, we consider $\mathcal{T}_T = \{\phi \in L^2 : \ \text{supp}(\phi) \subseteq T\}$ and

$$\kappa \doteq \sup_{T \text{ tube of radius } K} \|P_{\mathcal{T}_T}(F)\|_2 \, ,$$

where $P_{\mathcal{T}_T}(F) = \operatorname{argmin}_{h \in \mathcal{T}_T} \|h - F\|_2^2$. We claim that

$$\inf_{f \in \mathcal{F}_N} \|f_{r_d,\mathbf{w},\mathbf{v}} - f\|_\varphi^2 \geq 1 - N\kappa^2 \,. \tag{B.2}$$

Indeed, given $f \in \mathcal{F}_N$, denote by $T_1, \dots T_N$ the associated N tubes, and by $\mathcal{T}_{T_1,\dots T_N} = \sum_{k \in [N]} \mathcal{T}_{T_k}$ the corresponding subspace spanned by $\mathcal{T}_{T_k}$, $k \in [N]$. Then, by using the isometry of the Fourier transform, we have that

$$\begin{aligned}
\inf_{f \in \mathcal{F}_N} \|f - f_{r_d,\mathbf{w},\mathbf{v}}\|_\varphi^2 &= \inf_{f \in \mathcal{F}_N} \|\widehat{f\varphi} - F\|_2^2 \\
&\geq \inf_{T_1,\dots T_N} \inf_{h \in \mathcal{T}_{T_1,\dots T_N}} \|h - F\|_2^2 \\
&= \inf_{T_1,\dots T_N} \|P_{\mathcal{T}_{T_1,\dots T_N}}(F) - F\|_2^2 \\
&= \inf_{T_1,\dots T_N} (\|F\|_2^2 - \|P_{\mathcal{T}_{T_1,\dots T_N}}(F)\|_2^2) \,.
\end{aligned} \tag{B.3}$$

Now, observe that $\sup_{T_1,\dots,T_N} \|P_{\mathcal{T}_{T_1,\dots T_N}}(F)\|_2^2 \leq N \sup_T \|P_{\mathcal{T}_T}(F)\|_2^2$. Equation (B.3) therefore becomes

$$\inf_{f \in \mathcal{F}_N} \|f - f_{r_d,\mathbf{w},\mathbf{v}}\|_\varphi^2 \;\geq\; \|F\|_2^2 - N \sup_T \|P_{\mathcal{T}_T}(F)\|_2^2 \,,$$

which proves (B.2) by plugging in the definition of κ and recalling that $\|F\|_2^2 = \|f_{r_d,\mathbf{w},\mathbf{v}}\|_\varphi^2 = 1$ by Parseval. To establish (B.1), it is therefore sufficient to prove that

$$\kappa^2 \leq \; \|\psi\|_1^2 2^{1-2\eta} K^{\,d} O(d \cdot \tau_d \cdot r_d) \,. \tag{B.4}$$

The rest of the proof will be devoted to establishing a sufficiently sharp upper bound for $\|P_{\mathcal{T}_T}(F)\|_2$. Observe that $P_{\mathcal{T}_T}(F)$ is simply obtained by setting to zero all frequencies of F outside T. We start by computing an upper bound on $|F(\boldsymbol{\xi})|$. We claim the following.

Lemma B.1. *It holds that*

$$|F(\boldsymbol{\xi})| \;\leq\; \frac{\|\varphi\|_1}{2^d} \sum_{S\subseteq[d]} \prod_{j=1}^{d} \min\;\; 1, \frac{2K}{\pi(|\xi_j - \xi_{S,j}| - K)_+} \quad .$$

Let $D(\boldsymbol{\xi}) \doteq \quad {}_S\, D_S(\boldsymbol{\xi})$, with $D_S(\boldsymbol{\xi}) \doteq \quad {}_{j=1}^{d} \min\;\; 1, \frac{2K}{\pi(|\xi_j - \xi_{S,j}| - K)_+}$, so that from Lemma B.1 we have

$$|F(\boldsymbol{\xi})| \leq 2^{-d} \|\varphi\|_1 D(\boldsymbol{\xi}) \,. \tag{B.6}$$

Recall that $\tau_d = \sup_{S\in[d]} \|\mathbf{v}_S\|_\infty$. Given $\boldsymbol{\xi}$ non-zero, we claim the following.

Lemma B.2. *It holds that*

$$D(\boldsymbol{\xi}) \leq C_{K,\gamma} 2^{d(1-\eta)} \min\;\; 1, 2K(\pi(\|\boldsymbol{\xi}\|_\infty - r_d\tau_d - K)_+)^{-1} \quad , \tag{B.7}$$

where $C_{K,\gamma} = 2\exp \quad \overline{\frac{8K}{\pi\gamma}}$.

Now, pick any arbitrary non-zero direction $\boldsymbol{\nu}$ such that $\|\boldsymbol{\nu}\|_\infty = 1$. Let

$$T = \{\boldsymbol{\xi} : \inf_{\alpha\in\mathbb{R}} \|\boldsymbol{\xi} - \alpha\boldsymbol{\nu}\|_\infty \leq K\} \tag{B.8}$$

denote the tube of radius K in the direction $\boldsymbol{\nu}$. It holds that

$$\int_T D(\boldsymbol{\xi})^2 d\boldsymbol{\xi} \;=\; \underbrace{\int_{T\cap\{\|\boldsymbol{\xi}\|_\infty \leq 2\tau_d r_d\}} D(\boldsymbol{\xi})^2 d\boldsymbol{\xi}}_{t_1} + \underbrace{\int_{T\cap\{\|\boldsymbol{\xi}\|_\infty > 2\tau_d r_d\}} D(\boldsymbol{\xi})^2 d\boldsymbol{\xi}}_{t_2} \,.$$

In order to control the two terms t_1 and t_2, we use the following lemma to upper bound the measure of a ℓ_∞-cylinder.

Lemma B.3. *Let T be an ℓ_∞-tube of radius K as defined in (B.8). If μ denotes the d-dimensional*

Lebesgue measure, then

$$\mu\ \ T \cap [-R,R]^d\ \ \leq 8e^2(d-1)(K+R)(2K)^{d-1}\ . \tag{B.10}$$

Moreover, if $g : \mathbb{R} \to \mathbb{R}$ *is in* $L^1(\mathbb{R})$ *and non-increasing, then*

$$\int_{T\cap\{\|\boldsymbol{\xi}\|_\infty > R\}} g(\|\boldsymbol{\xi}\|_\infty)\, d\boldsymbol{\xi} \leq 4e^2(d-1)(2K)^{d-1} \int_{R-K(2+3/(d-1))}^{\infty} g(u)du\ , \tag{B.11}$$

as long as $R > K(2 + 3/(d-1))$.

From (B.7) and (B.10), the first term of (B.9) can be bounded as

$$\begin{aligned} t_1 &\leq 8e^2 C_{K,\gamma}^2 2^{2d(1-\eta)+(d-1)} K^{d-1}(d-1)(K+2\tau_d r_d) \\ &\leq D_{K,\gamma}^{(1)} \cdot d \cdot (\tau_d r_d)\ \ 2^{2(1-\eta)+1} K^{\ \ d} \end{aligned} \tag{B.12}$$

for $D_{K,\gamma}^{(1)} = 16e^2K^{-1}C_{K,\gamma}^2$ and d large enough, such that $2\tau_d r_d \geq K$. Similarly, using (B.11), the second term t_2 in turn can be bounded as

$$\begin{aligned} t_2 &\leq 8e^2\pi^{-2}C_{K,\gamma}^2 d\ \ 2^{2(1-\eta)+1}K^{\ \ d} \int_{2\tau_d r_d - K(2+3/(d-1))} (u - \tau_d r_d - K)^{-2}\, du \\ &= 8e^2\pi^{-2}KC_{K,\gamma}^2 d\ \ 2^{2(1-\eta)+1}K^{\ \ d}(\tau_d r_d - 3K(1+1/(d-1)))^{-1} \\ &\leq D_{K,\gamma}^{(2)} \cdot d \cdot\ \ 2^{2(1-\eta)+1}K^{\ \ d}\ , \end{aligned} \tag{B.13}$$

for $D_{K,\gamma}^{(2)} = 16e^2\pi^{-2}C_{K,\gamma}^2$ and and d large enough, such that $\tau_d r_d \geq 10K$. Thus, collecting (B.12)

and (B.13) and using (B.6), we obtain

$$\begin{aligned}\int_T |F(\boldsymbol{\xi})|^2 d\boldsymbol{\xi} &\leq \|\varphi\|_1^2 \cdot 2^{-2d}(t_1 + t_2) \\ &\leq d \cdot \|\varphi\|_1^2 \ 2^{1-2\eta} K^{\ d} \ D_{K,\gamma}^{(1)} \tau_d r_d + D_{K,\gamma}^{(2)} \\ &\leq D_{K,\gamma} \cdot d \cdot \|\varphi\|_1^2 \ 2^{1-2\eta} K^{\ d} \max(1, \tau_d r_d) \ ,\end{aligned}$$

where

$$D_{K,\gamma} \doteq D_{K,\gamma}^{(1)} + D_{K,\gamma}^{(2)} = 32 \exp \ \ 2 + \sqrt{\frac{8K}{\pi\gamma}} \quad \pi^{-2} + K^{-1} \quad .$$

It follows that

$$\|P_{\mathcal{T}_T}(F)\|_2^2 = \int_T |F(\boldsymbol{\xi})|^2 d\boldsymbol{\xi} \leq D_{K,\gamma} \cdot (d \cdot \tau_d \cdot r_d) \cdot \ \|\psi\|_1^2 \, 2^{1-2\eta} K^{\ d} \ ,$$

as long as $d \geq [\beta^{-1} \max(1, 10K)]^{1/k}$ (where β and k satisfy $\tau_d r_d \geq \beta d^k$). We have just established (B.4), and this concludes the proof of the theorem. In the remaining part of this section we prove the auxiliary lemmas used above.

Proof of Lemma B.1. We start by computing $\hat{f}_{r_d,\mathbf{w},\mathbf{v}}$. From the definition of σ_r, it follows that

$$\hat{\sigma}_{r,S}(\boldsymbol{\xi}) = \delta(\boldsymbol{\xi} - r\mathbf{v}_S) \ ,$$

which combined with the definition of H yields

$$\hat{f}_{r_d,\mathbf{w},\mathbf{v}}(\boldsymbol{\xi}) = \sum_{S \subseteq [d]} \ \hat{H}_S * \hat{\sigma}_{r_d,S} \ \ (\boldsymbol{\xi}) = \sum_{S \subseteq [d]} \hat{H}_S(\boldsymbol{\xi} - r_d \mathbf{v}_S) \ .$$

Let $\boldsymbol{\xi}_S \doteq r_d \mathbf{v}_S$. It holds that

$$F(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} \hat{f}_{r_d,\mathbf{w},\mathbf{v}}(\boldsymbol{\nu})\hat{\varphi}(\boldsymbol{\xi}-\boldsymbol{\nu})\,d\boldsymbol{\nu} = \sum_{S\subseteq[d]} \int_{\mathbb{R}^d} \hat{H}_S(\boldsymbol{\nu}-\boldsymbol{\xi}_S)\hat{\varphi}(\boldsymbol{\xi}-\boldsymbol{\nu})\,d\boldsymbol{\nu}$$
$$= \sum_{S\subseteq[d]} \underbrace{\int_{\mathbb{R}^d} \hat{H}_S(\boldsymbol{\nu})\hat{\varphi}(\boldsymbol{\xi}-\boldsymbol{\xi}_S-\boldsymbol{\nu})\,d\boldsymbol{\nu}}_{\doteq F_S(\boldsymbol{\xi}-\boldsymbol{\xi}_S)} . \tag{B.14}$$

We can now bound each term F_S separately. It holds that

$$F_S(\boldsymbol{\xi}) = \int \hat{H}_S(\boldsymbol{\nu})\hat{\varphi}(\boldsymbol{\xi}-\boldsymbol{\nu})d\boldsymbol{\nu} = \int H_S(\mathbf{x})e^{2i\pi\boldsymbol{\xi}^T\mathbf{x}}\varphi(\mathbf{x})d\mathbf{x} = \prod_{j=1}^{d} F_j(\xi_j) \tag{B.15}$$

where

$$F_j(t) = \int_{\mathbb{R}} \mathbb{1}\{\epsilon_j x > 0\} e^{2i\pi t x}\psi(x)\,dx\,, \tag{B.16}$$

with $\epsilon_j = \pm 1$. Assume without loss of generality that $\epsilon_j = 1$. Observe that $F_j = \check{Q}$, where

$$Q(u) = \mathbb{1}\{u > 0\}\psi(u)\,.$$

Since $\psi \in L^1(\mathbb{R})$ and its Fourier transform $\hat{\psi}$ has compact support in $[-K, K]$, it holds that

$$|\hat{\psi}(\tau)| \le \|\psi\|_1 \text{ for } \tau \in [-K,K] \quad \text{and} \quad \hat{\psi}(\tau) = 0 \text{ for } |\tau| > K\,. \tag{B.17}$$

On the one hand, since ψ is even, it holds, by directly bounding (B.16), that

$$|F_j(t)| \le \frac{1}{2}\int_{\mathbb{R}} |\psi(u)|du = \frac{1}{2}\|\psi\|_1 \quad \text{for all } t\,,$$

and from (B.17) and the Hilbert transform of Q we deduce on the other hand that

$$|F_j(t)| = \frac{1}{2\pi}\int_{-K}^{K} \frac{\hat{\psi}(\tau)}{t-\tau}d\tau \le \frac{2K\|\psi\|_1}{(2\pi)(|t|-K)} \quad \text{for } |t| > K\,,$$

so that it follows that

$$|F_j(t)| \leq \frac{\|\psi\|_1}{2} \min \quad 1, \frac{2K}{\pi(|t| - K)_+} \quad . \tag{B.18}$$

Thus, from equations (B.14), (B.15) and (B.18) it follows that

$$\begin{aligned} |F(\boldsymbol{\xi})| &\leq \sum_{S \subseteq [d]} |F_S(\boldsymbol{\xi} - \boldsymbol{\xi}_S)| \\ &\leq \frac{\|\varphi\|_1}{2^d} \sum_{S \subseteq [d]} \prod_{j=1}^{d} \min \quad 1, \frac{2K}{\pi(|\xi_j - \xi_{S,j}| - K)_+} \quad , \end{aligned}$$

which proves Lemma B.1. □

Proof of Lemma B.2. Let define for any $\boldsymbol{\xi} \in \mathbb{R}^d$ and $\lambda > 0$

$$\mathsf{n}(\boldsymbol{\xi}, \lambda) \doteq |\{j \in [d] : |\xi_j| > \lambda\}| \, .$$

Recall that $\mathbf{v}_S = \mathbf{v} + \mathbf{I}_S \mathbf{w}$ and $\boldsymbol{\xi}_S = r_d \mathbf{v}_S$. Observe that $\boldsymbol{\xi}_S - \boldsymbol{\xi}_{S'} = r_d(\mathbf{I}_S - \mathbf{I}_{S'})\mathbf{w}$, so

$$|\xi_{S,j} - \xi_{S',j}| = \begin{cases} r_d |w_j| & \text{if } j \in (S \cup S\,) \setminus (S \cap S\,) \\ 0 & \text{otherwise} \end{cases} \quad . \tag{B.19}$$

If $\mathsf{d}(S, S\,)$ denotes the Hamming distance between two subsets S, S , then for all S, S , the following holds.

Lemma B.4. *It holds that*

$$\mathsf{n}(\boldsymbol{\xi}_S - \boldsymbol{\xi}_{S'}, \gamma d^2) = \mathsf{d}(S \cap \Omega_d, S \cap \Omega_d) \, . \tag{B.20}$$

This immediately implies that

$$\mathsf{n} \quad \boldsymbol{\xi} - \boldsymbol{\xi}_S, \frac{\gamma d^2}{2} \quad + \mathsf{n} \quad \boldsymbol{\xi} - \boldsymbol{\xi}_{S'}, \frac{\gamma d^2}{2} \quad \geq \mathsf{d}(S \cap \Omega, S \cap \Omega) \quad \text{for all } \boldsymbol{\xi} \text{ and } S \neq S \quad . \tag{B.21}$$

Indeed, if that was not the case, applying the triangle inequality coordinate-wise would contradict equation (B.20). The first upper bound is obtained by first noticing that, for $d > 2 \quad \overline{K/\gamma}$, it holds

$$D_S(\boldsymbol{\xi}) \leq \quad \pi(\gamma d^2/2 - K)/(2K) \quad ^{-\mathsf{n}(\boldsymbol{\xi} - \boldsymbol{\xi}_S, \gamma d^2/2)} \qquad \text{for all } S \text{ and } \xi \,.$$

Now, defining $S^*_{\boldsymbol{\xi}} = \arg\min_{S \subseteq [d]} \mathsf{n}(\boldsymbol{\xi} - \boldsymbol{\xi}_S, \gamma d^2/2)$, from (B.21) it follows that

$$\mathsf{n}(\boldsymbol{\xi} - \boldsymbol{\xi}_S, \gamma d^2/2) \geq \frac{\mathsf{d}(S \cap \Omega_d, S \ \cap \Omega_d)}{2} \qquad \text{for all } S \neq S^*_{\boldsymbol{\xi}}$$

and thus, for $d > 2 \quad \overline{K/\gamma}$, it holds

$$\begin{aligned}
D(\boldsymbol{\xi}) &= D_{S^*_{\boldsymbol{\xi}}}(\boldsymbol{\xi}) + \sum_{S \neq S^*_{\boldsymbol{\xi}}} D_S(\boldsymbol{\xi}) \\
&\leq D_{S^*_{\boldsymbol{\xi}}}(\boldsymbol{\xi}) + \sum_{s=1}^{|\Omega_d|} \sum_{S \,:\, \mathsf{d}(S \cap \Omega_d, S^*_{\boldsymbol{\xi}} \cap \Omega_d) = s} \pi(\gamma d^2/2 - K)/(2K) \quad ^{-s/2} \\
&\leq D_{S^*_{\boldsymbol{\xi}}}(\boldsymbol{\xi}) + 2^{d - |\Omega_d|} \sum_{s=1}^{|\Omega_d|} \quad \begin{matrix} |\Omega_d| \\ s \end{matrix} \quad \pi(\gamma d^2/2 - K)/(2K) \quad ^{-s/2} \\
&\leq 1 + 2^{d - |\Omega_d|} \quad 1 + \frac{1}{\overline{\pi(\gamma d^2/2 - K)/(2K)}} \quad ^{|\Omega_d|} \\
&\leq C_{K,\gamma} 2^{d(1-\eta)}
\end{aligned} \tag{B.22}$$

since $|\{S : \mathsf{d}(S \cap \Omega_d, S^*_{\boldsymbol{\xi}} \cap \Omega_d) = s\}| \leq 2^{d - |\Omega_d|} \quad \begin{matrix} |\Omega_d| \\ s \end{matrix}$. The term $C_{K,\gamma}$ is a constant that depends only on K and γ; in particular, we can choose $C_{K,\gamma} = 2 \exp \quad \overline{\frac{8K}{\pi\gamma}}$. The second upper bound is obtained using the above argument as follows. Let $q_{\boldsymbol{\xi}} = \arg\max_j |\xi_j|$. Since $\|\boldsymbol{\xi}_S\|_\infty \leq r_d \tau_d$ for

any $S \subseteq [d]$, it holds that

$$
\begin{aligned}
D(\boldsymbol{\xi}) &\leq \sum_{S \subseteq [d]} \frac{2K}{\pi(|\xi_{q_\xi} - \xi_{S,q_\xi}| - K)_+} \cdot \prod_{j \neq q_\xi} \min \quad 1, \frac{2K}{\pi(|\xi_j - \xi_{S,j}| - K)_+} \\
&\leq 2K(\pi(\|\boldsymbol{\xi}\|_\infty - \tau_d r_d - K)_+)^{-1} \sum_{S \subseteq [d]} \prod_{j \neq q_\xi} \min \quad 1, \frac{2K}{\pi(|\xi_j - \xi_{S,j}| - K)_+} \\
&\leq C_{K,\gamma} 2K(\pi(\|\boldsymbol{\xi}\|_\infty - \tau_d r_d - K)_+)^{-1} \cdot 2^{d(1-\eta)} \qquad \text{(B.23)}
\end{aligned}
$$

by noticing that the argument leading to (B.22) can now be repeated for the $(d-1)$-dimensional vector $\check{\boldsymbol{\xi}} = (\xi_1, \ldots, \xi_{q_{\boldsymbol{\xi}}-1}, \xi_{q_{\boldsymbol{\xi}}+1}, \ldots \xi_d)$, so that

$$
\mathsf{n}(\check{\boldsymbol{\xi}} - \check{\boldsymbol{\xi}}_S, \gamma d^2/2) \geq \frac{\mathsf{d}((S \cap \Omega_d) \setminus \{q_{\boldsymbol{\xi}}\}, (S \cap \Omega_d) \setminus \{q_{\boldsymbol{\xi}}\})}{2} \quad \text{for all } S \neq S^*_{\boldsymbol{\xi}}
$$

which proves (B.23) and concludes the proof of Lemma B.2. □

Proof of Lemma B.4. In fact, it holds that the two sets $A_1 := \{j \in [d] : |\xi_{S,j} - \xi_{S',j}| \geq \gamma d^2\}$ and $A_2 := \{j \in [d] : j \in (S \cap \Omega_d) \backslash (S \cap \Omega_d)\}$ are equal. Let $j \in A_1$. Then $|\xi_{S,j} - \xi_{S',j}| > \gamma d^2$. Since this quantity is nonzero, equation (B.19) indicates that therefore $j \in S \backslash S$ without loss of generality. Moreover, $|\xi_{S,j} - \xi_{S',j}| = r_d |w_j|$ which implies that $r_d |w_j| > \gamma d^2$ and $j \in \Omega_d$. We conclude that $j \in (S \cap \Omega_d) \backslash (S \cap \Omega_d)$ which implies that $j \in A_2$. Now, let $j \in A_2$. Then, without loss of generality, $j \in (S \cap \Omega_d) \backslash (S \cap \Omega_d)$. Then, it holds $r|w_j| > \gamma d^2$ since $j \in S \backslash S$ according to (B.19) and $|\xi_{S,j} - \xi_{S',j}| = r_d |w_j|$. Combining these two facts, it follows that $|\xi_{S,j} - \xi_{S',j}| > \gamma d^2$ which means that $j \in A_2$. □

Proof of Lemma B.3. Let

$$
\begin{aligned}
T_R(\boldsymbol{\nu}) &= T(\boldsymbol{\nu}) \cap [-R, R]^d \\
&= \{\boldsymbol{\xi} : \inf_{\alpha \in \mathbb{R}} \sup_{j \in [d]} |\xi_j - \alpha \nu_j| \leq K \text{ and } \|\boldsymbol{\xi}\|_\infty \leq R\} .
\end{aligned}
$$

The aim is to upper bound the volume of $T_R(\boldsymbol{\nu})$ for any $\boldsymbol{\nu}$. Assume, without loss of generality, that $\|\boldsymbol{\nu}\|_\infty = 1$. The cut-off tube $T_R(\boldsymbol{\nu})$ can be covered with ℓ_∞-balls of radius $K = \vartheta K$ centered along the ray defined by $\boldsymbol{\nu}$, that is

$$T_R(\boldsymbol{\nu}) \subseteq \bigcup_{j=-\lfloor (K+R)/s \rfloor}^{\lfloor (K+R)/s \rfloor} js\boldsymbol{\nu} + [-\vartheta K, \vartheta K]^d \quad . \tag{B.24}$$

Now, we optimize both the sampling rate $s \in (0, K)$ and the radius ratio $\vartheta \geq 1$ while satisfying (B.24). Given s, let us first compute the smallest admissible ϑ. Any $\mathbf{x} \in T_R(\boldsymbol{\nu})$ satisfies

$$\|\mathbf{x} - (j+y)s\boldsymbol{\nu}\|_\infty \leq K$$

for some $j \in \mathbb{N}$ and $|y| < 1$. This implies that $\|\mathbf{x} - js\boldsymbol{\nu}\|_\infty \leq K + ys \leq K + s$. Therefore an admissible ϑ is given by the solution of $K + s = \vartheta K$, that is $\vartheta = 1 + sK^{-1}$. Now, the volume of

$$S_R \doteq \bigcup_{j=-\lfloor (K+R)/s \rfloor}^{\lfloor (K+R)/s \rfloor} js\boldsymbol{\nu} + \left[- \left(1 + \frac{s}{K}\right) K, \left(1 + \frac{s}{K}\right) K \right]^d$$

is upper bounded by

$$l(s) \doteq 4 \frac{K+R}{s} \left(2(K+s)\right)^d \ .$$

Minimizing over s gives $s = \frac{K}{d-1}$. Therefore, for all $\boldsymbol{\nu} \in \mathbb{R}^d$, it holds

$$T_R(\boldsymbol{\nu}) \leq (K+R)K^{d-1}(d-1) \left(1 + \frac{1}{d-1}\right)^d \leq (K+R)(d-1)K^{d-1}e^2 \ ,$$

which proves (B.10). Equation (B.11) is established analogously. Let $T_{>R}(\boldsymbol{\nu}) = T(\nu) \cap \{\boldsymbol{\xi} : \|\boldsymbol{\xi}\|_\infty > R\}$. Then we have that

$$T_{>R}(\boldsymbol{\nu}) \subseteq \bigcup_{j \geq \lfloor \frac{R-K}{s} \rfloor} js\boldsymbol{\nu} + [-(K+s), (K+s)]^d \quad ,$$

where we set $s = K/(d-1)$. Since g is non-increasing, it follows that

$$\begin{aligned}
\int_{T_{>R}(\boldsymbol{\nu})} g(\|\boldsymbol{\xi}\|_\infty)\, d\boldsymbol{\xi} &\leq \sum_{|j|\geq \lfloor \frac{R-K}{s} \rfloor} \int_{\|\boldsymbol{\xi}-js\boldsymbol{\nu}\|_\infty \leq K+s} g(\|\boldsymbol{\xi}\|_\infty)\, d\boldsymbol{\xi} \\
&\leq 2(2(K+s))^d \sum_{j\geq \lfloor \frac{R-K}{s} \rfloor} g(js-(K+s)) \\
&\leq 2(2(K+s))^d \sum_{j\geq \lfloor \frac{R-K}{s} \rfloor} \frac{1}{s} \int_{(j-1)s-(K+s)}^{js-(K+s)} g(u)\, du \\
&\leq \frac{2(2(K+s))^d}{s} \int_{R-K-2s-(K+s)}^{\infty} g(u)\, du \\
&\leq \frac{2e^2(d-1)(2K)^d}{K} \int_{R-K(2+3/(d-1))}^{\infty} g(u)\, du\,.
\end{aligned}$$

