Deep Linear Networks with Arbitrary Loss: All Local Minima Are Global

Thomas Laurent [*1] **James H. von Brecht** [*2]

Abstract

We consider deep linear networks with arbitrary convex differentiable loss. We provide a short and elementary proof of the fact that all local minima are global minima if the hidden layers are either 1) at least as wide as the input layer, or 2) at least as wide as the output layer. This result is the strongest possible in the following sense: If the loss is convex and Lipschitz but not differentiable then deep linear networks **can** have sub-optimal local minima.

1. Introduction

Deep linear networks (DLN) are neural networks that have multiple hidden layers but have no nonlinearities between layers. That is, for given data points $\{\mathbf{x}^{(i)}\}_{i=1}^N$ the outputs of such networks are computed via a series

$$\hat{\mathbf{y}}^{(i)} = W_L W_{L-1} \cdots W_1 \mathbf{x}^{(i)}$$

of matrix multiplications. Given a target $\mathbf{y}^{(i)}$ for the i^{th} data point and a pairwise loss function $\ell(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)})$, forming the usual summation

$$\mathcal{L}(W_1, \ldots, W_L) = \frac{1}{N} \sum_{i=1}^{N} \ell(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}) \qquad (1)$$

then yields the total loss.

Such networks have few direct applications, but they frequently appear as a class of toy models used to understand the loss surfaces of deep neural networks (Saxe et al., 2014; Kawaguchi, 2016; Lu & Kawaguchi, 2017; Hardt & Ma, 2017). For example, numerical experiments indicate that DLNs exhibit some behavior that resembles the behavior of deep nonlinear networks during training (Saxe et al., 2014). Results of this sort provide a small piece of evidence that DLNs can provide a decent simplified model of more realistic networks with nonlinearities.

From an analytical point-of-view, the simplicity of DLNs allows for a rigorous, in-depth study of their loss surfaces. These models typically employ a convex loss function $\ell(\hat{\mathbf{y}}, \mathbf{y})$, and so with one layer (i.e. $L = 1$) the loss $\mathcal{L}(W_1)$ is convex and the resulting optimization problem (1) has no sub-optimal local minimizers. With multiple layers (i.e. $L \geq 2$) the loss $\mathcal{L}(W_1, \ldots, W_L)$ is not longer convex, and so the question of paramount interest concerns whether this addition of depth and the subsequent loss of convexity creates sub-optimal local minimizers. Indeed, most analytical treatments of DLNs focus on this question.

We resolve this question in full for arbitrary convex differentiable loss functions. Specifically, we consider deep linear networks satisfying the two following hypotheses:

(i) The loss function $\hat{\mathbf{y}} \mapsto \ell(\mathbf{y}, \hat{\mathbf{y}})$ is convex and differentiable.

(ii) The thinnest layer is either the input layer or the output layer.

Many networks of interest satisfy both of these hypotheses. The first hypothesis (i) holds for nearly all network criteria, such as the mean squared error loss, the logistic loss or the cross entropy loss, that appear in applications. In a classification scenario, the second hypothesis (ii) holds whenever each hidden layer has more neurons than the number of classes. Thus both hypotheses are often satisfied when using a deep linear network (1) to model its nonlinear counterpart. In any such situation we resolve the deep linear problem in its entirety.

Theorem 1. *If hypotheses (i) and (ii) hold then* (1) *has no sub-optimal minimizers, i.e. any local minimum is global.*

We provide a short, transparent proof of this result. It is easily accessible to any reader with a basic understanding of the singular value decomposition, and in particular, it does not rely on any sophisticated machinery from either optimization or linear algebra. Moreover, this theorem is the strongest possible in the following sense —

Theorem 2. *There exists a convex, Lipschitz but not differentiable function* $\hat{\mathbf{y}} \mapsto \ell(\mathbf{y}, \hat{\mathbf{y}})$ *for which* (1) *has sub-optimal*

[*]Equal contribution [1]Department of Mathematics, Loyola Marymount University, Los Angeles, CA 90045, USA [2]Department of Mathematics and Statistics, California State University, Long Beach, Long Beach, CA 90840, USA. Correspondence to: Thomas Laurent <tlaurent@lmu.edu>, James H. von Brecht <james.vonbrecht@csulb.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

local minimizers.

In other words, we have a (perhaps surprising) hard limit on how far "local equals global" results can reach; differentiability of the loss is essential.

