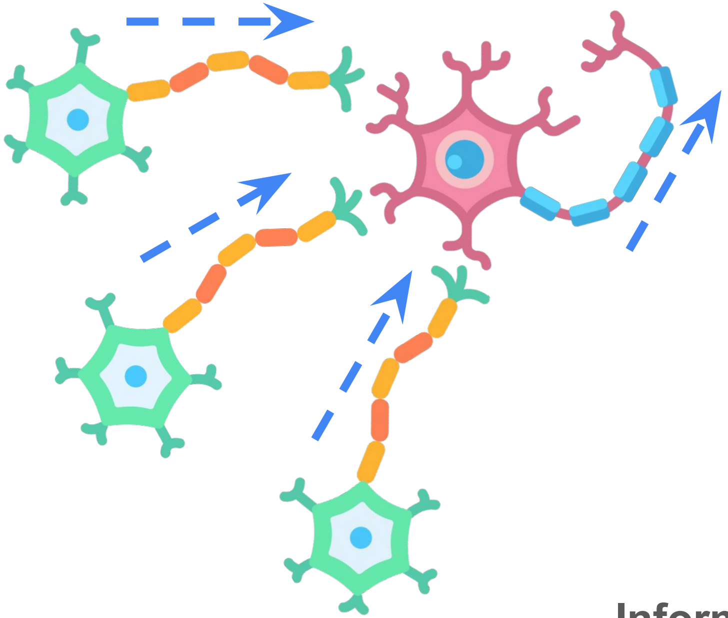
How to Train Your Wide Neural Network Without Backprop: An Input-Weight Alignment Perspective

Akhilan Boopathy, Ila Fiete

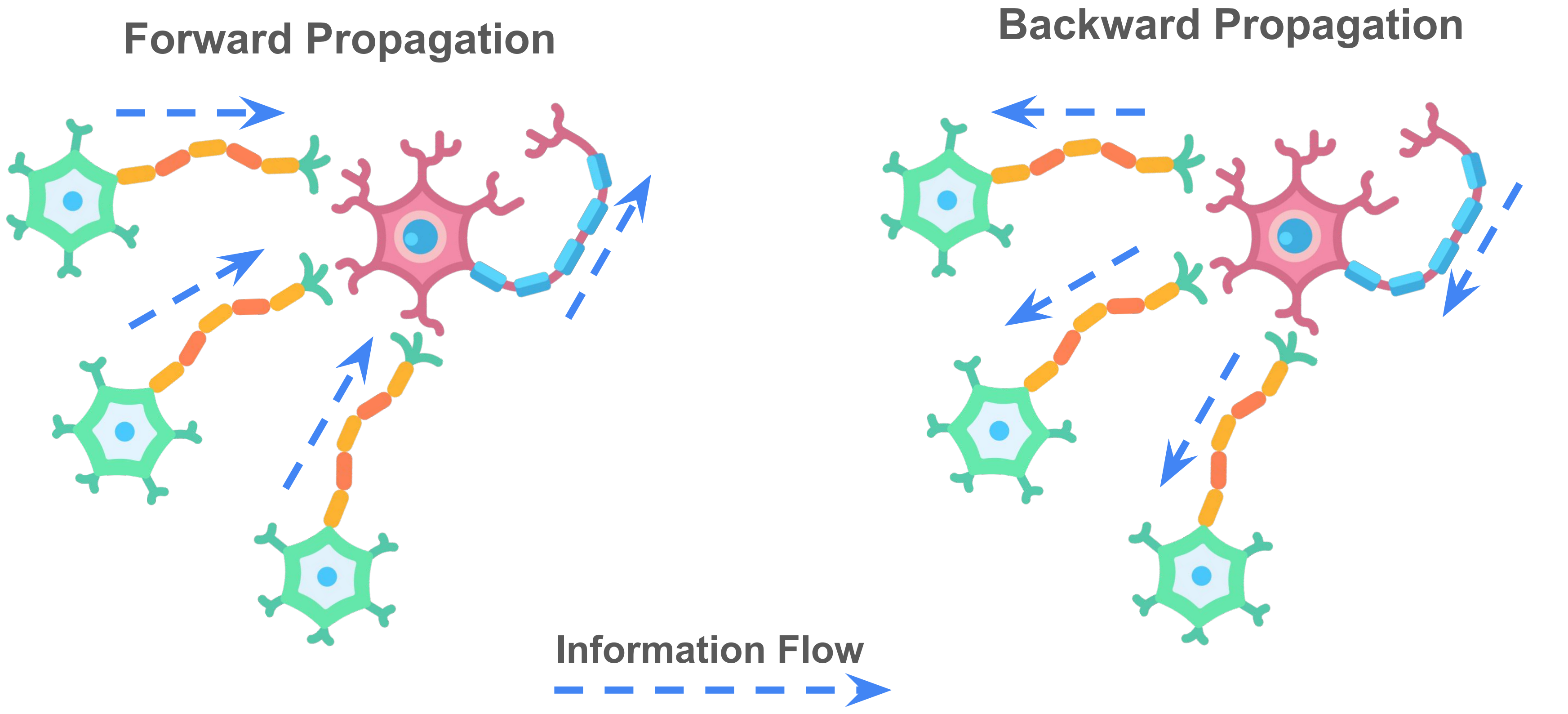
Backpropagation is biologically difficult to implement

