Layer-Wise Feedback Alignment is Conserved in Deep Neural Networks

Zachary Robertson Oluwasanmi Koyejo

Computer Science, Stanford

Introduction and Motivation

Problem Statement:

- Backpropagation, the default learning rule for training neural networks, faces criticism for its inefficiency and biological implausibility.
- Feedback Alignment (FA), a biologically inspired learning rule, presents an alternative. The understanding of its dynamics is important in the ongoing research on efficient neural network training.

Contributions:

- We derived novel conservation laws to elucidate the dynamics under feedback alignment.
- Revealed that the conservation laws manifest an implicit bias analogous to gradient descent.
- Presented sufficient conditions that ensure layer-wise alignment in feedback alignment, offering a pathway for efficient learning in neural networks.

Background: Feedback Alignment

Background:

- Feedback Alignment (FA) emerged as a more efficient and biologically plausible alternative to backpropagation.
- Unlike backpropagation, which requires backward pass weights to be the transpose of the feed-forward weights, FA uses fixed random matrices, simplifying the computational process and solving the transport problem in backpropagation.
- In FA, the feedforward computation is used for activations and a feedback pass with a distinct matrix computes errors.
- The fixed random matrix in the backward pass can still guide the network to learn useful representations, even though it does not directly mirror the forward weights.

Feedback Alignment:

- Consider an L-layer neural network with weight matrices $W_l \in \mathbb{R}^{n_{l+1}\times n_l}$ where n_l is the number of neurons in layer l.
- Feedforward: $h_l = W_l a_{l-1}$, $a_l = \phi(h_l)$ where ϕ is a nonlinear activation function
- Feedback: $\delta_l = \phi'(h_l) \odot B_{l+1}\delta_{l+1}$, $\delta_L = \nabla_{a_L}\mathcal{L}(f)$, where B_l are random matrices and $\mathcal{L}(f)$ is the loss function
- Weight update: $\Delta W_l = -\eta \cdot (a_l)^\top \delta_{l+1}^\top$, η is the learning rate.

Layer-Wise Alignment

Key Challenges:

- How to align random feedback weights with the changing forward weights during training?
- How to understand the theoretical reasons behind the empirical alignment observed in prior work?

Our Approach:

- We formulate novel conservation laws that model learning dynamics for ReLU networks.
- Our laws demonstrate that layer-wise alignment emerges from an implicit bias of the learning rule.
- We devise initialization schemes ensuring layer-wise alignment, thus mitigating the first challenge.

Theoretical Results

We provide two main theoretical results regarding layer-wise alignment:

Theorem 1: Suppose that we apply feedback alignment to a scalar output ReLU network with differentiable loss function. Then the flow of the layer weights under feedback alignment for all $t \in \mathbb{R}_{\geq 0}$ maintains,

$$\frac{1}{2}\|W_i(t)\|_F^2 - \langle W_{i+1}(t), B_{i+1}\rangle = \frac{1}{2}\|W_i(0)\|_F^2 - \langle W_{i+1}(0), B_{i+1}\rangle$$

The conservation law implies an implicit bias analogous to gradient descent. If we initialize $W_{i+1}(0) = B_{i+1}$ such that $\|W_i(0)\| \leq \|W_{i+1}(0)\|$ then we guarantee layer-wise alignment. By exploiting the conservation law, we show that these networks are capable of converging to a global optimum.

Theorem 2: Assume that we are to fit data y with squared-loss and an over-parameterized two-layer network $f_{w_t}(X) = Xw_t = XW_1(t)W_2(t)$ with data X such that rows are linearly-independent. Assume we may pick $w_0 \in \text{span}(X^T)$ such that we have positive alignment for all time. If we run (direct) feedback alignment flow then we have the following,

