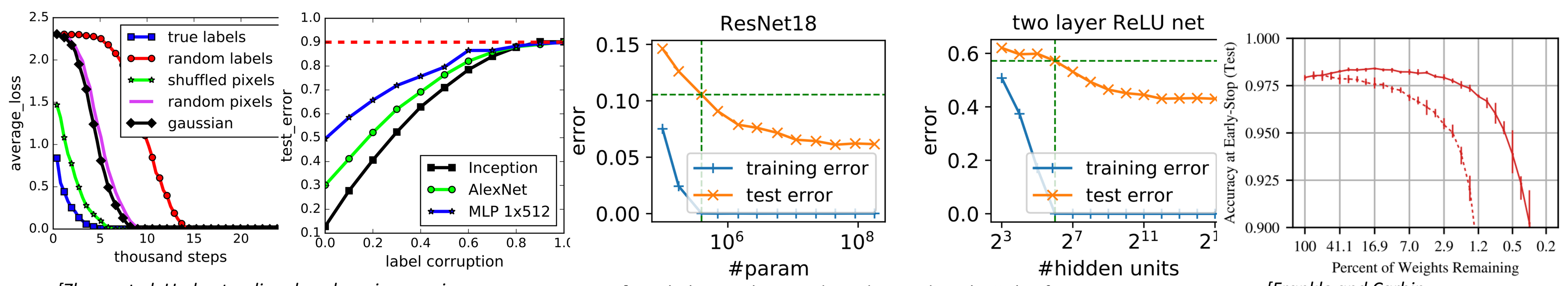
Luck Matters: Understanding the Dynamics of Training Deep ReLU Neural Networks

Yuandong Tian, Tina Jiang, Qucheng Gong, Ari Morcos

https://github.com/facebookresearch/luckmatters

facebook Artificial Intelligence Research

Motivation



[Zhang et al. Understanding deep learning requires rethinking generalization, ICLR'17] [Neyshabur et al. Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks, ICLR'19] [Frankle and Carbin. The Lottery Ticket Hypothesis, ICLR'19]

Over-Parameterization
- More parameters, better test performance.
- Network can be pruned substantially
- Training with models with intrinsic capacity gives poor performance.

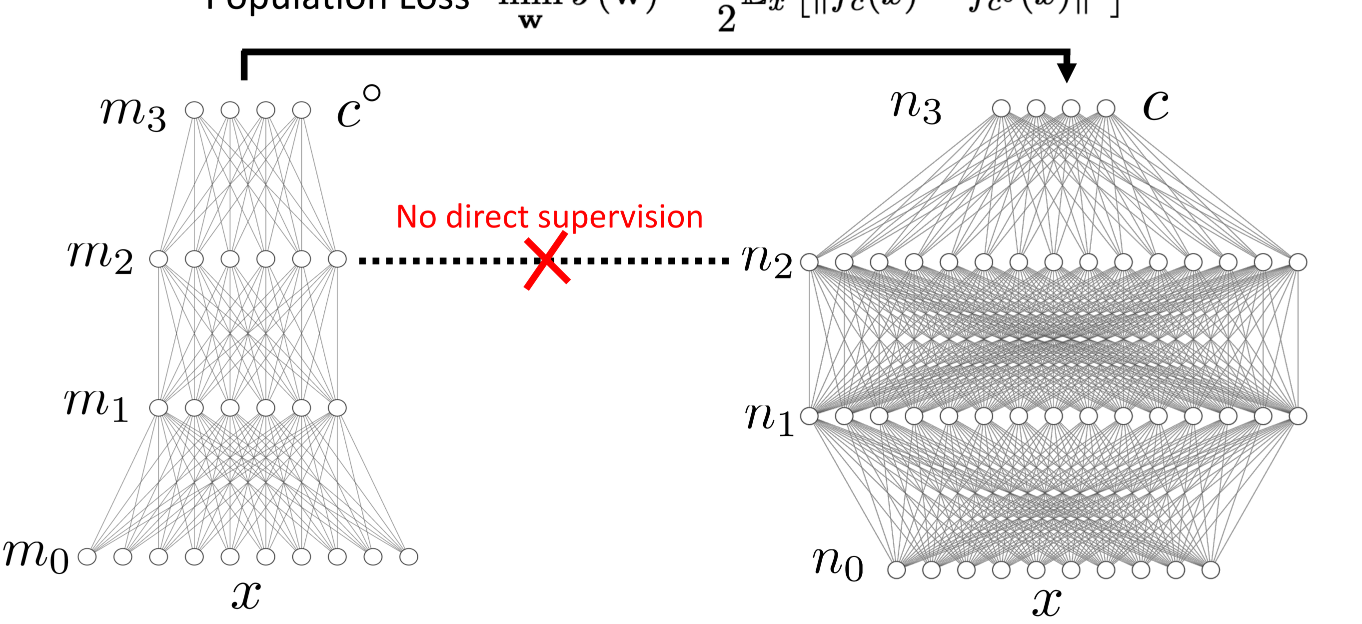
Lottery tickets
- Use salient weights restarted to initialization gives lower test error
- Use salient weights reinitialized gives poor performance.