Many prior analytical treatments of DLNs focus on similar questions. For instance, both (Baldi & Hornik, 1989) and (Baldi & Lu, 2012) consider "deep" linear networks with two layers (i.e. $L = 2$) and a mean squared error loss criterion. They provide a "local equals global" result under some relatively mild assumptions on the data and targets. More recently, (Kawaguchi, 2016) proved that deep linear networks with arbitrary number of layers and with mean squared error loss do not have sub-optimal local minima under certain structural assumptions on the data and targets. The follow-up (Lu & Kawaguchi, 2017) futher simplifies the proof of this result and weakens the structural assumptions. Specifically, this result shows that the loss (1) associated with a deep linear network has no sub-optimal local minima provided all of assumptions

(i) The loss $\ell(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}) = \|\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\|^2$ is the mean squared error loss criterion;

(ii) The data matrix $X = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}]$ has full row rank;

(iii) The target matrix $Y = [\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)}]$ has full row rank;

are satisfied. Compared to our result, (Lu & Kawaguchi, 2017) therefore allows for the hidden layers of the network to be thinner than the input and output layers. However, our result applies to network equipped with any differentiable convex loss (in fact any differentiable loss $\mathcal{L}$ for which first-order optimality implies global optimality) and we do not require any assumption on the data and targets. Our proof is also shorter and much more elementary by comparison.

2. Proof of Theorem 1

Theorem 1 follows as a simple consequence of a more general theorem concerning real-valued functions that take as input a product of matrices. That is, we view the deep linear problem as a specific instance of the following more general problem. Let $\mathbb{M}_{m\times n}$ denote the space of $m \times n$ real matrices, and let $f : \mathbb{M}_{d_L \times d_0} \to \mathbb{R}$ denote any differentiable function that takes $d_L \times d_0$ matrices as input. For any such function we may then consider both the single-layer optimization

$$\text{(P1)} \quad \begin{cases} \text{Minimize} \quad f(A) \\ \text{over all } A \text{ in } \mathbb{M}_{d_L \times d_0} \end{cases}$$

as well as the analogous problem

$$\text{(P2)} \quad \begin{cases} \text{Minimize} \quad f(W_L W_{L-1} \cdots W_1) \\ \text{over all } L\text{-tuples } (W_1, \dots, W_L) \\ \text{in } \mathbb{M}_{d_1 \times d_0} \times \cdots \times \mathbb{M}_{d_L \times d_{L-1}} \end{cases}$$

that corresponds to a multi-layer deep linear optimization. In other words, in (P2) we consider the task of optimizing f over those matrices $A \in \mathbb{M}_{d_L \times d_0}$ that can be realized by an L-fold product

$$A = W_L W_{L-1} \cdots W_1 \qquad W_\ell \in \mathbb{M}_{d_\ell \times d_{\ell-1}} \tag{2}$$

of matrices. We may then ask how the parametrization (2) of A as a product of matrices affects the minimization of f, or in other words, whether the problems (P1) and (P2) have similar structure. At heart, the use of DLNs to model nonlinear neural networks centers around this question.

Any notion of structural similarity between (P1) and (P2) should require that their global minima coincide. As a matrix of the form (2) has rank at most $\min\{d_0, \dots, d_L\}$, we must impose the structural requirement

$$\min\{d_1, \dots, d_{L-1}\} \geq \min\{d_L, d_0\} \tag{3}$$

in order to guarantee that (2) does, in fact, generate the full space of $d_L \times d_0$ matrices. Under this assumption we shall prove the following quite general theorem.

Theorem 3. *Assume that $f(A)$ is any differentiable function and that the structural condition* (3) *holds. Then at **any** local minimizer $\left(\hat{W}_1, \dots, \hat{W}_L\right)$ of (P2) the optimality condition*

$$\nabla f\left(\hat{A}\right) = 0 \qquad \hat{A} := \hat{W}_L \hat{W}_{L-1} \cdots \hat{W}_1$$

is satisfied.

Theorem 1 follows as a simple consequence of this theorem. The first hypothesis (i) of theorem 1 shows that the total loss (1) takes the form

$$\mathcal{L}(W_1, \dots, W_L) = f(W_L \cdots W_1)$$

for $f(A)$ some convex and differentiable function. The structural hypothesis (3) is equivalent to the second hypothesis (ii) of theorem 1, and so we can directly apply theorem 3 to conclude that a local minimum $\left(\hat{W}_1, \dots, \hat{W}_L\right)$ of $\mathcal{L}$ corresponds to a critical point $\hat{A} = \hat{W}_L \cdots \hat{W}_1$ of $f(A)$, and since $f(A)$ is convex, this critical point is necessarily a global minimum.

