Discussion on Several Basic Problems within Non-convex Optimization

YWJ
HUST-IDC / No Address
ywj@hust.edu.cn

Abstract

1 Introduction

This article begin with proof of two statements below:

1. All neural networks with more than one hidden layer can only fitting non-convex functions.
2. For any linear neural network with square loss function, every local minimum is a global minimum [Kawaguchi(2016)].

2 All Neural Networks with more than One Hidden Layer can Only Fitting Non-convex Functions

Firstly, take a special two layers Neural Network without activate function into consideration. It's defined by $f(w_1, w_2, w_y) = x \cdot w_1 \cdot w_2 \cdot w_y$. Assume that we have ideally perfect weights (w_a, w_b, w_y) to fit our object $LOSS(f(w_1, w_2, w_y))$. if $LOSS(f(w_1, w_2, w_y))$ is convex, for any other weights $(w_1, w_2, w_y) \neq (w_a, w_b, w_y)$ have $LOSS(f(w_1, w_2, w_y)) > LOSS(f(w_a, w_b, w_y))$. Because convex functions have a unique optimal point. By opposite we give this:

$$f(w_1, w_2, w_y) = x \cdot w_a \cdot w_b \cdot w_y = x \cdot w_a \cdot I \cdot w_b \cdot w_y = x \cdot w_a \cdot A \cdot A^{-1} \cdot w_b \cdot w_y = x \cdot w_1 \cdot w_2 \cdot w_y$$

In fact, when (A, A^{-1}) is a pair of opposite elementary transformation, we can always find another set of weights to minimize $LOSS(f(w_1, w_2, w_y))$. To include the circumstance that (w_1, w_2, w_y) and (w_a, w_b, w_y) are exactly the same point, we give these two lemmas:

Lemma 1 *For every point meet $w_a = w_1$ or $w_b = w_2$, it's not a unique optimal point.*

Lemma 2 *If for arbitrary full rank matrix B has:$A \cdot B = A$, then A is a zeros matrix.*

Lemma 2 point out that when there isn't a full rank A let $w_a \neq w_a \cdot A$, w_a must be zeros matrix. Which means the following two conditions must be satisfied at most one: 1) There is a full rank A make $w_a \neq w_a \cdot A$, then (w_a, w_b, w_y) is not a unique one and the function is non-convex. 2) w_a or w_b is zeros matrix. If one of (w_a, w_b) is zeros, whatever the rest parameters (w_y) are, the network output a zeros vector. If (w_1, w_2, w_y) is optimal, then for arbitrary w_y, they are equally optimal. Which means (w_1, w_2, w_y) is not a unique one and the function is non-convex. The rest is proof of **Lemma 2**:

Proof 2.1 ***Condition 1****: A is a square matrix, note $shape(A) = (n, n)$:*

When $n = 1$, A and B are both constant, since B is full rank, $B \neq 0$, then $A = 0$ must be satisfied.

Assume that when $n = k$, ***Lemma 2*** *is true. When $n = k + 1$, $A \cdot B = A$ could be written in partitioned matrix format:*

$$\begin{bmatrix} A^{kk} & A^{k1} \\ A^{1k} & A^{11} \end{bmatrix} \cdot \begin{bmatrix} B^{kk} & B^{k1} \\ B^{1k} & B^{11} \end{bmatrix} = \begin{bmatrix} A^{kk}B^{kk} + A^{k1}B^{1k} & A^{kk}B^{k1} + A^{k1}B^{11} \\ A^{1k}B^{kk} + A^{11}B^{1k} & A^{1k}B^{k1} + A^{11}B^{11} \end{bmatrix} = \begin{bmatrix} A^{kk} & A^{k1} \\ A^{1k} & A^{11} \end{bmatrix}$$

The existence of solution $A = 0$ is obviously. Then a special case in set of B could be used to demonstrate $A = 0$ is the unique solution. Set $B^{k1} = 0$ and $B^{1k} = 0$, because B is full rank, must have $B^{11} \neq 0$ and B^{kk} is an arbitrary full rank one. The equation above could be written into:

$$\begin{bmatrix} A^{kk}B^{kk} & A^{k1}B^{11} \\ A^{1k}B^{kk} & A^{11}B^{11} \end{bmatrix} = \begin{bmatrix} A^{kk} & A^{k1} \\ A^{1k} & A^{11} \end{bmatrix}$$

