Critical Point Behavior of Deep Networks

Junzi Zhang, ICME

November 8, 2016

Today, we examine the behavior of critical points in general deep learning problems. We will begin with linear cases, and then proceed on to nonlinear networks.

1 Linear Deep Networks

1.1 Main Results

Consider a feedforward linear network with H hidden layers with sizes $d_1, \ldots, d_H$ respectively. The input/output sizes are $d_x = d_0/d_y = d_{H+1}$, and there are m samples. Our goal is to minimize the squared error loss function

$$\bar{\mathcal{L}}(W) = \frac{1}{2}\|W_{H+1}W_H \ldots W_2W_1X - Y\|_F^2 \tag{1}$$

where $W_k \in \mathbb{R}^{d_k \times d_{k-1}}$ are the weight matrices, while $X \in \mathbb{R}^{d_x \times m}$ and $Y \in \mathbb{R}^{d_y \times m}$ are the input and output data respectively.

Previous results state that $\bar{\mathcal{L}}$ is convex in each W_k when all others are fixed, which is obvious. Moreover, Baldi & Hornik [2] proved in 1988 that if $H = 1$ (a "shallow" network), $p < d_y = d_x$, and XX^T, XY^T are invertible, then every local minimum is a global minimum. Here $p = \min\{d_1, \ldots, d_H\}$ is the smallest width of a hidden layer. The proof is mainly done by using first order conditions and eigenvalue perturbation arguments.

For general H, the result is generalized by Kenji Kawaguchi [1] in NIPs 2016 as follows:

Theorem 1. *Suppose $d_y \leqslant d_x$, and assume that XX^T is invertible and XY^T is full rank. Then*

1. *It is non-convex and non-concave;*
2. *Every local minimum is a global minimum;*

3. Every critical point that is not a global minimum is a saddle point;

4. If rank($W_H \dots W_2$)= p, then Hessian at any saddle point has at least one negative eigenvalue.

Considering the noise in data, the full rank assumption is realistic. Notice that by 2, 3 is equivalent to saying that there is no local (or global) maximum. And as a corollary of 4, if $H = 1$, then $W_H \dots W_2 = I_{d_1} = I_p$, and hence the Hessian at any saddle point has at least one negative eigenvalue. Notice that here and below we adopt the notation that $W_k \dots W_{k'} = I_{d_k}$ if $k < k'$.

1.2 Proof Techniques

The proof is heavily relied on linear algebra stuffs.

Firstly, we introduce the following notations. Let $A \otimes B$ be the Kronecker product of matrices A and B. Let $vec(A)$ be the vectorization of A, and let $\mathcal{D}_{vec(W_k^T)} f(\cdot)$ be the partial derivative of f w.r.t. $vec(W_k^T)$ in the numerator layout (i.e. the gradient of a scalar function will be a row vector). Let $\mathcal{R}(A)$ be the column space of A, and let A^- be (any) generalized inverse. In deriving necessary conditions, we sometimes construct a specific generalized inverse, or require it to be a Moore-Penrose pseudo-inverse. Let $r = (W_{H+1} \dots W_1 X - Y)^T$, and let $C = W_{H+1} \dots W_2$.

We also define $\Sigma = YX^T(XX^T)^{-1}XY^T$ with eigenvalue decomposition $\Sigma = U\Lambda U^T$, where $\Lambda_{1,1} \geqslant \cdots \geqslant \Lambda_{d_y,d_y}$, and $U = [u_1, \dots, u_{d_y}]$. Moreover, let $\bar{p}$ =rank$(C)\leqslant \min\{d_y, p\}$, and for $\mathcal{I}_{\bar{p}} = \{i_1, \dots, i_{\bar{p}} | 1 \leqslant i_1 \leqslant \cdots \leqslant i_{\bar{p}} \leqslant \min\{d_y, p\}\}$, we define $U_{\mathcal{I}_{\bar{p}}} = [u_{i_1}, \dots, u_{i_{\bar{p}}}]$. In particular, we also define $U_{\bar{p}} = [u_1, \dots, u_{\bar{p}}]$.

