École Polytechnique Fédérale de Lausanne
Lausanne | Vaud | Switzerland

Institute of Mathematics

College of Basic Sciences

Report

— MATH-498 | Semester Project | (Spring 2023)

By:
Edmund Hofflin

Supervisors:
Prof. Nicolas Boumal
Chris Criscitiello

June 2023

Table of Contents

Acknowledgements I would like to thank Christopher Criscitiello for supervising the project. His direction was incredibly helpful, pointing me towards productive avenues of investigation while also preventing me from getting too stuck in needlessly abstract arguments. His Mathematica skills saved myself significant time and provided the confirmation that I was on the right track.

1. Introduction

Over the past decade, deep neural networks (DNNs) have seen significant success, being at the center of many ground-breaking applications in a variety of fields and domains [1]. However, despite their success, relatively little is known about loss landscapes of DNNs [15]. Deep linear networks (DLNs) are simplified versions of DNNs, lacking activation functions between the hidden layers. While DLNs are machine learning methods, their linear form limits their application. However, this form and their relation to DNNs makes analysis of them provide a glimpse into the function and loss landscapes of DNNs. Consequentially, significant research has been completed into properties and behaviour of DLNs [16; 7; 8; 3; 5].

In particular, in their paper 'Deep Linear Networks with Arbitrary Loss: All Local Minima are Global' [9], Laurent and von Brecht demonstrate that DLNs have no spurious local minima when the loss function is convex and the dimensions of the internal hidden layers is at least as big as the input or output layer. We can use a Riemannian optimisation framework developed by E. Levin et al. [11] to interpret this result in a more general setting: the smooth function

$$\phi : \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}} \to \mathbb{R}, \qquad \phi\left(W_1, \ldots, W_L\right) \to W_L \ldots W_1$$

which defines the DLN model, satisfies the "local $\implies$ 1" property when $\min\left(d_0, \ldots, d_L\right)$ is either d_0 or d_L. That is, given any smooth function f, if $\left(f \circ \phi\right)\left(W_1, \ldots, W_L\right)$ is a local minimum over $\mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}}$ then $f\left(W_L \ldots W_1\right)$ is a 1-critical point over $\mathbb{R}^{d_L \times d_0}$, provided that $\min\left(d_0, \ldots, d_L\right)$ is either d_0 or d_L.

So Laurent and von Brecht's already insightful result regarding DLNs can be re-interpreted as a result regarding a wide class of optimisation problems. Furthermore, the Riemannian optimisation framework also opens many interesting questions: Firstly, does ϕ still satisfy "local $\implies$ 1" without the restriction on the hidden dimensions? Secondly, do critical points of $\left(f \circ \phi\right)$ correspond to critical points of f? Finally, when do local minimum of $\left(f \circ \phi\right)$ map directly to local minima of f, thereby removing conditions on the loss function of the DLN while still removing spurious local minima? In this report, we seek to answer these questions.

This report begins by outlining the Background of the Parametrisation Framework used to interpret and consider the Problem describing the matrix multiplication lift. The Results section is split into four parts. In the first, we derive new Sufficient Conditions for "Local Implies One" with respect to the matrix multiplication lift. In the second, we use these conditions to construct Counter Examples for "Local Implies One", in the process also disproving that the lift is "3 $\implies$ 1". Motivated by these counter-examples, the third part of the results sections focusses on finding Sufficient and Necessary Conditions for Local Implies One, allowing us to completely describe when matrix multiplication lift is "local $\implies$ 1". Finally, we conclude the Results section by considering when "Local Implies Local". This report then closes with a discussion of potential avenues for Future Work and a summarising Conclusion.

Notation $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, while $\|\cdot\|_2$ denotes the 2-norm. Similarly, $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ and $\langle\cdot,\cdot\rangle_2$ denote the Frobenius and Euclidean inner products, respectively. An unspecified norm $\|\cdot\|$ or inner product $\langle\cdot,\cdot\rangle$ are used to signify that any arbitrary norm or inner product can be used.

We use $\succeq$ to denote the partial ordering over the cone of positive semi-definite matrices. In particular, $A \succeq 0$ denotes A being positive semi-definite. We use the operators $\operatorname{im}, \operatorname{rank}, \ker, \operatorname{nullity}$ to denote the image, rank, kernel, and nullity, i.e. dimension of the kernel, respectively.

Finally, we use shorthand $[n] := \{1, \ldots, n\}$.

2. Background

In this section, we formally outline the background of the problem in question. We split this into two parts. The first part outlines the abstract framework through which we interpret the problem, covering the fundamental definitions and a few critical results. The second part outlines the problem itself, defining the central model, motivation, and desired result.

2.1. Parametrisation Framework

We now outline the abstract framework through which we interpret the problem. We do note that this framework is not critical to the problem itself: the problem can be disconnected entirely the framework. However, as illustrated in the Introduction, this interpretation not only allows the results to generalise immediately to other optimisation problems, but also provides a framework for considering further properties and results. This part predominately follows 'The Effect of Smooth Parametrizations on Nonconvex Optimization Landscapes' by E. Levin et al.[11].

2.1.1. Setup

The general framework is setup as follows: Let $\mathcal{E}$ be a linear space with search space $\mathcal{X} \subseteq \mathcal{E}$. Let $\mathcal{M}$ be a smooth manifold and $\phi : \mathcal{M} \to \mathcal{E}$ a smooth function such that it parametrises $\mathcal{X}$, i.e. $\operatorname{im}(\phi) = \mathcal{X}$. Finally, let $f : \mathcal{X} \to \mathbb{R}$ be a smooth objective function and $g := f \circ \phi$. We consider two related optimisation problems:

$$\min_{x \in \mathcal{X}} f(x) \tag{P}$$

$$\min_{y \in \mathcal{M}} g(y) \tag{Q}$$

We call ϕ the *Lift* of $\mathcal{X}$ to $\mathcal{M}$. Figure 1 diagrammatically outlines the relations between the various spaces and functions.

$$\begin{tikzcd}
\mathcal{M} \arrow[d, "\phi"'] \arrow[dr, "g=f\circ\phi"] & \\
\mathcal{X} \subseteq \mathcal{E} \arrow[r, "f"'] & \mathbb{R}
\end{tikzcd}$$

Figure 1: Search Spaces and Objection Functions of Problems (P) and (Q).

2.1.2. Definitions

We now establish various constructions, properties, and results that relate to the framework. We separate these into three topics: Cones, Desirable Points, and Desirable Lifts.

Cones We are interested in two types of cones: the dual and tangent cones. We begin with the dual cone.

Definition 2.1 (Dual Cone [14, Def 2.23])**.** Let K be a cone in $\mathbb{R}^n$. The set

$$K^\circ := \left\{ y \in \mathbb{R}^n : \forall x \in K \left[\langle y, x \rangle \leq 0 \right] \right\}$$

is the *Polar Cone* of K. The *Dual Cone* of K is $K^\star := -K^\circ$.

Note that both the polar and dual cones are closed cones, although they are not necessarily convex [14, Lmm 2.25]. We move to the dual cone.

Definition 2.2 (Tangent Cone [14, Def 3.11])**.** A direction d is called *Tangent* to the set $\mathcal{X} \subseteq \mathbb{R}^n$ at the point $x \in \mathcal{X}$ if there exists a sequence $\{x_k\}_{k\in\mathbb{Z}_{\geq 0}} \subseteq \mathcal{X}$ and scalars $\tau_k > 0$ for all $k \in \mathbb{Z}_{\geq 0}$ such that $x_k \to x$, $\tau_k \to 0$, and

$$d = \lim_{k\to\infty} \frac{x_k - x}{\tau_k}$$

The *Tangent Cone* of $\mathcal{X}$ at x, denoted $\mathrm{T}_x\mathcal{X}$, is the set of all tangent directions. That is:

$$\mathrm{K}_x\mathcal{X} := \left\{ d = \lim_{k\to\infty} \frac{x_k - x}{\tau_k} : x_k \to x, \tau_k \downarrow 0 \right\}$$

The tangent cone is also known as the *Contingent* or *Bouligand* tangent cone.

It is immediate that if γ is a differentiable curve in $\mathcal{X}$ with $\gamma(0) = x$, then $\gamma'(0) \in \mathrm{K}_x\mathcal{X}$. Furthermore, if $\mathcal{X}$ is a smooth embedded submanifold of a linear space $\mathcal{E}$ near x, then $\mathrm{K}_x\mathcal{X} = \mathrm{T}_x\mathcal{X}$, i.e. the tangent cone agrees with the usual tangent space of $\mathcal{X}$ at x [13, Ex. 6.8].

Desirable Points We first outline the desirable and interesting points in the optimisation problem (P):

Definition 2.3 (Desirable Points for (P) [11, Def 2.2])**.** A point $x \in \mathcal{X}$ is a

(i) *Global Minimum* for (P) if $f(x) = \min_{x'\in\mathcal{X}} f(x')$;

(ii) *Local Minimum* for (P) if there is a neighbourhood $U \subseteq \mathcal{X}$ of x such that $f(x) = \min_{x'\in U} f(x')$; and,

(iii) *First-Order Stationary Point* for (P) if $\mathrm{D} f(x)[v] \geq 0$ for all $v \in \mathrm{T}_x\mathcal{X}$, or equivalently if $\nabla f(x)$ is in the dual of the tangent cone $(\mathrm{T}_x\mathcal{X})^\star$.

Note that all global minima of (P) are local minima, and all local minima are first-order stationary points [14, Thm 3.24].

