Report on *Spurious Valleys in one-hidden-layer Neural Networks, Optimization landscapes*

Leo Davy

March 11, 2022

Abstract

This report has been written as a companion paper for the Seminar in English course of the M2 Advanced Mathematics in the Probability and Statistics track. In this report are introduced and presented results from the article *Spurious Valleys in One-hidden-layer Neural Network, Optimization landscapes*, written by Luca Venturi, Alfonso S. Bandeira and Joan Bruna which has been published in *Journal of Machine Learning Research*. The goal of this report is to introduce the reader to the topic and provide a comprehensive review of the results and ideas provided in the article whilst emphasizing on an intuitive presentation through detailed examples.

The authors consider the problem of having continuous optimization paths of one-hidden-layer Neural Networks that do not increase the loss. An optimization problem where there exists initial parameters such that it is not possible to reach a global minimum with such a path is said to have *spurious valleys*. Some key quantities, namely the *lower* and *upper intrinsic dimension*, of a Neural Network are introduced. Those will allow to give simple explicit conditions to guarantee the absence, or existence, of spurious valleys. In a very concise way, if an intrinsic dimension is finite, then by having more parameters than this quantity guarantees the absence of spurious valleys.

Although over-parametrization is not feasible in general, it is shown that spurious valleys floors are confined to low risk levels and avoided with high probability, this phenomenon being more and more significant as parametrisation increases.

Contents

1 Introduction

1.1 Motivation to study the model

The empirical success of Neural Networks (NNs) has made them extremely popular for a wide variety of problems. However, before the recent surge of research on the topic of their effectiveness, no effective theoretical means were available to back-up their success, and even less to help designing Neural Networks. Indeed, the best tools that were available at the time were dedicated to convex or low-dimensional problems. On the other hand, NNs tackled several problems where neither of those properties were satisfied. Thus the surprise came from the fact that although there was no guarantee for optimization to reach a global minima of the loss, the optimization still reached parameters for which the loss was sufficiently low; in the sense that it improved on the state of the art methods for several problems.

Let's now say a few words on the optimization. The most used algorithm for such optimization has been Stochastic Gradient Descent (SGD), a refinement of Gradient Descent (Newton's method), or proximal methods, however for this report we won't need to know any details about those. The only thing we keep from their construction, is that at every step of optimization, they choose new parameters for which the loss doesn't increase (*no up-hill climb property*). Thus, in the present paper we will investigate the more general situation on whether or not a continuous path can exist that reaches a global minimum. More precisely we focus on the existence of local continuous path of parameters that do not increase the loss, we call such a path a *descent path*.

Since we ignore the details of the algorithms that are actually used for optimisation, the theory developed here gives a very partial account of the behaviour of optimisation. Many phenomenons will be missed (such as double descent) and more importantly, we will not know if the descent paths we will construct are the ones used by SGD; and for some of them this would be unlikely. However, the theory developed here will not be vain, since if we prove that there exists no descent path to a global (or good) minima, then it is not possible for an algorithm satisfying the no up-hill property to reach such a minima.

To introduce some notations, although they will be used mostly in the third section, we consider maps $\Phi_\theta : \mathbb{R}^n \to \mathbb{R}^m$ parametrized by $\theta \in \Theta \subset \mathbb{R}^P$, which for NNs consists of the weights across all layers. We denote the data as (X, Y) a random variable taking values in $\mathbb{R}^n \times \mathbb{R}^m$ following some, usually, unknown data distribution.

Our goal in this setting is to find some parameter θ that minimizes the loss

$$L(\theta) = \mathbb{E}_{(X,Y)}[l(\Phi_\theta(X), Y)]$$

where l is some real-valued function *convex in its first variable*.

1.2 Outline of the report

Here we describe the structure of this report. In a nutshell, the second section is about functional spaces, the third is about optimisation in those spaces.

In the first section the model will be introduced in a detailed way by first going through explicit example. The hope being that the examples will provide useful ideas and constructions that will be used later in the paper and to which the reader can refer. Examples will lead to the key definitions of *lower* and *upper* intrinsic dimensions of one-hidden-layer NNs (for a given activation function and data-distribution).
This first section will be concluded by reviewing the Universal Approximation Theorem that will allow us to characterize the activation functions for which the lower and upper intrinsic dimension will be finite.

The second part starts by an introduction to spurious valleys and what is exactly the optimization problem under consideration. This will be followed by a presentation of situations where spurious valleys do not exist and results from the preceding section will allow to obtain corollaries corresponding to explicit situations. Then some other results will be introduced on specific situations for which it is possible to improve over previously obtained results. Those upgrades, and some discussions, will hint at possible (conjectured) improvements on existing results.
The next situation investigated will be on the situation where networks are under-parametrized, or of infinite intrinsic dimension, for which it is possible to construct adversarial data distribution for which arbitrarily bad spurious valleys can exist.

The last situation considered will be of a more probabilistic nature and use results from Reproducing Kernel Hilbert Space to show that there are solid grounds for networks with a large number of parameters, as they will usually have spurious valleys but for which the floor is not too far from the lowest possible loss.

Note that the second and third sections can be read almost independently, for the third section only the definitions of intrinsic dimensions should be known, results from the second section are used to obtain corollaries.

2 Intrinsic dimensions

2.1 Detailed introduction to the setting

We consider Neural Networks $\Phi_{\theta,\sigma} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ given by a parameter $\theta = (W, U)$ and an activation function σ that represents a chain of maps

$$x \mapsto W(x) \mapsto \sigma(Wx) \mapsto U\sigma(Wx). \tag{1}$$

This defines a one-hidden-layer Neural Network with a layer of size p where W and U are linear maps of euclidean spaces with support and range of respective dimensions $n \to p$ and $p \to m$. Also, σ is an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ which defines a map $\mathbb{R}^p \to \mathbb{R}^p$ by acting element-wise on vectors, when there is no ambiguity on the activation function σ we omit it from notations for brevity. Also, σ is always assumed continuous.