The following properties are the major instruments in proving theorem 1.

1. $vec(ABC) = (C^T \otimes A)vec(B)$
2. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$, $(A \otimes B)^T = A^T \otimes B^T$, $A \otimes B = 0 \Rightarrow A = 0$ or $B = 0$
3. $(A \otimes B)^- = A^- \otimes B^-$, $A(A^TA)^-A^TA = A$, $(A^TA)^-A^TAA^T = A^T$
4. $Ax = b \iff \exists w, x = A^-b + (I - A^-A)w$, and solution exists iff $AA^-b = b$.
5. $\mathcal{R}(A) \subseteq \mathcal{R}(B) \iff BB^-A = A \iff \exists \mathcal{I}_{\text{rank}(A)}$ and invertible G, $A = [B_{\mathcal{I}_{\text{rank}(A)}}, 0]G$

With these properties, the following lemmas can be proved.

Lemma 1. *(First Order) W is a critical point of $\bar{\mathcal{L}}(W)$ iff $\forall k = 1, \dots, H+1$,*

$$\left(\mathcal{D}_{vec(W_k^T)} \bar{\mathcal{L}}(W)\right)^T = \left((W_{H+1} \dots W_{k+1}) \otimes (W_{k-1} \dots W_1 X)^T\right)^T vec(r) = 0 \tag{2}$$

which implies that $W_{H+1} \dots W_1 = C(C^TC)^-C^TYX^T(XX^T)^{-1}$.

Lemma 2. *(Second Order) If $\nabla^2\bar{\mathcal{L}}(W)$ is positive/negative semidefinite at a critical point, then for any $k \in \{2, \dots, H+1\}$, $\mathcal{R}\left((W_{k-1}\dots W_2)^T\right) \subseteq \mathcal{R}(C^TC)$ or $XrW_{H+1}\dots W_{k+1} = 0$. This implies that either rank$(W_{H+1}\dots W_k) \geqslant$ rank$(W_{k-1}\dots W_2)$ or $XrW_{H+1}\dots W_{k+1} = 0$.*

Furthermore, if $\nabla^2\bar{\mathcal{L}}(W)$ is positive semidefinite at a critical point, then

$$C(C^TC)^-C^T = U_{\bar{p}}U_{\bar{p}}^T \text{ or } Xr = 0 \tag{3}$$

In addition to these main lemmas and proof techniques, Kenji Kawaguchi [1] also provides a explicit calculation of the Hessian $\nabla^2\bar{\mathcal{L}}(W)$ in block forms, i.e. $\mathcal{D}_{vec(W_k^T)}\left(\mathcal{D}_{vec(W_j^T)}\bar{\mathcal{L}}(W)\right)^T$.

1.3 Proof of Theorem 1

The proof of claim 1 of theorem 1 is then done by using the rank inequality in lemma 2 with $k = H+1$ and considering $W_{H+1} = W_1 = 0$ and $W_2, \dots, W_H \neq 0$.

The proof of claim 2 of theorem 1 is done by a detailed case analysis. The first case is when rank$(W_H\dots W_2) = p$ (which is then subdivided into two cases $d_y \leqslant p$ and $d_y > p$), in which we show that as long as the Hessian is positive semidefinite at a critical point W, then it's a global minimum. When $d_y \leqslant p$, this is done by proving that $Xr = 0$, while when $d_y > p$, this is done by proving that $W_{H+1}\dots W_1 = U_{\bar{p}}U_{\bar{p}}^TYX^T(XX^T)^{-1}$. As a byproduct, claim 4 of theorem 1 is also proved.