We now outline the desirable and interesting points in the optimisation problem (Q):

Definition 2.4 (Desirable Points for (Q) [11, Def 2.4])**.** A point $y \in \mathcal{M}$ is a

(i) *Global Minimum* for (Q) if $g(y) = \min_{y'\in\mathcal{M}} g(y')$;

(ii) *Local Minimum* for (Q) if there is a neighbourhood $V \subseteq \mathcal{M}$ of y such that $g(y) = \min_{y'\in V} g(y')$;

(iii) *First-Order Stationary Point*, also called 1-*Critical*, for (Q) if for each smooth curve $c : I \to \mathcal{M}$ satisfying $c(0) = y$ we have $(g\circ c)'(0) \geq 0$, or equivalently $(g\circ c)'(0) = 0$; and,

(iv) *Second-Order Stationary Point*, also called 2-*Critical*, for (Q) if it is 1-Critical and $(g\circ c)''(0) \geq 0$ for all smooth curves satisfying $c : I \to \mathcal{M}$ with $c(0) = y$.

(v) *Third-Order Stationary Point*, also called 3-*Critical*, for (Q) if it is 1- and 2-Critical and

$(g \circ c)''' (0) = 0$ for all smooth curves satisfying $c : I \to \mathcal{M}$ with $c(0) = y$.

If $\mathcal{M}$ is embedded in a linear space, first-order stationary in Definition 2.4 (iii) coincides with Definition 2.3 (iii) by [13, Ex. 6.8].

The following proposition shows that 1- and 2-Critical points have the usual sufficient and necessary conditions with respect to their Riemannian gradients.

Proposition 2.1 ([11, Prop 2.5])**.** *A point $y \in \mathcal{M}$ is*

(i) 1*-Critical for* (Q) *if and only if $\nabla g(y) = 0$;*

(ii) 2*-Critical if and only if $\nabla g(y) = 0$ and $\nabla^2 g(y) \succeq 0$; and,*

(iii) 3*-Critical if and only if $\nabla g(y) = 0$, $\nabla^2 g(g) \succeq 0$, and $\nabla^3 g(y)[\dot{y}, \dot{y}, \dot{y}] = 0$ for all $\dot{y}$ such that $\nabla^2 g(y)[\dot{y}] = 0$.*

Desirable Lifts We now focus on lifts: $\phi : \mathcal{M} \to \mathcal{E}$ with $\operatorname{im}(\phi) = \mathcal{X}$. In our setup, we assumed that the lift is smooth. Formally, that is

Definition 2.5 (Smooth Lift [11, Def 2.3])**.** A map $\phi : \mathcal{M} \to \mathcal{E}$ is a *Smooth Lift* of $\mathcal{X} \subseteq \mathcal{E}$ if $\mathcal{M}$ is a smooth manifold, ϕ is a smooth map, and $\operatorname{im}(\phi) = \mathcal{X}$.

Smoothness is required to guarantee that $g = f \circ \phi$ is still continuous and has the same number of higher-order derivatives as f. With these properties guaranteed, we can analyse ϕ according to how it impacts desirable points of (Q).

Definition 2.6 (Desirable Properties of Lifts [11, Def 2.7])**.** Suppose $\phi : \mathcal{M} \to \mathcal{X}$ is a smooth lift.

(i) The lift satisfies *"local $\implies$ local"* property at $y \in \mathcal{M}$ if, for all continuous functions $f : \mathcal{X} \to \mathbb{R}$, y is a local minimum for (Q) then $x = \phi(y)$ for a local minimum for (P). We say that ϕ satisfies the *"local $\implies$ local"* property if its does so at all $y \in \mathcal{M}$.

(ii) The lift ϕ satisfies the *"local $\implies$ 1"* property at $y \in \mathcal{M}$, if, for all differentiable $f : \mathcal{X} \to \mathbb{R}$, y is a local minimum of (Q) then $x = \phi(y)$ is stationary for (P). We say that ϕ satisfies the *"local $\implies$ 1"* property if it does so at all $y \in \mathcal{M}$.

(iii) The lift ϕ satisfies the *"k $\implies$ 1"* property at $y \in \mathcal{M}$ for $k = 1, 2$, if, for all k times differentiable $f : \mathcal{X} \to \mathbb{R}$, y is k-Critical for (Q) then $x = \phi(y)$ is stationary for (P). We say that ϕ satisfies the *"k $\implies$ 1"* property if it does so at all $y \in \mathcal{M}$.

Ideally, a smooth lift is both "local $\implies$ local" and "1 $\implies$ 1". However, this is often not the case and so we must consider smooth lifts that are "local $\implies$ 1" or "2 $\implies$ 1".

Fortunately, openness is a necessary and sufficient condition of a smooth lift being "local $\implies$ local".

Definition 2.7 (Open Map [11, Def A.1])**.** Let $\mathcal{M}$ be a topological space and $\mathcal{X}$ a metric space. Let $\phi : \mathcal{M} \to \mathcal{X}$ such that $\phi(y) = x$. ϕ is *Open* at y if $\phi(U)$ is a neighbourhood of x in $\mathcal{X}$ for all (open or closed) neighbourhoods of y in $\mathcal{M}$.

Theorem 2.1 (Open Equivalence with "Local $\implies$ Local" [11, Thm A.2]). *If $\mathcal{M}$ is a Hausdorff, second-countable, and locally compact topological space, then a smooth lift ϕ is open at $y \in \mathcal{M}$ if and only if it satisfies the "local $\implies$ local" property at y.*

Note that if $\mathcal{M}$ is a topological manifold, then it is Hausdorff, second-countable, and locally compact, and so the theorem applies immediately. Therefore, in such setting, openness provides us with a direct method for determining whether a smooth lift is "'local $\implies$ local".

2.2. Problem

We now outline the specific problem we consider. We split this outline into two parts: the Model and Machine Learning Problem and Matrix Multiplication Lift.

2.2.1. Model and Machine Learning Problem

The problem we consider centers on the following model.

Definition 2.8 (Deep Linear Network). Let $L \in \mathbb{Z}_{\geq 0}$ be the number of layers in the network. Let $d_0, \ldots, d_L \in \mathbb{Z}_{>0}$ be the dimensions of the layers, i.e. d_0 and d_L are the dimensions of the input and output, respectively. A *Deep Linear Network* (DLN) with these hyperparameters is a function of the form

$$\begin{aligned} \mathrm{DLN} : \mathbb{R}^{d_0} \times \mathbb{R}^{d_1 \times d_0} \times \mathbb{R}^{d_2 \times d_1} \times \ldots \times \mathbb{R}^{d_n \times d_{n-1}} \to \mathbb{R}^{d_L} \\ \mathrm{DLN}\left(\mathbf{x}; W_1, \ldots, W_L\right) := W_L W_{L-1} \ldots W_1 \mathbf{x} \end{aligned}$$

We denote the minimum width of a layer by $d_{\min} = \min\left\{d_0, \ldots, d_n\right\}$.

Deep linear networks are almost always associated with the optimisation problem

$$\underset{\{W_i\}_{i \in [L]} \in \mathcal{N}}{\arg\min} \; \mathcal{L}\left(W_1, \ldots, W_L\right) := \frac{1}{N} \sum_{i=1}^{n} \ell\left(\hat{\mathbf{y}}^{(\mathbf{i})}, \mathbf{y}^{(\mathbf{i})}\right) \tag{ML}$$

where $\mathcal{N} = \mathbb{R}^{d_1, \times d_0} \times \ldots \times \mathbb{R}^{d_n \times d_{n-1}}$ is the *Weight Space*, $\ell : \mathbb{R}^{d_L} \times \mathbb{R}^{d_L} \to \mathbb{R}$ is a *Loss Function*, $\left\{\left(\mathbf{x}^{(\mathbf{i})}, \mathbf{y}^{(\mathbf{i})}\right)\right\}_{i \in [N]}$ is a *Dataset* of inputs and target outputs, and we compute $\hat{\mathbf{y}}^{(\mathbf{i})} := \mathrm{DLN}\left(\mathbf{x}^{(\mathbf{i})}; W_1, \ldots, W_L\right)$ for each $i \in [N]$.

While this optimisation problem (ML) sits within the standard supervised learning and risk minimisation paradigm of machine learning [4], deep linear networks themselves are not commonly used in practice. Instead, they are primarily used as toy models to anaylse and understand deep neural networks, allowing their loss landscapes, hyperparameters, convergence properties, and more to be examined in a simpler setting [16; 7; 8; 3; 5].

2.2.2. Matrix Multiplication Lift

The paper 'Deep Linear Networks with Arbitrary Loss: All Local Minima are Global' by T. Laurent and J. von Brecht [9] centers on the following result:

Theorem 2.2 (Absence of Supurious Local Minimisers for DLN with Higher Hidden Dimensions [9, Thm 1]). *For problem ML, if*

(i) the loss function $y \mapsto \ell\left(y, \hat{y}\right)$ is convex and differentiable;

(ii) the thinnest layer is either the input layer of the output layer, i.e. $d_{\min} \geq \min(d_0, d_N)$*;*

then the problem has no sub-optimal minimizers, i.e. any local minimum is global.

This result is a direct consequence of the more general theorem:

Theorem 2.3 ([9, Thm 3])**.** *Let* $L \in \mathbb{Z}_{\geq 0}$ *and* $d_0, \ldots, d_L \in \mathbb{Z}_{>0}$*. Suppose we have differentiable functions*

$$g : \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}} \to \mathbb{R}, \qquad f : \mathbb{R}^{d_L \times d_0} \to \mathbb{R}$$

such that $g(W_1, \ldots, W_L) = f(W_L \ldots W_1)$*. If* $\min(d_1, \ldots, d_{L-1}) \geq \min(d_0, d_L)$*, then at any local minimiser* $\left(\hat{W}_1, \ldots, \hat{W}_L\right) \in \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}}$ *of* g *the optimality condition*

$$\nabla f\left(\hat{W}\right) = 0, \qquad \hat{W} := \hat{W}_L \ldots \hat{W}_1$$

is satisfied.

