

Orthogonal Skip Connections

Filip Morawiec

June 2025

Abstract

ResNet architecture was a breakthrough in deep learning. Its success has been attributed to the use of skip connections, which in turn mitigated the problem of exploding gradients. It has been shown that ResNet overcame exploding gradients because its skip connections (identity matrices) are norm preserving. In this paper, we replace these identity matrices with orthogonal matrices, which are by definition norm preserving. First, following the previous formulation of ResNet, we interestingly show that as the number of layers grows, the orthogonal matrices need to converge to the identity matrix. Then, we explore the performance of orthogonal skip connections across multiple regression tasks, using the recent efficient algorithm for gradient descent on orthogonal matrices [Bernstein(2025)]. In contrary to previous works, we show that orthogonal skip connections do not outperform identity skip connections and require more training time to converge. We also analyze the validity of the theoretical construction, showing that the orthogonal matrices do not converge to identity matrices in practice as fast as the theory suggests. Finally, we explore alternative training schemes that could make use of orthogonal skip connections, but find that they do not outperform the standard training scheme. *TODO: apply different training schemes, optimizers.*

Contents

1	Introduction	3
2	Related works	3
3	Theoretical considerations	3
3.1	Notation and definitions	3
3.2	Decomposing a function into near-orthogonal functions	3
3.2.1	Proof setup	4
3.2.2	Bounding the first term $\ h_1 - W_1\ _L$	5
3.2.3	Bounding the general term $\ h_i - W_i\ _L$	5
3.2.4	Combining the estimates	9
3.3	Optimality Conditions via Zero Fréchet Derivatives	9
3.4	Analyzing the convergence of orthogonal connections during training	9
4	Empirical results	9
4.1	Performance	9
4.2	Validity of main theorem: distance of orthogonal weights to identity	10
4.3	Different optimizers	11

1 Introduction

Our main contributions:

- we show that under a certain mathematical construction, orthogonal skip connections converge to identity connections (as network depth increases);
- we explore if the theoretical construction actually matches the trained model;
- we explore alternative schemes of trainings that could make use of orthogonal skip connections;
- we argue that orthogonal skip connections aren't a good replacement for identity skip connections

2 Related works

This paper has already done it, but they claim the method works, but it is 1 percentage point better: <https://arxiv.org/abs/1707.05974>

Recent paper with a different idea: <https://arxiv.org/abs/2505.11881>.

3 Theoretical considerations

This part wouldn't have been possible if not the work of [Bartlett et al.(2018)Bartlett, Evans, and Long], which came up with a construction of the decomposition of a function into near-identity functions. Their proof has been adapted to orthogonal matrices.

3.1 Notation and definitions

3.2 Decomposing a function into near-orthogonal functions

Theorem 1. Fix $R > 0$, denote $\mathcal{X} = B_R(\mathbb{R}^d)$. Let $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a differentiable, invertible map satisfying the following properties:

1. **Smoothness:** for some $\alpha > 0$ and all $x, y, u \in \mathcal{X}$,

$$\|(Dh(y) - Dh(x))u\| \leq \alpha \|y - x\| \|u\|; \quad (1)$$

2. **Lipschitz inverse:** for some $M > 0$, $\|h^{-1}\|_L \leq M$;

3. **Positive orientation:** For some $x_0 \in \mathcal{X}$, $\det(Dh(x_0)) > 0$.

4. **Zero at origin:** $h(0) = 0$

Then for all m , there exist m functions $h_1, \dots, h_m : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and m orthogonal matrices $W_1, \dots, W_m \in \mathbb{R}^{n \times n}$ satisfying, for all $x \in \mathcal{X}$,

$$h_m \circ h_{m-1} \circ \dots \circ h_1(x) = h(x),$$

and, on $h_{i-1} \circ \dots \circ h_1(\mathcal{X})$, the following terms decrease in $O\left(\frac{\log m}{m}\right)$ rate: $\|h_i - W_i\|_L$, $\|I - W_i\|$ for $i = 1, \dots, m$.