Before turning to the proof of theorem 3 we recall a bit of notation and provide a calculus lemma. Let

$$\langle A, B\rangle_{\text{fro}} := \text{Tr}(A^T B) = \sum_i \sum_j A_{ij} B_{ij} \quad \text{and}$$

$$\|A\|_{\text{fro}}^2 := \langle A, A\rangle_{\text{fro}}$$

denote the Frobenius dot product and the Frobenius norm, respectively. Also, recall that for a differentiable function $\phi : \mathbb{M}_{m\times n} \mapsto \mathbb{R}$ its gradient $\nabla\phi(A) \in \mathbb{M}_{m\times n}$ is the unique matrix so that the equality

$$\phi(A+H) = \phi(A) + \langle \nabla\phi(A), H\rangle_{\rm fro} + o\left(\|H\|_{\rm fro}\right) \quad (4)$$

holds. If $F(W_1,\ldots,W_L) := f(W_L\cdots W_1)$ denotes the objective of interest in (P2) the following lemma gives the partial derivatives of F as a function of its arguments.

Lemma 1. *The partial derivatives of F are given by*

$$\begin{aligned}
\nabla_{W_1}F(W_1,\ldots,W_L) &= W_{2,+}^T\nabla f\big(A\big),\\
\nabla_{W_k}F(W_1,\ldots,W_L) &= W_{k+1,+}^T\nabla f(A)W_{k-1,-}^T,\\
\nabla_{W_L}F(W_1,\ldots,W_L) &= \nabla f\big(A\big)W_{L-1,-}^T,
\end{aligned}$$

where A stands for the full product $A := W_L\cdots W_1$ and $W_{k,+}, W_{k,-}$ are the truncated products

$$\begin{aligned}
W_{k,+} &:= W_L\cdots W_k,\\
W_{k,-} &:= W_k\cdots W_1.
\end{aligned} \quad (5)$$

Proof. The definition (4) implies

$$\begin{aligned}
&F(W_1,\ldots,W_{k-1},W_k+H,W_{k+1},\ldots,W_L)\\
&= f\big(A + W_{k+1,+}HW_{k-1,-}\big)\\
&= f(A) + \langle \nabla f(A), W_{k+1,+}HW_{k-1,-}\rangle_{\rm fro} + o\big(\|H\|_{\rm fro}\big).
\end{aligned}$$

Using the cyclic property $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB)$ of the trace then gives

$$\begin{aligned}
\langle \nabla f(A)\ ,\ W_{k+1,+}HW_{k-1,-}\ \rangle_{\rm fro}&\\
&= \mathrm{Tr}\left(\ \nabla f(A)^T W_{k+1,+}HW_{k-1,-}\ \right)\\
&= \mathrm{Tr}\left(\ W_{k-1,-}\nabla f(A)^T W_{k+1,+}H\ \right)\\
&= \langle\ W_{k+1,+}^T\nabla f(A)W_{k-1,-}^T\ ,\ H\ \rangle_{\rm fro}
\end{aligned}$$

which, in light of (4), gives the desired formula for $\nabla_{W_k}F$. The formulas for $\nabla_{W_1}F$ and $\nabla_{W_L}F$ are obtained similarly. □

Proof of Theorem 3: To prove theorem 3 it suffices to assume that $d_L \geq d_0$ without loss of generality. This follows from the simple observation that

$$g\big(A\big) := f\big(A^T\big)$$

defines a differentiable function of $d_0\times d_L$ matrices for $f(A)$ any differentiable function of $d_L\times d_0$ matrices. As a point $\big(W_1,\ldots,W_L\big)$ defines a local minimum of $f\big(W_LW_{L-1}\cdots W_1\big)$ if and only if $\big(W_1^T,\ldots,W_L^T\big)$ defines a minimum of $g\big(V_1\cdots V_{L-1}V_L\big)$, the theorem for the case $d_L < d_0$ follows by appealing to its $d_L \geq d_0$ instance. It therefore suffices to assume that $d_L \geq d_0$, and by the structural assumption that $d_k \geq d_0$, throughout the remainder of the proof.

