Journal of Machine Learning Research 20 (2019) 1-34 Submitted 9/18; Revised 7/19; Published 8/19

Spurious Valleys in One-hidden-layer Neural Network Optimization Landscapes

Luca Venturi VENTURI@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012

Afonso S. Bandeira BANDEIRA@CIMS.NYU.EDU
Joan Bruna BRUNA@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012
Center for Data Science, 60 5th Avenue, New York, NY 10011

Editor: Animashree Anandkumar

Abstract

Neural networks provide a rich class of high-dimensional, non-convex optimization problems. Despite their non-convexity, gradient-descent methods often successfully optimize these models. This has motivated a recent spur in research attempting to characterize properties of their loss surface that may explain such success.

In this paper, we address this phenomenon by studying 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, we identify a notion of intrinsic dimension and show that it provides necessary and sufficient conditions for the absence of spurious valleys. More concretely, finite intrinsic dimension guarantees that for sufficiently overparametrised models no spurious valleys exist, independently of the data distribution. Conversely, infinite intrinsic dimension implies that spurious valleys do exist for certain data distributions, independently of model overparametrisation. Besides these positive and negative results, we show that, although spurious valleys may exist in general, they are confined to low risk levels and avoided with high probability on overparametrised models.

1. Introduction

Modern machine learning applications involve datasets of increasing dimensionality, complexity and size, which in turn motivate the use of high-dimensional, non-linear models, as illustrated in many deep learning algorithms across computer vision, speech and natural language understanding. The prevalent strategy for learning is to rely on Stochastic Gradient Descent (SGD) methods, that typically operate on non-convex objectives. In this context, an outstanding goal is to provide a theoretical framework that explains under what conditions – relating input data distribution, choice of architecture and choice of optimization scheme – this setup will be successful.

More precisely, let $\mathbf{\Phi}_{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}^m$ denote a model class parametrized by $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^P$, which in the case of Neural Networks (NNs) contains the aggregated weights across all layers. In a supervised learning setting, this model is deployed on some data $(\mathbf{X}, \mathbf{Y})$ random variable

taking values in $\mathbb{R}^n \times \mathbb{R}^m$, to predict targets $\mathbf{Y}$ given input $\mathbf{X}$, and its risk for a given $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim P}[\ell(\boldsymbol{\Phi}_{\boldsymbol{\theta}}(\mathbf{X}), \mathbf{Y})] \tag{1}$$

where ℓ is a convex loss, such as a square loss or a logistic regression loss. In the following we refer to (1) as the risk, the energy or the loss interchangeably. The aim is to find $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$ and this is attempted in practice by running SGD iteration. Under some technical conditions, the expected gradient is known to converge to zero (Bottou et al., 2016). Understanding the nature of such stationary points - and therefore the landscape of the loss function - is a task of fundamental importance to understand performance of SGD.

Whereas there is a growing literature in analyzing the behavior of SGD on non-convex objectives (Soudry et al., 2017; Ji and Telgarsky, 2018; Gunasekar et al., 2018; Wilson et al., 2017), we focus here on properties of the optimization problem above that are algorithm independent. A common factor shared in the above cited works (and in common practice) is that *overparametrisation* of the model class (i.e. $P \gg 1$) often leads to improved performance, despite the potential increase in generalization error.

Our analysis focuses mostly on the class of one-hidden-layer neural networks, with a hidden layer of size p, 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.

Summary of contributions More specifically, we show that:

- For Empirical Risk Minimization 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, 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).

Related works A considerable amount of literature has attempted to characterize the landscape of the loss function (1) by studying its critical points. Global optimality results have been obtained for NN architectures with linear activations (Hardt and Ma, 2016; Kawaguchi, 2016; Yun et al., 2018), quadratic activations (Soltanolkotabi et al., 2017; Du and Lee, 2018) and some more general non-linear activations, under appropriate regularity assumptions (Soudry and Carmon, 2016; Nguyen and Hein, 2017; Feizi et al., 2017). Some other insights have been obtained by leveraging tools for complexity analysis of spin glasses

(Choromanska et al., 2015) and random matrix theory (Pennington and Bahri, 2017). Other analysis involved studying goodness of the initialization of the parameter values $\boldsymbol{\theta}_0$ (Daniely et al., 2016; Safran and Shamir, 2016; Du et al., 2017) or other topological properties of the loss (1), such as connectivity of sub-level sets (Draxler et al., 2018; Freeman and Bruna, 2017).

Several other type of analysis of the convergence of NNs gradient-based optimization algorithms have been considered in the literature. For example, (Ge et al., 2017b) proved convergence of GD on a modified loss; (Shamir, 2018) compared optimization properties of residual networks with respect to linear models; in (Dauphin et al., 2014) it is argued that the issues arising in the optimization of NN architectures are due to the presence of saddle points in the loss function rather than spurious local minima. Optimization landscapes have also been studied in other contexts than from NNs training, such as non-convex low rank problems (Ge et al., 2017a), matrix completion (Ge et al., 2016), problems arising in semidefinite programming (Boumal et al., 2016; Bandeira et al., 2016) and implicit generative modeling (Bottou et al., 2017).

Structure of the paper The rest of the paper is structured as follows. Section 2 formally introduces the notion of spurious valleys and explains why this is a relevant concept from the optimization point of view. It also defines the intrinsic dimensions of a network (Section 2.2). In Section 3 we state our main positive results (Theorem 8) and we discuss two settings where they bear fruit: polynomial activation functions and empirical risk minimization. Section 4 is dedicated to constructions of worst case scenarios for activation with infinite lower intrinsic dimension. We then show, in Section 5, that, even if spurious valleys may exist, they tend to be confined to regimes of low risk. Some conclusive discussion is reported in Section 6.

1.1. Notation

We introduce notation we use throughout the rest of the paper. For any integers $n \leq m$ we denote $[n, m] = \{n, n+1, \dots, m\}$ and, if $n > 0$, $[n] = [1, n]$. We denote scalar valued variables as lowercase non-bold; vector valued variables as lowercase bold; matrix and tensor valued variables and multivariate random variables (r.v.'s) as uppercase bold. Given a vector $\mathbf{v} \in \mathbb{R}^n$, we denote its components as v_i; given a matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$, we denote its rows as $\mathbf{w}_i$; given a tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times \cdots \times n_k}$, we denote its components as $T_{i_1 \cdots i_k}$. Given some vectors $\mathbf{v}_i \in \mathbb{R}^{n_i}$, $i \in [k]$, the tensor product $\mathbf{v}_1 \otimes \cdots \otimes \mathbf{v}_k$ denotes the $n_1 \times \cdots \times n_k$ dimensional tensor $\mathbf{T}$ whose components are given by $T_{i_1 \cdots i_k} = v_{i_1} \cdots v_{i_k}$; given a vector $\mathbf{v}$, we denote $\mathbf{v}^{\otimes k} = \otimes_{i=1}^{k} \mathbf{v}$. We denote by $\mathrm{S}^k(\mathbb{R}^n)$ the space of order k symmetric tensors on $\mathbb{R}^n$. For any $\mathbf{T} \in \mathrm{S}^k(\mathbb{R}^n)$, we define the symmetric rank (Comon et al., 2008) as $\mathrm{rk}_{\mathrm{S}}(\mathbf{T}) = \min\Big\{p \geq 1 : \mathbf{T} = \sum_{i=1}^{p} u_i \mathbf{w}_i^{\otimes k} \text{ for some } \mathbf{u} \in \mathbb{R}^p, \mathbf{w}_1, \dots, \mathbf{w}_p \in \mathbb{R}^n\Big\}$. We define $\mathrm{rk}_{\mathrm{S}}(k, n) = \max\{\mathrm{rk}_{\mathrm{S}}(\mathbf{T}) : \mathbf{T} \in \mathrm{S}^k(\mathbb{R}^n)\}$. Finally, $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ denotes the $(n-1)$-dimensional sphere $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$.

2. Preliminaries

2.1. Problem setup

Let $(\mathbf{X}, \mathbf{Y})$ be two r.v.'s. These r.v.'s take values in $\mathbb{R}^n$ 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

$$L(\boldsymbol{\theta}) \doteq \mathbb{E}[\ell(\boldsymbol{\Phi}(\mathbf{X}; \boldsymbol{\theta}), \mathbf{Y})] \tag{2}$$

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 $\boldsymbol{\Phi}(\cdot; \boldsymbol{\theta}) : \mathbb{R}^n \to \mathbb{R}^m$ models the dependence of the output on the input as $\mathbf{Y} \simeq \boldsymbol{\Phi}(\mathbf{X}; \boldsymbol{\theta})$. We focus on one-hidden-layer NN functions $\boldsymbol{\Phi}$, i.e. $\boldsymbol{\Phi}$ of the form

$$\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x}) \tag{3}$$

where $\boldsymbol{\theta} = (\mathbf{U}, \mathbf{W}) \in \Theta \doteq \mathbb{R}^{m \times p} \times \mathbb{R}^{p \times n}$. Here p represents the width of the hidden layer and $\sigma : \mathbb{R} \to \mathbb{R}$ is a continuous element-wise *activation* function.

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

Definition 1 *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 (2) 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 (Freeman and Bruna, 2017), 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 Appendix E).

Lemma 2 *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.*

In the following, we prove absence of spurious valleys by proving that property **P.1** holds. Intuitively, we 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 positive 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 (2) 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.

2.2. Intrinsic dimension of a network

The main result of this work is to exploit that the property of absence of spurious valleys is related to the complexity of the functional space $V_\sigma = \{f = \Phi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ 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 $\mathbf{v} \in \mathbb{R}^n$ we denote $\psi_{\sigma,\mathbf{v}}$ to be the function $\psi_{\sigma,\mathbf{v}} : \mathbf{x} \in \mathbb{R}^n \mapsto \sigma(\langle \mathbf{v}, \mathbf{x} \rangle) \in \mathbb{R}$. We refer to each $\psi_{\sigma,\mathbf{v}}$ as a *filter* function. If $\mathbf{X}$ is a r.v. taking values in $\mathbb{R}^n$, we denote by $L^2_{\mathbf{X}}$ the space of square integrable function on $\mathbb{R}^n$ w.r.t. the probability measure induced by the r.v. $\mathbf{X}$. We then define the two following functional spaces:

$$V_{\sigma,p} = \left\{ f = \Phi(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} = (\mathbf{u}^T, \mathbf{W}) \in \Theta = \mathbb{R}^p \times \mathbb{R}^{p \times n} \right\}$$
$$\mathcal{R}_2(\sigma, n) = \left\{ \mathbf{X} \text{ r.v. taking values in } \mathbb{R}^n : \psi_{\sigma,\mathbf{v}} \in L^2_{\mathbf{X}} \text{ for every } \mathbf{v} \in \mathbb{R}^n \right\}$$

$V_{\sigma,p}$ represents the space of (one-dimensional output) functions modeled by the network architecture and $\mathcal{R}_2(\sigma, n)$ to be the space of (n-dimensional) input data distributions for which the filter functions have finite second moment. We finally define

$$V_\sigma = \operatorname{span}(\{f : f \in V_{\sigma,1}\}) = \bigcup_{p=1}^{\infty} V_{\sigma,p}$$

as the linear space spanned by the functions $\psi_{\mathbf{v},\sigma}$ for $\mathbf{v} \in \mathbb{R}^n$.

Definition 3 *Let σ be a continuous activation function and $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$ a r.v. We define*[1]

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

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 $L^2_{\mathbf{X}}$.

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

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 $\mathbf{X}$ is discrete, i.e. it is concentrated on a finite number of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^n$, 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 n level upper intrinsic dimension $\dim^*(\sigma, n)$ is defined as the dimension of the functional linear space V_σ. We note that if $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$ is a r.v. with almost surely (a.s.) positive density w.r.t. the Lebesgue measure dx, then $\dim^*(\sigma, n) = \dim^*(\sigma, \mathbf{X})$.

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

Lemma 4 *Let* σ *be a continuous activation function and* $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$ *such that* $\dim\big(L^2_{\mathbf{X}}\big) = \infty$. *If* $\sigma(z) = \sum_{k=0}^d a_k z^k$ *is a polynomial, then*

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

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 universal approximation theorem (Leshno et al., 1993). 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 V_σ'.

Definition 5 *Let* σ *be a continuous activation function and* $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$ *a r.v. We define*[2]

$$\dim_*(\sigma, \mathbf{X}) = \max\Big\{p \geq 1 \ : \ V_{\sigma,p-1} \subsetneq_{L^2_{\mathbf{X}}} V_{\sigma,p}\Big\}$$

as the lower dimension of the pair $(\sigma, \mathbf{X})$. *We define the level* n *lower dimension of* σ *as* $\dim_*(\sigma, n) = \max\{p \geq 1 \ : \ V_{\sigma,p-1} \subsetneq V_{\sigma,p}\} = \sup\{\dim_*(\sigma, \mathbf{X}) : \mathbf{X} \in \mathcal{R}_2(\sigma, n)\}$.