According to the assumption about $n = k$, $A^{kk} = 0$. Because B^{11} is a arbitrary constant and $B^{11} \neq 0$, must have $A^{k1} = 0$ and $A^{11} = 0$. As A^{1k}, padding with zeros and making it a square matrix with shape (n, n). We have the equivalent equation below. According to the assumption about $n = k$, $A^{1k} = 0$.

$$A^{1k} \cdot B^{kk} = A^{1k} \Leftrightarrow \begin{bmatrix} A^{1k} \\ O \end{bmatrix}^{kk} \cdot B^{kk} = \begin{bmatrix} A^{1k} \\ O \end{bmatrix}^{kk}$$

The special case of B demonstrate that any other A except $A = 0$ can't fit this case. That means $A = 0$ is the only solution of precondition of ***Lemma 2****.*

Condition 2*: A is not a square.*

It's easy to prove ***Lemma 2*** *when padding A into a square one with zeros like dealing with A^{1k}. The detail is omitted.*

The proof above works even when w_1 and w_2 are not square matrices. And because of two equally optimal point could be always found in the same function, I proved the function must be a non-convex one. Then consider the general networks with an activate function $\alpha()$, for any kind of activate functions we have:

$$\alpha(x \cdot w_1) \cdot A = \alpha(x \cdot w_1 \cdot A)$$

Noted that A denote a elementary transformation. Only when w_1 is squared can A be written as a elementary matrix. Then in a similar way:

$$f(x) = \alpha(x \cdot w_a) \cdot w_b \cdot w_y = \alpha(x \cdot w_a) \cdot I \cdot w_b \cdot w_y = \alpha(x \cdot w_a) \cdot A \cdot A^{-1} \cdot w_b \cdot w_y$$
$$= \alpha(x \cdot w_a \cdot A) \cdot A^{-1} \cdot w_b \cdot w_y = \alpha(x \cdot w_1) \cdot w_2 \cdot w_y$$

End here, proposition in the title has been proved.

3 For Any Linear Neural Network with Square Loss Function, Every Local Minimum is a global minimum.

Before the main proof, a series of lemma are given below:

Lemma 3 *Denote A_i as column of matrix A, then:*

$$\sum_i ||A_i||_2^2 = ||A||_F^2$$

Lemma 4 *$vec()$ denote the function unfold a matrix into a vector according to the column order.*

$$vec(AXB) = (B^T \otimes A) \cdot vec(X)$$

Before give the rest lemmas, some notations are given here. A linear neural network with square loss function is defined by $L(W) = \sum_i ||\overline{Y}_i(W,X) - Y_i||_2^2$, the forward function is $\overline{Y}(W,X) = W_y W_n \cdots W_2 W_1 X$. d_k denotes the dimension of kth layer, thus W_k has a shape (d_{k-1}, d_k). Let $p = \min(d_1, d_2 \cdots d_n)$.

For simplicity, denote the error transposed matrix by r and a constant $\frac{1}{2}$ is added to the loss. By using **Lemma 3**, the loss could be written as matrix product:

$$L(W) = \frac{1}{2}\sum_i ||\overline{Y}_i(W,X) - Y_i||_2^2 = \frac{1}{2}||\overline{Y}(W,X) - Y||_F^2 = \frac{1}{2}tr(r \cdot r^T) = \frac{1}{2}sum(r^T \circ r^T) = \frac{1}{2}vec(r)^T vec(r)$$

If W is a critical point, then for $k \in [1, 2, \cdots, n]$, the derivatives of $L(W)$ with respect of W_k equal to 0:

$$\frac{\partial L(W)}{\partial vec(W_k^T)} = \frac{1}{2}\frac{\partial vec(r)^T vec(r)}{\partial vec(r)^T} \cdot \frac{\partial vec(r)^T}{\partial vec(W_k^T)}$$

$$= vec(r) \cdot \frac{\partial vec(W_y W_n \cdots W_2 W_1 X - Y)^T}{vec(W_k^T)}$$

$$= vec(r)^T \cdot \frac{\partial vec(X^T W_1^T \cdots W_n^T W_y^T)}{vec(W_k^T)}$$

$$= vec(r)^T \cdot \frac{\partial (W_y W_n \cdots W_{k+1} \otimes X^T W_1^T \cdots W_{k-1}^T) \cdot vec(W_k^T)}{vec(W_k^T)}$$

$$= vec(r)^T \cdot (W_y W_n \cdots W_{k+1} \otimes X^T W_1^T \cdots W_{k-1}^T) = O$$

