Feature Learning and Signal Propagation in DNNs

Y. Lou, C. Mingard, S. Hayou

Department of Statistics, University of Oxford
Department of Mathematics, National University of Singapore

Goal

Given an input $x \in \mathbb{R}^d$, the network is described by the equations

$$z_l(x) = \mathcal{F}_l(\theta_l, z_{l-1}(x)), \quad 1 \leq l \leq L,$$

where $\mathcal{F}_l$ is a mapping that defines the l^{th} layer, e.g. fully-connected, convolutional, etc.

Goal

Given an input $x \in \mathbb{R}^d$, the network is described by the equations

$$z_l(x) = \mathcal{F}_l(\theta_l, z_{l-1}(x)), \quad 1 \leq l \leq L,$$

where $\mathcal{F}_l$ is a mapping that defines the l^{th} layer, e.g. fully-connected, convolutional, etc.

- How does depth in neural networks affect feature learning?

Goal

Given an input $x \in \mathbb{R}^d$, the network is described by the equations

$$z_l(x) = \mathcal{F}_l(\theta_l, z_{l-1}(x)), \quad 1 \leq l \leq L,$$

where $\mathcal{F}_l$ is a mapping that defines the l^{th} layer, e.g. fully-connected, convolutional, etc.

- How does depth in neural networks affect feature learning?
- But first, how do we measure feature learning?

Tangent Features

Let f_θ denote the network output with with weights θ. The Neural Tangent Kernel (NTK, Jacot et al. (2018)) is given by

$$K_\theta^L(x, x') = \nabla_\theta f_\theta(x) \nabla_\theta f_\theta(x')^T \in \mathbb{R}^{o \times o}. \tag{1}$$

Tangent Features

Let f_θ denote the network output with with weights θ. The Neural Tangent Kernel (NTK, Jacot et al. (2018)) is given by

$$K_\theta^L(x, x') = \nabla_\theta f_\theta(x) \nabla_\theta f_\theta(x')^T \in \mathbb{R}^{o \times o}. \tag{1}$$

The tangent features are the feature maps of the NTK, given by the output gradients w.r.t the network parameters, namely

$$\Psi_\theta(x) = \nabla_\theta f_\theta(x)^T \in \mathbb{R}^{P \times o}. \tag{2}$$

Centered Kernel Alignment

The *centered kernel alignment* between two kernel matrices $\mathbf{K}, \mathbf{K}' \in \mathbb{R}^{on \times on}$ is defined by

$$A\left(\mathbf{K}, \mathbf{K}'\right) = \frac{\mathrm{Tr}\left[\mathbf{K}_c \mathbf{K}'_c\right]}{\left\|\mathbf{K}_c\right\|_F \left\|\mathbf{K}'_c\right\|_F} \tag{3}$$

where $\mathbf{K}_c = \mathbf{CKC}$, $\mathbf{C} = \mathbf{I} - \frac{1}{on}\mathbf{1}\mathbf{1}^T$ is the centering matrix ($\mathbf{1}$ is a vector with all entries being 1), and $\|.\|_F$ is the Frobenius norm. The CKA was used by (Baratin et al. 2021) as a measure of feature learning.