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 V_σ with the NN architecture (3). Clearly, this implies that

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

for every continuous activation function σ and any $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$. As with the upper instrinsic dimension, we note that if $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$ is a r.v. with a.s. positive density w.r.t. the Lebesgue measure dx, then $\dim_*(\sigma, n) = \dim_*(\sigma, \mathbf{X})$.

In the case of homogeneus polynomial activations $\sigma(z) = z^k$ with $k \geq 1$ integer, the level n lower dimension of σ coincides with the notion of (maximal) symmetric tensor rank.

2. For any subsets $V, W \subseteq L^2_{\mathbf{X}}$, we say that $V \subsetneq_{L^2_{\mathbf{X}}} W$ if $V \subsetneq W$ as subsets of $L^2_{\mathbf{X}}$ (and similar with other inclusions or equalities).

Lemma 6 *Let $\sigma(z) = z^k$, with k positive integer. Then*

$$\dim_*(\sigma, n) = \mathrm{rk}_{\mathrm{S}}(k, n)$$

Finally, the next lemma implies that for most non-polynomial activation functions practical interest, the lower intrinsic dimension $\dim_*(\sigma, n)$ is infinite.

Lemma 7 *Let σ be a continuous activation function such that $\sigma \in L^2(\mathbb{R}, e^{-x^2/2}\,dx)$ and $n > 1$. Then $\dim_*(\sigma, n) = \infty$ if and only if σ is not a polynomial.*

The proof of the above Lemma is based on Hermite decomposition and on the correspondence between one-hidden-layer nets and symmetric tensors (Mondelli and Montanari, 2018).

3. Finite intrinsic dimension and absence of spurious valleys

In this section we provide our positive results. Essentially they state that if the width of the network matches the dimension of the functional space V_σ spanned by its filter functions, then no spurious valleys exist. We first provide the main result (Theorem 8) in a general form, which allows a straight-forward derivation of two cases of interest: empirical risk minimization (Corollary 9) and polynomial activations (Corollary 10).

Theorem 8 *For any continuous activation function σ and r.v. $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$ 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 NNs $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$ admits no spurious valleys in the over-parametrised regime $p \geq \dim^*(\sigma, \mathbf{X})$.

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 $\left\{\mathbf{x} \mapsto \sigma(\langle \mathbf{w}_{1/2,i}, \mathbf{x}\rangle)\right\}_{i\in[p]}$ form a basis of V_σ. It follows that there exists a minimal function $\mathbf{f} \in V_\sigma^m \doteq \{(f_1, \dots, f_m) : f_i \in V_\sigma\}$, i.e.

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

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 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 V_σ, 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 2.2 describing the cases of finite intrinsic dimension, we immediately get the following corollaries.

Corollary 9 (ERM) *Consider N data points $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \subset \mathbb{R}^n \times \mathbb{R}^m$. For one-hidden-layer NNs $\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$, 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 $p \geq N$.

Comparison with existing results This results was already shown in (Livni et al., 2014). The only difference with our result is that we allow for rank degeneracy in the matrix $\sigma(\mathbf{W}[\mathbf{x}_1|\cdots|\mathbf{x}_N])$. 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, (Soudry and Carmon, 2016) showed that (almost surely) there exists no differentiable spurious minima if one of the layer weights $\mathbf{W}_i \in \mathbb{R}^{p_i \times p_{i-1}}$ satisfy $p_i p_{i-1} \geq N$. (Nguyen and Hein, 2017) showed that no spurious minima occur for multi-layer NNs 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 10 (Polynomial activations) *For one-hidden-layer NNs $\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$ with polynomial activation function $\sigma(z) = a_0 + a_1 z + \cdots + a_d z^d$, the loss function $L(\boldsymbol{\theta}) = \mathbb{E}[\ell(\mathbf{\Phi}(\mathbf{X}; \boldsymbol{\theta}), \mathbf{Y})]$ admits no spurious valleys in the over-parametrized regime*

$$p \geq \sum_{i=1}^{d} \binom{n+i-1}{i} \mathbf{1}_{\{a_i \neq 0\}} = O(n^d)$$

Under the hypothesis of Corollary 10 with $p = O(n^d)$, a generic function of V_σ, $\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{u}^T\sigma(\mathbf{W}\mathbf{x})$, can be also represented, for some $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\boldsymbol{\theta})$, in the generalized linear form

$$\mathbf{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \langle \boldsymbol{\gamma}, \boldsymbol{\varphi}(\mathbf{x}) \rangle$$

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

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

One would therefore like Corollary 10 to hold also (at least) for $p \geq O(n^{d-1})$. In the next section we address this problem for the linear activation $\sigma(z) = z$ and the quadratic activation $\sigma(z) = z^2$.

3.1. Improved over-parametrization bounds for homogeneous polynomial activations

The over-parametrization bounds obtained in Corollary 10 are quite non-desiderable in practical applications. We show that they can indeed be improved, for the case of linear and quadratic networks.

3.1.1. Linear networks case

Linear networks have been considered as a first order approximation of feed-forward multi-layers networks (Kawaguchi, 2016). It was shown, in several works (Kawaguchi, 2016; Freeman and Bruna, 2017; Yun et al., 2018), that, for linear networks of any depth

$$\mathbf{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{W}_{K+1}\cdots\mathbf{W}_1\mathbf{x} \tag{4}$$

with $\boldsymbol{\theta} = (\mathbf{W}_{K+1}, \mathbf{W}_K, \dots, \mathbf{W}_2, \mathbf{W}_1) \in \mathbb{R}^{m\times p_K} \times \mathbb{R}^{p_K\times p_{K-1}} \times \cdots \mathbb{R}^{p_2\times p_1} \times \mathbb{R}^{p_1\times n}$, the loss function (2) has no spurious local minima, if $\min_{i\in[K]} p_i \geq \min\{n, m\}$. This corresponds exactly with over-parametrization regime in Corollary 10, for the case of one-hidden-layer networks. The following theorem improves on Corollary 10 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 11 (Linear networks) *For linear NNs* (4) *of any depth* $K \geq 1$ *and of any layer widths* $p_k \geq 1$, $k \in [K]$, *and any input-output dimensions* $n, m \geq 1$, *the square loss function* $L(\boldsymbol{\theta}) = \mathbb{E}\|\mathbf{\Phi}(\mathbf{X};\boldsymbol{\theta}) - \mathbf{Y}\|^2$ *admits no spurious valleys.*

3.1.2. Quadratic networks case

Quadratic activations $\sigma(z) = z^2$ have been considered in the literature (Livni et al., 2014; Du and Lee, 2018; Soltanolkotabi et al., 2017) as second order approximation of general non-linear activations. Corollary 10 says that, if $p \geq n(n+1)/2$, the loss function (2) admits no spurious valleys. In the following theorem we relax the over-parametrisation requirement and show that $p > 2n$ is sufficient for the statement to hold, in the case of square loss functions and one dimensional output ($m = 1$).

Theorem 12 (Quadratic networks) *For one-hidden-layer NNs* $\Phi(\mathbf{x};\boldsymbol{\theta}) = \mathbf{u}^T\sigma(\mathbf{W}\mathbf{x})$ *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* $p \geq 2n + 1 = O(n)$.

Sketch of the proof The proof (reported in Section A) consists in constructing a path satisfying property **P.1** and improves upon the proof of Theorem 8 by leveraging the special linearized structure of the network for quadratic activation. For every parameter $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W}) \in \mathbb{R}^p \times \mathbb{R}^{p\times n}$, we can write

$$\Phi(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{p} u_i(\langle \mathbf{w}_i, \mathbf{x}\rangle)^2 = \Big\langle \sum_{i=1}^{p} u_i \mathbf{w}_i\mathbf{w}_i^T, \mathbf{x}\mathbf{x}^T \Big\rangle_F$$

We notice that $\Phi(\cdot;\boldsymbol{\theta})$ can also be represented by a NN $\Phi(\cdot;\hat{\boldsymbol{\theta}})$ with n hidden units; indeed, if $\sum_{i=1}^n \sigma_i\mathbf{v}_i\mathbf{v}_i^T$ is the SVD of $\sum_{i=1}^p u_i\mathbf{w}_i\mathbf{w}_i^T$, then $\Phi(\mathbf{x};\boldsymbol{\theta}) = \langle \sum_{i=1}^n \sigma_i\mathbf{v}_i\mathbf{v}_i^T, \mathbf{x}\mathbf{x}^T\rangle_F$. Therefore $p \geq n$ is sufficient to describe any element in V_σ. 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 should be able to extend the result to the regime $p \geq n$. 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 left for future work.

Comparison with previous works The same optimization landscape has been considered in the works (Soltanolkotabi et al., 2017) and (Du and Lee, 2018). In the first work, the authors show absence of spurious minima for the case of $p \geq 2n$ 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 $p \geq n$, if $n \leq N \leq O(n^2)$. This last condition can be removed by considering the regularized loss with non-zero weight decay, as shown in (Du and Lee, 2018); in the same work, the authors also proved absence of spurious minima in the case $p < n$ and $p(p+1) \geq 2N$ 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 $p > 2n$.

3.1.3. Lower to upper intrinsic dimension gap

As observed in Lemma 6 $\dim_*(\sigma(z) = z, n) = 1$ and $\dim_*(\sigma(z) = z^2, n) = n$ for all integer $n \geq 1$. Therefore, Theorem 11 and Theorem 12 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 $p \geq O(\dim_*(\sigma, n))$. We conjecture that this hold for any (sufficiently regular) activation function with finite intrinsic lower dimension.

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 13) 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.

Theorem 13 *Consider the square loss function $L(\boldsymbol{\theta}) = \mathbb{E}\|\boldsymbol{\Phi}(\mathbf{X}; \boldsymbol{\theta}) - \mathbf{Y}\|^2$ for one-hidden-layer NNs $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$ with non-negative activation function $\sigma \geq 0$ such that $\sigma \in L^2(\mathbb{R}, e^{-x^2}\, dx)$. If $p \leq \frac{1}{2}\dim_*(\sigma, n-1)$, then there exists a r.v. $(\mathbf{X}, \mathbf{Y})$ such that the square loss function L admits spurious valleys. In particular, for any given $M > 0$, the r.v. $\mathbf{Y}$ can be chosen in such a way that there exists a (non-empty) open set $\Omega \subset \Theta$ such that*

$$M/2 + \min_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}) \geq \sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}) \geq \min_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}) \geq M + \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) \tag{5}$$

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

$$\max_{t \in [0,1]} L(\boldsymbol{\theta}_t) \geq \min_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}) + M \tag{6}$$

Equation (5) in Theorem 13 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 (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 13, 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.

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

Corollary 14 (Homogeneous even degree polynomial activations) *Consider the case of activation $\sigma(z) = z^{2k}$ with $k \geq 1$ integer. For one-hidden-layer NNs $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$, if $n \geq 2$ and the hidden layer width satisfies*

$$p \leq \begin{cases} n-1 & \text{if } k = 1 \\ \frac{1}{2}\,\mathrm{rk_S}(2k, n-1) & \text{if } k > 1 \end{cases}$$

then there exists a r.v. $(\mathbf{X}, \mathbf{Y})$ 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 13 and Lemma 6, since $\dim_*(\sigma(z) = z^{2k}, n) = \mathrm{rk_S}(2k, n)$. For the well known case $k = 1$ (symmetric matrices) it holds $\mathrm{rk_S}(2, n) = n$; therefore Corollary 14 implies that the bound provided in Corollary 10 is almost (up to a factor 2) tight. Notice that our result is indeed in line with the results discussed in Section 3.1.2.

Corollary 15 (Spurious valleys exist in generic architectures) *If $n \geq 2$, for one-hidden-layer NNs $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$ with any hidden layer width $p \geq 1$ and continuous non-negative non-polynomial activation function $\sigma \in L^2(\mathbb{R}, e^{-x^2/2})$, then there exists a r.v. $(\mathbf{X}, \mathbf{Y})$ 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 setting includes the following activation functions:*

- *The ReLU activation function $\sigma(z) = z_+$ and some relaxations of it, such as softplus activation functions $\sigma(z) = \beta^{-1}\log\big(1 + e^{\beta z}\big)$, with $\beta > 0$;*
- *The sigmoid activation function $\sigma(z) = (1 + e^{-z})^{-1}$ and the approximating erf function $\sigma(z) = 2/\pi \int_0^z e^{-u}\,du$, which represents an approximation to the sigmoid function.*

This follows by Theorem 13 by observing that $\dim_*(\sigma, n) = \infty$ if σ is one of the above activation functions.

