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 it's skip connections (identity matrices) are norm preserving. In this paper, we replace these identity matrices with orthogonal matrices, which are by defintion norm preserving. First, we interestingly show that as the number of layers grows, the orthogonal matrices need to converge to the identity matrix, following the previous formulation of ResNet. Then, we explore the performance of orthogonal skip connections multiple regression tasks, using the recent efficient algorithm for gradient descent on orthogonal matrices [Bernstein(2025)]

Contents

1 Introduction

Our main contribtions:

- 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 replacemenet 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 $h : \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable, invertible map satisfying the following properties:*

1. ***Smoothness:*** *for some $\alpha > 0$ and all $x, y, u \in \mathcal{X}$,*
$$\left\| \big(Dh(y) - Dh(x)\big)u \right\| \leq \alpha \|y - x\| \|u\|; \tag{1}$$
2. ***Lipschitz inverse:*** *for some $M > 0$, $\|h^{-1}\|_L \leq M$;*
3. ***Positive orientation:*** *For some $x_0 \in \mathcal{X}$, $\det\big(Dh(x_0)\big) > 0$.*
4. ***Zero at origin:*** $h(0) = 0$

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

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

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

TODO: We can probably replace the $\log m / m$ ***term by choosing any sequence in*** $(0, 1)$ ***that converges to*** 0

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

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

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

and moreover,

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

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 $i = 1, 2, \ldots, m$, define

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

Each $a_i > 0$ is a scalar parameter with

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

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

$$\begin{aligned}\|W_i - W_{i-1}\| &\le \|I - W_i\| + \|I - W_{i-1}\| && (2)\\ &\le C\|a_i - a_{i-1}\| + C\|a_{i-1} - a_{i-2}\| && (3)\\ &\le C\|a_i - a_{i-2}\| < C\|a_i - a_{i-1}\| && (4)\end{aligned}$$

and

$$\begin{aligned}\|a_iW_i - a_{i-1}W_{i-1}\| &\le \|a_iW_i - a_iW_{i-1}\| + \|a_iW_{i-1} - a_{i-1}W_{i-1}\| && (5)\\ &\le a_i\|W_i - W_{i-1}\| + \|a_i - a_{i-1}\| && (6)\\ &< a_iC\|a_i - a_{i-1}\| + \|a_i - a_{i-1}\| = \|a_i - a_{i-1}\|(Ca_i + 1) && (7)\end{aligned}$$

Then we set

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

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

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

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

We aim to show

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

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

We have

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

Hence

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

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

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

Using Lemma 1, expand

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

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

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

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

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

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 $\|h^{-1}\|_L \le M$, we have

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

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

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

Recall

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

Hence we can rewrite

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

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

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

and expanding further:

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

Thus

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

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

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

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

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

Note that we can bound D by:

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

Combining, we obtain:

$$\begin{aligned}
\big\| \big(h_i(x) - W_i x\big) - \big(h_i(y) - W_i y\big) \big\| &\leq r \;+\; \|u - v\| \Big(\alpha\, \|B - Y\|^2 + J_2 \Big) \leq \\
&\leq \tfrac{\alpha}{2} \left(a_i^2 \|u - v\|^2 + a_{i-1}^2 \|x - y\|^2 \right) + \\
&\|u - v\| \left(\alpha\, \|a_i W_i v - a_{i-1} W_{i-1} v\|^2 + \|a_{i-1} M R\| \left(2\|W_i - W_{i-1}\| + \|I - W_i\| \right) \right) \\
&\leq \tfrac{\alpha}{2} \left((a_i^2 M^2 + a_{i-1}^2) \|x - y\|^2 \right) + \\
&M\, \|x - y\| \left(\alpha\, \|a_i W_i - a_{i-1} W_{i-1}\|\, R^2 + \|a_{i-1} M R\| \left(2\|W_i - W_{i-1}\| + \|I - W_i\| \right) \right) \\
\leq \|x - y\| \left[\tfrac{\alpha}{2} M^2 \|x - y\| + \alpha\, \|a_i W_i - a_{i-1} W_{i-1}\|\, R^2 + \|a_{i-1} M R\| \left(2\|W_i - W_{i-1}\| + \|I - W_i\| \right) \right]&
\end{aligned}$$

