Spurious Valleys in One-hidden-layer Neural Networks, Optimization Landscapes

By L. VENTURI, A. BANDEIRA and J. BRUNA

In *JMLR, 2019*

Leo Davy

ENS Lyon
M2 Advanced Mathematics

March 2022

Current situation of ML

- There exists random variables (X, Y) such that $Y = f(X)$
- There exists models $\Phi_\theta : X \mapsto \Phi_\theta(X)$
- There exists some optimisation methods $\Phi_\theta \mapsto \Phi_{\tilde{\theta}}$
- Such that $L(\Phi_{\tilde{\theta}}, Y) \sim 0$ (L could be MSE, log-likelihood, ...)

A lot of blackboxes... and very few guarantees...

Optimization landscape

For l a convex function in its first variable, we define the loss as:

$$\theta \in \Theta \mapsto L(\Phi_\theta(X); Y) := \mathbb{E}_X l(\Phi_\theta(X), Y) := L(\theta).$$

Θ the parameter space ($\mathbb{R}^P$, $P \gg 1$)

Goal

Understanding the optimization landscape for simple models

Optimization paths

Starting from some initial parameter $\theta_0 \in \Theta$

- discrete : find $\theta_1, \cdots, \theta_N$ s.t. $L(\theta_{k+1}) \leq L(\theta_k)$
- continuous : find a continuous path $t \in [0,1] \mapsto \theta_t \in \Theta$ is non-increasing

Definition (descent path)

We call a descent path, a path $t : [0,1] \to \Theta$ that satisfies the two assumptions

- $t \mapsto \theta_t$ is continuous
- $t \mapsto L(\theta_t)$ is not increasing

The last property is called *no up-hill climb property*.

Problem

Depending on $(X, Y), \theta \mapsto \Phi_\theta$ and l, is there for any initial parameter θ_0 a descent path that reaches a global minima ?

- Does the optimization landscape contain a spurious valley?

Definition (spurious valley)

A *spurious valley* is a maximally descent-path-connected component that doesn't contain a global minima.

Model considered

One-hidden layer Neural Networks (NNs) with continuous activation function σ.

$$X \mapsto Wx \mapsto \sigma(Wx) \mapsto U\sigma(Wx) = \Phi_{\theta=(U,W)}(X)$$

- *activation function*: $\sigma : \mathbb{R} \to \mathbb{R}$ continuous and acts component-wise on $\mathbb{R}^p$
- *filter functions*: $\psi_{\sigma,w}(x) \mapsto \sigma(\langle w, x\rangle)$
- *parameters*: $\theta = (U, W) \in \mathbb{R}^{m\times p} \times \mathbb{R}^{p\times n}$

Additional assumptions : $m = 1$, $l : \mathbb{R}^m \times \mathbb{R}^m$ *convex in its first variable*, $X \in R_2(\sigma, n) = \{X : ||\sigma(w, \cdot)||_{L^2(X)} < \infty, \forall\theta\}$

Functions expressible by:

- p parameters:

$$V_{\sigma,p} = \{\Phi_{\sigma,\theta} : \theta \in \Theta_p\} = \left\{\sum_{i=1}^{p} u_i \psi_{\sigma,w} : (U, W) \in \theta_p\right\}.$$

$V_{\sigma,p}$ is not a vector space in general.[1]

- an arbitrary number of parameters:

$$V_\sigma = \bigcup_{p=1}^{\infty} V_{\sigma,p} \quad \text{Usually a (big) vector space}$$

[1] Take $\sigma(z) = z^2$ and $X = (x, y)$, then $xy \in V_\sigma$ but $xy \notin V_{\sigma,1}$

Intrinsic dimensions

- lower intrinsic dimension:

$$\dim_*(\sigma, n) = \inf\{p : f \in V_\sigma \implies f \in V_{\sigma,p}\}$$

  i.e. the minimal number of parameters to express any function in V_σ

- upper intrinsic dimension:

$$\dim^*(\sigma, n) = \sup_{X \in R_2(\sigma,n)} \dim_{L^2(X)} V_\sigma$$

  i.e. the minimal number of parameters for $V_{\sigma,p}$ to be a linear space.

Examples

For general distribution X:

$$\sigma(z) = z \longrightarrow \dim^*(\sigma, n) = n$$
$$\longrightarrow \dim_*(\sigma, n) = 1$$

For finitely supported X on N atoms, i.e. $\mathbb{P}(X \in \{x_1, \cdots, x_N\}) = 1$:

$$V_\sigma \subseteq L^2(X) \cong \mathbb{R}^N$$
$$\longrightarrow \dim^*(\sigma, X) \leq N$$

Polynomial activation functions

- If $\sigma(z) = z^d$, then

  $V_\sigma = \{\text{homogeneous polynomial of degree } d \text{ in } X_1, \cdots, X_n\}$

  so,

$$dim^*(\sigma, n) = \binom{n+d-1}{d} = \mathcal{O}(n^d)$$

- If $\sigma(z) = \sum_{i=1}^{d} a_k z^k$, then

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

In particular, V_σ is of finite dimension if σ is a polynomial.