We want to consider the class of functions expressible by a 1-layer NN defined as above with one-dimensional output ($m = 1$)

$$V_{\sigma,p} := \{f = \Phi(\cdot, \theta) : \theta = (U^T, W) \in \mathbb{R}^{p\times 1} \times \mathbb{R}^{p\times n}\}. \tag{2}$$

By re-expressing the parameter θ by its value we obtain

$$V_{\sigma,p} = \{f = \sum_{i=1}^{p} \lambda_i \Phi(\cdot, \theta = (e_i, W)) : \lambda_1, \cdots, \lambda_p \in \mathbb{R}; W \in \mathbb{R}^{p\times n}\} \tag{3}$$

where $(e_i)_{i=1,\cdots,p}$ is the canonical basis of $\mathbb{R}^p$. Hence, for any $i \in [p]$

$$V_{\sigma,p} = \bigcup_{W\in\mathbb{R}^{p\times n}} span\{f = \Phi(\cdot, \theta = (e_i, W))\} \tag{4}$$

and finally since

$$\Phi(x, \theta = (e_i, W)) = (\sigma(W(x)))_i = \sigma(\langle W_i, x\rangle) = \Phi(x, \theta = (1, W_i)) \tag{5}$$

where W_i is the i-th line of W

$$V_{\sigma,p} = \{\sum_{i=1}^{p} \lambda_i \Phi(\cdot, \theta = (1, w_i)) : \lambda_1, \cdots, \lambda_p \in \mathbb{R}^p; w_1, \cdots, w_p \in \mathbb{R}^n\}. \tag{6}$$

Since the last expression is a 1-layer NN with one parameter (i.e. it belongs to $V_{\sigma,1}$) which can be expressed as the map

$$\psi_{\sigma,v}(x) := \sigma(\langle v, x\rangle) = \Phi(\cdot, \theta = (1, v)) \tag{7}$$

and therefore

$$V_{\sigma,p} = \{\sum_{i=1}^{p} \lambda_i \psi_{\sigma,v_i}(\cdot) : \lambda_1, \cdots, \lambda_p \in \mathbb{R}; v_1, \cdots, v_p \in \mathbb{R}^n\} \tag{8}$$

$$= \{\sum_{i=1}^{p} \psi_{\sigma,v_i} : v_i \in \mathbb{R}^n\}. \tag{9}$$

What we get from this construction is that the functions expressible by a 1-layer NN with p parameters corresponds to p-sums of elements of the form $\psi_{\sigma,v}$ (with a free choice -unconstrained- on the weights v). We also notice that the functional space $V_{\sigma,p}$ is not in general a linear space.
This can be seen as follows, given two 1-layer NN defined by W and W' with 1 parameter, then their sum network $\Phi_W + \Phi_{W'}$ is not necessarily in $V_{\sigma,1}$, i.e., there might not exist a W'' such that $\Phi_W + \Phi_{W'} = \Phi_{W''}$.

2.2 Example: quadratic and polynomial networks

An explicit example is $\sigma(z) = z^2$, $x = (x_1, x_2)$, $W = (1, 0)$, $W' = (0, 1)$ then $\Phi_W + \Phi_{W'} = ||\cdot||_2^2$. Suppose that $||\cdot||_2^2$ is given by some W'', then $(\langle W'', x\rangle)^2 = w_1^2x_1^2 + 2w_1w_2x_1x_2 + w_2^2x_2^2 = x_1^2 + x_2^2$ which raises a contradiction. However, the sum network can be written as a two parameters 1-layer NN with $W'' = W \otimes W'$ and $U = (1, 1)$.

However, as is the case in the example detailed above, it can be the case that when the number of parameters increases, the space turns into a linear space. For instance in the case of quadratics $\sigma(z) = z^2$, then let $\lambda \in \mathbb{R}$, $f, g \in V_{\sigma,4}$ and write $h := f + \lambda g = h_1x_1^2 + h_2x_2^2 + h_3x_1x_2$ and we want to check if there exists parameters $U^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and W such that

$$h = U^T\sigma(Wx) = \alpha_1\langle W_1, x\rangle^2 + \alpha_2\langle W_2, x\rangle^2 + \alpha_3\langle W_3, x\rangle^2 + \alpha_4\langle W_4, x\rangle^2$$

. A solution is given by taking the first two terms to get the coefficients in front of the squares and the last term will express the cross term (x_1x_2). So we can pick $W_1 = (1, 0), W_2 = (0, 1)$ and $\alpha_1 = h_1, \alpha_2 = h_2$. For the products we can pick $W_3 = (1, 1), W_4 = (1, -1)$, and $\alpha_3 = \frac{1}{4}h_3, \alpha_4 = -\frac{1}{4}h_3$.
This gives us that for any scalar $\lambda \in \mathbb{R}$ and 1-layer NN $\Phi_1, \Phi_2 \in V_{\sigma,4}$, there exists at least one parameter $\theta = (U^T, W)$ constructed as above such that the 1-layer NN $\Phi_\theta \in V_{\sigma,4}$ satisfies

$$\Phi_\theta = \Phi_1 + \lambda\Phi_2.$$

Hence, $V_{\sigma,4}$ is a linear space for $\sigma(z) = z^2$. This construction also gives us for any $p \geq 4$,

$$V_{\sigma,4} = V_{\sigma,p}.$$

Note that here we only have a sufficient lower parametrization for the space to be a linear space, it is in fact not a minimal parametrization. It should also be noted that there is no unique representation for any element in any of the parametrisations since σ is homogeneous. This is an important thing to notice as having (continuous) group actions on the parameters that do not change the resulting Neural Network will be of great use for optimisation. A more explicit use of those symmetries seems to be a promising direction of research to improve, and perhaps generalize, some results from the third section.