This establishes (B.11) and concludes the proof. □

B.1.2 Proof of Theorem 3.5

The proof consists in approximating the activation σ_r using Assumption 1.2 on σ. Since σ_r is $(2\pi r)$-Lipschitz, we obtain that there exists, for any $r, Q > 0$, $\alpha_k, \beta_k \in \mathbb{R}$ such that over the interval $[-Q, Q]$ it holds

$$\sup_{|t|\leq Q} \left| \sigma_r(t) - \sum_{k=1}^{N} \alpha_k \sigma(t-\beta_k) \right| \leq \frac{2Qr}{N}$$

as well as

$$\left| \sum_{k=1}^{N} \alpha_k \sigma(t-\beta_k) \right| \leq 1 + 2Qr/N \quad \text{for } t \in \mathbb{R}\,.$$

Let $f_N \in \mathcal{F}_N^\sigma$ be defined as

$$f_N(\mathbf{x}) = \sum_{k=1}^{N} \alpha_k \sigma\left(r_d \left| \mathbf{v}_d^T \mathbf{x} + \mathbf{w}_d^T \mathbf{x}_+ \right| - \beta_k \right)$$

Now, let $\gamma_d = \|\mathbf{v}_d\|_1 + \|\mathbf{w_d}\|_1$ and $\tilde{Q}_d = \frac{Q_d}{\gamma_d}$, so that by definition when $\|\mathbf{x}\|_\infty \leq \tilde{Q}_d$ it holds that

$$|\mathbf{v}_d^T \mathbf{x} + \mathbf{w}_d^T \mathbf{x}_+| \leq Q_d \, .$$

The approximation error can be decomposed as follows:

$$\begin{aligned}
&\int_{\mathbb{R}^d} (f_{r_d,\mathbf{w}_d,\mathbf{v}_d}(\mathbf{x}) - f_N(\mathbf{x}))^2 \varphi(\mathbf{x})^2 \, d\mathbf{x} = \\
&= \int_{\|\mathbf{x}\|_\infty \leq \tilde{Q}_d} (f_{r_d,\mathbf{w}_d,\mathbf{v}_d}(\mathbf{x}) - f_N(\mathbf{x}))^2 \varphi(\mathbf{x})^2 \, d\mathbf{x} + \int_{\|\mathbf{x}\|_\infty > \tilde{Q}_d} (f_{r_d,\mathbf{w}_d,\mathbf{v}_d}(\mathbf{x}) - f_N(\mathbf{x}))^2 \varphi(\mathbf{x})^2 \, d\mathbf{x} \\
&\leq \frac{4Q_d^2 r_d^2}{N^2} \|\varphi \cdot \mathbb{1}_{B^d_{\tilde{Q}_d,\infty}}\|_2^2 + 4 \left(1 + \frac{Q_d r_d}{N}\right)^2 (\|\varphi\|_2^2 - \|\varphi \cdot \mathbb{1}_{B^d_{\tilde{Q}_d,\infty}}\|_2^2) \\
&\leq \frac{4\tilde{Q}_d^2 \gamma_d^2 r_d^2}{N^2} \|\varphi\|_2^2 + 4 \left(1 + \frac{Q_d r_d}{N}\right)^2 \left(1 - (1 - \alpha \, \tilde{Q}_d^{-1})^d\right) \\
&\leq \|\varphi\|_2^2 \left(\frac{4\tilde{Q}_d^2 \gamma_d^2 r_d^2}{N^2} + 16\alpha d \tilde{Q}_d^{-1}\right) ,
\end{aligned}$$

since $|\psi(x)|^2 \leq \alpha |x|^{-2}/2$ for some $\alpha > 0$, as long as $\tilde{Q}_d > \alpha$ and $N > r_d \tilde{Q}_d$. Optimizing this upper bound with respect to $\tilde{Q}_d$ gives

$$\tilde{Q}_d = \left(2\alpha d \frac{N^2}{r_d^2 \gamma_d^2}\right)^{1/3} ,$$

which results in

$$\|f_{r,\mathbf{w},\mathbf{v}} - f\|_\varphi^2 \lesssim \left(\frac{d\gamma_d r_d}{N}\right)^{2/3} ,$$

as long as $N > \alpha r_d \gamma_d$. This concludes the proof.

B.2 Proofs of $\mathrm{poly}(d)$ upper bounds

B.2.1 Proof of Lemma 3.1

We show this for the case $L(f^{(d)}) = 2$, but the proof it is analogous for the other cases. The function $f^{(d)}$ has the form

$$f^{(d)}(\mathbf{x}) = \boldsymbol{\gamma}_d^T \boldsymbol{\rho}_2(\mathbf{W}_d \boldsymbol{\rho}_1(\mathbf{U}_d \mathbf{x}))$$

where $\boldsymbol{\rho}_1^{(d)}, \boldsymbol{\rho}_2^{(d)}$ are component-wise activations satisfying Assumption 1, and $\boldsymbol{\gamma}_d \in \mathbb{R}^{q_d}$, $\mathbf{W} \in \mathbb{R}^{q_d \times p_d}$, $\mathbf{U} \in \mathbb{R}^{p_d \times d}$, with

$$p_d, q_d, \|\boldsymbol{\gamma}\|_\infty, \|\mathbf{W}\|_{F,\infty}, \|\mathbf{U}\|_{F,\infty} \leq \mathrm{poly}(d) \ .$$

Thanks to Assumption 1.2, there exists $\mathbf{A} \in \mathbb{R}^{Np_d \times d}$, $\mathbf{B} \in \mathbb{R}^{p_d \times Np_d}$, $\mathbf{c} \in \mathbb{R}^{Np_d}$ such that

$$\sup_{\mathbf{x} \in K} \boldsymbol{\gamma}^T \boldsymbol{\rho}_2(\mathbf{W} \boldsymbol{\rho}_1(\mathbf{U}\mathbf{x})) - \boldsymbol{\gamma}^T \boldsymbol{\rho}_2(\mathbf{W}\mathbf{B}\, \boldsymbol{\sigma}(\mathbf{A}\mathbf{x} + \mathbf{c})) \ \leq \frac{\epsilon}{2}$$

and

$$N, \|\mathbf{c}\|_\infty, \|\mathbf{B}\|_{F,\infty}, \|\mathbf{A}\|_{F,\infty} \leq \epsilon^{-1} \cdot \mathrm{poly}(d) \ .$$

Let $K_1 = \{\mathbf{B}\boldsymbol{\sigma}(\mathbf{A}\mathbf{x} + \mathbf{c}) : \mathbf{x} \in K\}$; it holds $\mathrm{diam}(K_1) \leq \mathrm{poly}(d)$. Similarly as before, we get that there exists $\mathbf{D} \in \mathbb{R}^{Mq_d \times p_d}$, $\mathbf{E} \in \mathbb{R}^{q_d \times Mq_d}$, $\mathbf{f} \in \mathbb{R}^{Mq_d}$ such that

$$\sup_{\mathbf{y} \in K_1} \boldsymbol{\gamma}^T \boldsymbol{\rho}_2(\mathbf{W}\mathbf{y}) - \boldsymbol{\gamma}^T \mathbf{E}\, \boldsymbol{\sigma}(\mathbf{D}\mathbf{y} + \mathbf{f}) \ \leq \frac{\epsilon}{2}$$

and

$$M, \|\mathbf{f}\|_\infty, \|\mathbf{E}\|_{F,\infty}, \|\mathbf{D}\|_{F,\infty} \leq \epsilon^{-1} \cdot \mathrm{poly}(d) \ .$$

By calling $\tilde{\boldsymbol{\gamma}} = \mathbf{E}^T\boldsymbol{\gamma}$, $\tilde{\mathbf{W}} = \mathbf{DWB}$ and $\tilde{\mathbf{U}} = \mathbf{UA}$, we get that

$$g^{\sigma}(\mathbf{x}) \doteq \tilde{\boldsymbol{\gamma}}^T\boldsymbol{\sigma}(\tilde{\mathbf{W}}\boldsymbol{\sigma}(\tilde{\mathbf{U}}\mathbf{x} + \mathbf{c}) + \mathbf{f})$$

satisfies the statement of the theorem.

B.2.2 Preliminary lemmas

The first lemma is a known results in approximation theory.

Lemma B.5. *(Jackson's Theorem, Theorem 1.4 in [Riv81]) Let $f : [a, b] \to \mathbb{R}$ with modulus of continuity ω. Then there exists a polynomial $p_n(t) = \sum_{k=0}^{n} p_k t^k$, $p_k \in \mathbb{R}$, such that*

$$\sup_{t\in[-r,r]} |f(t) - p_n(t)| \leq 6\,\omega\left(\frac{b-a}{2n}\right) .$$

The next lemma yields a worst approximation rate but allows us to control the coefficients of the polynomial. It is a small modification of Lemma 4 in [SES19].

Lemma B.6. *Let $f : [-r, r] \to \mathbb{R}$ $(1, \alpha)$-Holder. Then for any $\epsilon > 0$ there exists a polynomial $p_n(t) = \sum_{k=0}^{n} r_k t^k$, $r_k \in \mathbb{R}$, of degree $n = \left\lceil \frac{4^{\frac{1}{\alpha}} r^{\alpha}}{1+\frac{2}{\alpha}} \right\rceil$ such that*

$$\sup_{t\in[-r,r]} |f(t) - p_n(t)| \leq \epsilon .$$

Moreover, p_n can be chosen such that $|r_k| \leq 2^n r^{\alpha-k}$, $k \in [n]$, and $|r_0| \leq r^{\alpha} + |f(0)|$.

Proof. Notice that we can assume $f(0) = 0$ without loss of generality. Define $g(t) = f(r(2t-1))$ for $t \in [0, 1]$ and notice that g is $((2r)^{\alpha}, \alpha)$-Holder. Also, define the n Bernstein polynomial $b_{n,i}$, $i \in [0, n]$, as

$$b_{n,i}(t) = \binom{n}{i} t^i (1-t)^{n-i}$$

for $t \in [0,1]$. Notice that they form a partition of unity. We define

$$g_n(t) = \sum_{i=0}^{n} g\left(\frac{i}{n}\right) b_{n,i}(t) \, .$$

We have that

$$\begin{aligned} |g_n(t) - g(t)| &\leq \sum_{i=0}^{n} b_{n,i}(t) \; \left| g(t) - g\left(\frac{i}{n}\right) \right| \\ &= \sum_{i \, : \, \left|\frac{i}{n} - t\right| < \epsilon} b_{n,i}(t) \; \left| g(t) - g\left(\frac{i}{n}\right) \right| + \sum_{i \, : \, \left|\frac{i}{n} - t\right| \geq} b_{n,i}(t) \; \left| g(t) - g\left(\frac{i}{n}\right) \right| \\ &\leq \epsilon^{\alpha} + 2r^{\alpha} \sum_{i \, : \, \left|\frac{i}{n} - t\right| \geq} b_{n,i}(t) \leq \epsilon^{\alpha} + \frac{r^{\alpha}}{2n\epsilon^2} \, . \end{aligned}$$

In particular $\frac{r^{\alpha}}{2n\epsilon^2} \leq \epsilon^{\alpha}$ if

$$n \geq \frac{r^{\alpha}}{2\epsilon^{2+\alpha}} \, .$$

If we define $p_n(t) = g_n\left(\frac{t}{2r} + \frac{1}{2}\right)$, then we have that

$$\sup_{x \in [-r,r]} |f(t) - p_n(t)| \leq \epsilon$$

if

$$n \geq \frac{4^{\frac{1}{\alpha}} r^{\alpha}}{\epsilon^{1+\frac{2}{\alpha}}} \, .$$

Finally, we want to upper bound the coefficients of p_n. Notice that we have

$$p_n(t) = (2r)^{-n} \sum_{i=0}^{n} \binom{n}{i} g\left(\frac{i}{n}\right) (t+r)^i (t-r)^{n-i} \, .$$

It follows that the coefficients of p_n can be bounded by those of

$$(2r)^{-n} \sum_{i=0}^{n} \binom{n}{i} g\left(\frac{i}{n}\right) (t+r)^n \leq r^{\alpha-n}(t+r)^n \,.$$

Let r_k the k-th coefficients of $r^{\alpha-n}(t+r)^n$. Then

$$r_k = r^{\alpha-n} \binom{n}{k} r^{n-k} \leq 2^n r^{\alpha-k} \,.$$

This concludes the proof. □

B.2.3 Approximation by shallow Fourier neural networks

We start by reporting a known result.

Lemma B.7. *Let $g : [-\pi, \pi] \to \mathbb{R}$ 2π-periodic with modulus of continuity ω. Then there exists a trigonometric polynomial $q_n(t) = \sum_{k=-n}^{n} b_k e^{ikt}$, $b_k \in \mathbb{C}$, with real values (i.e. $q_n(t) \in \mathbb{R}$ for all $t \in [-\pi, \pi]$), such that*

$$\sup_{t\in[-\pi,\pi]} |g(t) - q_n(t)| \leq \frac{2}{\pi}\omega\left(\frac{2}{n}\right) \left(2 + \omega(\pi) - \log \omega\left(\frac{2}{n}\right)\right) .$$

Moreover, it holds that

$$|b_k| \leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |g(t)| \, dt \,.$$

Proof. The polinomyal q_n is given by the Fejer sum of the Fourier series of g, that is

$$q_n(t) = \frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=-j}^{j} \hat{g}_k e^{ikt} = \sum_{k=-(n-1)}^{n-1} \frac{n-|k|}{n} \hat{g}_k e^{ikt}$$

where

$$\hat{g}_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(t) e^{-ikt} \, dt \,.$$

The proof of the upper bound can be found in [Bur59], Theorem 18. Finally, notice that q_n is real-valued since

$$\hat{g}_k e^{ikt} + \hat{g}_{-k} e^{-ikt} = 2\mathrm{Re}\ \hat{g}_k e^{ikt}$$

because $\hat{g}_{-k} = \overline{\hat{g}_k}$ since g takes values in $\mathbb{R}$. □

The above result immediately implies a convergence rate for univariate approximation by shallow Fourier networks (that is, with activation $\sigma_1(t) = e^{2\pi i t}$).

Lemma B.8. *Let $f : [-r, r] \to \mathbb{R}$ be L-Lipschitz. Then there exists a real-valued Fourier shallow network $q_n(t) = \sum_{k=-n}^{n} b_k e^{iw_k t}$, $b_k \in \mathbb{C}$, $w_k \in \mathbb{R}$, such that*

$$\sup_{x\in[-r,r]} |f(x) - q_n(x)| \leq 3\ \ 1 + 2L^2 r^2\ \ \frac{\log n}{n}$$

for any $n \geq 2$. Moreover q_n can be chosen such that $|w_k| \leq \frac{\pi|k|}{r}$ and $|b_k| \leq \|f\|_\infty$ for any $k \in [-n, n]$.

Proof. Assume, w.l.o.g., that $f(r) \leq f(-r)$ (otherwise we can consider $f(-x)$ in place of $f(x)$). First, we want to transform f into a 2-pi periodic function on $[-\pi, \pi]$. To do this we consider $\tilde{g}$ defined as

$$\tilde{g}(x) = \begin{cases} L(x + r) + f(-r) & \text{if } x \in \ -r - \frac{c}{2L}, -r \\ f(x) & \text{if } x \in [-r, r] \\ L(x - r) + f(r) & \text{if } x \in \ r, r + \frac{c}{2L} \end{cases}$$

where $c = f(-r) - f(r)$. Notice that $\tilde{g}$ is L-Lipschitz and $2\ r + \frac{c}{2L}$ -periodic. Finally, let $g : [-\pi, \pi] \to \mathbb{R}$ defined as

$$g(x) = \tilde{g}\ \ \frac{2Lr + c}{2L\pi} x \quad .$$

We have that g is 2π-periodic and ℓ-Lipschitz for

$$\ell = \frac{2Lr + c}{2\pi} \leq \frac{2Lr}{\pi} .$$

Therefore, we can apply Lemma B.7 to g. This gives us a (real-valued) trigonometric polynomial $r_n(t) = \sum_{k=-n}^{n} b_k e^{ikt}$ such that

$$\begin{aligned} \sup_{x\in[-\pi,\pi]} |g(x) - r_n(x)| &\leq \frac{4\ell}{\pi n}\left(2 + \ell\pi - \log\frac{2\ell}{n}\right) \\ &\leq 3\left(1 + 2L^2r^2\right)\frac{\log n}{n} \end{aligned}$$

for $n \geq 2$. Since

$$\sup_{x\in[-r,r]} \left| f(x) - r_n\left(\frac{L}{\ell}x\right)\right| \leq \sup_{x\in\left[-r-\frac{c}{2L}, r+\frac{c}{2L}\right]} \left|\tilde{g}(x) - r_n\left(\frac{L}{\ell}x\right)\right| = \sup_{x\in[-\pi,\pi]} |g(x) - r_n(x)|$$

the thesis follows. □

To conclude we make some remarks about shallow Fourier networks. Note that a generic shallow Fourier network f_N with N units can be represented as

$$f(\mathbf{x}) = \sum_{k=1}^{N} u_k e^{i\mathbf{w}_k^T\mathbf{x}} . \tag{B.25}$$

Indeed we have that

$$\sum_{k=1}^{N} u_k e^{i(\mathbf{w}_k^T\mathbf{x} + b_k)} + b = \sum_{k=1}^{N} \left(u_k e^{ib_k}\right) e^{i\mathbf{w}_k^T\mathbf{x}} + b \cdot e^{i\mathbf{0}^T\mathbf{x}}$$

for any b, $b_k \in \mathbb{C}$. Let $\mathcal{F}_N^f$ be the space of networks as in equation (B.25). Notice that a universal approximation theorem holds for shallow Fourier networks as well. This is because the universal approximation theorem holds for shallow networks with activation $\sigma(t) = \cos(t)$ and since

$\cos(t) = (e^{it} + e^{-it})/2$, the thesis follows. Finally, the following lemma will be used in the proof of Theorem 3.6.

Lemma B.9. *If f is a (real-valued) shallow Fourier neural network, then so is f^k, for k non-negative integer. Moreover, if f has n units, then the number of units of f^k is upper bounded by*

$$\binom{n+k-1}{k}.$$

Proof. Let $f(\mathbf{x}) = \sum_{j=1}^{n} u_j e^{i\mathbf{w}_j^T \mathbf{x}}$ be a shallow Fourier neural network. Then, by the multinomial formula, we have that

$$\begin{aligned} f^k(\mathbf{x}) = \left(\sum_{j=1}^{n} u_j e^{i\mathbf{w}_j^T \mathbf{x}} \right)^k &= \sum_{p_1+\cdots+p_n=k} \binom{k}{p_1, \ldots, p_n} \prod_{j=1}^{n} u_j^{p_j} \left(e^{i\mathbf{w}_j^T \mathbf{x}} \right)^{p_j} \\ &= \sum_{p_1+\cdots+p_n=k} \binom{k}{p_1, \ldots, p_n} \prod_{j=1}^{n} u_j^{p_j} e^{i\left(\sum_{j=1}^{n} p_j \mathbf{w}_j \right)^T \mathbf{x}}. \end{aligned}$$

Clearly, if f is real-valued, so is f^k. Finally notice that by the formula above, the number of units of f^k is upper bounded by $|\{(p_1, \ldots, p_n) : p_1 + \cdots + p_n = k\}|$. □

B.2.4 $\mathrm{poly}(d)$ upper bounds for two-hidden-layers networks

Consider a two-hidden-layers neural network f defined as

$$f : \mathbf{x} \in \mathbb{R}^d \mapsto \boldsymbol{\gamma}^T \mathbf{g}\left(\mathbf{W}^T \mathbf{h}\left(\mathbf{U}^T \mathbf{x}\right)\right) \in \mathbb{C},$$

where $\mathbf{h} : \mathbb{R}^p \to \mathbb{R}^p$ and $\mathbf{g} : \mathbb{R}^o \to \mathbb{R}^o$ are, respectively, component-wise 1-Lipschitz and $(1, \alpha)$-Holder activation functions, and $\mathbf{U} \in \mathbb{R}^{d \times p}$, $\mathbf{W} \in \mathbb{R}^{p \times o}$, $\boldsymbol{\gamma} \in \mathbb{C}^o$. We wish to approximate f with a one-hidden-layer neural network with a given activation σ satisfying Assumption 1.2, for some constant $\nu_\sigma > 0$. We start by proving a result for approximation by shallow Fourier networks at a $\mathrm{poly}(d)$ rate.

Proposition B.10. *Let $K \subset \mathbb{R}^d$ be a compact set. There exist $f_N \in \mathcal{F}_N^f$ such that*

$$f - f_N^f \ _{K,\infty} \leq \epsilon$$

with

$$f_N^f(\mathbf{x}) = \sum_{\nu=1}^{N} b_\nu e^{i\mathbf{v}_\nu^T \mathbf{x}} \,,$$

for

$$N = (2np+1)^m$$

with

$$n = \left\lceil \frac{9 \cdot 4^{\frac{1}{\alpha}} \|\boldsymbol{\gamma}\|_1^2 \|\mathbf{W}\|_\infty^2 (1+2C^2)^2}{\epsilon^{\frac{2}{\alpha}}} \right\rceil \quad \textit{and} \quad m = \left\lceil \frac{2 \cdot 16^{\frac{1}{\alpha}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}} \left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1} \right)^{\frac{1}{\alpha}} + M \right)^{\alpha} \right\rceil ,$$

where we denoted

$$C = \sup_{x \in K} \|\mathbf{U}^T \mathbf{x}\|_\infty \quad \textit{and} \quad M = \sup_{x \in K} \ \mathbf{W}^T h \ \mathbf{U}^T \mathbf{x} \ _\infty .$$

Moreover f_N^f can be chosen such that it holds

$$\sup_{\mathbf{x} \in K} \mathbf{v}_\nu^T \mathbf{x} \ \leq \pi mn \quad \textit{and} \quad |b_\nu| \leq 2\|\boldsymbol{\gamma}\|_1 \left[1 + \left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1} \right)^{\frac{1}{\alpha}} + M \right)^{\alpha} (4npH\|\mathbf{W}\|_{F,\infty})^m \tag{B.26}$$

where $H = \sup_{\mathbf{x} \in [-C,C]^d} \|h(\mathbf{x})\|_\infty$.