Forward Propagation



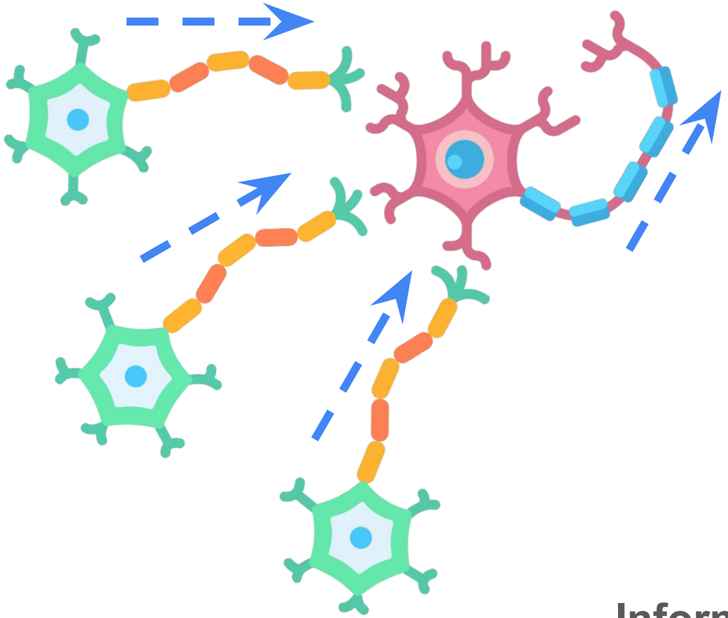
Information Flow

Backpropagation is biologically difficult to implement

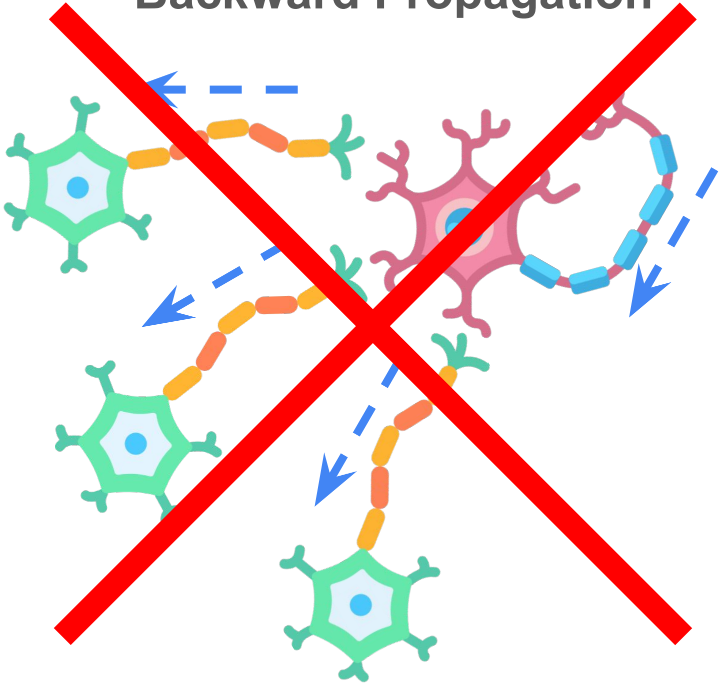


Backpropagation is biologically difficult to implement

Forward Propagation



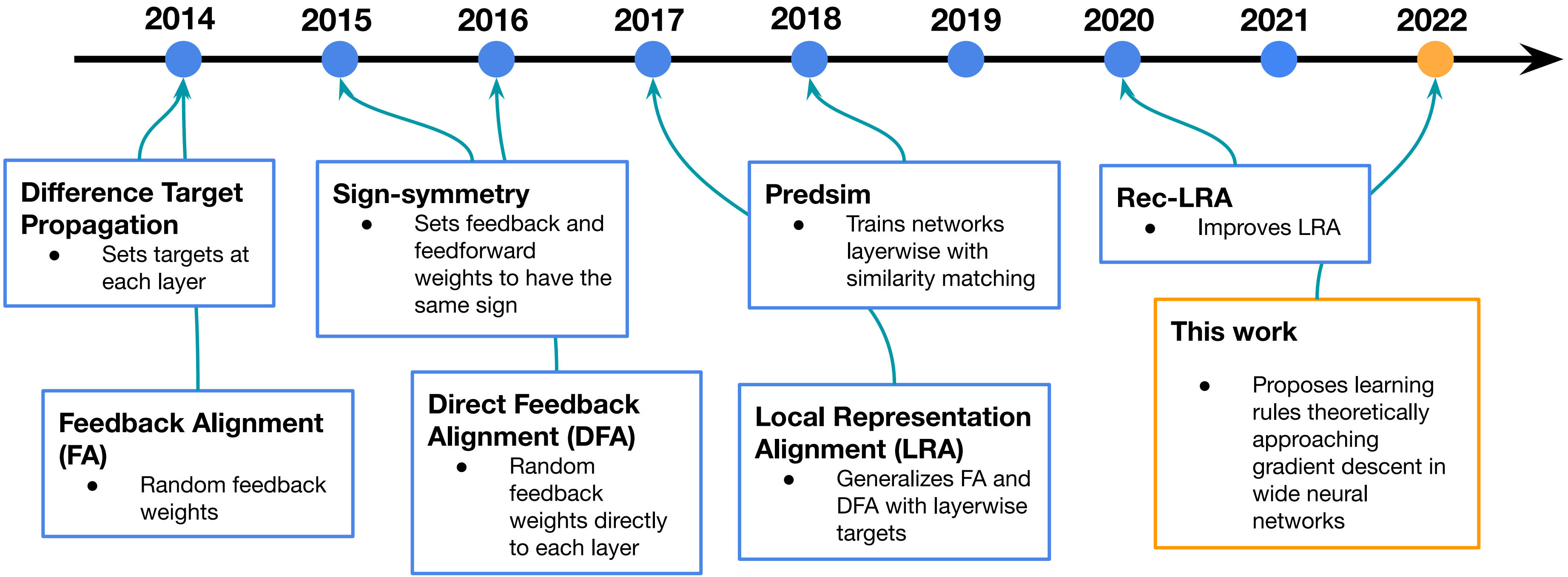
Backward Propagation