To help prove the theorem, we will note the following:

Lemma 1. *For h satisfying the conditions of Theorem 1 and any $x, y \in \mathcal{X}$,*

$$\|h(y) - h(x) - Dh(x)(y - x)\| \leq \frac{\alpha}{2} \|y - x\|^2,$$

and moreover,

$$\|h\|_L \leq 1 + \alpha R.$$

Now we can proceed to prove Theorem 1.

Proof. To organize the proof, we will first set up our notation and then prove two main bounds (Case 1 and Case 2) that together imply the final estimate.

3.2.1 Proof setup

For $i = 1, 2, \dots, m$, define

$$g_i(x) := \frac{1}{a_i} W_i^\top h(a_i W_i x), \quad g_i : \mathcal{X} \longrightarrow \mathbb{R}^d.$$

Each $a_i > 0$ is a scalar parameter with

$$0 < a_1 < a_2 < \dots < a_m = 1, \quad a_i = (1 - c)^{m-i} \text{ for some } c \in (0, 1).$$

The orthogonal matrices $W_i \in \mathbb{R}^{d \times d}$ satisfy $\|I - W_i\| \leq C\|a_i - a_{i-1}\|$ for some constant $C > 0$ and all $i = 1, \dots, m$. Note that from this property we obtain bounds on closeness of consecutive W_i 's:

$$\|W_i - W_{i-1}\| \leq \|I - W_i\| + \|I - W_{i-1}\| \tag{2}$$

$$\leq C\|a_i - a_{i-1}\| + C\|a_{i-1} - a_{i-2}\| \tag{3}$$

$$\leq C\|a_i - a_{i-2}\| < C\|a_i - a_{i-1}\| \tag{4}$$

and

$$\|a_i W_i - a_{i-1} W_{i-1}\| \leq \|a_i W_i - a_i W_{i-1}\| + \|a_i W_{i-1} - a_{i-1} W_{i-1}\| \tag{5}$$

$$\leq a_i \|W_i - W_{i-1}\| + \|a_i - a_{i-1}\| \tag{6}$$

$$< a_i C \|a_i - a_{i-1}\| + \|a_i - a_{i-1}\| = \|a_i - a_{i-1}\| (Ca_i + 1) \tag{7}$$

Then we set

$$h_i := g_i \circ g_{i-1}^{-1} \quad (i = 1, \dots, m),$$

with the convention $g_0 = \text{Id}$ (the identity map). Observe that

$$h = h_m \circ \dots \circ h_1 = g_m \circ g_{m-1}^{-1} \circ \dots \circ g_1 \circ g_0^{-1} = g_m \circ g_{m-1} \circ \dots \circ g_1,$$

so indeed the composition of h_1, \dots, h_m yields h .

We aim to show

$$\|h_i - W_i\|_L := \sup_{x \neq y} \frac{\|(h_i(x) - W_i x) - (h_i(y) - W_i y)\|}{\|x - y\|} \text{ is } O\left(\frac{\log m}{m}\right).$$