Discussion and comparison with previous works Several works showed existence of spurious minima: (Safran and Shamir, 2017) showed counterexamples under Gaussian input distributions, for $p = n - 1 \in \{8, \ldots, 19\}$, using a computer-assisted proof; (Swirszcz et al., 2016) and (Zhou and Liang, 2017) provided a few numerical examples; (Yun et al., 2018) showed existence of spurious minima for ReLU-like activations under non-realizability, and provided counterexamples for smooth activations. For any number of hidden neurons p, 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 (2). 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.

5. Typical spurious valleys and low-energy barriers

In the previous section it was shown that whenever the number of hidden units p 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?).

In this section, we study the energy landscape under generic data distributions in case of homogeneous activation, and show that, although spurious valleys may appear, they tend do so below a certain energy level, controlled by the decay of the spectral decomposition of the kernel defined by the activation function and by the amount of parametrisation p. This offers a first glimpse at the empirical success of local descent algorithms in conditions where p is indeed below the intrinsic dimension.

We consider oracle square loss functions of the form

$$L(\boldsymbol{\theta}) = \mathbb{E}|\Phi(\mathbf{X};\boldsymbol{\theta}) - Y|^2 \tag{7}$$

for one-dimensional output one-hidden-layer NNs $\Phi(\mathbf{x};\boldsymbol{\theta}) = \mathbf{u}^T\sigma(\mathbf{W}\mathbf{x})$, with $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W}) \in \mathbb{R}^p \times \mathbb{R}^{p\times n}$, σ a positively homogeneous function, and $\mathbf{X}, Y$ square integrable r.v. Notice that we can write

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

for some measurable $f^* : \mathbb{R}^n \to \mathbb{R}$ such that $f^*(\mathbf{X}) = \mathbb{E}[Y \,|\, \mathbf{X}]$. In particular this implies that

$$\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) \geq \mathcal{R}(\mathbf{X}, Y) \doteq \mathbb{E}|Y - f^*(\mathbf{X})|^2$$

If f^* can be written as a one-hidden-layer neural network with an arbitrary number of hidden units, that is

$$f^*(\mathbf{x}) = \int_{\mathbb{R}^n} \sigma(\langle \mathbf{x}, \mathbf{w}\rangle)\rho(\mathbf{w})\, d\mu(\mathbf{w})$$

for some measure μ and weight function ρ, then a possible approach to find a proper approximation of f^* is through random features sampling (Rahimi and Recht, 2008). Applying some recent results (Bach, 2017b) relating random features expansions with kernel quadrature rules, we show that this implies the following statement: *as the network width increases, spurious valleys tend to be confined to decreasingly low loss value.* In this regime, large loss barriers are therefore avoided with high probability over initialization of the parameters. The statement is made more rigorous in the following:

Theorem 16 *Let $d\tau$ be the uniform distribution over the unit sphere $\mathbb{S}^n$ and consider an initial parameter $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{u}}, \tilde{\mathbf{W}})$ with $\tilde{\mathbf{w}}_i \sim d\tau$ sampled i.i.d. Then the following hold:*

1. *There exists a path $t \in [0,1] \mapsto \boldsymbol{\theta}_t$ such that $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}$, the function $t \in [0,1] \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing, and*

$$L(\boldsymbol{\theta}_1) \leq \mathcal{R}(\mathbf{X}, Y) + \lambda \quad \text{if} \quad p \geq O\big(-\lambda^{-1}\log(\lambda\delta)\big)$$

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

2. *If f^* is sufficiently regular[3], there exists a path $t \in [0,1] \mapsto \boldsymbol{\theta}_t$ such that $\boldsymbol{\theta}_0 = \tilde{\boldsymbol{\theta}}$, the function $t \in [0,1] \mapsto L(\boldsymbol{\theta}_t)$ is non-increasing, and*

$$L(\boldsymbol{\theta}_1) \leq \mathcal{R}(\mathbf{X}, Y) + O(p^{-1+\delta})$$

with probability greater or equal then $1 - e^{-O(p^\delta)}$ for every $\delta \in (0,1)$.

Sketch of the proof Assume that f^* admits the representation

$$f^*(\mathbf{x}) = \int_\Theta \rho(\mathbf{w})\sigma(\langle \mathbf{x}, \mathbf{w}\rangle)\, d\tau(\mathbf{w})$$

for some density ρ. If $\mathbf{w}_i \sim d\tau$, $i \in [p]$, are drawn i.i.d., we have

$$\mathbb{E}\left(\frac{1}{p}\sum_{i=1}^{p}\rho(\mathbf{w}_i)\sigma(\langle \mathbf{w}_i, \mathbf{x}\rangle) - f^*(\mathbf{x})\right)^2 = O\left(\frac{1}{p}\right)$$

Notice that by only moving the second layer, we can construct a (linear) descent path from $(\tilde{\mathbf{u}}, \tilde{\mathbf{W}})$ to $(\mathbf{u}, \tilde{\mathbf{W}})$, where $u_i = \rho(\mathbf{w}_i)$. The proof is then concluded by applying an Hoeffding's-type inequality to get property 2. if it holds $\rho \in L^\infty(\mathbb{S}^n\, d\tau)$ or by applying Proposition 1 in (Bach, 2017b) to obtain property 1.

Related works Many recent works leveraged arguments based on random features to explain the empirical success of local descent algorithms to train neural networks (see e.g. (Jacot et al., 2018; Allen-Zhu et al., 2018; Oymak and Soltanolkotabi, 2019; Yehudai and Shamir, 2019; Ma et al., 2019; Du et al., 2018)). In Theorem 16, we used this type of technique to show properties of the optimization landscape. The main limitation shared by our and the cited results is the gap between the regimes in which the apply (high over-parametrized NNs) and the regimes attained in practice. A current important direction is to understand the dynamics of neural networks training over kernel approximation and to extend such results to *moderatly* over-parametrized architectures.

Remark 17 *Notice that in the previous description of the problem, we dropped bias terms from the neural network architectures for sake of simplicity, as we can immediately generalize to the biases case by stacking the bias in the weights and input random variables. As a bias term is needed in order to use universal approximation results, with abuse of notation, in the above theorem we wrote $\langle \mathbf{w}, \mathbf{x}\rangle \doteq \langle \mathbf{w}^{(n)}, \mathbf{x}\rangle + w_{n+1}$ for $\mathbf{w} \in \mathbb{S}^n, \mathbf{x} \in \mathbb{R}^n$, where $\mathbf{w}^{(n)} = (w_1, \ldots, w_n)$ represents a neuron weight and w_{n+1} a bias term; again, note that this can be done by simply considering the r.v. $\tilde{\mathbf{X}} \doteq (\mathbf{X}, 1)$ in place of $\mathbf{X}$.*

3. More precisely, if the function f^* can be written as $f^*(\mathbf{x}) = \int_{\mathcal{W}} g^*(\mathbf{w})\psi_{\mathbf{w}}(\mathbf{x})\, d\tau(\mathbf{w})$ for some $g^* \in L^\infty_{d\tau}$.

6. Future directions

We considered the problem of characterizing the loss surface of neural networks from the perspective of optimization, with the goal of deriving weak certificates that enable - or prevent - the existence of descent paths towards global minima.

The topological properties studied in this paper, however, do not yet capture fundamental aspects that are necessary to explain the empirical success of deep learning methods. We identify a number of different directions that deserve further attention.

The positive results presented above rely on being able to reduce the network to the case when (convex) optimization over the output layer is sufficient to reach optimal weight values. A better understanding of first layer dynamics needs to be carried out. Moreover, in such positive results we only proved non-existence of (high) energy barriers. While this is an interesting property from the optimization point of view, it is also not sufficient to guarantee convergence of local descent algorithms. Another informative property of the loss function that should be addressed in future works is the existence of local descents in non optimal points: for every $\boldsymbol{\theta}_0 \in \Theta$ non optimal and any neighborhood $\mathcal{U} \subseteq \Theta$ of $\boldsymbol{\theta}_0$, there exists $\boldsymbol{\theta} \in \mathcal{U}$ such that $L(\boldsymbol{\theta}) < L(\boldsymbol{\theta}_0)$. More generally, our present work is not informative on the performance of gradient descent in the regimes with no spurious valley.

The other very important point to be addressed in future is how to extend the above results to architectures of more practical interest. Depth and the specific linear structure of Convolutional Neural Networks, critical to explain the excellent empirical performance of deep learning in computer vision, text or speech, need to be exploited, as well as specific design choices such as Residual connections and several normalization strategies – as done recently in (Shamir, 2018) and (Santurkar et al., 2018) respectively. This also requires making specific assumptions on the data distribution, and is left for future work.

Acknowledgements We would like to thank Gérard Ben Arous and Léon Bottou for fruitful discussions, and Jean Ponce for valuable comments and corrections of the original version of this manuscript. LV would also like to thank Jumageldi Charyyev for fruitful discussions on the proofs of several propositions and Andrea Ottolini for valuable comments on a previous version of this manuscript. LV was partially supported by NSF grant DMS-1719545. ASB was partially supported by NSF grants DMS-1712730 and DMS-1719545, and by a grant from the Sloan Foundation. JB acknowledges the partial support by the Alfred P. Sloan Foundation, NSF RI-1816753, NSF CAREER CIF 1845360, and Samsung Electronics.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017a.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017b.

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.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2016.

Leon Bottou, Martin Arjovsky, David Lopez-Paz, and Maxime Oquab. Geometrical insights for implicit generative modeling. *arXiv preprint arXiv:1712.07822*, 2017.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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.

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.

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.

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.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A Hamprecht. Essentially no barriers in neural network energy landscape. *arXiv preprint arXiv:1803.00885*, 2018.

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.

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.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

Soheil Feizi, Hamid Javadi, Jesse Zhang, and David Tse. Porcupine neural networks:(almost) all local optima are global. *arXiv preprint arXiv:1710.02196*, 2017.

Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *ICLR 2017*, 2017.

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.

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*, 2017a.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017b.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

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

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

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.

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.

Chao Ma, Lei Wu, et al. A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics. *arXiv preprint arXiv:1904.04326*, 2019.

Marco Mondelli and Andrea Montanari. On the connection between learning two-layers neural networks and tensor decomposition. *arXiv preprint arXiv:1802.07301*, 2018.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.

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.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

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.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization?(no, it is not about internal covariate shift). *arXiv preprint arXiv:1805.11604*, 2018.

Ohad Shamir. Are resnets provably better than linear predictors? *arXiv preprint arXiv:1804.06739*, 2018.

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.

Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Grzegorz Swirszcz, Wojciech Marian Czarnecki, and Razvan Pascanu. Local minima in training of neural networks. *arXiv preprint arXiv:1611.06310*, 2016.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687*, 2019.

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.

Yi Zhou and Yingbin Liang. Critical points of neural networks: Analytical forms and landscape properties. *arXiv preprint arXiv:1710.11205*, 2017.

Appendix A. Proofs of Section 3

Notations For any r.v.'s $\mathbf{X}$ and $\mathbf{Y}$ with values in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, we denote $\mathbf{\Sigma_X} = \mathbb{E}\big[\mathbf{X}\mathbf{X}^T\big]$ and $\mathbf{\Sigma_{XY}} = \mathbb{E}\big[\mathbf{X}\mathbf{Y}^T\big]$. For every integer $n \geq 1$, we denote by $GL(n)$, $O(n)$ and $SO(n)$, respectively, the general linear group, the orthogonal group and the special orthogonal group of real $n \times n$ matrices. $\mathbf{I}$ denotes the identity matrix and $\mathbf{e}_1, \ldots, \mathbf{e}_n$ the standard basis in $\mathbb{R}^n$.