Implicit Regularization
- Same network trained with SGD can fit both random and structured data
- Network generalizes on structured data

Flat Minima
- Many small eigenvalues in Hessian after convergence

Teacher-student setting

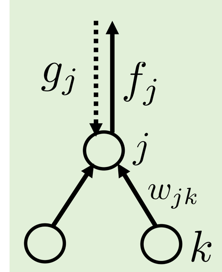
Population Loss $\min_{\mathbf{w}} J(\mathbf{w}) = \frac{1}{2}\mathbb{E}_x\left[\|f_c(x) - f_{c^\circ}(x)\|^2\right]$



Teacher Network **(Fixed parameters)**

(Over-parameterized) Student Network **(Learnable Parameters)**

Notation



$f_j(x)$ Activation

$f'_j(x)$ Gating

$g_j(x)$ Gradient

Gradient Update: $\dot{w}_{jk} = \mathbb{E}_x[g_j f_k]$

Recursive Gradient Rule

For the top-layer we have: $g_c(x) = f_{c^\circ}(x) - f_c(x)$

Is this condition apply to lower layers?

Base case:
$$\beta^*_{cc^\circ}(x) = \delta(c - c^\circ)$$
$$\beta_{cc'}(x) = \delta(c - c')$$

Theorem 1 Assuming for every node *j* in a layer, the gradient is:

$$g_j(x) = f'_j(x)\left[\sum_{j^\circ}\beta^*_{jj^\circ}(x)f_{j^\circ}(x) - \sum_{j'}\beta_{jj'}(x)f_{j'}(x)\right]$$

$\beta^*_{kk^\circ}(x)$ Compatibility between teacher k^0 and student k

Then for the lower layer we have the same form with

$$\beta^*_{kk^\circ}(x) \equiv \sum_{jj^\circ} w_{jk}f'_j(x)\beta^*_{jj^\circ}(x)f'_{j^\circ}(x)w^*_{j^\circ k^\circ} \qquad \beta_{kk'}(x) \equiv \sum_{jj'} w_{jk}f'_j(x)\beta_{jj'}(x)f'_{j'}(x)w_{j'k'}$$