3.2.2 Bounding the first term $\|h_1 - W_1\|_L$

We have

$$h_1(x) = g_1(x) = \frac{1}{a_1} W_1^\top h(a_1 W_1 x).$$

Hence

$$h_1(x) - W_1 x = \frac{1}{a_1} W_1^\top h(a_1 W_1 x) - W_1 x.$$

Consider two inputs $x, y \in \mathcal{X}$. We want to bound

$$\|(h_1(x) - W_1 x) - (h_1(y) - W_1 y)\|.$$

Using Lemma 1, expand

$$h(a_1 W_1 x) = h(a_1 W_1 y) + Dh(a_1 W_1 y)(a_1 W_1(x - y)) + r,$$

where $\|r\| \leq (\alpha/2) \|a_1 W_1(x - y)\|^2$. Hence

$$\begin{aligned} & \|(h_1(x) - W_1 x) - (h_1(y) - W_1 y)\| = \\ & \left\| \frac{1}{a_1} W_1^\top [h(a_1 W_1 x) - h(a_1 W_1 y)] - W_1(x - y) \right\| \leq \\ & \left\| \frac{1}{a_1} W_1^\top Dh(a_1 W_1 y)(a_1 W_1(x - y)) - W_1(x - y) \right\| + \left\| \frac{1}{a_1} W_1^\top r \right\| \leq \\ & \left\| Dh(W_1 y) - I \right\| a_1 \|x - y\| + \frac{\alpha a_1}{2} \|x - y\|^2 = \\ & a_1 \|x - y\| \left[\left\| Dh(W_1 y) - I \right\| - \frac{\alpha}{2} \|x - y\| \right] \end{aligned}$$

3.2.3 Bounding the general term $\|h_i - W_i\|_L$

For $i > 1$, recall

$$h_i = g_i \circ g_{i-1}^{-1},$$

so

$$h_i(x) - W_i x = g_i(g_{i-1}^{-1}(x)) - W_i x.$$

Given $x, y \in \mathcal{X}$, define

$$u := g_{i-1}^{-1}(x), \quad v := g_{i-1}^{-1}(y).$$

Then

$$x = g_{i-1}(u), \quad y = g_{i-1}(v).$$

Note that since $\|h^{-1}\|_L \leq M$, we have

$$\begin{aligned}
\|x - y\| &= \frac{1}{a_{i-1}} \|a_{i-1} W_{i-1} g_{i-1}(u) - a_{i-1} W_{i-1} g_{i-1}(v)\| \geq \\
&\frac{1}{M a_{i-1}} \|h^{-1}(a_{i-1} W_{i-1} g_{i-1}(u)) - h^{-1}(a_{i-1} W_{i-1} g_{i-1}(v))\| = \\
&\frac{1}{M a_{i-1}} \|h^{-1}(h(a_{i-1} W_{i-1} u)) - h^{-1}(h(a_{i-1} W_{i-1} v))\| = \\
&\frac{1}{M a_{i-1}} \|a_{i-1} W_{i-1} u - a_{i-1} W_{i-1} v\| = \\
&\frac{1}{M} \|u - v\|
\end{aligned}$$

and also $\|v\| \leq MR$ and $\|u\| \leq MR$. **TODO: Update the following to use this**
The difference we need to bound is

$$\|(h_i(x) - W_i x) - (h_i(y) - W_i y)\| = \|g_i(u) - g_i(v) - W_i [g_{i-1}(u) - g_{i-1}(v)]\|.$$

Recall

$$g_i(u) = \frac{1}{a_i} W_i^\top h(a_i W_i u), \quad g_{i-1}(u) = \frac{1}{a_{i-1}} W_{i-1}^\top h(a_{i-1} W_{i-1} u).$$

Hence we can rewrite

$$\begin{aligned}
&g_i(u) - g_i(v) - W_i [g_{i-1}(u) - g_{i-1}(v)] \\
&= \frac{1}{a_i} W_i^\top [h(a_i W_i u) - h(a_i W_i v)] - \frac{1}{a_{i-1}} W_i W_{i-1}^\top [h(a_{i-1} W_{i-1} u) - h(a_{i-1} W_{i-1} v)].
\end{aligned}$$

To shorten notation, let

$$A := a_i W_i u, \quad B := a_i W_i v, \quad X := a_{i-1} W_{i-1} u, \quad Y := a_{i-1} W_{i-1} v.$$

By Lemma 1, we can write

$$\begin{aligned}
h(A) - h(B) &= Dh(B) (A - B) + r_1, \quad \text{where } \|r_1\| \leq \frac{\alpha}{2} \|A - B\|^2, \\
h(X) - h(Y) &= Dh(Y) (X - Y) + r_2, \quad \text{where } \|r_2\| \leq \frac{\alpha}{2} \|X - Y\|^2.
\end{aligned}$$

and expanding further:

$$\begin{aligned}
\|r_1\| &\leq \frac{\alpha}{2} \|A - B\|^2 \leq \frac{\alpha}{2} a_i^2 \|u - v\|^2 \\
\|r_2\| &\leq \frac{\alpha}{2} \|X - Y\|^2 \leq \frac{\alpha}{2} a_{i-1}^2 \|x - y\|^2
\end{aligned}$$