Alignment Hierarchy

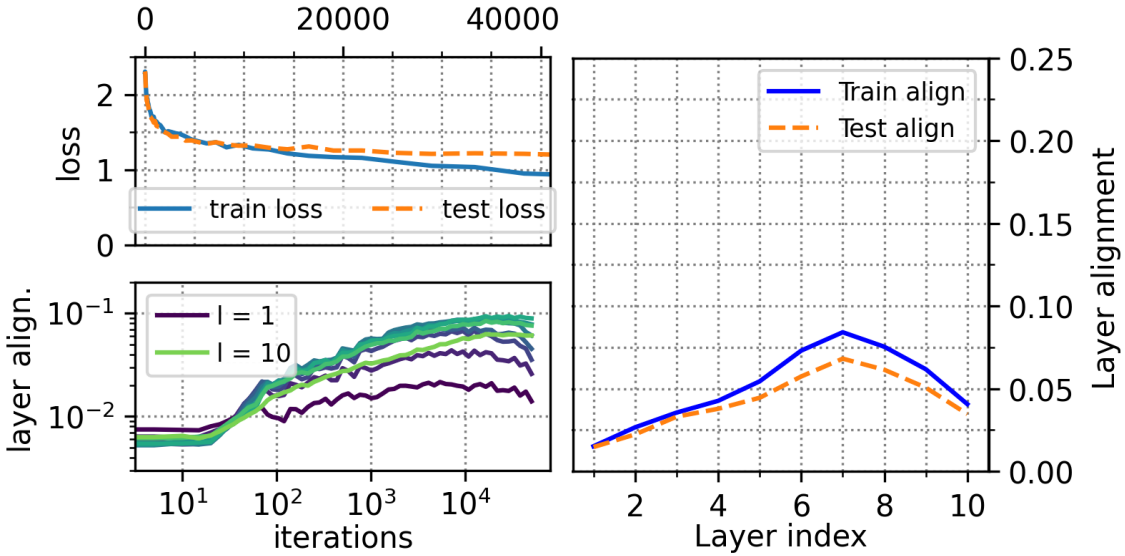


Figure: (CIFAR10)Layerwise alignment hierarchy for the MNIST and CIFAR10 datasets when trained on an FFNN with depth 10 and width 256. Left hand panels show progression of loss and layer alignment with iterations of SGD. Right hand panel shows layer alignment at the end of training.

Alignment Hierarchy

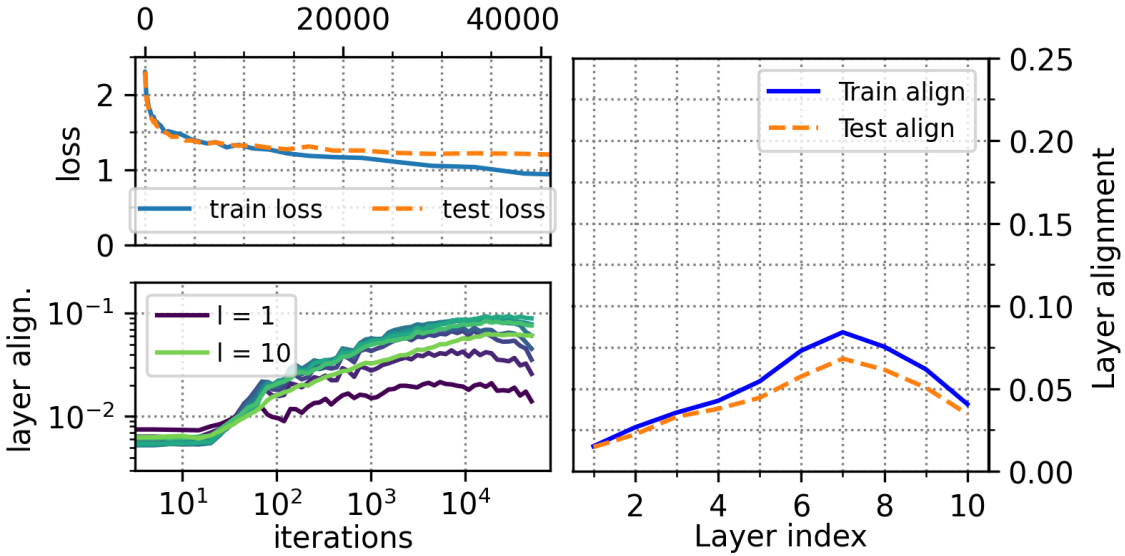


Figure: (CIFAR10)Layerwise alignment hierarchy for the MNIST and CIFAR10 datasets when trained on an FFNN with depth 10 and width 256. Left hand panels show progression of loss and layer alignment with iterations of SGD. Right hand panel shows layer alignment at the end of training.

