# Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks


Peter L. Bartlett
UC Berkeley
bartlett@cs.berkeley.edu

David P. Helmbold
UC Santa Cruz
dph@soe.ucsc.edu

Philip M. Long
Google
plong@google.com


February 23, 2018


**Abstract**

We analyze algorithms for approximating a function $f(x) = \Phi x$ mapping $\Re^d$ to $\Re^d$ using deep linear neural networks, i.e. that learn a function $h$ parameterized by matrices $\Theta_1, ..., \Theta_L$ and defined by $h(x) = \Theta_L \Theta_{L-1} ... \Theta_1 x$. We focus on algorithms that learn through gradient descent on the population quadratic loss in the case that the distribution over the inputs is isotropic.

We provide polynomial bounds on the number of iterations for gradient descent to approximate the optimum, in the case where the initial hypothesis $\Theta_1 = ... = \Theta_L = I$ has loss bounded by a small enough constant. On the other hand, we show that gradient descent fails to converge for $\Phi$ whose distance from the identity is a larger constant, and we show that some forms of regularization toward the identity in each layer do not help.

If $\Phi$ is symmetric positive definite, we show that an algorithm that initializes $\Theta_i = I$ learns an $\epsilon$-approximation of $f$ using a number of updates polynomial in $L$, the condition number of $\Phi$, and $\log(d/\epsilon)$. In contrast, we show that if the target $\Phi$ is symmetric and has a negative eigenvalue, then all members of a class of algorithms that perform gradient descent with identity initialization, and optionally regularize toward the identity in each layer, fail to converge.

We analyze an algorithm for the case that $\Phi$ satisfies $u^\top \Phi u > 0$ for all $u$, but may not be symmetric. This algorithm uses two regularizers: one that maintains the invariant $u^\top \Theta_L \Theta_{L-1} ... \Theta_1 u > 0$ for all $u$, and another that "balances" $\Theta_1 ... \Theta_L$ so that they have the same singular values.


## 1 Introduction

Residual networks (He et al., 2016) are deep neural networks in which, roughly, subnetworks determine how a feature transformation should differ from the identity, rather than how it should differ from zero. After enabling the winning entry in the ILSVRC 2015 classification task, they have become established as a central idea in deep networks.

Hardt & Ma (2017) provided a theoretical analysis that shed light on residual networks. They showed that

- any linear transformation with a positive determinant and a bounded condition number can be approximated by a "deep linear network" of the form $f(x) = \Theta_L \Theta_{L-1} ... \Theta_1 x$, where, for large $L$, each layer $\Theta_i$ is close to the identity, and
- for networks that compose near-identity transformations this way, if the loss is large, then the gradient is steep.

These results are interesting because they suggest that, in many cases, this non-convex objective may be efficiently optimized through gradient descent if the layers stay close to the identity, possibly with the help of a regularizer.

This paper describes and analyzes such algorithms for linear regression with $d$ input variables and $d$ response variables with respect the quadratic loss, the same setting analyzed by Hardt and Ma. We abstract away sampling issues by analyzing an algorithm that performs gradient descent with respect to the population loss, as done by Hardt and Ma. We also assume that $y = \Phi x$; it is not hard to see that this is equivalent to the case that $y = \Phi x + \xi$ for noise $\xi$ that is independent of $x$ and has zero mean. Finally, we focus on the case that the distribution on the input patterns is isotropic. (The data may be transformed through a preprocessing step to approximately satisfy this constraint.)

The traditional analysis of convex optimization algorithms (see Boyd & Vandenberghe, 2004) provides a bound in terms of the quality of the initial solution, together with bounds on the eigenvalues of the Hessian of the loss. For the non-convex problem of this paper, we show that if gradient descent starts at the identity in each layer, and if the loss of that initial solution is bounded by a constant, then the Hessian remains well-conditioned enough throughout training for successful learning. Specifically, there is a constant $c_0$ such that, if the loss of the identity is at most $c_0$, then back-propagation initialized at the identity in each layer achieves loss at most $\epsilon$ in time polynomial in $\log(1/\epsilon)$, $d$, and $L$ (Section 3). On the other hand, we show that there is a constant $c_1$ and a $\Phi$ such that the identity has loss $c_1$ with respect to $\Phi$, but backpropagation with identity initialization fails to learn $\Phi$ (Section 6).

We also show that if the target $\Phi$ is symmetric positive definite then gradient descent with identity initialization achieves loss at most $\epsilon$ in a number of steps bounded by a polynomial in $\log(d/\epsilon)$, $L$ and the condition number of $\Phi$ (Section 4).

In contrast, for any $\Phi$ that is symmetric but has a negative eigenvalue, we show that no such guarantee is possible for a wide variety of algorithms of this type: the loss is forever bounded below by the square of this negative eigenvalue. This holds for step-and-project algorithms, and also algorithms that initialize at the identity, and regularize by early stopping or penalizing $\sum_i ||\Theta_i - I||_F^2$ (Section 6). Both this and the previous impossibility result can be proved using a target $\Phi$ with a positive determinant and a good condition number. Recall that such $\Phi$ were proved by Hardt and Ma to have a good approximation as a product of near-identity matrices – we prove that gradient descent cannot learn them, even with the help of regularizers that reward near-identity representations.

In Section 5 we provide a convergence guarantee for $\Phi$ that may not be symmetric, but satisfies $u^\top \Phi u > \gamma$ for a $\gamma > 0$ that appears in the bounds. We say that such matrices *have a margin* $\gamma$. Such $\Phi$ include rotations by acute angles. In this case, we consider an algorithm that regularizes in addition to a near-identity initialization. After the gradient update, the algorithm performs what we call *power projection*, projecting its hypothesis $\Theta_L \Theta_{L-1} ... \Theta_1$ onto the set of matrices with a margin $\gamma$. Second, it "balances" $\Theta_1, ..., \Theta_L$ so that, informally, they contribute equally to $\Theta_L \Theta_{L-1} ... \Theta_1$. (See Section 5 for the details.) We view this regularizer as a theoretically tractable proxy for regularizers that promote a large margin and balance between layers by adding penalties.

While, in practice, deep networks are non-linear, analysis of the linear case can provide a tractable way to gain insight through rigorous theoretical analysis (Saxe et al., 2013; Kawaguchi, 2016; Hardt & Ma, 2017). We might view back-propagation in the non-linear case as an approximation to a procedure that locally modifies the function computed by each layer in a manner that reduces the loss as fast as possible. If a non-linear network is obtained by composing transformations, each of which is chosen from a Hilbert space of functions (as in Daniely et al. (2016)), then a step in "function space" corresponds to a step in an (infinite-dimensional) linear space of functions.

**Related work.** The most closely related previous work is the paper by Hardt & Ma (2017) mentioned above. Saxe et al. (2013) studied the dynamics of a continuous-time process obtained by taking the step size of backpropagation applied to deep linear neural networks to zero. Kawaguchi (2016) showed that deep linear neural networks have no suboptimal local minima. In the case that $L = 2$, the problem studied here has a similar structure as problems arising from low-rank approximation of matrices, especially as regards algorithms that approximate a matrix $A$ by iteratively improving an approximation of the form $UV$. For an interesting survey on the rich literature on these algorithms, please see Ge et al. (2017a). Taghvaei et al. (2017) studied the properties of critical points on the loss when learning deep linear neural networks in the presence of a weight decay regularizer; they studied networks that transform the input to the output through a process indexed by a continuous variable, instead of through discrete layers. Lee et al. (2016) showed that, given regularity conditions, for a random initialization, gradient descent converges to a local minimizer almost surely; while their paper yields useful insights, their regularity condition does not hold for our problem. Many papers have analyzed learning of neural networks with non-linearities. The papers most closely related to this work analyze algorithms based on gradient descent. Some of these (Andoni et al., 2014; Brutzkus & Globerson, 2017; Ge et al., 2017b; Li & Yuan, 2017; Zhong et al., 2017; Zhang et al., 2018; Brutzkus et al., 2018; Ge et al., 2018) analyze constant-depth networks. Daniely (2017) showed that stochastic gradient descent learns a subclass of functions computed by log-depth networks in polynomial time; this class includes constant-degree polynomials with polynomially bounded coefficients. Other theoretical treatments of neural network learning algorithms include Lee et al. (1996); Arora et al. (2014); Livni et al. (2014); Janzamin et al. (2015); Safran & Shamir (2016); Zhang et al. (2016); Nguyen & Hein (2017); Zhang et al. (2017); Orhan & Pitkow (2018), although these are less closely related.

All three upper bound analyses use Hardt and Ma's lemmas, together with a new upper bound on the operator norm of the Hessian of a deep linear network. They otherwise have different outlines. Roughly, the bound in terms of the loss of the initial solution proceeds by showing that the distance from each layer to the identity grows slowly enough that the loss is reduced before the layers stray far enough to harm the conditioning of the Hessian. The bound for symmetric positive definite matrices proceeds by showing that, in this case, all of the layers are the same, and each of their eigenvalues converges to the $L$th root of a corresponding eigenvalue of $\Phi$. As mentioned above, the bound for matrices $\Phi$ with a positive margin is for an algorithm that achieves favorable conditioning through regularization.

We expect that the theoretical analysis reported here will inform the design of practical algorithms for learning non-linear deep networks. One potential avenue for this arises from the fact that the leverage provided by regularizing toward the identity appears to already be provided by a weaker policy of promoting the property that the composition of layers is (potentially asymmetric) positive definite. Also, balancing singular values of the layers of the network aided our analysis; an analogous balancing of Jacobians associated with various layers may improve conditioning in practice in the non-linear case.