Thus

$$\begin{aligned}
&\left\| \frac{1}{a_i} W_i^\top [h(A) - h(B)] - \frac{1}{a_{i-1}} W_i W_{i-1}^\top [h(X) - h(Y)] \right\| \\
&= \left\| \frac{1}{a_i} W_i^\top [Dh(B) (A - B) + r_1] - \frac{1}{a_{i-1}} W_i W_{i-1}^\top [Dh(Y) (X - Y) + r_2] \right\| \\
&\leq \left\| \frac{1}{a_i} W_i^\top Dh(B) (A - B) - \frac{1}{a_{i-1}} W_i W_{i-1}^\top Dh(Y) (X - Y) \right\| + \underbrace{\left\| \frac{1}{a_i} W_i^\top r_1 - \frac{1}{a_{i-1}} W_i W_{i-1}^\top r_2 \right\|}_{r :=} \\
&\leq \left\| \underbrace{\frac{1}{a_i} W_i^\top Dh(B) a_i W_i - \frac{1}{a_{i-1}} W_i W_{i-1}^\top Dh(Y) a_{i-1} W_{i-1}}_{J :=} \right\| \|u - v\| + r.
\end{aligned}$$

We can split and bound the term J in the following way:

$$\begin{aligned}
\|J\| &= \left\| \frac{1}{a_i} W_i^\top Dh(B) a_i W_i - \frac{1}{a_{i-1}} W_i W_{i-1}^\top Dh(Y) a_{i-1} W_{i-1} \right\| = \\
&= \left\| W_i^\top Dh(B) W_i - W_i W_{i-1}^\top Dh(Y) W_{i-1} \right\| = \\
&= \left\| \left[W_i^\top Dh(B) W_i - W_i^\top Dh(Y) W_i \right] + \left[W_i^\top Dh(Y) W_i - W_i W_{i-1}^\top Dh(Y) W_{i-1} \right] \right\| \leq \\
&\leq \alpha \|B - Y\|^2 + \underbrace{\left\| W_i^\top Dh(Y) W_i - W_i W_{i-1}^\top Dh(Y) W_{i-1} \right\|}_{J_2 :=}
\end{aligned}$$

Then, abusing notation and writing $Dh(Y)$ as D

$$\begin{aligned}
\|J_2\| &= \left\| \left[W_i^\top D W_i - W_i^\top D W_{i-1} \right] + \left[W_i^\top D W_{i-1} - W_i W_{i-1}^\top D W_{i-1} \right] \right\| \leq \\
&\leq \|D\| \|W_i - W_{i-1}\| + \left\| W_i^\top - W_i W_{i-1}^\top \right\| \|D\| = \\
&= \|D\| \left(\|W_i - W_{i-1}\| + \left\| W_i^\top - W_i W_{i-1}^\top + W_i W_{i-1}^\top - W_i W_{i-1}^\top \right\| \right) \leq \\
&\leq \|D\| \left(\|W_i - W_{i-1}\| + \left\| W_i^\top - W_i W_{i-1}^\top \right\| + \left\| W_i W_{i-1}^\top - W_i W_{i-1}^\top \right\| \right) = \\
&= \|D\| \left(\|W_i - W_{i-1}\| + \|I - W_i\| + \left\| W_i^\top - W_{i-1}^\top \right\| \right) = \\
&= \|D\| (\|W_i - W_{i-1}\| + \|I - W_i\| + \|W_i - W_{i-1}\|) .
\end{aligned}$$

Note that we can bound D by:

$$\|Dh(a_{i-1} W_{i-1} v)\| \leq \alpha \|a_{i-1} W_{i-1} v\| \leq \alpha a_{i-1} M R \quad (8)$$