We can interpret this theorem within the optimisation Parametrisation Framework established in Background: Let $\mathcal{E} = \mathbb{R}^{d_L \times d_0}$ be the linear space and $\mathcal{X} = \mathcal{E}$ our search space. Let $\mathcal{M} = \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}}$ be a smooth manifold and we define

$$\phi : \mathcal{M} \to \mathcal{E}, \qquad \phi(W_1, \ldots, W_L) := W_L \ldots W_1 \tag{Z}$$

which trivially satisfies $\operatorname{im}(\phi) = \mathcal{X}$ and is smooth. Finally, let $f : \mathcal{X} \to \mathbb{R}$ be a differentiable objective function and $g = f \circ \phi$. We can now consider two optimisation problems:

$$\min_{W \in \mathcal{X}} \ f(W) \tag{A}$$

$$\min_{\{W_i\}_{i \in [L]} \in \mathcal{M}} \ g(W_1, \ldots, W_L) \tag{B}$$

Recalling Proposition 2.1, we can interpret Theorem 3.2 using this framework to infer the following: Given optimisation problems (A) and (B), the smooth lift ϕ satisfies the “local $\implies$ 1” property if $\min(d_1, \ldots, d_{L-1}) \geq \min(d_0, d_L)$. Note that we make no assumptions on f, in particular we do not assume that f is a loss function $y \mapsto \ell(y, \hat{y})$ defined using a deep linear network.

Given this interpretation, we can relate Theorem 3.2 to the following proposition in ‘The Effect of Smooth Parametrizations on Nonconvex Optimization Landscapes’ by E. Levin et al.[11]:

Proposition 2.2 ([11, Prop 3.6])**.** *The lift* $\psi(L, R) = LR^\top$ *from* $\mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$ *to* $\mathbb{R}^{m \times n}_{\leq r}$*, with* $r < \min(m, n)$*, satisfies:*

(i) “local $\implies$ *local” at* (L, R) *if and only if* $\operatorname{rank}(L) = \operatorname{rank}(R) = \operatorname{rank}\left(LR^\top\right)$*;*

(ii) “1 $\implies$ *1” at* (L, R) *if and only if* $\operatorname{rank}(L) = \operatorname{rank}(R) = r$*; and,*

(iii) “2 $\implies$ *1” everywhere on* $\mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$*.*

This propositions removes the condition on the internal dimensions but also requires that $L = 2$, so it is neither stronger nor weaker than Theorem 3.2. However, together these two results suggest that the matrix multiplication lift ϕ may satisfy various of the Desirable Properties of Lifts [11, Def 2.7] in more general settings. We seek to establish further results in pursuit of resolving this question.

3. Results

We present a variety of results concerning the matrix multiplication map ϕ. They are split by four parts: Sufficient Conditions for "Local Implies One", Counter Examples for "Local Implies One", Sufficient and Necessary Conditions for Local Implies One, and finally "Local Implies Local"

3.1. Sufficient Conditions for "Local Implies One"

In this part, we prove new sufficient conditions on $(W_1, \ldots, W_L)$ such that ϕ, as defined in (Z), is "local $\implies$ 1". This result requires some intermediate lemmas, the first of which is a simple calculation:

Lemma 3.1 ([9, Lemma 1]). *The partial derivatives of g are given by*

$$\begin{aligned}
\nabla_{W_1} g(W_1, \ldots, W_L) &= {W_{2,+}}^\top \nabla f(W) \\
\nabla_{W_k} g(W_1, \ldots, W_L) &= {W_{k+1,+}}^\top \nabla f(W) {W_{k-1,-}}^\top, && \forall k \in \{2, \ldots, L-1\} \\
\nabla_{W_L} g(W_1, \ldots, W_L) &= \nabla f(W) {W_{L-1,-}}^\top
\end{aligned}$$

where $W := W_L \ldots W_1$, $W_{i,+} := W_L \ldots W_i$ and $W_{i,-} = W_i \ldots W_1$.

The next lemma first requires some notation:

$$R^{m\times n}_{\le r} := \left\{ M \in \mathbb{R}^{m\times n} : \operatorname{rank}(M) \le r \right\}$$

for some $r \le \min(m, n)$. Generally, sets of form $R^{m\times n}_{\le r}$ are called *Low-Rank Matrices*.

Lemma 3.2 ([6, 2.1.1]). *Consider the optimisation problem*

$$\min_{X \in \mathbb{R}^{m\times n}_{\le r}} h(X)$$

for some differentiable function $h : \mathbb{R}^{m\times n} \to \mathbb{R}$. Suppose $X \in \mathbb{R}^{m\times n}_{\le r}$ and has (possibly non-unique) singular value decomposition $X = U_X \operatorname{diag}(\sigma_1, \ldots, \sigma_r) {V_X}^\top$, with $\sigma \ge 1 \ldots \ge \sigma_r$. Then X is a first-order stationary point of (1) if and only if

$$\begin{cases} \nabla h(X)^\top U_X = 0 \ \wedge \ \nabla h(X) V_X = 0, & \operatorname{rank}(X) = r \\ \nabla h(X) = 0, & \operatorname{rank}(X) < r \end{cases}$$

This lemma has the following corollary:

Corollary 3.1. *Given the same setup as Lemma 3.2, X is a first-order stationary point of (1) if and only if*

$$\begin{cases} \nabla h(X)^\top X = 0 \ \wedge \ \nabla h(X) X^\top = 0, & \operatorname{rank}(X) = r \\ \nabla h(X) = 0, & \operatorname{rank}(X) < r \end{cases}$$

Proof. If $\operatorname{rank}(X) = r$, then there exists $\hat{U}_X \in \mathbb{R}^{r\times m}$ and $\hat{V}_X \in \mathbb{R}^{r\times n}$ such that

$$U_X \hat{U}_X = I_r = \hat{V}_X V_X$$

Furthermore, all $\sigma_i > 0$ for $i \in [r]$ and so Σ is non-singular. Therefore, by Lemma 3.2, when $\mathrm{rank}\,(X) = 0$, X is a first-order stationary point if and only if:

$$\nabla h\,(X)^\top U_X = 0 \iff \nabla h\,(X)^\top U_X \Sigma V_X{}^\top = 0 \iff \nabla h\,(X)^\top X = 0$$
$$\nabla h\,(X)\, V_X = 0 \iff \nabla h\,(X)\, V_X \Sigma U_X{}^\top = 0 \iff \nabla h\,(X)\, X^\top = 0$$

as claimed. □

With Lemma 3.2 and Corollary 3.1, we prove the following

Proposition 3.1 (Nullity Conditions for ϕ to Satisfy "local $\implies$ 1")**.** *Suppose* $(W_1, \ldots, W_L) \in \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}}$ *and also*

$$W := W_L \ldots W_1, \qquad W_{i,+} := W_L \ldots W_i, \qquad W_{i,-} := W_i \ldots W_1$$

We define $k_1 \in [L]$ *such that*

$$\begin{aligned} \ker\left(\hat{W}_{k,-}\right) &= \{\mathbf{0}_{d_0}\}, \quad \forall k < k_1 \\ \ker\left(\hat{W}_{k,-}\right) &\neq \{\mathbf{0}_{d_0}\}, \quad \forall k \geq k_1 \end{aligned} \tag{1}$$

and $k_2 \in [L]$ *such that*

$$\begin{aligned} \mathrm{rank}\left(\hat{W}_{k,-}\right) &> \mathrm{rank}\left(\hat{W}\right), \quad \forall k < k_2 \\ \mathrm{rank}\left(\hat{W}_{k,-}\right) &= \mathrm{rank}\left(\hat{W}\right), \quad \forall k \geq k_2 \end{aligned} \tag{2}$$

If either

(i) $\ker\left(W_{L-1,-}{}^\top\right)$ *is trivial or* $\mathrm{rank}\left(\hat{W}\right) = d_{\min}$.

(ii) $\ker\left(W_{L-1,-}{}^\top\right)$ *is non-trivial,* $\mathrm{rank}\left(\hat{W}\right) < d_{\min}$*, and* $k_1 = k_2$.

then ϕ *satisfies "local* $\implies$ *1" at* $(W_1, \ldots, W)$.

Proof of Proposition 2.2. Suppose $\left(\hat{W}_1, \ldots, \hat{W}_L\right) \in \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}}$ is a local minimiser of (B). Given that local minima of (B) are also 1-critical, Proposition 2.1 and Lemma 3.1 together imply that:

$$0 = \hat{W}_{2,+}^\top \nabla f\left(\hat{W}\right) \tag{3}$$
$$0 = \hat{W}_{k+1,+}^\top \nabla f\left(\hat{W}\right) \hat{W}_{k-1,-}^\top, \qquad \forall k \in \{2, \ldots, L-1\} \tag{4}$$
$$0 = \nabla f\left(\hat{W}\right) \hat{W}_{L-1,-}^\top \tag{5}$$

where $\hat{W} := \hat{W}_L \ldots \hat{W}_1$, $\hat{W}_{i,+} := \hat{W}_L \ldots \hat{W}_i$ and $\hat{W}_{i,-} := \hat{W}_i \ldots \hat{W}_1$. We wish to show that $\hat{W}$ is stationary for (A).

If $\hat{W}_{L-1,-}^\top$ has trivial kernel, then (4) immediately implies that

$$0 = \nabla f\left(\hat{W}\right) \hat{W}_{L-1,-}^\top \implies \nabla f\left(\hat{W}\right) = 0$$

By Corollary 3.1, we can infer that $\hat{W}$ is a stationary point of f over $\mathbb{R}^{d_L \times d_0}_{\leq d_{\min}}$. As $\mathbb{R}^{d_L \times d_0}_{\leq d_{\min}} \subseteq \mathbb{R}^{d_L \times d_0}$, we can therefore conclude that $\hat{W}$ is a stationary point of (A). So hereafter, we assume that $\hat{W}_{L-1,-}^\top$ has non-trivial kernel, i.e.

$$\ker\left(\hat{W}_{L-1,-}^\top\right) \neq \{\mathbf{0}_{d_0}\} \tag{6}$$

Given that $\hat{W}$ is a low-rank matrix, we split into two cases:

$$(1)\ \ \text{rank}\left(\hat{W}\right) = d_{\min} \qquad\qquad (2)\ \ \text{rank}\left(\hat{W}\right) < d_{\min}$$

We first consider case (1). Equations (2) and (4) imply that

$$\begin{aligned}
&0 = \hat{W}_{2,+}^{\top} \nabla f\left(\hat{W}\right) \implies 0 = \hat{W}_1 \hat{W}_{2,+}^{\top} \nabla f\left(\hat{W}\right) \implies 0 = \nabla f\left(\hat{W}\right)^{\top} \hat{W} \\
&0 = \nabla f\left(\hat{W}\right) \hat{W}_{L-1,-}^{\top} \implies 0 = \nabla f\left(\hat{W}\right) \hat{W}_{L-1,-}^{\top} \hat{W}_L \implies 0 = \nabla f\left(\hat{W}\right) \hat{W}^{\top} = 0
\end{aligned}$$

Therefore, by Corollary 3.1 and the same reasoning as for the trivial kernel case, we can conclude that $\hat{W}$ is a stationary point of (A).