# 2 Preliminaries

## 2.1 Setting

For a joint distribution $P$ with support contained in $\Re^d \times \Re^d$ and $g : \Re^d \to \Re^d$, define $\ell_P(g) = \mathbb{E}_{(X,Y)\sim P}(||g(X) - Y||^2/2)$. We focus on the case that, for $(X, Y)$ drawn from $P$, the marginal on $X$ is isotropic, with $\mathbb{E}XX^\top = I_d$, and $Y = \Phi X$ for $\Phi \in \Re^{d\times d}$. (As mentioned in the introduction, this is equivalent to the case of zero-mean noise that is independent of $X$.)

We study algorithms that learn linear mappings parameterized by deep networks. The network with $L$ layers and parameters $\Theta = (\Theta_1, \dots, \Theta_L)$ computes the parameterized function $f_\Theta(x) = \Theta_L \Theta_{L-1} \cdots \Theta_1 x$, where $x \in \Re^d$ and $\Theta_i \in \Re^{d\times d}$.

We use the notation $\Theta_{i:j} = \Theta_j \Theta_{j-1} \cdots \Theta_i$ for $i \le j$, so that we can write $f_\Theta(x) = \Theta_{1:L}x = \Theta_{i+1:L}\Theta_i\Theta_{1:i-1}x$.

When there is no possibility of confusion, we will sometimes refer to loss $\ell(f_\Theta)$ simply as $\ell(\Theta)$. Because the distribution of $X$ is isotropic, $\ell(\Theta) = \frac{1}{2}||\Theta_{1:L} - \Phi||_F^2$ with respect to target matrix $\Phi$. When $\Theta$ is produced by an iterative algorithm, will we also refer to loss of the $t$th iterate by $\ell(t)$.

**Definition 1.** *For $\gamma > 0$, a matrix $A \in \Re^{d\times d}$ has* a margin $\gamma$ *if, for all unit length $u$, we have $u^\top A u > \gamma$.*

## 2.2 Tools and background

We use $||A||_F$ for the Frobenius norm of matrix $A$, $||A||_2$ for its operator norm, and $\sigma_{\min}(A)$ for its least singular value. For vector $v$, we use $||v||$ for its Euclidian norm.

For a matrix $A$ and a matrix-valued function $B$, define $D_A B(A)$ to be the matrix with

$$(D_A B(A))_{i,j} = \frac{\partial \text{vec}(B(A))_i}{\partial \text{vec}(A)_j},$$

where vec($A$) is the column vector constructed by stacking the columns of $A$. We use $T_{d,d}$ to denote the $d^2 \times d^2$ permutation matrix mapping vec($A$) to vec($A^\top$) for $A \in \Re^{d\times d}$. For $A \in \Re^{n\times m}$ and $B \in \Re^{p\times q}$, $A \otimes B$ denotes the Kronecker product, that is, the $np \times mq$ matrix of $n \times m$ blocks, with the $i, j$th block given by $A_{ij}B$.

We will need the gradient and Hessian of $\ell$. (The gradient, which can be computed using backprop, is of course well known.) The proof is in the Appendix A.

**Lemma 1.** *For $i < j$,*

$$D_{\Theta_i}\ell\left(f_\Theta\right) = (\mathrm{vec}(I_d))^\top \left(\left(\Theta_{1:i-1}^\top \otimes (\Theta_{1:L} - \Phi)^\top \Theta_{i+1:L}\right)\right)$$
$$= \mathrm{vec}(G)^\top,$$

*for the $d \times d$ matrix given by*

$$G \overset{\mathrm{def}}{=} \Theta_{i+1:L}^\top \left(\Theta_{1:L} - \Phi\right) \Theta_{1:i-1}^\top. \tag{1}$$

$$D_{\Theta_j} D_{\Theta_i}\ell\left(f_\Theta\right) = \left(I_{d^2} \otimes (\mathrm{vec}(I_d))^\top\right)\left(I_d \otimes T_{d,d} \otimes I_d\right)\left(\mathrm{vec}(\Theta_{1:i-1}^\top) \otimes I_{d^2}\right)$$
$$((\Theta_{i+1:L}^\top \Theta_{j+1:L} \otimes \Theta_{1:j-1}^\top)T_{d,d}$$
$$+ (\Theta_{i+1:j-1}^\top \otimes (\Theta_{1:L} - \Phi)^\top \Theta_{j+1:L})).$$

$$D_{\Theta_i} D_{\Theta_i}\ell\left(f_\Theta\right) = \left(I_{d^2} \otimes (\mathrm{vec}(I_d))^\top\right)\left(I_d \otimes T_{d,d} \otimes I_d\right)\left(\mathrm{vec}(\Theta_{1:i-1}^\top) \otimes I_{d^2}\right)$$
$$\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}.$$

# 3 Targets near the identity

In this section, we prove an upper bound for gradient descent in terms of the loss of the initial solution.

## 3.1 Procedure and upper bound

First, set $\Theta^{(0)} = (I, I, ..., I)$, and then iteratively update

$$\Theta_i^{(t+1)} = \Theta_i^{(t)} - \eta(\Theta_{i+1:L})^\top \left(\Theta_{1:L}^{(t)} - \Phi\right)(\Theta_{1:i-1}^{(t)})^\top.$$

**Theorem 1.** *There are positive constants $c_1$ and $c_2$ and polynomials $p_1$ and $p_2$ such that, if $\ell(\Theta_{1:L}^{(0)}) \leq c_1$, $L \geq c_2$, and $\eta \leq \frac{1}{p_1(L,d,||\Phi||_2)}$, then the above gradient descent procedure achieves $\ell(f_{\Theta^{(t)}}) \leq \epsilon$ within $t = p_2(1/\eta)\ln\left(\frac{\ell(0)}{\epsilon}\right)$ iterations.*

## 3.2 Proof of Theorem 1

Hardt and Ma showed in the proof of their Theorem 2.2 that the gradient is steep if the loss is large and the singular values of the layers are not too small.

**Lemma 2** (Hardt & Ma 2017)**.** *Let $\nabla_\Theta \ell(\Theta)$ be the gradient of $\ell(\Theta)$ with respect to any flattening of $\Theta$. If, for all layers $i$, $\sigma_{\min}(\Theta_i) \geq 1 - a$, then $||\nabla_\Theta \ell(\Theta)||^2 \geq 4\ell(\Theta)L(1-a)^{2L}$.*

Next, we show that, if $\Theta^{(t)}$ and $\Theta^{(t+1)}$ are both close to the identity, then the gradient is not changing very fast between them, so that rapid progress continues to be made. We prove this through an upper bound on the operator norm of the Hessian that holds uniformly over members of a ball around the identity, which in turn can be obtained through a bound on the Frobenius norm. The proof is in Appendix B.

**Lemma 3.** *Choose an arbitrary* $\Theta$ *with* $||\Theta_i||_2 \le 1+z$ *for all* $i$*, and target* $\Phi$ *with* $||\Phi||_2 \le (1+z)^L$*. Let* $\nabla^2$ *be the Hessian of* $\ell(f_\Theta)$ *with respect to an arbitrary flattening of the parameters of* $\Theta$*. We have*

$$||\nabla^2||_F \le 3Ld^5(1+z)^{2L}.$$

Armed with Lemmas 2 and 3, let us now analyze gradient descent. Very roughly, our strategy will be to show that the distance from the identity to the various layers grows slowly enough for the leverage from Lemmas 2 and 3 to enable successful learning. Let $\mathcal{R}(\Theta) = \max_i ||\Theta_i - I||_2$. From the update, we have

$$\begin{aligned}
||\Theta_i^{(t+1)} - I||_2 &\le ||\Theta_i^{(t)} - I||_2 + \eta||(\Theta_{i+1:L}^{(t)})^\top \left(\Theta_{1:L}^{(t)} - \Phi\right)(\Theta_{1:i-1}^{(t)})^\top||_2 \\
&\le ||\Theta_i^{(t)} - I||_2 + \eta(1+\mathcal{R}(\Theta^{(t)}))^L ||\Theta_{1:L}^{(t)} - \Phi||_2 \\
&\le ||\Theta_i^{(t)} - I||_2 + \eta(1+\mathcal{R}(\Theta^{(t)}))^L ||\Theta_{1:L}^{(t)} - \Phi||_F.
\end{aligned}$$

If $\mathcal{R}(t) = \max_{s\le t} \mathcal{R}(\Theta^{(s)})$ (so $\mathcal{R}(0)=0$) and $\ell(t) = \frac{1}{2}||\Theta_{1:L}^{(t)} - \Phi||_F^2$, this implies

$$\mathcal{R}(t+1) \le \mathcal{R}(t) + \eta(1+\mathcal{R}(t))^L\sqrt{2\ell(t)}. \tag{2}$$

By Lemma 3, for all $\Theta$ on the line segment from $\Theta^{(t)}$ to $\Theta^{(t+1)}$, we have:

$$||\nabla_\Theta^2||_2 \le ||\nabla_{\Theta^{(t)}}^2||_F \le 3Ld^5 \max\{(1+\mathcal{R}(t+1)^{2L}, ||\Phi||_2^2\},$$

so that

$$\ell(t+1) \le \ell(t) - \eta||\nabla_{\Theta^{(t)}}||^2 + \frac{3}{2}\eta^2 Ld^5 \max\{(1+\mathcal{R}(t+1))^{2L}, ||\Phi||_2^2\}||\nabla_{\Theta^{(t)}}||^2.$$

Thus, if we ensure

$$\eta \le \frac{1}{3Ld^5(\max\{(1+\mathcal{R}(t+1))^{2L}, ||\Phi||_2^2\}}, \tag{3}$$

we have $\ell(t+1) \le \ell(t) - (\eta/2)||\nabla_{\Theta^{(t)}}||^2$, which, using Lemma 2, gives

$$\ell(t+1) \le \left(1 - 2\eta L(1-\mathcal{R}(t))^{2L}\right)\ell(t). \tag{4}$$

Pick any $c \ge 1$. Assume that $L \ge (4/3)\ln c = c_2$, $\ell(\Theta_{1:L}^{(0)}) \le \frac{\ln(c)^2}{8c^{10}} = c_1$ and $\eta \le \frac{1}{3Ld^5 \max\{c^2, ||\Phi||_2^2\}}$. We claim that, for all $t \ge 0$,

1. $\mathcal{R}(t) \le \eta c\sqrt{2\ell(0)} \sum_{0\le s<t} \exp\left(-\frac{s\eta L}{c^4}\right)$
2. $\ell(t) \le \left(\exp\left(-\frac{2t\eta L}{c^4}\right)\right)\ell(0)$.