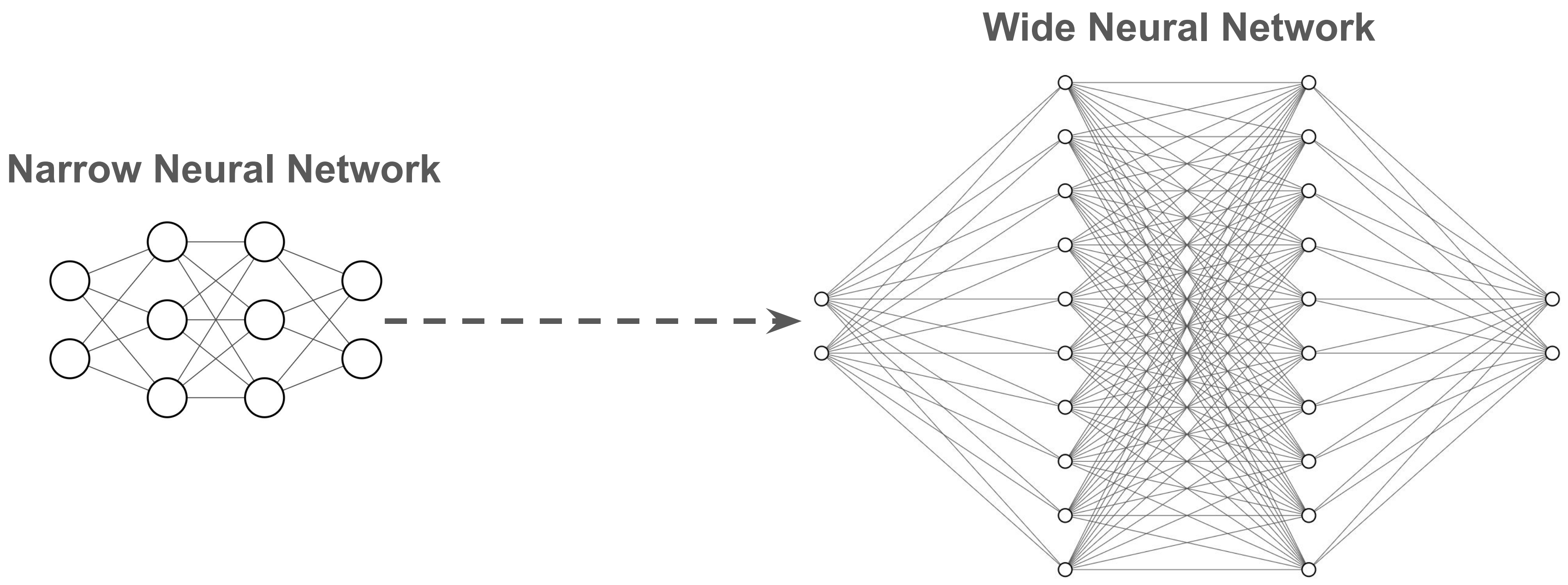
Information Flow

Implausible…

Many biologically-motivated learning rules have been proposed

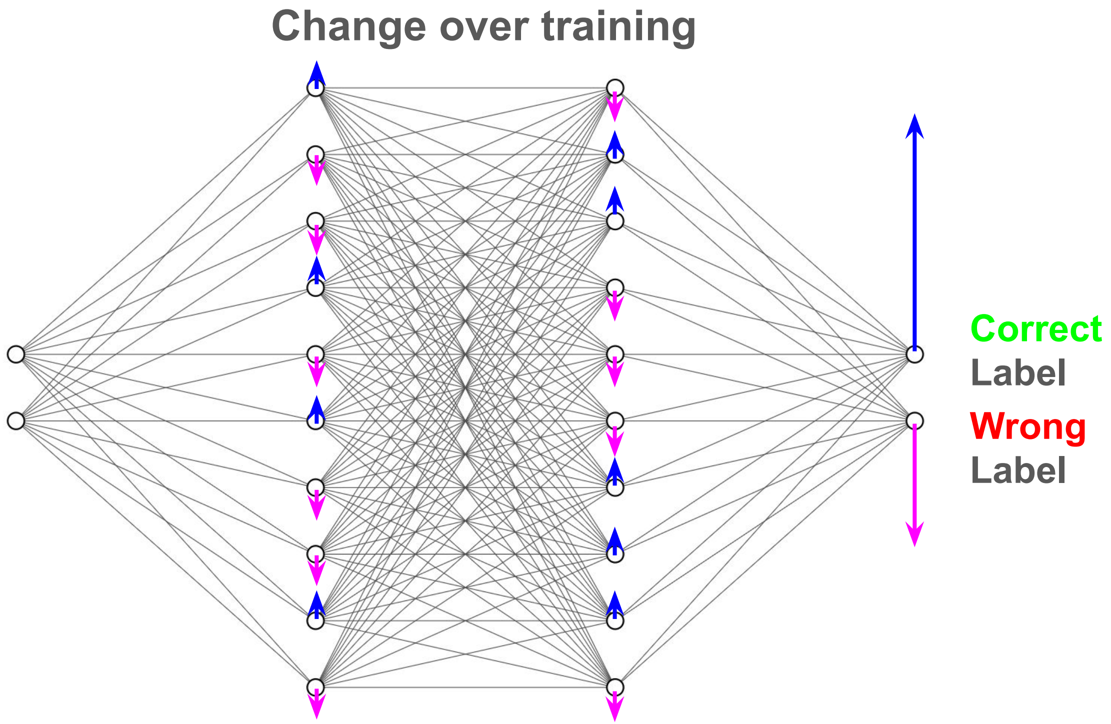


Neural Tangent Kernel (NTK) theory allows for theoretical analysis of infinite width neural networks



Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. NeurIPS, 2018.