Consider any local minimizer $\big(\hat W_1,\ldots,\hat W_L\big)$ of F and denote by $\hat A$, $\hat W_{k,+}$ and $\hat W_{k,-}$ the corresponding full and truncated products (c.f. (5)). By definition of a local minimizer there exists some $\varepsilon_0 > 0$ so that

$$F(W_1,\ldots,W_L) \geq F(\hat W_1,\ldots,\hat W_L) = f\big(\hat A\big) \quad (6)$$

whenever the family of inequalities

$$\|W_\ell - \hat W_\ell\|_{\rm fro} \leq \varepsilon_0 \quad \text{for all} \quad 1\leq \ell\leq L$$

all hold. Moreover, lemma 1 yields

$$\begin{aligned}
&\text{(i)}\ \ 0 = \hat W_{2,+}^T\nabla f\big(\hat A\big),\\
&\text{(ii)}\ \ 0 = \hat W_{k+1,+}^T\nabla f\big(\hat A\big)\hat W_{k-1,-}^T \quad \forall\ 2\leq k\leq L-1,\\
&\text{(iii)}\ \ 0 = \nabla f\big(\hat A\big)\hat W_{L-1,-}^T.
\end{aligned} \quad (7)$$

since all partial derivatives must vanish at a local minimum. If $\hat W_{L-1,-}$ has a trivial kernel, i.e. $\ker(\hat W_{L-1,-}) = \{\mathbf{0}\}$, then the theorem follows easily. The critical point condition (7) part (iii) implies

$$\hat W_{L-1,-}\nabla f\big(\hat A\big)^T = 0,$$

and since $\hat W_{L-1,-}$ has a trivial kernel this implies $\nabla f\big(\hat A\big) = \nabla f\big(\hat W_L\hat W_{L-1}\cdots\hat W_1\big) = 0$ as desired.

The remainder of the proof concerns the case that $\hat W_{L-1,-}$ has a nontrivial kernel. The main idea is to use this nontrivial kernel to construct a family of infinitesimal perturbations of the local minimizer $\big(\hat W_1,\ldots,\hat W_L\big)$ that leaves the overall product $\hat W_L\cdots\hat W_1$ unchanged. In other words, the family of perturbations $\big(\tilde W_1,\ldots,\tilde W_L\big)$ satisfy

$$\|\tilde W_\ell - \hat W_\ell\|_{\rm fro} \leq \epsilon_0/2 \quad \forall \ell = 1,\ldots,L, \quad (8)$$

$$\tilde W_L\tilde W_{L-1}\cdots\tilde W_1 = \hat W_L\hat W_{L-1}\cdots\hat W_1. \quad (9)$$

Any such perturbation also defines a local minimizer.

Claim 1. *Any tuple of matrices $\big(\tilde W_1,\ldots,\tilde W_L\big)$ satisfying* (8) *and* (9) *is necessarily a local minimizer F.*

Proof. For any matrix W_ℓ satisfying $\|W_\ell - \tilde W_\ell\|_{\rm fro} \leq \varepsilon_0/2$, inequality (8) implies that

$$\|W_\ell - \hat W_\ell\|_{\rm fro} \leq \|W_\ell - \tilde W_\ell\|_{\rm fro} + \|\tilde W_\ell - \hat W_\ell\|_{\rm fro} \leq \varepsilon_0$$

Equality (9) combined to (6) then leads to

$$\begin{aligned}
F\big(W_1,\ldots,W_L\big) &\geq f\big(\hat A\big) = f(\hat W_L\cdots\hat W_1)\\
&= f(\tilde W_L\cdots\tilde W_1) = F\big(\tilde W_1,\ldots,\tilde W_L\big)
\end{aligned}$$

for any W_ℓ with $\|W_\ell - \tilde W_\ell\|_{\rm fro} \leq \varepsilon_0/2$ and so the point $(\tilde W_1,\ldots,\tilde W_L)$ defines a local minimum. □