Combining, we obtain:

$$\begin{aligned}
&\left\| (h_i(x) - W_i x) - (h_i(y) - W_i y) \right\| \leq r + \|u - v\| \left(\alpha \|B - Y\|^2 + J_2 \right) \leq \\
&\leq \frac{\alpha}{2} (a_i^2 \|u - v\|^2 + a_{i-1}^2 \|x - y\|^2) + \\
&\|u - v\| \left(\alpha \|a_i W_i v - a_{i-1} W_{i-1} v\|^2 + \|a_{i-1} M R\| (2\|W_i - W_{i-1}\| + \|I - W_i\|) \right) \\
&\leq \frac{\alpha}{2} ((a_i^2 M^2 + a_{i-1}^2) \|x - y\|^2) + \\
&M \|x - y\| \left(\alpha \|a_i W_i - a_{i-1} W_{i-1}\| R^2 + \|a_{i-1} M R\| (2\|W_i - W_{i-1}\| + \|I - W_i\|) \right) \\
&\leq \|x - y\| \left[\frac{\alpha}{2} M^2 \|x - y\| + \alpha \|a_i W_i - a_{i-1} W_{i-1}\| R^2 + \|a_{i-1} M R\| (2\|W_i - W_{i-1}\| + \|I - W_i\|) \right] \\
&\quad \text{using the equations (4) and (7), we can write:} \\
&< \|x - y\| \left[\frac{\alpha}{2} M^2 \|x - y\| + \alpha (C a_i + 1) (a_i - a_{i-1}) R^2 + \|a_{i-1} M R\| (2C(a_i - a_{i-1}) + C(a_i - a_{i-1})) \right] \\
&\leq \|x - y\| \left[\frac{\alpha}{2} M^2 \|x - y\| + \alpha (C a_i + 1) (a_i - a_{i-1}) R^2 + 3C \|a_{i-1} M R\| (a_i - a_{i-1}) \right].
\end{aligned}$$

Case 1: Arguments are close. If $\|x - y\| \leq \|a_i - a_{i-1}\|$, then we can bound the above expression by

$$\|h_i - W_i\|_L < \left[\frac{\alpha}{2} M^2 + \alpha (C a_i + 1) R^2 + 3C \|a_{i-1} M R\| \right] \|a_i - a_{i-1}\|$$

Case 2: Arguments are far. If $\|x - y\| \geq a_i - a_{i-1}$, we can again use the fact that the increments $a_i - a_{i-1}$, as well as the orthogonal matrix differences $\|W_i - W_{i-1}\|$ and $\|I - W_i\|$, are all $O(a_i - a_{i-1})$, but this time also use the global Lipschitz properties of h .

We can rewrite the main term, substituting $S_k(z) = g_k(z) - g_{k-1}(z)$:

$$\begin{aligned} (h_i(x) - W_i x) - (h_i(y) - W_i y) &= [g_i(u) - g_{i-1}(u) - (g_i(v) - g_{i-1}(v))] + (I - W_i)[g_{i-1}(u) - g_{i-1}(v)] \\ &= S_i(u) + S_i(v) + (I - W_i)[g_{i-1}(u) - g_{i-1}(v)] = \\ &= S_i(u) - S_i(v) + (I - W_i)[x - y] \end{aligned}$$

We have already shown the term $I - W_i = O(a_i - a_{i-1})$. Now we will show the function $S_i(z)$ is Lipschitz with constant $O(a_i - a_{i-1})$.