We have now proven that condition (i) implies that ϕ is "local $\implies$ 1". So we now consider condition (ii), which corresponds to case (2) and assumption (6). We first must prove the existence of k_1 and k_2, as defined by (1) and (2). We begin by noting that for any $(W_1, \ldots, W_L) \in \mathbb{R}^{d_1 \times d_0} \times \ldots \times R^{d_L \times d_{L-1}}$ we have

$$\ker\left(W_{1,-}\right) \subseteq \ker\left(W_{2,-}\right) \subseteq \ldots \subseteq \ker\left(W_{L,-}\right) \tag{7}$$
$$\text{rank}\left(W_{1,-}\right) \geq \text{rank}\left(W_{2,-}\right) \geq \ldots \geq \text{rank}\left(W_{L-1,-}\right) \tag{8}$$

where $W_{k,-} = W_k \ldots W_1$. By $\text{rank}\left(\hat{W}\right) < d_{\min}$ and the Rank-Nullity Theorem, we know that $\ker\left(\hat{W}_{L,-}\right) = \ker\left(\hat{W}\right)$ is non-trivial. So by the the chain of inclusions (6), we know that k_1 exists, i.e. there must be a $k_1 \in [L]$ such that

$$\ker\left(\hat{W}_{k,-}\right) = \left\{\mathbf{0}_{d_0}\right\}, \qquad \forall k < k_1 \tag{9}$$
$$\ker\left(\hat{W}_{k,-}\right) \neq \left\{\mathbf{0}_{d_0}\right\}, \qquad \forall k \geq k_1 \tag{10}$$

That is, $\hat{W}_{k_1}$ is the first rank-deficient matrix. Furthermore, by (7), we know that k_2 as defined by (2) exists, i.e. there exists $k_2 \in [L]$ such that

$$\begin{aligned}
&\text{rank}\left(\hat{W}_{k,-}\right) > \text{rank}\left(\hat{W}\right), \qquad \forall k < k_2 \\
&\text{rank}\left(\hat{W}_{k,-}\right) = \text{rank}\left(\hat{W}\right), \qquad \forall k \geq k_2 \qquad (11)
\end{aligned}$$

That is, $\hat{W}_{k_2,-}$ is the first matrix has rank equal to $\hat{W}$. In fact, as $\text{rank}\left(\hat{W}\right) < d_{\min}$, $\hat{W}$, we know that $\hat{W}_{k_2,-}$ is rank-deficient. Therefore, $k_1 \leq k_2$, i.e. $k_2 \in \{k_1, \ldots, L\}$. Additionally, as column rank equals row rank, (11) implies

$$\text{rank}\left(\hat{W}_{k,-}^{\top}\right) = \text{rank}\left(\hat{W}\right), \qquad \forall k \geq k_2$$

Given that $\text{rank}\left(\hat{W}\right) < d_{\min}$, we can apply the Rank-Nullity Theorem to yield

$$\ker\left(\hat{W}_{k,-}^{\top}\right) \neq \left\{\mathbf{0}_{d_k}\right\}, \qquad \forall k \geq k_2 \tag{12}$$

So, for all $k \geq k_2$, there exists unit vector $\hat{\mathbf{u}}_{\mathbf{k}} \in \mathbb{R}^{d_k}$ such that

$$\hat{W}_{k,-}^{\top} \hat{\mathbf{u}}_{\mathbf{k}} = \mathbf{0}_{d_k} \tag{13}$$

Finally, as $\left(\hat{W}_1, \ldots, \hat{W}_L\right)$ is a local minimum, we know that there exists $\varepsilon_0 > 0$ such that

$$g\left(W_1, \ldots, W_L\right) \geq g\left(\hat{W}_1, \ldots, \hat{W}_L\right) = f\left(\hat{W}\right) \tag{14}$$

for all $(W_1, \dots, W_L) \in \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}}$ satisfying

$$\left\| W_k - \hat{W}_k \right\|_{\mathrm{F}} \leq \varepsilon_0 \tag{15}$$

for all $k \in [L]$.

We now claim the following:

Claim 3.1. *Suppose we have unit vectors* $\mathbf{w_k} \in \mathbb{R}^{d_k}$ *and scalars* δ_k *satisfying* $0 \leq \delta_k \leq \frac{\varepsilon_0}{2}$*, for all* $k_2 < k \leq L$*. The tuple of matrices* $\left(\tilde{W}_1, \dots, \tilde{W}_L\right)$ *defined by*

$$\tilde{W}_k := \begin{cases} \hat{W}_k, & 1 \leq k \leq k_2 \\ \hat{w}_k + \delta_k \mathbf{w_k} \hat{\mathbf{u}}_{\mathbf{k-1}}^\top, & \textit{otherwise} \end{cases} \tag{16}$$

satisfies

$$\left\| \tilde{W}_k - \hat{W}_k \right\|_{\mathrm{F}} \leq \frac{\varepsilon_0}{2} \tag{17}$$

$$\tilde{W} = \hat{W} \tag{18}$$

where $\tilde{W} := \tilde{W}_L \dots \tilde{W}_1$*. Furthermore,* $\left(\tilde{W}_1, \dots, \tilde{W}_L\right)$ *is a local minimiser of* (B)*.*

Proof. We first consider inequality (17). For $k \leq k_2$, we have that

$$\left\| \tilde{W}_k - \hat{W}_k \right\|_{\mathrm{F}} = \left\| \hat{W}_k - \hat{W}_k \right\|_{\mathrm{F}} = 0$$

by definition of $\tilde{W}_k$ for $k \leq k_2$. For $k_2 < k \leq L$, consider:

$$\left\| \tilde{W}_K - \hat{W}_k \right\|_{\mathrm{F}} = \delta_k \left\| \mathbf{w_k} \hat{\mathbf{u}}_{\mathbf{k-1}}^\top \right\|_{\mathrm{F}} = \delta \left\| \mathbf{w_k} \right\|_2 \left\| \hat{\mathbf{u}}_{\mathbf{k-1}} \right\|_2 = \delta \cdot 1 \cdot 1 \leq \frac{\varepsilon_0}{2}$$

where the first equality follows from definition of $\tilde{W}_k$ (16), the final equality follows from $\mathbf{w_k}$ and $\hat{\mathbf{u}}_{\mathbf{k}}$ being unit vectors, and the inequality follows from the condition on δ_k. So inequality (17) holds for the $\tilde{W}_k$.

We now consider condition (18). Let $\tilde{W}_{k,-} := \tilde{W}_k \dots \tilde{W}_1$. We will prove that

$$\tilde{W}_{k,-} = \hat{W}_{k,-} \tag{19}$$

by induction on k. For $1 \leq k \leq k_2$, (19) is immediately true by definition of $\tilde{W}_k$ (16). So now suppose (19) holds for some $k \geq k_2$. Consider:

$$\begin{aligned} \tilde{W}_{k+1,-} &= \tilde{W}_{k+1} \tilde{W}_{k,-} \\ &= \left(\hat{W}_{k+1} + \delta_{k+1} \mathbf{w_{k+1}} \hat{\mathbf{u}}_{\mathbf{k}}^\top \right) \hat{W}_{k,-} \qquad (20) \\ &= \hat{W}_{k+1,-} + \delta_{k+1} \mathbf{w_{k+1}} \hat{\mathbf{u}}_{\mathbf{k}}^\top \hat{W}_{k,-} \\ &= \hat{W}_{k+1,-} + \delta_{k+1} \mathbf{w_{k+1}} \mathbf{0}_{d_k}{}^\top \qquad (21) \\ &= \hat{W}_{k+1,-} \end{aligned}$$

where (20) follows from the definition of $\tilde{W}_k$(16) and the induction hypothesis and (21) follows from (13). Therefore, by induction, the $\tilde{W}_k$ also satisfy condition (18).