Proof. Let q_n^j given by Lemma B.8 to approximate h_j over $[-C, C]$ and

$$q_k^{(n)}(\mathbf{x}) = \sum_{j=1}^{p} w_{k,j} q_n^j (\mathbf{u}_j^T \mathbf{x})$$

for $k \in [o]$. We have that

$$\begin{aligned} q_k^{(n)}(\mathbf{x}) - \mathbf{w}_k^T h\ \mathbf{U}^T\mathbf{x} \quad &\leq \sum_{j=1}^{p} |w_{k,j}|\ q_n^j(\mathbf{u}_j^T\mathbf{x}) - h_j(\mathbf{u}_j^T\mathbf{x}) \\ &\leq 3\|\mathbf{W}\|_\infty\ 1 + 2C^2\ \frac{\log n}{n} \doteq \|\mathbf{W}\|_\infty(1+2C^2)\epsilon_n \end{aligned}$$

for $\mathbf{x} \in K$. It holds that $q_k^{(n)}$ is a real-valued shallow Fourier network with $(2n-1)p$ terms and first layers weights given by $\frac{\pi k}{C}\mathbf{u}_j$ for $k \in [-(n-1), n-1]$. Moreover, it holds that

$$q_k^{(n)}(\mathbf{x})\ \leq\ q_k^{(n)}(\mathbf{x}) - \mathbf{w}_k^T h\ \mathbf{U}^T\mathbf{x} \quad +\ \mathbf{w}_k^T h\ \mathbf{U}^T\mathbf{x} \quad \leq \|\mathbf{W}\|_\infty(1+2C^2)\epsilon_n + M \doteq L\ .$$

Let $p_m^k(t) = \sum_{h=0}^{m} \beta_h^k t^h$ given by Corollary 3 to approximate g_k over the interval $[-L, L]$ and ϵ_m the relative error. Let then

$$f_{n,m}(\mathbf{x}) = \sum_{k=1}^{o} \gamma_k p_m^k(q_k^n(\mathbf{x}))\ .$$

It holds that

$$\begin{aligned} |f(\mathbf{x}) - f_{n,m}(\mathbf{x})| &\leq \sum_{k=1}^{o} |\gamma_k|\ g_k(\mathbf{w}_k^T h(\mathbf{U}^T\mathbf{x})) - p_m^k(q_k^{(n)}(\mathbf{x})) \\ &\leq \sum_{k=1}^{o} |\gamma_k|\ g_k(\mathbf{w}_k^T h(\mathbf{U}^T\mathbf{x})) - g_k(q_k^n(\mathbf{x}))\ + \sum_{k=1}^{o} |\gamma_k|\ g_k(q_k^{(n)}(\mathbf{x})) - p_m^k(q_k^{(n)}(\mathbf{x})) \\ &\leq \|\boldsymbol{\gamma}\|_1 \sup_{k\in[o]}\ \mathbf{w}_k^T h(\mathbf{U}^T\mathbf{x}) - q_k^{(n)}(\mathbf{x})\ ^\alpha + \|\boldsymbol{\gamma}\|_1\epsilon_m \\ &\leq \|\boldsymbol{\gamma}\|_1\|\mathbf{W}\|_\infty^\alpha(1+2C^2)^\alpha\epsilon_n^\alpha + \|\boldsymbol{\gamma}\|_1\epsilon_m\ . \end{aligned}$$

It holds that

$$\|\boldsymbol{\gamma}\|_1\|\mathbf{W}\|_\infty^\alpha(1+2C^2)^\alpha\epsilon_n^\alpha \leq \frac{\epsilon}{2}$$

as long as

$$n \geq \frac{9 \cdot 4^{\frac{1}{\alpha}}\|\boldsymbol{\gamma}\|_1^2\|\mathbf{W}\|_\infty^2(1+2C^2)^2}{\epsilon^{\frac{2}{\alpha}}}\ . \tag{B.27}$$

Similarly

$$\|\boldsymbol{\gamma}\|_1 \epsilon_m \leq \frac{\epsilon}{2}$$

as long as

$$m \geq L \left(\frac{12\|\boldsymbol{\gamma}\|_1}{\epsilon} \right)^{\frac{1}{\alpha}} = \left(\frac{12\|\boldsymbol{\gamma}\|_1}{\epsilon} \right)^{\frac{1}{\alpha}} \left(\|\mathbf{W}\|_\infty (1+2C^2)\epsilon_n + M \right) .$$

Moreover, by Lemma B.6, $p_m^k(t) = \sum_{h=0}^{m} \beta_h^k t^h$ can be chosen with

$$m \geq \frac{2 \cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}} L^\alpha = \frac{2 \cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}} \left(\|\mathbf{W}\|_\infty (1+2C^2)\epsilon_n + M \right)^\alpha$$

such that its coefficients β_h^k, $k \in [m]$, are bounded by

$$\begin{aligned} |\beta_k| &\leq \max \left(2^m L^{\alpha-k}, L^\alpha + |g(0)| \right) \leq 2^m (1 + L^\alpha) + |g(0)| \\ &= 2^m \left(1 + \left(\|\mathbf{W}\|_\infty (1+2C^2)\epsilon_n + M \right)^\alpha \right) + |g(0)| . \end{aligned}$$

Notice that we can assume $g(0) = 0$ without loss of generality. Therefore

$$\sup_{x \in K} |f(\mathbf{x}) - f_{n,m}(\mathbf{x})| \leq \epsilon$$

as long as (B.27) holds and

$$m \geq \left(\frac{12\|\boldsymbol{\gamma}\|_1}{\epsilon} \right)^{\frac{1}{\alpha}} \left[\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1} \right)^{\frac{1}{\alpha}} + M \right] = 6^{\frac{1}{\alpha}} \left(1 + M \left(\frac{2\|\boldsymbol{\gamma}\|_1}{\epsilon} \right)^{\frac{1}{\alpha}} \right) . \tag{B.28}$$

If we further assume that

$$m \geq \frac{2 \cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}} \left[\left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1} \right)^{\frac{1}{\alpha}} + M \right]^\alpha$$

we can also assume that

$$\beta_h^k \leq 2^{1+\frac{2\cdot 16^{\frac{1}{\alpha}}\|\gamma\|_1^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}}\left[\left(\frac{\epsilon}{2\|\gamma\|_1}\right)^{\frac{1}{\alpha}}+M\right]^{\alpha}} \quad 1+\left[\frac{\epsilon}{2\|\gamma\|_1}^{\frac{1}{\alpha}}+M\right]^{\alpha}$$

for $k \in [m]$. Finally, notice that, by Lemma B.9, $f_{n,m}$ is a shallow Fourier neural network with number of units upper bounded by

$$\begin{aligned} N &= \sum_{k=0}^{m} \binom{(2n-1)p+k-1}{k} = \binom{(2n-1)p+m}{m} \\ &= \frac{1}{m!}((2n-1)p+k+m)\cdots((2n-1)p+1) \\ &\leq ((2n-1)p+1)^m \,. \end{aligned}$$

Therefore, it holds that

$$\inf_{f_N \in \mathcal{F}_N^f} \sup_{\mathbf{x}\in K} |f(\mathbf{x}) - f_N(\mathbf{x})| \leq \epsilon$$

as long as

$$N \geq (2np+1)^m$$

with n and m given by (B.27) and (B.28) respectively. Finally, notice that the first layer weights of $f_{n,m}$ are given by

$$\sum_{j=1}^{p}\sum_{k=-(n-1)}^{n-1} s_{k,j}\frac{\pi k}{C}u_j$$

over all non-negative integers $s_{k,j}$ such that $\sum_{j=1}^{p}\sum_{k=-(n-1)}^{n-1} s_{k,j} \leq m$. Therefore, if

$$f_{n,m}(\mathbf{x}) = \sum_{\nu=1}^{N} b_\nu e^{i\mathbf{v}_\nu^T\mathbf{x}} \,,$$

then

$$\mathbf{v}_\nu^T\mathbf{x} \leq m\frac{\pi(n-1)}{C}\max_{j\in[p]}|\mathbf{u}_j^T\mathbf{x}| \leq mn\pi \,.$$

On the other hand, the coefficients b_k have the form

$$b_\nu = \begin{matrix} h \\ s \end{matrix} \sum_{k=1}^{o} \gamma_k \beta_h^k \; w_{k,j} (q_n^j)_l \; \begin{matrix} \\ s_{l,j} \end{matrix}$$

for all non-negative integers $s = (s_{l,j})_{l,j}$ such that $\begin{matrix} p \\ j=1 \end{matrix} \quad \begin{matrix} n-1 \\ l=-(n-1) \end{matrix} s_{l,j} = h \leq m$, where $(q_n^j)_l$ denotes the l-th coefficients of q_n^j. By Lemma B.7, we know that

$$(q_n^j)_l \; \leq \sup_{t \in [-C,C]} |h_j(t)| \, .$$

Therefore

$$\begin{aligned} |b_\nu| &\leq ((2n-1)p)^h \sup_{t \in [-C,C]} |h_j(t)|^{s_{l,j}} \sum_{k=1}^{o} |\gamma_k| |\beta_h^k| |w_{k,j}|^{s_{l,j}} \\ &\leq [(2n-1)p \, H \, \|\mathbf{W}\|_{F,\infty}]^m \|\boldsymbol{\gamma}\|_1 \|\boldsymbol{\beta}\|_{F,\infty} \, . \end{aligned}$$

This concludes the proof. □

We can now conclude with a detailed version of Theorem 3.6.

Theorem B.11. *Let K be a compact set and*

$$C = \sup_{\mathbf{x} \in K} \|\mathbf{U}^T \mathbf{x}\|_\infty \, , \quad M = \sup_{\mathbf{x} \in K} \; \mathbf{W}^T \mathbf{h} \; \mathbf{U}^T \mathbf{x} \quad _\infty \quad \textit{and} \quad H = \sup_{\mathbf{x} \in [-C,C]^d} \|\mathbf{h}(\mathbf{x})\|_\infty \, .$$

It holds that

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f(\mathbf{x}) - f_N^\sigma(\mathbf{x})\|_{K,\infty} \leq \epsilon \tag{B.29}$$

for some

$$N \leq \frac{16\pi\nu_\sigma}{\epsilon} \|\boldsymbol{\gamma}\|_1 mn(4np+1)^{2m} (H\|\mathbf{W}\|_{F,\infty})^m \left[1 + \quad \frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1} \quad ^{\frac{1}{\alpha}} + M \quad ^{\alpha} \right. \quad ,$$

where

$$n = \frac{9 \cdot 4^{\frac{1}{\alpha}} \|\boldsymbol{\gamma}\|_1^2 \|\mathbf{W}\|_\infty^2 (1+2C^2)^2}{\epsilon^{\frac{2}{\alpha}}} \quad \textit{and} \quad m = \frac{2 \cdot 16^{\frac{1}{\alpha}}}{\epsilon^{1+\frac{2}{\alpha}}} \|\boldsymbol{\gamma}\|_1^{\frac{1}{\alpha}} \left(\frac{\epsilon}{2\|\boldsymbol{\gamma}\|_1} \right)^{\frac{1}{\alpha}} + M \Bigg)^{\alpha}.$$

Moreover, it is possible to choose f_N^σ attaining (B.29) *with $m_\infty(f_N^\sigma)$ satisfying a bound similar to the one on N, for example $m_\infty(f_N^\sigma) \leq (1+N^2)$.*

Proof. Let f_N given by Proposition B.10 such that

$$\sup_{\mathbf{x} \in K} |f(\mathbf{x}) - f_N(\mathbf{x})| \leq \frac{\epsilon}{2}.$$

We know that

$$f_N(\mathbf{x}) = \sum_{k=1}^{N} b_k e^{i \mathbf{v}_k^T \mathbf{x}} = f_N^c(\mathbf{x}) + i f_N^s(\mathbf{x})$$

where

$$f_N^c(\mathbf{x}) = \sum_{k=1}^{N} b_k \cos(\mathbf{v}_k^T \mathbf{x}) \quad \text{and} \quad f_N^s(\mathbf{x}) = \sum_{k=1}^{N} b_k \sin(\mathbf{v}_k^T \mathbf{x})$$

and $|b_k| \leq B$ and $\mathbf{v}_k^T \mathbf{x} \leq V$ for $\mathbf{x} \in K$, where B and V are given by (B.26). Using the assumption on σ, we know that, for each $k \in [N]$, there exist shallow networks f_k^c and f_k^s with activation σ and number of units

$$n \leq c_\sigma \frac{4VBN}{\epsilon}$$

such that

$$\sup_{\mathbf{x} \in K} f_k^c(\mathbf{x}) - \cos(\mathbf{v}_k^T \mathbf{x}) \leq \frac{\epsilon}{4NB} \quad \text{and} \quad \sup_{\mathbf{x} \in K} f_k^s(\mathbf{x}) - \sin(\mathbf{v}_k^T \mathbf{x}) \leq \frac{\epsilon}{4NB}.$$

Letting $f_{\mathcal{N}}(\mathbf{x}) = \sum_{k=1}^{N} b_k f_k^c(\mathbf{x}) + i \sum_{k=1}^{N} b_k f_k^s(\mathbf{x})$ it holds that

$$\begin{aligned}
\sup_{\mathbf{x}\in K}|f_{\mathcal{N}}(\mathbf{x}) - f_N(\mathbf{x})| &\leq \sup_{\mathbf{x}\in K} \sum_{k=1}^{N} b_k \ f_k^c(\mathbf{x}) - \cos(\mathbf{w}_k^T\mathbf{x}) \quad + \sup_{\mathbf{x}\in K} \sum_{k=1}^{N} b_k \ f_k^s(\mathbf{x}) - \sin(\mathbf{w}_k^T\mathbf{x}) \\
&\leq \sum_{k=1}^{N} |b_k| \sup_{\mathbf{x}\in K} \ f_k^c(\mathbf{x}) - \cos(\mathbf{w}_k^T\mathbf{x}) \ + \sum_{k=1}^{N} |b_k| \sup_{\mathbf{x}\in K} \ f_k^s(\mathbf{x}) - \sin(\mathbf{w}_k^T\mathbf{x}) \\
&\leq NB\frac{\epsilon}{4NB} + NB\frac{\epsilon}{4NB} = \frac{\epsilon}{2}
\end{aligned}$$

which implies that

$$\sup_{\mathbf{x}\in K}|f_{\mathcal{N}}(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon \ .$$

Moreover notice that we can assume that all second layer weights of $f_{\mathcal{N}}$ are real; indeed, if this is not the case, one can replace them by the real part, and upper bound above can only decrease. Finally, we have that the number of units of $f_{\mathcal{N}}$ is given by

$$\mathcal{N} \leq \frac{8c_\sigma}{\epsilon} \cdot V \cdot B \cdot N \ .$$

Applying Proposition B.10 concludes the proof. □

B.3 Proofs of special cases

B.3.0.1 Radial functions

Let $f(\mathbf{x}) = \varphi(\|\mathbf{x}\|)$ with φ 1-Lipschitz. Then it holds that $f(\mathbf{x}) = g(\mathbf{1}^T\mathbf{h}(\mathbf{x}))$ where $g(t) = \varphi(\sqrt{t})$ and $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^d$ is defined as $h_i(\mathbf{x}) = x_i^2$. Clearly, $\sup_{\mathbf{x}\in B_{1,2}^d}\|\mathbf{x}\|_\infty = 1$, $\sup_{\mathbf{x}\in B_{1,2}^d} \mathbf{1}^T\mathbf{h}(\mathbf{x}) = \sup_{\mathbf{x}\in B_{1,2}^d}\|\mathbf{x}\|^2 = 1$ and $\sup_{\mathbf{x}\in[-1,1]^d}\|\mathbf{h}(\mathbf{x})\|_\infty = \sup_{x\in[-1,1]}|x|^2 = 1$. Moreover, $\|\mathbf{1}\|_1 = d$ and g is $(1, 1/2)$-Holder. Then, by applying Theorem B.11, we get the following.

Corollary B.12 (Radial functions)**.** *It holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f_N^\sigma - f\|_{B^d_{1,2,\infty}} \leq \epsilon$$

for some

$$N \leq \nu_\sigma \alpha \cdot d^2 \cdot \frac{(4+\epsilon)^2}{\epsilon^{10}} \left(\alpha \frac{d^3}{\epsilon^4} + 1 \right)^{\frac{\alpha}{\epsilon^5}(2+\ )}$$

where $\alpha > 0$ is a numerical constant.

B.3.0.2 Shallow approximation of (3.1)

Consider $f_{\mathbf{w},\mathbf{U}} : \mathbf{x} \in \mathbb{R}^d \mapsto e^{i\mathbf{w}^T(\mathbf{U}\mathbf{x})_+}$ for some $\mathbf{w} \in \mathbb{R}^p$, $\mathbf{U} \in \mathbb{R}^{p\times d}$. Then Theorem B.11 implies the following.

Corollary B.13 (Approximation of (3.1) by shallow networks)**.** *It holds that*

$$\inf_{f_N^\sigma \in \mathcal{F}_N^\sigma} \|f_{\mathbf{w},\mathbf{U}} - f_N^\sigma\|_{B^d_{r,p},\infty} \leq \epsilon$$

for some

$$N \leq \frac{\nu_\sigma \beta}{\epsilon^6} \cdot (2 + \epsilon + 2r\|\mathbf{w}\|_1 \|\mathbf{U}\|_{p,\infty})^2 \cdot \left[r\|\mathbf{w}\|_\infty \|\mathbf{U}\|_{p,\infty} \left(\frac{4p\beta}{\epsilon^2} + 1 \right)^{2} \right]^{\frac{\alpha}{\epsilon^2}(\ +2r\|\mathbf{w}\|_1\|\mathbf{U}\|_{p,\infty})}$$

where $\beta = \alpha\|\mathbf{w}\|_1^2 \cdot \left(1 + 2r^2\|\mathbf{U}\|_{p,\infty}^2\right)^2$ and α is a numerical constant.

B.3.0.3 Approximation bounds under the Gaussian metric

For sake of simplicity in this section we consider approximation bounds for the function of interest

$$f_{\mathbf{w},\mathbf{U}} : \mathbf{x} \in \mathbb{R}^d \mapsto e^{i\mathbf{w}^T(\mathbf{U}\mathbf{x})_+}$$

for some $\mathbf{w} \in \mathbb{R}^p$, $\mathbf{U} = [\mathbf{u}_1| \cdots |\mathbf{u}_p]^T \in \mathbb{R}^{p \times d}$. Notice that the following results can be naturally extended to any three-layer network target. We are interested in upper bounding the error

$$\inf_{f_N \in \mathcal{F}_N^f} \mathbb{E}|f_{\mathbf{w},\mathbf{U}}(\mathbf{X}) - f_N(\mathbf{X})|^2 \ ^{\frac{1}{2}}$$

where the expectation is taken over $\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. For sake of simplicity of notation, we denote

$$\|f - g\|_{\sigma,2} \doteq \ \mathbb{E}|f(\mathbf{X}) - g(\mathbf{X})|^2 \ ^{\frac{1}{2}} .$$

It is a well known fact that Gaussian vectors concentrates in a ball of radius $\sqrt{d}$. We recall a quantitative version of this fact in the following.

Lemma B.14. *Let $\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ a d-dimensional Gaussian vector. Then it holds that*

$$P \ \ \|\mathbf{X}\|_2 \geq \sigma\sqrt{d} + t \ \ \leq e^{-\frac{t^2}{2\sigma^2}} .$$

Thanks to Proposition B.10, the following holds.

Lemma B.15. *Let $r > 0$. Then it holds that*

$$\inf_{f_N \in \mathcal{F}_N^f} \|f_N - f_{\mathbf{w},\mathbf{U}}\|_{B^d_{r,2},\infty} \leq \delta \tag{B.30}$$

as long as

$$N \geq (2np + 1)^m$$

where

$$n = \frac{36}{\delta^2}\|\mathbf{w}\|_1^2 \ 1 + r^2\|\mathbf{U}\|_{2,\infty}^2 \ ^2 \quad \textit{and} \quad m \geq \frac{16}{\delta^3}(\delta + 2r\|\mathbf{w}\|_1\|\mathbf{U}\|_{2,\infty}) .$$

Moreover, under the same assumption, we can also assume that the function f_N that satisfies (B.30)

also satisfies

$$\|f_N\|_\infty \leq N(2+\delta+2r\|\mathbf{w}\|_1\|\mathbf{U}\|_{2,\infty})(4npr\|\mathbf{w}\|_\infty\|\mathbf{U}\|_{2,\infty})^m \ .$$

Thanks to these two lemmas, the following proposition follows.

Proposition B.16. *Let $\sigma = d^{-1/2}$ and assume that $\|\mathbf{U}\|_{2,\infty} \leq 1$. Then it holds*

$$\inf_{f_N \in \mathcal{F}_N^f} \|f_N - f_{\mathbf{w},\mathbf{U}}\|_{\sigma,2} \leq \epsilon \tag{B.31}$$

as long as

$$N \geq \quad Kp \quad 1+\frac{1}{\epsilon^s} \quad (1+\|\mathbf{w}\|_1^s)^{K\left(1+\left(\frac{\log p}{d}\right)^s\right)\left(1+\frac{1}{\epsilon^s}\right)\left(1+\|\mathbf{w}\|_1^s\right)}$$

where $K > 0$ and $s \geq 1$ are some numerical constant.

Proof. Let $c = \|\mathbf{w}\|_1$. First, notice that $\|f_{\mathbf{w},\mathbf{U}}\|_\infty = 1$. Let $\chi_r(\mathbf{x}) = \mathbb{1}\{\|\mathbf{x}\|_2 \leq r\}$ and f_N given by Lemma B.15 for a certain $\delta > 0$. Then it holds that

$$\begin{aligned}\|f_N - f_{\mathbf{w},\mathbf{U}}\|_{\sigma,2} &\leq \|(f_N - f_{\mathbf{w},\mathbf{U}})(1-\chi_r)\|_{\sigma,2} + \|(f_N - f_{\mathbf{w},\mathbf{U}})\chi_r\|_{\sigma,2} \\ &\leq \|f_N - f_{\mathbf{w},\mathbf{U}}\|_{B^d_{r,2},\infty} + P(\|\mathbf{x}\|_2 > r)(\|f_N\|_\infty + \|f_{\mathbf{w},\mathbf{U}}\|_\infty) \ .\end{aligned}$$

If $r = 1+t$ for $t > 0$, it follows

$$\|f_N - f_{\mathbf{w},\mathbf{U}}\|_{2,\sigma} \leq \delta + e^{-\frac{dt^2}{2}}(1+\|f_N\|_\infty)$$

as long as

$$N \geq \quad \frac{72p}{\delta^2}c^2 \ \ 1+r^2 \ \ ^2+1 \ \ ^{\frac{1}{\delta^3}(\delta+2rc)} \ .$$

Moreover, one can assume

$$\begin{aligned}\|f_N\|_\infty &\leq (2+\delta+2rc)\ \ \frac{72p}{\delta^2}c^2\ \ 1+r^2\ \ ^2+1\ \ ^{\frac{16}{\delta^3}(\delta+2rc)}\ \ 144\frac{pr}{\delta^2}c^3\ \ 1+r^2\ \ ^2\ \ ^{\frac{16}{\delta^3}(\delta+2r)}\\ &\leq (2+\delta+2r\omega)\ \ 144\frac{p}{\delta^2}\omega^3 r(1+r^2)^2+1\ \ ^{\frac{32}{\delta^3}(\delta+2r\omega)}\end{aligned}$$

where $\omega=\max(1,c)$. Let $\delta=\frac{}{2}$. If $t\geq 1$, it holds that

$$\begin{aligned}\|f_N\|_\infty &\leq (4\omega+\epsilon+2\omega t)\ \ 576\frac{p}{\epsilon^2}\omega^3(1+t)\ \ 1+(1+t)^2\ \ ^2+1\ \ ^{\frac{256}{\epsilon^3}(\ +2\omega+2\omega t)}\\ &\leq K(\epsilon+\omega+\omega t)\ \ K\frac{p}{\epsilon^2}\omega^2 t^5+1\ \ ^{\frac{K}{\epsilon^3}(\ +\omega+\omega t)}\ .\end{aligned}$$

In the equation above above and in the following, K denotes a (large enough) numerical constant. Therefore

$$e^{-\frac{dt^2}{2}}(1+\|f_N\|_\infty)\leq\frac{\epsilon}{2} \tag{B.32}$$

as long as

$$\frac{dt^2}{2}-\log\ \ 1+K(\epsilon+\omega+\omega t)\ \ K\frac{p}{\epsilon^2}\omega^2t^5+1\ \ ^{\frac{K}{\epsilon^3}(\ +\omega+\omega t)}\ \ +\log\frac{\epsilon}{2}\geq 0\ .$$

Since $\log(1+Cs^\alpha)\leq\log(1+C)+\alpha\log(s)$ if $s\geq 1$, $C>0$ and $\alpha>0$, the above is implied by

$$\frac{dt^2}{2}-\log(1+K(\epsilon+\omega+\omega t))-\frac{K}{\epsilon^3}(\epsilon+\omega+\omega t)\log\ \ K\frac{p}{\epsilon^2}\omega^2t^5+1\ \ +\log\frac{\epsilon}{2}\geq 0\ .$$

Since

$$\log(1+K(\epsilon+\omega+\omega t))\leq K(\epsilon+\omega+\omega t)$$

and

$$\log\ \ K\frac{p}{\epsilon^2}\omega^2t^5+1\ \ \leq\log\ \ 1+K\frac{p\omega^2}{\epsilon^2}\ \ +5\log t\leq\log\ \ 1+K\frac{p\omega^2}{\epsilon^2}\ \ +5\sqrt{t}$$

equation (B.32) holds if

$$\frac{dt^2}{2} - \alpha - \beta t^{1/2} - \gamma t - \eta t^{3/2} \geq 0$$

where

$$\begin{aligned}
\alpha &= K(\epsilon+\omega) + \frac{K}{\epsilon^3}(\epsilon+\omega)\log\left(1+K\frac{p\omega^2}{\epsilon^2}\right) - \log\frac{\epsilon}{2} > 0\,, \\
\beta &= \frac{K}{\epsilon^3}(\epsilon+\omega) > 0\,, \\
\gamma &= K\omega t + \frac{K}{\epsilon^3}\omega\log\left(1+K\frac{p\omega^2}{\epsilon^2}\right) > 0\,, \\
\eta &= \frac{K}{\epsilon^3}\omega t > 0\,.
\end{aligned}$$

It follows that eq. (B.32) holds if

$$t \geq 1 + 4\left(\frac{\alpha+\beta+\gamma+\eta}{d}\right)^2.$$

It follows that the error bound (B.31) holds as long as

$$N \geq \left(\frac{Kp}{\epsilon^2}(1+c)^2\left(1+4\left(\frac{\alpha+\beta+\gamma+\eta}{d}\right)^2\right)^4+1\right)^{\frac{K}{\epsilon^3}+c\left(1+\left(\frac{\alpha+\beta+\gamma+\eta}{d}\right)^2\right)}.$$

The thesis follows. □

B.3.1 Extension to generic L-layers networks

The results presented in the previous section can be generalized to hold for approximating generic multi-layer neural networks. In this section we present an analogous result to Theorem 3.6 for this more general case. Consider a multi-layer neural network f defined as

$$f : \mathbf{x} \in \mathbb{R}^d \to x^{(L)}(\mathbf{x}) \in \mathbb{C}$$

where $x^{(L)}$ is defined by recursion by $\mathbf{x}^{(0)}(\mathbf{x}) = \mathbf{x}$,

$$\mathbf{x}^{(k)}(\mathbf{x}) = \boldsymbol{\sigma}^{(k)}(\mathbf{A}^{(k)}\mathbf{x}^{(k-1)}(\mathbf{x})) \text{ for } k \in [L] \quad \text{and} \quad x^{(L+1)}(\mathbf{x}) = \mathbf{a}^{(L+1)\ T}\mathbf{x}^{(L)}(\mathbf{x}) \ ,$$

where $\mathbf{A}^{(k)} = [\mathbf{a}_1^{(k)}| \cdots |\mathbf{a}_{d_k}^{(k)}]^T \in \mathbb{R}^{d_k \times d_{k-1}}$ for $k \in [L]$ (with $d_0 = d$), $\mathbf{a}^{(L+1)} \in \mathbb{C}^{d_L}$ and $\boldsymbol{\sigma}^{(k)} : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ are $\frac{1}{6}$-Lipschitz component-wise activation functions and verify $\boldsymbol{\sigma}^k(\mathbf{0}) = \mathbf{0}$ for $k \in [L]$. In the following we also assume that $\|\mathbf{A}^{(k)}\|_\infty \leq 1$ for $k \in [L]$ and $\|\mathbf{a}_{L+1}\|_1 \leq 1$. Note that these assumption can easily be relaxed, but we adopt them here for sake of simplicity.

