LEARNING DEEP MODELS: CRITICAL POINTS AND LOCAL OPENNESS

Maher Nouiehed **Meisam Razaviyayn**[*]

ABSTRACT

In this paper we present a unifying framework to study the local/global optima equivalence of the optimization problems arising from training non-convex deep models. Using the *local openness* property of the underlying training models, we provide simple sufficient conditions under which any local optimum of the resulting optimization problem is globally optimal. We first *completely characterize the local openness of matrix multiplication mapping in its range*. Then we use our characterization to: 1) show that every local optimum of two layer linear networks is globally optimal. Unlike many existing results, our result requires no assumption on the target data matrix $\boldsymbol{Y}$, and input data matrix $\boldsymbol{X}$. 2) develop *almost complete* characterization of the local/global optima equivalence of multi-layer linear neural networks. 3) show global/local optima equivalence of non-linear deep models having certain pyramidal structure. Unlike some existing works, our result requires no assumption on the differentiability of the activation functions.

1 INTRODUCTION

Deep learning models have recently led to significant practical successes in various fields ranging from computer vision to natural language processing. Despite these significant empirical successes, the theoretical understanding of the behavior of these models is still very limited. To understand the landscape of these non-convex models, we study the general optimization problem

$$\min_{\boldsymbol{w}\in\mathcal{W}} \quad \ell(\mathcal{F}(\boldsymbol{w})), \tag{1}$$

where $\ell : \mathcal{Z} \mapsto \mathbb{R}$ is the loss function and $\mathcal{F} : \mathcal{W} \mapsto \mathcal{Z}$ represents a statistical model with parameter $\boldsymbol{w}$ that needs to be learned by solving the above optimization problem. Here we assume that the set $\mathcal{W}$ is closed and the mapping $\mathcal{F}$ is continuous. In this paper, we use local openness of $\mathcal{F}$ to provide sufficient conditions under which every local optimum of (1) is in fact a global optimum.

To proceed, let us define the auxiliary optimization problem

$$\min_{\boldsymbol{z}\in\mathcal{Z}} \quad \ell(\boldsymbol{z}), \tag{2}$$

where $\mathcal{Z}$ is the range of the mapping $\mathcal{F}$. Since problem (2) minimizes the function $\ell(\cdot)$ over the range of the mapping $\mathcal{F}$, there is a clear relation between the global optimal points of the two optimization problem through the mapping $\mathcal{F}$. However, the connection between the local optima of the two optimization problems is not clear. This connection, in particular, is important when the local optima of (2) are "nice" (e.g. globally optimal or close to optimal). In what follows, we establish the connection between the local optima of the optimization problems (1) and (2) under some simple sufficient conditions. This connection is then used to study the relation between local and global optima of (1) and (2) for various non-convex learning models.

Before proceeding, we define the following concepts. A mapping $\mathcal{F} : \mathcal{W} \to \mathcal{Z}$ is said to be open, if for every open set $U \in \mathcal{W}$, $\mathcal{F}(U)$ is (relatively) open in $\mathcal{Z}$. Moreover, a mapping $\mathcal{F}(\cdot)$ is said to be locally open at $\boldsymbol{w}$ if for every $\epsilon > 0$, there exists $\delta > 0$ such that $\mathcal{B}_\delta\big(\mathcal{F}(\boldsymbol{w})\big) \subseteq \mathcal{F}\big(\mathcal{B}_\epsilon(\boldsymbol{w})\big)$. We call a point $\boldsymbol{W} = (\boldsymbol{W}_h, \ldots, \boldsymbol{W}_1)$, with $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$, *non-degenerate* if $\text{rank}(\boldsymbol{W}_h \cdots \boldsymbol{W}_1) = \min_{0\le i\le h}\, d_i$, and degenerate if $\text{rank}(\boldsymbol{W}_h \cdots \boldsymbol{W}_1) < \min_{0\le i\le h}\, d_i$. We also say a point $\mathbf{W}$ is a *second order saddle point* of an unconstrained optimization problem if the gradient of the objective function is zero at $\mathbf{W}$ and its hessian at $\mathbf{W}$ has a negative eigenvalue. The following simple intuitive observation, which establishes the connection between the local optima of (1) and (2), is a major building block of our analyses.