The base case holds as $\mathcal{R}(0) = 0$ and $\ell(0) = \ell(0)$.

Before starting the inductive step, notice that for any $t \geq 0$,

$$\begin{aligned}
&\eta c\sqrt{2\ell(0)} \sum_{0\leq s<t} \exp\left(-\frac{s\eta L}{c^4}\right) \\
&\leq \eta c\sqrt{2\ell(0)} \times \frac{1}{1-\exp\left(-\frac{\eta L}{c^4}\right)} \\
&\leq \eta c\sqrt{2\ell(0)} \times \frac{2c^4}{\eta L} \quad (\text{since } \tfrac{\eta L}{c^4} \leq 1) \\
&\leq \frac{2c^5\sqrt{2\ell(0)}}{L} \\
&\leq \frac{\ln c}{L} \leq 3/4
\end{aligned}$$

where the last two inequalities follow from the constraints on $\ell(0)$ and $L$.
Using (2),

$$\begin{aligned}
\mathcal{R}(t+1) &\leq \mathcal{R}(t) + \eta(1+\mathcal{R}(t))^L\sqrt{2\ell(t)} \\
&\leq \mathcal{R}(t) + \eta\left(1+\frac{\ln c}{L}\right)^L \sqrt{2\ell(t)} \\
&\leq \mathcal{R}(t) + \eta c\sqrt{2\ell(t)} \\
&\leq \mathcal{R}(t) + \eta c\sqrt{2\ell(0)}\exp\left(-\frac{t\eta L}{c^4}\right) \\
&\leq \eta c\sqrt{2\ell(0)} \sum_{0\leq s<t+1} \exp\left(-\frac{s\eta L}{c^4}\right).
\end{aligned}$$

Since $\mathcal{R}(t+1) \leq \frac{\ln c}{L}$, the choice of $\eta$ satisfies (3), so

$$\ell(t+1) \leq \left(1 - 2\eta L(1-\mathcal{R}(t))^{2L}\right)\ell(t).$$

Now consider $(1-\mathcal{R}(t))^{2L}$:

$$\begin{aligned}
\ln\left((1-\mathcal{R}(t))^{2L}\right) &= 2L\ln(1-\mathcal{R}(t)) \\
&\geq 2L(-2\mathcal{R}(t)) && \text{since } \mathcal{R}(t) \in [0, 3/4] \\
&\geq 2L(-2\frac{\ln c}{L}) && \text{since } \mathcal{R}(t) \leq \frac{\ln c}{L} \\
(1-\mathcal{R}(t))^{2L} &\geq 1/c^4.
\end{aligned}$$

Using this in the bound on $\ell(t+1)$:

$$\begin{aligned}\ell(t+1) &\le \left(1 - 2\eta L(1-\mathcal{R}(t))^{2L}\right)\ell(t) \\ &\le \left(1 - \frac{2\eta L}{c^4}\right)\ell(t) \\ &\le \left(\exp\left(-\frac{2\eta L}{c^4}\right)\right)\left(\exp\left(-\frac{2t\eta L}{c^4}\right)\right)\ell(0) \\ &= \left(\exp\left(-\frac{2(t+1)\eta L}{c^4}\right)\right)\ell(0).\end{aligned}$$

Solving $\ell(0)\exp\left(-\frac{2t\eta L}{c^4}\right) \le \epsilon$ for $t$ and recalling that $\eta < 1/c^2$ completes the proof of the theorem.

# 4 Symmetric positive definite targets

In this section, we analyze the procedure of Section 3.1 when the target $\Phi$ is symmetric and positive definite.

**Theorem 2.** *There is an absolute positive constant $c_3$ such that, if $\Phi$ is symmetric with margin $0 < \gamma < 1$, and $L \ge c_3 \ln(||\Phi||_2/\gamma)$, then for all $\eta \le \frac{1}{L(1+||\Phi||_2^2)}$, gradient descent achieves $\ell(f_{\Theta^{(t)}}) \le \epsilon$ in $\mathrm{poly}(L, ||\Phi||_2/\gamma, 1/\eta)\log(d/\epsilon)$ iterations.*

## 4.1 Proof of Theorem 2

Let $\Phi$ be a symmetric real matrix with margin $\gamma > 0$, and let $\Theta^{(0)}, \Theta^{(1)}, ...$ be the iterates of gradient descent with a step size $\eta > 0$ that will be determined later by the analysis.

**Definition 2.** *Symmetric matrices $\mathcal{A} \subseteq \Re^{d\times d}$ are* commuting normal matrices *if there is a single unitary matrix $U$ such that for all $A \in \mathcal{A}$, $U^\top A U$ is diagonal.*

We will use the following well-known facts about commuting normal matrices.

**Lemma 4** (Horn & Johnson 2013)**.** *If $\mathcal{A} \subseteq \Re^{d\times d}$ is a set of symmetric commuting normal matrices and $A, B \in \mathcal{A}$, the following hold:*

- *$AB = BA$;*
- *for all scalars $\alpha$ and $\beta$, $\mathcal{A} \cup \{\alpha A + \beta B, AB\}$ are commuting normal;*
- *there is a unitary matrix $U$ such that $U^\top A U$ and $U^\top B U$ are real and diagonal;*
- *the set of singular values of $A$ is the same as the set of magnitudes of its eigenvalues;*
- *$||A - I||_2$ is the largest value of $|z-1|$ for an eigenvalue $z$ of $A$.*

**Lemma 5.** *The matrices $\{\Phi\} \cup \{\Theta_i^{(t)} : i \in \{1, ..., L\}, t \in \mathbb{Z}^+\}$ are commuting normal. For all $t$, $\Theta_1^{(t)} = ... = \Theta_L^{(t)}$.*

*Proof.* The proof is by induction. The base case follows from the fact that $\Phi$ and $I$ are commuting normal.

For the induction step, the fact that

$$\{\Phi\} \cup \{\Theta_i^{(s)} : i \in \{1, ..., L\}, s \leq t\} \cup \{\Theta_i^{(s+1)} : i \in \{1, ..., L\}, s \leq t\}$$

are commuting normal follows from Lemma 4. The update formula now reveals that $\Theta_1^{(t+1)} = ... = \Theta_L^{(t+1)}$. □

Now we are ready to analyze the dynamics of the learning process. Let $\Phi = U^\top D^L U$ be a diagonalization of $\Phi$. Let $\Gamma = \max\{1, ||\Phi||_2\}$. We next describe a sense in which gradient descent learns each eigenvalue independently.

**Lemma 6.** *For each $t$, there is a real diagonal matrix $\hat{D}^{(t)}$ such that, for all $i$, $\Theta_i^{(t)} = U^\top \hat{D}^{(t)} U$ and*

$$\hat{D}^{(t+1)} = \hat{D}^{(t)} - \eta(\hat{D}^{(t)})^{L-1}((\hat{D}^{(t)})^L - D^L). \tag{5}$$

*Proof.* Lemma 5 implies that there is a single real $U$ such that $\Theta_i^{(t)} = U^\top \hat{D}^{(t)} U$ for all $i$. Applying Lemma 1, recalling that $\Theta_1^{(t)} = ... = \Theta_L^{(t)}$, and applying the fact that $\Theta_i^{(t)}$ and $\Phi$ commute, we get

$$\Theta_i^{(t+1)} = \Theta_i^{(t)} - \eta(\Theta_i^{(t)})^{L-1} \left( (\Theta_i^{(t)})^L - \Phi \right).$$

Replacing each matrix by its diagonalization, we get

$$\begin{aligned} U^\top \hat{D}^{(t+1)} U &= U^\top \hat{D}^{(t)} U - \eta(U^\top (\hat{D}^{(t)})^{L-1} U) \left( U^\top (\hat{D}^{(t)})^L U - U^\top D^L U \right) \\ &= U^\top \hat{D}^{(t)} U - \eta U^\top (\hat{D}^{(t)})^{L-1} \left( (\hat{D}^{(t)})^L - D^L \right) U, \end{aligned}$$

and left-multiplying by $U$ and right-multiplying by $U^\top$ gives (5). □

We will now analyze the convergence of each $\hat{D}_{kk}^{(t)}$ to $D_{kk}$ separately. Let us focus for now on an arbitrary single index $k$, let $\lambda = D_{kk}$ and $\hat{\lambda}^{(t)} = \hat{D}_{kk}^{(t)}$.