We now prove the last part of lemma. That is, we wish to show that $\left(\tilde{W}_1, \ldots, \tilde{W}_k\right)$ is a local minimiser of (B). Consider tuple of matrices $(W_1, \ldots, W_L) \in \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}}$ satisfying

$$\left\|W_k - \tilde{W}_k\right\|_{\mathrm{F}} \leq \frac{\varepsilon_0}{2}$$

for all $k \in [L]$. Condition (17) implies that

$$\left\|W_k - \hat{W}_k\right\|_{\mathrm{F}} \leq \left\|W_k - \tilde{W}_k\right\|_{\mathrm{F}} + \left\|\tilde{W}_k - \hat{W}_k\right\|_{\mathrm{F}} \leq \varepsilon$$

Furthermore, (14) and condition (18) together yield

$$g\left(W_1, \ldots, W_L\right) \geq f\left(\hat{W}\right) = f\left(\tilde{W}\right) = g\left(\tilde{W}_1, \ldots, \tilde{W}_L\right)$$

Therefore, $\left(\tilde{W}_1, \ldots, \tilde{W}_L\right)$ is a local minimiser of (B) in an ε neighbourhood around each component. □

So any $\left(\tilde{W}_1, \ldots \tilde{W}_L\right)$ satisfying (16) for some unit vectors $\mathbf{w_k} \in \mathbb{R}^{d_k}$ and scalars δ_k such that $0 \leq \delta_k \leq \frac{\varepsilon_0}{2}$, for all $k_2 < k \leq L$, is a local minimiser of (B). Consequentially, again using that local minima of (B) are also 1-critical and then applying Proposition 2.1 and Lemma 3.1, we can infer that

$$0 = \tilde{W}_{2,+}^{\top} \nabla f\left(\tilde{W}\right) \tag{22}$$

$$0 = \tilde{W}_{k+1,+}^{\top} \nabla f\left(\tilde{W}\right) \tilde{W}_{k-1,-}^{\top}, \qquad \forall k \in \{2, \ldots, L-1\} \tag{23}$$

$$0 = \nabla f\left(\tilde{W}\right) \tilde{W}_{L-1,-}^{\top} \tag{24}$$

holds for all unit vectors $\mathbf{w_k} \in \mathbb{R}^{d_k}$ and scalars δ_k satisfying $0 \leq \delta_k \leq \frac{\varepsilon_0}{2}$, for all $k_2 < k \leq L$.

Now make another claim:

Claim 3.2. *Suppose* $\left(\tilde{W}_1, \ldots \tilde{W}_L\right)$ *satisfies* (9) *for some unit vectors* $\mathbf{w_k} \in \mathbb{R}^{d_k}$ *and scalars* δ_k *such that* $0 \leq \delta_k \leq \frac{\varepsilon_0}{2}$. *Then*

$$\nabla f\left(\hat{W}\right)^{\top} \tilde{W}_{k_2+1,+} = 0 \tag{25}$$

for k_2 *as defined above.*

Proof. Firstly, $\tilde{W}_{k_2,-} = \hat{W}_{k_2,-}$ and $\tilde{W} = \hat{W}$ hold by definition of $\tilde{W}_k$ (16) and condition (18) in Claim 3.1. We will use these equalities often and so will not explicitly note them.

We split into cases for k_2. For $k_2 = 1$, the transpose of (22), we can immediately conclude (25).

We now consider $1 < k_2 < L$. The transpose of (23) yields

$$0 = \tilde{W}_{k_2-1,-} \nabla f\left(\tilde{W}\right)^{\top} \tilde{W}_{k_2+1,+} \tag{26}$$

We must show that $\tilde{W}_{k_2-1,-}$ has trivial kernel. By (9) and (10), we know that this is only true if $k_2 - 1 < k_1$. However, we already noted that $k_1 \leq k_2$. So $k_2 - 1 < k_1$ if and only if $k_1 = k_2$. We now apply condition (ii): given the case we are in, $\ker\left(W_{L-1,-}{}^{\top}\right)$

being non-trivial and rank $\left(\hat{W}\right) < d_{\min}$, we can assume $k_1 = k_2$. So $\tilde{W}_{k_2-1,-}$ has trivial kernel and so can conclude (25).

Finally, we do not need to consider the case where $k_2 = L$, as (6) and (12) imply that $k_2 \leq L - 1$. □

We can now use Claim 3.2 to complete the proof. Specifically, we know that (25) holds for all choices of unit vectors $\mathbf{w_k} \in \mathbb{R}^{d_k}$ and scalars δ_k such that $0 \leq \delta_k \leq \frac{\varepsilon_0}{2}$. So set $\delta_{k_2+1} = 0$, thereby setting $\tilde{W}_{k_2+1} = \hat{W}_{k_2+1}$. Making this substitution into (25) yields

$$\nabla f\left(\hat{W}\right)^\top \tilde{W}_{k_2+2,+}\hat{W}_{k_2+1} = 0 \tag{27}$$

We can subtract (26) from (25) and then using the definition of $\tilde{W}_k$ (16), we get

$$0 = \nabla f\left(\hat{W}\right)^\top \tilde{W}_{k_2+2,+}\left(\hat{W}_{k_2+1} - \tilde{W}_{k_2+1}\right) = \nabla f\left(\hat{W}\right)^\top \tilde{W}_{k_2+2,+}\left(\delta_{k_2+1}\mathbf{w}_{k_2+1}\hat{\mathbf{u}}_{\mathbf{k_2}}^\top\right)$$

for all unit vectors $\mathbf{w}_{k_2+1}$ and $0 \leq \delta_{k_2+1} \leq \frac{\varepsilon_0}{2}$. Choosing $\delta_{k_2+1} \neq 0$, we can divide by it and also right-multiply by $\hat{\mathbf{u}}_{\mathbf{k_2}}$, which we recall is also a unit vector and so satisfies $\hat{\mathbf{u}}_{\mathbf{k_2}}^\top\hat{\mathbf{u}}_{\mathbf{k_2}} = 1$, to infer that

$$\begin{aligned} 0 &= \frac{1}{\delta_{k_2+1}}\nabla f\left(\hat{W}\right)^\top \tilde{W}_{k_2+2,+}\left(\delta_{k_2+1}\mathbf{w}_{k_2+1}\hat{\mathbf{u}}_{\mathbf{k_2}}^\top\right)\hat{\mathbf{u}}_{\mathbf{k_2}} \\ &= \nabla f\left(\hat{W}\right)^\top \tilde{W}_{k_2+2,+}\mathbf{w}_{k_2+1} \end{aligned} \tag{28}$$

for all unit vectors $\mathbf{w}_{k_2+1}$. As (27) holds for all unit vectors $\mathbf{w}_{k_2+1}$, we can infer that

$$0 = \nabla f\left(\hat{W}\right)^\top \tilde{W}_{k_2+2,+}$$

for all choices of unit vectors $\mathbf{w_k} \in \mathbb{R}^{d_k}$ and scalars δ_k such that $0 \leq \delta_k \leq \frac{\varepsilon_0}{2}$, for $k \geq k_2 + 2$. Notice (28) has eliminated the final $\tilde{W}_{k_2+1}$ from (25), without changing setting any $\mathbf{w_k}$ or δ_k for $k \geq k_2 + 2$. Therefore, we can repeat this process until we get

$$\nabla f\left(\hat{W}\right) = 0$$

Finally, by Corollary 3.1 and the same reasoning as the trivial kernel case, we can infer that $\hat{W}$ is a stationary point of (A).

Overall, we have now shown that $\hat{W}$ is a stationary point of (A) for all cases and so we can conclude that ϕ is "local $\implies$ 1". □

3.2. Counter Examples for "Local Implies One"

The conditions of Proposition 3.1 outline a situation where ϕ is "local $\implies$ 1". We construct counter-example that break these conditions to determine further properties of ϕ.

We define the following matrices and calculate their respective ranks and nullities:

$$\begin{aligned}
W_1 &:= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, & \operatorname{rank}(W_1) = 2, && \operatorname{nullity}(W_1) = 1 \\
W_2 &:= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, & \operatorname{rank}(W_2) = 1, && \operatorname{nullity}(W_1) = 1 \\
W_3 &:= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, & \operatorname{rank}(W_3) = 2, && \operatorname{nullity}(W_1) = 0 \\
W := W_3 W_2 W_1 &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, & \operatorname{rank}(W) = 1, && \operatorname{nullity}(W) = 2 \\
W_2 W_1 &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, & \operatorname{rank}(W_2 W_1) = 1, && \operatorname{nullity}(W_2 W_1) = 2 \\
W_3 W_2 &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, & \operatorname{rank}(W_3 W_2) = 1, && \operatorname{nullity}(W_3 W_2) = 1
\end{aligned} \tag{29}$$

Note that $(W_2 W_1)^\top = W_3 W_2$. We now define the function

$$f : \mathbb{R}^{3\times 3} \to \mathbb{R}, \qquad X \mapsto \langle A, X \rangle_{\mathrm{F}} + \frac{c}{2} \| X - W \|_{\mathrm{F}}^2 \tag{30}$$

for some $A \in \mathbb{R}^{3\times 3}$ and $c \in \mathbb{R}$. We now define g according to (A) and (B):

$$\begin{aligned}
& g : \mathbb{R}^{3\times 2} \times \mathbb{R}^{2\times 2} \times \mathbb{R}^{2\times 3} \to \mathbb{R} \\
& \qquad (X_1, X_2, X_3) \mapsto (f \circ \phi)(X_1, X_2, X_3) = \langle A, X_3 X_2 X_1 \rangle_{\mathrm{F}} + \frac{c}{2} \| X_3 X_2 X_1 - W \|_{\mathrm{F}}^2
\end{aligned} \tag{31}$$

As illustrated by calculations (29), (W_1, W_2, W_3) do not satisfy the conditions outlined by Proposition 3.1. So ϕ may not be "local $\implies$ 1" at (W_1, W_2, W_3) for g and f as defined in (30) and (31), respectively.