Matrix Form of Gradient Descent

Student intermediate nodes mimics teacher

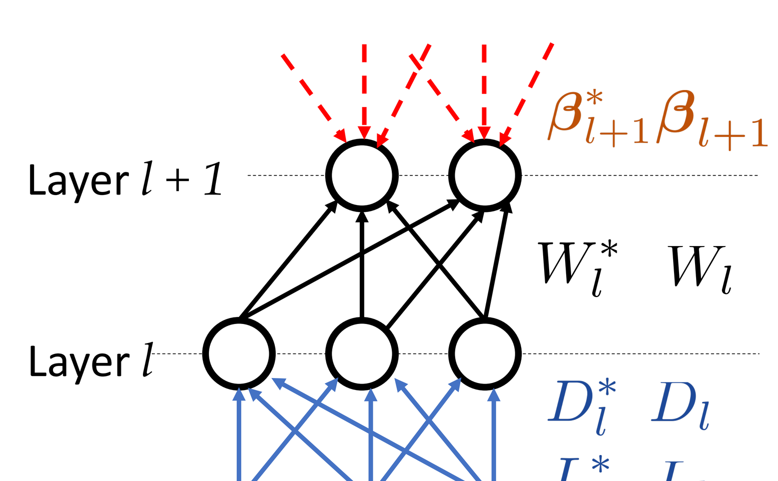
$$\dot{W}_l = L^*_l W^*_l H^*_{l+1} - L_l W_l H_{l+1}$$

$[L^*]_{jj^\circ} = l^*_{jj^\circ} = \mathbb{E}_x[f_j(x)f_{j^\circ}(x)]$ $\quad [L]_{jj'} = l_{jj'} = \mathbb{E}_x[f_j(x)f_{j'}(x)]$

$[D^*]_{jj^\circ} = d^*_{jj^\circ} = \mathbb{E}_x[f'_j(x)f'_{j^\circ}(x)]$ $\quad [D]_{jj'} = d_{jj'} = \mathbb{E}_x[f'_j(x)f'_{j'}(x)]$

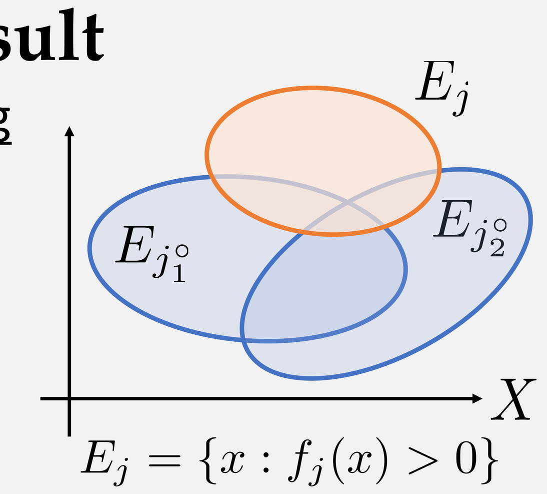
$[\beta^*]_{jj^\circ} = \mathbb{E}_x[\beta^*_{jj^\circ}(x)]$ $\quad [\beta]_{jj'} = \mathbb{E}_x[\beta_{jj'}(x)]$