Observation 1. *Suppose $\mathcal{F}(\cdot)$ is locally open at $\bar{\boldsymbol{w}}$. If $\bar{\boldsymbol{w}}$ is a local minimum of problem (1), then $\bar{\boldsymbol{z}} = \mathcal{F}(\bar{\boldsymbol{w}})$ is a local minimum of problem (2).*

Observation 1 can be used to map multiple local optima of the original problem (1) to one local optimum of the auxiliary problem (2); and potentially simplify the problem. Moreover, this observation motivates us to study the local openness of the matrix multiplication mapping defined by

$$\mathcal{M} : \mathbb{R}^{m\times k} \times \mathbb{R}^{k\times n} \mapsto \mathcal{R}_{\mathcal{M}} \quad \text{with} \quad \mathcal{M}(\boldsymbol{W}_1, \boldsymbol{W}_2) \triangleq \boldsymbol{W}_1\boldsymbol{W}_2, \tag{3}$$

where $\mathcal{R}_{\mathcal{M}} \triangleq \{\, \boldsymbol{Z} \in \mathbb{R}^{m\times n} \mid \text{rank}(\boldsymbol{Z}) \leq \min(m, n, k)\}$ is the range of the mapping $\mathcal{M}$. We study in the next section the local openness/openness of the mapping $\mathcal{M}$. We later use these results to analyze the behavior of local optima of deep neural networks.

2 Local Openness of the Matrix Multiplication Mapping

When $\boldsymbol{W}_1 \in \mathbb{R}^{m\times k}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{k\times n}$ with $k \geq \min\{m, n\}$, the range of the mapping $\mathcal{M}(\mathbf{W}_1, \mathbf{W}_2) = \mathbf{W}_1\mathbf{W}_2$ is the entire space $\mathbb{R}^{m\times n}$. In this case, (Behrends, 2017, Theorem 2.5) provides a complete characterization of the pairs $(\boldsymbol{W}_1, \boldsymbol{W}_2)$ for which the mapping is locally open. However, when $k < \min\{m, n\}$ the characterization of the set of points for which the mapping is locally open has not been resolved before. We settle this question in Theorem 2 stated below

Theorem 2. *Let $\mathcal{M}(\mathbf{W}_1, \mathbf{W}_2) = \mathbf{W}_1\mathbf{W}_2$ denote the matrix multiplication mapping with $\mathbf{W}_1 \in \mathbb{R}^{m\times k}$ and $\mathbf{W}_2 \in \mathbb{R}^{k\times n}$. Assume $k < \min\{m, n\}$. Then if $\text{rank}(\bar{\mathbf{W}}_1) \neq \text{rank}(\bar{\mathbf{W}}_2)$, $\mathcal{M}(\cdot, \cdot)$ is not locally open at $(\bar{\mathbf{W}}_1, \bar{\mathbf{W}}_2)$. Else, if $\text{rank}(\bar{\mathbf{W}}_1) = \text{rank}(\bar{\mathbf{W}}_2)$, then the following statements are equivalent:*

i) $\exists\, \widetilde{\mathbf{W}}_1 \in \mathbb{R}^{m\times k}$ *such that* $\widetilde{\mathbf{W}}_1\bar{\mathbf{W}}_2 = \mathbf{0}$ *and* $\bar{\mathbf{W}}_1 + \widetilde{\mathbf{W}}_1$ *is full column rank.*

ii) $\exists\, \widetilde{\mathbf{W}}_2 \in \mathbb{R}^{k\times n}$ *such that* $\bar{\mathbf{W}}_1\widetilde{\mathbf{W}}_2 = \mathbf{0}$ *and* $\bar{\mathbf{W}}_2 + \widetilde{\mathbf{W}}_2$ *is full row rank.*

iii) $\dim\big(\mathcal{N}(\bar{\mathbf{W}}_1) \cap \mathcal{C}(\bar{\mathbf{W}}_2)\big) = 0.$

iv) $\dim\big(\mathcal{N}(\bar{\mathbf{W}}_2^T) \cap \mathcal{C}(\bar{\mathbf{W}}_1^T)\,\big) = 0.$

v) $\mathcal{M}(\cdot, \cdot)$ *is locally open at* $(\bar{\mathbf{W}}_1, \bar{\mathbf{W}}_2)$ *in its range* $\mathcal{R}_{\mathcal{M}}$*.*