Since the equation is true statement for every k in $[1, 2, \cdots, n]$, choose $k = 1$ then have:

$$O^T = (W_y W_n \cdots W_2 \otimes X^T)^T vec(r) = vec(X r W_y W_n \cdots W_2)$$

Denote $W_y W_n \cdots W_2$ by C, then $r = C W_1 X$:

$$vec(X r W_y W_n \cdots W_2) = vec(X X^T W_1^T C^T C - X Y^T C) = O$$

$$C^T C W_1 X X^T - C^T Y X^T = O$$

With decorrelated X, XX^T is invertible:

$$C^T C W_1 = C^T Y X^T (X X^T)^{-1}$$

For any generalized inverse g the linear equation has a solution:

$$W_1 = (C^T C)^g C^T Y X^T (X X^T)^{-1} + (I - (C^T C)^g C^T C) M$$

M is a matrix determined by the certain generalized inverse.According to the theory of Schur Complement [Zhang(2005)], for any generalized inverse satisfied $A(A^T A)^g (A^T A) = A$, thus:

$$C W_1 = C(C^T C)^g C^T Y X^T (X X^T)^{-1} + (CI - C(C^T C)^g C^T C) M = C(C^T C)^g C^T Y X^T (X X^T)^{-1}$$

For later using, denote the conclusion above as **Lemma 5**:

Lemma 5 *For any critical point W in $L(W)$, has:*

$$W_y W_n \cdots W_2 W_1 = C(C^T C)^g C^T Y X^T (X X^T)^{-1}$$

Lemma 6 *For any critical point W in $L(W)$, it's hessian matrix could be written in format:*

$$\nabla^2 L(W) = \begin{bmatrix} \frac{\partial^2 L(W)}{\partial vec(W_y^T)\partial vec(W_y^T)} & \frac{\partial^2 L(W)}{\partial vec(W_y^T)\partial vec(W_n^T)} & \cdots & \frac{\partial^2 L(W)}{\partial vec(W_y^T)\partial vec(W_1^T)} \\ \frac{\partial^2 L(W)}{\partial vec(W_n^T)\partial vec(W_y^T)} & \frac{\partial^2 L(W)}{\partial vec(W_n^T)\partial vec(W_n^T)} & \cdots & \frac{\partial^2 L(W)}{\partial vec(W_n^T)\partial vec(W_1^T)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 L(W)}{\partial vec(W_1^T)\partial vec(W_y^T)} & \frac{\partial^2 L(W)}{\partial vec(W_1^T)\partial vec(W_n^T)} & \cdots & \frac{\partial^2 L(W)}{\partial vec(W_1^T)\partial vec(W_1^T)} \end{bmatrix}$$

Lemma 7 *Denote the range of matrix M as $\Re(M)$. If $\nabla^2 L(W)$ is is positive semidefinite or negative semidefinite at a critical point, then for any $k \in [2, 3, \cdots, n, y]$:*

$$\Re(W_{k-1} \cdots W_2) \subseteq \Re(C^T C)$$

or:

$$XrW_y W_n \cdots W_{k+1} = O$$

Lemma 8 *If $\nabla^2 L(W)$ is is positive semidefinite or negative semidefinite at a critical point, then for any $k \in [2, 3, \cdots, n, y]$:*

$$rank(W_y W_n \cdots W_k) \geq rank(W_{k-1} \cdots W_2)$$

or:

$$XrW_y W_n \cdots W_{k+1} = O$$

A little more notation are make here before the last lemma. Let $\Sigma = YX^T(XX^T)^{-1}XY^T$. Denote the eigendecomposition of Σ as $\Sigma = U\Lambda U^T$, where $\Lambda_{1,1} > \Lambda_{2,2} > \cdots \Lambda_{d_y,,d_y}$ are the eigenvalues in descending order, with eigenvectors $U = [u_1, u_2, \cdots, u_{d_y}]$ note that Σ is symmetrical. Let $\overline{p} = rank(C)$, remember that p has been defined as $p = \min(d_1, d_2, \cdots, d_n)$, $C = W_y W_n \cdots W_2$, W_2 has shape (d_2, d_1), then $rank(C)$ has range: $\overline{p} \in [1, 2, \cdots \min(d_1, d_2, \cdots d_y)] = [1, 2, \cdots \min(p, d_y)]$. Define the $\overline{p}$ largest eigenvectors as $U_{\overline{p}} = [u_1, u_2, \cdots, u_{\overline{p}}]$.