Proposition B.17. *Let f as above. It holds that*

$$\inf_{f_N \in \mathcal{F}_N^f} \|f - f_N\|_{B^d_{1,\infty},\infty} \leq \epsilon$$

as long as

$$N \geq \quad 2^L C \quad 1 + \frac{1}{\epsilon^2} \quad d_1 \quad ^{CL\left(1+\frac{1}{\epsilon}\right)^{L-1}}$$

where C is a numerical constant.

Before proving the above proposition, we prove two preliminary lemmas.

Lemma B.18. *Let $\mathcal{W} = \{\mathbf{w}_\ell\}_{\ell \in [K]} \subset \mathbb{R}^d$ and $\mathbf{h} : \mathbb{R}^d \to \mathbb{R}^p$ such that h_j is a shallow Fourier neural networks with first layer weights given by $\mathcal{W}$, for all $j \in [p]$. Consider $\mathbf{q} : \mathbb{R}^p \to \mathbb{R}^m$ of the form*

$$\mathbf{q}(\mathbf{x}) = \mathbf{B}\boldsymbol{\sigma}(\mathbf{x})$$

where $\boldsymbol{\sigma} : \mathbb{R}^p \to \mathbb{R}^p$ is a component-wise polynomial activation function of degree at most D and $\mathbf{B} \in \mathbb{C}^{m \times p}$. Then there exists $\mathcal{V} \subset \mathbb{R}^d$ finite such that $\mathbf{f} \doteq \mathbf{q} \circ \mathbf{h}$ is such that f_j is a Fourier neural nets with first layer weights given by $\mathcal{V}$ for each $j \in [p]$ and such that

$$|\mathcal{V}| \leq (2K)^D \ .$$

Proof. The functions f_j have the form

$$f_j(\mathbf{x}) = \sum_{k=1}^{p} b_{jk} \sum_{l=0}^{D} \alpha_{k,l} (h_k(\mathbf{x}))^l = \sum_{k=1}^{p} b_{jk} \sum_{l=0}^{D} \alpha_{k,l} \left(\sum_{\nu=1}^{K} \beta_{k,\nu} e^{i \mathbf{w}_\nu^T \mathbf{x}} \right)^l .$$

By Lemma B.9, we see that each f_j is a Fourier neural network with the same set of first layer weights of size at most

$$\sum_{l=0}^{D} \binom{K+l-1}{l} = \binom{K+D}{D} \leq (K+1)^D \leq (2K)^D .$$

This concludes the proof. □

Lemma B.19. *Consider the same assumption as Proposition B.17. Then, there exists a polynomial*

$$f_{N_1,\ldots,N_L} : \mathbf{x} \in \mathbb{R}^d \to y^{(L+1)}(\mathbf{x}) \in \mathbb{C}$$

given by the recursion $\mathbf{y}^{(0)}(\mathbf{x}) = \mathbf{x}$*,*

$$\begin{aligned} \mathbf{y}^{(k)}(\mathbf{x}) &= \mathbf{p}_{N_k}^k (\mathbf{A}^{(k)} \mathbf{y}^{(k-1)}(\mathbf{x})) \quad \textit{for } k \in [L] \\ y^{(L+1)}(\mathbf{x}) &= \mathbf{a}^{(L+1)\ T} \mathbf{y}^{(L)}(\mathbf{x}) \end{aligned}$$

where $\mathbf{p}_{N_k}^k$ *are component-wise polynomial activation functions of degree* N_k*, such that*

$$\| f - f_{N_1,\ldots,N_L} \|_{B^d_{1,\infty},\infty} \leq \epsilon \tag{B.33}$$

as long as $N_k \geq \frac{L}{} + (L-1)$ *for* $k \in [L]$*. In particular,* f *is a polynomial of degree* $\prod_{k=1}^{L} N_k$*.*

Proof. We can show this by induction over L. First, consider the case $L = 1$. By Lemma B.5, for

each $j \in [d_1]$, there exist polynomials $p_{N,j} : \mathbb{R} \to \mathbb{R}$ of degree N which verify

$$p_{N,j}((\mathbf{a}_i^{(1)})^T \mathbf{x}) - \sigma_j^{(1)}((\mathbf{a}_j^{(1)})^T \mathbf{x}) \ \leq \frac{1}{N}$$

since $(\mathbf{a}_i^{(1)})^T \mathbf{x} \ \leq 1$ by assumption. Since $\|\mathbf{a}^{(2)}\|_1 \leq 1$, it follows that

$$(\mathbf{a}^{(2)})^T \mathbf{p}_N(\mathbf{A}^{(1)} \mathbf{x}) - (\mathbf{a}^{(2)})^T \boldsymbol{\sigma}^{(1)}(\mathbf{A}^{(1)} \mathbf{x}) \ \leq \frac{1}{N} \, .$$

This implies the thesis for the case $L = 1$. Now consider the induction step, that is, assume that, for every $\delta > 0$ and j, there exists a certain $f^j_{N_1,\dots,N_{L-1}}$ such that

$$x_j^{(L-1)}(\mathbf{x}) - f^j_{N_1,\dots,N_{L-1}}(\mathbf{x}) \ \leq \delta$$

as long as $N_k \geq \frac{L-1}{\delta} + (L-2)$ for $k \in [L-1]$. Notice that this implies that

$$(\mathbf{a}_j^{(L)})^T \mathbf{f}_{N_1,\dots,N_{L-1}}(\mathbf{x}) \ \leq 1 + \delta \, ,$$

where $\mathbf{f}_{N_1,\dots,N_{L-1}} = (f^1_{N_1,\dots,N_{L-1}}, \dots, f^{d_{L-1}}_{N_1,\dots,N_{L-1}})$. Therefore for each $j \in [d_L]$, by Lemma B.5, there exist polynomials $p_{N,j}$ of degree N such that

$$p_{N,j}((\mathbf{a}_j^{(L)})^T \mathbf{f}_{N_1,\dots,N_{L-1}}(\mathbf{x})) - \sigma_j^{(L)}((\mathbf{a}_j^{(L)})^T \mathbf{f}_{N_1,\dots,N_{L-1}}(\mathbf{x})) \ \leq \frac{1+\delta}{N} \, .$$

Let then $f_{N_1,\dots,N_{L-1},N}$ be defined as

$$f_{N_1,\dots,N_{L-1},N}(\mathbf{x}) = \sum_{j=1}^{N} a_j^{(L+1)} p_{N,j}((\mathbf{a}_j^{(L)})^T \mathbf{f}_{N_1,\dots,N_{L-1}}(\mathbf{x})) \, .$$

Since $\|\mathbf{a}^{(L+1)}\|_1 \leq 1$, it holds that

$$\begin{aligned} f_{N_1,\dots,N_{L-1},N}(\mathbf{x}) - f(\mathbf{x}) \;&\leq f_{N_1,\dots,N_{L-1},N}(\mathbf{x}) - \mathbf{a}_{L+1}^T \boldsymbol{\sigma}^{L+1}\; f_{N_1,\dots,N_{L-1},N}(\mathbf{x}) \\ &\quad + \mathbf{a}_{L+1}^T \boldsymbol{\sigma}^{L+1}\; f_{N_1,\dots,N_{L-1},N}(\mathbf{x}) \;- f(\mathbf{x}) \\ &\leq \frac{1+\delta}{N} + \delta \,. \end{aligned}$$

If $\delta = \frac{L-1}{L}\epsilon$ then equation (B.33) holds as long as

$$N \geq \frac{1 + \frac{L-1}{L}\epsilon}{\overline{L}} = \frac{L}{\epsilon} + (L-1) \,.$$

This concludes the proof of the lemma. □

Proof of Proposition B.17. It holds that

$$f(\mathbf{x}) = g(\boldsymbol{\sigma}^{(1)}(\mathbf{A}^{(1)}\mathbf{x}))$$

where g is a $(L-1)$-hidden-layers neural network with input dimension d_1. By Lemma B.7, for every $\delta > 0$ and $j \in [d_1]$, there exists Fourier networks $q_{N_1,j}(\mathbf{x})$ with $2N_1 - 1$ units such that

$$\sigma_j^{(1)}((\mathbf{a}_j^{(1)})^T\mathbf{x}) - q_{N_1,j}((\mathbf{a}_j^{(1)})^T\mathbf{x}) \;\leq \frac{C}{\sqrt{N_1}}$$

where $C > 0$ is a numerical constant. Notice that this implies that, for $N_1 \geq 4C^2$, it holds

$$\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x}) \;_\infty \leq 1 \,.$$

Now, we can approximate g with a polynomial neural network $g_{N_L,\dots,N_2}$ as given by Lemma B.19.

In particular, for any $\delta > 0$, there exist $g_{N_L,\ldots,N_2}$ such that

$$\sup_{\mathbf{x}\in[-1,1]^d} |g_{N_L,\ldots,N_2}(\mathbf{x}) - g(\mathbf{x})| \leq \delta$$

as long as $N_k \geq \frac{L-1}{\delta} + (L-2)$ for $k \in [2, L]$. It follows that

$$g_{N_L,\ldots,N_2}(\mathbf{q}_{N_1}(\mathbf{A}^1\mathbf{x})) - f(\mathbf{x}) \ \leq \delta + \frac{C}{\sqrt{N_1}} \, .$$

Let $f_N(\mathbf{x}) = g_{N_L,\ldots,N_2}(\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x}))$. By choosing $\delta = \epsilon/2$, it holds that

$$\sup_{\mathbf{x}\in[-1,1]^d} |f_N(\mathbf{x}) - f(\mathbf{x})| \leq \epsilon$$

as long as $N_k \geq 2^{\frac{L-1}{}} + (L-2)$ for $k \in [2, L]$ and $N_1 \geq C^2 \ 1 + \frac{4}{2}$. We claim that f_N is a Fourier network with at most

$$N = \ 2^L N_1 d_1 \ \prod_{k=2}^{L} N_k \tag{B.34}$$

units. We can prove this by induction over $L \geq 2$. Remember that $g_{N_L,\ldots,N_2}$ is is the form

$$g_{N_L,\ldots,N_2}(\mathbf{x}) = \ \mathbf{a}^{(L+1)\ T}\mathbf{g}^L_{N_L} \ \ \mathbf{A}^{(L)}\mathbf{g}^{L-1}_{N_{L-1}} \ \ \mathbf{A}_{(L-1)} \cdots \mathbf{g}^2_{N_2}(\mathbf{A}^{(2)}\mathbf{x})$$

where $\mathbf{g}^k_{N_k}$ is a component-wise polynomial of degree at most N_k, for $k \in [2, L]$. We start by the case $L = 2$. Notice that each component of $\mathbf{A}^{(2)}\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x})$ is a Fourier network with the same set of first layer weights, of size at most $(2N-1)d_1$. Then, by Lemma B.18, we have that each component of

$$\mathbf{f}^2_{N_2,N_1}(\mathbf{x}) \doteq \mathbf{A}^{(3)}\mathbf{g}^2_{N_2}(\mathbf{A}^{(2)}\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x}))$$

is a Fourier network with the same set of first layer weights of size at most

$$(2(2N_1 - 1)d_1)^{N_2} .$$

Finally, consider the induction step. By the assumption hypothesis, the function

$$\mathbf{f}^{L-1}_{N_{L-1},\dots,N_1}(\mathbf{x}) \doteq \mathbf{A}^{(L)}\mathbf{g}^{L-1}_{N_{L-1}}(\mathbf{A}^{(L-1)} \cdots \mathbf{g}^{2}_{N_2}(\mathbf{A}^{(2)}\mathbf{q}_{N_1}(\mathbf{A}^{(1)}\mathbf{x})))$$

is such that each component is a Fourier network with the same set of first layer weights of size at most

$$2^{L-2}(2N_1 - 1)d_1 \ ^{\prod_{k=2}^{L-1} N_k} .$$

Then, by Lemma B.18, the function

$$f_N(\mathbf{x}) = \ \mathbf{a}^{(L+1)\ T}\mathbf{g}^{L}_{N_L}(\mathbf{f}^{L-1}_{N_{L-1},\dots,N_1}(\mathbf{x}))$$

is a Fourier network with at most

$$2 \cdot \ 2^{L-2}(2N_1 - 1)d_1 \ ^{\prod_{k=2}^{L-1} N_k \quad N_L} = 2^{N_L} 2^{(L-2)\prod_{k=2}^{L-1} N_k}((2N_1 - 1)d_1)^{\prod_{k=2}^{L} N_k}$$

which implies equation (B.34). Plugging in the lower bounds on N_k in terms of ϵ, the thesis follows. □

B.3.2 Fixed-dimension approximation

The results of Section 3.3 on fixed-threshold approximation can be complemented by the following result on fixed-dimension approximation. The proposition below is a straight-forward generalization of Theorem 3 in [SES19].

Proposition B.20. *Let σ be an activation satisfying Assumption 1. Then there exists a constant $\beta > 0$ such that for any $f : B^d_{1,2} \to \mathbb{C}$ 1-Lipschitz function and $\epsilon > 0$ there exists a network $f_N \in \mathcal{F}^\sigma_N$ such that*

$$\|f - f_N\|_{B^d_{1,\infty},\infty} \leq \epsilon$$

for some $N \leq 2 + \beta d^7 (\beta \epsilon^{-1})^d \epsilon^{-6}$.

Proof. The result is proved by noticing that the proof of Theorem 3 in [SES19] actually holds for any function f as in the statement. Moreover, using Assumption 1, f_N can also be chosen so that an equivalent bound holds for $m_\infty(f_N)$. □

B.4 Proofs related to spherical harmonics analysis of shallow networks

B.4.1 Low-coherence zonal harmonics frames

In this section, we wish to quantify how much incoherent can a frame composed of zonal harmonics be. More specifically, we wish to find a lower bound for

$$N(d,k,\epsilon) = \sup \quad N \geq 1 : \exists\, \mathbf{w}_1, \ldots, \mathbf{w}_N \in \mathbb{S}^{d-1} : \sup_{i \neq j} P^d_k \ \mathbf{w}_i^T \mathbf{w}_j \quad \leq \epsilon$$

for $\epsilon \in (0,1)$.

Lemma B.21. *It holds that*

$$N(d,k,\epsilon) \geq \sup \quad N \geq 1 : \exists\, \mathbf{w}_1, \ldots, \mathbf{w}_N \in \mathbb{S}^{d-1} : \sup_{i \neq j} \mathbf{w}_i^T \mathbf{w}_j \ \leq \quad \overline{1 - \frac{d}{k\epsilon^{4/d}}}$$

for $k > d \geq 5$ and $\left(\frac{d}{k}\right)^{d/4} \leq \epsilon < 1$.

Proof. We recall that it holds

$$P_k^d(t) \leq \frac{1}{\sqrt{\pi}}\Gamma\left(\frac{d-1}{2}\right)\left(\frac{4}{k(1-t^2)}\right)^{(d-2)/2}$$

for $d \geq 2$ and $t \in (-1,1)$ (cfr. eq. (2.117) in [AH12]) and that

$$\Gamma(x) \leq \left(\frac{x}{2}\right)^{x-1}$$

for $x \geq 2$. Therefore it holds that

$$\begin{aligned} P_k^d(t) &\leq \frac{1}{\sqrt{\pi}}\left(\frac{d-1}{4}\right)^{(d-3)/2}\left(\frac{4}{k(1-t^2)}\right)^{(d-2)/2} \\ &\leq \frac{1}{\sqrt{\pi}}\left(\frac{d}{4}\right)^{-1/2}\left(\frac{d}{k(1-t^2)}\right)^{(d-2)/2} \leq \left(\frac{d}{k(1-t^2)}\right)^{(d-2)/2} \end{aligned}$$

for $d \geq 5$ and $|t| < 1$. In particular, for $\epsilon \in (0,1)$, it holds that $P_k^d(t) \leq \epsilon$ if

$$\frac{d}{k(1-t^2)} \leq \epsilon^{4/d}$$

that is if

$$|t| \leq \sqrt{1 - \frac{d}{k\epsilon^{4/d}}}\,.$$

The thesis follows. □

Define

$$N(d,\delta) = \sup\left\{N \geq 1 : \exists\, \mathbf{w}_1,\ldots,\mathbf{w}_N \in \mathbb{S}^{d-1} : \sup_{i\neq j} \mathbf{w}_i^T\mathbf{w}_j \leq \delta\right\}$$

for $\delta \in (0,1)$. The previous lemma says that

$$N(d,k,\epsilon) \geq N \quad d, \quad \overline{1 - \frac{d}{k\epsilon^{4/d}}} \quad .$$

Example 9. Taking

$$\{\mathbf{w}_i\}_{i=1}^{N} = \quad \epsilon \in \quad \pm\frac{1}{\sqrt{d}} \quad ^{d} : \epsilon_1 > 0 \tag{B.35}$$

it holds that $N = 2^{d-1}$ and

$$\max_{i \neq j} \mathbf{w}_i^T \mathbf{w}_j \quad = 1 - \frac{2}{d} \, .$$

Therefore

$$N \quad d, 1 - \frac{2}{d} \quad \geq 2^{d-1} \, .$$

Taking $\epsilon = 2^{-d}$, it holds that, if $k \geq 8d^2$, then

$$N \;\; d, k, 2^{-d} \quad \geq 2^{d-1} \, .$$

Using this fact it is possible to explicitly construct a high energy sparse function.

Lemma B.22. *Take $k \geq 16d^2$ even and let*

$$\hat{P}(\mathbf{x}) = \beta_d \sum_{i=1}^{2^{d-1}} (N_k^d)^{1/2} P_k^d(\mathbf{w}_i^T \mathbf{x})$$

with $\beta_d = 2(2^d + 2)^{-1/2}$ and $\mathbf{w}_i$ as in equation (B.35). Then $\|\hat{P}\|_2 = \Theta_d(1)$ and it is exponentially spread, that is $\ell_{\infty,2}(\hat{P}) \leq O_d(2^{-d/2}) \quad \overline{N_k^d}$.

Proof. It holds that

$$\begin{aligned}\|\hat{P}\|_2^2 &= \beta_d^2 \Bigg[2^{d-1} + \sum_{i \neq j} P_k^d \ \mathbf{w}_i^T \mathbf{w}_j \\ &\leq \frac{2}{2^{d-1}+1} \ 2^{d-1} + \ 2^{2d-2} - 2^{d-1} \ 2^{-d} \\ &= \frac{2}{2^{d-1}+1} \ 2^{d-1} + 2^{d-2} - 2^{-1} \ \leq 3\end{aligned}$$

and that

$$\begin{aligned}\|\hat{P}\|_2^2 &\geq \frac{2}{2^{d-1}+1} \ 2^{d-1} - \ 2^{2d-2} - 2^{d-1} \ 2^{-d} \\ &= \frac{2}{2^{d-1}+1} \ 2^{d-1} - 2^{d-2} + 2^{-1} \ \geq 1 \, .\end{aligned}$$

On the other hand, it holds that

$$\|\hat{P}\|_\infty \leq \beta_d (N_k^d)^{1/2} \sup_{x \in \mathbb{S}^{d-1}} \sum_{i=1}^{2^{d-1}} P_k^d(\mathbf{w}_i^T \mathbf{x}) \ \ .$$

By definition of the vectors $\{\mathbf{w}_i\}_{i=1}^{2^{d-1}}$, it holds

$$\begin{aligned}\sup_{x \in \mathbb{S}^{d-1}} \sum_{i=1}^{2^{d-1}} P_k^d(\mathbf{w}_i^T \mathbf{x}) \ &= \frac{1}{2} \sup_{x \in \mathbb{S}^{d-1}, x>0} \sum_{\boldsymbol{\epsilon} \in \{\pm d^{-1/2}\}^d} P_k^d(\mathbf{x}^T \boldsymbol{\epsilon}) \\ &\leq 1 + \frac{1}{2} \sup_{\mathbf{x} \in \mathbb{S}^{d-1}, \mathbf{x}>0} \sum_{\boldsymbol{\epsilon} \in \{\pm d^{-1/2}\}^d \, : \, |\mathbf{1}^T \boldsymbol{\epsilon}| < \sqrt{d}} \ \frac{1}{16d \ 1 - \left|\mathbf{x}^T \boldsymbol{\epsilon}\right|^2}^{(d-2)/2} \\ &\leq 1 + \frac{1}{2}(2^d - 2) \ \ \frac{1}{16d \ 1 - \frac{d-1}{d}}^{(d-2)/2} \leq 1 + \frac{2^{d-1}-1}{4^{d-2}} \leq 2 \, .\end{aligned}$$

This proves the claim. □

Appendix C

Appendix to chapter 4

In this appendix we make use of the following notation. For any random variables $\mathbf{X}$ and $\mathbf{Y}$ with values in $\mathbb{R}^d$ and $\mathbb{R}^m$ respectively, we denote $\mathbf{\Sigma_X} = \mathbb{E}\ \mathbf{X}\mathbf{X}^T$ and $\mathbf{\Sigma_{XY}} = \mathbb{E}\ \mathbf{X}\mathbf{Y}^T$. For every integer $d \geq 1$, we denote by $GL(d)$, $O(d)$ and $SO(d)$, respectively, the general linear group, the orthogonal group and the special orthogonal group of real $d \times d$ matrices. $\mathbf{I}$ denotes the identity matrix and $\mathbf{e}_1, \ldots, \mathbf{e}_n$ the standard basis in $\mathbb{R}^d$.