By definition, $\mathcal{M}(\cdot, \cdot)$ is locally open at $(\boldsymbol{W}_1, \boldsymbol{W}_2)$ if for a given $\epsilon > 0$, there exists $\delta > 0$ such that for any $\tilde{\boldsymbol{Z}} = \boldsymbol{Z} + \boldsymbol{R}_\delta \in \mathcal{R}_{\mathcal{M}}$ with $\|\boldsymbol{R}_\delta\| \leq \delta$, there exists $\tilde{\boldsymbol{W}}_1$, $\tilde{\boldsymbol{W}}_2$ with $\|\tilde{\boldsymbol{W}}_1\| \leq \epsilon$, $\|\tilde{\boldsymbol{W}}_2\| \leq \epsilon$, such that $\tilde{\boldsymbol{Z}} = (\boldsymbol{W}_1 + \tilde{\boldsymbol{W}}_1)(\boldsymbol{W}_2 + \tilde{\boldsymbol{W}}_2)$. As a perturbation bound on δ, we show that for any locally open pair $(\boldsymbol{W}_1, \boldsymbol{W}_2)$, given an $\epsilon > 0$, the chosen δ is of order ϵ, i.e., $\delta = \mathcal{O}(\epsilon)$. In the next sections, we use our local openness result to characterize the cases where the local optima of various training optimization problem of the form (2) are globally optimal.

3 Non-linear Deep Neural Network with a Pyramidal Structure:

Consider the non-linear deep neural network optimization problem with a pyramidal structure

$$\min_{\boldsymbol{W}} \; \ell\big(\mathcal{F}_h(\boldsymbol{W})\big) \quad \text{with} \quad \mathcal{F}_i(\boldsymbol{W}) \triangleq \boldsymbol{\sigma}_i\big(\boldsymbol{W}_i\mathcal{F}_{i-1}(\boldsymbol{W})\big), \text{ for } i \in \{2, \ldots, h\}, \tag{4}$$

and $\mathcal{F}_1(\boldsymbol{W}) \triangleq \boldsymbol{\sigma}_1(\boldsymbol{W}_1\mathbf{X})$ where $\boldsymbol{\sigma}_i : \mathbb{R} \mapsto \mathbb{R}$ is a continuous and strictly monotone activation function applied component-wise to the entries of each layer, i.e., $\boldsymbol{\sigma}_i(\boldsymbol{A}) = [\boldsymbol{\sigma}_i(\boldsymbol{A}_{jk})]_{j,k}$. Here $\boldsymbol{W} = \big(\boldsymbol{W}_i\big)_{i=1}^h$ where $\boldsymbol{W}_i \in \mathbb{R}^{d_i\times d_{i-1}}$ is the weight matrix of layer i, $\boldsymbol{X} \in \mathbb{R}^{d_0\times n}$ is the input training data. In this section, we consider the pyramidal network structure with $d_0 > n$ and $d_i \leq d_{i-1}$ for $1 \leq i \leq h$; see Nguyen & Hein (2017) for more details on these types of networks.

First notice that when $\boldsymbol{X}$ is full column rank and the functions $\boldsymbol{\sigma}_i$'s are all continuous and strictly monotone, the image of the mapping $\mathcal{F}_h$ is convex. We now show that when $\boldsymbol{W}_i$'s are all full row rank and the functions $\boldsymbol{\sigma}_i$'s are all strictly monotone, the mapping $\mathcal{F}_h$ is locally open at $\boldsymbol{W}$.

Lemma 3. *Assume the functions $\boldsymbol{\sigma}_i(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ are all continuous strictly monotone. Then the mapping $\mathcal{F}_h$ defined in (4) is locally open at the point $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_h)$ if $\boldsymbol{W}_i$'s are all full row rank.*

Lemma 3 in conjunction with Observation 1 implies that if $\bar{\mathbf{W}}$ is a local optimum of problem (4) with $\bar{\mathbf{W}}_i$'s being full row rank, then $\bar{\boldsymbol{Z}} = \mathcal{F}_h(\bar{\boldsymbol{W}})$ is a local optimum of the corresponding auxiliary problem $\underset{\boldsymbol{Z} \in \mathcal{Z}}{\text{minimize}}\ \ell(\boldsymbol{Z})$ where $\mathcal{Z}$ is convex. Consequently, $\bar{\boldsymbol{Z}}$ is a global optimum of problem (4) when the loss function $\ell(\cdot)$ is convex. A popular activation function that is strictly monotonic and not differentiable is the Leaky ReLU, for which our result follows.