Recalling that $||\Phi||_2 \leq \Gamma$, we have $\gamma^{1/L} \leq \lambda \leq \Gamma^{1/L}$. Also, $\Gamma^{1/L} = e^{\frac{1}{L} \ln \Gamma} \leq e^{1/a} \leq 1 + 2/a$ whenever $a \geq 1$ and $L \geq a \ln \Gamma$. Similarly, $\gamma^{1/L} \geq 1 - a$ whenever $L \geq a \ln(1/\gamma)$. Thus, there are absolute constants $c_3$ and $c_4$ such that $|1 - \lambda| \leq \frac{c_4 \ln(\Gamma/\gamma)}{L} < 1$ for all $L \geq c_3 \ln(\Gamma/\gamma)$.

We claim that, for all $t$, $\hat{\lambda}^{(t)}$ lies between 1 and $\lambda$ inclusive, so that $|\hat{\lambda}^{(t)} - \lambda| \leq \frac{c_4 \ln(\Gamma/\gamma)}{L}$. The base case holds because $\hat{\lambda}^{(t)} = 1$ and $|1 - \lambda| \leq \frac{c_4 \ln(\Gamma/\gamma)}{L}$. Now let us work on the induction step. Applying (5) together with Lemma 1, we get

$$\hat{\lambda}^{(t+1)} = \hat{\lambda}^{(t)} + \eta(\hat{\lambda}^{(t)})^{L-1}(\lambda^L - (\hat{\lambda}^{(t)})^L). \tag{6}$$

By the induction hypothesis, we just need to show that $|\hat{\lambda}^{(t+1)} - \lambda| \leq |\hat{\lambda}^{(t)} - \lambda|$. This would follow from $\text{sign}(\hat{\lambda}^{(t+1)} - \hat{\lambda}^{(t)}) = \text{sign}(\lambda - \hat{\lambda}^{(t)})$ and $|\hat{\lambda}^{(t+1)} - \hat{\lambda}^{(t)}| \leq |\lambda - \hat{\lambda}^{(t)}|$. (That is, the step is in the correct direction, and the algorithm does not "overshoot".) First, to see that the step is in the right direction, note that $\lambda^L \geq (\hat{\lambda}^{(t)})^L$ if and only if $\lambda \geq (\hat{\lambda}^{(t)})$, and the inductive hypothesis implies that $\hat{\lambda}^{(t)}$, and therefore $(\hat{\lambda}^{(t)})^{L-1}$, is non-negative. To show that $|\hat{\lambda}^{(t+1)} - \hat{\lambda}^{(t)}| \leq |\lambda - \hat{\lambda}^{(t)}|$, it suffices to show that $\eta(\hat{\lambda}^{(t)})^{L-1} \left|\lambda^L - (\hat{\lambda}^{(t)})^L\right| \leq |\lambda - \hat{\lambda}^{(t)}|$, which, in turn would be implied by $\eta \leq \left|\frac{1}{(\hat{\lambda}^{(t)})^{L-1}\left(\sum_{i=0}^{L-1}(\hat{\lambda}^{(t)})^i \lambda^{L-1-i}\right)}\right|$ (since $\lambda^L - (\hat{\lambda}^{(t)})^L = (\lambda - \hat{\lambda}^{(t)}) \sum_{i=0}^{L-1}(\hat{\lambda}^{(t)})^i \lambda^{L-1-i}$), which follows from the inductive hypothesis and $\eta \leq \frac{1}{L\Gamma^2}$.

We have proved that each $\hat{\lambda}^{(t)}$ lies between $\lambda$ and 1, so that $|1 - \hat{\lambda}^{(t)}| \leq |1 - \lambda| \leq c_4 \ln(\Gamma/\gamma)$.

Now, since the step is in the right direction, and does not overshoot,

$$\begin{aligned}
|\hat{\lambda}^{(t+1)} - \lambda| &\leq |\hat{\lambda}^{(t)} - \lambda| - \eta(\hat{\lambda}^{(t)})^{L-1}|\lambda^L - (\hat{\lambda}^{(t)})^L| \\
&\leq |\hat{\lambda}^{(t)} - \lambda| \left(1 - \eta(\hat{\lambda}^{(t)})^{L-1}\left(\sum_{i=0}^{L-1}(\hat{\lambda}^{(t)})^i \lambda^{L-1-i}\right)\right) \\
&\leq |\hat{\lambda}^{(t)} - \lambda| \left(1 - \eta L \gamma^2\right),
\end{aligned}$$

since the fact that $\hat{\lambda}^{(t)}$ lies between 1 and $\lambda$ implies that $\hat{\lambda}^{(t)} \geq \gamma^{1/L}$. Thus, $|\hat{\lambda}^{(t)} - \lambda| \leq \left(1 - \eta L \gamma^2\right)^t c_4 \ln(\Gamma/\gamma)$. This implies that, for any $\epsilon \in (0, 1)$, for any absolute constant $c_5$, there is a constant $c_6$ such that, after $c_6 \frac{1}{\eta L \gamma^2} \ln\left(\frac{dL \ln \Gamma}{\gamma \epsilon}\right)$ steps, we have $|\hat{\lambda}^{(t)} - \lambda| \leq \frac{c_5 \gamma \sqrt{\epsilon}}{L\Gamma\sqrt{d}}$. Writing $r = \hat{\lambda}^{(t)} - \lambda$, this implies, if $c_5$ is small enough, that

$$((\hat{\lambda}^{(t)})^L - \lambda^L)^2 = ((\lambda+r)^L - \lambda^L)^2 \leq \Gamma^2\left(\left(1+\frac{r}{\lambda}\right)^L - 1\right)^2 \leq \Gamma^2\left(\frac{2c_5 r L}{\lambda}\right)^2 \leq \Gamma^2\left(\frac{2c_5 r L}{\gamma}\right)^2 \leq \frac{\epsilon}{d}.$$

Thus, after $O\left(\frac{1}{\eta L \gamma^2}\ln\left(\frac{dL\ln\Gamma}{\gamma\epsilon}\right)\right)$ steps, $(D_{kk} - \hat{D}_{kk}^{(t)})^2 \leq \epsilon/d$ for all $k$, and therefore $\ell(\Theta^{(t)}) \leq \epsilon$, completing the proof.

# 5 Asymmetric positive definite matrices

In this section, we consider $\Phi$ that may be asymmetric, but which are positive definite in the sense that $u^\top \Phi u \geq \gamma \geq 0$ for all unit length $u$. This includes both rotations by an acute angle and "partial reflections" of the form $ax + b\,\mathrm{refl}(x)$ where $\mathrm{refl}(\cdot)$ is a length-preserving reflection and $0 \leq |b| < a$.

Since $\left(u^\top A u\right)^\top = u^\top A^\top u$, a matrix $A$ satisfies the above condition if and only if $u^\top(A + A^\top)u \geq 2\gamma$ for all unit length $u$, i.e. $A + A^\top$ is positive definite with eigenvalues at least $2\gamma$.

## 5.1 Balanced factorizations

The algorithm analyzed in this section uses a construction that is new, as far as we know, that we call a *balanced factorization.* This factorization may be of independent interest.

Recall that a *polar decomposition* of a matrix $A$ consists of a unitary matrix $R$ and a positive semidefinite matrix $P$ such that $A = RP$. The *principal Lth root* of a complex number whose expression in polar coordinates is $re^{\theta i}$ is $r^{1/L}e^{\theta i/L}$. The *principal* $L$*th root* of a matrix $A$ is the matrix $B$ such that $B^L = A$, and each eigenvalue of $B$ is the principal $L$th root of the corresponding eigenvalue of $A$.

**Definition 3.** *If $A$ be a matrix with polar decomposition $RP$, then $A$ has the* balanced factorization *$A = A_1, ..., A_L$ where for each $i$,*

$$A_i = R^{1/L} P_i, \text{ with } P_i = R^{(L-i)/L} P^{1/L} R^{-(L-i)/L},$$

*and each of the $L$th roots is the principal $L$th root.*

The motivation for balanced factorization is as follows. We want each factor to do a $1/L$ fraction of the total amount of rotation, and a $1/L$ fraction of the total amount of scaling. However, the scaling done by the $i$th factor should be done in directions that take account of the partial rotations done by the other factors. The following is the key property of the balanced factorization; it is proved in Appendix C.

**Lemma 7.** *If $\sigma_1, ..., \sigma_d$ are the singular values of $A$, and $A_1, ..., A_L$ is a balanced factorization of $A$, then the following hold: (a) $A = \prod_{i=1}^{L} A_i$; (b) for each $i \in \{1, ..., L\}$, $\sigma_1^{1/L}, ..., \sigma_d^{1/L}$ are the singular values of $A_i$.*

## 5.2 Procedure and upper bound

The following is the *power projection algorithm.* It has a (margin) parameter $\gamma > 0$, and uses

$$\mathcal{H} = \{A : \forall u \text{ s.t. } ||u|| = 1,\ u^\top A u \geq \gamma\}$$

as its "hypothesis space". First, it initializes $\Theta_i^{(0)} = \gamma^{1/L} I$ for all $i \in \{1, ..., L\}$. Then, for each $t$, it does the following.