C.1 Proofs of result on intrinsic dimension

C.1.1 Proof of Lemma 4.2

If σ is a polynomial of any degree k, then it holds that $\dim^*(\sigma, d) < \infty$. Indeed, let $\sigma(z) = a_0 + a_1 z + \cdots + a_k z^k$, for some $a_i \in \mathbb{R}$. If $I = \{i \in [0, d] : a_i \neq 0\}$, then

$$\mathcal{F}^\sigma \subseteq \mathbb{R}_I[\mathbf{x}] \doteq \{\mathbf{x} \mapsto \sum_{k \in I} \sum_{|\boldsymbol{\beta}|=k} \alpha_{\boldsymbol{\beta}} \mathbf{x}^{\boldsymbol{\beta}} : \alpha_{\boldsymbol{\beta}} \in \mathbb{R}\} .$$

It follows that

$$\dim^*(\sigma, d) = \dim(\mathcal{F}^\sigma) \leq \dim(\mathbb{R}_I[\mathbf{x}]) = \sum_{i=0}^{k} \binom{d+i-1}{i} \mathbf{1}_{\{a_i \neq 0\}} = O(d^k) \, .$$

This proves one implication. We prove the other one by contradiction. Assume now that σ is not a polynomial and that $\dim(\mathcal{F}^\sigma) = q < \infty$. Thanks to Theorem 1.1, for every continuous function $g : \mathbb{R}^d \to \mathbb{R}$, any compact set $K \subset \mathbb{R}^d$, and any $\varepsilon > 0$ there exist $h \in \mathcal{F}^\sigma$ such that

$$\sup_{\mathbf{x} \in K} |h(\mathbf{x}) - g(\mathbf{x})| < \varepsilon \, . \tag{C.1}$$

Now, let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuous function supported on a compact set $C \subset \mathbb{R}^d$. We call $C_c(\mathbb{R}^d)$ the set of the real-valued continuous functions from $\mathbb{R}^d$ with compact support. Thanks to (C.1), we can find a sequence of compact sets $\{K_m\}_{m \geq 1}$ of $\mathbb{R}^d$ such that

$$C \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m \subseteq \cdots \subseteq \cup_{m=1}^{\infty} K_m = \mathbb{R}^d$$

and a sequence of functions $\{h_m\}_{m \geq 1} \subset \mathcal{F}^\sigma$ such that

$$\|g - h_m \mathbb{1}_{K_m}\|_{L^2(\mathbf{X})} = \|(g - h_m)\mathbb{1}_{K_m}\|_{L^2(\mathbf{X})} < 2^{-m}$$

In particular this implies that

$$\|h_n \mathbb{1}_{K_n} - h_m \mathbb{1}_{K_m}\|_{L^2(\mathbf{X})} < 2^{1-\min\{n,m\}} \to 0$$

as $n, m \to \infty$, i.e. $\{h_m \mathbb{1}_{K_m}\}_{m \geq 1}$ is a Cauchy sequence in $L^2(\mathbf{X})$ and therefore it admits a limit $\lim_{m \to \infty} h_m \mathbb{1}_{K_m} = g \in L^2(\mathbf{X})$. Since $\dim(V_\sigma) = q < \infty$, there exists $\mathbf{w}_1, \ldots, \mathbf{w}_q \in \mathbb{R}^n$ such that every $h \in \mathcal{F}^\sigma$ can be written as

$$h(\mathbf{x}) = \mathbf{u}^T \boldsymbol{\gamma}(\mathbf{x})$$

for some $\mathbf{u} \in \mathbb{R}^q$, where $\boldsymbol{\gamma}(\mathbf{x}) = (\sigma(\mathbf{w}_1^T\mathbf{x}), \dots, \sigma(\mathbf{w}_q^T\mathbf{x}))$. Let $\{\mathbf{u}_m\}_{m\geq 1} \subset \mathbb{R}^q$ such that $h_m(\mathbf{x}) = \mathbf{u}_m^T\boldsymbol{\gamma}(\mathbf{x})$. Thanks to the above calculations, we know that the sequence $\{\|h_m \mathbb{1}_K\|_{L^2(\mathbf{X})}\}_{m\geq 1}$ is bounded for any arbitrary compact set $K \subseteq \mathbb{R}^d$. Since

$$\|h_m \mathbb{1}_K\|^2_{L^2(\mathbf{X})} = \mathbf{u}_m^T \mathbf{M} \mathbf{u}_m \ ,$$

where $\mathbf{M} = \mathbb{E}\ \boldsymbol{\gamma}(\mathbf{X})\boldsymbol{\gamma}(\mathbf{X})^T \mathbb{1}_{\{\mathbf{X}\in K\}} \ \in \mathbb{R}^{q\times q}$, this implies that the sequence $\{\mathbf{u}_m\}_{m\geq 1}$ is bounded (unless $g = 0$). Therefore (up to extracting a sub-sequence) we can assume that it has a limit $\mathbf{u} \in \mathbb{R}^q$. If we call $h \in \mathcal{F}^\sigma$ the function defined as $h(\mathbf{x}) = \mathbf{u}^T\boldsymbol{\gamma}(\mathbf{x})$, it is easy to check (from the above calculations) that $h = g$ in $L^2(\mathbf{X})$. This shows that $C_c(\mathbb{R}^n) \subseteq \mathcal{F}^\sigma$, which in turn implies that $\mathcal{F}^\sigma$ is dense in $L^2(\mathbf{X})$ (since $C_c(\mathbb{R}^n)$ is dense in $L^2(\mathbf{X})$). But this is impossible, since $\dim(\mathcal{F}^\sigma) = q < \infty = \dim(L^2(\mathbf{X}))$. Therefore, it must hold $\dim(\mathcal{F}^\sigma) = \infty$.

C.1.2 Proof of Lemma 4.3 and Lemma 4.4

Assume that $\sigma \in L^2_\varphi$ is a continuous activation (any polynomial σ satisfies this) and let $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$ be a standard Gaussian random variable. Then, we can write $\sigma(z) = \sum_{k=0}^{\infty} \hat{\sigma}_k h_k(z)$, where h_k is the degree-k Hermite polynomial. It follows that, for $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W})$,

$$\mathbb{E}|\Phi(\mathbf{X};\boldsymbol{\theta})|^2 = \sum_{k=1}^{\infty} \hat{\sigma}_k^2 \left\| \sum_{i=1}^{N} u_i \mathbf{w}_i^{\otimes k} \right\|_F^2 ,$$

as shown in Lemma 1 in [MM18]. It follows that

$$\min\ \ r \geq 0 : \Phi(\cdot;\boldsymbol{\theta}) \in \overline{\mathcal{F}_r^\sigma}^{L^2(\mathbf{X})} \ \ \geq \mathrm{rk}^*_{\mathrm{S}} \left\| \sum_{i=1}^{N} u_i \mathbf{w}_i^{\otimes k} \right\|$$

for all k such that $\hat{\sigma}_k \neq 0$. This implies that

$$\dim_*(\sigma, d) \geq \sup_{k \geq 0 \,:\, \hat{\sigma}_k \neq 0} \mathrm{rk}_{\mathrm{S}}^*(k, d)$$

which implies that

$$\dim_*(\sigma, d) \geq \mathrm{rk}_{\mathrm{S}}^*(k, d) \geq \frac{1}{2}\mathrm{rk}_{\mathrm{S}}(k, d)$$

for any σ polynomial of degree k, and that

$$\dim_*(\sigma, d) = \infty$$

for any $\sigma \in L^2_\varphi$ non-polynomial, thanks to Lemma C.10. For $\sigma(z) = z^k$, it is also easy to see that

$$\dim_*(\sigma, d) \leq \mathrm{rk}_{\mathrm{S}} \left(\sum_{k=1}^{N} u_i \mathbf{w}_i^{\otimes k} \right) \leq \mathrm{rk}_{\mathrm{S}}(k, d)$$

C.2 Proofs of results regarding absence of spurious valleys

C.2.1 Proof of Theorem 4.5

First, notice that, under the assumptions of Theorem 4.5, the same optimal neural networks $\Phi_i(\cdot; \boldsymbol{\theta})$ could also be obtained using a generalized linear model, where the representation function has the linear form $\Phi_i(\mathbf{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}_i, \boldsymbol{\varphi}(\mathbf{x}) \rangle$, for some parameter independent function $\boldsymbol{\varphi} : \mathbb{R}^n \to \mathbb{R}^{\dim^*(\sigma, \mathbf{X})}$. The main difference between the two models is that the former requires the choice of a non-linear activation function σ, while the latter implies the choice of a kernel functions. This is the content of the following lemma.

Lemma C.1. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function and let $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$. Let $\mathcal{F}^\sigma_{\mathbf{X}}$ denote the embedding of $\mathcal{F}^\sigma$ in $L^2_{\mathbf{X}}$, and assume that $\mathcal{F}^\sigma_{\mathbf{X}}$ is finite dimensional. Then there exists a scalar*

product $\langle \cdot, \cdot \rangle$ *on* $\mathcal{F}_{\mathbf{X}}^{\sigma}$ *and a map* $\mathbf{x} \in \mathbb{R}^n \mapsto \varphi(\mathbf{x}) \in \mathcal{F}_{\mathbf{X}}^{\sigma}$ *such that*

$$\langle \psi_{\sigma,\mathbf{w}}, \varphi(\mathbf{x}) \rangle = \psi_{\sigma,\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}) \tag{C.2}$$

for all $\mathbf{w} \in \mathbb{R}^n$. *Moreover, the function* $\mathbf{w} \in \mathbb{R}^n \mapsto \psi_{\sigma,\mathbf{w}} \in \mathcal{F}_{\mathbf{X}}^{\sigma}$ *is continuous.*

Proof. For sake of simplicity, in the following we write $\psi_{\mathbf{w}}$ for $\psi_{\sigma,\mathbf{w}}$ and $\mathcal{F}$ for $\mathcal{F}_{\mathbf{X}}^{\sigma}$. Let $\psi_{\mathbf{w}_1}, \dots, \psi_{\mathbf{w}_q}$ be a basis of $\mathcal{F}$. If $\psi_{\mathbf{w}} = \sum_{i=1}^{q} \alpha_i \psi_{\mathbf{w}_i}$ and $\psi_{\mathbf{v}} = \sum_{j=1}^{q} \beta_j \psi_{\mathbf{w}_j}$, then we can define a scalar product on $\mathcal{F}$ as

$$\langle \psi_{\mathbf{w}}, \psi_{\mathbf{v}} \rangle \doteq \sum_{i=1}^{q} \alpha_i \beta_i \ .$$

If we define the map $\mathbf{x} \in \mathbb{R}^n \mapsto \varphi(\mathbf{x}) \in \mathcal{F}$ as

$$\varphi(\mathbf{x}) = \sum_{i=1}^{q} \psi_{\mathbf{w}_i}(\mathbf{x}) \psi_{\mathbf{w}_i} \ ,$$

then property (C.2) follows directly by the definition of the function $\psi_{\mathbf{w}}$. Moreover, we can choose $\mathbf{x}_1, \dots, \mathbf{x}_q$ such that $\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_q)$ is a basis of $\mathcal{F}$. Now we need to show that, for $i \in [q]$, the map $\mathbf{w} \mapsto \langle \psi_{\mathbf{w}}, \psi_{\mathbf{w}_i} \rangle$ is continuous. Let $\mathbf{M}$ be the matrix $\mathbf{M} \doteq (\psi_{\mathbf{w}_j}(\mathbf{x}_i))_{ij} \in \mathbb{R}^{q \times q}$ and $\mathbf{z}(\mathbf{w})$ be the vector $\mathbf{z}(\mathbf{w}) \doteq (\psi_{\mathbf{w}}(\mathbf{x}_i))_i \in \mathbb{R}^q$. Then $\langle \psi_{\mathbf{w}}, \psi_{\mathbf{w}_i} \rangle = (\mathbf{M}^{-1} \mathbf{z}(\mathbf{w}))_i$, which is continuous in $\mathbf{w}$. This shows that the map $\mathbf{w} \in \mathbb{R}^n \mapsto \psi_{\mathbf{w}} \in \mathcal{F}$ is continuous. □

The non-trivial fact captured by Theorem 4.5 is the following: when the capacity of network is large enough to match a generalized linear model, but still finite, then the problem of optimizing the loss function (4.1), which is in general a highly non-convex object, satisfies an interesting optimization property in view of the local descent algorithms which are used to solve it in practice.

Proof of Theorem 4.5. Thanks to Lemma C.1, there exist two continuous maps $\boldsymbol{\varphi}, \boldsymbol{\psi} : \mathbb{R}^n \to \mathbb{R}^q \simeq \mathcal{F}_{\mathbf{X}}^{\sigma}$, with $q = \dim^*(\sigma, \mathbf{X})$, such that $\sigma(\mathbf{w}^T \mathbf{x}) = \langle \boldsymbol{\psi}(\mathbf{w}), \boldsymbol{\varphi}(\mathbf{x}) \rangle$ for every $\mathbf{w}, \mathbf{x} \in \mathbb{R}^n$. Therefore, every one-hidden-layer neural network $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U} \sigma(\mathbf{W}^T \mathbf{x})$ can be written as $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) =$

$\mathbf{U}[\boldsymbol{\psi}(\mathbf{W})]^T\boldsymbol{\varphi}(\mathbf{x})$, where, if $\mathbf{W} \in \mathbb{R}^{d\times N}$, then $\boldsymbol{\psi}(\mathbf{W}) \in \mathbb{R}^{q\times N}$ (that is $\boldsymbol{\psi}$ is applied row-wise).

The proof of the theorem consists in exploiting the above 'linearized' representation of $\boldsymbol{\Phi}$ to show that property **P.1** holds (remind that this is equivalent to saying that the loss function has no spurious valleys). Given an initial parameter $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{U}}, \tilde{\mathbf{W}})$, we want to construct a continuous path $t \in [0,1] \mapsto \boldsymbol{\theta}_t = (\mathbf{U}_t, \mathbf{W}_t)$, such that the function $t \in [0,1] \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing and such that $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}$, $\boldsymbol{\theta}_1 \in \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, where $L(\boldsymbol{\theta}) = \mathbb{E}[\ell(\boldsymbol{\Phi}(\mathbf{X};\boldsymbol{\theta}), \mathbf{Y})]$. The construction of such a path can be articulated in two main steps.

Step 1. The first part of the path consist showing that we can assume that $\mathrm{rk}(\boldsymbol{\psi}(\tilde{\mathbf{W}})) = q$ without loss of generality. Let $\mathbf{w}_1, \dots, \mathbf{w}_N \in \mathbb{R}^d$ be the columns of $\tilde{\mathbf{W}}$; suppose that $\mathrm{rk}(\boldsymbol{\psi}(\tilde{\mathbf{W}})) = r < q$ (otherwise there is nothing to show) and that $\boldsymbol{\psi}(\mathbf{w}_{i_1}), \dots, \boldsymbol{\psi}(\mathbf{w}_{i_r})$ are linearly independent. Denote $I = \{i_1, \dots, i_r\}$, $J = [N] \setminus I = \{j_1, \dots, j_{N-r}\}$ and $\mathbf{u}_1, \dots, \mathbf{u}_N$ the columns of $\tilde{\mathbf{U}}$. For $j \in J$, we can write

$$\boldsymbol{\psi}(\mathbf{w}_j) = \sum_{k=1}^{r} a_j^k \, \boldsymbol{\psi}(\mathbf{w}_{i_k}) \quad \text{for some } a_j^k \in \mathbb{R}\,.$$

If we define $\mathbf{U}_1$ such that (denoting $\mathbf{u}_{1,i}$ the i-th column of $\mathbf{U}_1$)

$$\mathbf{u}_{1,i} = \mathbf{u}_i + \sum_{k=1}^{N-r} a_k^i \, \mathbf{u}_{j_k} \quad \text{for } i \in I, \quad \mathbf{u}_{1,j} = 0 \quad \text{for } j \in J$$

then $\mathbf{U}_1\tilde{\mathbf{W}} = \tilde{\mathbf{U}}\tilde{\mathbf{W}}$. The path $t \in [0, 1/2] \mapsto \boldsymbol{\theta}_t = (2t\,\mathbf{U}_1 + (1-2t)\tilde{\mathbf{U}}, \tilde{\mathbf{W}})$ leaves the network unchanged, i.e. $\boldsymbol{\Phi}(\cdot;\tilde{\boldsymbol{\theta}}) = \boldsymbol{\Phi}(\cdot;\boldsymbol{\theta}_t)$ for $t \in [0, 1/2]$. At this point, we can select $\mathbf{w}_{1,j_1}, \dots, \mathbf{w}_{1,j_{N-r}} \in \mathbb{R}^n$ such that the matrix $\mathbf{W}_1$ with columns $\mathbf{w}_{1,i} = \mathbf{w}_i$ for $i \in I$ and $\mathbf{w}_{1,j}$ for $j \in J$, verifies $\mathrm{rk}(\boldsymbol{\psi}(\mathbf{W}_1)) = q$. Notice that the existence of such vectors $\mathbf{w}_{1,j_k}$, $k \in [p-r]$, is guaranteed by the definition of $q = \dim^*(\sigma, \mathbf{X})$. The path $t \in [1/2, 1] \mapsto \boldsymbol{\theta}_t = (\mathbf{U}_1, (2t-1)\mathbf{W}_1 + (2-2t)\tilde{\mathbf{W}})$ leaves the network unchanged, i.e. $\boldsymbol{\Phi}(\cdot;\boldsymbol{\theta}_0) = \boldsymbol{\Phi}(\cdot;\boldsymbol{\theta}_t)$ for $t \in [0, 1]$. The new parameter value $\boldsymbol{\theta}_1 = (\mathbf{U}_1, \mathbf{W}_1)$ satisfies $\mathrm{rk}(\boldsymbol{\psi}(\mathbf{W}_1)) = q$.

Step 2. By step 1, we can assume that $\mathrm{rk}(\tilde{\mathbf{W}}) = q$. Since the network has the form $\boldsymbol{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{U}[\boldsymbol{\psi}(\mathbf{W})]^T\boldsymbol{\varphi}(\mathbf{x})$ and since the function ℓ is convex, there exists $\mathbf{U}^* \in \mathbb{R}^{m\times p}$ such that $\boldsymbol{\theta} = (\mathbf{U}^*, \tilde{\mathbf{W}}) \in \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. The proof is therefore concluded by selecting the path $t \in [0,1] \mapsto \boldsymbol{\theta}_t = (t\mathbf{U}^* + (1-t)\tilde{\mathbf{U}}, \tilde{\mathbf{W}})$.

This shows that property **P.1** holds and therefore it proves the theorem. □

C.2.2 Proof of Theorem 4.8

The first step for proving Theorem 4.8 consists in extending the result of Theorem 4.5 to the case of one-hidden-layer linear neural networks $\boldsymbol{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{U}\mathbf{W}^T\mathbf{x}$ with $\mathbf{U} \in \mathbb{R}^{m\times N}$, $\mathbf{W} \in \mathbb{R}^{d\times N}$ with $N < d$ and square loss functions $L(\boldsymbol{\theta}) = \mathbb{E}\|\boldsymbol{\Phi}(\mathbf{X};\boldsymbol{\theta}) - \mathbf{Y}\|^2$. We start by pointing out a symmetry property of this type of networks: for every $\mathbf{G} \in GL(N)$ it holds that

$$\boldsymbol{\Phi}(\mathbf{x};(\mathbf{U},\mathbf{W})) = \mathbf{U}\mathbf{W}^T\mathbf{x} = (\mathbf{U}\mathbf{G}^{-1})(\mathbf{W}\mathbf{G}^T)^T\mathbf{x} = \boldsymbol{\Phi}(\mathbf{x};(\mathbf{U}\mathbf{G}^{-1},\mathbf{W}\mathbf{G}^T)) .$$

This means that the map $\boldsymbol{\theta} \mapsto \boldsymbol{\Phi}(\cdot;\boldsymbol{\theta})$ is defined up to an action of the group $GL(N)$ over the parameter space $\Theta = \mathbb{R}^{m\times N} \times \mathbb{R}^{N\times d}$; the same remark holds for the loss function $L(\boldsymbol{\theta})$. We can therefore think about the loss function as defined over the topological quotient $\Theta/GL(N)$. We denote the orbit of an element $\boldsymbol{\theta} = (\mathbf{U},\mathbf{W}) \in \Theta$ as

$$[\boldsymbol{\theta}] = [\mathbf{U},\mathbf{W}] = \{\mathbf{G}\cdot\boldsymbol{\theta} = (\mathbf{U}\mathbf{G}^{-1},\mathbf{W}\mathbf{G}^T) : \mathbf{G} \in GL(N)\} .$$

If g is a real-valued function defined on Θ such that $g(\mathbf{G}\cdot\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ for all $\mathbf{G} \in GL(N)$ and $\boldsymbol{\theta} \in \Theta$, then one can equivalently consider g as defined on $\Theta/GL(N)$ as $g([\boldsymbol{\theta}]) = g(\boldsymbol{\theta})$; for simplicity we denote $g[\boldsymbol{\theta}] = g([\boldsymbol{\theta}])$. This is exactly the case for the loss function $L(\boldsymbol{\theta})$. In the proof of Theorem 4.5, we describe how to construct a path from an initial parameter value $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{U}},\tilde{\mathbf{W}})$ to a parameter

value $\boldsymbol{\theta}_1 = (\mathbf{Q}(\mathbf{W}_1), \mathbf{W}_1)$, with $\mathrm{rk}(\mathbf{W}_1) = N$ and $\mathbf{Q} : \mathbb{R}^{d\times N} \to \mathbb{R}^{m\times N}$ the function defined by

$$\mathbf{Q}(\mathbf{W}) = \boldsymbol{\Sigma}_{\mathbf{YX}}\mathbf{W}(\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{W})^\dagger \in \arg\min_{\mathbf{U}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=(\mathbf{U},\mathbf{W})}$$

(see Lemma C.11). Therefore, let $\tilde{\boldsymbol{\theta}} = (\mathbf{Q}(\tilde{\mathbf{W}}), \tilde{\mathbf{W}})$ with $\mathrm{rk}(\tilde{\mathbf{W}}) = N$, be an initial parameter. Since an optimal parameter is given by $\boldsymbol{\theta} = (\mathbf{Q}(\mathbf{W}), \mathbf{W})$ for some $\mathbf{W}$, we seek for a path in the form $\boldsymbol{\theta}_t = (\mathbf{Q}(\mathbf{W}_t), \mathbf{W}_t)$ with $\mathrm{rk}(\mathbf{W}_t) = N$ for all $t \in [0, 1]$. This path must be such that $t \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing. If we assume that $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}$, it holds

$$L(\boldsymbol{\theta}_t) = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{Y}}) - \mathrm{tr}(\mathbf{M}\mathbf{P}_{\mathbf{W}_t})$$

where $\mathbf{M}$ is a positive semi-definite matrix and, for every matrix $\mathbf{W}$, $\mathbf{P}_{\mathbf{W}}$ denotes the orthogonal projection on space spanned by the columns of $\mathbf{W}$, that is $\mathbf{P}_{\mathbf{W}} = (\mathbf{W}\mathbf{W}^\dagger)^T$ (see Lemma C.11). Therefore it is equivalent for the path $\boldsymbol{\theta}_t = (\mathbf{q}(\mathbf{W}_t), \mathbf{W}_t)$ to be such that the function

$$t \in [0, 1] \mapsto f(\mathbf{W}_t) \doteq \mathrm{tr}(\mathbf{M}\mathbf{P}_{\mathbf{W}_t})$$

is non-decreasing. In particular, the function f is defined up to the action of the group $GL(N)$ on Θ. Since we look for $\mathbf{W}_t$ of rank N, we can consider f as defined on $G(N, d)$, the Grassmanian of N dimensional linear subspaces of $\mathbb{R}^d$. The proof below for the linear one-hidden-layer case is articulated as follows. We first construct a path $[\mathbf{W}_t] \in G(N, d)$ such that $[\mathbf{W}_0] = [\tilde{\mathbf{W}}]$, $[\mathbf{W}_1]$ maximizes f and such that the function $t \in [0, 1] \mapsto f[\mathbf{W}_t]$ is non-decreasing (Lemma C.2). We then show that such a path can be lifted to a corresponding path $\mathbf{W}_t \in \mathbb{R}^{N\times d}$ (Lemma C.3). Finally, we show that we can drop the assumption $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}$ and the result still holds (Lemma C.4).