Following the same strategy, of reconstructing each possible monomial at a time, it is possible to show the following result for general polynomial activation functions:

Proposition 2.1. *Let σ be a continuous activation function and $X \in \mathcal{R}_2(\sigma, n)$ such that $\dim(L^2_X) = \infty$. If $\sigma(z) = \sum_{k=0}^d a_kz^k$, then*

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

Let's also consider the following result that will be proved, and used, later in the report

Proposition 2.2. *Let $\sigma(z) = z^k$ with k a positive integer. Then*

$$\dim_*(\sigma, n) = rk_S(k, n)$$

where rk_S is the maximal symmetric tensor rank. In particular, if $\dim_*(z \mapsto z, n) = 1$ and $\dim_*(z \mapsto z^2, n) = n$.

2.3 Definitions

Now we introduce some useful definitions to study the functional spaces encountered above. We define

$$V_\sigma = \bigcup_{p=1}^{\infty} V_{\sigma,p}$$

as the linear space by the filter functions $\psi_{\sigma,v} = \sigma(\langle\cdot, v\rangle)$. Now we introduce the most general family of distributions for X that we can consider in this setting as

$$R_2(\sigma, n) = \{X \text{ r.v. taking values in}\mathbb{R}^n : \psi_{\sigma,v} \in L^2(\mathbb{R}^n) \text{ for every } v \in \mathbb{R}^n\}.$$

Using these we can define

- the upper intrinsic dimension of a pair (σ, X) as

$$dim^*(\sigma, X) = dim_{L^2(X)}(V_\sigma) \tag{10}$$

- the upper intrinsic dimension of a pair (σ, n) as

$$dim^*(\sigma, n) = dim(V_\sigma) = sup\{dim^*(\sigma, X) : X \in R_2(\sigma, n)\} \tag{11}$$

- the lower dimension of a pair (σ, X) as

$$dim_*(\sigma, X) = max\{p \geq 1 : V_{\sigma,p-1} \neq V_{\sigma,p}\} \tag{12}$$

- the lower dimension of a pair (σ, n) as

$$dim_*(\sigma, n) = sup\{dim_*(\sigma, X) : X \in R_2(\sigma, n)\} \tag{13}$$

Hence, dim^* corresponds to the (minimal) number of coordinates needed to represent any finite linear combination of $\psi_{\sigma,v}$ (for any choices of v), hence for any $X \in R_2(\sigma, n)$ there exists I such that $card(I) = dim^*(\sigma, X)$ and

$$\exists (e_i)_{i\in I} \in L^2(X) : \forall f \in V_\sigma, \exists (\lambda_i(f))_{i\in I} \in \mathbb{R}, f = \sum_i \lambda_i(f) e_i.$$

On the other hand, dim_* corresponds to the smallest number of parameters needed to represent any 1-layer NN with an arbitrary number of parameters, explicitly for any $X \in R_2(\sigma, n)$ there exists I such that $card(I) = dim_*(\sigma, X)$ and

$$\forall f \in V_\sigma, \exists (\psi_{\sigma,v_i})_{i\in I} \in L^2(X) : \exists (\lambda_i(f))_{i\in I} \in \mathbb{R}, f = \sum_i \lambda_i(f) \psi_{\sigma,v_i}.$$

It is then clear that dim_* is the smallest number such that any element of V_σ can be written as a linear combination of dim_* elements from $V_{\sigma,1}$, whereas dim^* claims the existence of dim^* elements from $L^2(X)$ such that any $f \in V_\sigma$ can be written as a linear combination of those elements. This gives us the inequality

$$dim_*(\sigma, X) \leq dim^*(\sigma, X). \tag{14}$$

Another fact we can obtain is that if $V_{\sigma,dim_*(\sigma,X)} := V_*$ is a linear space then $dim^*(\sigma, X) \leq dim_*(\sigma, X)$ so that we have the equality $dim^*(\sigma, X) = dim_*(\sigma, X)$.
Another point of view on this situation is that V_* is by definition the space of $dim_*(\sigma, X)$-sparse elements of V_σ and it contains all of V_σ.

2.4 Universal approximation theorem

The following result is an essential result in Neural Networks as it states that even the simplest networks will have the capacity to express a large range of functions when the activation function is not a polynomial.

Theorem 2.1. *$dim^*(\sigma, n) < \infty \iff \sigma$ is a polynomial.*

Proof. The fact that if σ is a polynomial then $dim^*(\sigma, n) < \infty$ is easy. It suffices to observe that if $deg(\sigma) < d$, then $V_\sigma \subset \mathbb{R}_d[X = (X_1, \cdots, X_n)]$ as a vector space, where $\mathbb{R}_d[X]$ is the vector space of polynomials of degree $\leq d$, since the RHS is generated by the elements $(X_1^{\alpha_1} \cdots X_n^{\alpha_n})_{\sum \alpha_i \leq d}$ we have that $dim^*(\sigma, n) \leq C \sum_{i=1}^d n^i$.

The other direction is proved as follows[1], where the claim is "if σ is not a polynomial then $dim^*(\sigma, n) = \infty$", equivalently, "$dim^*(\sigma, n) < \infty$ only if σ is a polynomial":

1. If $V_{\sigma,n=1}$ is dense in $C^0(\mathbb{R})$, then $V_{\sigma,n=n}$ is dense in $C^0(\mathbb{R}^n)$.[2]

[1] Leshno and al., Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, *Neural Networks*, 1993

[2] This allows us to ignore the input dimension for the rest of the proof.