How can we explain this pattern?

Tangent Kernel Decomposition

The tangent kernel at hidden layer l can be expressed as

$$K_l(x, x') = \nabla_{\theta_l} f(x) \cdot \nabla_{\theta_l} f(x') = \sum_{i,j} \phi(z^j_{l-1}(x))\phi(z^j_{l-1}(x')) \frac{\partial f}{\partial z^i_l}(x) \frac{\partial f}{\partial z^i_l}(x').$$

In matrix form, $\bar{K}_l$ can be written as the Hadamard product of two kernels

$$\bar{\mathbf{K}}_l \propto \overrightarrow{\mathbf{K}}_l \circ \overleftarrow{\mathbf{K}}_l, \tag{4}$$

where

- $\overrightarrow{K}_l(x, x') = \frac{1}{N}\phi(z_{l-1}(x)) \cdot \phi(z_{l-1}(x'))$ is the *forward* features kernel
- $\overleftarrow{K}_l(x, x') = \frac{1}{N}\frac{\partial f_{l:L}}{\partial z}(z_l(x)) \cdot \frac{\partial f_{l:L}}{\partial z}(z_l(x'))$ is the *backward* tangent features kernel, where $f_{l:L}$ is the function that maps the l^{th} layer to the network output.

The Equilibrium Hypothesis

The Equilibrium Hypothesis (EH) (Informal). *The layers with the highest alignments with data labels are the ones that satisfy an equilibrium between forward information and backward information.*

EH for FC nets

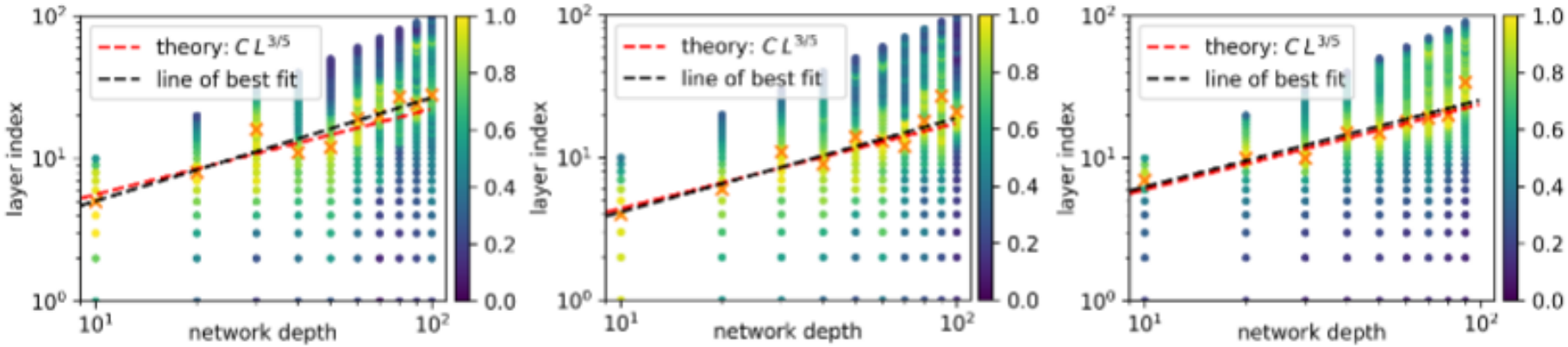
Theorem (Equilibrium)

The equilibrium for a fully connected NN is achieved for layers with index

$$l = \Theta_L(L^{3/5})$$

where L is the network depth.

EH for FC nets



Data with $x = 10j$ in the plot corresponds to layer alignments for a FFNN with depth $10j$ trained on the MNIST/FashionMNIST/CIFAR10 datasets. The brighter the color, the closer the corresponding layer's alignment is to the maximum alignment across all layers. x indicates the layer where largest alignment occurs.

More theoretical and empirical details are provided in the paper! Check it out by scanning the QR code below! Thanks.