Universal approximation property

Theorem

Let σ a continuous activation function, then the following statements are equivalent:

- *For any continuous compactly supported f ($f \in \mathcal{C}_c(\mathbb{R}^n)$) and any $\varepsilon > 0$, there exists a number of parameters $p \geq 1$ and a one hidden-layer $\Phi_\theta \in V_{\sigma,p}$ satisfying*

$$||f - \Phi_\theta||_\infty < \varepsilon$$

- *σ is not a polynomial*

Corollary

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

Spurious valleys

Recall:

- goal: minimize $L(\theta) = \mathbb{E}l(\Phi_\theta(X), Y)$
- using descent path: $t \in [0, 1] \mapsto \theta_t = \gamma(t)$ s.t. $t_2 \geq t_1 \implies L(\theta_2) \leq L(\theta_1)$.

Denote $\Omega_{\theta_0} = \{\gamma(1) \in \Theta : \gamma \text{ descent path starting at } \theta_0\}$ (a "rooted valley")

Definition/Theorem

If L is continuous, then t.f.a.e.:

1. There is no spurious valley
2. $\forall C > 0$ and any maximal descent-path-connected component
$$U \subset \Omega_C = \{\theta : L(\theta) \leq C\},$$
U contains a global minima
3. $\forall \theta_0 \in \Theta$, Ω_{θ_0} contains a global minima

Theorem

If σ is continuous, $X \in R_2(\sigma, n)$, l convex in its first argument with $dim^*(\sigma, n) < \infty$, then

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

for one hidden-layer NNs Φ_θ has no spurious valley in the overparametrised regime

$$p \geq dim^*(\sigma, n)$$

Corollary

If σ is a polynomial, or if X is supported on a finite number of atoms, then overparametrisation is feasible.

Proof by constructing a path to a global minima in two parts

1. Treat V_σ as a finite dimensional vector space
   - Pick a basis $(w_i) = W_1$
   - Construct a path γ such that $\gamma(0) = \theta_0$ and $\gamma(1) = (U_1, W_1)$ for some U_1
   - Make this path such that $\forall t_1, t_2, \Phi_{t_1}(x) = \Phi_{t_2}(x)$
2. Optimize (very easily) using the last layer only
   - Pick a global minima and write it in the basis W_1

$$\Phi_{\theta^*=(U^*,W_1)} = U^*\sigma(W_1\cdot) = \sum_{i=1}^{p} u_i \psi_{\sigma,w_i}$$

   - Translate the coefficients of U_1 to those of U^*

$$\begin{aligned} L(\theta_t = ((1-t)U_1 + tU^*, W_1)) &= \mathbb{E}l((1-t)\Phi_{\theta_1} + t\Phi_{\theta^*}, X), Y) \\ &\leq (1-t)L(\theta_1) + tL(\theta^*), \quad \forall t \in [0,1]. \end{aligned}$$

Using *convexity in its first variable of the loss function l*, we have a descent path to a global minima

Treating V_σ as a f.d. vector space ?

It is not straightforward to consider V_σ as a finite dimensional vector space through W, the only interaction we can have with V_σ is through σ ! This problem is solved by using a Reproducing Kernel Hilbert Space (RKHS)

Lemma

If V_σ is finite dimensional, then there exist $\langle \cdot, \cdot \rangle$ and $\phi : \mathbb{R}^n \to V_\sigma \cong R^q$, where $q = \dim^*(\sigma, n)$, such that

$$\langle \psi_{\sigma,w}, \phi(x) \rangle = \psi_{\sigma,w}(x) = \sigma(\langle w, x \rangle).$$

Also, the map $w \in \mathbb{R}^n \to \psi_{\sigma,w}$ is continuous.

This gives us two maps $\phi, \psi : \mathbb{R}^n \to \mathbb{R}^q$ such that $\sigma(\langle w, x \rangle) = \langle \psi(w), \phi(x) \rangle$. Thus, we can rewrite $\Phi_\theta(X) = U\sigma(Wx)$ as

$$\Phi_\theta(x) = U\psi(W)\phi(x).$$

Now that we can rewrite our network in a *linearized* way:

$$\Phi_\theta(x) = U\psi(W)\phi(x)$$

where $\psi(W) \in \mathbb{R}^{p \times q}$.
From this, we don't want W to be a basis, but we want $\psi(W)$ to be a generating family (since $p \geq q = \dim^*(\sigma, n)$, we want $\operatorname{rank}(\psi(W)) = q$), i.e., we want the p rows of $\psi(w_i)$ of $\psi(W)$ to contain q linearly independent rows.
We can do as follows, with constant output:

- If $\operatorname{rank}(\psi(W)) < q$, W can be continuously mapped to $\psi(\tilde{W})$ that has zeroes on the $p - q$ dependent rows.
- Then modifying U to have zeros on the zeros of $\psi(\tilde{W})$ we can ignore the degenerate part of W.
- Finally, we are free to do what we want in $\tilde{W}$ to get a matrix of full rank.