A.1. Proof of Theorem 8

We note that, under the assumptions of Theorem 8, the same optimal NN functions $\Phi_i(\cdot;\boldsymbol{\theta})$ 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 18 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function and $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$ a r.v. Assume that the linear space*

$$V_{\sigma,\mathbf{X}} \doteq \mathrm{span}(\{f : f \in V_{\sigma,1}\}) \subseteq L^2_{\mathbf{X}}$$

is finite dimensional. Then there exists a scalar product $\langle \cdot, \cdot \rangle$ on $V_{\sigma,\mathbf{X}}$ and a map $\mathbf{x} \in \mathbb{R}^n \mapsto \boldsymbol{\varphi}(\mathbf{x}) \in V_{\sigma,\mathbf{X}}$ such that

$$\langle \boldsymbol{\psi}_{\sigma,\mathbf{w}}, \boldsymbol{\varphi}(\mathbf{x})\rangle = \psi_{\sigma,\mathbf{w}}(\mathbf{x}) = \sigma(\langle \mathbf{w}, \mathbf{x}\rangle) \tag{8}$$

for all $\mathbf{w} \in \mathbb{R}^n$. Moreover, the function $\mathbf{w} \in \mathbb{R}^n \mapsto \boldsymbol{\psi}_{\sigma,\mathbf{w}} \in V_{\sigma,\mathbf{X}}$ is continuous.

Proof For sake of simplicity, in the following we write $\psi_{\mathbf{w}}$ for $\psi_{\sigma,\mathbf{w}}$ and V for $V_{\sigma,\mathbf{X}}$. Let $\boldsymbol{\psi}_{\mathbf{w}_1}, \ldots, \boldsymbol{\psi}_{\mathbf{w}_q}$ be a basis of V. If $\boldsymbol{\psi}_{\mathbf{w}} = \sum_{i=1}^q \alpha_i \boldsymbol{\psi}_{\mathbf{w}_i}$ and $\psi_{\mathbf{v}} = \sum_{j=1}^q \beta_j \boldsymbol{\psi}_{\mathbf{w}_j}$, then we can define a scalar product on V as

$$\langle \boldsymbol{\psi}_{\mathbf{w}}, \boldsymbol{\psi}_{\mathbf{v}}\rangle \doteq \sum_{i=1}^q \alpha_i \beta_i$$

If we define the map $\mathbf{x} \in \mathbb{R}^n \mapsto \boldsymbol{\varphi}(\mathbf{x}) \in V$ as

$$\boldsymbol{\varphi}(\mathbf{x}) = \sum_{i=1}^q \psi_{\mathbf{w}_i}(\mathbf{x}) \boldsymbol{\psi}_{\mathbf{w}_i}$$

then property (8) follows directly by the definition of the function $\psi_{\mathbf{w}}$. Moreover, we can choose $\mathbf{x}_1, \ldots, \mathbf{x}_q$ such that $\boldsymbol{\varphi}(\mathbf{x}_1), \ldots, \boldsymbol{\varphi}(\mathbf{x}_q)$ is a basis of V. Now we need to show that, for $i \in [q]$, the map $\mathbf{w} \mapsto \langle \boldsymbol{\psi}_{\mathbf{w}}, \boldsymbol{\psi}_{\mathbf{w}_i}\rangle$ is continuous. Let $\mathbf{M}$ be the matrix $\mathbf{M} \doteq (\psi_{\mathbf{w}_j}(\mathbf{x}_i))_{i,j} \in \mathbb{R}^{q\times q}$ 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 $\langle \boldsymbol{\psi}_{\mathbf{w}}, \boldsymbol{\psi}_{\mathbf{w}_i}\rangle = (\mathbf{M}^{-1}\mathbf{z}(\mathbf{w}))_i$, which is continuous in $\mathbf{w}$. This shows that the map $\mathbf{w} \in \mathbb{R}^n \mapsto \boldsymbol{\psi}_{\mathbf{w}} \in V$ is continuous. ■

The non-trivial fact captured by Theorem 8 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 (2), which is in general a highly non-convex object, satisfies an

interesting optimization property in view of the local descent algorithms which are used in practice to solve it.

Proof [Proof of Theorem 8] Thanks to Lemma 18, there exist two continuous maps $\boldsymbol{\varphi}, \boldsymbol{\psi} : \mathbb{R}^n \to \mathbb{R}^q \simeq V_{\sigma,\mathbf{X}}$, with $q = \dim^*(\sigma, \mathbf{X})$, such that $\sigma(\langle \mathbf{w}, \mathbf{x}\rangle) = \langle \boldsymbol{\psi}(\mathbf{w}), \boldsymbol{\varphi}(\mathbf{x})\rangle$ for every $\mathbf{w}, \mathbf{x} \in \mathbb{R}^n$. Therefore, every one-hidden-layer NN $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\sigma(\mathbf{W}\mathbf{x})$ can be written as $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\boldsymbol{\psi}(\mathbf{W})\boldsymbol{\varphi}(\mathbf{x})$, where, if $\mathbf{W} \in \mathbb{R}^{p\times n}$, then $\boldsymbol{\psi}(\mathbf{W}) \in \mathbb{R}^{p\times q}$ (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 $\mathrm{rk}(\boldsymbol{\psi}(\tilde{\mathbf{W}})) = q$ w.l.o.g. Let $\mathbf{w}_1^T, \dots, \mathbf{w}_p^T \in \mathbb{R}^n$ be the rows 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 $I = \{i_1, \dots, i_r\}$, $J = [1,p] \setminus I = \{j_1, \dots, j_{p-r}\}$ and $\mathbf{u}_1, \dots, \mathbf{u}_p$ the columns of $\tilde{\mathbf{U}}$. For $j \in J$, we can write

$$\boldsymbol{\psi}(\mathbf{w}_j) = \sum_{k=1}^{r} a_j^k \, \boldsymbol{\psi}(\mathbf{w}_{i_k}) \quad \text{for some } a_j^k \in \mathbb{R} \tag{9}$$

If we define $\mathbf{U}_1$ such that (denoting $\mathbf{u}_{1,i}$ the i-th row 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 $\mathbf{w}_{1,j_1}, \dots, \mathbf{w}_{1,j_{p-r}} \in \mathbb{R}^n$ such that the matrix $\mathbf{W}_1$ with rows $\mathbf{w}_{1,i} = \mathbf{w}_i$ 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 $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\boldsymbol{\psi}(\mathbf{W})\boldsymbol{\varphi}(\mathbf{x})$ 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. ■

A.2. Proof of Theorem 11

The first step for proving Theorem 11 consists in extending the result of Theorem 8 to the case of one-hidden-layer linear NNs $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{U}\mathbf{W}\mathbf{x}$ with $\mathbf{U} \in \mathbb{R}^{m\times p}$, $\mathbf{W} \in \mathbb{R}^{p\times n}$ with $p < n$

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 $\mathbf{G} \in GL(p)$ it holds that

$$\boldsymbol{\Phi}(\mathbf{x};(\mathbf{U},\mathbf{W})) = \mathbf{U}\mathbf{W}\mathbf{x} = (\mathbf{U}\mathbf{G}^{-1})(\mathbf{G}\mathbf{W})\mathbf{x} = \boldsymbol{\Phi}(\mathbf{x};(\mathbf{U}\mathbf{G}^{-1},\mathbf{G}\mathbf{W})) \tag{10}$$

This means that the map $\boldsymbol{\theta} \mapsto \boldsymbol{\Phi}(\cdot;\boldsymbol{\theta})$ is defined up to an action of the group $GL(p)$ over the parameter space $\Theta = \mathbb{R}^{m\times p} \times \mathbb{R}^{p\times n}$; 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 $\Theta/GL(p)$. We denote the orbit of an element $\boldsymbol{\theta} = (\mathbf{U},\mathbf{W}) \in \Theta$ as

$$[\boldsymbol{\theta}] = [\mathbf{U},\mathbf{W}] = \{\mathbf{G}\cdot\boldsymbol{\theta} = (\mathbf{U}\mathbf{G}^{-1},\mathbf{G}\mathbf{W}) : \mathbf{G} \in GL(p)\}$$

If g is a real-valued function defined on Θ such that $g(\mathbf{G}\cdot\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ for all $\mathbf{G} \in GL(p)$ and $\boldsymbol{\theta} \in \Theta$, then one can equivalently consider g as defined on $\Theta/GL(p)$ 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 8, 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 $\mathrm{rk}(\mathbf{W}_1) = p$ and $\mathbf{q} : \mathbb{R}^{p\times n} \to \mathbb{R}^{m\times p}$ the function defined by

$$\mathbf{q}(\mathbf{W}) = \boldsymbol{\Sigma}_{\mathbf{YX}}\mathbf{W}^T(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{W}^T)^\dagger \in \arg\min_{\mathbf{U}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=(\mathbf{U},\mathbf{W})}$$

(see Lemma 28). Therefore, let $\tilde{\boldsymbol{\theta}} = (\mathbf{q}(\tilde{\mathbf{W}}),\tilde{\mathbf{W}})$ with $\mathrm{rk}(\tilde{\mathbf{W}}) = p$, 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 $\mathrm{rk}(\mathbf{W}_t) = p$ 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 PSD matrix and, for every matrix $\mathbf{W}$, $\mathbf{P}_{\mathbf{W}}$ denotes the orthogonal projection on the rows of $\mathbf{W}$, that is $\mathbf{P}_{\mathbf{W}} = \mathbf{W}^\dagger\mathbf{W}$ (see Lemma 28). 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 $GL(p)$ on Θ. Since we look for $\mathbf{W}_t$ of rank p, we can consider f as defined on $G(p,n)$, the Grassmanian of p dimensional linear subspaces of $\mathbb{R}^n$. The proof below for the linear one-hidden-layer case is articulated as follows. We first construct a path $[\mathbf{W}_t] \in G(p,n)$ 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 (Lemma 19). We then show that such a path can be *lifted* to a corresponding path $\mathbf{W}_t \in \mathbb{R}^{p\times n}$ (Lemma 20). Finally, we show that we can drop the assumption $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}$ and the result still holds (Lemma 21).

Lemma 19 *Let $[\tilde{\mathbf{W}}] \in G(p,n)$ and assume $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}$. Then there exists a continuous path $t \in [0,1] \mapsto [\mathbf{W}_t] \in G(p,n)$ 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 While it is geometrically intuitive that the results should hold, we derive a constructive proof. We start by noticing that if $[\mathbf{W}] \in G(p,n)$ and $\mathbf{w}_1, \dots, \mathbf{w}_p$ is an orthonormal basis of $[\mathbf{W}]$, then

$$f[\mathbf{W}] = \sum_{i=1}^{p} \mathbf{w}_i^T \mathbf{M} \mathbf{w}_i \tag{11}$$

Moreover, if $\mathbf{M} = \sum_{j=1}^{n} \sigma_i \mathbf{v}_j \mathbf{v}_j^T$ is the SVD of $\mathbf{M}$, where $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, then (11) can be written as

$$f[\mathbf{W}] = \sum_{j=1}^{n} \sigma_j \sum_{i=1}^{p} \langle \mathbf{v}_j, \mathbf{w}_i \rangle^2$$

In particular the maximum of f is obtained for $[\mathbf{W}] = [\mathbf{V}] \doteq [\mathbf{v}_1, \dots, \mathbf{v}_p]$ (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 $i \in [p]$, with $[\mathbf{W}^0] = [\tilde{\mathbf{W}}]$, $[\mathbf{W}^p] = [\mathbf{V}]$ and

$$\mathbf{W}^i = [\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{w}_{i+1}^{i-1}, \dots, \mathbf{w}_p^{i-1}] \quad \text{for } i \in [p]$$

where $\mathbf{w}_1^j = \mathbf{v}_1, \dots, \mathbf{w}_j^j = \mathbf{v}_j, \mathbf{w}_{j+1}^j, \dots, \mathbf{w}_p^j$ is an orthonormal basis of $[\mathbf{W}^j]$, for $j \in [0,p]$. 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 $i \in [0, p-1]$ and consider

$$[\mathbf{W}^i] = [\mathbf{w}_1^i = \mathbf{v}_1, \dots, \mathbf{w}_i^i = \mathbf{v}_i, \mathbf{w}_{i+1}^i, \dots, \mathbf{w}_p^i]$$

We define

$$\mathbf{u}_{i+1}^i = \begin{cases} \frac{\mathbf{P}_{\mathbf{W}^i} \mathbf{v}_{i+1}}{\|\mathbf{P}_{\mathbf{W}^i} \mathbf{v}_{i+1}\|} & \text{if } \mathbf{P}_{\mathbf{W}^i} \mathbf{v}_{i+1} \neq \mathbf{0} \\ \mathbf{0} & \text{o.w.} \end{cases}$$

Then we complete $\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{u}_{i+1}^i$ to an orthonormal basis of $[\mathbf{W}^i]$:

$$\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{u}_{i+1}^i, \dots, \mathbf{u}_p^i$$

We call $\mathbf{w}_j^{i+1} = \mathbf{u}_j^i$ for $j \in [i+2, p]$ and we define

$$[\mathbf{W}^{i+1}] = [\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{w}_{i+1}^{i+1} = \mathbf{v}_{i+1}, \mathbf{w}_{i+2}^{i+1}, \dots, \mathbf{w}_p^{i+1}]$$