$H^* = D^* \circ \beta^*$ $\quad H = D \circ \beta$



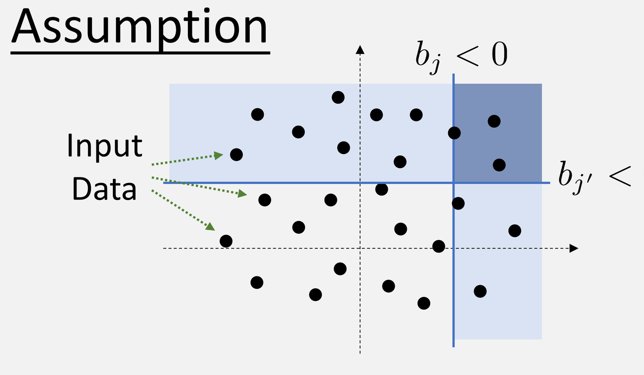
Main Result

Setting



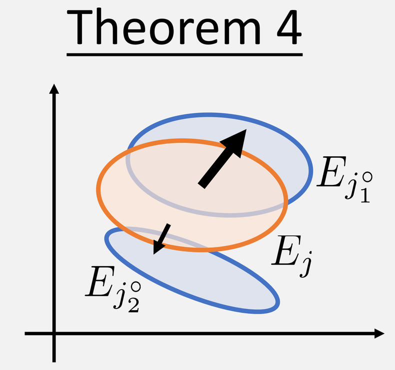
$E_j = \{x : f_j(x) > 0\}$

Assumption



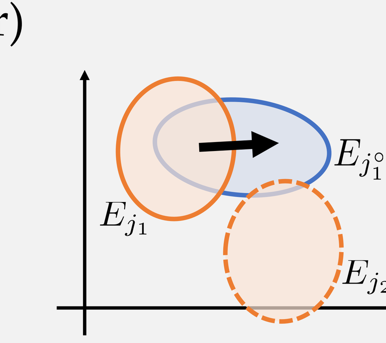
Small overlap (negative bias on whitened data)

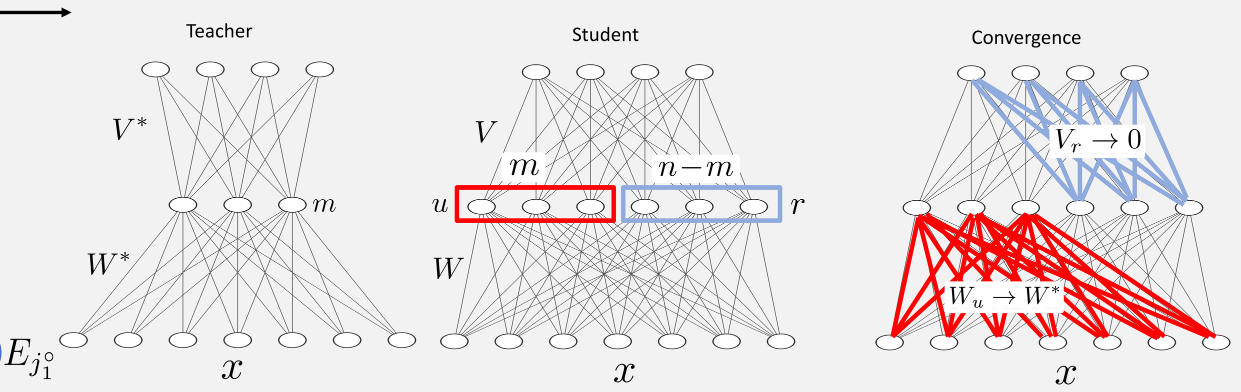
Theorem 4



Theorem 4. *For dynamics $\dot{\mathbf{w}}_j = P^{\perp}_{\mathbf{w}_j}(W^*\mathbf{h}^*_j - W\mathbf{h}_j)$, where $P^{\perp}_{\mathbf{w}_j} \equiv I - \mathbf{w}_j\mathbf{w}_j^T$ is a projection matrix into the orthogonal complement of $\mathbf{w}_j$. $\mathbf{h}^*_j$, $\mathbf{h}_j$ are corresponding j-th column in H^* and H. Denote $\theta_j = \angle(\mathbf{w}_j, \mathbf{w}^*_j)$ and assume $\theta_j \le \theta_0$. If $\gamma = \cos\theta_0 - (m-1)\epsilon_d M_d > 0$, then $\mathbf{w}_j \to \mathbf{w}^*_j$ with the rate $1 - \eta\bar{d}\gamma$ (η is learning rate). Here $\bar{d} = [1 + 2K_d\sin(\theta_0/2)]\min_j d^{*0}_{jj}$ and $M_d = (1 + K_d)[1 + 2K_d\sin(\theta_0/2)]^2 / \cos\frac{\theta_0}{2}$.*