2. If σ is not a polynomial and $\sigma \in C^\infty$, then V_σ is dense in C^0.[3]

3. If $\varphi \in C_c^\infty$ (i.e. φ is compactly supported and C^∞) and σ is not a polynomial, then $\sigma \star \varphi \in \bar{V_\sigma} := closure(V_\sigma)$.[4]

4. If for some $\varphi \in C_c^\infty$ we have that $\sigma \star \varphi$ is not a polynomial, then V_σ is dense in C^0.[5]

5. If for all $\varphi \in C_c^\infty$, $\sigma \star \varphi$ is a polynomial, then there exists an $m \in \mathbb{N}$ such that $\sigma \star \varphi$ is a polynomial of degree at most m for all $\varphi \in C_c^\infty$.[6]

6. If $\sigma \star \varphi$ is a polynomial of degree at most m for all $\varphi \in C_c^\infty$, then σ is a polynomial of degree at most m almost everywhere.[7]

□

The following result shows that for many continuous activation functions the previous result also holds for the lower intrinsic dimension, which is even stronger.

Theorem 2.2. *Let σ be a continuous activation function such that $\sigma \in L^2(\mathbb{R}, e^{-\frac{x^2}{2}}dx)$ and $n > 1$. Then $dim_*(\sigma, n) = \infty \iff \sigma$ is not a polynomial.*

3 Spurious valleys

3.1 Introduction to spurious valleys

One of the main current goals of the mathematical study of neural networks resides in understanding its good behaviour when performing optimization. It has indeed been observed since early 2010's that Machine Learning could provide a much better performance on specific tasks (audio/video processing, Natural Language Processing) compared to state of the art algorithms (based in a way on "classical statistics"). However that performance is even today not fully backed by theory.

Prior to the arrival of modern machine learning, based on (convolutional) neural networks, when one had a problem similar to a regression problem, the theory would usually provide solutions (with provable guarantees) when the problem under study was convex. Indeed, when assuming convexity the problem of finding minimum values is relatively easy, the existence is guaranteed as well as the fact that if one keeps decreasing the loss, the minimal value for the loss will be attained. In such situation the difficulty resided in designing algorithms that could potentially converge fast to a solution.

In a way, the "unreasonable effectiveness of neural networks" came from their simplicity, in construction and in optimization. The construction has been described above, NN generate a functional space, and it has been known for a long time that these could be very rich for most (non polynomial) activation function. However, a drawback of these large functional spaces makes the task of optimizing parameters much harder. The first (minor) difficulty comes from the task of handling very large dimensional spaces. The main difficulty that has to be tackled is the fact that the activation functions are usually neither linear nor convex. Hence the fact that NN worked well meant that optimizing a neural network consisted in a non-convex optimization problem, but the final solution was still close from a true solution.

In order to provide an explanation of this success, and possibly gain insight on the problem, the studied paper gives a description of the landscape under exploration during optimization. When performing minimization, the goal is to minimize the loss function

$$L(\theta) = \mathbb{E}[l(\Phi(X;\theta), Y)] \tag{15}$$

[3] Because V_σ contains the algebra of polynomials (of all degrees)

[4] Even if σ is not continuous, $\sum_{i=1}^m \sigma(x - y_i)\varphi(y_i)\Delta_i$ uniformly converges to $\sigma \star \varphi$ and at each step is in V_σ (more precisely in $V_{\sigma,p=m}$)

[5] Because $C^0 = V_{\sigma\star\varphi}^- \subset \bar{V_\sigma}$ where the first equality comes from step 2, and the last inequality comes from the previous step.

[6] We consider this because the previous step implies that if we want $dim^*(\sigma, n)$ to be finite then for all $\varphi \in C_c^\infty$, $\sigma \star \varphi$ has to be a polynomial.

[7] With the previous step, this implies that if $dim^*(\sigma, n) < \infty$ then σ is a polynomial.

where the parameters θ are in some space Θ (usually $\Theta = \mathbb{R}^M$), and l is *convex in its first variable.* Then the landscape where we perform optimization is the graph given by $(\theta, L(\theta))_\theta$. When exploring this landscape (similarly to gradient descent or its variant) we will consider the naive way of performing loss minimization, namely optimization is performed by continuously moving along paths of decreasing energy.

To get a picture of the situation, one can think of the optimization landscape as made of mountains, hills, valleys, plains and the sea. Using the optimization process, the only way we may move in this landscape from an initial position is only by either staying at the same altitude, or going downhill. The questions under investigation are of the form :

1. Starting from some initial position θ_0, is it possible to reach θ^* the optimal solution (the sea) ?
2. Starting from some initial position θ_0, how far (in "altitude") can we get from θ^* ?

for varying settings of the problem (e.g. choice of the activation function, of the random variables,...). The main goal of the article is to propose (partial) answers to the above questions.