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 $S^{n-1} \subset \mathbb{R}^n$, i.e.

$$[\mathbf{W}_t^{i+1}] = [\mathbf{v}_1, \dots, \mathbf{v}_i, \mathbf{u}_{i+1}^i(t), \mathbf{u}_{i+2}^i, \dots, \mathbf{u}_p^i]$$

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 $\mu_{i+1} = \langle \mathbf{u}_{i+1}^i, \mathbf{v}_{i+1} \rangle$. 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

$$f[\mathbf{W}_t^{i+1}] - f[\mathbf{W}^i] = \sum_{j=i+1}^{n} \sigma_j \langle \mathbf{u}_{i+1}^i(t), \mathbf{v}_j \rangle^2$$

and by showing that the derivative of the RHS is greater or equal than 0. This concludes the proof of the lemma. ■

Lemma 20 *Let $\tilde{\mathbf{W}} \in \mathbb{R}^{p \times n}$ and assume $\mathbf{\Sigma}_X = \mathbf{I}$. Then there exists a continuous path $t \in [0,1] \mapsto \mathbf{W}_t \in \mathbb{R}^{p \times n}$ 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 $[\mathbf{W}_t^i] \in G(p,n)$ from the proof of Lemma 19 to continuous paths $\mathbf{W}_t^i \in \mathbb{R}^{n \times p}$. We first notice that if the basis $\{\mathbf{w}_1^i, \dots, \mathbf{w}_p^i\}$ and $\{\mathbf{w}_1^i, \dots, \mathbf{w}_i^i, \mathbf{u}_{i+1}^i, \dots, \mathbf{u}_p^i\}$ are defined as above, then we can assume (up to changing some signs) that they have all the same orientation, for all $i \in [0,p]$. Therefore we can define the matrices $\mathbf{W}^i \in \mathbb{R}^{p \times n}$ with rows $\mathbf{w}_1^i, \dots, \mathbf{w}_p^i$ and the matrices $U^i \in \mathbb{R}^{p \times n}$ with rows $\mathbf{w}_1^i, \dots, \mathbf{w}_i^i, \mathbf{u}_{i+1}^i, \dots, \mathbf{u}_p^i$, for $i \in [0,p]$. The paths $\mathbf{W}_t^{i+1}$ are defined in the same way as in the proof of Lemma 19. 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

$$\mathbf{O}^i = \mathbf{W}_i^T \mathbf{U}_i \in SO(n)$$

Notice that $\mathbf{W}^i \mathbf{O}^i = \mathbf{U}^i$. In particular there exist $\mathbf{A}^i$ real skew-symmetric such that $\mathbf{O}^i = e^{\mathbf{A}^i}$. Therefore the paths $t \in [0,1] \mapsto \mathbf{U}_t^i = \mathbf{W}^i e^{t\mathbf{A}^i}$ 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 $\tilde{\mathbf{W}} \in \mathbb{R}^{n \times p}$ with columns $\mathbf{w}_1, \dots, \mathbf{w}_p$, 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 $\tilde{\mathbf{W}} = \mathbf{O}\mathbf{\Lambda}\mathbf{U}$ is the SVD of $\tilde{\mathbf{W}}$ with $\mathbf{U} = \mathbf{W}^0$, $\mathbf{\Lambda} = \mathrm{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{p \times p}$ (with $\sigma_i > 0$, $i \in [p]$) and $\mathbf{O} \in SO(p)$, there exists $\mathbf{A}$ real skew-symmetric such that $\mathbf{O} = e^{\mathbf{A}}$. Thus the path $t \in [0,1] \mapsto \mathbf{W}_t = e^{(1-t)\mathbf{A}} \mathbf{\Lambda}^{1-t} \mathbf{W}^0$ 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 21 *Lemma 20 holds even if we drop the assumption $\mathbf{\Sigma_X} = \mathbf{I}$.*

Proof For sake of simplicity we distinguish two cases.
Case 1: $\mathrm{rk}(\mathbf{\Sigma_X}) = n$. Let $\mathbf{K} = (\mathbf{\Sigma_X})^{1/2}$. 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 20 for the case $\mathbf{X} = \tilde{\mathbf{X}}$, the sought path (for $\mathbf{X} = \mathbf{X}$) is given by $t \in [0,1] \mapsto (\mathbf{U}_t, \mathbf{W}_t \mathbf{K}^{-1})$.
Case 2: $\mathrm{rk}(\mathbf{\Sigma_X}) < n$. In this case, if $r = \mathrm{rk}(\mathbf{\Sigma_X})$, $\mathbf{X}$ belongs to a r-dimensional subspace of $\mathbb{R}^n$ (a.s.), call it V. If $\mathbf{O} \in \mathbb{R}^{n \times r}$ 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 $\mathrm{rk}(\mathbf{\Sigma}_{\tilde{\mathbf{X}}}) = r$. 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 $t \in [0,1] \mapsto (\mathbf{U}_t, \mathbf{W}_t \mathbf{O}^T)$. ■

This concludes the proof of non-existence of spurious valleys for the square loss function of linear one-hidden-layer NNs $\mathbf{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{UWx}$. The fact that such proof does not require any assumptions on the dimensions of the layers n, p, 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 NNs of any depth $K \geq 1$:

$$\mathbf{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \mathbf{W}_{K+1} \cdots \mathbf{W}_1 \mathbf{x} \tag{12}$$

We start by proving a simple lemma.

Lemma 22 *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 (w.l.o.g.) 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 8, 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 11.

Proof [Proof of Theorem 11] Consider a linear network $\mathbf{\Phi}(\mathbf{x};\boldsymbol{\theta})$ as in (12), where

$$\mathbf{W}_k \in \mathbb{R}^{p_k \times p_{k-1}} \quad \text{for } k \in [K+1]$$

We select $p_s = \min_{i \in [K]} p_k$. Then the network can be written as

$$\mathbf{\Phi}(\mathbf{x};\boldsymbol{\theta}) = \hat{\mathbf{W}}^2 \hat{\mathbf{W}}^1 \mathbf{x} \quad \text{where} \quad \hat{\mathbf{W}}^2 = \mathbf{W}^{K+1} \cdots \mathbf{W}^{s+1}, \quad \hat{\mathbf{W}}^1 = \mathbf{W}^s \cdots \mathbf{W}^1 \tag{13}$$

Now we want to prove property that given an initial parameter $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{W}}^{K+1}, \dots, \tilde{\mathbf{W}}^1)$, there exists a continuous path $\boldsymbol{\theta}_t = (\mathbf{W}_t^{K+1}, \dots, \mathbf{W}_t^1)$ 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 (13) for $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, then by Lemma 21 there exists a path $(\hat{\mathbf{W}}_t^2, \hat{\mathbf{W}}_t^1)$ satisfying the above. Thanks to Lemma 22, we can decompose

$$\hat{\mathbf{W}}_t^2 = \mathbf{W}_t^{K+1} \cdots \mathbf{W}_t^{s+1}, \quad \hat{\mathbf{W}}_t^1 = \mathbf{W}_t^s \cdots \mathbf{W}_t^1 \tag{14}$$

in a continuous way. Since p_s was to chosen as the minimum, it also holds that

$$\min_{\boldsymbol{\theta} = (\hat{\mathbf{W}}^2, \hat{\mathbf{W}}^1)} L(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta} = (\mathbf{W}^{K+1}, \dots, \mathbf{W}^1)} L(\boldsymbol{\theta})$$

Therefore this is a suitable path and this concludes the proof of the theorem. ■

A.3. Proof of Theorem 12

Proof [Proof of Theorem 12] 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 $\mathbf{A} = \sum_{k=1}^{p} \tilde{u}_k \tilde{\mathbf{w}}_k \tilde{\mathbf{w}}_k^T$ and $\sum_{k=1}^{n} \mathbf{u}_k^* \mathbf{w}_k^* (\mathbf{w}_k^*)^T$ be the SVD of $\mathbf{A}$. We define the parameters value $\boldsymbol{\theta}^* = (\mathbf{u}^*, \mathbf{W}^*)$ where $\mathbf{u}^* = (u_1^*, \dots, u_n^*, 0, \dots, 0)$ and $\mathbf{W}^*$ is the $p \times n$ matrix with rows $\mathbf{w}_i^*$ for $i \in [n]$ and $\mathbf{0}$ for $i \in [n+1, p]$. 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 23.

Step 2. As noticed above, the network can be written as $\boldsymbol{\Phi}(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{u}^T \sigma(\mathbf{W}\mathbf{x}) = \langle \mathbf{A}, \mathbf{M} \rangle_F$, where $\mathbf{A} = \sum_{k=1}^{p} u_k \mathbf{w}_k \mathbf{w}_k^T$ 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 $\sum_{i=1}^{n} \bar{u}_k \bar{\mathbf{w}}_k \bar{\mathbf{w}}_k^T$ be the SVD of $\bar{\mathbf{A}}$; also let $\bar{\mathbf{u}} = (0, \dots, 0, \bar{u}_1, \dots, \bar{u}_n)$ and $\bar{\mathbf{W}}$ be the $p \times n$ matrix with rows $\mathbf{0}$ for $i \in [p-n]$ and $\bar{\mathbf{w}}_i$ for $i \in [p-n+1, p]$. 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 $i \in [n+1, p]$. 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 12. ■

To conclude the proof we just need to prove the following lemmas.

Lemma 23 *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 12. 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 row $\mathbf{w}_k$ of $\mathbf{W}$ by $\sqrt{|u_k|}$. Assume first that $\mathbf{u} \in \{\pm 1\}^p$. The general case ($u_k = 0$ for some k) is addressed in Remark 26. The sought path $\boldsymbol{\theta}_t$ can be constructed by iterating two steps (a finite amount of times). First we select a row $\mathbf{w}_k$ and construct a continuous path that maps this row to one of the $\mathbf{w}_i^*$; then we orthogonalize (w.r.t. such $\mathbf{w}_i^*$) the rest of rows $\mathbf{w}_j$, $j \neq k$. These two steps are constructed so that $\mathbf{A}$ never changes and therefore the loss is constant. The first step is described in Lemma 24, while the second is detailed in Lemma 25. At this point the parameter $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W})$ verifies $u_i = u_i^*$, $\mathbf{w}_i = \mathbf{w}_i^*$ and $\mathbf{w}_j \in \langle \{\mathbf{w}_i^*\} \rangle^{\perp}$ for $j \neq k$. In particular it holds

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} u_j^* \mathbf{w}_j^* (\mathbf{w}_j^*)^T = \sum_{\substack{j=1 \\ j \neq k}}^{p} u_k \mathbf{w}_k \mathbf{w}_k^T$$

Therefore, an induction step applied on the reduced parameter values

$$\mathbf{u}_{-k} = (u_1, \dots, \widehat{u_k}, \dots, u_p)$$

and $\mathbf{W}_{-k} = [\mathbf{w}_1, \dots, \widehat{\mathbf{w}_k}, \dots, \mathbf{w}_p]^T P$, where $\mathbf{P} = \sum_{j=1, j \neq i}^{n} \mathbf{w}_j^* \mathbf{e}_j^T \in \mathbb{R}^{n \times (n-1)}$, concludes the proof. The fact that the non-zero components of $\mathbf{u}$ and $\mathbf{W}$ coincide with the first n is not

necessary, but we can clearly assume it to hold w.l.o.g. ■

Lemma 24 *The first step described in the Proof of Lemma 23 can be performed when $p > 2n$.*

Proof Let $E_+ = \{k \in [p] : u_k = 1\}$, $E_- = \{k \in [p] : u_k = -1\}$ and $p_+ = |E_+|$, $p_- = |E_-|$. Accordingly we define

$$\mathbf{W}_+ = ([\mathbf{w}_k]_{k\in E_+})^T \in \mathbb{R}^{p_+\times n} \quad \text{and} \quad \mathbf{W}_- = ([\mathbf{w}_k]_{k\in E_-})^T \in \mathbb{R}^{p_-\times n}$$