Lemma 9 *If $\nabla^2 L(W)$ is positive semidefinite at a critical point, ,then:*

$$C(C^T C)^g C^T = U_{\overline{p}} U_{\overline{p}}^T$$

or:

$$Xr = O$$

Finally comes to our main proposition.

Theorem 1 *Assume that XX^T and XY^T are of full rank with $d_y \leq d_x$ and Σ has d_y distinct eigenvalues. Then, for any depth $n \geq 1$ and for any layer widths and any input-output dimensions $d_y, d_n, ..., d_1, dx \geq 1$, the loss function $L(W)$ has the following properties: Every local minimum is a global minimum.*

Proof 3.1 ***Condition 1**: $rank(W_n ... W_2) = p$ and $d_y \leq p$.*

*In a critical point, **Lemma 8** with $k = y$ implies:*

$$rank(W_y) \geq rank(W_n \cdots W_2) \qquad \text{or} \qquad Xr = O$$

If $d_y < p$, then $rank(W_y) < p = min(d_n, ..., d_1) = rank(W_n \cdots W_2)$, $Xr = O$ would be the necessary condition.

*Else if $d_y = p$, from **Lemma 7** with $k = y$ we know $\Re(W_n \cdots W_2) \subseteq \Re(C^T C)$, then we have*

$$p = \min(d_n, \cdots, d_1) = rank(W_n \cdots W_2) \leq \Re(C^T C) = rank(C)$$

*From **Lemma 8** with $k = 2$ we have:*

$$rank(W_n \cdots W_2) = rank(C) > rank(I_{d_1}) \qquad or \qquad XrW_yW_n \cdots W_3 = O$$

*Note that when $k = 2$ $rank(W_{-1}, ...W_2)$ is invalid. This could be solved by adding unit matrix to both side of the inequation. Denoted unit matrix with dimension of d_1 as I_{d_1}, then the function becomes $\overline{Y}(W, X) = W_yW_n \cdots W_2I_{d_1}W_1$, apply to **Lemma 8** get the result $rank(W_n \cdots W_2) > rank(I_{d_1})$. Consider the latter, since:*

$$rank(W_yW_n \cdots W_3) \geq rank(W_yW_n \cdots W_2) = rank(C) \geq p \qquad and \qquad p = \min(d_n, \cdots, d_1) = d_y$$

$W_yW_n \cdots W_3$ has shape (d_y, d_2) and $d_y \leq d_2$. If $rank(W_yW_n \cdots W_3) \geq d_y$, then it's full rank. So we have $Xr = O$. Consider the former, C has shape (d_n, d_1) and $rank(C) \geq d_1$, so have C^TC is full rank. And the generalized inverse for a invertible matrix is $g = -1$, we have:

$$Xr = X(CW_1X - Y) = XC(C^TC)^{-1}C^TYX^T(XX^T)^{-1}X - XY = XY - XY = O$$

*In total we have $Xr = O$ in **Condition 1**, as X is full rank, always have $r = O$. And the error matrix of square loss is zeros matrix guarantees W is a global minimum.*

***Condition 2**: $rank(W_n...W_2) = p$ and $d_y > p$.*

*From **Lemma 9** we have $C(C^TC)^gC^T = I$, then:*

$$W = W_yW_n \cdots W_1 = C(C^TC)^{-1}C^TYX^T(XX^T)^{-1} = U_{\overline{p}}U_{\overline{p}}^TYX^T(XX^T)^{-1}$$

Since $\overline{p} = rank(C) = min(d_y, d_n, ...d_1)$ and $d_y > p$, $\overline{p} = rank(C) = min(d_n, ...d_1) = p$. the equation can be rewrite as:

$$W_yW_n \cdots W_1 = U_pU_p^TYX^T(XX^T)^{-1}$$

According to the result in [Baldi and Lu(2012)], this is indeed the expression of a global minimum.

***Condition 3**: $rank(W_n...W_2) < p$.*

Omit.