From the description of the situation made above it is very natural to define from a given parameter θ_0 the set of parameters that can be reached as

$$\Omega_{\theta_0} := \{\gamma(1) : \gamma \in \Gamma_{\theta_0}\} \tag{16}$$

where

$$\begin{aligned}\Gamma_{\theta_0} :&= \{\gamma \in \mathcal{C}^0([0,1], \Theta) : \gamma(0) = \theta_0 \text{ and } L \circ \gamma \text{ is non-increasing}\} \\ &= \{\text{Descent paths starting at } \theta_0\}.\end{aligned}$$

Thus, we can rephrase the two previous questions as:

1. $\theta^* \in \Omega_{\theta_0}$?
2. $\inf_{\theta \in \Omega_{\theta_0}} |L(\theta^*) - L(\theta)| =?$

An alternate way of considering the problem consists in studying sub-level sets, i.e. for $C \geq 0$ we consider $\Omega_C := \{\theta : L(\theta) \leq C\} = L^{-1}([0, C])$. With this perspective, when given some initial parameter θ_0, we denote $C_0 = L(\theta_0)$, the set of feasible parameters consists of the path connected component of Ω_{C_0} that contains θ_0, i.e. Ω_{θ_0}. Remembering that we are interested on whether or not we can reach global minimum from some initial position motivates the following definition:

Definition 3.1. *A* spurious valley *is a path connected component*[8] *of a sub-level set which does not contain a global minima.*

Using the above discussion we can obtain the following result:

Theorem 3.1. *If L is continuous, then the following statements are equivalent:*

1. *There exists no spurious valley*
2. *For all $C > 0$ and any (maximal) connected component U of Ω_C, U contains a global minima*
3. *For any initial parameter, Ω_{θ_0} contains a global minima*

[8] In the maximal sense, i.e. a path connected component U of Ω_C is such that adding another point implies either that the loss at this point is strictly larger than C, or that there exists no *continuous* descent path between the new point and any point of U

3.2 No spurious valley in the over-parametrised regime

With this result we can then state the main result of the article and sketch the idea of the proof:

Theorem 3.2. *For any continuous activation function σ and r.v. $X \in \mathcal{R}_2(\sigma, n)$ with finite upper intrinsic dimension $dim^*(\sigma, X) < \infty$, the loss function*

$$L(\theta) = \mathbb{E}[l(\Phi(X;\theta), Y)] \tag{17}$$

for one hidden layer NNs $\Phi(x;\theta) = U\sigma(Wx)$ admits no spurious valleys in the over-parametrised regime $p \geq dim^*(\sigma, X)$.

The idea of the proof is as follows, since the upper intrinsic dimension is finite, it is possible to view V_σ as a finite dimensional space. Then, since the activation function is continuous we know that for a given θ_0 we can consider the set Ω_{θ_0} and we want to check whether it contains a global minima.
Thus, we construct a first half of a path that turns θ_0 in $\theta_{1/2} = (U_{1/2}, W_{1/2})$, on which the loss doesn't increase, such that the neurons of the network $(\{\psi_{\sigma,(W)_{i=1,\cdots,p}}\})$ form a basis of V_σ. In order to construct the second half of the part, it is sufficient to observe that any minimizer f^* can be written as a linear combination of elements from the basis above, i.e. $f^* = U_1\sigma W_{1/2}$, then to remark that since l is convex in its first variable the path $t \mapsto ((1-t)U_{1/2} + tU_1, W_{1/2}$ is a descent path. The latter remark being true since we are acting on the vectors $(\psi_{\sigma,w})$ ignoring the potential source of difficulties, σ.

Some remarks are due on this sketch of proof. The first half of the path works by "aligning"[9] the neural network with a basis of V_σ; whilst modifying the coefficients of W to align with the basis is not difficult, work has to be done to check that by choosing an appropriate path for U, the path can be a descent path. On the other hand the construction of the second half of the path is straightforward, since *any* minimizer can be written in any basis, reaching a minimizer can be simply done (in a linear space) by translation of the coefficients of the previously obtained coefficients to the coefficients of a minimizer. The fact that the translation is a descent path comes from the hypothesis that l is convex in its first variable.

A consequence of the simplicity of the second half of the path is that *any* minimizer can be attained from *any* initial choice of parameters. However this should not be mistaken as meaning that the energy landscape is trivial, i.e. if $L(\theta_0) = C_0$, then $\Omega_{\theta_0} = \Omega_{C_0}$, which is not true in general. Rather, what is indicated is that once our parameters are aligned with a basis, let's call them $\theta_{1/2}$, optimization is straightforward. Observing that $\Omega_{C=0}$, the set of minimizers, is a convex set, then the whole convex hull defined by $(\theta_{1/2}, \Omega_{C=0})$ is a valid area for optimization, i.e. contained in Ω_{θ_0}.

From this discussion, the role played by a finite intrinsic dimension is clear, it allows to turn the problem into a simple finite dimensional linear algebra problem. Using this fact, we can combine what we learned when introducing the spaces V_σ and $\mathcal{R}_2(\sigma, n)$ into the following corollaries:

Corollary 3.1. *Let $\sigma(z) = a_0 + a_1 z + \cdots + a_d z^d$ a polynomial and $V_{\sigma,p}$ the space of one layer NNs with p parameters. Then the loss function $L(\theta) = \mathbb{E}[l(\Phi(X;\theta), Y)]$ admits no spurious valleys in the over-parametrized regime*

$$p \geq \sum_{i=1}^{d} \binom{n+i-1}{i} 1_{a_i \neq 0} = \mathcal{O}(n^d). \tag{18}$$

Corollary 3.2. *Let $V_{\sigma,p}$ the space of one layer NNs with p parameters and σ is any continuous activation function and $(x_i, y_i)_{i=1,\cdots,N}$ be N data points[10]. Then the empirical loss function*

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} l(\Phi(x_i;\theta), y_i) \tag{19}$$

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

[9] We loosely call a set of vectors "aligned", here with a basis, in the sense that each element of the basis is a scalar multiple of exactly one of the "aligned" vector.The idea of alignment arising from the fact that when constructing the first half of the path the, usually already spanning family, has to be turned in a family which is here also free. In the process vectors vectors are aligned, or merged, into each other, in order to cancel some of the vectors.

