Loss landscape: SGD can have a better view than GD

Tomaso Poggio

CBMM, MIT

Yaim Cooper

Institute for Advanced Study, School of Mathematics

July 1, 2020

Abstract

Consider a loss function $L = \sum_{i=1}^{n} \ell_i^2$ with $\ell_i = f(x_i) - y_i$, where $f(x)$ is a deep feed-forward network with R layers, no bias terms and scalar output. Assume the network is overparametrized that is, $d >> n$, where d is the number of parameters and n is the number of data points. The networks are assumed to interpolate the training data (e.g. the minimum of L is zero). If GD converges, it will converge to a critical point of L, namely a solution of $\sum_{i=1}^{n} \ell_i \nabla \ell_i = 0$. There are two kinds of critical points - those for which each term of the above sum vanishes individually, and those for which the expression only vanishes when all the terms are summed. The main claim in this note is that while GD can converge to both types of critical points, SGD can only converge to the first kind, which include all global minima.

We review other properties of the loss landscape:

- As shown rigorously by [1] for the case of smooth RELUs the global minima in the Ws, when not empty, are highly degenerate with dimension $d - n$ and for them $\ell_i = 0 \quad \forall i = 1, \cdots, N$ (see also [2]).
- Under additional assumptions all of the global minima are connected within a unique and potentially very large global valley ([3], based on [4]).

1 Introduction

Understanding properties of the loss landscape can help us understand properties of trained networks. In the case of the square loss, for overparametrized networks, experiments suggest that SGD, unlike GD, may avoid critical points of the gradient which are not global minima.

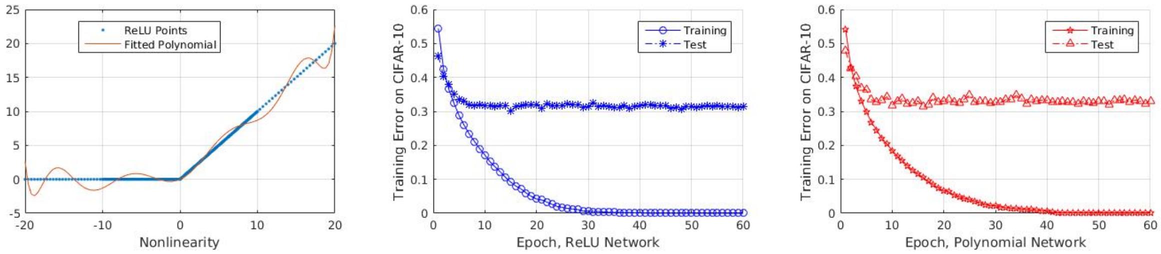

Figure 1: One can convert a deep network into a polynomial function by using polynomial nonlinearity. As long as the nonlinearity approximates ReLU well (especially near 0), the "polynomial net" performs similarly to a ReLU net. From [2].

Here we assume that the deep networks have an activation function which is a RELU. In some of the statements we assume either a smooth RELU or a polynomial approximation of a RELU as in Figure 1 .

2 Definitions

We consider the case in which f takes scalar values, implying that the last layer matrix W^R is has size 1 x h_{R-1}, where h_k denotes the size of layer k. The weights of hidden layer k has size $h_k \times h_{k-1}$. In the case of of binary classification the labels are $y \in \{-1, 1\}$.

For non-smooth ReLU activations the following important positive one-homogeneity property holds $\forall z, \forall a \geq 0, \sigma(\alpha z) = \alpha\sigma(z)$. Thus, $\sigma(z) = \frac{\partial \sigma(z)}{\partial z} z$. A consequence of one-homogeneity is a structural lemma (Lemma 2.1 of [5], closely related to Euler's theorem for homogeneous functions) $\sum_{i,j} W_k^{i,j} \left(\frac{\partial f(W;x)}{\partial W_k^{i,j}} \right) = f(W;x)$ where W_k is here the vectorized representation of the weight matrices W_k for layer k.

For the network, homogeneity of the ReLU implies $f(W;x) = \prod_{k=1}^{K} \rho_k f(V_1, \cdots, V_K; x)$, where $W_k = \rho_k V_k$ with the matrix norm $||V_k||_p = 1$. Note that $\frac{\partial f(W;x)}{\partial \rho_k} = \frac{\rho}{\rho_k} f(V;x)$ and that the definitions of ρ_k and V_k all depend on the choice of the norm used in normalization.

There are no bias terms but one of the input components is set to be a constant.

3 Background: gradient descent and SGD

Consider

$$\min_W L(f(W)) = \min_W \sum_{i=1}^{N} \ell_i^2 \tag{1}$$

with $\ell_i = y_i - f(W; x_i)$.