4 Mathematical Tools for Analysing First and Second Order Based Optimization

An convex minimization problem with finite number of variables could be abstracted into the form below:

$$\min_{\alpha} E(\alpha)$$

In which α denotes the vector of variables, $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_2)^T$. Analogy to the taylor expansion within unary function:

$$f(x) = \sum_{i=0}^{i=2} \frac{1}{i!} f^i(x_0)(x - x_0)^i + O((x - x_0)^2)$$

We have the expansion form within high-dimensional condition:

$$E(\alpha) = E(\alpha_0) + E'(\alpha_0)(\alpha - \alpha_0) + \frac{1}{2}E''(\alpha_0)(\alpha - \alpha_0)^2 + O((\alpha - \alpha_0)^2)$$

For the left of equation is constant, each item in the right should be constant too. Reform the right side we have:

$$E(\alpha) = E(\alpha_0) + (\alpha - \alpha_0)^T E'(\alpha_0) + \frac{1}{2}(\alpha - \alpha_0)^T E''(\alpha_0)(\alpha - \alpha_0) + O((\alpha - \alpha_0)^2)$$

For the gradient based method which update α in direction β with a step size of ε as below:

$$\begin{aligned}
&\Delta E = E(\alpha + \varepsilon \cdot \beta) - E(\alpha) \\
&= E(\alpha) + \varepsilon \cdot \beta^T E'(\alpha) + \frac{1}{2}\varepsilon \cdot \beta^T E''(\alpha_0) \cdot \varepsilon \cdot \beta + O((\varepsilon \cdot \beta)^2) - E(\alpha) \\
&= \varepsilon \cdot \beta^T E'(\alpha) + \frac{1}{2}\varepsilon \cdot \beta^T E''(\alpha_0) \cdot \varepsilon \cdot \beta + O((\varepsilon \cdot \beta)^2)
\end{aligned}$$

Following the definitions, the first order derivation of multivariate function is the gradient vector of it, $E'(\alpha) = \nabla E(\alpha)$. And the second order derivation is its Hessian matrix, $E''(\alpha) = \mathcal{H}_E(\alpha)$. We have:

$$\Delta E = \varepsilon \cdot \beta^T \nabla E(\alpha) + \frac{1}{2}\varepsilon \cdot \beta^T \mathcal{H}_E(\alpha) \cdot \varepsilon \cdot \beta + O((\varepsilon \cdot \beta)^2)$$

As a lemma, a proof is going below:

Lemma 10 *If $\mathcal{H}$ is positive definite, then for all α guarantee $\alpha^T \mathcal{H} \alpha > 0$*

Proof 4.1 *Because $\mathcal{H}$ is a Hermitian, $\mathcal{H} = U\Sigma U^T$. U denote a set of orthogonal basis and let α has coordinate C under U, $\alpha = UC$:*

$$\alpha^T \mathcal{H} \alpha = (UC)^T U \Sigma U^T UC = C^T UC = \sum_{i=1}^{n} c_1^2 \lambda_i$$

Because $\mathcal{H}$ is positive definite, which guarantees $\lambda_i > 0$ for all i. So have $\alpha^T \mathcal{H} \alpha > 0$

Consider the optimal object $\min_{\alpha} E(\alpha)$, we hope $E(\alpha)$ goes down with each step, which requires:

$$\Delta E < 0 \Rightarrow \varepsilon \cdot \beta^T \nabla E(\alpha) + \frac{1}{2}\varepsilon \cdot \beta^T \mathcal{H}_E(\alpha) \cdot \varepsilon \cdot \beta + O((\varepsilon \cdot \beta)^2) < 0$$

Since the second order item is positive definite, ignore the infinitesimal item of higher order,

References

[Baldi and Lu(2012)] Pierre Baldi and Zhiqin Lu. 2012. Complex-Valued Autoencoders. *Neural Networks* 33:136–147. ArXiv: 1108.4135. https://doi.org/10.1016/j.neunet.2012.04.011.

[Kawaguchi(2016)] Kenji Kawaguchi. 2016. Deep Learning without Poor Local Minima. *arXiv:1605.07110 [cs, math, stat]* ArXiv: 1605.07110. http://arxiv.org/abs/1605.07110.

[Zhang(2005)] Fuzhen Zhang, editor. 2005. *The Schur complement and its applications*. Number v. 4 in Numerical methods and algorithms. Springer, New York.