[10] One could also pick any finitely supported distribution

Both corollaries are straightforward consequences of the theorem. The first one since for arbitrary distributions the upper intrinsic dimension $dim^*(\sigma, n) = \sup_{X \in \mathcal{R}_2(\sigma,n)} dim^*(\sigma, X)$ is finite (only) for polynomial activation functions, and the dimension of this space is (upper) bounded by the quantity given in the corollary. The second one is true since for empirical distributions on N points we have that $dim^*(\sigma, X) \leq dim^*(L^2(X)) \leq N$.

Thus the corollaries are true because in those specific situations, and only on them, we can find that the upper intrinsic dimension is finite, making it possible to enter the over-parametrized regime.

3.3 Improved bounds on linear and quadratic networks

We have seen in the previous part that when (and only when) the activation function is a polynomial there exists no spurious valleys for arbitrary distributions whenever the number of parameters p is larger than the upper intrinsic dimension $dim^*(\sigma, n)$. We will now see that in some situations it is possible for specific networks to have a smaller bound than the upper intrinsic dimension to guarantee that there exists no spurious valley.

The first result we discuss concerns the simplest activation function $\sigma(z) = z$ but for an arbitrary number of layers. This setting corresponds to networks of the form

$$\Phi(x, \theta) = M_1 \cdots M_N x \tag{20}$$

where $\theta = (U_1, W_1, \cdots, U_N, W_N)$ with matching matrix dimensions and writing $M_i = U_i W_i$. This setting which is more a toy model than a model that should be used in practice will help us to see how to use previously obtained results and improve upon them.

First, because of the linear activation function, it is possible to turn the model 20 into any 1 hidden layer NN with activation function $\sigma(z) = z$ as

$$\Phi(x, \theta) = \tilde{U}\tilde{W}x \tag{21}$$

for any number of parameter p_i matching the number of parameters of one of the hidden layers in 20. Hence, it is possible to apply the main theorem 3.2 for 1 hidden layer NNs to get that if one of the hidden layer of 20 is over-parametrised, i.e. there exists p_i such that $p_i \geq dim^*(\sigma, n) = n$, then the linear model admits no spurious valley.

However, it is possible to obtain a much stronger result

Proposition 3.1. *Linear models 20 with any depths, number of parameters and output dimensions all ≥ 1, admit no spurious valley for the square loss $L(\theta) = \mathbb{E}||\Phi(X, \theta) - Y||_2^2$.*

In the article, authors propose a detailed and lengthy proof of this result. The version presented here is much shorter and inspired of the proof of the main theorem 3.2.

Restricting to one-dimensional output, the first observation is that the lower intrinsic dimension of such a network is equal to 1. The previous sentence corresponding simply to the fact that any real valued linear function can be written as a dot product with a well chosen vector (Riesz theorem). Because of the linearity of the network, any minimal solution in V_σ is also written as a dot product with a vector (Hahn-Banach theorem). Hence, the linear translation between those two vectors is a descent path (that remains in $V_{\sigma,1}$).

The second improvement on the bound of 3.2 is for 1 hidden layer NNs with a quadratic activation function $\sigma(z) = z^2$ with one dimensional output, which are called *quadratic neural networks*. For such networks, 3.1 gives us that whenever $p \geq \frac{n(n+1)}{2}$ then there exists no spurious valley. It is in fact possible to prove the following result:

Proposition 3.2. *For one-hidden-layer with quadratic activation function $\sigma(z) = z^2$ and one dimensional output, the square loss function admits no spurious valley in the regime $p \geq 2n + 1$.*

We present here the sketch of the proof that uses the idea of the proof of the main theorem together with the singular value decomposition (SVD). The first step consists in rewriting the output of the NN as

$$\Phi(x, \theta) = \sum_{i=1}^{p} u_i(\langle w_i, x \rangle)^2 = \langle \sum_{i=1}^{p} u_i w_i w_i^T, x x^T \rangle_F \tag{22}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. We now use the SVD to rewrite

$$\sum_{i=1}^{p} u_i w_i w_i^T = \sum_{i=1}^{n} \sigma_i v_i v_i^T \tag{23}$$

thus any output of the network can be written using n vectors (which proves that the lower intrinsic dimension of quadratic network is bounded above by the input dimension). From there it is easy to conclude, since an optimal solution can also be expressed using n units for some coefficients, it is sufficient that $p \geq 2n$ to first translate coefficients of a p parameter network to a n parameter network while keeping the output constant, then to translate the n nonzero coefficients to the optimal coefficients/units, which is decreasing by convexity of the loss in U.

In fact, both those results and proofs hint at the following conjecture:

Conjecture 3.1. *One-hidden-layer networks with activation function σ admit no spurious valleys in the regime $p \geq \mathcal{O}(dim_*(\sigma, n))$.*

Some words should be said about this conjecture. The first remark is that it corresponds to previous results, also the authors conjecture that for quadratic networks, the constant 2 is probably not necessary and is more likely to be an artefact of the proof used. Also, as mentioned by the authors, it could be the case that the conjecture could be settled by exploiting network symmetries more carefully.
A more important remark is that, like theorem 3.2, for general distributions, it only gives useful information for polynomial activation functions. Indeed, if $dim^*(\sigma, n) = \infty$, then V_{σ,dim^*} is dense in the space of continuous function, so that V_{σ,dim_*} is also dense in continuous functions, hence $dim_* = \infty$. This result points to the direction that more sensitive definitions of intrinsic dimensions could be needed for more general results.

3.4 Arbitrarily bad spurious valleys exist in general

Now that we have detailed some results for one-hidden-layer networks with finite upper intrinsic dimension, equivalently with polynomial activation functions, we will consider until the end of this report networks with non polynomial activation functions. It is important to note that this situation is probably the most important to consider for application purposes, since all activations functions used in practice are non polynomial.

Whilst in the first part of the report we have mostly seen positive results, stating that it is possible to minimize the loss starting with arbitrary initial parameters. We will now present some negative results that show that it is possible to construct distributions with arbitrarily high spurious valleys in some regimes.