The second case is when rank$(W_H\dots W_2) < p$, which is more troublesome. In this case, we first prove that if rank$(C) \geqslant \min\{p, d_y\}$, then we are fine. When rank$(C) < \min\{p, d_y\}$, we prove by induction that we can have rank$(W_k\dots W_1) \geqslant \min\{p, d_y\}$ for $k = 1, \dots, H+1$ by **arbitrarily small perturbation** of $W_1, \dots, W_k$ without changing the value of $\bar{\mathcal{L}}(W)$. Here in the k-th induction step, only W_k is modified. Moreover, *it is in this case where the difference between a local minimum and a saddle point with semidefinite Hessian is explicit – after the perturbation of W, only the local minimum still maintains a positive semidefinite Hessian since its value is still a local minimum (in some small neighbor of the original point).* This also indicates an example of "bad" saddle points when $H > 1$ by considering $W_1 = \dots = W_H = 0$ and arbitrary W_{H+1}.

The proof of claim 3 of theorem 1 is done by first showing that if $W_{H+1}\dots W_2 \neq 0$, then the loss function has a strictly increasing direction w.r.t W_1, and then in the case where $W_{H+1}\dots W_2 = 0$ but the Hessian is negative semidefinite, we prove by induction as in the proof of claim 2 that we

can have $W_k \dots W_2 \neq 0$ for $k = 2, \dots, H+1$ by **arbitrarily small perturbation** of $W_1, \dots, W_k$ without changing the value of $\bar{\mathcal{L}}(W)$.

2 Nonlinear Deep Networks

In this section, we move on to nonlinear deep networks. Here we consider the ReLU activator $\sigma(x) = \max\{0, x\}$, and the output of the neural network becomes

$$\hat{Y}(W, X) = q\sigma_{H+1}(W_{H+1}\sigma_H(W_H \dots \sigma_2(W_2\sigma_1(W_1 X)) \dots)) \tag{4}$$

where q is some constant while the other notations remain the same as in section 1.

The key observation here is to notice that the (j, i)-th entry of $\hat{Y}$ can be written as

$$\hat{Y}(W, X)_{j,i} = q \sum_{p=1}^{\Psi} [X_i]_{j,p} [Z_i]_{j,p} \prod_{k=1}^{H+1} w_{j,p}^{(k)} \tag{5}$$

Here Ψ is the total number of paths from the inputs to the j-th output in the neural network, $[X_i]_{j,p}$ is the entry of the i-th sample input datum that is used in the p-th path of the j-th output, $w_{j,p}^{(k)}$ is the entry of W_k that is used in the p-th path of the j-th output, and $[Z_i]_{j,p} \in \{0, 1\}$ is the indicator of whether the p-th path is active or not for each sample i. This can be easily understood in analogy with Markov Chain transition probabilities, just with an additional activation σ along the paths.

Previously, Choromanska *et al* [3] proved in 2015 that under the realistic assumptions **A1p** (activations have the same probability of success), **A2p** (X_i's are Gaussian), **A3p** (no redundancy in parameters), **A4p** (uniformity of non-redundant parameters), **A7p** (spherical constraints), and the unrealistic assumptions **A5u** (activation independent of input data) and **A6u** (all paths have independent inputs), both the expected hinge loss and the expected absolute loss can be rewritten as (after regrouping X and Y together and re-indexing)

$$\mathcal{L}_{previous}(W) = C_1 + C_2 \frac{\mathbf{1}}{\boldsymbol{\Lambda^{H/2}}} \sum_{\boldsymbol{i_1, \dots, i_{H+1}=1}}^{\boldsymbol{\Lambda}} \boldsymbol{X_{i_1, \dots, i_{H+1}}} \prod_{\boldsymbol{k=1}}^{\boldsymbol{H+1}} \boldsymbol{w_{i_k}} \text{ subject to } \frac{1}{\Lambda} \sum_{i=1}^{\Lambda} w_i^2 = 1 \tag{6}$$

where $\Lambda = \Psi^{1/(H+1)}$ is the number of non-redundant parameters, and C_1, C_2 are constants with $C_2 > 0$. Notice that here the output data / true labels Y's are also assumed to be randomly taking values in $\{S, -S\}$ for some constant S in the above assumptions, but its absorbed into X due to the Gaussian assumption **A2p** . Now the bold term is exactly the Hamiltonian of the spherical

spin-glass model, whose asymptotic behavior as $\Lambda \to \infty$ is well studied by Auffinger *et al* [5] in 2010. Correspondingly, Choromanska *et al* [4] arrives at the following conclusions for large networks when $m = d_y = 1$ (i.e. $\Lambda \to \infty$):