We know that if (W_1, W_2, W_3) is a local minimum for g, then it is also 1- and 2-critical for g. So first we wish to find an A and c such that (W_1, W_2, W_3) is 2-critical for g, but not stationary for f. So we calculate:

$$\begin{aligned}
\nabla f(X) &= A + c(X - W) \\
\nabla g(X_1, X_2, X_3) &= \left(A {X_1}^\top {X_2}^\top, {X_3}^\top A {X_1}^\top, {X_2}^\top {X_3}^\top A \right) \\
&\quad + \frac{c}{2} \left(\frac{\mathrm{d} \| X_1 X_2 X_3 - W \|_{\mathrm{F}}^2}{\mathrm{d} X_1}, \frac{\mathrm{d} \| X_1 X_2 X_3 - W \|_{\mathrm{F}}^2}{\mathrm{d} X_2}, \frac{\mathrm{d} \| X_1 X_2 X_3 - W \|_{\mathrm{F}}^2}{\mathrm{d} X_3} \right)
\end{aligned}$$

We do not calculate the partial derivatives of $\| X_1 X_2 X_3 - W \|_{\mathrm{F}}^2$, or later their the higher-order partials, as we already know that they are zero at $(X_1, X_2, X_3) = (W_1, W_2, W_3)$. Substituting in our previous calculations of the matrices W_1, W_2, W_3, and their compositions (29), we have

$$\nabla g(W_1, W_2, W_3) = \left(\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} a_{1,1} & a_{2,1} \\ a_{2,1} & a_{2,2} \end{pmatrix}, \begin{pmatrix} a_{1,1} & 0 \\ a_{2,1} & 0 \\ a_{3,1} & 0 \end{pmatrix} \right)$$

where $a_{i,j}$ is the (i,j)-th entry of A. Therefore, we conclude that

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & x \\ 0 & y & z \end{pmatrix} \implies \nabla g(W_1, W_2, W_3) = 0 \ \wedge \nabla f(W) \neq 0$$

for all $x, y, z \in \mathbb{R}$. So ϕ is not "1 $\implies$ 1". Furthermore, we computed $\nabla^2 g\left(W_1, W_2, W_3\right)$ and $\nabla^3 g\left(W_1, W_2, W_3\right)$ for the same A using Mathematica and found that both also equal 0: see Appendix A for details of the Mathematica computations. Therefore, W_1, W_2, W_3, A, f, and g are a counter-example to ϕ satisfying "3 $\implies$ 1". So we can conclude that:

Theorem 3.1 (ϕ is not "3 $\implies$ 1")**.** *The smooth lift $\phi : \mathcal{M} \to \mathcal{E}$ does not satisfy "3 $\implies$ 1" and therefore also does not satisfy "2 $\implies$ 1".*

Note that as all derivatives of $\|X_1X_2X_3 - W\|_{\mathrm{F}}^2$ are zero at $(X_1, X_2, X_3) = (W_1, W_2, W_3)$, we did not need to set c. We now use this free variable to make the following claim:

Claim 3.3. *Suppose*

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & x \\ 0 & y & z \end{pmatrix} \tag{32}$$

for some $x, y, z \in \mathbb{R}$. Then there exists $c > 0$ such that (W_1, W_2, W_3), as defined in (29), is a local minimum of g.

Proof. Firstly, we calculate

$$\begin{aligned} g\left(W_1, W_2, W_3\right) &= \langle A, W_3W_2W_1\rangle_{\mathrm{F}} + \frac{c}{2}\left\|W_3W_2W_1 - W\right\|_{\mathrm{F}}^2 \\ &= \langle A, W\rangle_{\mathrm{F}} + \frac{c}{2}\left\|W - W\right\|_{\mathrm{F}}^2 \\ &= \operatorname{Tr}\left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & x \\ 0 & y & z \end{pmatrix}^{\top} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\right) + 0 \qquad (33) \\ &= 0 \qquad (34) \end{aligned}$$

where (33) follows from the assumption on A (32) and our calculation of W (29). Note that this calculation also shows that

$$g\left(X_1, X_2, X_3\right) = 0 \tag{35}$$

if $X_3X_2X_1 = W$. Therefore, we have that

$$\begin{aligned} \lim_{c\to\infty} g\left(X_1, X_2, X_3\right) &= \lim_{c\to\infty} \langle A, X_3X_2X_1\rangle_{\mathrm{F}} + \frac{c}{2}\left\|X_3X_2X_1 - W\right\|_{\mathrm{F}}^2 \\ &= \langle A, X_3X_2X_1\rangle_{\mathrm{F}} + \frac{1}{2}\left\|X_3X_2X_1 - W\right\|_{\mathrm{F}}^2 \left(\lim_{c\to\infty} c\right) \\ &= \begin{cases} \infty, & X_3X_2X_1 \neq W \\ 0, & \text{otherwise} \end{cases} \qquad (36) \end{aligned}$$

where (36) follows from (35). Additionally, we note that the map

$$c \mapsto \langle A, X_3X_2X_1\rangle_{\mathrm{F}} + \frac{c}{2}\left\|X_3X_2X_1 - W\right\|_{\mathrm{F}}^2 \tag{37}$$

is continuous and strictly increasing for all (X_1, X_2, X_3) such that $X_3X_2X_1 \neq W$.

Let $\varepsilon > 0$. We now define

$$\mathcal{C} = \left\{\left(X_1, X_2, X_3\right) \in B_{\mathrm{F}}\left(\left(W_1, W_2, W_3\right), \varepsilon\right) : X_3X_2X_1 \neq W\right\}$$

where $B_{\mathrm{F}}\left(\left(W_1, W_2, W_3\right), \varepsilon\right)$ denotes the open ball around $\left(W_1, W_2, W_3\right)$ of radius ε as defined by the Frobenius norm. From (36) and the properties of the map (37), we can infer that there must exist a $c' \in \mathbb{R}_>$ such that

$$g\left(X_1, X_2, X_3\right) > 0 \tag{38}$$

for all $\left(X_1, X_2, X_3\right) \in \mathcal{C}$. Therefore, for this c', we have that

$$g\left(X_1, X_2, X_3\right) \geq 0 = g\left(W_1, W_2, W_3\right)$$

for all $\left(X_1, X_2, X_3\right) \in B_{\mathrm{F}}\left(\left(W_1, W_2, W_3\right), \varepsilon\right)$, where the inequality follows from (35) and (38) and the equality follows from (34). Therefore, $\left(W_1, W_2, W_3\right)$ is a local minimum of g given $c = c'$. □

Given this, we can conclude the following:

Theorem 3.2 (ϕ is not "local $\implies$ local"). *The smooth lift $\phi : \mathcal{M} \to \mathcal{E}$ as defined in (Z) does not satisfy "local $\implies$ 1".*

Additionally, given that the counter-example used to prove this theorem does not satisfy the conditions of Proposition 3.1, we can conclude that those conditions are strongest in the sense that breaking them may lead to ϕ not being "local $\implies$ 1".

3.3. Sufficient and Necessary Conditions for Local Implies One

While Proposition 3.1 outlines sufficient conditions for ϕ to be "local $\implies$ 1", Theorem 3.2 demonstrates that ϕ is not "local $\implies$ 1" everywhere. This motivated us to find the following sufficient and necessary conditions:

Theorem 3.3 (Rank Conditions for ϕ to Satisfy "local $\implies$ 1"). *Suppose $d_{\min} < d_0, d_L$ and let $k \in \operatorname*{arg\,min}_{i \in [L-1]} d_i$, i.e. $d_k = d_{\min}$. Then ϕ, as defined in (Z), satisfies "local $\implies$ 1" at $\left(\hat{W}_1, \ldots, \hat{W}\right)$ if and only if*

$$\operatorname{rank}\left(\hat{W}_k \ldots \hat{W}_1\right) = \operatorname{rank}\left(\hat{W}_L \ldots \hat{W}_{k+1}\right) = d_{\min} \tag{39}$$

Proof. Suppose $\left(\hat{W}_1, \ldots, \hat{W}_L\right) \in \mathbb{R}^{d_1 \times d_0} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}}$ is a local minimiser of (B). We want to show that $\phi\left(\hat{W}_1, \ldots, \hat{W}_L\right) = \hat{W}_L \ldots \hat{W}_1$ is 1-critical for (B) if and only if $\left(\hat{W}_1, \ldots, \hat{W}_L\right)$ satisfy condition (39).

Firstly, we use notation

$$\hat{W}_{l,-} := \hat{W}_l \ldots \hat{W}_1, \qquad \hat{W}_{l,+} := \hat{W}_L \ldots \hat{W}_l$$

for all $l \in [L]$. Given the function $\psi(A, B) = AB$, we can get an alternative representation of g from (B):

$$\begin{aligned} g\left(W_1, \ldots, W_L\right) &= f\left(W_L \ldots W_1\right) \\ &= f\left(\left(W_L \ldots W_{k+1}\right)\left(W_k \ldots W_1\right)\right) \\ &= f\left(\psi\left(\phi\left(W_{k+1}, \ldots, W_L\right), \phi\left(W_1, \ldots, W_k\right)\right)\right) \end{aligned}$$

Given this, we define functions:

$$\begin{aligned} f_- &: \mathbb{R}^{d_k \times d_0} \to \mathbb{R} \\ f_-\left(W_{k,-}\right) &:= f\left(\psi\left(\phi\left(\hat{W}_{k+1}, \ldots, \hat{W}_L\right), W_{k,-}\right)\right) \end{aligned}$$

$$\begin{aligned} g_- : \mathbb{R}^{d_0 \times d_1} \times \ldots \times \mathbb{R}^{d_k \times d_{k-1}} &\to \mathbb{R} \\ g_-\left(W_1, \ldots, W_k\right) &:= \left(f_- \circ \phi\right)\left(W_1, \ldots, W_k\right) \\ &= f\left(\psi\left(\phi\left(\hat{W}_{k+1}, \ldots, \hat{W}_L\right), \phi\left(W_1, \ldots, W_k\right)\right)\right) \end{aligned}$$

$$\begin{aligned} f_+ &: \mathbb{R}^{d_L \times d_k} \to \mathbb{R} \\ f_+\left(W_{k+1,+}\right) &:= f\left(\psi\left(W_{k+1,+}, \phi\left(\hat{W}_1, \ldots, \hat{W}_k\right)\right)\right) \end{aligned}$$

$$\begin{aligned} g_+ : \mathbb{R}^{d_{k+1} \times d_k} \times \ldots \times \mathbb{R}^{d_L \times d_{L-1}} &\to \mathbb{R} \\ g_+\left(W_{k+1}, \ldots, W_L\right) &:= \left(f_+ \circ \phi\right)\left(W_1, \ldots, W_k\right) \\ &= f\left(\psi\left(\phi\left(W_{k+1}, \ldots, W_L\right), \phi\left(\hat{W}_1, \ldots, \hat{W}_k\right)\right)\right) \end{aligned}$$

i.e. f_- and g_- fix the matrix above W_k, while f_+ and g_+ fix the matrices below W_{k+1}. As $\left(\hat{W}_1, \ldots, \hat{W}_L\right)$ is a local minimiser of (B), we can infer that $\left(\hat{W}_1, \ldots, \hat{W}_k\right)$ and $\left(\hat{W}_{k+1}, \ldots, \hat{W}_L\right)$ for g_- and g_+, respectively. Given the definition of k, we know that both $\left(\hat{W}_1, \ldots, \hat{W}_k\right)$ and $\left(\hat{W}_{k+1}, \ldots, \hat{W}_L\right)$ satisfy the conditions for Theorem 2.3, i.e.

$$\begin{aligned} d_k &= \min\left(d_0, d_k\right) = \min\left(d_0, \ldots, d_k\right) \\ d_k &= \min\left(d_k, d_L\right) = \min\left(d_k, \ldots, d_L\right) \end{aligned}$$

Therefore, we can infer that $\hat{W}_{k,-}$ and $\hat{W}_{k+1,+}$ are 1-critical points for f_- and f_+, respectively. Note that Theorem 2.3 on all sequences of matrices with higher hidden dimensions, and so, by our assumption that $d_{\min} < d_0, d_L$ and our construction of k, this inference is reversible.