Theorem 3.3. *Consider the square loss function for one-hidden-layer NNs 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 random variable (X, Y) such that the square loss function L admits spurious valleys. In particular, for any given $M > 0$, the random variable Y can be chosen in such a way that there exists a (non-empty) open set $\Omega \subset \Theta$ such that*

$$M/2 + \min_{\theta \in \Omega} L(\theta) \geq \sup_{\theta \in \Omega} L(\theta) \geq \min_{\theta \in \Omega} L(\theta) \geq M + \min_{\theta \in \Theta} L(\theta) \tag{24}$$

and any path $\theta : [0,1] \to \Theta$ such that $\theta_0 \in \Omega$ and θ_1 is a global minima verifies

$$\max_{t \in [0,1]} L(\theta_t) \geq \min_{\theta \in \Omega} + M. \tag{25}$$

This theorem can be summarized as "arbitrarily high spurious valleys can exist" and "spurious valleys can be arbitrarily hard to escape", we refer to such a situation as having "arbitrarily bad spurious valleys". As mentioned, this results applies to most of the one-hidden-layer neural network models that could be used in practice (like with ReLU or sigmoid activation function) when the input dimension satisfies $n \geq 2$.

Another consequence that can be derived is that even for models with finite lower intrinsic dimension, if the network doesn't have enough parameters, it can have arbitrarily bad spurious valleys. For instance it is possible to get the following result for quadratic neural networks.

Corollary 3.3. *For any one-hidden-layer NN with quadratic activation function $\sigma(z) = z^2$ with input dimension n, is the number of parameters satisfies $p \leq n - 1$, it exists a random variable (X, Y) with arbitrarily bad spurious valleys for the square loss function.*

It is useful to compare this result with the proposition 3.2, or conjecture 3.1, to see that, up to the constant 2, results obtained are tight.

3.5 Spurious valleys get lower when parameterisation increases

In previous sections we have seen that in situations where over-parametrisation is feasible it is possible to prevent the existence of spurious, and we have also seen a converse statement, if over-parametrisation is not feasible, or if the network is under-parametrised, then it is possible to construct specific distributions for which spurious valleys exist.

The latter being the usual situation (since 3.3 applies for instance to ReLU, softplus or sigmoid activation functions) it is of high interest to obtain some information on how likely it is to fall in a spurious from randomly chosen parameters. Another key quantity in such a situation would be to obtain information on how high such a spurious valley would be (i.e. how much higher the loss for the best choice of parameters in the spurious is from the minimal loss among all choices of parameters).

We will consider in this part the square loss function for one dimensional output one-hidden-layer NNs and X, Y square integrable distributions. The first thing we do, which is classical in statistics, is to split the loss as

$$L(\theta) = \mathbb{E}|\Phi(X;\theta) - f^*(X)|^2 + \mathbb{E}|Y - f^*(X)|^2. \tag{26}$$

In this decomposition, we set $f^*(X) = \mathbb{E}[Y|X]$ as the best solution, so that we have one term which corresponds to the distance between best solution and the output distribution (bias), and the other is the distance between the best solution and the network output (variance).

In particular, the previous decomposition gives a global lower bound on the loss as

$$\min_{\theta\in\Theta} L(\theta) \geq \mathcal{R}(X, Y) := \mathbb{E}|Y - f^*(X)|^2. \tag{27}$$

Hence, in order to know how accurate our network is, we only need to focus on the first term, and then check how far apart $L(\theta)$ is from $\mathcal{R}(X, Y)$.

Now, if we assume that f^* can be represented as a one-hidden-layer NN with an arbitrary number of parameters, it can be written as

$$f^*(x) = \int_\Theta \sigma(\langle x, w\rangle)\rho(w)d\mu(w) \tag{28}$$

with an appropriate choice of measure μ and weight function ρ. Those choices of μ, ρ corresponding to the choice of layers of arbitrary size where ρ corresponds to U and μ to the choice of optimal weights.

From this expression of f^*, the works of Francis Bach on random features expansion[11] can be applied. The idea behind random feature expansions (in non technical terms) is that instead of trying to pick the best solution at once, one instead chooses random parameters and depending on how well they improve the resulting approximation (through a well chosen dot product), they are added in the reconstruction formula. Allowing an arbitrary number of units allows to reach an arbitrary precision, thus in our setting, the non-asymptotic results from Bach will allow us to quantify how much increasing parametrisation increases accuracy.

If we assume that f^* can be written as above for some density ρ, then if we draw p weights w_i i.i.d on the unit sphere we have

$$\mathbb{E}\left(\frac{1}{p}\sum_{i=1}^{p}\rho(w_i)\sigma(\langle w_i, x\rangle) - f^*(x)\right)^2 = \mathcal{O}(\frac{1}{p}). \tag{29}$$

Now, it is possible to notice that even if in practice ρ is unknown, from any coefficients in the second layer it is possible to translate them to the "true" coefficients $\rho(w_i)$, and this consists of a linear descent path to the optimal solution.

[11] F. Bach, On the equivalence between Kernel Quadrature Rules and Random Feature Expansions, *Journal of Machine Learning Research*, 2017

We will now give two results that can help to quantify the needed number of neurons for a desired accuracy with high probability.

Theorem 3.4. *Let $d\tau$ be the uniform distribution over the unit sphere $\mathbb{S}^n$ and consider an initial parameter $\tilde{\theta} = (\tilde{u}, \tilde{W})$ with $w_i \sim d\tau$ sampled i.i.d. Then the following hold*

1. *There exists a path $t \in [0,1] \mapsto \theta_t$ such that $\theta_0 = \tilde{\theta}$, the function $t \mapsto L(\theta_t)$ is non-increasing and*

$$L(\theta_1) \leq \mathcal{R}(X,Y) + \lambda \tag{30}$$

   if $p \geq \mathcal{O}(-\frac{1}{\lambda}\log(\lambda\delta))$ with probability $\geq 1-\delta$ for every $\lambda, \delta \in (0,1)$.