Lemma C.2. *Let $[\tilde{\mathbf{W}}] \in G(N, d)$ and assume $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}$. Then there exists a continuous path $t \in [0, 1] \mapsto [\mathbf{W}_t] \in G(N, d)$ such that $[\mathbf{W}_0] = [\tilde{\mathbf{W}}]$, $[\mathbf{W}_1]$ maximizes f and such that the function $t \in [0, 1] \mapsto f[\mathbf{W}_t]$ is non-decreasing.*

Proof. While it is geometrically intuitive that the results should hold, we derive a constructive proof. We start by noticing that if $[\mathbf{W}] \in G(d, N)$ and $\mathbf{w}_1, \ldots, \mathbf{w}_N$ is an orthonormal basis of $[\mathbf{W}]$, then

$$f[\mathbf{W}] = \sum_{i=1}^{N} \mathbf{w}_i^T \mathbf{M} \mathbf{w}_i \ . \tag{C.3}$$

Moreover, if $\mathbf{M} = \sum_{j=1}^{d} \sigma_i \mathbf{v}_j \mathbf{v}_j^T$ is the SVD of $\mathbf{M}$, where $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$, then (C.3) can be written as

$$f[\mathbf{W}] = \sum_{j=1}^{d} \sigma_j \sum_{i=1}^{N} \langle \mathbf{v}_j, \mathbf{w}_i \rangle^2 \ .$$

In particular the maximum of f is obtained for $[\mathbf{W}] = [\mathbf{V}] \doteq [\mathbf{v}_1, \ldots, \mathbf{v}_N]$ (with some abuse of notation, we identify a subspace with one of its basis). To prove the result is therefore sufficient to show a path $[\mathbf{W}_t]$ from any $[\mathbf{W}_0] = [\tilde{\mathbf{W}}]$ to $[\mathbf{W}_1] = [\mathbf{V}]$, such that the function $t \in [0, 1] \mapsto f[\mathbf{W}_t]$ is non-decreasing. To do this we construct a finite sequence of paths

$$[\mathbf{W}_t^i] \quad \text{such that} \quad [\mathbf{W}_0^i] = [\mathbf{W}^{i-1}] \quad \text{and} \quad [\mathbf{W}_1^i] = [\mathbf{W}^i]$$

for $i \in [N]$, with $[\mathbf{W}^0] = [\tilde{\mathbf{W}}]$, $[\mathbf{W}^N] = [\mathbf{V}]$ and

$$\mathbf{W}^i = [\mathbf{v}_1, \ldots, \mathbf{v}_i, \mathbf{w}_{i+1}^{i-1}, \ldots, \mathbf{w}_N^{i-1}] \quad \text{for } i \in [N] \ ,$$

where $\mathbf{w}_1^j = \mathbf{v}_1, \ldots, \mathbf{w}_j^j = \mathbf{v}_j, \mathbf{w}_{j+1}^j, \ldots, \mathbf{w}_N^j$ is an orthonormal basis of $[\mathbf{W}^j]$, for $j \in [0, N]$. Moreover, the paths $[\mathbf{W}_t^i]$ are such that the functions $t \in [0, 1] \mapsto f[\mathbf{W}_t^i]$ are non-decreasing. Such paths are defined as follows. Let $i \in [0, N-1]$ and consider

$$[\mathbf{W}^i] = [\mathbf{w}_1^i = \mathbf{v}_1, \ldots, \mathbf{w}_i^i = \mathbf{v}_i, \mathbf{w}_{i+1}^i, \ldots, \mathbf{w}_N^i] \ .$$

We define

$$\mathbf{u}_{i+1}^i = \begin{cases} \frac{\mathbf{P}_{\mathbf{W}^i}\mathbf{v}_{i+1}}{\|\mathbf{P}_{\mathbf{W}^i}\mathbf{v}_{i+1}\|} & \text{if } \mathbf{P}_{\mathbf{W}^i}\mathbf{v}_{i+1} \neq \mathbf{0} \\ \mathbf{w}_{i+1}^i & \text{o.w.} \end{cases} .$$

Then we complete $\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{u}_{i+1}^i$ to an orthonormal basis of $[\mathbf{W}^i]$:

$$\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{u}_{i+1}^i, \dots, \mathbf{u}_N^i .$$

We call $\mathbf{w}_j^{i+1} = \mathbf{u}_j^i$ for $j \in [i+2, N]$ and we define

$$[\mathbf{W}^{i+1}] = [\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{w}_{i+1}^{i+1} = \mathbf{v}_{i+1}, \mathbf{w}_{i+2}^{i+1}, \dots, \mathbf{w}_N^{i+1}]$$

The path $[\mathbf{W}_t^i]$ is then obtained by moving $\mathbf{u}_{i+1}^i$ to $\mathbf{v}_{i+1}$ on a geodesic on the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, i.e.

$$[\mathbf{W}_t^{i+1}] = [\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{u}_{i+1}^i(t), \mathbf{u}_{i+2}^i, \dots, \mathbf{u}_N^i] .$$

where we defined

$$\mathbf{u}_{i+1}^i(t) \;=\; (1-(1-\mu_{i+1})t)\mathbf{u}_{i+1}^i \;+\; \sqrt{1-(1-(1-\mu_{i+1})t)^2} \cdot \frac{\mathbf{v}_{i+1} - \mu_{i+1}\mathbf{u}_{i+1}^i}{\sqrt{1-\mu_{i+1}^2}}$$

for $\mu_{i+1} = [\mathbf{u}_{i+1}^i]^T \mathbf{v}_{i+1}$. The fact that the function $t \in [0,1] \mapsto f[\mathbf{W}_t^{i+1}]$ is non-decreasing can be proved by noticing that

$$f[\mathbf{W}_t^{i+1}] - f[\mathbf{W}^i] = \sum_{j=i+1}^{d} \sigma_j \langle \mathbf{u}_{i+1}^i(t), \mathbf{v}_j \rangle^2$$

and by showing that the derivative of the RHS is greater or equal than 0. This concludes the proof of the lemma. □

Lemma C.3. *Let $\tilde{\mathbf{W}} \in \mathbb{R}^{d \times N}$ and assume $\mathbf{\Sigma}_X = \mathbf{I}$. Then there exists a continuous path $t \in$*

$[0,1] \mapsto \mathbf{W}_t \in \mathbb{R}^{d\times N}$ *such that* $\mathbf{W}_0 = \tilde{\mathbf{W}}$, $\mathbf{W}_1$ *maximizes* f *and such that the function* $t \in [0,1] \mapsto f(\mathbf{W}_t)$ *is non-decreasing.*

Proof. The only thing we need to prove in this case is that we can lift the paths $[\mathbf{W}_t^i] \in G(N,d)$ from the proof of Lemma C.2 to continuous paths $\mathbf{W}_t^i \in \mathbb{R}^{d\times N}$. We first notice that if the basis $\{\mathbf{w}_1^i, \dots, \mathbf{w}_N^i\}$ and $\{\mathbf{w}_1^i, \dots, \mathbf{w}_i^i, \mathbf{u}_{i+1}^i, \dots, \mathbf{u}_N^i\}$ are defined as above, then we can assume (up to changing some signs) that they have all the same orientation, for all $i \in [0,N]$. Therefore we can define the matrices $\mathbf{W}^i \in \mathbb{R}^{d\times N}$ with columns $\mathbf{w}_1^i, \dots, \mathbf{w}_N^i$ and the matrices $\mathbf{U}^i \in \mathbb{R}^{d\times N}$ with colmuns $\mathbf{w}_1^i, \dots, \mathbf{w}_i^i, \mathbf{u}_{i+1}^i, \dots, \mathbf{u}_N^i$, for $i \in [0,N]$. The paths $\mathbf{W}_t^{i+1}$ are defined in the same way as in the proof of Lemma C.2. Notice that such paths go from $\mathbf{W}_0^{i+1} = \mathbf{U}^i$ to $\mathbf{W}_1^{i+1} = \mathbf{W}^{i+1}$. It remains to construct paths from $\mathbf{W}^i$ to $\mathbf{U}^i$. Consider the matrix

$$\mathbf{O}^i = \mathbf{U}^i[\mathbf{W}^i]^T \in SO(d) \, .$$

Notice that $\mathbf{O}^i\mathbf{W}^i = \mathbf{U}^i$. In particular there exist $\mathbf{A}^i$ real skew-symmetric such that $\mathbf{O}^i = e^{\mathbf{A}^i}$. Therefore the paths $t \in [0,1] \mapsto \mathbf{U}_t^i = e^{t\mathbf{A}^i}\mathbf{W}^i$ go from $\mathbf{U}_0^i = \mathbf{W}^i$ to $\mathbf{U}_1^i = \mathbf{U}^i$. Moreover $f(\mathbf{U}_t^i)$ is constant in t (since the underlying linear subspace does not change). The only thing that remains to prove is that, given the matrix $\tilde{\mathbf{W}} \in \mathbb{R}^{d\times N}$ with columns $\mathbf{w}_1, \dots, \mathbf{w}_N$, there is a path from $\tilde{\mathbf{W}}$ to $\mathbf{W}^0$. Now, $\mathbf{W}^0$ was chosen as a matrix with orthonormal columns such that $[\tilde{\mathbf{W}}] = [\mathbf{W}^0]$. Therefore if $\tilde{\mathbf{W}} = \mathbf{U}\mathbf{\Lambda}\mathbf{O}$ is the SVD of $\tilde{\mathbf{W}}$ with $\mathbf{U} = \mathbf{W}^0$, $\mathbf{\Lambda} = \mathrm{diag}(\sigma_1, \dots, \sigma_N) \in \mathbb{R}^{N\times N}$ (with $\sigma_i > 0$, $i \in [N]$) and $\mathbf{O} \in SO(N)$, there exists $\mathbf{A}$ real skew-symmetric such that $\mathbf{O} = e^{\mathbf{A}}$. Thus the path $t \in [0,1] \mapsto \mathbf{W}_t = \mathbf{W}^0\mathbf{\Lambda}^{1-t}e^{(1-t)\mathbf{A}}$ is a path between $\mathbf{W}_0 = \tilde{\mathbf{W}}$ and $\mathbf{W}_1 = \mathbf{W}^0$. This concludes the proof of the lemma. □

Lemma C.4. *Lemma C.3 holds even if we drop the assumption* $\mathbf{\Sigma}_\mathbf{X} = \mathbf{I}$.

Proof. For sake of simplicity we distinguish two cases.

Case 1: $\mathrm{rk}(\mathbf{\Sigma}_\mathbf{X}) = d$. Let $\mathbf{K} = (\mathbf{\Sigma}_\mathbf{X})^{1/2}$. Then $\tilde{\mathbf{X}} = \mathbf{K}^{-1}\mathbf{X}$ is such that $\mathbf{\Sigma}_{\tilde{\mathbf{X}}} = \mathbf{I}$. Therefore, if

$t \in [0,1] \mapsto \boldsymbol{\theta}_t = (\mathbf{U}_t, \mathbf{W}_t)$ is the path given by Lemma C.3 for the case $\mathbf{X} = \tilde{\mathbf{X}}$, the sought path (for $\mathbf{X} = \mathbf{X}$) is given by $t \in [0,1] \mapsto (\mathbf{U}_t, \mathbf{K}^{-T}\mathbf{W}_t)$.

Case 2: $\mathrm{rk}(\boldsymbol{\Sigma}_{\mathbf{X}}) < n$. In this case, if $r = \mathrm{rk}(\boldsymbol{\Sigma}_{\mathbf{X}})$, $\mathbf{X}$ belongs to a r-dimensional subspace of $\mathbb{R}^d$ (a.s.), call it V. If $\mathbf{O} \in \mathbb{R}^{d \times r}$ is a matrix with an orthonormal basis of V as columns, then $\mathbf{O}\mathbf{O}^T\mathbf{X} = \mathbf{X}$ (a.s.), and, if $\tilde{\mathbf{X}} = \mathbf{O}^T\mathbf{X}$ then $\tilde{\mathbf{X}} \in \mathbb{R}^r$ and $\mathrm{rk}(\boldsymbol{\Sigma}_{\tilde{\mathbf{X}}}) = r$. Therefore, if $t \in [0,1] \mapsto \boldsymbol{\theta}_t = (\mathbf{U}_t, \mathbf{W}_t)$ is the path given by case 1 for $\mathbf{X} = \tilde{\mathbf{X}}$, the sought path (for $\mathbf{X} = \mathbf{X}$) is given by $t \in [0,1] \mapsto (\mathbf{U}_t, \mathbf{O}\mathbf{W}_t\mathbf{O})$. □

This concludes the proof of non-existence of spurious valleys for the square loss function of linear one-hidden-layer neural networks $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\mathbf{W}^T\mathbf{x}$. The fact that such proof does not require any assumptions on the dimensions of the layers d, N, m neither on the rank of the initial layers, allows us to prove non-existence of spurious valleys for the square loss function of linear neural networks of any depth $L \geq 1$:

$$\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{W}_{L+1}^T \cdots \mathbf{W}_1^T \mathbf{x} \tag{C.4}$$

We start by proving a simple lemma.

Lemma C.5. *Let* $\tilde{\mathbf{U}} = \tilde{\mathbf{M}}^1 \cdots \tilde{\mathbf{M}}^n$, *where* $\tilde{\mathbf{U}} \in \mathbb{R}^{r_0 \times r_n}$ *and* $\tilde{\mathbf{M}}^i \in \mathbb{R}^{r_{i-1} \times r_i}$. *Suppose that* $t \in [0,1] \mapsto \mathbf{U}_t$ *is a given continuous path between* $\mathbf{U}_0 = \tilde{\mathbf{U}}$ *and another matrix* $\mathbf{U}_1 \in \mathbb{R}^{r_0 \times r_n}$. *If* $r_i \geq \min\{r_0, r_n\}$ *for all* i, *then there exist continuous paths* $\mathbf{M}_t^i$ *such that* $\mathbf{M}_0^i = \tilde{\mathbf{M}}^i$ *and such that* $\mathbf{U}_t = \mathbf{M}_t^1 \dots \mathbf{M}_t^n$.

Proof. The statement can be proved by induction. If $n = 1$ there is nothing to prove. Assume now (by induction) that it holds for all decompositions of $\mathbf{U}_0$ with size less than n. Let $r = r_h = \min_{i \in [n-1]} r_i$ and assume (without loss of generality) that $r_n = \min\{r_0, r_n\}$. We want to describe two paths $t \in [0,1] \mapsto \mathbf{V}_t \in \mathbb{R}^{r_0 \times r}$, $t \in [0,1] \mapsto \mathbf{W}_t \in \mathbb{R}^{r \times r_n}$ such that $\mathbf{U}_t = \mathbf{V}_t\mathbf{W}_t$ and $\mathbf{V}_0 = \tilde{\mathbf{M}}^1 \cdots \tilde{\mathbf{M}}^h$, $\mathbf{W}_0 = \tilde{\mathbf{M}}^{h+1} \cdots \tilde{\mathbf{M}}^n$. By operating as in step 1 in the proof of Theorem 4.5,

we can assume $\mathrm{rk}(\mathbf{W}_0) = r_n$. Moreover (up to adding a linear path in $\mathbf{V}_t$) we can assume that $\mathbf{V}_0 = \mathbf{U}_0\mathbf{W}_0^\dagger$. We can then define $\mathbf{V}_t = \mathbf{U}_t\mathbf{W}_0^\dagger$ and $\mathbf{W}_t = \mathbf{W}_0$ for $t \in (0, 1]$. We thus factorized $\mathbf{U}_t$ as $\mathbf{U}_t = \mathbf{V}_t\mathbf{W}_t$. By induction, we can assume that we can factorize $\mathbf{V}_t = \mathbf{M}_t^1 \cdots \mathbf{M}_t^h$ and $\mathbf{W}_t = \mathbf{M}_t^{h+1} \cdots \mathbf{M}_t^n$. This concludes the proof. □

We can now conclude the proof of Theorem 4.8.

Proof of Theorem 4.8. Consider a linear network $\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta})$ as in (C.4), where

$$\mathbf{W}_k \in \mathbb{R}^{d_{k-1} \times d_k} \quad \text{for } k \in [L+1]$$

We select $d_s = \min_{i \in [L]} d_k$. Then the network can be written as

$$\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \hat{\mathbf{W}}^2\hat{\mathbf{W}}^1\,\mathbf{x} \quad \text{where} \quad \hat{\mathbf{W}}^2 = \mathbf{W}_{L+1}^T \cdots \mathbf{W}_{s+1}^T, \quad \hat{\mathbf{W}}^1 = \mathbf{W}_s^T \cdots \mathbf{W}_1^T \tag{C.5}$$

Now we want to prove property that given an initial parameter $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{W}}_{L+1}, \dots, \tilde{\mathbf{W}}_1)$, there exists a continuous path $\boldsymbol{\theta}_t = (\mathbf{W}_{L+1,t}, \dots, \mathbf{W}_{1,t})$ such that $L(\boldsymbol{\theta}_t)$ is non-increasing and such that $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}$ and $L(\boldsymbol{\theta}_1) = \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. If we call $\hat{\tilde{\mathbf{W}}}^i$, $i = 1, 2$, the matrices defined in (C.5) for $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, then by Lemma C.4 there exists a path $(\hat{\mathbf{W}}_t^2, \hat{\mathbf{W}}_t^1)$ satisfying the above. Thanks to Lemma C.5, we can decompose

$$\hat{\mathbf{W}}_t^2 = \mathbf{W}_{K+1,t}^T \cdots \mathbf{W}_{s+1,t}^T, \quad \hat{\mathbf{W}}_t^1 = \mathbf{W}_{s,t}^T \cdots \mathbf{W}_{1,t}^T$$

in a continuous way. Since d_s was to chosen as the minimum, it also holds that

$$\min_{\boldsymbol{\theta} = (\hat{\mathbf{W}}^2, \hat{\mathbf{W}}^1)} L(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta} = (\mathbf{W}^{L+1}, \dots, \mathbf{W}^1)} L(\boldsymbol{\theta})$$

Therefore this is a suitable path and this concludes the proof of the theorem. □

C.2.3 Proof of Theorem 4.9

Proof of Theorem 4.9. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{u}}, \tilde{\mathbf{W}})$ be a starting parameter value. We aim to construct a continuous path $t \in [0,1] \mapsto \boldsymbol{\theta}_t \in \Theta$ starting in $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}$ and such that $L(\boldsymbol{\theta}_1) = \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ and such that the function $t \in [0,1] \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing. Such a path can be constructed in two steps.

Step 1. Let $\mathbf{A} = \sum_{k=1}^{N} \tilde{u}_k \tilde{\mathbf{w}}_k \tilde{\mathbf{w}}_k^T$ and let $\sum_{k=1}^{d} u_k^* \mathbf{w}_k^* (\mathbf{w}_k^*)^T$ be the SVD of $\mathbf{A}$. We define the parameters value $\boldsymbol{\theta}^* = (\mathbf{u}^*, \mathbf{W}^*)$ where $\mathbf{u}^* = (u_1^*, \dots, u_d^*, 0, \dots, 0) \in \mathbb{R}^N$ and $\mathbf{W}^*$ is the $d \times N$ matrix with columns $\mathbf{w}_i^*$ for $i \in [d]$ and $\mathbf{0}$ for $i \in [d+1, N]$. The first step consists in continuously mapping $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{u}}, \tilde{\mathbf{W}})$ to $\boldsymbol{\theta}^* = (\mathbf{u}^*, \mathbf{W}^*)$ with a path $\boldsymbol{\theta}_t$ such that $L(\boldsymbol{\theta}_t)$ is constant; the construction of such a path is detailed in Lemma C.6.

Step 2. As noticed above, the network can be written as $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{u}^T \boldsymbol{\sigma}(\mathbf{W}^T \mathbf{x}) = \langle \mathbf{A}, \mathbf{M} \rangle_F$, where $\mathbf{A} = \sum_{k=1}^{N} u_k \mathbf{w}_k \mathbf{w}_k^T$ and $\mathbf{M} = \mathbf{x}\mathbf{x}^T$. The square loss $L(\boldsymbol{\theta})$ is convex in the parameter $\mathbf{A}$. Be $\bar{\mathbf{A}}$ a minima of L as function of $\mathbf{A}$ and let $\sum_{i=1}^{d} \bar{u}_k \bar{\mathbf{w}}_k \bar{\mathbf{w}}_k^T$ be the SVD of $\bar{\mathbf{A}}$; also let $\bar{\mathbf{u}} = (0, \dots, 0, \bar{u}_1, \dots, \bar{u}_d)$ and $\bar{\mathbf{W}}$ be the $d \times N$ matrix with columns $\mathbf{0}$ for $i \in [N-d]$ and $\bar{\mathbf{w}}_i$ for $i \in [N-d+1, N]$. By the previous step we can assume that the initial parameter $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{u}}, \tilde{\mathbf{W}})$ is such that $\tilde{u}_i = 0$ and $\tilde{\mathbf{w}}_i = \mathbf{0}$ for $i \in [d+1, N]$. Then the path $\boldsymbol{\theta}_t = (1-t)(\mathbf{u}, \mathbf{W}) + t(\bar{\mathbf{u}}, \bar{\mathbf{W}})$ verifies property **P.1**. This indeed follows from the fact that $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}_t) = (1-t)\langle \mathbf{A}, \mathbf{M} \rangle_F + t \langle \bar{\mathbf{A}}, \mathbf{M} \rangle_F$ and from the convexity of the loss L as function of $\mathbf{A}$.

This shows that property **P.1** holds and so it concludes the proof of Theorem 4.9. □

To conclude the proof we just need to prove the following lemmas.

Lemma C.6. *Let $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W})$ be an initial parameter and $\boldsymbol{\theta}^* = (\mathbf{u}^*, \mathbf{W}^*)$ be as in step 1 of the proof of Theorem 4.9. Then there exists a continuous path $\boldsymbol{\theta}_t$ from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$ such that the loss $L(\boldsymbol{\theta}_t)$ is constant (as a function of t).*

Proof. Notice that we can assume $\mathbf{u} \in \{-1, 0, 1\}^p$. This can be done simply scaling (continuously) each column $\mathbf{w}_k$ of $\mathbf{W}$ by $\overline{|u_k|}$. Assume first that $\mathbf{u} \in \{\pm 1\}^N$. The general case ($u_k = 0$ for some k) is addressed in Remark 9. The sought path $\boldsymbol{\theta}_t$ can be constructed by iterating two steps (a finite amount of times). First we select a column $\mathbf{w}_k$ and construct a continuous path that maps this column to one of the $\mathbf{w}_i^*$; then we orthogonalize (with respect to such $\mathbf{w}_i^*$) the rest of the columns $\mathbf{w}_j$, $j \neq k$. These two steps are performed so that $\mathbf{A}$ never changes and therefore the loss is constant. The first step is described in Lemma C.7, while the second is detailed in Lemma C.8. At this point the parameter $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W})$ verifies $u_i = u_i^*$, $\mathbf{w}_i = \mathbf{w}_i^*$ and $\mathbf{w}_j \in (\mathrm{span}(\{\mathbf{w}_i^*\}))^\perp$ for $j \neq k$. In particular it holds that

$$\sum_{\substack{j=1 \\ j \neq i}}^{d} u_j^* \mathbf{w}_j^* (\mathbf{w}_j^*)^T = \sum_{\substack{j=1 \\ j \neq k}}^{N} u_k \mathbf{w}_k \mathbf{w}_k^T \, .$$

Therefore, an induction step applied on the reduced parameter values

$$\mathbf{u}_{-k} = (u_1, \dots, \widehat{u_k}, \dots, u_N)$$

and $\mathbf{W}_{-k} = \mathbf{P}[\mathbf{w}_1 | \dots | \widehat{\mathbf{w}_k} | \dots | \mathbf{w}_N]$, where $\mathbf{P} = \sum_{j=1, j \neq i}^{d} \mathbf{e}_j (\mathbf{w}_j^*)^T \in \mathbb{R}^{(d-1) \times d}$, concludes the proof. The fact that the non-zero components of $\mathbf{u}$ and $\mathbf{W}$ coincide with the first d is not necessary, but we can clearly assume it to hold without loss of generality. □

Lemma C.7. *The first step described in the Proof of Lemma C.6 can be performed when $N > 2d$.*

Proof. Let $E_+ = \{k \in [N] : u_k = 1\}$, $E_- = \{k \in [N] : u_k = -1\}$ and $N_+ = |E_+|$, $N_- = |E_-|$. Accordingly we define

$$\mathbf{W}_+ = [\mathbf{w}_k]_{k \in E_+} \in \mathbb{R}^{d \times N_+} \quad \text{and} \quad \mathbf{W}_- = [\mathbf{w}_k]_{k \in E_-} \in \mathbb{R}^{d \times N_-} \, .$$

Notice that then we can write

$$\mathbf{A} = \mathbf{W}_+\mathbf{W}_+^T - \mathbf{W}_-\mathbf{W}_-^T \, .$$

The main step of the proof is to observe that $\mathbf{A}$ (and therefore the loss) is invariant to the action of orthogonal matrices $\mathbf{Q}_+ \in SO(N_+)$ and $\mathbf{Q}_- \in SO(N_-)$. So, if $\mathbf{Q}_+(t)$ (resp. $\mathbf{Q}_-(t)$) is a continuous paths in $SO(N_+)$ (resp. in $SO(N_-)$) starting at the identity, acting on $\mathbf{W}$ as

$$\mathbf{W}_+(t) \doteq \mathbf{W}_+\mathbf{Q}_+(t), \quad \mathbf{W}_-(t) \doteq \mathbf{W}_-\mathbf{Q}_-(t) \, ,$$

we have that

$$\mathbf{A} = \mathbf{W}_+(t)\mathbf{W}_+(t)^T - \mathbf{W}_-(t)\mathbf{W}_-(t)^T$$

is constant for all t. Now, since $N = N_+ + N_- > 2d$, it follows that either $N_+ > d$ or $N_- > d$. Assume without loss of generality that $N_+ > d$. Since $N_+ > d$, we can rotate the subspace generated by the rows of $\mathbf{W}_+$ so that its first column is $\mathbf{0}$. That is, there exist $\mathbf{h} \in \mathbb{R}^{p_+}$ non-zero such that $\mathbf{W}_+\mathbf{h} = \mathbf{0}$ and $\|\mathbf{h}\|_2 = 1$. It then suffices to choose a path $\mathbf{Q}(t)$ in $SO(p_+)$ whose first column equals $\mathbf{h}$ at $t = 1$. It follows that $\mathbf{W}_+\mathbf{Q}(1)$ has a first column equal to $\mathbf{0}$. We then set the corresponding $u_1 = 0$, which does not change the loss, and finally set $\mathbf{w}_1$ to the desired eigenvector $\mathbf{w}_1^*$. □

Lemma C.8. *Assume that after the step in Lemma C.7, the first column of $\mathbf{W}_+$ (resp. $\mathbf{W}_-$) is given by $\mathbf{w}_i^*$. Then we can map all the other columns of $\mathbf{W}$ to be orthogonal to $\mathbf{w}_i^*$, while keeping $\mathbf{A}$ constant.*

Proof. To simplify the notation we assume, without loss of generality, that $\mathbf{w}_i^* = \mathbf{w}_1^*$ and that

$$\mathbf{W} = [\mathbf{w}_1^*|\mathbf{w}_2|\cdots|\mathbf{w}_N] \, .$$

Now we want to construct a path

$$\mathbf{u}_t = (u_{1,t}, u_2, \ldots, u_N)$$
$$\mathbf{W}_t = [\mathbf{w}_1^* | \mathbf{w}_{2,t} | \cdots | \mathbf{w}_{N,t}]$$

such that $\mathbf{w}_{2,1}, \ldots, \mathbf{w}_{N,1} \in (\mathrm{span}(\{\mathbf{w}_1^*\}))^\perp$. To do this we simply take

$$\mathbf{w}_{k,t} \doteq \mathbf{w}_k - t(\mathbf{w}_k^T \mathbf{w}_1^*)\mathbf{w}_1^* \,.$$

If $\mathbf{A}_t = \sum_{k=1}^N u_{k,t}\mathbf{w}_{k,t}\mathbf{w}_{k,t}^T$, we can show that there exists a choice of $u_{1,t}$ such that $\mathbf{A}_t = \mathbf{A}$ for all $t \in [0,1]$. It holds that

$$\begin{aligned}\mathbf{A}_t &= u_{1,t}\, \mathbf{w}_1^*(\mathbf{w}_1^*)^T \\ &\quad + \sum_{k=2}^N u_k \;(1-t)^2(w_k^1)^2 \mathbf{w}_1^*(\mathbf{w}_1^*)^T + (1-t)w_k^1 \;\tilde{\mathbf{w}}_k(\mathbf{w}_1^*)^T + \mathbf{w}_1^*\tilde{\mathbf{w}}_k^T \;\; + \tilde{\mathbf{w}}_k\tilde{\mathbf{w}}_k^T\end{aligned}$$

where $w_k^1 \doteq \mathbf{w}_k^T\mathbf{w}_1^*$ and $\tilde{\mathbf{w}}_k = \mathbf{w}_k - w_k^1\mathbf{w}_1^*$. In particular

$$\mathbf{A}_t = \mathbf{V}^* \left[\begin{array}{c|c} a_t & \mathbf{b}_t^T \\ \hline \mathbf{b}_t & \mathbf{A}_{2:d,2:d} \end{array}\right] (\mathbf{V}^*)^T \,,$$

where $\mathbf{V}^* = [\mathbf{w}_1^*, \cdots, \mathbf{w}_d^*] \in O(d)$. Since $\sum_{k=2}^N u_k w_k^1 \tilde{\mathbf{w}}_k = 0$, it follows that

$$\mathbf{b}_t = (1-t)\sum_{k=2}^N u_k w_k^1 \tilde{\mathbf{w}}_k = 0 \quad \text{for all } t \in [0,1] \,.$$

If we take

$$u_{1,t} = \lambda_1 - (1-t)^2 \sum_{k=2}^N u_k (w_k^1)^2 \,,$$

it holds that

$$a_t = u_{1,t} + (1-t)^2 \sum_{k=2}^{N} u_k (w_k^1)^2 = \lambda_1 \quad \text{for all } t \in [0,1] \,.$$

Therefore, $\mathbf{A}_t = \mathbf{A}$ constant. This concludes the proof of the lemma. □

Remark 9. In the proof of Lemma C.6, we assumed that (after rescaling) $\mathbf{u} \in \{\pm 1\}^N$. In general, it could be that $u_k = 0$ for some k. In this case we can first map the corresponding vectors $\mathbf{w}_k$ to $\mathbf{0}$ and the map such u_k to 1, without affecting the loss.