using the equations (4) and (7), we can write:

$$\begin{aligned}
&< \|x - y\| \left[\tfrac{\alpha}{2} M^2 \|x - y\| + \alpha (C a_i + 1)(a_i - a_{i-1}) R^2 + \|a_{i-1} M R\| \left(2C(a_i - a_{i-1}) + C(a_i - a_{i-1}) \right) \right] \\
&\leq \|x - y\| \left[\tfrac{\alpha}{2} M^2 \|x - y\| + \alpha (C a_i + 1)(a_i - a_{i-1}) R^2 + 3C\, \|a_{i-1} M R\|\, (a_i - a_{i-1}) \right].
\end{aligned}$$

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

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

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 $S_k(z) = g_k(z) - g_{k-1}(z)$:

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

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

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

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

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

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

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

Now we can bound the Lipschitz constant of S_i:

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

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

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

3.2.4 Combining the estimates

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

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

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 $h_m \circ h_{m-1} \circ \cdots \circ h_1$ 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

Training Cifar10, Cifar100 doesn't show any difference, as seen on Figure 1. However, training on synthetic data shows that orthogonal skip connections are better than residual blocks, as seen on the left plot of Figure 1.

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/`.

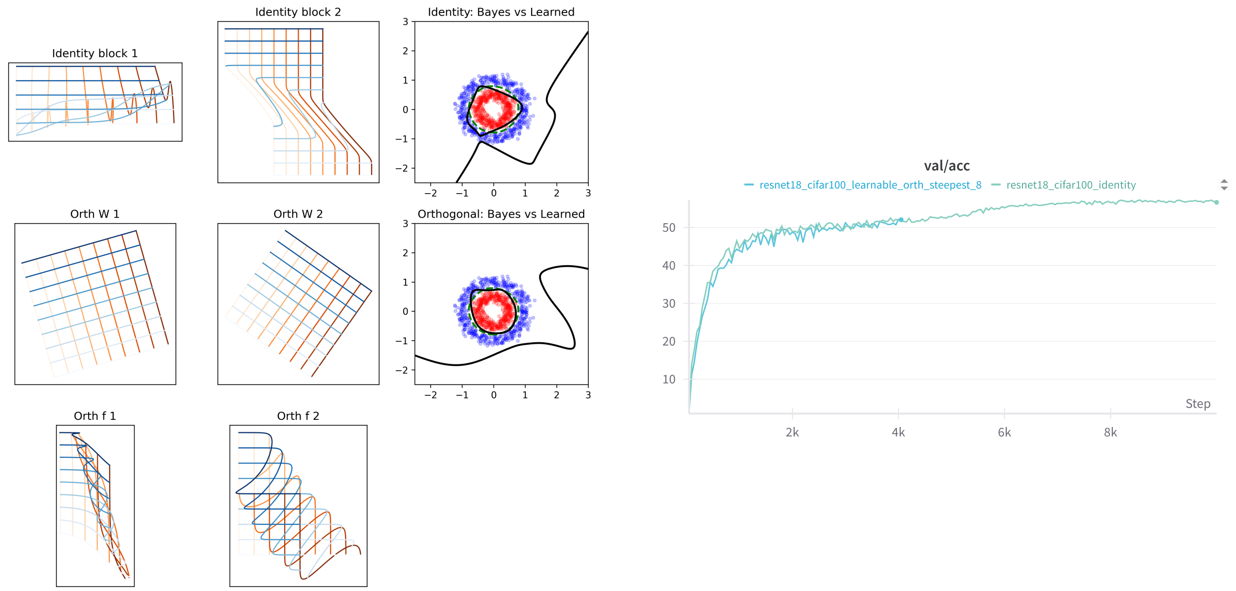

Figure 1: Left: Comparison of residual blocks on synthetic data. Right: Comparison of ResNet-18 variants on CIFAR-100.