Wide neural networks weights and activations move little during training

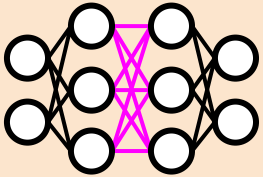


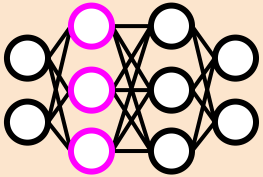
Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. NeurIPS, 2018.

Weights of wide neural networks are *aligned* to simple statistics of network inputs

Proposition 1 (informal)

As network width $\longrightarrow \infty$:





$$\underbrace{(\overbrace{W^{(t)} - W^{(0)}}^{\text{Weight change}})^T (W^{(t)} - W^{(0)})}_{\text{Weight change correlation}} \propto \underbrace{\mathbb{E}[\overbrace{\sigma(z^{(0)}(x_1))}^{\text{Layer value at initialization}} q(x_1, x_2) \sigma(z^{(0)}(x_2))^T]}_{\text{Layer input correlation}}$$

Alignment at each layer can be quantified using an alignment score

- Alignment score is cosine distance between weight correlation and data correlation

$$\frac{tr(\Lambda_l \Sigma_l)}{\sqrt{tr(\Lambda_l^2)tr(\Sigma_l^2)}}$$

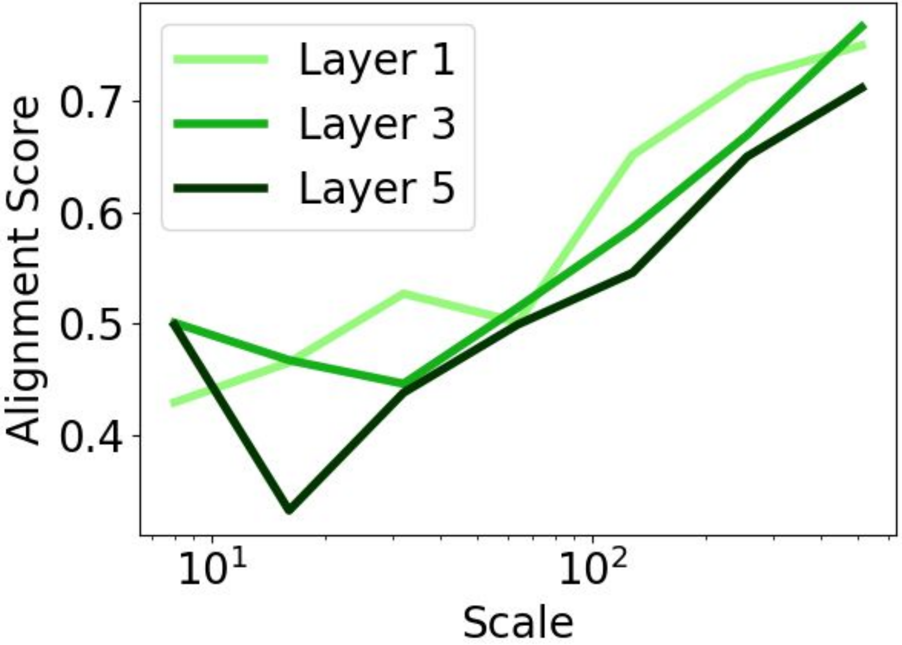
Alignment Score

Weight change correlation $\left\{ \Lambda_l = (W_l^{(t)} - W_l^{(0)})^T (W_l^{(t)} - W_l^{(0)}) \right.$

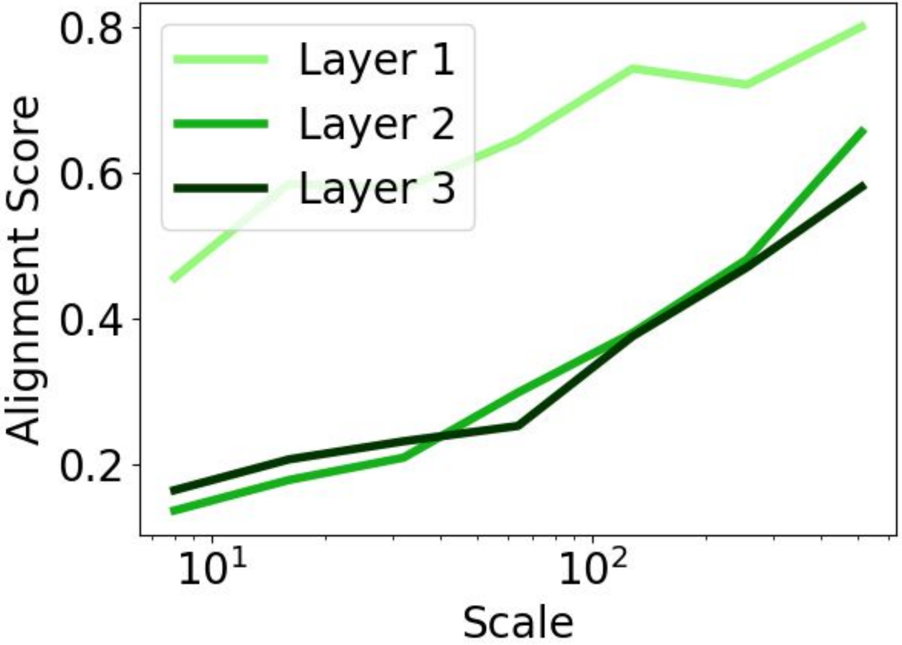
Layer input correlation $\left\{ \Sigma_l = \mathbb{E}[\sigma(z_{l-1}^{(0)}(x_1))q(x_1, x_2)\sigma(z_{l-1}^{(0)}(x_2))^T] \right.$