We now define

$$\begin{aligned} f_0 : \mathbb{R}^{d_0 \times d_k} \times \mathbb{R}^{d_L \times d_k} &\to \mathbb{R} \\ f_0\left(W_{k,-}, W_{k+1,+}\right) &:= \left(f \circ \psi\right)\left(W_{k+1,+}, W_{k,-}\right) \\ &= f\left(W_{k+1,+} W_{k,-}\right) \end{aligned}$$

Given that

$$f_0\left(\hat{W}_{k,-}, \hat{W}_{k+1,+}\right) = f_-\left(\hat{W}_{k,-}\right) = f_+\left(\hat{W}_{k+1,+}\right)$$

we can calculate

$$\nabla f_0\left(\hat{W}_{k,-}, \hat{W}_{k+1,+}\right) = \left(\nabla f_-\left(\hat{W}_{k,-}\right), \nabla f_+\left(\hat{W}_{k+1,+}\right)\right)$$

So, by Proposition 2.1, $\left(\hat{W}_{k,-}, \hat{W}_{k+1,+}\right)$ is a 1-critical point of f_0 if and only if $\hat{W}_{k,-}$ and $\hat{W}_{k+1,+}$ are 1-critical for f_- and f_+, respectively.

Finally, we can apply Proposition 2.2 (ii) to conclude that

$$\psi\left(\hat{W}_{k+1,+}, \hat{W}_{k,-}\right) = \hat{W}_L \dots \hat{W}_1$$

is 1-critical for (B) if and only if condition 39 holds. □

Combining this result with Theorem 2.3, we can exactly describe when ϕ is "local $\implies$ 1":

Theorem 3.4 (ϕ and "Local $\implies$ 1"). *ϕ, as defined in (Z), satisfies "local $\implies$ 1" at $\left(\hat{W}_1, \dots, \hat{W}\right)$ if and only if either*

(i) $\min(d_0, \dots, d_L) \in \{d_0, d_L\}$*; or*

(ii) $\operatorname{rank}\left(\hat{W}_k \dots \hat{W}_1\right) = \operatorname{rank}\left(\hat{W}_L \dots \hat{W}_{k+1}\right) = d_k$ *where* $k \in \underset{i\in[L-1]}{\arg\min}\ d_i$.

Proof. Follows by splitting into the two cases $d_{\text{min}} = \min(d_0, d_L)$ and $d_{\text{min}} < \min(d_0, d_L)$ and applying Theorem 2.3 and Theorem 3.3 to each case, respectively. □

3.4. "Local Implies Local"

Given we have a complete description for when ϕ is "local $\implies$ 1", we now turn our attention to "local $\implies$ local". As such, we present one final result outlining sufficient and necessary conditions for ϕ being "local $\implies$ local". This final result leverages the following theorem from 'Where is matrix multiplication locally open' by E. Behrends[2]:

Theorem 3.5 (Characterising Openness of Matrix Multiplication [2, Thm 2.5]). *Let X, Y, and Z be real or complex normed spaces with dimensions n, k, and m, respectively. $S_0 \in L(Y, Z)$ and $T_0 \in L(X, Y)$ are linear operators, we denote s and t the rank of S_0 and T_0, respectively. The number t is written as $t = t_1 + t_2$, where t_2 is the dimension of* $\operatorname{im}(T_0) \cap \ker(S_0)$. *The following conditions are equivalent:*

(i) Multiplication is locally open at (S_0, T_0);

(ii) $t_2 \leq k - m$, or $n - t_1 \leq k - s$.

(iii) There exists $T \in L(X, Y)$ with $S_0T = 0$ such that $T_0 + \alpha T$ is one-to-one for every $\alpha \neq 0$, or there exists $S \in L(Y, Z)$ such that $ST_0 = 0$ and $S_0 + \alpha S$ is onto for every $\alpha \neq 0$.

Combining this theorem with Theorem 2.1, we get the following:

Corollary 3.2 (Conditions for ϕ to Satisfy "Local $\implies$ Local"). *The smooth lift $\phi : \mathcal{M} \to \mathcal{E}$ as defined in (Z) is "local $\implies$ local" if and only if the following equivalent conditions are satisfied*

(i) Either

$$\dim\left(\operatorname{im}(W_{k,-}) \cap \ker(W_{k+1})\right) \leq d_k - d_{k+1}$$

or

$$d_0 - \operatorname{rank}(W_{k,-}) - \dim\left(\operatorname{im}(W_{k,-}) \cap \ker(W_{k+1})\right) \leq d_k - \operatorname{rank}(W_{k+1})$$

(ii) There exists $V \in \mathbb{R}^{d_k \times d_0}$ with $W_{k+1}V = 0$ such that $W_{k,-} + \alpha V$ is one-to-one for every $\alpha \neq 0$, or there exists $U \in \mathbb{R}^{d_{k+1} \times d_k}$ such that $UW_{k,-} = 0$ and $W_{k+1} + \alpha U$ is onto for every $\alpha \neq 0$.

for all $k \in [L-1]$, where $W_{k,-} := W_k \dots W_1$.

Proof. Let $(W_1, \dots, W_L) \in \mathbb{R}^{d_1 \times d_0} \times \dots \times \mathbb{R}^{d_L \times d_{L-1}}$. By Theorem 2.1, ϕ satisfies "local $\implies$ local" at $(W_1, \dots, W_L)$ if and only if ϕ is open at $(W_1, \dots, W_L)$. ϕ is open at this point if each matrix multiplication in the sequence

$$W_1 \xrightarrow{W_2 \times} W_2 W_1 \xrightarrow{W_3 \times} \dots \xrightarrow{W_L \times} W_L \dots W_1$$

is open. Theorem 3.5 states that these matrix multiplications are open if and only if either

(i) Either

$$\dim\left(\operatorname{im}\left(W_{k,-}\right) \cap \ker\left(W_{k+1}\right)\right) \leq d_k - d_{k+1}$$

or

$$d_0 - \operatorname{rank}\left(W_{k,-}\right) - \dim\left(\operatorname{im}\left(W_{k,-}\right) \cap \ker\left(W_{k+1}\right)\right) \leq d_k - \operatorname{rank}\left(W_{k+1}\right)$$

(ii) There exists $V \in \mathbb{R}^{d_k \times d_0}$ with $W_{k+1} V = 0$ such that $W_{k,-} + \alpha V$ is one-to-one for every $\alpha \neq 0$, or there exists $U \in \mathbb{R}^{d_{k+1} \times d_k}$ such that $U W_{k,-} = 0$ and $W_{k+1} + \alpha U$ is onto for every $\alpha \neq 0$.

for all $k \in [L-1]$. □

Finally, we note that $\operatorname{rank}(AB) = \operatorname{rank}(B) - \dim(\operatorname{im}(B) \cap \ker(A))$. In some cases, this may make Corollary 3.2 easier to apply.

4. Future Work

One possible avenue of further investigation is finding the conditions for ϕ to satisfy "$k \implies 1$". We have outlined sufficient and necessary conditions for "local $\implies$ 1" and "local $\implies$ local". However, in [9] Laurent and von Brecht construct counter-examples that satisfy the conditions of Theorem 2.2 but do not exhibit "1 $\implies$ 1" and so remark that stronger conditions are needed. Therefore, the conditions for theses desirable properties is unknown and so further investigation could yield interesting results.

Another potentially worthwhile avenue of investigation is generalising Theorem 2.1. Again, in [9] Laurent and von Brecht construct counter examples that demonstrate that f must be differentiable, not merely globally Lipschitz, to ensure the absence of spurious local minima. Furthermore, they demonstrate that restrictions on the constraint set may also introduce spurious local minima. This problem could potentially be approached using the Restricted Isometry Property, which has been shown to eliminate spurious local minima in some settings [17; 12]. Strong results in this direction could provide insight into the varying performance of machine learning methods across different models and problems.

Finally, determining how these results translate from deep linear networks to deep neural networks would be incredibly useful. This would require considering a lift that employs activation functions:

$$\phi' : \mathbb{R}^{d_1 \times d_0} \times \dots \times \mathbb{R}^{d_L \times d_{L-1}} \to \mathbb{R}, \qquad \phi'(W_1, \dots, W_L) = \sigma\left(W_L \sigma\left(W_{L-1} \sigma\left(\dots \sigma\left(W_1\right)\right)\right)\right)$$

for some activation function $\sigma : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$. Unfortunately, many of the most used activations are non-differentiable, e.g. ReLU, and modern training methods are often dynamic and stochastic, e.g. employing the dropout regulariser [10]. These features would lead to ϕ' being non-differentiable, let alone non-smooth, as well as potentially stochastic and non-continuous. That said, results that only consider smooth activation functions, e.g. the sigmoid function, and vanilla training regimes would still be invaluable for understanding the loss landscapes of deep neural networks and consequentially their strong performance across many problem domains.