Notice that then we can write

$$\mathbf{A} = \mathbf{W}_+^T\mathbf{W}_+ - \mathbf{W}_-^T\mathbf{W}_-$$

The main step of the proof is to observe that $\mathbf{A}$ (and therefore the loss) is invariant to the action of orthogonal matrices $\mathbf{Q}_+ \in SO(p_+)$ and $\mathbf{Q}_- \in SO(p_-)$. So, if $\mathbf{Q}_+(t)$ (resp. $\mathbf{Q}_-(t)$) is a continuous paths in $SO(p_+)$ (resp. in $SO(p_-)$) starting at the identity, acting on $\mathbf{W}$ as

$$\mathbf{W}_+(t) \doteq \mathbf{Q}_+(t)\mathbf{W}_+, \quad \mathbf{W}_-(t) \doteq \mathbf{Q}_-(t)\mathbf{W}_-$$

we have that

$$\mathbf{A} = \mathbf{W}_+(t)^T\mathbf{W}_+(t) - \mathbf{W}_-(t)^T\mathbf{W}_-(t)$$

is constant for all t. Now, since $p = p_+ + p_- > 2n$, it follows that either $p_+ > n$ or $p_- > n$. Assume w.l.o.g. that $p_+ > n$. Since $p_+ > n$, we can rotate the subspace generated by the columns of $\mathbf{W}_+$ so that its first row is $\mathbf{0}$. That is, there exist $\mathbf{h} \in \mathbb{R}^{p_+}$ non-zero such that $\mathbf{h}^T\mathbf{W}_+ = 0$ and $\|\mathbf{h}\| = 1$. It suffices to choose a path $\mathbf{Q}(t)$ in $SO(p_+)$ whose first row equals $\mathbf{h}$ at $t = 1$. It follows that $\mathbf{Q}(1)\mathbf{W}_+$ has a first row 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 25 *Assume that after the step in Lemma 24, the first row of $\mathbf{W}_+$ (resp. $\mathbf{W}_-$) is given by $\mathbf{w}_i^*$. Then we can map all the other rows of $\mathbf{W}$ to be orthogonal to $\mathbf{w}_i^*$, while keeping $\mathbf{A}$ constant.*

Proof To simplify the notation we assume (w.l.o.g.) that $\mathbf{w}_i^* = \mathbf{w}_1^*$ and that

$$\mathbf{W} = [\mathbf{w}_1^*, \mathbf{w}_2, \cdots, \mathbf{w}_p]^T$$

Now we want to construct a path

$$\mathbf{u}_t = (u_{1,t}, u_2, \ldots, u_p)$$
$$\mathbf{W}_t = [\mathbf{w}_1^*, \mathbf{w}_{2,t}, \cdots, \mathbf{w}_{p,t}]^T$$

such that $\mathbf{w}_{2,1}, \ldots, \mathbf{w}_{p,1} \in \langle\{\mathbf{w}_1^*\}\rangle^\perp$. To do this we simply take

$$\mathbf{w}_{k,t} \doteq \mathbf{w}_k - t\langle\mathbf{w}_1^*, \mathbf{w}_k\rangle\mathbf{w}_1^*$$

If $\mathbf{A}_t = \sum_{k=1}^p u_{k,t}\mathbf{w}_{k,t}\mathbf{w}_{k,t}^T$, 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

$$\begin{aligned}\mathbf{A}_t &= u_{1,t}\,\mathbf{w}_1^*(\mathbf{w}_1^*)^T \\ &\quad + \sum_{k=2}^p u_k\big[(1-t)^2(w_k^1)^2\mathbf{w}_1^*(\mathbf{w}_1^*)^T + (1-t)w_k^1\big(\tilde{\mathbf{w}}_k(\mathbf{w}_1^*)^T + \mathbf{w}_1^*\tilde{\mathbf{w}}_k^T\big) + \tilde{\mathbf{w}}_k\tilde{\mathbf{w}}_k^T\big]\end{aligned}$$

where $w_k^1 \doteq \langle \mathbf{w}_k, \mathbf{w}_1^*\rangle$ and $\tilde{\mathbf{w}}_k = \mathbf{w}_k - w_k^1\mathbf{w}_1^*$. In particular

$$\mathbf{A}_t = \mathbf{V}^*\left[\begin{array}{c|c} a_t & \mathbf{b}_t^T \\ \hline \mathbf{b}_t & \mathbf{A}_{2:n,2:n}\end{array}\right](\mathbf{V}^*)^T$$

where $\mathbf{V}^* = [\mathbf{w}_1^*, \cdots, \mathbf{w}_n^*] \in O(n)$. Since $\sum_{k=2}^p u_k w_k^1\,\tilde{\mathbf{w}}_k = 0$, it follows

$$\mathbf{b}_t = (1-t)\sum_{k=2}^p u_k w_k^1\,\tilde{\mathbf{w}}_k = 0 \quad \text{for all } t \in [0,1]$$

If we take

$$u_{1,t} = \lambda_1 - (1-t)^2\sum_{k=2}^p u_k(w_k^1)^2$$

it holds that

$$a_t = u_{1,t} + (1-t)^2\sum_{k=2}^p u_k(w_k^1)^2 = \lambda_1 \quad \text{for all } t \in [0,1]$$

Therefore, $\mathbf{A}_t = \mathbf{A}$ constant. This concludes the proof of the lemma. ■

Remark 26 *In the proof of Lemma 23, we assumed that (after rescaling)* $\mathbf{u} \in \{\pm 1\}^p$. *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.*

Appendix B. Proofs of Section 4

Proof [Proof of Theorem 13] We consider here the case $m = 1$, but the same proof can be extended to the case $m > 1$. We start by properly choosing a r.v. $(\mathbf{X}, \mathbf{Y})$. Be $\bar{\mathbf{X}} \in \mathcal{R}_2(\sigma, n-1)$ a $(n-1)$ dimensional r.v. and $\bar{X}_n \in \mathcal{R}_2(\sigma, 1)$ a one dimensional r.v. We consider $\tilde{\mathbf{X}} = Z\bar{\mathbf{X}}$, $X_n = (1-Z)\bar{X}_n$ and $\mathbf{X} = (\tilde{\mathbf{X}}, X_n)$, where $Z \sim \mathrm{Ber}(1/2)$ and $\bar{\mathbf{X}}, \bar{X}_n, Z$ are independent. By hypothesis, $p \le 2^{-1}\dim_*(\sigma, \tilde{\mathbf{X}})$. The proof is based on the fact that (for a proper choice of $\tilde{\mathbf{X}}$) this implies that $\overline{V_{\sigma,p-1}^+} \neq V_{\sigma,p}^+$, where we defined

$$V_{\sigma,p}^+ = \{\Phi(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} \in [0,\infty)^p \times \mathbb{R}^{p\times n}\} \subseteq L^2_{\mathbf{X}}$$

(see the remark at the end of the proof). The r.v. Y is taken to be $Y = g_1(\mathbf{X}) - g_2(\mathbf{X})$, where $g_2 = \beta\psi_{\sigma,\mathbf{v}} \in V_{\sigma,1}^+$, $\beta > 0$, $\mathbf{v} = \mathbf{e}_n$, and $g_1 = \sum_{i=1}^p \alpha_i\psi_{\sigma,\mathbf{v}_i} \in V_{\sigma,p}^+$, $\boldsymbol{\alpha} \in (0,\infty)^p$, $\mathbf{v}_i \in \langle\{\mathbf{e}_n\}\rangle^\perp$, $i \in [p]$, is such that

$$\inf_{f \in V_{\sigma,p-1}^+} \mathbb{E}|f(\mathbf{X}) - g_1(\mathbf{X})|^2 = \epsilon > 0$$

We define

$$V_{\sigma,(p-1,1)} = \left\{ f = f_1 - f_2 : f_1 \in V^+_{\sigma,p-1}, f_2 \in V^+_{\sigma,1} \right\}$$

Notice that, for every path $\boldsymbol{\theta} : t \in [0,1] \mapsto \boldsymbol{\theta}_t \in \Theta$ such that $\Phi(\cdot;\boldsymbol{\theta}_0) \in V^+_{\sigma,p}$ and $\Phi(\cdot;\boldsymbol{\theta}_1) \in V_{\sigma,(p-1,1)}$, there exists $t_0 \in (0,1)$ such that $\Phi(\cdot;\boldsymbol{\theta}_{t_0}) \in V^+_{\sigma,p-1}$. Consider the *lifted* square loss function $L : V_{\sigma,p} \to [0,\infty)$ defined as

$$L(f) = \mathbb{E}|f(\mathbf{X}) - g(\mathbf{X})|^2 \qquad \text{for } f \in V_{\sigma,p}$$

We want to show that

$$L_{(p-1,0)} \doteq \min_{f \in V^+_{\sigma,p-1}} L(f) > L_{(p,0)} \doteq \min_{f \in V^+_{\sigma,p}} L(f) > L_{(p-1,1)} \doteq \min_{f \in V_{\sigma,(p-1,1)}} L(f)$$

It holds that

$$\begin{aligned} L_{(p-1,0)} = & \min_{f \in V^+_{\sigma,p-1}} \left\{ \mathbb{E}|f(\mathbf{X}) - g_1(\mathbf{X})|^2 \right\} + 2 \min_{f \in V^+_{\sigma,p-1}} \{\mathbb{E}[f(\mathbf{X}) g_2(\mathbf{X})]\} \\ & + \mathbb{E}|g_2(\mathbf{X})|^2 - C\sigma(0) \\ & \geq \epsilon + L_{(p,0)} - C\sigma(0) \end{aligned}$$

where $C = \mathbb{E}[g_1(\tilde{\mathbf{X}})] + \mathbb{E}[g_2(X_n)]$,and that

$$\begin{aligned} L_{(p,0)} = & \min_{f \in V^+_{\sigma,p}} \left\{ \mathbb{E}|f(\mathbf{X}) - g_1(\mathbf{X})|^2 \right\} + 2 \min_{f \in V^+_{\sigma,p}} \{\mathbb{E}[f(\mathbf{X}) g_2(\mathbf{X})]\} \\ & + \mathbb{E}|g_2(\mathbf{X})|^2 - C\sigma(0) \\ & \geq \beta^2 \, \mathbb{E}|\psi_{\sigma,\mathbf{v}}(X_n)|^2 - C\sigma(0) \end{aligned}$$

Finally, it holds that

$$L_{(p-1,1)} \leq \min_{i \in [1,p]} \alpha_i^2 \, \mathbb{E}|\psi_{\sigma,\mathbf{v}_i}(\mathbf{X})|^2$$

Given $M > 0$, up to multiply g_1 by a positive constant, it holds that

$$\begin{aligned} \epsilon &\geq M + C\sigma(0) \\ \beta^2 &\geq \frac{M + C\sigma(0) + \min_{i \in [1,p]} \alpha_i^2 \, \mathbb{E}|\psi_{\sigma,\mathbf{v}_i}(\mathbf{X})|^2}{\mathbb{E}|\psi_{\sigma,\mathbf{v}}(X_n)|^2} \end{aligned}$$

To finish the proof, consider $\mathcal{U} = \{\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W}) \in \Theta : \mathbf{u} \in (0,\infty)^p\}$ 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.

■

Remark 27 *In the proof of Theorem 13 we used the fact that, if $p \geq 1$ verifies $p \leq \frac{1}{2}\dim_*(\sigma, n)$, then there exist $\mathbf{X} \in \mathcal{R}_2(\sigma, n)$ such that $\overline{V^+_{\sigma,p-1}} \neq V^+_{\sigma,p}$ (in the $L^2_{\mathbf{X}}$ metric). Assume $\mathbf{X}$ is a n-dimensional standard Gaussian variable. Consider first the case $\sigma(z) = z^k$. If $k = 2$, then*

$$V^+_{\sigma,p} \simeq \{\mathbf{M} \in \mathrm{S}^2(\mathbb{R}^n) : \mathbf{M} \textit{ is PSD}\}$$

In particular, this implies that $V^+_{\sigma,p-1}$ is not dense in $V^+_{\sigma,p}$ if $p \leq n = \dim_*(\sigma, n)$ (which justifies the statement of Corollary 14). If $k > 2$, let $p \leq \frac{1}{2}\dim_*(\sigma, n)$ and assume that $\overline{V^+_{\sigma,p-1}} = V^+_{\sigma,p}$. This implies that every tensor $\mathbf{T} = \sum_{i=1}^p \mathbf{v}_i^{\otimes k}$ can be approximated up to any accuracy by $\mathbf{T} = \sum_{i=1}^{p-1} \tilde{\mathbf{v}}_i^{\otimes k}$ for some $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_{p-1} \in \mathbb{R}^n$. But this also implies that every tensor $\mathbf{T} \in \mathbb{S}^k(\mathbb{R}^n)$ has border rank less or equal than $2\big(\frac{1}{2}\dim_*(\sigma, n) - 1\big) = \mathrm{rk}_{\mathrm{S}}(k, n) - 2$, which contradicts the definition of $\mathrm{rk}_{\mathrm{S}}(k, n)$. For non-polynomial σ, we can get the same result, by using the decomposition (18) *and proceeding as above.*