Since $Dg_k(z) = W_k^\top Dh(a_k W_k z) W_k$, we have

$$\begin{aligned} DS_i(z) &= W_i^\top Dh(a_i W_i z) W_i - W_{i-1}^\top Dh(a_{i-1} W_{i-1} z) W_{i-1}, \\ &= W_i^\top [Dh(a_i W_i z) - Dh(a_{i-1} W_{i-1} z)] W_i + \\ &\quad + (W_i^\top - W_{i-1}^\top) Dh(a_{i-1} W_{i-1} z) W_{i-1} \\ &\quad + W_i^\top Dh(a_{i-1} W_{i-1} z) (W_i - W_{i-1}). \end{aligned}$$

Smoothness (1) and $\|z\| \leq R$ give

$$\|Dh(a_i W_i z) - Dh(a_{i-1} W_{i-1} z)\| \leq \alpha \|a_i W_i z - a_{i-1} W_{i-1} z\| < \alpha R \|a_i - a_{i-1}\| (Ca_i + 1).$$

Therefore, combining the bounds for the terms of $DS_i(z)$ and using the triangle inequality, we obtain

$$\begin{aligned} \|DS_i(z)\| &\leq \|Dh(a_i W_i z) - Dh(a_{i-1} W_{i-1} z)\| + \|W_i^\top - W_{i-1}^\top\| \|Dh(a_{i-1} W_{i-1} z)\| + \\ &\quad + \|Dh(a_{i-1} W_{i-1} z)\| \|W_i - W_{i-1}\| \\ &\leq \alpha R \|a_i - a_{i-1}\| (Ca_i + 1) + 2C \|a_i - a_{i-1}\| \|Dh(a_{i-1} W_{i-1} z)\| \\ &\leq \|a_i - a_{i-1}\| (\alpha R (Ca_i + 1) + 2C \alpha a_{i-1} R) = \\ &\leq \alpha R \|a_i - a_{i-1}\| (Ca_i + 2Ca_{i-1} + 1). \end{aligned}$$

Now we can bound the Lipschitz constant of S_i :

$$\begin{aligned} \|S_i(v) - S_i(u)\| &\leq \|DS_i(z)\| \cdot \|v - u\| \\ &\leq \alpha R \|a_i - a_{i-1}\| (Ca_i + 2Ca_{i-1} + 1) \cdot \|v - u\| \\ &= \alpha R M^{-1} \|a_i - a_{i-1}\| (Ca_i + 2Ca_{i-1} + 1) \cdot \|x - y\| \end{aligned}$$

Going back to the difference we need to bound, we have

$$\begin{aligned} \|(h_i(x) - W_i x) - (h_i(y) - W_i y)\| &\leq \|S_i(u) - S_i(v)\| + \|I - W_i\| \cdot \|x - y\| \\ &< \|x - y\| \|a_i - a_{i-1}\| \cdot [\alpha R M^{-1} (Ca_i + 2Ca_{i-1} + 1) + C] \end{aligned}$$

3.2.4 Combining the estimates

We have shown that for all $i = 1, \dots, m$:

$$\begin{aligned} \|h_i - W_i\|_L &\leq \|a_i - a_{i-1}\| \cdot \max\left\{\frac{\alpha}{2}M^2 + \alpha(Ca_i + 1)R^2 + 3C\|a_{i-1}MR\|, \right. \\ &\quad \left. \alpha RM^{-1}(Ca_i + 2Ca_{i-1} + 1) + C\right\} \\ &\leq \|a_i - a_{i-1}\| \cdot \max\left\{\frac{\alpha}{2}M^2 + \alpha(C + 1)R^2 + 3CMR, \right. \\ &\quad \left. \alpha RM^{-1}(3C + 1) + C\right\}. \end{aligned}$$

The choice of $a_i = (1 - c)^{m-i}$ guarantees $a_i - a_{i-1} = O\left(\frac{\log m}{m}\right)$, which implies that $\|h_i - W_i\|_L = O\left(\frac{\log m}{m}\right)$ for all $i = 1, \dots, m$.

Finally, we can conclude that the composition of $h_m \circ h_{m-1} \circ \dots \circ h_1$ converges to h in the sense of the theorem. □

3.3 Optimality Conditions via Zero Fréchet Derivatives

NOTE: Again reference to [Bartlett et al.(2018)Bartlett, Evans, and Long]

3.4 Analyzing the convergence of orthogonal connections during training

TODO: Refer to the papers where authors show singular vectors of consecutive layers converge to each other.