Wide, finite-width networks exhibit high alignment

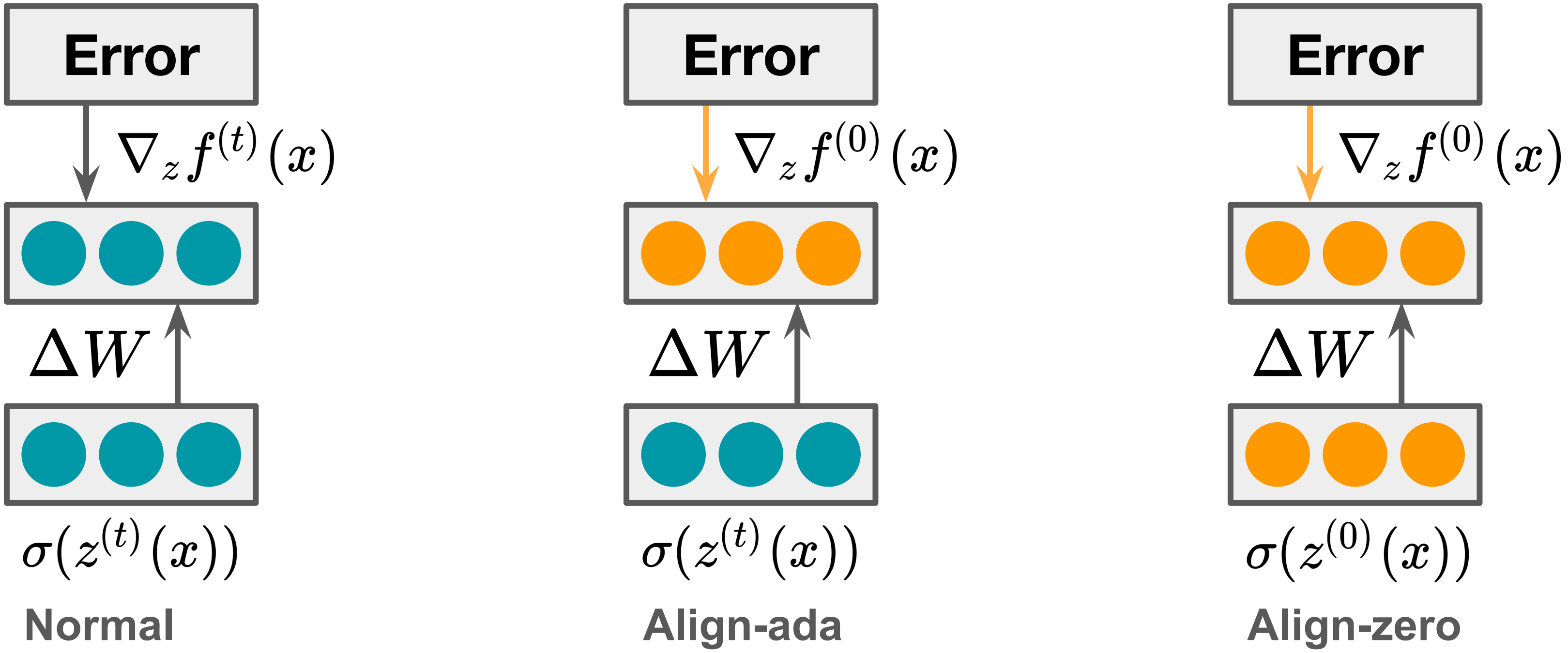
CIFAR-10



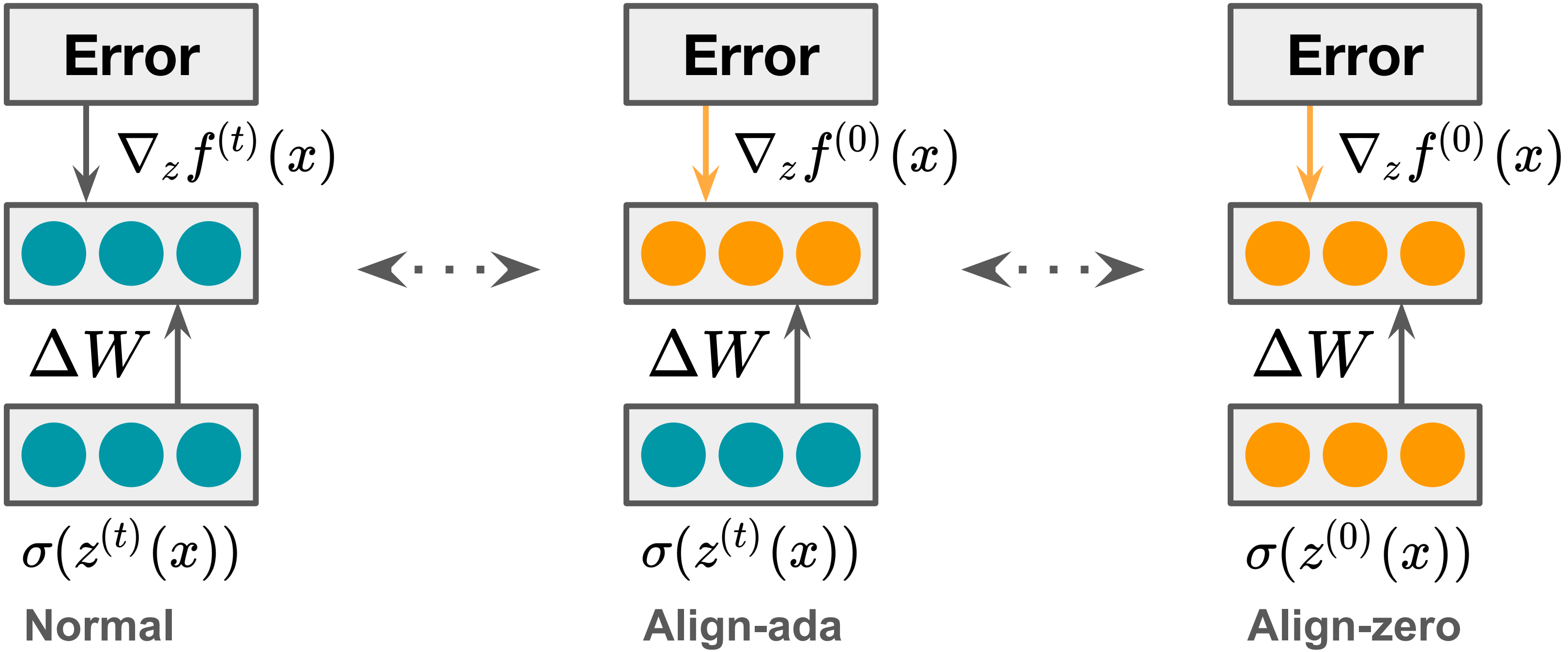
KMNIST



Simplified learning rules are equivalent to backpropagation in wide neural networks



Simplified learning rules are equivalent to backpropagation in wide neural networks



Simplified learning rules are equivalent to backpropagation in wide neural networks