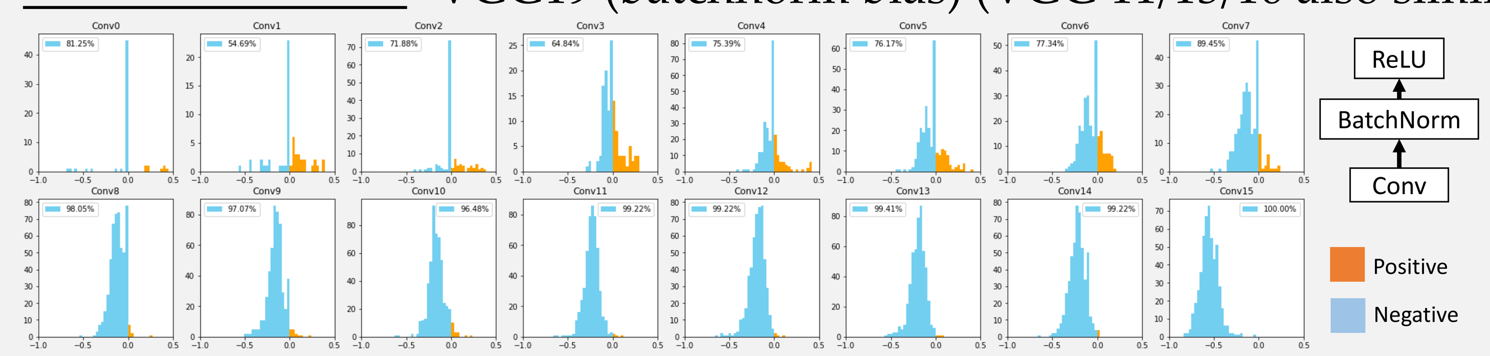
Theorem 5





Theorem 5 (Over-Parameterization and Top-down Modulation). *Consider $\dot{W} = W^*H^* - WH$ with over-parameterization ($n > m$) and its upper-layer dynamics $\dot{V} = L^*V^* - LV$. Assume that initial value W^0_u is close to W^*: $\theta_j = \angle(\mathbf{w}_j, \mathbf{w}^*_j) \le \theta_0$ for $j \in [u]$. If (1) Assumption 3 holds for all pairwise combination of columns of W^* **and** W^0_r, and (2) there exists $\gamma = \gamma(\theta_0, m) > 0$ and $\bar{\lambda}$ so that Eqn. 41 and Eqn. 42 holds, then $W_u \to W^*$, $V_u \to V^*$ and $V_r \to 0$ with rate $1 - \eta\bar{\lambda}\gamma$.*

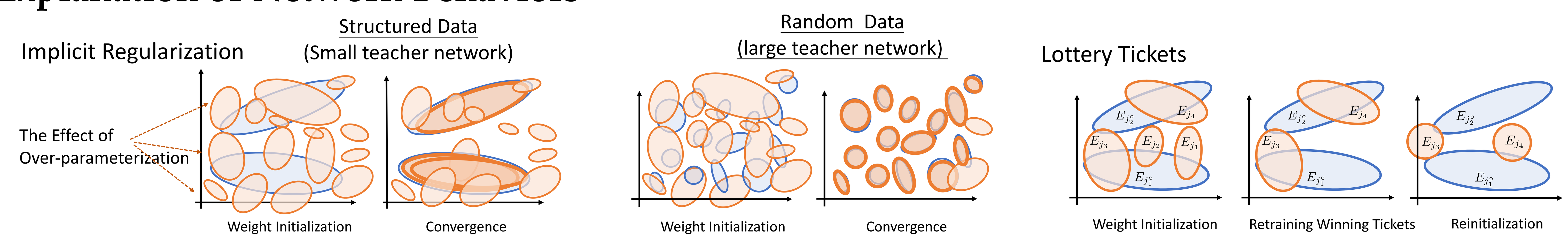
Initial Verification VGG19 (batchnorm bias) (VGG 11/13/16 also similar)