5. Conclusion

TODO

References

[1] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018. [Cited on page 1.]

[2] Ehrhard Behrends. Where is matrix multiplication locally open? *Linear Algebra and its Applications*, 517:167–176, 2017. [Cited on page 17.]

[3] Alberto Bernacchia, Máté Lengyel, and Guillaume Hennequin. Exact natural gradient in deep linear networks and its application to the nonlinear case. *Advances in Neural Information Processing Systems*, 31, 2018. [Cited on pages 1 and 5.]

[4] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49, 2008. [Cited on page 5.]

[5] Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pages 1655–1664. PMLR, 2019. [Cited on pages 1 and 5.]

[6] Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955, 2020. [Cited on page 7.]

[7] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018. [Cited on pages 1 and 5.]

[8] Andrew K Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *arXiv preprint arXiv:1809.10374*, 2018. [Cited on pages 1 and 5.]

[9] Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pages 2902–2907. PMLR, 2018. [Cited on pages 1, 5, 6, 7, and 18.]

[10] Gian Paolo Leonardi and Matteo Spallanzani. Analytical aspects of non-differentiable neural networks. *arXiv preprint arXiv:2011.01858*, 2020. [Cited on page 18.]

[11] Eitan Levin, Joe Kileel, and Nicolas Boumal. The effect of smooth parametrizations on nonconvex optimization landscapes. *arXiv preprint arXiv:2207.03512*, 2022. [Cited on pages 1, 2, 3, 4, 5, and 6.]

[12] Ziye Ma and Somayeh Sojoudi. Noisy low-rank matrix optimization: Geometry of local minima and convergence rate. In *International Conference on Artificial Intelligence and Statistics*, pages 3125–3150. PMLR, 2023. [Cited on page 18.]

[13] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009. [Cited on pages 3 and 4.]

[14] Andrzej Ruszczynski. *Nonlinear optimization*. Princeton university press, 2011. [Cited on pages 2 and 3.]

[15] Ruoyu Sun, Dawei Li, Shiyu Liang, Tian Ding, and Rayadurgam Srikant. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37(5):95–108, 2020. [Cited on page 1.]

[16] Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66 (6):86–93, 2023. [Cited on pages 1 and 5.]

[17] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *J. Mach. Learn. Res.*, 20(114):1–34, 2019. [Cited on page 18.]

A. Computations

We computed the ∇^2 and ∇^3 of g, as defined in (31), at (W_1, W_2, W_3), as defined in (**??**), using Mathematica. The script can be found in this Github Repo. For cursory inspection, we also display the script on the following pages.

```
In[∘]:= W1 = {{1, 0, 0}, {0, 1, 0}};
        W3 = Transpose[W1];
        W2 = {{1, 0}, {0, 0}};
        W1 // MatrixForm
        W2 // MatrixForm
        W3 // MatrixForm
Out[∘]//MatrixForm=
        ( 1  0  0 )
        ( 0  1  0 )

Out[∘]//MatrixForm=
        ( 1  0 )
        ( 0  0 )

Out[∘]//MatrixForm=
        ( 1  0 )
        ( 0  1 )
        ( 0  0 )
```

```
In[∘]:= A = {{0, 0, 0}, {0, 0, x}, {0, y, z}};
        A // MatrixForm
Out[∘]//MatrixForm=
        ( 0  0  0 )
        ( 0  0  x )
        ( 0  y  z )
```

```
In[∘]:= W1dot = {{w1dot11, w1dot12, w1dot13}, {w1dot21, w1dot22, w1dot23}};
        W2dot = {{w2dot11, w2dot12}, {w2dot21, w2dot22}};
        W3dot = {{w3dot11, w3dot12}, {w3dot21, w3dot22}, {w3dot31, w3dot32}};
```

```
In[∘]:= D[Tr[Transpose[A].(W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot)] +
           c Tr[Transpose[((W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot) - W3.W2.W1)].
               ((W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot) - W3.W2.W1)], t] /. t → 0

Out[∘]= 0
```

```
In[∘]:= D[Tr[Transpose[A].(W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot)] +
           c Tr[Transpose[((W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot) - W3.W2.W1)].
               ((W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot) - W3.W2.W1)], {t, 2}] /. t → 0

Out[∘]= c (2 w1dot13^2 + 2 (w1dot12 + w2dot12)^2 + 2 w2dot22^2 +
            2 (w1dot11 + w2dot11 + w3dot11)^2 + 2 (w2dot21 + w3dot21)^2 + 2 w3dot31^2) +
          2 w1dot23 w2dot22 x + 2 w1dot12 w3dot31 y + 2 w2dot12 w3dot31 y +
          2 w2dot22 w3dot32 y + 2 w1dot13 (w2dot21 x + w3dot21 x + w3dot31 z)
```

```
In[∘]:= D[Tr[Transpose[A].(W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot)] +
          c Tr[Transpose[((W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot) - W3.W2.W1)].
             ((W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot) - W3.W2.W1)], {t, 3}] /. t → 0

Out[∘]= c (6 w1dot13 (2 w1dot23 w2dot12 + 2 w1dot13 (w2dot11 + w3dot11)) +
            6 (w1dot11 + w2dot11 + w3dot11) (2 w1dot21 w2dot12 +
               2 w2dot11 w3dot11 + 2 w1dot11 (w2dot11 + w3dot11) + 2 w2dot21 w3dot12) +
            6 (w1dot12 + w2dot12) (2 w1dot22 w2dot12 + 2 w2dot12 w3dot11 +
               2 w1dot12 (w2dot11 + w3dot11) + 2 w2dot22 w3dot12) +
            6 (w2dot21 + w3dot21) (2 w1dot21 w2dot22 + 2 w2dot11 w3dot21 +
               2 w1dot11 (w2dot21 + w3dot21) + 2 w2dot21 w3dot22) + 6 w2dot22 (2 w1dot22 w2dot22 +
               2 w2dot12 w3dot21 + 2 w1dot12 (w2dot21 + w3dot21) + 2 w2dot22 w3dot22) +
            6 w3dot31 (2 w1dot11 w3dot31 + 2 w2dot11 w3dot31 + 2 w2dot21 w3dot32)) +
         3 w1dot12 (2 w2dot11 w3dot31 y + 2 w2dot21 w3dot32 y) +
         3
          w1dot22
          (2 w2dot12 w3dot31 y + 2 w2dot22 w3dot32 y) +
         3 w1dot13 (2 w2dot11 (w3dot21 x + w3dot31 z) + 2 w2dot21 (w3dot22 x + w3dot32 z)) +
         3 w1dot23
          (2 w2dot12 (w3dot21 x + w3dot31 z) + 2 w2dot22 (w3dot22 x + w3dot32 z))
```

```
In[•]:= D[Tr[Transpose[A].(W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot)] +
          c Tr[Transpose[((W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot) - W3.W2.W1)].
             ((W3 + t W3dot).(W2 + t W2dot).(W1 + t W1dot) - W3.W2.W1)], {t, 4}] /. t → 0

Out[•]= c (6 (2 w1dot23 w2dot12 + 2 w1dot13 (w2dot11 + w3dot11))^2 +
          6 (2 w1dot21 w2dot12 + 2 w2dot11 w3dot11 +
                2 w1dot11 (w2dot11 + w3dot11) + 2 w2dot21 w3dot12)^2 +
          6 (2 w1dot22 w2dot12 + 2 w2dot12 w3dot11 + 2 w1dot12 (w2dot11 + w3dot11) +
                2 w2dot22 w3dot12)^2 +
          8 (w1dot11 + w2dot11 + w3dot11) (3 w1dot11 (2 w2dot11 w3dot11 + 2 w2dot21 w3dot12) +
             3 w1dot21 (2 w2dot12 w3dot11 + 2 w2dot22 w3dot12)) +
          8 (w1dot12 + w2dot12) (3 w1dot12 (2 w2dot11 w3dot11 + 2 w2dot21 w3dot12) +
             3 w1dot22 (2 w2dot12 w3dot11 + 2 w2dot22 w3dot12)) +
          8 w1dot13 (3 w1dot13 (2 w2dot11 w3dot11 + 2 w2dot21 w3dot12) +
             3 w1dot23 (2 w2dot12 w3dot11 + 2 w2dot22 w3dot12)) +
          6 (2 w1dot23 w2dot22 + 2 w1dot13 (w2dot21 + w3dot21))^2 +
          6 (2 w1dot21 w2dot22 + 2 w2dot11 w3dot21 + 2 w1dot11 (w2dot21 + w3dot21) +
                2 w2dot21 w3dot22)^2 + 6 (2 w1dot22 w2dot22 + 2 w2dot12 w3dot21 +
                2 w1dot12 (w2dot21 + w3dot21) + 2 w2dot22 w3dot22)^2 +
          8 (w2dot21 + w3dot21) (3 w1dot11 (2 w2dot11 w3dot21 + 2 w2dot21 w3dot22) +
             3 w1dot21 (2 w2dot12 w3dot21 + 2 w2dot22 w3dot22)) +
          8 w2dot22 (3 w1dot12 (2 w2dot11 w3dot21 + 2 w2dot21 w3dot22) +
             3 w1dot22 (2 w2dot12 w3dot21 + 2 w2dot22 w3dot22)) + 24 w1dot13^2 w3dot31^2 +
          6 (2 w1dot11 w3dot31 + 2 w2dot11 w3dot31 + 2 w2dot21 w3dot32)^2 +
          6 (2 w1dot12 w3dot31 + 2 w2dot12 w3dot31 + 2 w2dot22 w3dot32)^2 +
          8 w3dot31 (3 w1dot11 (2 w2dot11 w3dot31 + 2 w2dot21 w3dot32) +
             3 w1dot21 (2 w2dot12 w3dot31 + 2 w2dot22 w3dot32)))
```