4 Empirical results

4.1 Performance

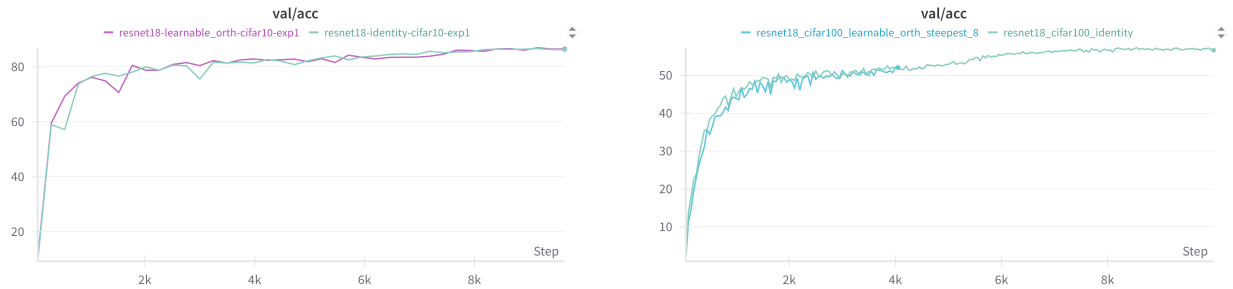


Figure 1: Comparison of validation accuracy of ResNets with identity and orthogonal skip connections. *Left*: CIFAR-10. *Right*: CIFAR-100.

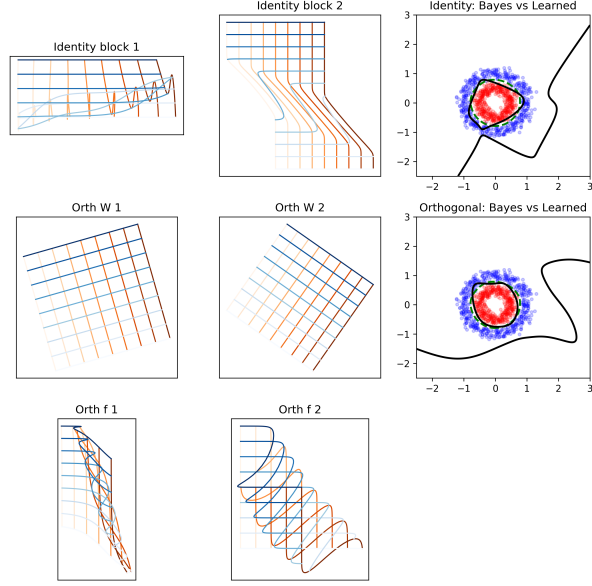


Figure 2: Comparison of validation accuracy of ResNets with identity and orthogonal skip connections on a synthetic dataset. *Upper row*: Identity blocks. Learned non-linear transformations. *Middle row*: Orthogonal blocks. Learned non-linear transformations. *Lower row*: Orthogonal blocks. Learned orthogonal matrices.

In contrast to [Wang et al.(2017)Wang, Xing, Zhang, and Zhang], using orthogonal skip connections instead of the identity on CIFAR-10, CIFAR-100 doesn't show any difference, as seen on Figure 1.

Similarly for training on synthetic data, as seen on the left plot of Figure 2 *Note: add a different plot here.*