4 Linear Deep Neural Network

Consider the training problem of multi-layer deep linear neural networks:

$$\min_{\boldsymbol{W}} \ \frac{1}{2}\|\boldsymbol{W}_h \cdots \boldsymbol{W}_1 \boldsymbol{X} \ - \ \boldsymbol{Y}\|^2. \tag{5}$$

Here $\boldsymbol{W} = \left(\boldsymbol{W}_i\right)_{i=1}^h$, $\boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ are the weight matrices, $\boldsymbol{X} \in \mathbb{R}^{d_0 \times n}$ is the input training data, and $\boldsymbol{Y} \in \mathbb{R}^{d_h \times n}$ is the target training data. Based on our general framework, the corresponding auxiliary optimization problem is given by

$$\min_{\boldsymbol{Z} \in \mathbb{R}^{d_h \times n}} \frac{1}{2}||\boldsymbol{Z}\boldsymbol{X} \ - \ \boldsymbol{Y}||^2 \quad \text{s.t. } \operatorname{rank}(\boldsymbol{Z}) \leq d_p \triangleq \min_{0 \leq i \leq h} d_i. \tag{6}$$

(Lu & Kawaguchi, 2017, Theorem 2.2) shows that when $\boldsymbol{X}$ is full rank, every local minimum of problem (6) is global. By using local openness, we relax the full rankness assumption on $\boldsymbol{X}$.

Lemma 4. *Every local minimum of problem (6) is global.*

(Yun et al., 2017, Theorem 2.2) shows that when $\boldsymbol{X}\boldsymbol{X}^T$, $\boldsymbol{Y}\boldsymbol{X}^T$, and $\boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y}^T$ are full rank, every local optimum of a linear deep network is global. Moreover, they provide necessary and sufficient conditions for a critical point to be a global minimum. In another result, Lu & Kawaguchi (2017) showed that when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are full row rank, every local minimum of (5) is global. We now relax the full rankness assumptions and reproduce similar results. First consider a two layer linear model, i.e. problem (5) with $h = 2$. Theorem 5, shows, without any assumptions on both $\boldsymbol{X}$ and $\boldsymbol{Y}$, that every local minimum of a two layer linear model is global. Furthermore, Theorem 5 and Corollary 6, show that, even when the square loss error is replaced by a general convex loss function, every degenerate critical point of a two layer linear model is either a global minimum or a second order saddle.

Theorem 5. *Every local minimum of a two layer linear deep model, problem (5) with $h = 2$, is global. Moreover, every degenerate saddle point of problem (5), with $h = 2$, is a second order saddle.*

Corollary 6. *Replace the square loss error in (5) by a general convex loss function $\ell(\cdot)$. Then, for $h = 2$, every degenerate critical point is either a global minimum or a second-order saddle.*

Now consider the general case of multi-layer linear models. Due to a simple counterexample, we cannot in general relax the full rankness assumption on $\mathbf{Y}$. However we determine a set of necessary conditions under which every local minimum of problem (5) is global. The results are stated in Lemma 7 and Theorem 8. We note that Theorem 8 holds when replacing the square error loss by a general convex and differentiable function $\ell(\cdot)$.

Lemma 7. *Every non-degenerate local minimum of (5) is global minimum.*

Theorem 8. *If there exist $1 \leq p_1 < p_2 \leq h - 1$ with $d_h > d_{p_2}$ and $d_0 > d_{p_1}$, we can find a rank deficient $\boldsymbol{Y}$ such that problem (5) has a local minimum that is not global. Otherwise, given any $\boldsymbol{X}$ and $\boldsymbol{Y}$, every local minimum of problem (5) is a global minimum.*

REFERENCES

E. Behrends. Where is matrix multiplication locally open? *Linear Algebra and its Applications*, 517:167–176, 2017.

H. Lu and K. Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

Q. Nguyen and M. Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

C. Yun, S. Sra, and A. Jadbabaie. Global optimality conditions for deep neural networks. *arXiv preprint arXiv:1707.02444*, 2017.