Appendix C. Proof of Theorem 16

Proof If we denote by $d\mu$ the probability distribution of $\mathbf{X}$, the continuous function

$$\psi : (\mathbf{w}, \mathbf{x}) \in \mathbb{S}^n \times \mathbb{R}^n \mapsto \psi_{\mathbf{w}}(\mathbf{x}) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$$

belongs to $L^2(\mathbb{S}^n \times \mathbb{R}^n, d\tau \otimes d\mu)$. We consider the kernel associated with the neural network architecture

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathcal{W}} \psi_{\mathbf{w}}(\mathbf{x})\psi_{\mathbf{w}}(\mathbf{y})\, d\tau(\mathbf{w}) \tag{15}$$

The above defines a continuous symmetric, positive semi-definite kernel k, along with $\mathbb{H}$, the RKHS associated, and the integral operator $\Sigma : L^2(\mathbb{R}^n, d\mu) \to \mathbb{H} \subseteq L^2(\mathbb{R}^n, d\mu)$ defined as

$$f \mapsto \left(\Sigma f : \mathbf{x} \mapsto \int_{\mathbb{R}^n} f(\mathbf{y})k(\mathbf{x}, \mathbf{y})\, d\mu(\mathbf{y})\right)$$

The operator Σ admits a spectral decomposition in $L^2(\mathbb{R}^n, d\mu)$: $\Sigma e_k = \lambda_k e_k$ for an orthonormal basis $\{e_k\}_{k\geq 1}$ of $L^2(\mathbb{R}^n, d\mu)$ and non-increasing sequence of non-negative eigenvalues $\{\lambda_k\}_{k\geq 1}$. Moreover the RKHS $\mathbb{H}$ is dense in $L^2(\mathbb{R}^n, d\mu)$ (see Lemma 30), which is equivalent to have $\lambda_k > 0$ for all $k \geq 1$. The expectation in (15) provides a singular value decomposition for Σ in terms of functions in $L^2(\mathbb{S}^n, d\tau)$. Indeed, given $g \in L^2(\mathbb{S}^n, d\tau)$, the linear operator $T : L^2(\mathbb{S}^n, d\tau) \to L^2(\mathbb{R}^n, d\mu)$ defined as

$$g \mapsto \left(Tg : \mathbf{x} \mapsto \int_{\mathbb{S}^n} g(\mathbf{w})\psi_{\mathbf{w}}(\mathbf{x})\, d\tau(\mathbf{w})\right)$$

satisfies $\Sigma = TT^*$. It follows that there exists an orthonormal basis of $L^2(\mathbb{S}^n, d\tau)$, $\{f_k\}_{k\geq 1}$ 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 (Bach, 2017a) that in fact $\mathbb{H} = \mathrm{Im}(T)$, and thus $\mathbb{H}$ consists of functions f that can be written, for some $g \in L^2(\mathbb{S}^n, d\tau)$ as

$$f(\mathbf{x}) = \int_{\mathbb{S}^n} g(\mathbf{w})\psi_{\mathbf{w}}(\mathbf{x}) d\tau(\mathbf{w}) = \langle g, \psi(\cdot, \mathbf{x}) \rangle_{L^2(\mathbb{S}^n, d\tau)} \quad \text{for } \mathbf{x} \in \mathbb{R}^n$$

For an account of these properties, we refer to Bach (Bach, 2017b). Thanks to the density of $\mathbb{H}$ in $L^2(\mathbb{R}^n, d\mu)$, we can assume, without loss of generality, that

$$f^*(\mathbf{x}) = \int_{\mathbb{S}^n} g^*(\mathbf{w})\psi_{\mathbf{w}}(\mathbf{x})d\tau(\mathbf{w})$$

for some $g^* \in L^2(\mathbb{S}^n, d\tau)$. Now, given an initial set of first layer weights $\mathbf{w}_1, \dots, \mathbf{w}_p \in \mathbb{S}^n$ sampled i.i.d. from $d\tau$, and $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_p]^T$, we define the empirical kernel

$$k_{\mathbf{W}}(\mathbf{x}, \mathbf{y}) = \frac{1}{p}\sum_{i=1}^{p} \sigma(\langle \mathbf{x}, \mathbf{w}_i\rangle)\sigma(\langle \mathbf{y}, \mathbf{w}_i\rangle)$$

which in turn defines an empirical RKHS $\mathbb{H}_{\mathbf{W}}$. Keeping the first layer weights fixed and optimizing the output layer weights thus gives us the ability to find a function $f^*_{\mathbf{W}} \in \mathbb{H}_{\mathbf{W}}$ that best approximates f^*:

$$\|f^*_{\mathbf{W}} - f^*\|_{L^2(\mathbb{R}^n, d\mu)} = \min_{f \in \mathbb{H}_{\mathbf{W}}} \|f - f^*\|_{L^2(\mathbb{R}^n, d\mu)} \doteq R(\mathbf{W})$$

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

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

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 (Bach, 2017b), 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

$$\begin{aligned} d(\lambda) &= \max_{\mathbf{w} \in \mathbb{S}^n} \mathbb{E}\big[\varphi_{\mathbf{w}}(\mathbf{X})((\Sigma + \lambda I)^{-1}\psi_{\mathbf{w}})(\mathbf{X})\big] \\ &= \max_{\mathbf{w} \in \mathbb{S}^n} \sum_{k=1}^{\infty} \frac{\lambda_k}{\lambda_k + \lambda} f_k(\mathbf{w})^2 \leq \lambda^{-1} \max_{\mathbf{w} \in \mathbb{S}^n} \sum_{k=1}^{\infty} \lambda_k f_k(\mathbf{w})^2 = \lambda^{-1} \max_{\mathbf{w} \in \mathbb{S}^n} \|\psi_{\mathbf{w}}\|^2_{L^2(\mathbb{R}^n, d\mu)} \end{aligned}$$

This shows part 1 of the statement. To prove part 2, notice that $f^*_{\mathbf{W}} = \Phi(\cdot; \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (\mathbf{u}^*_{\mathbf{W}}, \mathbf{W})$ for some $\mathbf{u}^*_{\mathbf{W}} \in \mathbb{R}^p$. By taking $u_k = \frac{1}{p} g^*(\mathbf{w}_k)$ for $k \in [p]$ and denoting $Z(\mathbf{w}) := g^*(\mathbf{w})\psi_{\mathbf{w}}$ and by Z the r.v. $Z = Z(\mathbf{v})$, for $\mathbf{v} \sim d\tau$, with values in $L^2(\mathbb{R}^n, d\mu)$, it holds

$$R(\mathbf{W}) \leq \Big\| \frac{1}{p}\sum_{k=1}^{p} Z(\mathbf{w}_k) - \mathbb{E}_\tau[Z] \Big\|_{L^2(\mathbb{R}^n, d\mu)} \tag{16}$$

Note that $C \doteq \sup_{\mathbf{w} \in \mathbb{S}^n} \|Z(\mathbf{w})\|_{L^2(\mathbb{R}^n, d\mu)} \leq \|g^*\|_{L^\infty(\mathbb{S}^n, d\tau)} \max_{\mathbf{w} \in \mathbb{S}^n} \|\psi_{\mathbf{w}}\|_{L^2(\mathbb{R}^n, d\mu)} < \infty$ if $\|g^*\|_{L^\infty(\mathbb{S}^n, d\tau)} < \infty$. Then, applying Lemma 29 to the bound (16), we get that

$$\mathbb{P}_\tau\{R(\mathbf{W}) \leq \varepsilon\} \geq 1 - \exp\Big\{ -\Big(\varepsilon - \sqrt{v(p)}\Big)^2 / (2v(p)) \Big\}$$

for every $\varepsilon \geq \sqrt{v(p)}$, with $v(p) = C^2/p$. The result follows by taking $\varepsilon = v(p)^{1/2} + p^{\delta/2 - 1/2}$. ■

Appendix D. Proofs of Section 2.2

Proof [Proof of Lemma 4] If σ is a polynomial of any degree d, then it holds that $\dim^*(\sigma, n) < \infty$. Indeed, let $\sigma(z) = a_0 + a_1 z + \cdots + a_d z^d$, for some $a_k \in \mathbb{R}$. If $I = \{k \in [0, d] : a_k \neq 0\}$, then

$$V_\sigma \subseteq \mathbb{R}_I[\mathbf{x}] \doteq \{\mathbf{x} \mapsto \sum_{k \in I} \sum_{|\boldsymbol{\beta}|=k} \alpha_{\boldsymbol{\beta}} \mathbf{x}^{\boldsymbol{\beta}} : \alpha_{\boldsymbol{\beta}} \in \mathbb{R}\}$$

It follows that

$$\dim^*(\sigma, n) = \dim(V_\sigma) \leq \dim(\mathbb{R}_I[\mathbf{x}]) = \sum_{k=0}^{d} \binom{n+k-1}{k} \mathbf{1}_{\{a_k \neq 0\}} = O(n^d)$$

This proves one implication. We prove the other one by contradiction. Assume now that σ is not a polynomial and that $\dim(V_\sigma) = q < \infty$. Thanks to Theorem 1 in Leshno et al. (Leshno et al., 1993), for every continuous function $g : \mathbb{R}^n \to \mathbb{R}$, any compact set $K \subset \mathbb{R}^n$, and any $\varepsilon > 0$ there exist $h \in V_\sigma$ such that

$$\sup_{\mathbf{x} \in K} |h(\mathbf{x}) - g(\mathbf{x})| < \varepsilon \tag{17}$$

Now, let $g : \mathbb{R}^n \to \mathbb{R}$ be a continuous function supported on a compact set $C \subset \mathbb{R}^n$. We call $C_c(\mathbb{R}^n)$ the set of the real-valued continuous functions from $\mathbb{R}^n$ with compact support. Thanks to (17), we can find a sequence of compact sets $\{K_m\}_{m \geq 1}$ of $\mathbb{R}^n$ such that

$$C \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m \subseteq \cdots \subseteq \cup_{m=1}^{\infty} K_m = \mathbb{R}^n$$

and a sequence of functions $\{h_m\}_{m \geq 1} \subset V_\sigma$ 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 $\lim_{n \to \infty} h_m \mathbb{1}_{K_m} = g \in L^2_{\mathbf{X}}$. Since $\dim(V_\sigma) = q < \infty$, there exists $\mathbf{w}_1, \ldots, \mathbf{w}_q \in \mathbb{R}^n$ such that every $h \in V_\sigma$ can be written as

$$h(\mathbf{x}) = \langle \mathbf{u}, \boldsymbol{\gamma}(\mathbf{x}) \rangle$$

for some $\mathbf{u} \in \mathbb{R}^q$, where $\boldsymbol{\gamma}(\mathbf{x}) = (\sigma(\langle \mathbf{w}_1, \mathbf{x} \rangle), \ldots, \sigma(\langle \mathbf{w}_q, \mathbf{x} \rangle))$. Let $\{\mathbf{u}_m\}_{m \geq 1} \subset \mathbb{R}^q$ such that $h_m(\mathbf{x}) = \langle \mathbf{u}_m, \boldsymbol{\gamma}(\mathbf{x}) \rangle$. 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 $K \subseteq \mathbb{R}^n$. Since

$$\|h_m \mathbb{1}_K\|^2_{L^2_{\mathbf{X}}} = \mathbf{u}_m^T M \mathbf{u}_m$$

where $M = \mathbb{E}\big[\boldsymbol{\gamma}(\mathbf{X})\boldsymbol{\gamma}(\mathbf{X})^T \mathbb{1}_{\{\mathbf{X} \in K\}}\big] \in \mathbb{R}^{q \times q}$, 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 $h \in V_\sigma$ the function defined as $h(\mathbf{x}) = \langle \mathbf{u}, \boldsymbol{\gamma}(\mathbf{x}) \rangle$, it is easy to check (from the above calculations) that $h = g$ in $L^2_{\mathbf{X}}$. This shows that $C_c(\mathbb{R}^n) \subseteq V_\sigma$, which

in turn implies that V_σ 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 $\dim(V_\sigma) = q < \infty = \dim(L^2_{\mathbf{X}})$. Therefore, it must hold $\dim(V_\sigma) = \infty$. ■

Proof [Proof of Lemma 6] Let $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W}) \in [0,\infty)^p \times \mathbb{R}^{p\times n}$. For every $\mathbf{x} \in \mathbb{R}^n$ it holds

$$\Phi(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^p u_i(\langle \mathbf{w}_i, \mathbf{x}\rangle)^k = \sum_{i=1}^p u_i \langle \mathbf{w}_i^{\otimes k}, \mathbf{x}^{\otimes k}\rangle_F = \Big\langle \sum_{i=1}^p u_i \mathbf{w}_i^{\otimes k}, \mathbf{x}^{\otimes k}\Big\rangle_F$$