GD can be used to minimize $L(f(W))$ by running the following dynamical system (e.g. gradient flow)

$$\dot{W} = \nabla_W L(f(W)) = \sum_i^N \nabla_W f(W; x_i)(y_i - f(W; x_i)). \tag{2}$$

SGD can be formulated as follows. First define

Definition 1 *A random vector $v \in R^d$ drawn from a distribution $\mathcal{D}$ is a sampling vector if* $\mathbb{E}_{\mathcal{D}}[v_i] = 1 \quad \forall i$

Then the stochastic version of Equation 1 is

$$\min_W \mathbb{E}_{\mathcal{D}}[L(f(W))] = \min_W \mathbb{E}_{\mathcal{D}} \sum_i^n v_i \ell_i \tag{3}$$

Usually the distribution over $\mathcal{D}$ is assumed to be random v with independent components v_i, satisfying condition 1. This implies that in expectation SGD is equal to GD (???).

4 Critical points

Finding the interpolating global minimizers of $L = \sum \ell_i^2$ is equivalent to finding the set of network weights W^* that solve the system of equations $\ell_i(W^*) = 0 \quad \forall i = 1, \cdots, N$. Thus instead of finding all the critical points of the gradient of L, we would like to find the joint minimizers – that is the W – that minimize $\ell_i^2 \quad \forall i = 1, \cdots, n$.

We distinguish two sets of solution to $\nabla L = 0$:

1. solutions of $\ell_i \nabla \ell_i = 0, \forall i$

2. solutions of $\sum_i \ell_i \nabla \ell_i = 0$ that are not solutions of $\ell_i \nabla \ell_i = 0, \forall i$

The solutions 1) of $\ell_i \nabla \ell_i = (f(x_i) - y_i) \nabla_W f(x_i) = 0, \forall i$ consist of the global minima that is $f(x_i) - y_i = 0, \forall i$ and of other points, which we call here "spurious" critical points, for which $\ell_i \nabla \ell_i = 0, \forall i$ but $L \neq 0$. This can happen if $\nabla_W \ell_i = 0$ for some i and $\ell_i = 0$ for the other i. Notice that $\nabla_W f(x_i) = 0$ is a vector of D components – as many as there are weights in the network – all identically $= 0$ for a non-xero x_i.

5 Why SGD is less likely than GD to get stuck in critical points which are not global minimizers

The global minima and the critical points satisfying $\nabla_w \ell_i = 0, \quad \forall i$ are critical points of the loss for any subset of the training points, that is for any of the batches used in SGD. This is *not true* for the solutions 2) of $\sum_i \ell_i \nabla \ell_i = 0$: they are not (generically) critical points of any random subset of the training points. This means that SGD will never stop after it reaches them (generically).

Consider

$$\dot{W} = \sum_{i}^{N} \ell_i \nabla \ell_i \tag{4}$$

where $v_i(t)$ is the i-th component of v which is a random binary vector with $(n - n_{SGD})$ zero's and n_{SGD} one's, where n_{SGD} is the size of the minibatches used by SGD. For simplicity, assume the basic form of SGD, in which the minibatch size is 1. The random vector v (all zero with a single component equal to one) changes at every interation. Suppose now that $W(t)$ has reached a critical point of the gradient which is a global minimum. Then at $t+1$, $\dot{W} = 0$ independently of the choice of $v(t+1)$ since $\nabla f(x_i)$ is zero for each of the data points. Suppose instead that $W(t)$ has reached a critical point of the gradient which is not a global minimum (and is not a spurious critical point). Then at $t+1$, $\dot{W} \neq 0$ for a random choice of $v(t+1)$; if $\dot{W} = 0$ for that particular choice, in one of the subsequent iterations the random v will yield $\dot{W} \neq 0$ and the dynamical system will move out of the critical point.

Theorem 2 *(informal) Consider a noiseless situation.The dynamical systems defined by SGD and GD stops at global minima. GD will also get stuck at all other critical points of the gradient. SGD will not get stuck at other critical points – apart from the spurious ones, which are easy to detect, because all the components of gradient vector $\nabla f(x_i)$ are zero for at least some of data points i.*

6 Degeneracy of global minima

We are interested in finding the global minimizers achieving zero loss of

$$L(f(W)) = \sum_{i=1}^{n} \ell_i^2 \tag{5}$$

with $\ell_i = y_i - f(W; x_i)$. The network f is assumed to be overparametrized with a number of weights $d >> n$ and to be able to interpolate the training data achieving $L(f(W^*)) = 0$ which implies $\ell_i = 0 \quad \forall i = 1, \cdots, N$.

If we assume overparametrized networks with $d >> n$, where d is the number of parameters and n is the number of data points, [1] proved that the global minima of $L(w)$ are highly degenerate[1] with dimension $d - n$.