4.2 Validity of main theorem: distance of orthogonal weights to identity

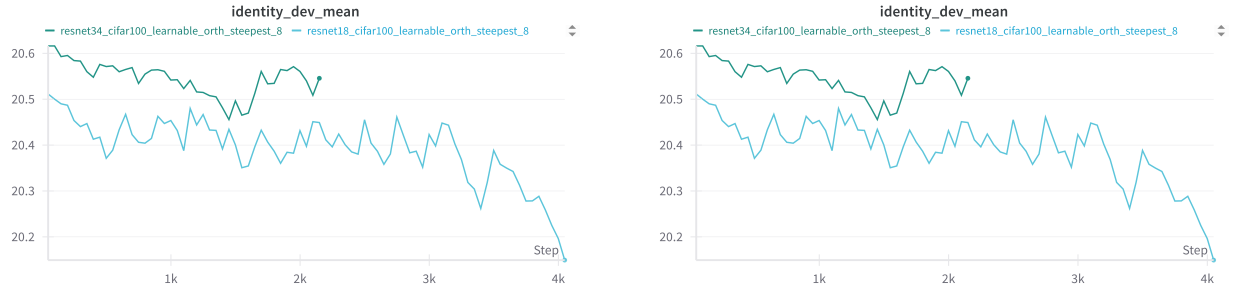


Figure 3: Trajectory of orthogonal skip connections of ResNet18 and ResNet50. *Left*: CIFAR-10. *Right*: CIFAR-100. *Note: add a different plot here.*

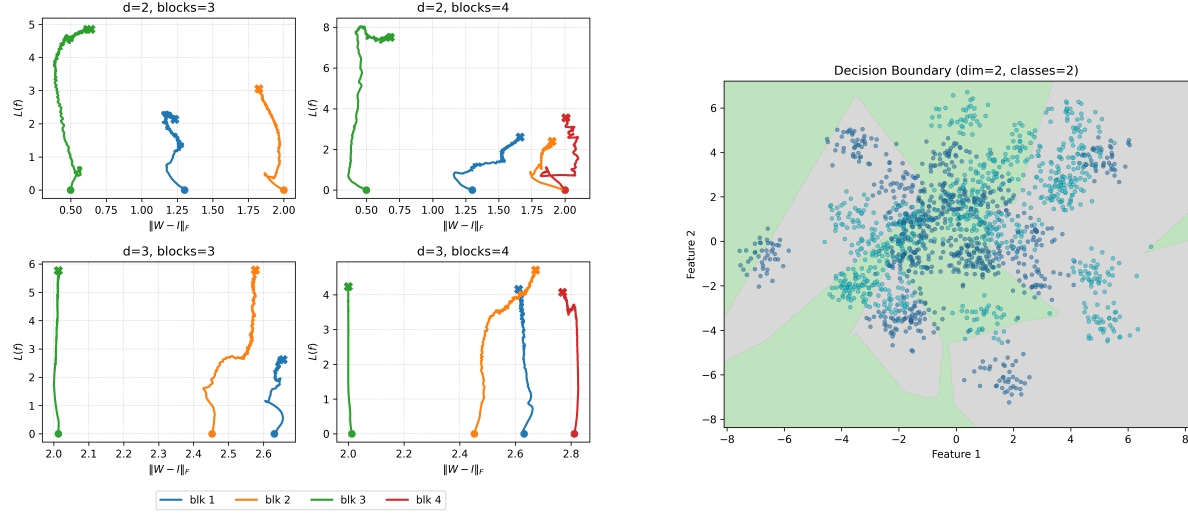


Figure 4: Trajectory of orthogonal skip connections on a synthetic dataset. *Left*: Trajectories of different blocks. *Right*: Example dataset and predictions.

It appears that the main theorem doesn't hold as much

4.3 Different optimizers

References

- [Bartlett et al.(2018)Bartlett, Evans, and Long] Peter L. Bartlett, Steven N. Evans, and Philip M. Long. Representing smooth functions as compositions of near-identity functions with implications for deep network optimization, 2018. URL <https://arxiv.org/abs/1804.05012>.
- [Bernstein(2025)] Jeremy Bernstein. The modula docs, 2025. URL <https://docs.modula.systems/>.
- [Wang et al.(2017)Wang, Xing, Zhang, and Zhang] Jingdong Wang, Yajie Xing, Kexin Zhang, and Cha Zhang. Orthogonal and idempotent transformations for learning deep neural networks, 2017. URL <https://arxiv.org/abs/1707.05974>.