Proposition 2 (informal)

Normal: $\dot{W}_l^{(t)} = \frac{\eta}{\sqrt{m_{l-1}}} \times \mathbb{E}_{p_x}[\nabla_{z_l} f^{(t)}(x) \nabla_{z_N} L(f^{(t)}(x), y(x)) \sigma(z_{l-1}^{(t)}(x))^T]$

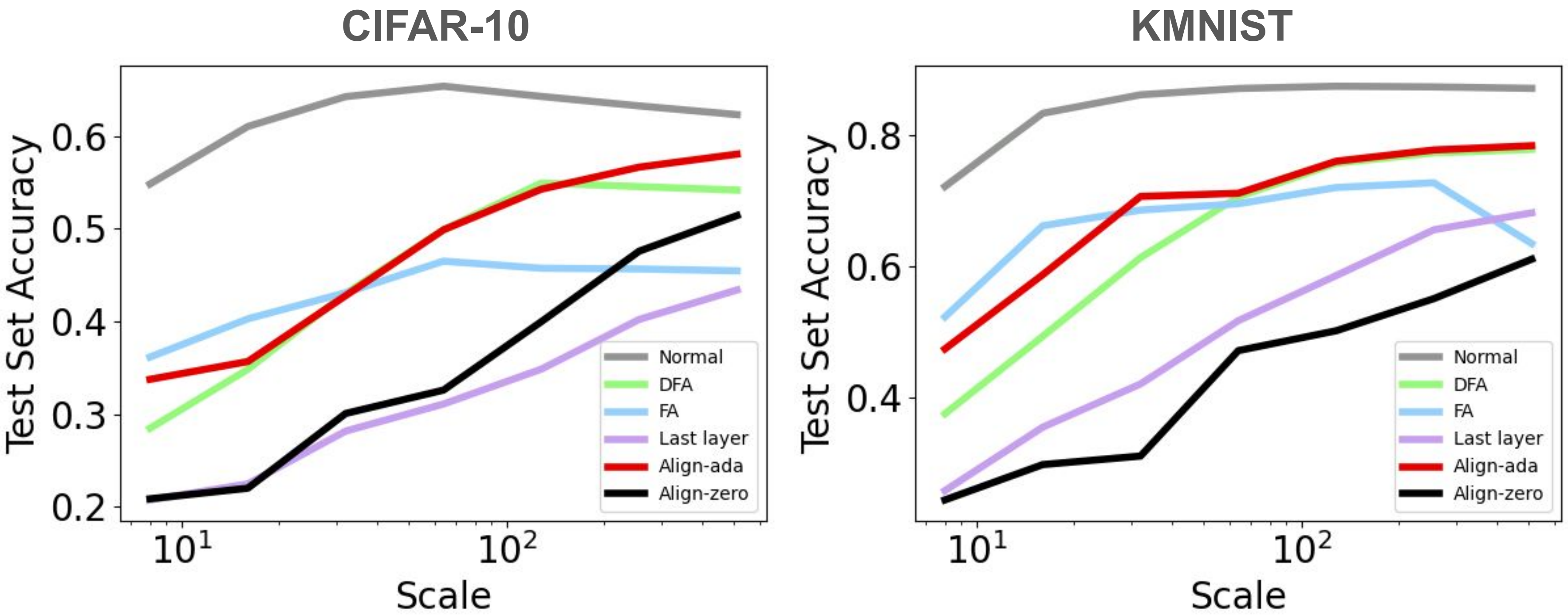
Align-zero: $\dot{W}_{l,al\cdot 0}^{(t)} = \frac{\eta}{\sqrt{m_{l-1}}} \times \mathbb{E}_{p_x}[\nabla_{z_l} f^{(0)}(x) \nabla_{z_N} L(f_{al\cdot 0}^{(t)}(x), y(x)) \sigma(z_{l-1}^{(0)}(x))^T]$