Explanation of Network Behaviors

Implicit Regularization

The Effect of Over-parameterization

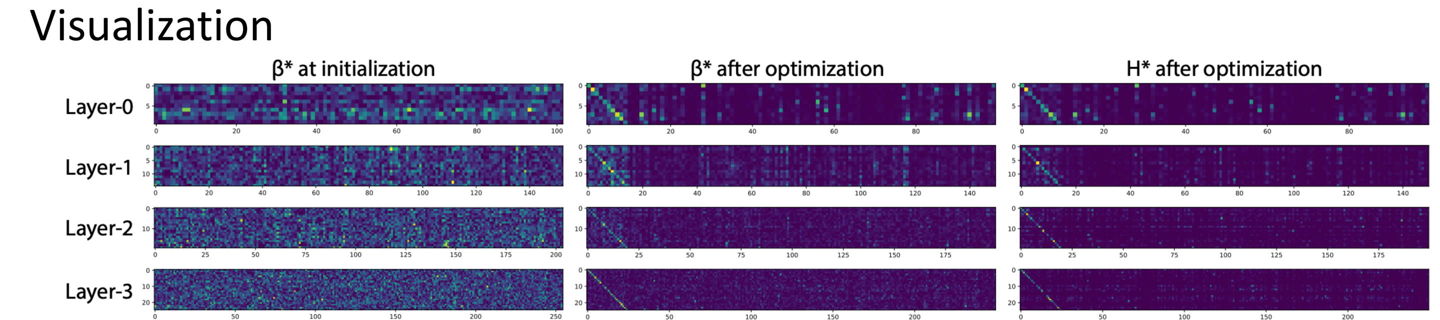


Experiments

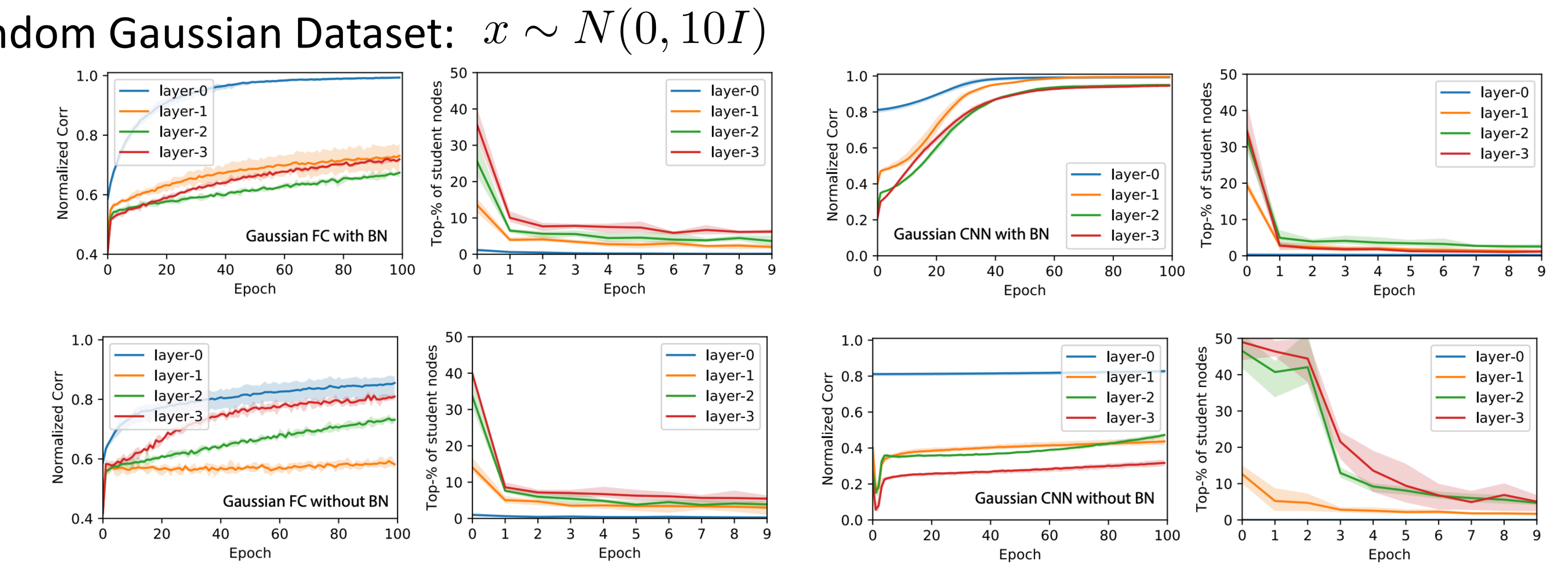
Setup
- FC Teacher: 50-75-100-125
- Convolution Teacher: 64-64-64-64
- Student: 10x teacher