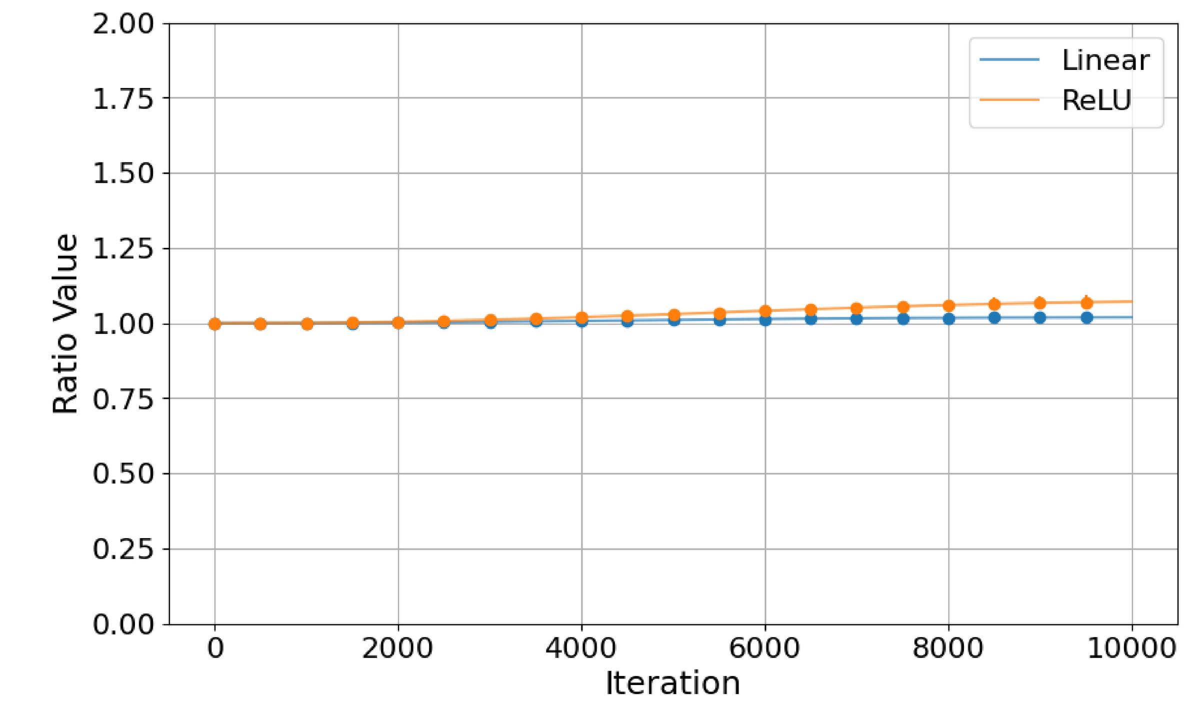
$$\lim_{t\to\infty} e^{rt} \cdot \|y - Xw_t\|_2 \to 0$$

for some $r > 0$. Moreover, $w_\infty = W_1(\infty)W_2(\infty)$ is the minimum-norm solution.

These results collectively suggest that Feedback Alignment exhibits an implicit regularization effect, guiding towards solutions that generalize similarly to gradient descent in the over-parameterized regime.

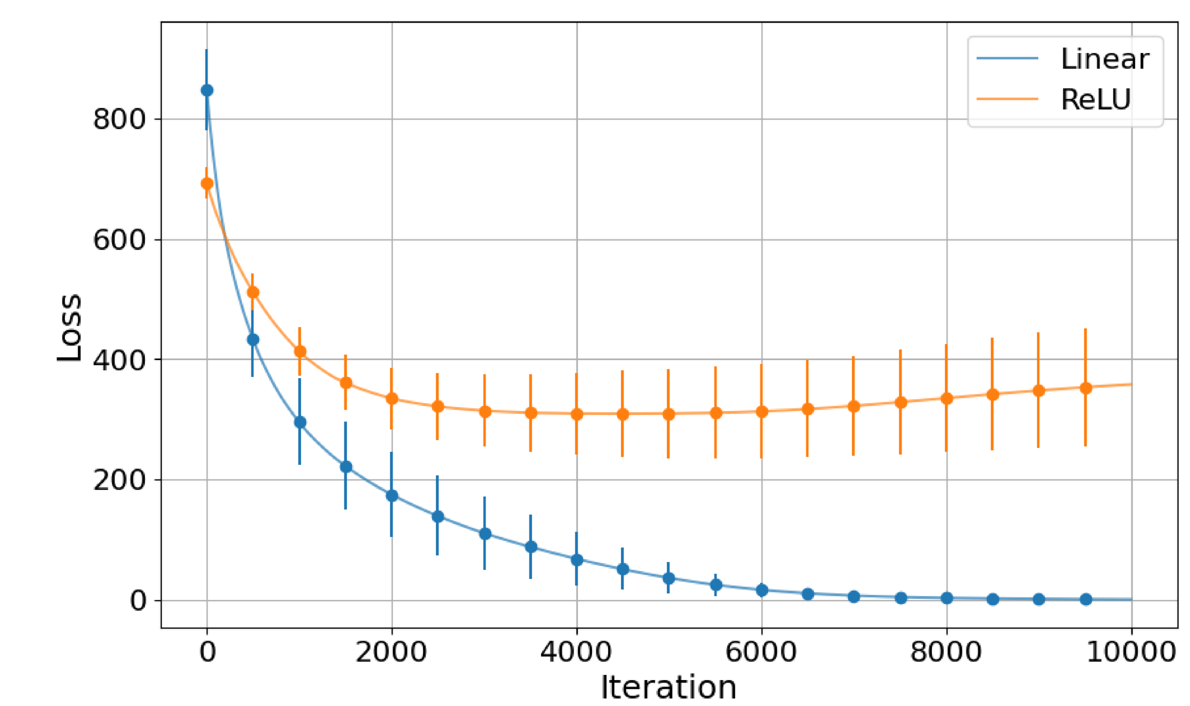
Experimental Results

Conservation of Alignment:



- Track conserved quantity from Theorem 1
- Remains nearly invariant in linear and ReLU nets
- Provides empirical evidence for theoretical results

Convergence Analysis:



- Compare FA-learned weights to minimum norm solution
- Verify weights converge to global optimum in 2-layer linear nets
- Aligns with prediction from conservation laws

Conclusions:

- Conservation laws hold empirically
- FA provably converges in overparameterized linear nets
- Results relate FA dynamics to gradient descent
- Further work needed extending to nonlinear nets