The construction of such perturbations requires a preliminary observation and then an appeal to the singular value decomposition. Due to the definition of $\hat{W}_{k,-}$ it follows that $\ker(\hat{W}_{k+1,-}) = \ker(\hat{W}_{k+1}\hat{W}_{k,-}) \supseteq \ker(\hat{W}_{k,-})$, and so the chain of inclusions

$$\ker(\hat{W}_{1,-}) \subseteq \ker(\hat{W}_{2,-}) \subseteq \cdots \subseteq \ker(\hat{W}_{L-1,-}) \quad (10)$$

holds. Since $\hat{W}_{L-1,-}$ has a nontrivial kernel, the chain of inclusions (10) implies that there exists $k_* \in \{1, \ldots, L-1\}$ such that

$$\ker(\hat{W}_{k,-}) = \{\mathbf{0}\} \quad \text{if } k < k_* \quad (11)$$

$$\ker(\hat{W}_{k,-}) \neq \{\mathbf{0}\} \quad \text{if } k \geq k_* \quad (12)$$

In other words, $\hat{W}_{k_*,-}$ is the first matrix appearing in (10) that has a nontrivial kernel.

The structural requirement (3) and the assumption that $d_L \geq d_0$ imply that $d_k \geq d_0$ for all k, and so the matrix $\hat{W}_{k,-} \in \mathbb{M}_{d_k \times d_0}$ has more rows than columns. As a consequence its full singular value decomposition

$$\hat{W}_{k,-} = \hat{U}_k \hat{\Sigma}_k \hat{V}_k^T \quad (13)$$

has the shape depicted in figure 1. The matrices $\hat{U}_k \in \mathbb{M}_{d_k \times d_k}$ and $\hat{V}_k \in \mathbb{M}_{d_0 \times d_0}$ are orthogonal, whereas $\hat{\Sigma}_k \in \mathbb{M}_{d_k \times d_0}$ is a diagonal matrix containing the singular values of $\hat{W}_{k,-}$ in descending order. From (12) $\hat{W}_{k,-}$ has a nontrivial kernel for all $k \geq k_*$, and so in particular its least singular value is zero. In particular, the (d_0, d_0) entry of $\hat{\Sigma}_k$ vanishes if $k \geq k_*$. Let $\hat{\mathbf{u}}_k$ denote the corresponding d_0^{th} column of $\hat{U}_k$, which exists since $d_k \geq d_0$.

Claim 2. *Let $\mathbf{w}_{k_*+1}, \ldots, \mathbf{w}_L$ denote any collection of vectors and $\delta_{k_*+1}, \ldots, \delta_L$ any collection of scalars satisfying*

$$\mathbf{w}_k \in \mathbb{R}^{d_k}, \quad \|\mathbf{w}_k\|_2 = 1 \quad \textit{and} \quad (14)$$

$$0 \leq \delta_k \leq \epsilon_0/2 \quad (15)$$

for all $k_* + 1 \leq k \leq L$. Then the tuple of matrices $(\tilde{W}_1, \ldots, \tilde{W}_L)$ defined by

$$\tilde{W}_k := \begin{cases} \hat{W}_k & \textit{if} \quad 1 \leq k \leq k_* \\ \hat{W}_k + \delta_k \mathbf{w}_k \hat{\mathbf{u}}_{k-1}^T & \textit{else}, \end{cases} \quad (16)$$

satisfies (8) *and* (9).

Proof. Inequality (8) follows from the fact that

$$\|\tilde{W}_k - \hat{W}_k\|_{\text{fro}} = \delta_k \|\mathbf{w}_k \hat{\mathbf{u}}_{k-1}^T\|_{\text{fro}} = \delta_k \|\mathbf{w}_k\|_2 \|\hat{\mathbf{u}}_{k-1}\|_2$$

for all $k > k^*$, together with the fact that $\hat{\mathbf{u}}_{k-1}$ and $\mathbf{w}_k$ are unit vectors and that $0 \leq \delta_k \leq \epsilon_0/2$.

To prove (9) let $\tilde{W}_{k,-} = \tilde{W}_k \cdots \tilde{W}_1$ and $\hat{W}_{k,-} = \hat{W}_k \cdots \hat{W}_1$ denote the truncated products of the matrices

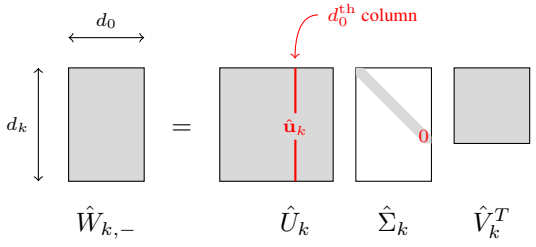

Figure 1. Full singular value decomposition of $\hat{W}_{k,-} \in \mathbb{M}_{d_k \times d_0}$. If $k \geq k_*$ then $\hat{W}_{k,-}$ does not have full rank and so the (d_0, d_0) entry of $\hat{\Sigma}_k$ is 0. The d_0^{th} column of $\hat{U}_k$ exists since $d_k \geq d_0$.