- **Gradient Step.** For each $i \in \{1, ..., L\}$, update:

$$\Theta_i^{(t+1/2)} = \Theta_i^{(t)} - \eta(\Theta_{i+1:L}^{(t)})^\top \left(\Theta_{1:L}^{(t)} - \Phi\right) (\Theta_{1:i-1}^{(t)})^\top.$$

- **Power Project.** Compute the projection $\Psi^{(t+1/2)}$ (w.r.t. the Frobenius norm) of $\Theta_{1:L}^{(t+1/2)}$ onto $\mathcal{H}$.
- **Factor.** Let $\Theta_1^{(t+1)}, ..., \Theta_L^{(t+1)}$ be the balanced factorization of $\Psi^{(t+1/2)}$, so that $\Psi^{(t+1/2)} = \Theta_{1:L}^{(t+1)}$.

**Theorem 3.** *For any $\Phi$ such that $u^\top \Phi u > \gamma$ for all unit-length $u$, the power projection algorithm produces $\Theta^{(t)}$ with $\ell(\Theta^{(t)}) \leq \epsilon$ in* $\text{poly}(d, ||\Phi||_F, \frac{1}{\gamma}) \log(1/\epsilon)$ *iterations.*

### 5.3 Proof of Theorem 3

**Lemma 8.** *For all $t$, $\Theta_{1:L}^{(t)} \in \mathcal{H}$.*

*Proof.* $\Theta_{1:L}^{(0)} = \gamma I \in \mathcal{H}$, and, for all $t$, $\Psi^{(t+1/2)}$ is obtained by projection onto $\mathcal{H}$, and $\Theta_{1:L}^{(t+1)} = \Psi^{(t+1/2)}$. □

**Definition 4.** *The exponential of a matrix $A$ is $\exp(A) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} \frac{1}{k!} A^k$, and $B$ is a logarithm of $A$ if $A = \exp(B)$.*

**Lemma 9** (Culver 1966)**.** *A real matrix has a real logarithm if and only if it is invertible and each Jordan block belonging to a negative eigenvalue occurs an even number of times.*

**Lemma 10.** *For all $t$, $\Theta_{1:L}^{(t)}$ has a real $L$th root.*

*Proof.* Since $\Theta_{1:L}^{(t)} \in \mathcal{H}$ implies $u^\top \Theta_{1:L}^{(t)} u > 0$ for all $u$, $\Theta_{1:L}^{(t)}$ does not have a negative eigenvalue and is invertible. By Lemma 9, $\Theta_{1:L}^{(t)}$ has a real logarithm. Thus, its real $L$th root can be constructed via $\exp(\log(\Theta_{1:L}^{(t)})/L)$. □

The preceding lemma implies that the algorithm is well-defined, since all of the required roots can be calculated.

**Lemma 11.** *$\mathcal{H}$ is convex.*

*Proof.* Suppose $A$ and $B$ are in $H$ and $\lambda \in (0, 1)$. We have

$$u^\top (\lambda A + (1 - \lambda)B)u = \lambda u^\top Au + (1 - \lambda)u^\top Bu \geq \gamma.$$

□

**Lemma 12.** *For all $A \in \mathcal{H}$, $\sigma_{\min}(A) \geq \gamma$.*

*Proof.* Let $u$ and $v$ be singular vectors such that $uAv^\top = \sigma_{\min}(A)$.

$$\gamma \leq v^\top Av = \sigma_{\min}(A)v^\top u \leq \sigma_{\min}(A).$$

□

**Lemma 13.** *For all $t$, $\sigma_{\min}(\Theta_i^{(t)}) \geq \gamma^{1/L}$.*

*Proof.* First, $\sigma_{\min}(\Theta_i^{(0)}) = \gamma^{1/L} \geq \gamma^{1/L}$.

Now consider $t > 0$. Since $\Psi^{(t-1/2)}$ was projected into $\mathcal{H}$, we have $\sigma_{\min}(\Psi^{(t-1/2)}) \geq \gamma$. Lemma 7 then completes the proof. □

Define $U(t) = \max\left\{\max_{s\leq t}\max_i ||\Theta_i^{(s)}||_2, ||\Phi||_2^{1/L}\right\}$, $B(t) = \min_{s\leq t}\min_i \sigma_{\min}(\Theta_i^{(s)})$, and recall that $\ell(t) = ||\Theta_{1:L}^{(t)} - \Phi||_F^2$.

Arguing as in the initial portion of Section 3.2, as long as

$$\eta \leq \frac{1}{3Ld^5U(t)^{2L}} \tag{7}$$

we have $\ell(t+1/2) \leq \left(1 - \eta LB(t)^{2L}\right)\ell(t)$ (see Equation 4). Lemma 13 gives $B(t) \geq \gamma^{1/L}$, so $\ell(t+1/2) \leq \left(1-\eta L\gamma^2\right)\ell(t)$. Since $\Psi^{(t+1/2)}$ is the projection of $\Theta_{1:L}^{(t+1/2)}$ onto a convex set $\mathcal{H}$ that contains $\Phi$, and $\Theta_{1:L}^{(t+1)} = \Psi^{(t+1/2)}$, (7) implies

$$\ell(t+1) \leq \ell(t+1/2) \leq \left(1-\eta L\gamma^2\right)\ell(t). \tag{8}$$

Next, we prove an upper bound on $U$.

**Lemma 14.** *For all $t$, $U(t) \leq \left(\sqrt{\ell(t)} + ||\Phi||_F\right)^{1/L}$.*

*Proof.* Recall that $\ell(t) = ||\Theta_{1:L}^{(t)} - \Phi||_F^2$. By the triangle inequality, we have $||\Theta_{1:L}^{(t)}||_F \leq \sqrt{\ell(t)} + ||\Phi||_F$. Thus $||\Theta_{1:L}^{(t)}||_2 \leq \sqrt{\ell(t)} + ||\Phi||_F$. By Lemma 7, for all $i$, we have $||\Theta_i^{(t)}||_2 \leq \left(\sqrt{\ell(t)} + ||\Phi||_F\right)^{1/L}$. Since $||\Phi||_2 \leq ||\Phi||_F$, this completes the proof. □

Note that the triangle inequality implies that $\ell(0) \leq ||\Theta_{1:L}^{(0)}||_F^2 + ||\Phi||_F^2 \leq \gamma^2 d + ||\Phi||_F^2$. Since $\sigma_{\min}(\Phi) \geq \gamma$, we have $||\Phi||_F^2 \geq \gamma^2 d$, so $\ell(t) \leq 2||\Phi||_F^2$ and $U(t) \leq (3||\Phi||_2)^{1/L}$. Now, if we set $\eta = \frac{1}{cLd^5||\Phi||_F^2}$, for a large enough absolute constant $c$, then (7) is satisfied, so that (8) gives $\ell(t+1) \leq \left(1 - \frac{\gamma^2}{cd^5||\Phi||_F^2}\right)\ell(t)$ and the power projection algorithm achieves $\ell(t+1) \leq \epsilon$ after

$$O\left(\frac{d^5||\Phi||_F^2}{\gamma^2}\log\left(\frac{\ell(0)}{\epsilon}\right)\right) = O\left(\frac{d^5||\Phi||_F^2}{\gamma^2}\log\left(\frac{||\Phi||_F^2}{\epsilon}\right)\right)$$

updates.

# 6 Failure

In this section, we show that positive definite targets are necessary for several gradient descent algorithms with different kinds of regularization to minimize the loss. One family of algorithms that we will analyze is parameterized by a function $\psi$ mapping the number of inputs $d$ and the number of layers $L$ to a radius $\psi(d, L)$, step sizes $\eta_t$ and initialization parameter $\gamma \geq 0$. In particular, a *$\psi$-step-and-project* algorithm is any instantiation of the following algorithmic template.

Initialize each $\Theta_i^{(0)} = \gamma^{1/L} I$ for some $\gamma \geq 0$ and iterate:

- **Gradient Step.** For each $i \in \{1, ..., L\}$, update:

$$\Theta_i^{(t+1/2)} = \Theta_i^{(t)} - \eta_t(\Theta_{i+1:L}^{(t)})^\top \left(\Theta_{1:L}^{(t)} - \Phi\right)(\Theta_{1:i-1}^{(t)})^\top.$$

- **Project.** Set each $\Theta_i^{t+1}$ to the projection of $\Theta_i^{t+1/2}$ onto $\{A : ||A - I||_2 \leq \psi(d, L)\}$.

We will also show that *Penalty Regularized Gradient Descent* which uses gradient descent with any step sizes $\eta_t$ on the regularized objective $\ell(\Theta) + \frac{\kappa}{2}\sum_i ||I - \Theta||_F^2$ also fails to minimize the loss.

Both results use the simple observation that when $\Theta_{1:L}$ and $\Phi$ are mutually diagonalizable then

$$||\Theta_{1:L} - \Phi||_F^2 = ||U^\top \hat{D} U - U^\top D U||_F^2 = \sum_{j=1}^{d} (\hat{D}_{jj} - D_{jj})^2,$$

where the $D_{ii}$ are the eigenvalues of $\Phi$.

**Theorem 4.** *If the target $\Phi$ is symmetric then Penalty Regularized Gradient Descent produces hypotheses $\Theta_{1:L}^{(t)}$ that are commuting normal with $\Phi$.*

*In addition, if $\Phi$ has a negative eigenvalue $-\lambda$ and $L$ is even, then $\ell(\Theta^{(t)}) \geq \lambda^2/2$ for all $t$.*

*Proof.* For all $t$, Penalty Regularized Gradient Descent produces

$$\Theta_i^{(t+1)} = (1-\kappa)\Theta_i^{(t)} + \kappa I - \eta_t (\Theta_{i+1:L}^{(t)})^\top \left(\Theta_{1:L}^{(t)} - \Phi\right) (\Theta_{1:i-1}^{(t)})^\top.$$

Thus, by induction, the $\Theta_i^{(t)}$ are matrix polynomials of $\Phi$, and therefore they are all commuting normal. As in Lemmas 5 and 6 each $\Theta_i^{(t)}$ is the same $U^\top \tilde{D}^{(t)} U$ and $\Theta_{1:L}^{(t)} = U^\top (\tilde{D}^{(t)})^L U$. Since $L$ is even, each $(\tilde{D}^{(t)})_{jj}^L \geq 0$, so $\ell(\Theta^{(t)}) = \frac{1}{2}||\Theta_{1:L}^{(t)} - \Phi||_F^2 \geq \lambda^2/2$. □

To analyze step-and-project algorithms, it is helpful to first characterize the project step.