Performance Metric
- Normalized Corr: $\bar{\rho} = \text{mean}_{j^\circ}\max_j \mathbf{f_j}^T\mathbf{f_{j^\circ}}$
- Ranking of winner student node in early epochs

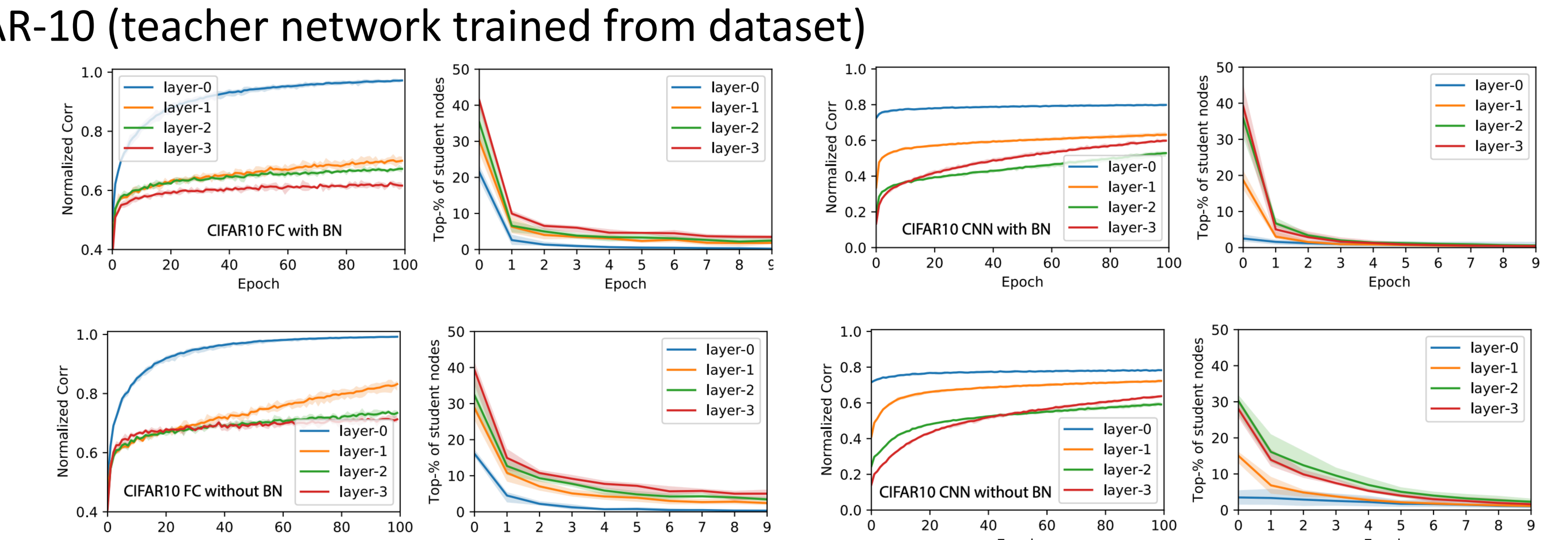