$\tilde{W}_k$ and $\hat{W}_k$. The equality $\tilde{W}_{k_*,-} = \hat{W}_{k_*,-}$ is immediate from the definition (16). The equality (9) will then follow from showing that

$$\tilde{W}_{k,-} = \hat{W}_{k,-} \quad \text{for all} \quad k_* \leq k \leq L. \quad (17)$$

Proceeding by induction, assume that $\tilde{W}_{k,-} = \hat{W}_{k,-}$ for a given $k \geq k_*$. Then

$$\begin{aligned} \tilde{W}_{k+1,-} &= \tilde{W}_{k+1}\tilde{W}_{k,-} \\ &= \tilde{W}_{k+1}\hat{W}_{k,-} \quad \text{(induction hypothesis)} \\ &= \left(\hat{W}_{k+1} + \delta_{k+1}\mathbf{w}_{k+1}\hat{\mathbf{u}}_k^T\right)\hat{W}_{k,-} \\ &= \hat{W}_{k+1,-} + \delta_{k+1}\mathbf{w}_{k+1}\mathbf{u}_k^T\hat{W}_{k,-} \end{aligned}$$

The second term of the last line vanishes, since

$$\mathbf{u}_k^T\hat{W}_{k,-} = \mathbf{u}_k^T\hat{U}_k\hat{\Sigma}_k\hat{V}_k^T = \mathbf{e}_{d_0}^T\hat{\Sigma}_k\hat{V}_k^T = \mathbf{0}$$

with $\mathbf{e}_{d_0} \in \mathbb{R}^{d_k}$ the d_0^{th} standard basis vector. The second equality comes from the fact that the columns of $\hat{U}_k$ are orthonormal, and the last equality comes from the fact that $\mathbf{e}_{d_0}^T\Sigma_{k_*} = \mathbf{0}$ since the d_0^{th} row of $\hat{\Sigma}_{k_*}$ vanishes. Thus (17) holds, and so (9) holds as well. □

Claims 1 and claim 2 show that the perturbation $\left(\tilde{W}_1, \ldots, \tilde{W}_L\right)$ defined by (16) is a local minimizer of F. The critical point conditions

$$\begin{aligned} &\text{(i)} \quad 0 = \tilde{W}_{2,+}^T \nabla f\left(\tilde{A}\right), \\ &\text{(ii)} \quad 0 = \tilde{W}_{k+1,+}^T \nabla f\left(\tilde{A}\right)\tilde{W}_{k-1,-}^T \quad \forall\, 2 \leq k \leq L-1, \\ &\text{(iii)} \quad 0 = \nabla f\left(\tilde{A}\right)\tilde{W}_{L-1,-}^T \end{aligned}$$

therefore hold as well **for all choices of $\mathbf{w}_{k_*+1}, \ldots, \mathbf{w}_L$ and $\delta_{k_*+1}, \ldots, \delta_L$ satisfying** (14) **and** (15)**.**

The proof concludes by appealing to this family of critical point relations. If $k_* > 1$ the transpose of condition (ii) gives

$$\hat{W}_{k_*-1,-} \nabla f\left(\hat{A}\right)^T \tilde{W}_{k_*+1,+} = 0 \quad (18)$$

since the equalities $\tilde{W}_{k_*-1,-} = \hat{W}_{k_*-1,-}$ (c.f. (16)) and $\tilde{A} = \tilde{W}_L \cdots \tilde{W}_1 = \hat{W}_L \cdots \hat{W}_1 = \hat{A}$ (c.f. (9)) both hold. But $\ker(\hat{W}_{k_*-1,-}) = \{\mathbf{0}\}$ by definition of k_* (c.f. (11)), and so

$$\nabla f\big(\hat{A}\big)^T \tilde{W}_L \cdots \tilde{W}_{k_*+1} = 0. \tag{19}$$

must hold as well. If $k_* = 1$ then (19) follows trivially from the critical point condition (i). Thus (19) holds for all choices of $\mathbf{w}_{k_*+1}, \ldots, \mathbf{w}_L$ and $\delta_{k_*+1}, \ldots, \delta_L$ satisfying (14) and (15). First choose $\delta_{k_*+1} = 0$ so that $\tilde{W}_{k_*+1} = \hat{W}_{k_*+1}$ and apply (19) to find

$$\nabla f\big(\hat{A}\big)^T \tilde{W}_L \cdots \tilde{W}_{k_*+2} \hat{W}_{k_*+1} = 0. \tag{20}$$