2. *Assume f^* can be written as in 28 with $d\mu = d\tau$ and $\rho \in L^\infty_d \tau$, then there exists a path $t \in [0,1] \mapsto \theta_t$ such that $\theta_0 = \tilde{\theta}$ and the function $t \mapsto L(\theta_t)$ is non-increasing and*

$$L(\theta_1) \leq \mathcal{R}(X,Y) + \mathcal{O}(p^{-1+\delta}) \tag{31}$$

   with probability $\geq 1 - e^{-\mathcal{O}(p^\delta)}$ for every $\delta \in (0,1)$.

The proof of the latter statement using the fact that since ρ is bounded, one can apply a Hoeffding-type inequality to get the result. The former statement is a consequence of Proposition 1 of Bach in random features expansions.

4 Conclusion and future directions

In conclusion, the presented paper provides a comprehensive overview of the optimisation landscape for one-hidden-layer networks.

The first results give guarantees for success of reaching a globally optimal choice of parameters for empirical risk minimization or polynomial activation functions. Those results hint at the fact that increasing the number of parameters increases the number of global descent paths. A key fact of these results is that they only use two properties of Neural Networks, namely the upper and lower intrinsic dimensions.

The second kind of results concern situations where the network is under-parametrised, there it is proved that for the most common activation functions, it is possible to find data distributions such that spurious valleys are arbitrarily bad. A similar result for quadratic networks also hints at the fact that when the number of network parameters is smaller than its lower intrinsic dimension, then such bad data distributions can also exist.

The last part is perhaps more optimistic as it is shown that even if reaching a global minimal choice of parameters is not necessarily feasible from any initial parameters, it is still possible to get arbitrarily close to those with high probability for a sufficiently large number of hidden units.

Although those results provide a good description of the landscape of one-hidden-layer NNs from a theoretical perspective, a lot of work is still to be done. For instance, the work only concerns presence or absence of spurious valleys, which can certainly be a barrier for reaching a global minimum, but no discussion is made on the existence or absence of strict descent path. The latter being more frequently encountered in practice. Notably, the current work doesn't give any method that could be of use in practical application. An instance of complementary work would be to check under polynomial activation function for what range (if any) of parametrization does SGD converge to a global minima[12].

Another possible direction of future work would be, as mentioned several times above, to consider more explicitly the role of symmetries of the spaces and activation functions under consideration. For

[12]It is possible to give the following heuristic for a double descent phenomenon to happen in over-parametrized models if SGD follows a path similar to the descent paths constructed in the proof. In a first time, SGD would be attracted by a local minima not too far from the initial parameters. Then to get from this point convergence to a better minima, the path would have to align with an optimal solution. However during this alignment, in the proofs we only show that, having enough symmetries, it is possible to keep the loss constant. In practice, it could be hard to keep this loss constant and it might grow during alignment. However, once alignment has been performed there is an easy optimizing path, so that loss will again decrease rapidly. (This is purely conjectural and both numerical and theoretical studies should be performed to check if the claim is valid)

instance, in the case of positively homogeneous functions (like monomials), then it is possible for the multiplicative group $\mathbb{R}^{+*}$ (and even $\mathbb{R}_{+*}^{p}$ as diagonal matrices) to act on V_σ such that every element of V_σ has the same output before and after having acted.

There are two clear advantages to take into account these symmetries. First, when constructing the paths as in the proof of the main theorem, having a large dimensional symmetry group allows to move in the space of parameters in many different directions while keeping the output constant. Remembering that using this we construct the first half of the path to align our parameters with an optimal solution, we get the second advantage: there are more global solutions with which we can try to align. Indeed, since acting on the parameters with a group of symmetries as above allows to get new parameters that lead to the same output. An additional consequence of this, is that since l is convex in its first variable, the set of optimal solutions is convex. Overall, increasing symmetries allows the alignment process to be easier, and also the alignment path should be shorter as there are more optimal solutions that can be reached. This leads to the following conjecture :

Conjecture 4.1. *For an appropriate definition*[13] *of symmetric group G associated to V_σ, then $V_{\sigma,p}$ admits no spurious valley when*

$$p \geq \frac{\dim^*(\sigma, n)}{\dim(G)}.$$

[13]The group action under consideration is $g \in G$ acts on $\Phi_\theta \in V_\sigma$ as:

$$g \cdot \theta = (g_1, g_2) \cdot (U, W) = (Ug_1, g_2W)$$

and the group is symmetric for V_σ if $\Phi_\theta = \Phi_{g\cdot\theta}$ for all $\Phi_\theta \in V_\sigma$.
For positively k-homogeneous activation function σ (such as degree k homogeneous polynomial), then $G = \{(Diag(\lambda_1, \cdots, \lambda_p)^{-1}, Diag(\lambda_1, \cdots, \lambda_p)^{1/k}) : \lambda_i \in \mathbb{R}_{+*} \forall i\}$ is a symmetric group for V_σ of dimension p. To get a group defined without reference to the number of parameters, it is likely that the action should be restricted to $V_{\sigma,\dim_*(\sigma,n)}$, so that the symmetric group associated to V_σ is the group of diagonal matrices of size n with positive entries. If the conjecture is true this would lead to the result that over-parametrization of k-homogeneous polynomials happens for $p \geq \mathcal{O}(n^{k-1})$, which is conjectured in the article.