C.3 Proofs of results regarding existence of spurious valleys

C.3.1 Proof of Theorem 4.10

We consider here the case $m = 1$, but the same proof can be extended to the case $m > 1$. We start by proving the following fact.

Lemma C.9. *Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ continuous and $\mathbf{X} \in \mathcal{R}(\sigma, d)$. Define the spaces*

$$\mathcal{F}_N^{\sigma,+} \doteq \{\Phi(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in [0,\infty)^N \times \mathbb{R}^{d\times N}\} \subseteq L^2_{\mathbf{X}}$$

for $N \geq 1$. If it holds that

$$\mathcal{F}_{R+1}^{\sigma,+} \subseteq \overline{\mathcal{F}_R^{\sigma,+}} \tag{C.6}$$

for some $R \geq 1$, then $2R \geq \dim_*(\sigma, \mathbf{X})$.

Proof. Assume that equation (C.6) holds for a certain $R > 0$. Then, for every $k \geq 1$, it holds that

$$\mathcal{F}_k^{\sigma,+} \subseteq \overline{\mathcal{F}_R^{\sigma,+}}$$

This can be shown by induction over k, starting from $k = R$. Then, it holds that, for all $k \geq 1$,

$$\mathcal{F}_k^\sigma = \bigcup_{j=0}^{k} \mathcal{F}_j^{\sigma,+} - \mathcal{F}_{k-j}^{\sigma,+} \subseteq \bigcup_{j=0}^{k} \overline{\mathcal{F}_R^{\sigma,+}} - \overline{\mathcal{F}_R^{\sigma,+}}$$
$$= \overline{\mathcal{F}_R^{\sigma,+}} - \overline{\mathcal{F}_R^{\sigma,+}} \subseteq \overline{\mathcal{F}_R^{\sigma,+} - \mathcal{F}_R^{\sigma,+}} \subseteq \overline{\mathcal{F}_{2R}^{\sigma}} .$$

Since this holds for every $k \geq 1$, then $\mathcal{F}^\sigma \subseteq \overline{\mathcal{F}_{2R}^\sigma}$, which implies the thesis. □

We can now complete the proof of Theorem 4.10. We start by properly choosing a random vector $(\mathbf{X}, \mathbf{Y})$. Let $\bar{\mathbf{X}} \in \mathcal{R}_2(\sigma, d-1)$ a $(d-1)$ dimensional random variable and $\bar{X}_d \in \mathcal{R}_2(\sigma, 1)$ a one dimensional random variable. We consider $\tilde{\mathbf{X}} = Z\bar{\mathbf{X}}$, $X_d = (1-Z)\bar{X}_d$ and $\mathbf{X} = (\tilde{\mathbf{X}}, X_d)$, where $Z \sim \mathrm{Ber}(1/2)$ and $\bar{\mathbf{X}}, \bar{X}_d, Z$ are independent. By hypothesis, $N \leq 2^{-1}\dim_*(\sigma, \tilde{\mathbf{X}})$. By Lemma C.9, this implies that $\mathcal{F}_N^{\sigma,+} \subsetneq \overline{\mathcal{F}_{N-1}^{\sigma,+}}$. The random variable Y is taken to be $Y = g_1(\mathbf{X}) - g_2(\mathbf{X})$, where $g_2 = \beta\psi_{\sigma,\mathbf{v}} \in \mathcal{F}_1^{\sigma,+}$, $\beta > 0$, $\mathbf{v} = \mathbf{e}_d$, and $g_1 = \sum_{i=1}^{N} \alpha_i \psi_{\sigma,\mathbf{v}_i} \in \mathcal{F}_N^{\sigma,+}$, $\boldsymbol{\alpha} \in (0,\infty)^N$, $\mathbf{v}_i \in (\mathrm{span}(\{\mathbf{e}_d\}))^\perp$, $i \in [N]$, is such that

$$\inf_{f \in \mathcal{F}_{N-1}^{\sigma,+}} \mathbb{E}|f(\mathbf{X}) - g_1(\mathbf{X})|^2 = \epsilon > 0 .$$

We define

$$\mathcal{F}_{(N-1,1)}^\sigma = \left\{ f = f_1 - f_2 : f_1 \in \mathcal{F}_{N-1}^{\sigma,+}, f_2 \in \mathcal{F}_1^{\sigma,+} \right\} .$$

Notice that, for every path $\boldsymbol{\theta} : t \in [0,1] \mapsto \boldsymbol{\theta}_t \in \Theta$ such that $\Phi(\cdot;\boldsymbol{\theta}_0) \in \mathcal{F}_N^{\sigma,+}$ and $\Phi(\cdot;\boldsymbol{\theta}_1) \in \mathcal{F}_{(N-1,1)}^\sigma$, there exists $t_0 \in (0,1)$ such that $\Phi(\cdot;\boldsymbol{\theta}_{t_0}) \in \mathcal{F}_{N-1}^{\sigma,+}$. Consider the *lifted* square loss function $L : \mathcal{F}_N^\sigma \to [0,\infty)$ defined as

$$L(f) = \mathbb{E}|f(\mathbf{X}) - g(\mathbf{X})|^2 \qquad \text{for } f \in \mathcal{F}_N^\sigma .$$

We want to show that

$$L_{(N-1,0)} \doteq \min_{f \in \mathcal{F}_{N-1}^{\sigma,+}} L(f) > L_{(N,0)} \doteq \min_{f \in \mathcal{F}_{N}^{\sigma,+}} L(f) > L_{(N-1,1)} \doteq \min_{f \in \mathcal{F}_{(N-1,1)}^{\sigma}} L(f) \, .$$

It holds that

$$\begin{aligned} L_{(N-1,0)} &= \min_{f \in \mathcal{F}_{N-1}^{\sigma,+}} \mathbb{E}|f(\mathbf{X}) - g_1(\mathbf{X})|^2 \quad + 2 \min_{f \in \mathcal{F}_{N-1}^{\sigma,+}} \{\mathbb{E}[f(\mathbf{X}) g_2(\mathbf{X})]\} + \mathbb{E}|g_2(\mathbf{X})|^2 \\ &\geq \epsilon + L_{(N,0)} \end{aligned}$$

and that

$$\begin{aligned} L_{(N,0)} &= \min_{f \in \mathcal{F}_{N}^{\sigma,+}} \mathbb{E}|f(\mathbf{X}) - g_1(\mathbf{X})|^2 \quad + 2 \min_{f \in \mathcal{F}_{N}^{\sigma,+}} \{\mathbb{E}[f(\mathbf{X}) g_2(\mathbf{X})]\} + \mathbb{E}|g_2(\mathbf{X})|^2 \\ &\geq \beta^2 \, \mathbb{E}|\psi_{\sigma,\mathbf{v}}(X_d)|^2 \, . \end{aligned}$$

Finally, it holds that

$$L_{(N-1,1)} \leq \min_{i \in [N]} \alpha_i^2 \, \mathbb{E}|\psi_{\sigma,\mathbf{v}_i}(\mathbf{X})|^2 \, .$$

Given $M > 0$, up to multiply g_1 by a positive constant, there exists $\beta > 0$ such that

$$\epsilon \geq M \quad \text{and} \quad \beta^2 \geq \frac{M + \min_{i \in [N]} \alpha_i^2 \, \mathbb{E}|\psi_{\sigma,\mathbf{v}_i}(\mathbf{X})|^2}{\mathbb{E}|\psi_{\sigma,\mathbf{v}}(X_d)|^2} \, .$$

To finish the proof, consider $\mathcal{U} = \{\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W}) \in \Theta : \mathbf{u} \in (0, \infty)^N\}$ and $\boldsymbol{\theta}^* \in \mathcal{U}$ such that

$$L(\boldsymbol{\theta}^*) = \min_{\boldsymbol{\theta} \in \mathcal{U}} L(\boldsymbol{\theta}) \, .$$

Then, (by continuity of L) there exists a neighborhood $\boldsymbol{\theta}^* \in \Omega \subset \mathcal{U}$ such that $\sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}) \leq L(\boldsymbol{\theta}^*) + M/2$. The set Ω then verifies the statement of the theorem.

C.4 Proofs for Section 4.5.1

C.4.1 Proof of Theorem

If we denote by $d\mu$ the probability distribution of $\mathbf{X}$ and S the uniform measure over $\mathbb{S}^d$, the continuous function

$$\psi : (\mathbf{w}, \mathbf{x}) \in \mathbb{S}^d \times \mathbb{R}^d \mapsto \psi_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$$

belongs to $L^2(S \otimes \mu)$. We consider the kernel associated with the neural network architecture

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathcal{W}} \psi_{\mathbf{w}}(\mathbf{x})\psi_{\mathbf{w}}(\mathbf{y})\, dS(\mathbf{w}) \,. \tag{C.7}$$

The above defines a continuous symmetric, positive semi-definite kernel k, along with $\mathcal{H}^2$, the RKHS associated, and the integral operator $\Sigma : L^2(\mu) \to \mathcal{H}^2 \subseteq L^2(\mu)$ defined as

$$f \mapsto \quad \Sigma f : \mathbf{x} \mapsto \int_{\mathbb{R}^d} f(\mathbf{y})k(\mathbf{x}, \mathbf{y})\, d\mu(\mathbf{y}) \quad .$$

The operator Σ admits a spectral decomposition in $L^2(\mu)$: $\Sigma e_k = \lambda_k e_k$ for an orthonormal basis $\{e_k\}_{k\geq 1}$ of $L^2(\mu)$ and non-increasing sequence of non-negative eigenvalues $\{\lambda_k\}_{k\geq 1}$. Moreover the RKHS $\mathcal{H}^2$ is dense in $L^2(\mu)$ (see Lemma C.13), which is equivalent to have $\lambda_k > 0$ for all $k \geq 1$. The expectation in (C.7) provides a singular value decomposition for Σ in terms of functions in $L^2(S)$. Indeed, given $g \in L^2(S)$, the linear operator $T : L^2(S) \to L^2(\mu)$ defined as

$$g \mapsto \quad Tg : \mathbf{x} \mapsto \int_{\mathbb{S}^d} g(\mathbf{w})\psi_{\mathbf{w}}(\mathbf{x})\, dS(\mathbf{w})$$

satisfies $\Sigma = TT^*$. It follows that there exists an orthonormal basis of $L^2(S)$, $\{f_k\}_{k\geq 1}$ such that $Tf_k = \lambda_k^{1/2} e_k$ and therefore $\psi_{\mathbf{w}} = \sum_{k=1}^{\infty} \lambda_k^{1/2} f_k(\mathbf{w}) e_k$. Finally, it can be shown [Bac17a] that in

fact $\mathcal{H}^2 = \mathrm{Im}(T)$, and thus $\mathcal{H}^2$ consists of functions f that can be written, for some $g \in L^2(S)$ as

$$f(\mathbf{x}) = \int_{\mathbb{S}^d} g(\mathbf{w})\psi_{\mathbf{w}}(\mathbf{x})dS(\mathbf{w}) = \langle g, \psi(\cdot, \mathbf{x})\rangle_{L^2(\mathbb{S}^n, dS)} \quad \text{for } \mathbf{x} \in \mathbb{R}^d \,.$$

For an account of these properties, we refer to Bach [Bac17b]. Thanks to the density of $\mathcal{H}^2$ in $L^2(\mu)$, we can assume, without loss of generality, that

$$f^*(\mathbf{x}) = \int_{\mathbb{S}^d} g^*(\mathbf{w})\psi_{\mathbf{w}}(\mathbf{x})dS(\mathbf{w}) \,,$$

for some $g^* \in L^2(S)$. Now, given an initial set of first layer weights $\mathbf{w}_1, \ldots, \mathbf{w}_N \in \mathbb{S}^n$ sampled i.i.d. from S, and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]$, we define the empirical kernel

$$k_{\mathbf{W}}(\mathbf{x}, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} \sigma(\langle \mathbf{x}, \mathbf{w}_i\rangle)\sigma(\langle \mathbf{y}, \mathbf{w}_i\rangle) \,,$$

which in turn defines an empirical RKHS $\mathcal{H}^2_{\mathbf{W}}$. Keeping the first layer weights fixed and optimizing the output layer weights thus gives us the ability to find a function $f^*_{\mathbf{W}} \in \mathcal{H}^2_{\mathbf{W}}$ that best approximates f^*:

$$\|f^*_{\mathbf{W}} - f^*\|_{L^2(\mu)} = \min_{f \in \mathcal{H}^2_{\mathbf{W}}} \|f - f^*\|_{L^2(\mu)} \doteq R(\mathbf{W}) \,.$$

Given an initial parameter parameter value $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{u}}, \tilde{\mathbf{W}})$ (here we incorporated $\tilde{\mathbf{b}}$ in $\tilde{\mathbf{W}}$) as in the statement, consider the path

$$\boldsymbol{\theta}_t = (t\mathbf{q}(\tilde{\mathbf{W}}) + (1-t)\tilde{\mathbf{u}}, \tilde{\mathbf{W}}) \quad \text{where} \quad \mathbf{q}(\tilde{\mathbf{W}}) = \operatorname*{arg\,min}_{\mathbf{u} \in \mathbb{R}^N} L(\boldsymbol{\theta})|_{\boldsymbol{\theta} = (\mathbf{u}, \tilde{\mathbf{W}})} \,.$$

By convexity of L, the function $t \in [0,1] \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing and it holds that

$$L(\boldsymbol{\theta}_1) \leq \mathcal{R}(\mathbf{X}, Y) + R(\tilde{\mathbf{W}}) \,.$$

Applying Proposition 1 from Bach [Bac17b], it holds that

$$R(\mathbf{W}) \leq 4\lambda \quad \text{if} \quad p \geq 5d(\lambda)\log(16d(\lambda)/\delta)$$

with probability greater or equal than $1-\delta$, where

$$\begin{aligned} d(\lambda) &= \max_{\mathbf{w}\in\mathbb{S}^d} \mathbb{E}\ \varphi_{\mathbf{w}}(\mathbf{X})((\Sigma+\lambda I)^{-1}\psi_{\mathbf{w}})(\mathbf{X}) \\ &= \max_{\mathbf{w}\in\mathbb{S}^d} \sum_{k=1}^{\infty} \frac{\lambda_k}{\lambda_k+\lambda} f_k(\mathbf{w})^2 \leq \lambda^{-1} \max_{\mathbf{w}\in\mathbb{S}^d} \sum_{k=1}^{\infty} \lambda_k f_k(\mathbf{w})^2 = \lambda^{-1} \max_{\mathbf{w}\in\mathbb{S}^d} \|\psi_{\mathbf{w}}\|^2_{L^2(\mu)} \,. \end{aligned}$$

This concludes the proof.

C.5 Useful lemmas

Lemma 4.1. *Be $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta})$ a continuous function. Then, property **P.1** implies absence of spurious valleys. In particular, this implies absence of strict spurious minima, and of (generally non-strict) spurious minima if property **P.1** holds with strictly decreasing paths $t \mapsto L(\boldsymbol{\theta}_t)$. Conversely, presence of spurious valleys implies existence of spurious minima.*

Proof. Assume that property **P.1** holds. Consider any value $c > 0$ such that $\Omega_L(c)$ is non-empty and let $\mathcal{U}$ be a path-connected component of $\Omega_L(c)$. Given a point $\boldsymbol{\theta} \in \mathcal{U}$ there exists a path from $\boldsymbol{\theta}$ satisfying property **P.1**. This means that $\mathcal{U}$ contains a global minima, and therefore it can not be a spurious valley. Similarly, assume that property **P.1** holds with strictly decreasing paths and that the function L admits a strict local minima. This means that there exists a point $\boldsymbol{\theta}_0$ such that $\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}_0) < L(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ in $B\ (\boldsymbol{\theta})$, for some $\epsilon > 0$. But this implies that for any path $t \in [0,1] \mapsto \boldsymbol{\theta}_t$ it holds $L(\boldsymbol{\theta}_t) > L(\boldsymbol{\theta}_0)$ for some $t > 0$ sufficiently small, a contradiction. To see the last point, assume that there exist spurious valleys and consider $\mathcal{U}$ a path-connected component of $\Omega_L(c)$ for some $c > 0$. Then $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ is a spurious minima. □

Lemma C.10. *Let k, d be positive integers such that $k \geq 2(d-1)$. Then it holds that*

$$\mathrm{rk}_{\mathrm{S}}(k, d) \geq (1+r)^{d-1}$$

where $r = \lfloor k/(2(d-1)) \rfloor$.

Proof. Let k, d and r as in the statement. Every element of $\mathrm{S}^k(\mathbb{R}^d)$ is in one-to-one correspondence with a homogeneous polynomials of degree k over $\mathbb{R}^d$. It has been shown in [LT10], Theorem 1.1, that the tensor corresponding to the polynomial

$$\pi_r(\mathbf{x}) \doteq x_1^{k-(d-1)r} \cdot \prod_{j=2}^{d} x_j^r$$

has border rank equal to $(1+r)^{d-1}$, if $k-(d-1)r \geq (d-1)r$. Although, the notion of border rank considered in [LT10] is over the complex field. Since we are interested in the corresponding notion over the real field, we get the inequality of the statement in place of equality. □

Lemma C.11. *Consider the optimization problem*

$$\arg\min_{\mathbf{W} \in \mathbb{R}^{m \times d}} \ell(\mathbf{W}) \quad \textit{where} \quad \ell(\mathbf{W}) = \mathbb{E}\|\mathbf{W}\mathbf{X} - \mathbf{Y}\|^2$$

for two square integrable random variables $\mathbf{X}$ and $\mathbf{Y}$ with values in $\mathbb{R}^d$ and $\mathbb{R}^m$ respectively. Then one solution to (C.11) is given by

$$\mathbf{W} = \mathbf{\Sigma}_{\mathbf{YX}} \mathbf{\Sigma}_{\mathbf{X}}^{\dagger} .$$

Similarly, one solution to the optimization problem

$$\arg\min_{\mathbf{U} \in \mathbb{R}^{m \times p}} \ell(\mathbf{U}; \mathbf{W}) \quad \textit{where} \quad \ell(\mathbf{U}; \mathbf{W}) = \mathbb{E}\|\mathbf{U}\mathbf{W}\mathbf{X} - \mathbf{Y}\|^2$$

for any $\mathbf{W} \in \mathbb{R}^{p\times d}$ *is given by*

$$\mathbf{U} = \mathbf{Q}(\mathbf{W}) \doteq \mathbf{\Sigma_{YX}} \mathbf{W}^T (\mathbf{W} \mathbf{\Sigma_X} \mathbf{W}^T)^\dagger \, .$$

Assuming that $\mathbf{\Sigma_X}$ *is invertible, the minimal value obtained by* $\ell(\mathbf{U}; \mathbf{W})$ *is given by*

$$\ell(\mathbf{Q}(\mathbf{W}); \mathbf{W}) = \operatorname{tr}(\mathbf{\Sigma_Y}) - \operatorname{tr}((\mathbf{WK})^\dagger (\mathbf{WK}) \mathbf{M}) \tag{C.8}$$

where $\mathbf{K} = \mathbf{\Sigma_X}^{1/2}$ *and* $\mathbf{M} = \mathbf{K}^{-1} \mathbf{\Sigma_{XY}} \mathbf{\Sigma_{YX}} \mathbf{K}^{-1}$. *If* $\mathbf{M} = \sum_{i=1}^{d} \mathbf{n}_i \mathbf{n}_i \mathbf{v}_i^T$ *is the SVD of* $\mathbf{M}$, *the quantity* (C.8) *is minimized over* $\mathbf{W}$ *for* $(\mathbf{WK})^\dagger (\mathbf{WK}) = \sum_{i=1}^{p\wedge d} \mathbf{n}_i \mathbf{n}_i^T$.

Proof. The first part of the lemma can be shown by writing problem (C.11) as

$$\underset{\mathbf{W}\in\mathbb{R}^{m\times d}}{\arg\min}\, \ell(\mathbf{W}) \quad \text{where} \quad \ell(\mathbf{W}) = \operatorname{tr}(\mathbf{W} \mathbf{\Sigma_X} \mathbf{W}^T) - 2 \operatorname{tr}(\mathbf{\Sigma_{YX}} \mathbf{W}^T)$$

and by taking $\mathbf{W}$ as a stationary point of the above $\ell(\mathbf{W})$. Using this fact, one minima of the function $\ell(\mathbf{U}; \mathbf{W})$ is given by

$$\mathbf{U} = \mathbf{\Sigma_{YXW}} (\mathbf{\Sigma_W X})^\dagger = \mathbf{\Sigma_{YX}} \mathbf{W}^T (\mathbf{W} \mathbf{\Sigma_X} \mathbf{W}^T)^\dagger \, .$$

Now assume that $\mathbf{\Sigma_X}$ is invertible; let $\mathbf{K} = (\mathbf{\Sigma_X})^{1/2}$ and $\mathbf{M} = \mathbf{K}^{-1}\mathbf{\Sigma_{XY}}\mathbf{\Sigma_{YX}}\mathbf{K}^{-1}$. Then it holds

$$\begin{aligned}
\ell(\mathbf{Q}(\mathbf{W}); \mathbf{W}) &= \operatorname{tr}(\mathbf{Q}(\mathbf{W})\mathbf{W}\mathbf{\Sigma_X}\mathbf{W}^T\mathbf{Q}(\mathbf{W})^T) - 2\operatorname{tr}(\mathbf{\Sigma_{YX}}\mathbf{W}^T\mathbf{Q}(\mathbf{W})^T) + \operatorname{tr}(\mathbf{\Sigma_Y}) \\
&= \operatorname{tr}(\mathbf{\Sigma_{YX}}\mathbf{W}^T(\mathbf{W}\mathbf{\Sigma_X}\mathbf{W}^T)^\dagger\mathbf{W}\mathbf{\Sigma_X}\mathbf{W}^T(\mathbf{W}\mathbf{\Sigma_X}\mathbf{W}^T)^\dagger\mathbf{W}\mathbf{\Sigma_{XY}}) \\
&\quad - 2\operatorname{tr}(\mathbf{\Sigma_{YX}}\mathbf{W}^T(\mathbf{W}\mathbf{\Sigma_X}\mathbf{W}^T)^\dagger\mathbf{W}\mathbf{\Sigma_{XY}}) + \operatorname{tr}(\mathbf{\Sigma_Y}) \\
&= -\operatorname{tr}(\mathbf{\Sigma_{YX}}\mathbf{W}^T(\mathbf{W}\mathbf{\Sigma_X}\mathbf{W}^T)^\dagger\mathbf{W}\mathbf{\Sigma_{XY}}) + \operatorname{tr}(\mathbf{\Sigma_Y}) \\
&= -\operatorname{tr}(\mathbf{M}(\mathbf{WK})^T((\mathbf{WK})(\mathbf{WK})^T)^\dagger(\mathbf{WK})) + \operatorname{tr}(\mathbf{\Sigma_Y}) \\
&= \operatorname{tr}(\mathbf{\Sigma_Y}) - \operatorname{tr}((\mathbf{WK})^\dagger(\mathbf{WK})\mathbf{M}) \,.
\end{aligned}$$

Finally, we notice that the matrix $(\mathbf{WK})^\dagger(\mathbf{WK})$ is the orthogonal projection on the space spanned by the rows of $\mathbf{WK}$, which we denote by $\mathbf{P}_{(\mathbf{WK})^T}$. In particular $\mathbf{P}_{(\mathbf{WK})^T}$ has the form $\mathbf{P}_{(\mathbf{WK})^T} = \sum_{i=1}^{r} \mathbf{v}_i\mathbf{v}_i^T$ for some $\{\mathbf{v}_1, \dots, \mathbf{v}_r\} \subset \mathbb{R}^d$ orthonormal vectors and $r \leq p \wedge d$. Therefore, minimize $\ell(\mathbf{Q}(\mathbf{W}); \mathbf{W})$ over $\mathbf{W}$ it is equivalent to maximize the quantity

$$\sum_{i=1}^{r} \mathbf{v}_i^T\mathbf{M}\mathbf{v}_i$$

over the sets of $\mathbf{v}_1, \dots, \mathbf{v}_r$ orthonormal vectors of $\mathbb{R}^d$, $r \leq p \wedge n$. Clearly, this is for $\mathbf{v}_1 = \mathbf{n}_1, \dots, \mathbf{v}_{p\wedge n} = \mathbf{n}_{p\wedge n}$. This concludes the proof of the lemma. □

Lemma C.12. *Let $X_1, \dots, X_n$ be independent zero-mean random variable taking values in a separable Hilbert space such that $\|X_i\| \leq c_i$ with probability one and denote $v = \sum_{i=1}^{n} c_i^2$. Then, for all $t \geq v$, it holds*

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i > t\right] \leq e^{-(t-\sqrt{v})^2/(2v)} \,.$$

Proof. The proof can be found in [BLM13], Example 6.3. □

Lemma C.13. *Consider $\sigma : \mathbb{R} \to \mathbb{R}$ a positively homogeneous activation function. Let $\mathbf{X}$ be a random variable taking values in $\mathbb{R}^d$ and let $\mathcal{H}^2 \subset L^2(\mathbf{X})$ be the RKHS defined by the kernel*

function

$$k : (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \int_{\mathbb{S}^{d-1}} \sigma(\mathbf{w}^T\mathbf{x})\sigma(\mathbf{w}^T\mathbf{y})\, dS(\mathbf{w}) \,.$$

Then $\mathcal{H}^2$ is dense in $L^2(\mathbf{X})$.