Align-ada: $\dot{W}_{l,al\cdot ada}^{(t)} = \frac{\eta}{\sqrt{m_{l-1}}} \times \mathbb{E}_{p_x}[\nabla_{z_l} f^{(0)}(x) \nabla_{z_N} L(f_{al\cdot ada}^{(t)}(x), y(x)) \sigma(z_{l-1}^{(t)}(x))^T]$

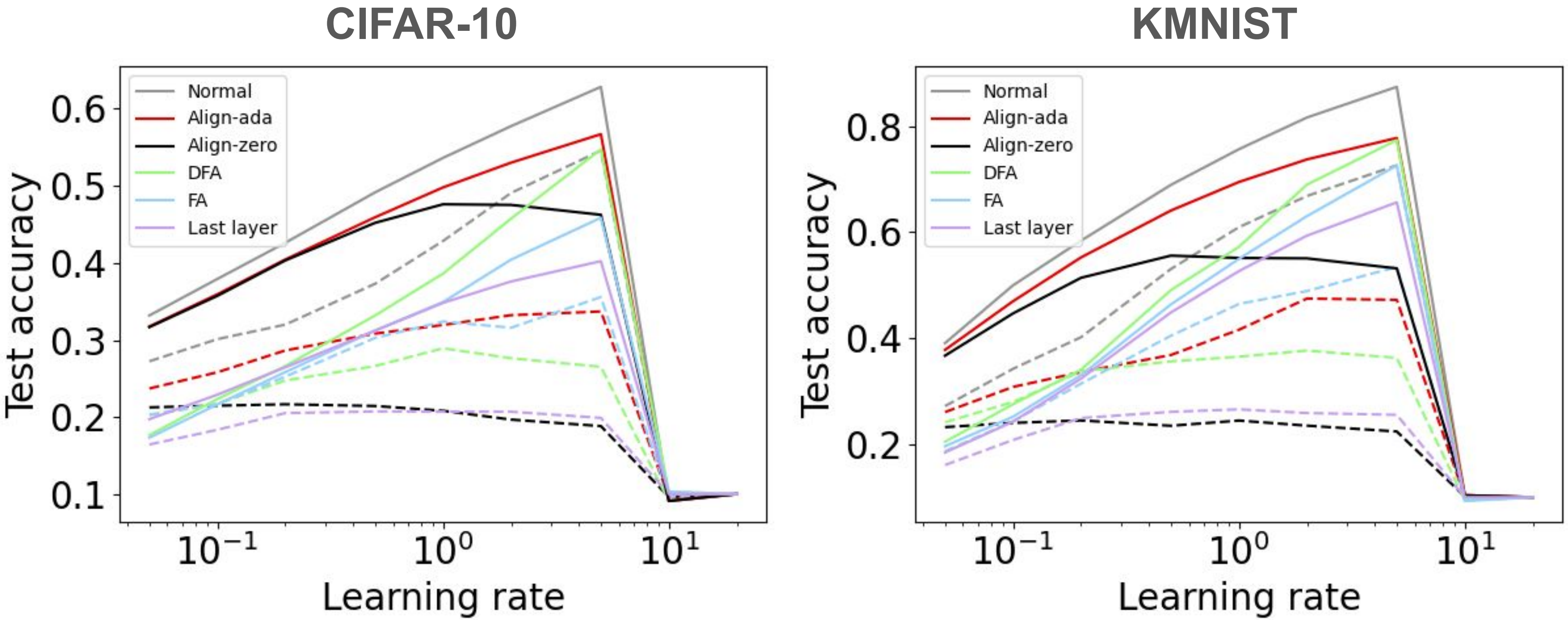
As network width $\longrightarrow \infty$ **:**