Then take any $\delta_{k_*+1} > 0$ and substract (20) from (19) to get

$$\begin{aligned}&\frac{1}{\delta_{k_*+1}} \nabla f\big(\hat{A}\big)^T \tilde{W}_L \cdots \tilde{W}_{k_*+2} \left(\tilde{W}_{k_*+1} - \hat{W}_{k_*+1}\right) \\ &= \nabla f\big(\hat{A}\big)^T \tilde{W}_L \cdots \tilde{W}_{k_*+2} \big(\mathbf{w}_{k_*+1} \hat{\mathbf{u}}_{k_*}^T\big) = 0\end{aligned}$$

for $\mathbf{w}_{k_*+1}$ an arbitrary vector with unit length. Right multiplying the last equality by $\hat{\mathbf{u}}_{k_*}$ and using the fact that $(\mathbf{w}_{k_*+1} \hat{\mathbf{u}}_{k_*}^T) \hat{\mathbf{u}}_{k_*} = \mathbf{w}_{k_*+1} \hat{\mathbf{u}}_{k_*}^T \hat{\mathbf{u}}_{k_*} = \mathbf{w}_{k_*+1}$ shows

$$\nabla f\big(\hat{A}\big)^T \tilde{W}_L \cdots \tilde{W}_{k_*+2} \mathbf{w}_{k_*+1} = 0 \tag{21}$$

for all choices of $\mathbf{w}_{k_*+1}$ with unit length. Thus (21) implies

$$\nabla f\big(\hat{A}\big)^T \tilde{W}_L \cdots \tilde{W}_{k_*+2} = 0$$

for all choices of $\mathbf{w}_{k_*+2}, \ldots, \mathbf{w}_L$ and $\delta_{k_*+2}, \ldots, \delta_L$ satisfying (14) and (15). The claim

$$\nabla f\big(\hat{A}\big) = 0$$

then follows by induction. □

3. Concluding Remarks

Theorem 3 provides the mathematical basis for our analysis of deep linear problems. We therefore conclude by discussing its limits.

First, theorem 3 fails if we refer to critical points rather than local minimizers. To see this, it suffices to observe that the critical point conditions for problem (P2),

$$\begin{aligned}&\text{(i)} \;\; 0 = \hat{W}_{2,+}^T \nabla f\big(\hat{A}\big), \\ &\text{(ii)} \;\; 0 = \hat{W}_{k+1,+}^T \nabla f\big(\hat{A}\big) \hat{W}_{k-1,-}^T \quad \forall\, 2 \le k \le L-1, \\ &\text{(iii)} \;\; 0 = \nabla f\big(\hat{A}\big) \hat{W}_{L-1,-}^T\end{aligned}$$

where $\hat{W}_{k,+} := \hat{W}_L \cdots \hat{W}_{k+1}$ and $\hat{W}_{k,-} := \hat{W}_{k-1} \cdots \hat{W}_1$, clearly hold if $L \ge 3$ and all of the $\hat{W}_\ell$ vanish. In other words, the collection of zero matrices always defines a critical point for (P2) but clearly $\nabla f\big(\mathbf{0}\big)$ need not vanish. To put it otherwise, if $L \ge 3$ the problem (P2) always has saddle-points even though all local optima are global.

Second, the assumption that $f(A)$ is differentiable is necessary as well. More specifically, a function of the form

$$F(W_1, \ldots, W_L) := f(W_L \cdots W_1)$$

can have sub-optimal local minima if $f(A)$ is convex and globally Lipschitz but is not differentiable. A simple example demonstrates this, and therefore proves theorem 2. For instance, consider the bi-variate convex function

$$f(x, y) := |x| + (1-y)_+ - 1, \quad (y)_+ := \max\{y, 0\}, \tag{22}$$

which is clearly globally Lipschitz but not differentiable. The set

$$\arg\min \; f := \{(x, y) \in \mathbb{R}^2 : x = 0,\, y \ge 1\}$$

furnishes its global minimizers while $f_{\text{opt}} = -1$ gives the optimal value. For this function even a two layer deep linear problem