Proof. Let μ denote the distribution of $\mathbf{X}$ and $\psi_{\mathbf{w}} : \mathbf{x} \in \mathbb{R}^d \to \sigma(\mathbf{w}^T\mathbf{x})$. First, note that the function $\mathbf{x} \in \mathbb{R}^d \mapsto k(\mathbf{x}, \mathbf{x})$ is in $L^1(\mu)$. Indeed

$$\begin{aligned}\int_{\mathbb{R}^d}\int_{\mathbb{S}^{d-1}} \psi_{\mathbf{w}}(\mathbf{x})^2\, dS(\mathbf{w})\, d\mu(\mathbf{x}) &= \int_{\mathbb{R}^d} \|\mathbf{x}\|_2^2 \int_{\mathbb{S}^{d-1}} \psi_{\mathbf{w}}(\mathbf{x}/\|\mathbf{x}\|_2)^2\, dS(\mathbf{w})\, d\mu(\mathbf{x}) \\ &\leq \mathbb{E}\|\mathbf{X}\|_2^2 \max_{\mathbf{w},\mathbf{y}\in\mathbb{S}^{d-1}} \psi_{\mathbf{w}}(\mathbf{y})^2\end{aligned}$$

This implies that $\mathcal{H}^2 \subseteq L^2(\mu)$. Now, we would like to show that $\mathcal{F}^\sigma$ is dense in $\overline{\mathcal{H}^2}$, where

$$V_\sigma = \left\{ \sum_{i=1}^k u_i \psi_{\mathbf{w}_i} : \mathbf{u} \in \mathbb{R}^d, \mathbf{w}_1, \ldots, \mathbf{w}_k \in \mathbb{S}^{d-1}, k \geq 1 \right\}$$

It suffices to show that, for every $\mathbf{w} \in \mathbb{S}^{d-1}$, there exists a sequence $\{f_n\}_{n\geq 1} \subset \mathcal{H}^2$ such that $f_n \to \psi_{\mathbf{w}}$ in $L^2(\mathbf{X})$. Choose $g_k \in L^2(S)$ such that $\operatorname{supp}(g_k) \subseteq B_{1/k}(\mathbf{w}) \doteq \{\mathbf{v} \in \mathbb{S}^{d-1} : \|\mathbf{v}-\mathbf{w}\| \leq 1/k\}$, $\int_{\mathbb{S}^{d-1}} g(\mathbf{v})\, dS(\mathbf{v}) = 1$ and $g_k \geq 0$, and define $f_k \in \mathcal{H}^2$ as $f_k(x) = \int_{\mathbb{S}^{d-1}} g_k(\mathbf{v})\psi_{\mathbf{v}}(x)\, dS(\mathbf{x})$. Then

$$\begin{aligned}\|f_k - \psi_{\mathbf{w}}\|_{\mu,2}^2 &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{S}^{d-1}} g_k(\mathbf{v})(\psi_{\mathbf{v}}(\mathbf{x}) - \psi_{\mathbf{w}}(\mathbf{x}))\, dS(\mathbf{v}) \right)^2 d\mu(\mathbf{x}) \\ &\leq \mathbb{E}\|\mathbf{X}\|_2^2 \max_{\substack{\mathbf{v}\in B_{1/k,2^d}(\mathbf{w}) \\ \mathbf{y}\in\mathbb{S}^{d-1}}} (\psi_{\mathbf{v}}(\mathbf{y}) - \psi_{\mathbf{w}}(\mathbf{y}))^2 \to 0\end{aligned}$$

as $k \to \infty$. This shows that $\mathcal{F}^\sigma$ is contained in $\overline{\mathcal{H}^2}$. As shown in the proof of Lemma 4.2, it holds that $\mathcal{F}^\sigma$ is dense in $L^2(\mu)$. This implies the statement of the lemma. □

Appendix D

Appendix to chapter 5

D.1 Proofs

D.1.1 Proof of Proposition 5.1

Let $\mathcal{Z}$ be the class of (centrally symmetric) zonoids in $\mathbb{R}^d$. As observed in [BLM89], it holds that

$$r_d\delta_d \leq \epsilon_d \leq R_d\delta_d \,.$$

where

$$\begin{aligned}
\delta_d &\doteq \inf\{\delta > 0 : \exists\, Z \in \mathcal{Z} : \Delta_d \subseteq Z \subseteq (1+\delta)\Delta_d\} \,, \\
r_d &= \sup\{r > 0 : B_r \subset \Delta_d\} = \frac{1}{\sqrt{d}} \,, \\
R_d &= \inf\{R > 0 : \Delta_d \subset B_R\} = 1 \,.
\end{aligned}$$

Notice that, if $\Delta_d \subseteq Z \subseteq (1+\delta)\Delta_d$, then

$$\mathrm{Vol}(\Delta_d) \leq \mathrm{Vol}(Z) \leq \mathrm{Vol}((1+\delta)\Delta_d) = (1+\delta)^d \mathrm{Vol}(\Delta_d) \,,$$

which implies that

$$(1+\delta)^d \geq \frac{\mathrm{Vol}(Z)}{\mathrm{Vol}(\Delta_d)} \,.$$

Putting these observations together, we get that

$$\epsilon_d \geq \frac{1}{\sqrt{d}} \left[\inf \left\{ \frac{\mathrm{Vol}(Z)}{\mathrm{Vol}(\Delta_d)} : Z \in \mathcal{Z}^+,\, \Delta_d \subseteq Z \right\}^{\frac{1}{d}} - 1 \right] .$$

It was proven in [HLW10] that

$$\inf \left\{ \frac{\mathrm{Vol}(Z)}{\mathrm{Vol}(\Delta_d)} : Z \in \mathcal{Z}^+,\, \Delta_d \subseteq Z \right\} \geq \frac{d!}{\mathrm{maxdet}(d)} \,,$$

where $\mathrm{maxdet}(d) \doteq \max \left\{ \det(H) : H \in \{\pm 1\}^{d \times d} \right\} \leq d^{\frac{d}{2}}$. It follows that

$$\begin{aligned}
\epsilon_d &\geq \frac{1}{\sqrt{d}} \left[\left(\frac{d!}{d^{d/2}} \right)^{\frac{1}{d}} - 1 \right] \\
&\geq \frac{1}{\sqrt{d}} \left[\left(\frac{d^{d+1/2} e^{\frac{1}{12d+1}} \sqrt{2\pi}}{e^d d^{d/2}} \right)^{\frac{1}{d}} - 1 \right] \\
&= \frac{1}{\sqrt{d}} \left[d^{\frac{1}{2}+\frac{1}{2d}} e^{\frac{1}{d(12d+1)}-1} (2\pi)^{\frac{1}{2d}} - 1 \right] \\
&\geq \frac{1}{\sqrt{d}} \left[\frac{\sqrt{d}}{e} - 1 \right] = \frac{1}{e} - \frac{1}{\sqrt{d}} \,.
\end{aligned}$$

Notice that this bound is in fact vacuous for $d \leq 7$, while a priori it should only be for $d \leq 2$.

Bibliography

[ABMM16] Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

[ADH+19] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[AGNZ18] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[AH12] Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.

[AS20] Emmanuel Abbe and Colin Sandon. Poly-time universality and limitations of deep learning. *arXiv preprint arXiv:2001.02992*, 2020.

[AVP21] Terrence Alsup, Luca Venturi, and Benjamin Peherstorfer. Multilevel stein vari-

ational gradient descent with applications to bayesian inverse problems. *arXiv preprint arXiv:2104.01945*, 2021.

[AZLL18] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[Bac17a] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[Bac17b] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[BALPO17] Leon Bottou, Martin Arjovsky, David Lopez-Paz, and Maxime Oquab. Geometrical insights for implicit generative modeling. *arXiv preprint arXiv:1712.07822*, 2017.

[Bar93] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

[BBV16] Afonso S Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Conference on Learning Theory*, pages 361–382, 2016.

[BCN18] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[BGC17] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Massachusetts, USA:, 2017.

[BH89] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[BL91] Edward K Blum and Leong Kwan Li. Approximation theory and feedforward networks. *Neural networks*, 4(4):511–515, 1991.

[BLM89] Jean Bourgain, Joram Lindenstrauss, and Vitali Milman. Approximation of zonoids by zonotopes. *Acta mathematica*, 162(1):73–141, 1989.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BMP+12] Annalisa Buffa, Yvon Maday, Anthony T Patera, Christophe Prud'homme, and Gabriel Turinici. A priori convergence of the greedy algorithm for the parametrized reduced basis method. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 46(3):595–603, 2012.

[BN20] Guy Bresler and Dheeraj Nagaraj. Sharp representation theorems for relu networks with precise dependence on depth. *arXiv preprint arXiv:2006.04048*, 2020.

[Bra98] Martin L Brady. A fast discrete approximation algorithm for the radon transform. *SIAM Journal on Computing*, 27(1):107–119, 1998.

[BS14] Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25(8):1553–1565, 2014.

[Bur59] John Charles Burkill. *Lectures on approximation by polynomials*, volume 16. Tata Institute of Fundamental Research, 1959.

[BVB16] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.

[BVB21] Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learningwith geometric stability. *preprint*, 2021.

[BWJ19] Christian Beck, E Weinan, and Arnulf Jentzen. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *Journal of Nonlinear Science*, 29(4):1563–1619, 2019.

[BZL20] Kaifeng Bu, Yaobo Zhang, and Qingxian Luo. Depth-width trade-offs for neural networks via topological entropy. *arXiv preprint arXiv:2010.07587*, 2020.

[C+47] Augustin Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[CB00] Gerald HL Cheang and Andrew R Barron. A better approximation for balls. *Journal of Approximation Theory*, 104(2):183–203, 2000.

[CB18] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

[CGLM08] Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

[CHM+15] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann

LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.

[CL55] Earl A Coddington and Norman Levinson. *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.

[CLQY20] Pierre Comon, Lek-Heng Lim, Yang Qi, and Ke Ye. Topology of tensor ranks. *Advances in Mathematics*, 367:107128, 2020.

[CLWY18] Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Theoretical linear convergence of unfolded ista and its practical weights and thresholds. *arXiv preprint arXiv:1808.10038*, 2018.

[CNPW19] Vaggos Chatziafratis, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. Depth-width trade-offs for relu networks via sharkovsky's theorem. *arXiv preprint arXiv:1912.04378*, 2019.

[COB18] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[CWT+15] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. In *International conference on machine learning*, pages 2285–2294. PMLR, 2015.

[CWZZ18] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136, 2018.

[Cyb89] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[Dan17a] Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, pages 690–696. PMLR, 2017.

[Dan17b] Amit Daniely. Sgd learns the conjugate kernel class of the network. *arXiv preprint arXiv:1702.08503*, 2017.

[DDDM04] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[DDF+19] Ingrid Daubechies, Ronald DeVore, Simon Foucart, Boris Hanin, and Guergana Petrova. Nonlinear approximation and (deep) relu networks. *arXiv preprint arXiv:1905.02199*, 2019.

[DFS16] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

[DFT16] Feng Dai, Han Feng, and Sergey Tikhonov. Reverse hölder's inequality for spherical harmonics. *Proceedings of the American Mathematical Society*, 144(3):1041–1051, 2016.

[DL89] Ronald J DiPerna and Pierre-Louis Lions. Ordinary differential equations, transport theory and sobolev spaces. *Inventiones mathematicae*, 98(3):511–547, 1989.

[DL18] Simon S Du and Jason D Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206*, 2018.

[DLL+19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient de-

scent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[DLT+17] Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *arXiv preprint arXiv:1712.00779*, 2017.

[DVSH18] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A Hamprecht. Essentially no barriers in neural network energy landscape. *arXiv preprint arXiv:1803.00885*, 2018.

[DX13] Feng Dai and Yuan Xu. *Approximation theory and harmonic analysis on spheres and balls*, volume 23. Springer, 2013.

[EEJ04] Kenneth Eriksson, Donald Estep, and Claes Johnson. Piecewise linear approximation. In *Applied Mathematics: Body and Soul*, pages 741–753. Springer, 2004.

[EF14] Costas Efthimiou and Christopher Frye. *Spherical harmonics in p dimensions*. World Scientific, 2014.

[ELMV20] Virginie Ehrlacher, Damiano Lombardi, Olga Mula, and François-Xavier Vialard. Nonlinear model reduction on metric spaces. application to one-dimensional conservative pdes in wasserstein spaces. *ESAIM. Mathematical Modelling and Numerical Analysis*, 54, 2020.

[ES16] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

[Eva98] Lawrence C Evans. Partial differential equations. *Graduate studies in mathematics*, 19(2), 1998.

[FB17] Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *ICLR 2017*, 2017.

[FJZT17] Soheil Feizi, Hamid Javadi, Jesse Zhang, and David Tse. Porcupine neural networks:(almost) all local optima are global. *arXiv preprint arXiv:1710.02196*, 2017.

[GHJVW18] Philipp Grohs, Fabian Hornung, Arnulf Jentzen, and Philippe Von Wurstemberger. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of black-scholes partial differential equations. *arXiv preprint arXiv:1809.02362*, 2018.

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[GL10] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pages 399–406, 2010.

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[GPR+20] Moritz Geist, Philipp Petersen, Mones Raslan, Reinhold Schneider, and Gitta Kutyniok. Numerical solution of the parametric diffusion equation by deep neural networks. *arXiv preprint arXiv:2004.12131*, 2020.

[GRK20] Ingo Gühring, Mones Raslan, and Gitta Kutyniok. Expressivity of deep neural networks. *arXiv preprint arXiv:2007.04759*, 2020.

[GU19] Constantin Greif and Karsten Urban. Decay of the kolmogorov n-width for wave problems. *Applied Mathematics Letters*, 96:216–222, 2019.

[Han19] Boris Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10):992, 2019.

[HJW18] Jiequn Han, Arnulf Jentzen, and E Weinan. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[HLW10] Martin Henk, Eva Linke, and Jörg M Wills. Minimal zonotopes containing the crosspolytope. *Linear Algebra and its Applications*, 432(11):2942–2952, 2010.

[HM16] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

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

[HRS+16] Jan S Hesthaven, Gianluigi Rozza, Benjamin Stamm, et al. *Certified reduced basis methods for parametrized partial differential equations*, volume 590. Springer, 2016.

[HS17] Boris Hanin and Mark Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.

[HSSVG21] Daniel Hsu, Clayton Sanford, Rocco A Servedio, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. *arXiv preprint arXiv:2102.02336*, 2021.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[JJL17] Yuling Jiao, Bangti Jin, and Xiliang Lu. Iterative soft/hard thresholding with homotopy continuation for sparse recovery. *IEEE Signal Processing Letters*, 24(6):784–788, 2017.

[JNG+19] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv preprint arXiv:1902.04811*, 2019.

[JNS19] Shirin Jalali, Carl Nuzman, and Iraj Saniee. Efficient deep learning of gmms. *arXiv preprint arXiv:1902.05707*, 2019.

[Kaw16] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[KB18] Jason M Klusowski and Andrew R Barron. Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

[Kol56] Andreui Nikolaevich Kolmogorov. The representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Doklady Akademii Nauk SSSR*, 108(2):179–182, 1956.

[KPRS19] Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider. A theoretical analysis of deep neural networks and parametric pdes. *arXiv preprint arXiv:1904.00377*, 2019.

[KTB19] Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. *arXiv preprint arXiv:1905.12207*, 2019.

[KV20] Dan Kushnir and Luca Venturi. Diffusion-based deep active learning. *arXiv preprint arXiv:2003.10339*, 2020.

[LB18] Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pages 2902–2907. PMLR, 2018.

[LC20] Kookjin Lee and Kevin T Carlberg. Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders. *Journal of Computational Physics*, 404:108973, 2020.

[Liu17] Qiang Liu. Stein variational gradient descent as gradient flow. *arXiv preprint arXiv:1704.07520*, 2017.

[LK17] Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

[LL18] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *arXiv preprint arXiv:1808.01204*, 2018.

[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.

[LP21] Fabian Laakmann and Philipp Petersen. Efficient approximation of solutions of parametric linear transport equations by relu dnns. *Advances in Computational Mathematics*, 47(1):1–32, 2021.

[LS16] Shiyu Liang and Rayadurgam Srikant. Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161*, 2016.

[LSSS14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

[LT10] Joseph M Landsberg and Zach Teitler. On the ranks and border ranks of symmetric tensors. *Foundations of Computational Mathematics*, 10(3):339–366, 2010.

[LW16] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

[Mha96] Hrushikesh N Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*, 8(1):164–177, 1996.

[MJSSS21] Eran Malach, Gilad Jehudai, Shai Shalev-Shwartz, and Ohad Shamir. The connection between approximation, depth separation and learnability in neural networks. *arXiv preprint arXiv:2102.00434*, 2021.

[MM00] VE Maiorov and Ron Meir. On the near optimality of the stochastic approximation of smooth functions by neural networks. *Advances in Computational Mathematics*, 13(1):79–103, 2000.

[MM18] Marco Mondelli and Andrea Montanari. On the connection between learning two-layers neural networks and tensor decomposition. *arXiv preprint arXiv:1802.07301*, 2018.

[MMM21] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. *arXiv preprint arXiv:2102.13219*, 2021.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[MPCB14] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

[MSS19] Eran Malach and Shai Shalev-Shwartz. Is deeper better only when shallow is good? In *Advances in Neural Information Processing Systems*, pages 6426–6435, 2019.

[NBS17] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[Nes98] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5, 1998.

[Ngu21] Quynh Nguyen. A note on connectivity of sublevel sets in deep learning. *arXiv preprint arXiv:2101.08576*, 2021.

[NH17] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

[NPOV15] Alexander Novikov, Dmitry Podoprikhin, Anton Osokin, and Dmitry Vetrov. Tensorizing neural networks. *arXiv preprint arXiv:1509.06569*, 2015.

[OR15] Mario Ohlberger and Stephan Rave. Reduced basis methods: Success, limitations and future challenges. *arXiv preprint arXiv:1511.02021*, 2015.

[OS20] Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

[OWSS19] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

[Par15] Matthew David Parno. *Transport maps for accelerated Bayesian computation*. PhD thesis, Massachusetts Institute of Technology, 2015.

[PB17] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning*, pages 2798–2806, 2017.

[Pet20] Philipp Christian Petersen. Neural network theory, 2020.

[Pin99] Allan Pinkus. Approximation theory of the mlp model. *Acta Numerica 1999: Volume 8*, 8:143–195, 1999.

[Pin12] Allan Pinkus. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

[PLR+16] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

[PMB13] Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*, 2013.

[PMR+17] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

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

[RBA+19] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[Riv81] Theodore J Rivlin. *An introduction to the approximation of functions*. Courier Corporation, 1981.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[RM15] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

[RPK+17] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl Dickstein. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2847–2854. JMLR. org, 2017.

[RPK19] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems

involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[RPM19] Donsub Rim, Benjamin Peherstorfer, and Kyle T Mandli. Manifold approximations via transported subspaces: Model reduction for transport-dominated problems. *arXiv preprint arXiv:1912.13024*, 2019.

[RR+07] Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, page 5. Citeseer, 2007.

[RT17] David Rolnick and Max Tegmark. The power of deeper networks for expressing natural functions. *arXiv preprint arXiv:1705.05502*, 2017.

[Rub98] Boris Rubin. Inversion of fractional integrals related to the spherical radon transform. *journal of functional analysis*, 157(2):470–487, 1998.

[RVBP20] Donsub Rim, Luca Venturi, Joan Bruna, and Benjamin Peherstorfer. Depth separation for reduced deep networks in nonlinear model reduction: Distilling shock waves in nonlinear hyperbolic problems. *arXiv preprint arXiv:2007.13977*, 2020.

[RVE18a] Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22, 2018.

[RVE18b] Grant M Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7146–7155, 2018.

[SC16] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training

error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[Sch14] Rolf Schneider. *Convex bodies: the Brunn–Minkowski theory*. Cambridge university press, 2014.

[SCP16] Grzegorz Swirszcz, Wojciech Marian Czarnecki, and Razvan Pascanu. Local minima in training of neural networks. *arXiv preprint arXiv:1611.06310*, 2016.

[SES19] Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: What is actually being separated? *arXiv preprint arXiv:1904.06984*, 2019.

[SJL17] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

[SS16] Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pages 774–782, 2016.

[SS17a] Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2979–2987. JMLR. org, 2017.

[SS17b] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[SS18] Justin Sirignano and Konstantinos Spiliopoulos. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 375:1339–1364, 2018.

[SYS20] Itay Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. *arXiv preprint arXiv:2006.01005*, 2020.

[SZ19] Christoph Schwab and Jakob Zech. Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in uq. *Analysis and Applications*, 17(01):19–55, 2019.

[Tel16] Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

[TKB19] Matthew Trager, Kathlén Kohn, and Joan Bruna. Pure and spurious critical points: a geometric study of linear networks. *arXiv preprint arXiv:1910.01671*, 2019.

[Tre19] Lloyd N Trefethen. *Approximation Theory and Approximation Practice, Extended Edition*. SIAM, 2019.

[TVE+10] Esteban G Tabak, Eric Vanden-Eijnden, et al. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[VBB19] Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133):1–34, 2019.

[VJOB21] Luca Venturi, Samy Jelassi, Tristan Ozuch, and Joan Bruna. Depth separation beyond radial functions. *arXiv preprint arXiv:2102.01621*, 2021.

[VRPS21] Gal Vardi, Daniel Reichman, Toniann Pitassi, and Ohad Shamir. Size and depth separation in approximating natural functions with neural networks. *arXiv preprint arXiv:2102.00314*, 2021.

[Wei76] Wolfgang Weil. Centrally symmetric convex bodies and distributions. *Israel Journal of Mathematics*, 24(3):352–367, 1976.

[Wel17] Gerrit Welper. Interpolation of functions with parameter dependent jumps by transformed snapshots. *SIAM Journal on Scientific Computing*, 39(4):A1225–A1250, 2017.

[Wel20] G Welper. Transformed snapshot interpolation with high resolution transforms. *SIAM Journal on Scientific Computing*, 42(4):A2037–A2061, 2020.

[WHR19] Qian Wang, Jan S Hesthaven, and Deep Ray. Non-intrusive reduced order modeling of unsteady flows using artificial neural networks with application to a combustion problem. *Journal of computational physics*, 384:289–307, 2019.

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

[YS19] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*, 2019.

[YSJ18] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. *arXiv preprint arXiv:1802.03487*, 2018.

[YSW95] Joseph E Yukich, Maxwell B Stinchcombe, and Halbert White. Sup-norm approximation bounds for networks through probabilistic methods. *IEEE Transactions on Information Theory*, 41(4):1021–1027, 1995.

[Zha19] Li Zhang. Depth creates no more spurious local minima. *arXiv preprint arXiv:1901.09827*, 2019.

[ZL17] Yi Zhou and Yingbin Liang. Critical points of neural networks: Analytical forms and landscape properties. *arXiv preprint arXiv:1710.11205*, 2017.