**Lemma 15.** *Let $X$ be a symmetric matrix and let $U^\top D U$ be its diagonalization.*

*For $a > 0$, let $Y$ be the Frobenius norm projection of $X$ onto*

$$\mathcal{B}_a = \{A : A \text{ is symmetric psd and } ||A - I||_2 \leq a\}.$$

*Then $Y = U^\top \tilde{D} U$ where $\tilde{D}$ is obtained from $D$ by projecting all of its diagonal elements onto $[1-a, 1+a]$.*

*Thus $\{X, Y\}$ are symmetric commuting normal matrices.*

*Proof.* First, if $X \in \mathcal{B}_a$, then $Y = X$ and we are done.

Assume $X \notin \mathcal{B}_a$. Clearly $U^\top \tilde{D} U \in \mathcal{B}_a$, so we just need to show that any member of $\mathcal{B}_a$ is at least as far from $X$ as $U^\top \tilde{D} U$ is. Let $\Lambda$ be the multiset of eigenvalues of $X$ (with repetitions) that are not in $[1-a, 1+a]$, and for each $\lambda \in \Lambda$, let $e_\lambda$ be the adjustment to $\lambda$ necessary to bring it to $[1-a, 1+a]$; i.e., so that $\lambda + e_\lambda$ is the projection of $\lambda$ onto $[1-a, 1+a]$.

If $u_\lambda$ is the eigenvector associated with $\lambda$, we have $U^\top \tilde{D} U - X = \sum_{\lambda \in \Lambda} e_\lambda u_\lambda u_\lambda^\top$, so that $||U^\top \tilde{D} U - X||_F^2 = \sum_{\lambda \in \Lambda} e_\lambda^2$.

Let $Z$ be an arbitrary member of $\mathcal{B}_a$. We would like to show that $||Z - X||_F^2 \geq \sum_{\lambda \in \Lambda} e_\lambda^2$. Since $Z \in \mathcal{B}_a$, we have $||Z - I||_2 \leq a$. $||Z - I||_2$ is the largest singular value of $Z - I$ so, for any unit length vector, in particular some $u_\lambda$ for $\lambda \in \Lambda$, $|u_\lambda^\top (Z - I) u_\lambda| = |u_\lambda^\top Z u_\lambda - 1| \leq a$, which implies $u_\lambda^\top Z u_\lambda \in [1-a, 1+a]$. Since $U$ is unitary $U^\top (X - Z) U$ has the same eigenvalues as $X - Z$, and, since the Frobenius norm is a function of the eigenvalues, $||U^\top (X - Z) U||_F = ||X - Z||_F$. But since $u_\lambda^\top Z u_\lambda \in [1-a, 1+a]$ for all $\lambda \in \Lambda$, just summing over the diagonal elements, we get $||U^\top (X - Z) U||_F^2 \geq \sum_{\lambda \in \Lambda} e_\lambda^2$, completing the proof. □

**Theorem 5.** *If the target $\Phi$ is symmetric then $\psi$-step-and-project algorithms produce hypotheses $\Theta_{1:L}^{(t)}$ that are commuting normal with $\Phi$.*

*In addition, if $\Phi$ has a negative eigenvalue $-\lambda$ and either $L$ is even or $\psi(L, d) \leq 1$, then $\ell(\Theta^{(t)}) \geq \lambda^2/2$ for all $t$.*

*Proof.* As in Lemmas 5 and 6, the $\Theta_i^{(t+1/2)}$ are identical and mutually diagonalizable with $\Phi$. Lemma 15 shows that this is preserved by the projection step. Thus there is a real diagonal $\tilde{D}^{(t)}$ such that each $\Theta_i^{(t)} = U^\top D_i^{(t)} U$, so $\Theta_{1:L}^{(t)} = U^\top (\tilde{D}^{(t)})^L U$.

When $L$ is even, each $(\tilde{D}^{(t)})^L)_{j,j} \geq 0$. When $\psi(d, L) \leq 1$ then the projection ensures that the elements of $\tilde{D}^{(t)}$ are non-negative, and thus each $(\tilde{D}^{(t)})^L)_{j,j} \geq 0$. In either case, $\ell(\Theta^{(t)}) = \frac{1}{2}||\Theta_{1:L}^{(t)} - \Phi||_F^2 \geq \lambda^2/2$. □

One choice of $\Phi$ that satisfies the requirements of Theorems 4 and 5 is $\Phi = \mathrm{diag}(-\lambda, 1, 1, ..., 1)$. For constant $\lambda$, the loss of $\Theta^{(0)} = (I, I, ..., I)$ is a constant for this target. Another choice is $\Phi = \mathrm{diag}(-\lambda, -\lambda, 1, 1, ..., 1)$, which has a positive determinant.


## Acknowledgements

We thank Yair Carmon, Nigel Duffy, Matt Feiszli, Roy Frostig, Vineet Gupta, Moritz Hardt, Tomer Koren, Antoine Saliou, Hanie Sedghi, Yoram Singer and Kunal Talwar for valuable conversations.

Peter Bartlett gratefully acknowledges the support of the NSF through grant IIS-1619362 and of the Australian Research Council through an Australian Laureate Fellowship (FL110100281) and through the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).


## References


Andoni, A., Panigrahy, R., Valiant, G., and Zhang, L. Learning polynomials with neural networks. In *ICML*, 2014.

Arora, Sanjeev, Bhaskara, Aditya, Ge, Rong, and Ma, Tengyu. Provable bounds for learning some deep representations. In *International Conference on Machine Learning*, pp. 584–592, 2014.

Boyd, S. P. and Vandenberghe, L. *Convex Optimization.* Cambridge University Press, Cambridge, UK, 2004. URL `http://www.stanford.edu/ boyd/cvxbook.html`.

Brutzkus, A. and Globerson, A. Globally optimal gradient descent for a convnet with gaussian inputs. In *ICML*, 2017.

Brutzkus, Alon, Globerson, Amir, Malach, Eran, and Shalev-Shwartz, Shai. SGD learns over-parameterized networks that provably generalize on linearly separable data. *ICLR*, 2018.

Culver, Walter J. On the existence and uniqueness of the real logarithm of a matrix. *Proceedings of the American Mathematical Society*, 17(5):1146–1151, 1966.

Daniely, A. SGD learns the conjugate kernel class of the network. *NIPS*, 2017.

Daniely, A., Frostig, R., and Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *NIPS*, 2016.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017a.

Ge, Rong, Lee, Jason D, and Ma, Tengyu. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017b.

Ge, Rong, Lee, Jason D, and Ma, Tengyu. Learning one-hidden-layer neural networks with landscape design. *ICLR*, 2018.

Hardt, M. and Ma, T. Identity matters in deep learning. *ICLR*, 2017.

Harville, D. A. *Matrix algebra from a statistician's perspective*, volume 1. Springer, 1997.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Horn, R. A. *Topics in Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1986. ISBN 0-521-30587-X.

Horn, Roger A and Johnson, Charles R. *Matrix analysis*. Cambridge University Press, 2013. Second edition.

Janzamin, Majid, Sedghi, Hanie, and Anandkumar, Anima. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

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

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *COLT*, 2016.

Lee, Wee Sun, Bartlett, Peter L, and Williamson, Robert C. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

Li, Yuanzhi and Yuan, Yang. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pp. 597–607, 2017.

Livni, Roi, Shalev-Shwartz, Shai, and Shamir, Ohad. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pp. 855–863, 2014.

Nguyen, Quynh and Hein, Matthias. The loss surface of deep and wide neural networks. In *ICML*, pp. 2603–2612, 2017.

Orhan, A Emin and Pitkow, Xaq. Skip connections eliminate singularities. *ICLR*, 2018.

Safran, Itay and Shamir, Ohad. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pp. 774–782, 2016.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Taghvaei, A., Kim, J. W., and Mehta, P. How regularization affects the critical points in linear networks. In *NIPS*, 2017.

Zhang, Q., Panigrahy, R., and Sachdeva, S. Electron-proton dynamics in deep learning. *ITCS*, 2018.

Zhang, Yuchen, Lee, Jason D, and Jordan, Michael I. l1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pp. 993–1001, 2016.

Zhang, Yuchen, Lee, Jason, Wainwright, Martin, and Jordan, Michael. On the learnability of fully-connected neural networks. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 83–91, 2017.

Zhong, Kai, Song, Zhao, Jain, Prateek, Bartlett, Peter L., and Dhillon, Inderjit S. Recovery guarantees for one-hidden-layer neural networks. In *ICML*, pp. 4140–4149, 2017.

# A Proof of Lemma 1

We rely on the following facts (Horn, 1986; Harville, 1997).