$$F\big(W_1, W_2\big) := f(W_2 W_1) \quad W_2 \in \mathbb{R}^2,\; W_1 \in \mathbb{R}$$

has sub-optimal local minimizers; the point

$$(\hat{W}_1, \hat{W}_2) = \left(0, \begin{bmatrix}1\\0\end{bmatrix}\right) \tag{23}$$

provides an example of a sub-optimal solution. The set of all possible points in $\mathbb{R}^2$

$$\begin{aligned}&\mathcal{N}(\hat{W}_1, \hat{W}_2) := \\ &\left\{W_2 W_1 : \|W_2 - \hat{W}_2\| \le \frac{1}{4},\; \|W_1 - \hat{W}_1\| \le \frac{1}{4}\right\}\end{aligned}$$

generated by a $1/4$-neighborhood of the optimum (23) lies in the two-sided, truncated cone

$$\mathcal{N}(\hat{W}_1, \hat{W}_2) \subset \left\{(x, y) \in \mathbb{R}^2 : |x| \le \frac{1}{2}, |y| \le \frac{1}{2}|x|\right\},$$

and so if we let $x \in \mathbb{R}$ denote the first component of the product $W_2 W_1$ then the inequality

$$f(W_2 W_1) \ge \frac{1}{2}|x| \ge 0 = f(\hat{W}_2 \hat{W}_1)$$

holds on $\mathcal{N}(\hat{W}_1, \hat{W}_2)$ and so $(\hat{W}_1, \hat{W}_2)$ is a sub-optimal local minimizer. Moreover, the minimizer $(\hat{W}_1, \hat{W}_2)$ is a *strict* local minimizer in the only sense in which strict optimality can hold for a deep linear problem. Specifically, the strict inequality

$$f(W_2 W_1) > f(\hat{W}_2 \hat{W}_1) \tag{24}$$

holds on $\mathcal{N}(\hat{W}_1, \hat{W}_2)$ unless $W_2 W_1 = \hat{W}_2 \hat{W}_1 = \mathbf{0}$; in the latter case (W_1, W_2) and $(\hat{W}_1, \hat{W}_2)$ parametrize the same

point and so their objectives must coincide. We may identify the underlying issue easily. The proof of theorem 3 requires a single-valued derivative $\nabla f(\hat{A})$ at a local optimum, but with $f(x, y)$ as in (22) its subdifferential

$$\partial f(\mathbf{0}) = \{(x, y) \in \mathbb{R}^2 : -1 \leq x \leq 1,\ y = 0\}$$

is multi-valued at the sub-optimal local minimum (23). In other words, if a globally convex function $f(A)$ induces sub-optimal local minima in the corresponding deep linear problem then $\nabla f(\hat{A})$ cannot exist at any such sub-optimal solution (assuming the structural condition, of course).

Third, the structural hypothesis

$$d_\ell \geq \min\{d_L, d_0\} \quad \text{for all } \ell \in \{1, \dots, L\}$$

is necessary for theorem 3 to hold as well. If $d_\ell < \min\{d_0, d_L\}$ for some ℓ the parametrization

$$A = W_L \cdots W_1$$

cannot recover full rank matrices. Let $f(A)$ denote any function where ∇f vanishes only at full rank matrices. Then

$$\nabla f\big(W_L \cdots W_1\big) \neq \mathbf{0}$$

at all critical points of (P2), and so theorem 3 fails.

Finally, if we do not require convexity of $f(A)$ then it is not true, in general, that local minima of (P2) correspond to minima of the original problem. The functions

$$f(x, y) = x^2 - y^2 \qquad F(W_1, W_2) = f(W_2 W_1)$$

and the minimizer (23) illustrate this point. While the origin is clearly a saddle point of the one layer problem, the argument leading to (24) shows that (23) is a local minimizer for the deep linear problem. So in the absence of additional structural assumptions on $f(A)$, we may infer that a minimizer of the deep linear problem satisfies first-order optimality for the original problem, but nothing more.

References

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Baldi, P. and Lu, Z. Complex-valued autoencoders. *Neural Networks*, 33:136–147, 2012.

Hardt, M. and Ma, T. Identity matters in deep learning. In *International Conference on Learning Representations*, 2017.

Kawaguchi, K. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2016.

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

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.