For any $p \geq 1$ and $(\mathbf{u}, \mathbf{W}) \in [0,\infty)^p \times \mathbb{R}^{p\times n}$, $\sum_{i=1}^p u_i \mathbf{w}_i^{\otimes k} \in S^k(\mathbb{R}^n)$. By definition of $\mathrm{rk}_{\mathrm{S}}(k,n)$, it follows that there exists $q \leq \mathrm{rk}_{\mathrm{S}}(k,n)$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\mathbf{u}}, \tilde{\mathbf{W}}) \in [0,\infty)^q \times \mathbb{R}^{q\times n}$ such that

$$\sum_{i=1}^p u_i \mathbf{w}_i^{\otimes k} = \sum_{i=1}^q \tilde{u}_i \tilde{\mathbf{w}}_i^{\otimes k} \quad \Rightarrow \quad \Phi(\cdot;\boldsymbol{\theta}) = \Phi(\cdot;\tilde{\boldsymbol{\theta}})$$

By definition of $\dim_*(\sigma,n)$, this implies that $\dim_*(\sigma,n) \leq \mathrm{rk}_{\mathrm{S}}(k,n)$. The equality follows by choosing $(\mathbf{u}, \mathbf{W}) \in [0,\infty)^p \times \mathbb{R}^{p\times p}$ such that $\mathrm{rk}_{\mathrm{S}}(\sum_{i=1}^p u_i \mathbf{w}_i^{\otimes k}) = \mathrm{rk}_{\mathrm{S}}(k,n)$. ■

Proof [Proof of Lemma 7] If σ is polynomial, then one implication follows by Lemma 4. Now, assume that $\sigma \in L^2(\mathbb{R}, e^{-x^2/2}\,dx)$ is a continuous non-polynomial activation and let $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$ be a r.v. in $\mathcal{R}_2(\sigma,n)$. Then, we can write $\sigma(z) = \sum_{k=0}^\infty \hat{\sigma}_k h_k(z)$, where h_k is the k-th Hermite polynomial. It follows that, for $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W})$,

$$\mathbb{E}|\Phi(\mathbf{X};\boldsymbol{\theta})|^2 = \sum_{k=1}^\infty \hat{\sigma}_k^2 \Big\| \sum_{i=1}^p u_i \mathbf{w}_i^{\otimes k} \Big\|_F^2 \tag{18}$$

(see Lemma 1 from (Mondelli and Montanari, 2018)). Since σ is not polynomial and $n > 1$, $V_{\sigma,p} \neq V_{\sigma,p+1}$, where

$$V_{\sigma,p} \doteq \left\{ \mathbf{x} \mapsto \sum_{k=1}^p u_k \sigma(\langle \mathbf{w}_k, \mathbf{x}\rangle) : (\mathbf{u}, \mathbf{W}) \in \mathbb{R}^p \times \mathbb{R}^{p\times n} \right\}$$

Indeed, if $V_{\sigma,p} = V_{\sigma,p+1}$, then $V_{\sigma,p} = V_{\sigma,q}$ for every $q > p$. Let k be a positive integer such that $\hat{\sigma}_k \neq 0$ and such that $\mathrm{rk}_{\mathrm{S}}(k,n) > p$. Let $\mathbf{G} = \sum_{i=1}^q \alpha_i \mathbf{v}_i^{\otimes k}$ a symmetric tensor with $\mathrm{rk}_{\mathrm{S}}(\mathbf{G}) = q = \mathrm{rk}_{\mathrm{S}}(k,n)$ and $g = \sum_{i=1}^q \alpha_i \psi_{\sigma,\mathbf{v}_i}$. If $g \in V_{\sigma,p}$, then there exists $\boldsymbol{\theta} = (\mathbf{u}, \mathbf{W}) \in \Theta = \mathbb{R}^p \times \mathbb{R}^{p\times n}$ such that

$$0 = \mathbb{E}|\Phi(\mathbf{X};\boldsymbol{\theta}) - g(\mathbf{X})|^2 = \sum_{k=1}^\infty \hat{\sigma}_k^2 \Big\| \sum_{i=1}^p u_i \mathbf{w}_i^{\otimes k} - \sum_{i=1}^q \alpha_i \mathbf{v}_i^{\otimes k} \Big\|_F^2$$

But this would imply that $\mathrm{rk}_{\mathrm{S}}(\mathbf{G}) \leq p$, which is a contradiction. This concludes the proof. ■

Appendix E. Proofs of Additional Lemmas

Lemma 2 *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 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 $B_\epsilon(\boldsymbol{\theta})$, for some $\epsilon > 0$. But this implies that for any path $t \in [0,1] \mapsto \boldsymbol{\theta}_t$ if 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 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 28 *Consider the optimization problem*

$$\arg\min_{\mathbf{W} \in \mathbb{R}^{m \times n}} \ell(\mathbf{W}) \quad \textit{where} \quad \ell(\mathbf{W}) = \mathbb{E}\|\mathbf{W}\mathbf{X} - \mathbf{Y}\|^2 \tag{19}$$

for two square integrable r.v.'s $\mathbf{X}$ and $\mathbf{Y}$ with values in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Then one solution to (19) is given by

$$\mathbf{W} = \boldsymbol{\Sigma}_{\mathbf{YX}} \boldsymbol{\Sigma}_{\mathbf{X}}^{\dagger} \tag{20}$$

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 $\mathbf{W} \in \mathbb{R}^{p \times n}$ is given by

$$\mathbf{U} = \mathbf{q}(\mathbf{W}) \doteq \boldsymbol{\Sigma}_{\mathbf{YX}} \mathbf{W}^T (\mathbf{W} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{W}^T)^{\dagger} \tag{21}$$

Assuming $\boldsymbol{\Sigma}_{\mathbf{X}}$ invertible, the minimal value obtained by $\ell(\mathbf{U}; \mathbf{W})$ is given by

$$\ell(\mathbf{q}(\mathbf{W}); \mathbf{W}) = \operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{Y}}) - \operatorname{tr}((\mathbf{W}\mathbf{K})^{\dagger} (\mathbf{W}\mathbf{K}) \mathbf{M}) \tag{22}$$

where $\mathbf{K} = (\boldsymbol{\Sigma}_{\mathbf{X}})^{1/2}$ and $\mathbf{M} = \mathbf{K}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}} \boldsymbol{\Sigma}_{\mathbf{YX}} \mathbf{K}^{-1}$. If $\mathbf{M} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ is the SVD of $\mathbf{M}$, the quantity (22) is minimized over $\mathbf{W}$ for $(\mathbf{W}\mathbf{K})^{\dagger}(\mathbf{W}\mathbf{K}) = \sum_{i=1}^{p \wedge n} \mathbf{v}_i \mathbf{v}_i^T$.

Proof The first part of the lemma can be shown by writing problem (19) as

$$\arg\min_{\mathbf{W} \in \mathbb{R}^{m \times n}} \ell(\mathbf{W}) \quad \text{where} \quad \ell(\mathbf{W}) = \operatorname{tr}(\mathbf{W} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{W}^T) - 2 \operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{YX}} \mathbf{W}^T) \tag{23}$$

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

$$\mathbf{U} = \boldsymbol{\Sigma}_{\mathbf{YXW}} (\boldsymbol{\Sigma}_{\mathbf{W}} \mathbf{X})^{\dagger} = \boldsymbol{\Sigma}_{\mathbf{YX}} \mathbf{W}^T (\mathbf{W} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{W}^T)^{\dagger}$$

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 $\mathbf{P_{WK}}$. In particular $\mathbf{P_{WK}}$ has the form $\mathbf{P_{WK}} = \sum_{i=1}^r \mathbf{w}_i\mathbf{w}_i^T$ for some $\{\mathbf{w}_1, \dots, \mathbf{w}_r\} \subset \mathbb{R}^n$ orthonormal vectors and $r \leq p \wedge n$. Therefore, minimize $\ell(\mathbf{q}(\mathbf{W}); \mathbf{W})$ over $\mathbf{W}$ it is equivalent to maximize the quantity

$$\sum_{i=1}^r \mathbf{w}_i^T\mathbf{M}\mathbf{w}_i$$

over the sets of $\mathbf{w}_1, \dots, \mathbf{w}_r$ orthonormal vectors of $\mathbb{R}^n$, $r \leq p \wedge n$. Clearly, this is for $\mathbf{w}_1 = \mathbf{v}_1, \dots, \mathbf{w}_{p\wedge n} = \mathbf{v}_{p\wedge n}$. This concludes the proof of the lemma. ■

Lemma 29 *Let $X_1, \dots, X_n$ be independent zero-mean r.v.'s 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*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n X_i\right\| > t\right\} \leq e^{-(t-\sqrt{v})^2/(2v)}$$

Proof The proof can be found in (Boucheron et al., 2013), Example 6.3. ■

Lemma 30 *Be $\mathbb{H} \subset L^2_{\mathbf{X}}$ the RKHS defined in the proof of Theorem 16. Then $\mathbb{H}$ is dense in $L^2_{\mathbf{X}}$.*

Proof First, note that the function $\mathbf{x} \in \mathbb{R}^n \mapsto k(\mathbf{x}, \mathbf{x})$ is in $L^1(\mathbb{R}^n, d\mu)$. Indeed

$$\begin{aligned}
\int_{\mathbb{R}^n}\int_{\mathbb{S}^n} \psi_{\mathbf{w}}(\mathbf{x})^2\, d\tau(\mathbf{w})\, d\mu(\mathbf{x}) &= \int_{\mathbb{R}^n}(1+\|\mathbf{x}\|^2)\int_{\mathbb{S}^n}\psi_{\mathbf{w}}(\mathbf{x}/\|(\mathbf{x},1)\|)^2\, d\tau(\mathbf{w})\, d\mu(\mathbf{x}) \\
&\leq (1+\mathbb{E}\|\mathbf{X}\|^2)\max_{\mathbf{w},\mathbf{y}\in\mathbb{S}^n}\psi_{\mathbf{w}}(\mathbf{y})^2
\end{aligned}$$

This implies that $\mathbb{H} \subseteq L^2(\mathbb{R}^n, d\mu)$. Now, we would like to show that V_σ is dense in $\overline{\mathbb{H}}$, where

$$V_\sigma = \left\{\sum_{i=1}^k u_i\psi_{\mathbf{w}_i} : \mathbf{u} \in \mathbb{R}^k, \mathbf{w}_1, \dots, \mathbf{w}_k \in \mathbb{S}^n, k \geq 1\right\}$$

It suffices to show that, for every $\mathbf{w} \in \mathbb{S}^{n-1}$, there exists a sequence $\{f_n\}_{n\geq 1} \subset \mathbb{H}$ such that $f_n \to \psi_{\mathbf{w}}$ in $L^2_{\mathbf{X}}$. Choose $g_k \in L^2(\mathbb{S}^n, d\tau)$ such that $\mathrm{supp}(g_k) \subseteq B_{1/k}(\mathbf{w}) \doteq \{\mathbf{v} \in \mathbb{S}^n : \|\mathbf{v}-\mathbf{w}\| \leq 1/k\}$, $\int_{\mathbb{S}^n} g(\mathbf{v})\, d\tau(\mathbf{v}) = 1$ and $g_k \geq 0$, and define $f_k \in \mathbb{H}$ as $f_k(x) = \int_{\mathbb{S}^n} g_k(\mathbf{v})\psi_{\mathbf{v}}(x)\, d\tau(\mathbf{x})$. Then

$$\begin{aligned}\|f_k - \psi_{\mathbf{w}}\|^2_{L^2(\mathbb{R}^n, d\mu)} &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{S}^n} g_k(\mathbf{v})(\psi_{\mathbf{v}}(\mathbf{x}) - \psi_{\mathbf{w}}(\mathbf{x}))\, d\tau(\mathbf{v}) \right)^2 d\mu(\mathbf{x}) \\ &\leq (1 + \mathbb{E}\|X\|^2) \max_{\substack{\mathbf{v}\in B_{1/k}(\mathbf{w}) \\ \mathbf{y}\in\mathbb{S}^n}} (\psi_{\mathbf{v}}(\mathbf{y}) - \psi_{\mathbf{w}}(\mathbf{y}))^2 \to 0\end{aligned}$$

as $k \to \infty$. This shows that V_σ is dense in $\overline{\mathbb{H}}$. Thanks to Theorem 1 in (Hornik, 1991), it holds that V_σ is dense in $L^2(\mathbb{R}^n, d\mu)$. This implies the statement of the lemma. ■