**Lemma 16.** *For compatible matrices (and, where $m, n, p, q, r, s$ are mentioned, $A \in \Re^{m \times n}$, $B \in$*

$\Re^{p\times q}$, $X \in \Re^{r\times s}$):

$$\begin{aligned}
A \otimes (B \otimes E) &= (A \otimes B) \otimes E, \\
AC \otimes BD &= (A \otimes B)(C \otimes D), \\
(A \otimes B)^\top &= A^\top \otimes B^\top, \\
\mathrm{vec}(AXB) &= (B^\top \otimes A)\mathrm{vec}(X), \\
T_{m,n}\mathrm{vec}(A) &\overset{\mathrm{def}}{=} \mathrm{vec}(A^\top), \\
T_{n,m}T_{m,n} &= I_{mn}, \\
T_{m,n} &= T_{n,m}^\top, \\
T_{1,n} &= T_{n,1} = I_n, \\
D_X(A(B(X))) &= D_B(A(B(X)))D_X(B(X)), \\
D_X(A(X)B(X)) &= (B(X)^\top \otimes I_m)D_X A(X) \\
&\quad + (I_q \otimes A(X))D_X B(X), \\
D_X(A(X)^T) &= T_{n,m}D_X(A(X)), \\
D_X(AXB) &= B^\top \otimes A, \\
D_A(A \otimes B) &= (I_n \otimes T_{q,m} \otimes I_p)(I_{mn} \otimes \mathrm{vec}(B)) \\
&= (I_{nq} \otimes T_{m,p})(I_n \otimes \mathrm{vec}(B) \otimes I_m), \\
D_B(A \otimes B) &= (I_n \otimes T_{q,m} \otimes I_p)(\mathrm{vec}(A) \otimes I_{pq}) \\
&= (T_{p,q} \otimes I_{mn})(I_q \otimes \mathrm{vec}(A) \otimes I_p).
\end{aligned}$$

Armed with Lemma 16, we now prove Lemma 1. We have

$$D_{\Theta_i} f_\Theta(x) = D_{\Theta_i}\left(\Theta_{i+1:L}\Theta_i\Theta_{1:i-1}x\right) = \left(\Theta_{1:i-1}x\right)^\top \otimes \Theta_{i+1:L}.$$

Again, from Lemma 16

$$\begin{aligned}
&D_{\Theta_i}\left(D_{\Theta_j} f_\Theta(x)\right) \\
&= D_{\Theta_i}\left(\left(\Theta_{1:j-1}x\right)^\top \otimes \Theta_{j+1:L}\right) \\
&= D_{\Theta_{1:j-1}x}\left(\left(\Theta_{1:j-1}x\right)^\top \otimes \Theta_{j+1:L}\right) D_{\Theta_i}\left(\Theta_{1:j-1}x\right) \\
&\qquad \text{(by the chain rule, since } i < j\text{)} \\
&= D_{\Theta_{1:j-1}x}\left(\left(\left(\Theta_{1:j-1}x\right) \otimes \Theta_{j+1:L}^\top\right)^\top\right)\left(\left(\Theta_{1:i-1}x\right)^\top \otimes \Theta_{i+1:j-1}\right). \qquad (9)
\end{aligned}$$

Define $P = \Theta_{1:j-1}x$ and $Q = \Theta_{j+1:L}$, so that $P \in \Re^{d\times 1}$ and $Q \in \Re^{d\times d}$. We have

$$\begin{aligned}
D_P\left(\left(P \otimes Q^\top\right)^\top\right) &= T_{d^2,d}D_P\left(P \otimes Q^\top\right) \\
&= T_{d^2,d}(I_1 \otimes T_{d,d} \otimes I_d)(I_d \otimes \mathrm{vec}(Q^T)) \\
&= T_{d^2,d}(T_{d,d} \otimes I_d)(I_d \otimes \mathrm{vec}(Q^\top)).
\end{aligned}$$

Substituting back into (9), we get

$$\begin{aligned}
&D_{\Theta_i}\left(D_{\Theta_j} f_\Theta(x)\right)\\
&= T_{d^2,d}(T_{d,d}\otimes I_d)(I_d\otimes \mathrm{vec}(\Theta_{j+1:L}^\top))\left((\Theta_{1:i-1}x)^\top\otimes\Theta_{i+1:j-1}\right).
\end{aligned}$$

The product rule in Lemma 16 gives, for each $i$,

$$\begin{aligned}
&D_{\Theta_i}\ell\left(f_\Theta\right)\\
&= \mathbb{E}(D_{\Theta_i}(\ell(f_\Theta(X)))\\
&= \mathbb{E}(D_{\Theta_i}(\frac{1}{2}(f_\Theta(X)-\Phi X)^\top(f_\Theta(X)-\Phi X)))\\
&= \mathbb{E}(((\Theta_{1:L}-\Phi)X)^\top D_{\Theta_i} f_\Theta(X))\\
&= \mathbb{E}\left(((\Theta_{1:L}-\Phi)X)^\top\left((\Theta_{1:i-1}X)^\top\otimes\Theta_{i+1:L}\right)\right)\\
&= \mathbb{E}\left((I_1\otimes((\Theta_{1:L}-\Phi)X)^\top)\left((\Theta_{1:i-1}X)^\top\otimes\Theta_{i+1:L}\right)\right)\\
&= \mathbb{E}\left(\left((\Theta_{1:i-1}X)^\top\otimes((\Theta_{1:L}-\Phi)X)^\top\Theta_{i+1:L}\right)\right)\\
&= \mathbb{E}\left(\left(X^\top\Theta_{1:i-1}^\top\right)\otimes\left(X^\top(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right)\right)\\
&= \mathbb{E}\left((X^\top\otimes X^\top)\left(\Theta_{1:i-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right)\right)\\
&= \mathbb{E}\left((X\otimes X)\mathrm{vec}(1)\right)^\top\left(\Theta_{1:i-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right)\\
&= \mathbb{E}\left(\mathrm{vec}(XX^\top)\right)^\top\left(\Theta_{1:i-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right)\\
&= (\mathrm{vec}(I_d))^T\left(\Theta_{1:i-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
(D_{\Theta_i}\ell\left(f_\Theta\right))^\top &= \left(\Theta_{1:i-1}\otimes\Theta_{i+1:L}^\top(\Theta_{1:L}-\Phi)\right)(\mathrm{vec}(I_d))\\
&= \mathrm{vec}\left(\Theta_{i+1:L}^\top(\Theta_{1:L}-\Phi)I_d\Theta_{1:i-1}^\top\right).
\end{aligned}$$

Also, recalling that $i<j$, we have

$$\begin{aligned}
&D_{\Theta_j}D_{\Theta_i}\ell\left(f_\Theta\right)\\
&= D_{\Theta_j}\left((\mathrm{vec}(I_d))^T\left(\Theta_{1:i-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right)\right)\\
&= (I_{d^2}\otimes(\mathrm{vec}(I_d))^T)D_{\Theta_j}\left(\Theta_{1:i-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right)\\
&= (I_{d^2}\otimes(\mathrm{vec}(I_d))^T)\left(I_d\otimes T_{d,d}\otimes I_d\right)\left(\mathrm{vec}(\Theta_{1:i-1}^\top)\otimes I_{d^2}\right)D_{\Theta_j}\left((\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right).
\end{aligned}$$

Continuing with the subproblem,

$$
\begin{aligned}
D_{\Theta_j}\left((\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right) &= (\Theta_{i+1:L}^\top\otimes I_d)D_{\Theta_j}\left((\Theta_{1:L}-\Phi)^\top\right) \\
&\quad + (I_d\otimes(\Theta_{1:L}-\Phi)^\top)D_{\Theta_j}\left(\Theta_{i+1:L}\right) \\
&= (\Theta_{i+1:L}^\top\otimes I_d)D_{\Theta_j}\left(\Theta_{1:L}^\top\right) \\
&\quad + (I_d\otimes(\Theta_{1:L}-\Phi)^\top)D_{\Theta_j}\left(\Theta_{i+1:L}\right) \\
&= (\Theta_{i+1:L}^\top\otimes I_d)\left(\Theta_{j+1:L}\otimes\Theta_{1:j-1}^\top\right)D_{\Theta_j}(\Theta_j^\top) \\
&\quad + (I_d\otimes(\Theta_{1:L}-\Phi)^\top)\left(\Theta_{i+1:j-1}^\top\otimes\Theta_{j+1:L}\right) \\
&= (\Theta_{i+1:L}^\top\otimes I_d)\left(\Theta_{j+1:L}\otimes\Theta_{1:j-1}^\top\right)T_{d,d} \\
&\quad + (I_d\otimes(\Theta_{1:L}-\Phi)^\top)\left(\Theta_{i+1:j-1}^\top\otimes\Theta_{j+1:L}\right) \\
&= \left(\Theta_{i+1:L}^\top\Theta_{j+1:L}\otimes\Theta_{1:j-1}^\top\right)T_{d,d} \\
&\quad + \left(\Theta_{i+1:j-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{j+1:L}\right).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
D_{\Theta_i}D_{\Theta_i}\ell(f_\Theta) &= D_{\Theta_i}\left((\text{vec}(I_d))^T\left(\Theta_{1:i-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right)\right) \\
&= (I_{d^2}\otimes(\text{vec}(I_d))^T)D_{\Theta_i}\left(\Theta_{1:i-1}^\top\otimes(\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right) \\
&= (I_{d^2}\otimes(\text{vec}(I_d))^T)\,(I_d\otimes T_{d,d}\otimes I_d) \\
&\qquad \left(\text{vec}(\Theta_{1:i-1}^\top)\otimes I_{d^2}\right)D_{\Theta_i}\left((\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
D_{\Theta_i}\left((\Theta_{1:L}-\Phi)^\top\Theta_{i+1:L}\right) &= (\Theta_{i+1:L}^\top\otimes I_d)D_{\Theta_i}\left((\Theta_{1:L}-\Phi)^\top\right) \\
&= (\Theta_{i+1:L}^\top\otimes I_d)D_{\Theta_i}\left(\Theta_{1:L}^\top\right) \\
&= (\Theta_{i+1:L}^\top\otimes I_d)\left(\Theta_{i+1:L}\otimes\Theta_{1:i-1}^\top\right)D_{\Theta_i}(\Theta_i^\top) \\
&= (\Theta_{i+1:L}^\top\otimes I_d)\left(\Theta_{i+1:L}\otimes\Theta_{1:i-1}^\top\right)T_{d,d} \\
&= \left(\Theta_{i+1:L}^\top\Theta_{i+1:L}\otimes\Theta_{1:i-1}^\top\right)T_{d,d}.
\end{aligned}
$$