Theorem 2. *We have the following observations (high level):*

1. *Critical points form an ordered structure such that there exists an energy/loss barrier $E_{-\infty}$ below which with overwhelming probability one can find only low-index critical points, most of which are concentrated close to the barrier;*
2. *Recovering the ground state, i.e. global minimum, takes exponentially long time;*
3. *With overwhelming probability one can find only high-index saddle points above energy $E_{-\infty}$ and there are exponentially many of those;*
4. *Low-index critical points are "geometrically" lying closer to the ground state than high-index critical points.*

Kenji Kawaguchi [1] proved the same result as in theorem 1 as a corollary of theorem 1 for linear deep networks, under weaker assumptions **A1p-m** and **A5u-m**. Here, assumption **A1p-m** assumes that $[Z_i]_{j,p}$'s are Bernoulli random variables with the same probability of success ρ, and assumption **A5u-m** assumes that $[Z_i]_{j,p}$'s are independent from the input X's and parameters w's. Under these assumptions, assuming no randomness in X, Y and W, if we consider the expected squared error loss $\mathcal{L}(W) = \frac{1}{2}\|\mathbb{E}_Z[\hat{Y}(W, X) - Y]\|_F^2$, and assuming $q = \rho^{-1}$ w.l.o.g., then we will arrive at

$$\mathbb{E}_Z[\hat{Y}(W, X)] = q \sum_{p=1}^{\Psi} [X_i]_{j,p} \rho \prod_{k=1}^{H+1} w_{j,p}^{(k)} = \sum_{p=1}^{\Psi} [X_i]_{j,p} \prod_{k=1}^{H+1} w_{j,p}^{(k)} = W_{H+1} \dots W_1 X \tag{7}$$

and thusly $\mathcal{L}(W) = \frac{1}{2}\|W_{H+1} \dots W_1 X - Y\|_F^2 = \bar{\mathcal{L}}(W)$, and hence $\mathcal{L}(W)$ behaves exactly the same as $\bar{\mathcal{L}}(W)$.

Notice that by this result, claim 4 of theorem 2 becomes trivial, and claim 2 of theorem 2 is also partially trivial (in the sense that there is no need to go from local minimum to saddle points lying above some loss threshold and then to another lower local minimum). But most part of theorem 2, which addresses the layout structure of critical points, remain unproved, despite the fact that Kenji has proved something new about saddle points too. Also the assumption **A5u-m** is still unrealistic. Deep nonlinear networks remain an almost unexploited area.

3 Two Additional Questions

I have two additional questions (may be trivial) for thoughts.

- Is it true that $\{W|W = W_{H+1} \dots W_1\} = \{W|W \in \mathbb{R}^{d_y \times d_x}, \text{rank}(W) \leqslant p\}$?
- Is it ensured that $\bar{\mathcal{L}}(W)$ attains its global minimum at some finite point W?

References

[1] Kawaguchi, Kenji. "Deep Learning without Poor Local Minima." arXiv preprint arXiv:1605.07110 (2016).

[2] Baldi, Pierre, and Kurt Hornik. "Neural networks and principal component analysis: Learning from examples without local minima." Neural networks 2.1 (1989): 53-58.

[3] Choromanska, Anna, et al. "The Loss Surfaces of Multilayer Networks." AISTATS. 2015.

[4] Choromanska, Anna, Yann LeCun, and Grard Ben Arous. "Open Problem: The landscape of the loss surfaces of multilayer networks." Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3. Vol. 6. No. 2015. 2015.

[5] Auffinger, Antonio, Grard Ben Arous, and Ji? ?ern. "Random matrices and complexity of spin glasses." Communications on Pure and Applied Mathematics 66.2 (2013): 165-201.