Visualization



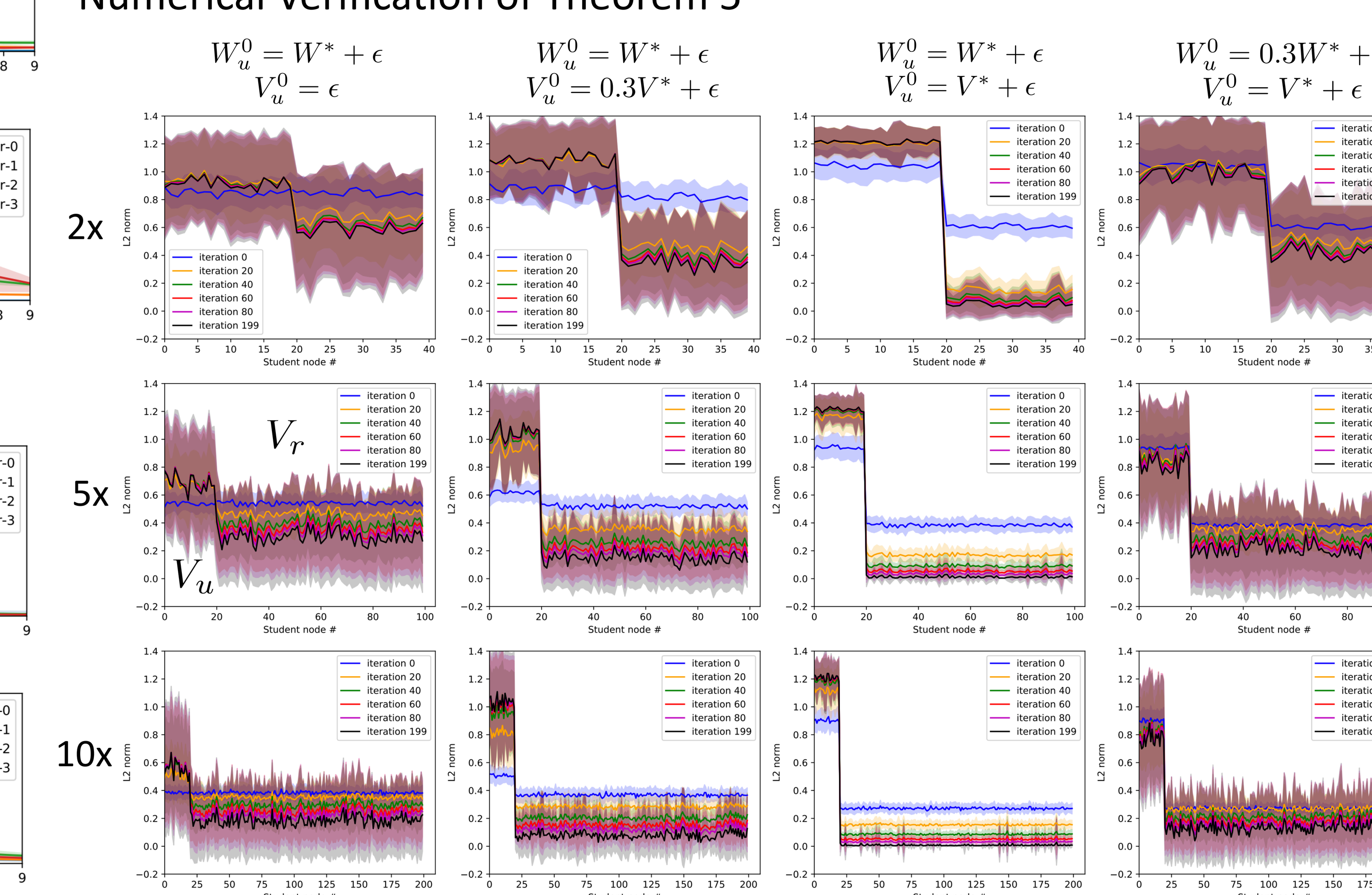
Random Gaussian Dataset: $x \sim N(0, 10I)$



CIFAR-10 (teacher network trained from dataset)



Numerical Verification of Theorem 5



Ablation Study