# B Proof of Lemma 3

We have

$$
||\nabla^2||_F^2 = 2\sum_{i<j}||D_{\Theta_j}D_{\Theta_i}\ell(f_\Theta)||_F^2 + \sum_i ||D_{\Theta_i}D_{\Theta_i}\ell(f_\Theta)||_F^2. \tag{10}
$$

Let's start with the easier term. Choose $\Theta$ such that $||\Theta_i - I||_2 \leq z$ for all $i$. We have

$$
\begin{aligned}
||D_{\Theta_i} D_{\Theta_i} \ell\left(f_\Theta\right)||_F &= \left|\left|\left(I_{d^2} \otimes (\operatorname{vec}(I_d))^\top\right)\left(I_d \otimes T_{d,d} \otimes I_d\right)\left(\operatorname{vec}(\Theta_{1:i-1}^\top) \otimes I_{d^2}\right)\right.\right. \\
&\quad \left.\left.\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}\right|\right|_F \\
&\leq \left|\left|\left(I_{d^2} \otimes (\operatorname{vec}(I_d))^\top\right)\left(I_d \otimes T_{d,d} \otimes I_d\right)\right|\right|_F \\
&\quad \times \left|\left|\left(\operatorname{vec}(\Theta_{1:i-1}^\top) \otimes I_{d^2}\right)\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}\right|\right|_F \\
&= d^{3/2} \left|\left|\left(\operatorname{vec}(\Theta_{1:i-1}^\top) \otimes I_{d^2}\right)\right.\right. \\
&\qquad \left.\left.\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}\right|\right|_F \\
&\leq d^{3/2} \left|\left|\left(\operatorname{vec}(\Theta_{1:i-1}^\top) \otimes I_{d^2}\right)\right|\right|_F \\
&\quad \times \left|\left|\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}\right|\right|_F \\
&= d^{7/2} \left|\left|\operatorname{vec}(\Theta_{1:i-1}^\top)\right|\right|_F \left|\left|\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}\right|\right|_F \\
&= d^{7/2} \, ||\Theta_{1:i-1}||_F \left|\left|\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}\right|\right|_F \\
&\leq d^4 \, ||\Theta_{1:i-1}||_2 \left|\left|\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}\right|\right|_F \\
&\leq d^4 (1+z)^{i-1} \left|\left|\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right) T_{d,d}\right|\right|_F \\
&= d^4 (1+z)^{i-1} \left|\left|\left(\Theta_{i+1:L}^\top \Theta_{i+1:L} \otimes \Theta_{1:i-1}^\top\right)\right|\right|_F \\
&= d^4 (1+z)^{i-1} \left|\left|\Theta_{i+1:L}^\top \Theta_{i+1:L}\right|\right|_F \times \left|\left|\Theta_{1:i-1}^\top\right|\right|_F \\
&\leq d^5 (1+z)^{i-1} \left|\left|\Theta_{i+1:L}^\top \Theta_{i+1:L}\right|\right|_2 \times \left|\left|\Theta_{1:i-1}^\top\right|\right|_2 \\
&\leq d^5 (1+z)^{2(L-1)}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
||D_{\Theta_j} D_{\Theta_i} \ell\left(f_{\Theta}\right)||_F = & \left|\left|\left(I_{d^2} \otimes (\operatorname{vec}(I))^{\top}\right)\left(I_d \otimes T_{d,d} \otimes I_d\right)\left(\operatorname{vec}(\Theta_{1:i-1}^{\top}) \otimes I_{d^2}\right)\right.\right. \\
& \left(\left(\Theta_{i+1:L}^{\top} \Theta_{j+1:L} \otimes \Theta_{1:j-1}^{\top}\right) T_{d,d}\right. \\
& \left.\left. + \left(\Theta_{i+1:j-1}^{\top} \otimes (\Theta_{1:L} - \Phi)^{\top} \Theta_{j+1:L}\right)\right)\right|\right|_F \\
\leq & \, d^4 (1+z)^{i-1} || \left(\Theta_{i+1:L}^{\top} \Theta_{j+1:L} \otimes \Theta_{1:j-1}^{\top}\right) T_{d,d} \\
& + \left(\Theta_{i+1:j-1}^{\top} \otimes (\Theta_{1:L} - \Phi)^{\top} \Theta_{j+1:L}\right) ||_F \\
\leq & \, d^4 (1+z)^{i-1} \left(\left|\left|\left(\Theta_{i+1:L}^{\top} \Theta_{j+1:L} \otimes \Theta_{1:j-1}^{\top}\right) T_{d,d}\right|\right|_F \right. \\
& \left. + \left|\left|\left(\Theta_{i+1:j-1}^{\top} \otimes (\Theta_{1:L} - \Phi)^{\top} \Theta_{j+1:L}\right)\right|\right|_F\right) \\
\leq & \, d^4 (1+z)^{i-1} \big(d(1+z)^{2L-1-i} \\
& + \left|\left|\left(\Theta_{i+1:j-1}^{\top} \otimes (\Theta_{1:L} - \Phi)^{\top} \Theta_{j+1:L}\right)\right|\right|_F \big) \\
= & \, d^4 (1+z)^{i-1} \big(d(1+z)^{2L-1-i} \\
& + ||\Theta_{i+1:j-1}||_F \times \left|\left|(\Theta_{1:L} - \Phi)^{\top} \Theta_{j+1:L}\right|\right|_F \big) \\
\leq & \, d^4 (1+z)^{i-1} \left(d(1+z)^{2L-1-i} + 2d(1+z)^{2L-1-i}\right) \\
= & \, 3d^5 (1+z)^{2L-2}.
\end{aligned}
$$

Putting these together with (10), we get $||\nabla^2||_F^2 \leq L^2 9 d^{10} (1+z)^{4L}$, so that

$$||\nabla^2||_F \leq 3Ld^5(1+z)^{2L}.$$

# C Proof of Lemma 7

Recall that a *polar decomposition* of a matrix $A$ consists of a unitary matrix $R$ and a positive semidefinite matrix $P$ such that $A = RP$.

**Lemma 17** ((Horn & Johnson, 2013))**.** *$A$ is a unitary matrix if and only if all of the (complex) eigenvalues $z$ of $A$ have magnitude* 1*.*

**Lemma 18** ((Horn & Johnson, 2013))**.** *If $A$ is unitary then $A$ is normal.*

**Lemma 19** ((Horn & Johnson, 2013))**.** *If $A$ is normal with eigenvalues $\lambda_1, ..., \lambda_d$, the singular values of $A$ are $|\lambda_1|, ..., |\lambda_d|$.*

**Lemma 20.** *If $A$ is unitary, then $A^{1/L}$ is unitary.*

*Proof.* By induction on $i$, $A^{i/L}$ is unitary for all $i$. □

**Lemma 21.** *If $A$ is invertible and normal with singular values $\sigma_1, ..., \sigma_d$, then, for any positive integer $L$, the singular values of $A^{1/L}$ are $\sigma_1^{1/L}, ..., \sigma_d^{1/L}$.*

*Proof.* Follows from Lemma 19 together with the fact that raising a non-singular matrix to a power results in raising its eigenvalues to the same power. □

**Lemma 22** ((Horn & Johnson, 2013))**.** *If $A = RP$ is the polar decomposition of $A$, then the singular values of $A$ are the same as the singular values of $P$.*

**Lemma 23.** *If $\sigma_1, ..., \sigma_d$ are the principal components of $A$, and $A = \prod_{i=1}^{L} A_i$ is a balanced factorization of $A$, then then $\sigma_1^{1/L}, ..., \sigma_d^{1/L}$ are the principal components of $A_i$, for each $i \in \{1, ..., L\}$.*

*Proof.* The singular values of $A_i = R_i P_i$ are the same as the singular values of $P_i$, which is similar to $P^{1/L}$, whose singular values are the $L$th roots of the singular values of $P$, which are the same as the singular values of $A$. □

**Lemma 24.** *If $A_1, ..., A_L$ is a balanced factorization of $A$, then*

$$A = \prod_{i=1}^{L} A_i.$$

*Proof.* We have

$$\begin{aligned}
A &= RP \\
&= R^{1/L} R^{1-1/L} P^{1/L} P^{1-1/L} \\
&= R^{1/L} R^{1-1/L} P^{1/L} R^{-(1-1/L)} R^{1-1/L} P^{1-1/L} \\
&= R_1 P_1 R^{1-1/L} P^{1-1/L} \\
&= A_1 R^{1-1/L} P^{1-1/L} \\
&= A_1 R^{1/L} R^{1-2/L} P^{1/L} P^{1-2/L}
\end{aligned}$$

and so on. □