$$f^{(t)}(x) = f_{al\cdot 0}^{(t)}(x) = f_{al\cdot ada}^{(t)}(x)$$

Align methods approach performance of backprop on wide networks

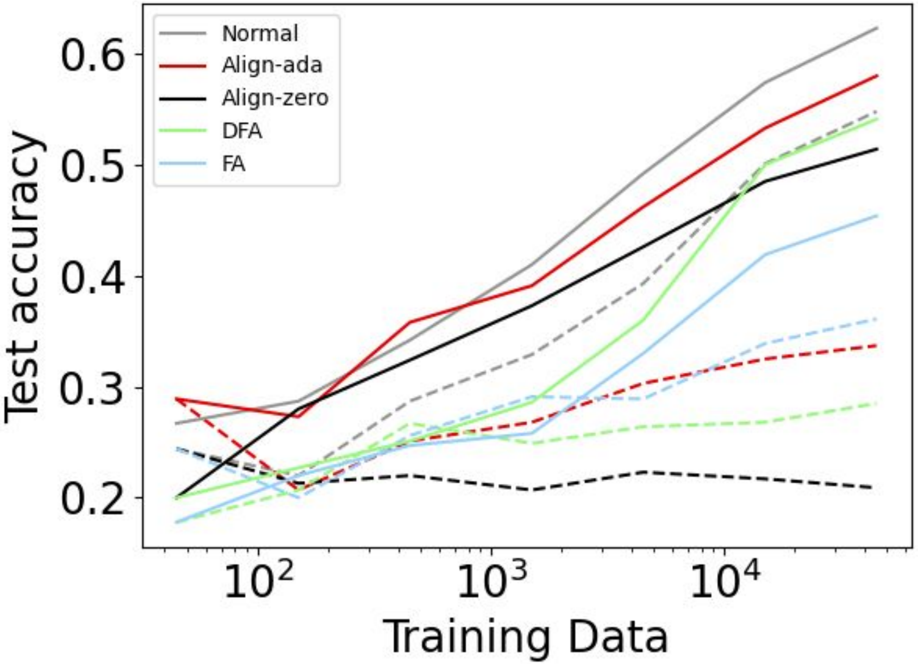


Align methods closely match backprop on wide networks at small learning rates



Align methods are particularly advantageous in low data regimes

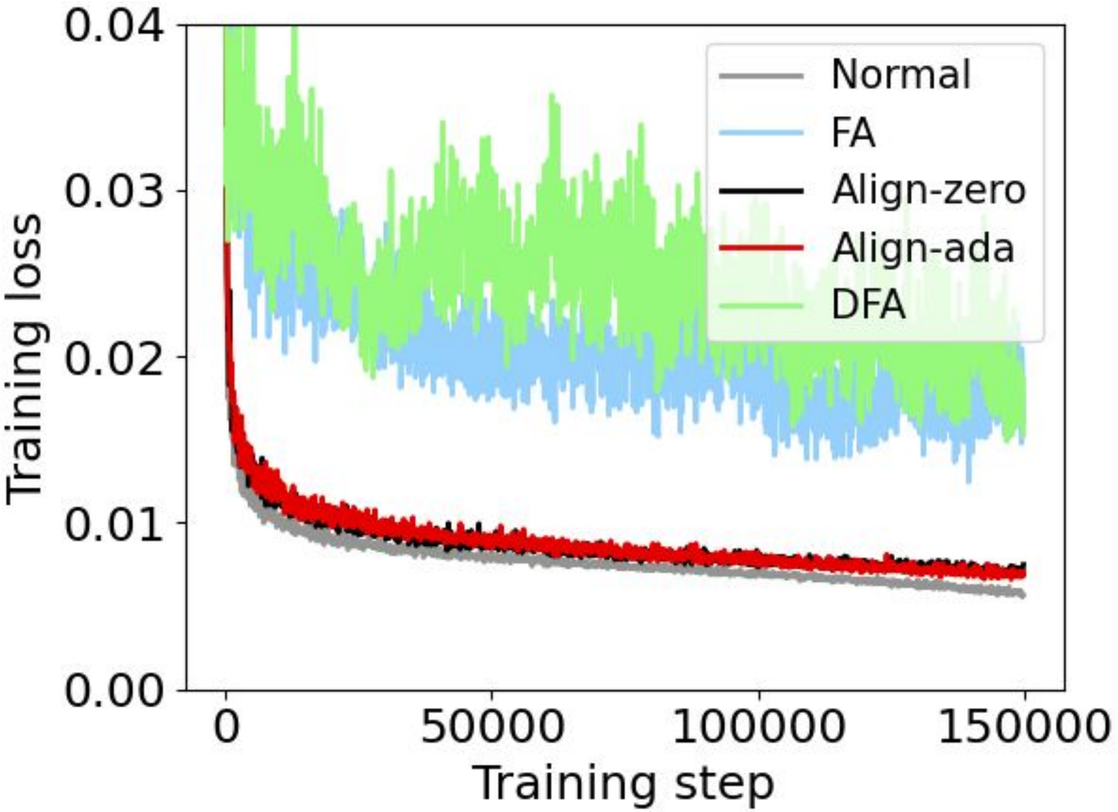
CIFAR-10



Wide network

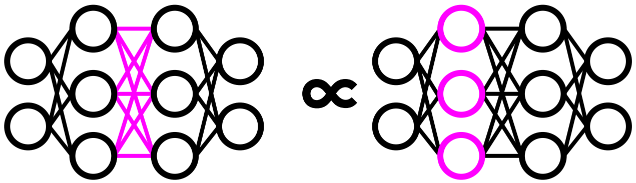
Narrow network

Align methods closely match backprop on ImageNet

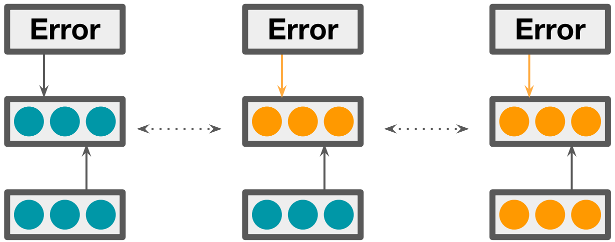


Conclusion

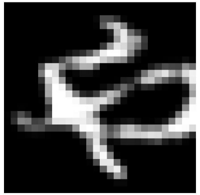
1) Wide neural network weights capture simple layerwise statistics of their inputs



2) Simplified Align learning rules are equivalent to backprop in infinite width networks



3) Empirically, Align rules approach the performance of backprop on wide, finite width networks in the following settings:



KMNIST

CIFAR-10

Low data CIFAR-10

ImageNet