1. Linearize the network (RKHS)
2. Ignore the degenerate part of $\psi(W)$ (technical)
3. Turn W into a full rank matrix (easy)
4. Reach a global minima

Only during the last step we decrease the loss, this is where we use the convexity in the first argument of l.

Underparametrisation

So far, if σ is a polynomial, or X has finitely many atoms, then

- $\dim^*(\sigma, n)$ or $\dim^*(\sigma, X)$ is less than ∞
- then $p \geq dim^* \implies$ no spurious valley

What if $p < \dim^*$?
Note that this is almost always the case:
σ = ReLU, sigmoid, softplus,...

Underparametrised networks can have arbitrarily bad spurious valleys

Theorem

For $n \geq 2$, the square loss and non-negative activation function σ.
If

$$p \leq \frac{1}{2} \dim^*(\sigma, n-1),$$

Then, $\forall M > 0$, there exists a non-empty open Ω and a random variable (X, Y) s.t. for any path $\theta : [0,1] \to \Theta$ such that $\theta(0) \in \Omega$ and $\theta(1)$ is a global minima satisfies

$$\max_t L(\theta_t) \geq \min_{\theta \in \Omega} L(\theta) + M.$$

With many parameters, spurious valleys are not so bad

Theorem

If the p initial units $\tilde{W}$ are initialized independently uniformly at random over the sphere $\mathbb{S}^n$. Let $f^*(X) = \mathbb{E}(Y|X)$ some measurable function that is minimal for the square loss, then there exists a descent path $t \mapsto \theta_t$ such that

$$L(\theta_1) \leq \mathbb{E}||f^*(X) - Y||_2^2 + \frac{1}{\lambda}$$

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

The floor of most valleys gets lower when parametrisation increases.

Proof

In the same spirit as for the overparametrised networks (turn the problem into one where optimization is easy to perform).
goal: Get filter vectors w_i not too far from some good vectors w_i^* (sample the w_i independently uniformly at random)

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

assuming $f^*(x) = \int \rho(w)\sigma(\langle w, x\rangle)d\tau(w)$.
There is a good approximation to f^* using the filters w_i. From last part of previous proof, using the last layer only we get a descent path to it from any initial parameter U.
Getting the right bound is tedious (see Bach 2017, quadrature rules). If ρ is assumed bounded, Hoeffding-type inequalities give exponential concentration.

A necessary and sufficient condition ?

Is $p \geq dim^*$ a necessary condition ?

- For $\sigma(z) = z$ (resp. z^2)

$$p \geq \mathcal{O}(\dim_*(\sigma, n)) \quad 1 \text{ (resp. n)}$$

is a sufficient parametrisation for the absence of spurious valleys.

Conjecture

If $p \geq \mathcal{O}(\dim_*(\sigma, n))$, then there is no spurious valley.

Idea to prove it: instead of getting $\psi(W)$ full rank to reach any global optima
choose a minima written as follows:

$$f^* = \sum_{i=1}^{\dim_*} u_i \psi_{\sigma, w_i^*} \quad \text{which is always possible}$$

and then generate the family ψ_{σ, w_i^*} from $\psi(W)$ with a constant output path.
This is not easy to do, getting a better use of symmetries of the form $\theta = (U, W) \mapsto (UG_1, G_2 W)$ for (G_1, G_2) in some group G that keep the output constant seems to be an important step...
and conjecture that one of the following is sufficient for the absence of spurious valley:

$$p \geq \mathcal{O}\left(\frac{\dim^*}{\dim(G)}\right) \quad \text{or } \mathcal{O}\left(\dim^* - \dim(G)\right)$$

...but nothing is clearly defined.

Conclusion

- If σ is a polynomial of degree $\geq \mathcal{O}(n^d)$, then there is no spurious valley.
- If the goal is empirical risk minimization and $p \geq N$, then there is no spurious valley.
- For general networks, $p \geq k \log(\frac{k}{\delta})$ is sufficient for having spurious valley with floor at most $\frac{1}{k}$ with probability at least $1 - \delta$

$\longrightarrow$ the largest the parametrisation, the less we have to worry about spurious valleys.