[1] This result is also what one expects from Bezout theorem for a deep polynomial network. As Terry Tao says in his blog “from the general “soft” theory of algebraic geometry, we know that the algebraic set V is a union of finitely many algebraic varieties, each of dimension at least d-n, with none of these components contained in any other. In particular, in the underdetermined case n<d , there are no zero-dimensional components of V , and thus V is either empty or infinite”.

Theorem 3 *([1]) For an overparametrized f with smooth activation function assuming a square loss, the minimizers $W*$ are highly degenerate with dimension $d - n$.*

6.1 Degeneracy of global critical points

We wish to understand whether the solutions of the global critical points, e.g. $\sum_i \ell_i \nabla \ell_i = 0$ that are neither global zeros nor solutions of $\nabla \ell_i = 0, \quad \forall i$ are degenerate. We hypothesize that the space of such critical points is lower dimensional than the space of global minima, and leave the study of this question to future work.

6.2 Non-smooth RELU

For exact RELU – as opposed to smooth RELU assumed so far – it turns out that $\nabla_W f(x_i) = 0$ *for some i then $f(x_i) = 0$ for the same i.*

The proof of this fact uses the structural lemma (Lemma 2.1 of [5], closely related to Euler's theorem for homogeneous functions)

$$\sum_{i,j} W_k^{i,j} \left(\frac{\partial f(W;x)}{\partial W_k^{i,j}} \right) = f(W;x) \tag{6}$$

where W_k is here the vectorized representation of the weight matrices W_k for layer k. Setting $\nabla_W f(x_i) = 0$ in Equation 6 gives

$$0 = f(W;x_i) \tag{7}$$

.

Taking second derivatives of Equation 6 shows that if $\nabla_W f(x_i) = 0$ then the Hessian $H(x_i) = 0$.

This property allows a better characterization of the spurious critical points. The spurious points are defined by $f(x_i) = y_i$ for some i and $f(x_i) = 0$ for the other i; this means that $L \neq 0$. Are the spurious points local minima or saddles? The structural lemma implies that for the i for which $f(x_i) = 0$ it holds $H(x_i) = 0$; for the other i H is positive semidefinite (because they are global minima of ℓ_i). Thus the Hessian associated with L, since L is the sum of the ℓ_i, is positive semidefinite. This means that the associated critical points are *degenerate local minima*.

Of course, the use of this property is tantamount to assuming non-smooth activation functions. This assumption is inconsistent with the smoothness we require for proving theorems on the degeneracy of the global and local minima. This is likely to be a technical problem that may be circumvented but at the present time represents a possibly fatal weakness in some of our arguments.

7 Discussion

Taken together, the results we summarize here may help explain some puzzling empirical observations of the past few years. In particular, minima found by GD can be unstable for SGD. According to [6], switching from GD to SGD at a point close to a global minimum, SGD often escapes from that minimum and converges to a better minimum [7]

7.1 GD and SGD "feel" different loss landscapes

Consider loss functions of the form $L = \sum_i \ell_i$. In the preceeding section we have shown that SGD cannot stop at critical points that exist because of the interaction between different ℓ_i whereas GD will stop at them. To visualize this point here is a simple example. Define $\ell(w, x_i)$ as $a(w - x_i)^2$. With $x_1 = -x_2$ consider $\sum_{i=1}^{2} \ell(w, x_i) = a(w - x_1)^2 + a(w + x_1)^2$. GD will "see" a loss of the form $2aw^2 + const$ with one minimum in zero, whereas SGD will see a loss with two minima, one in x_1 and one in x_2.

This shows that the analogy between SGD on one hand and GD + noise at the other hand,, can be quite misleading.

Acknowledgments We thank Andrzej Banburski for the observation on exact RELU. This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216, and part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), and AFSOR-THRL (FA8650-05-C-7262).

References

[1] Yaim Cooper. The loss landscape of overparameterized neural networks. *CoRR*, abs/1804.10200, 2018.

[2] T. Poggio and Q. Liao. Theory II: Landscape of the empirical risk in deep learning. *arXiv:1703.09833, CBMM Memo No. 066*, 2017.

[3] Quynh Nguyen. On connected sublevel sets in deep learning. *CoRR*, abs/1901.07417, 2019.

[4] J. C. Evard and F. Jafari. The set of all rectangular real matrices is connected by analytic regular arcs. *Proceedings of the American Mathematical Society*, 120(2):413–419, February 1994.

[5] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. *CoRR*, abs/1711.01530, 2017.

[6] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. *arXiv e-prints*, page arXiv:1803.00195, February 2018.

[7] Lei Wu, Zhanxing Zhu, and Weinan E. Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes. *arXiv e-prints*, page arXiv:1706.10239, June 2017.