Student Specialization in Deep ReLU Networks With Finite Width and Input Dimension

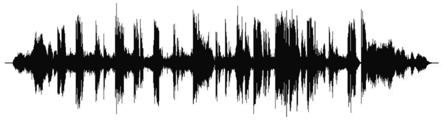
Yuandong Tian

Research Scientist and Manager
Facebook AI Research

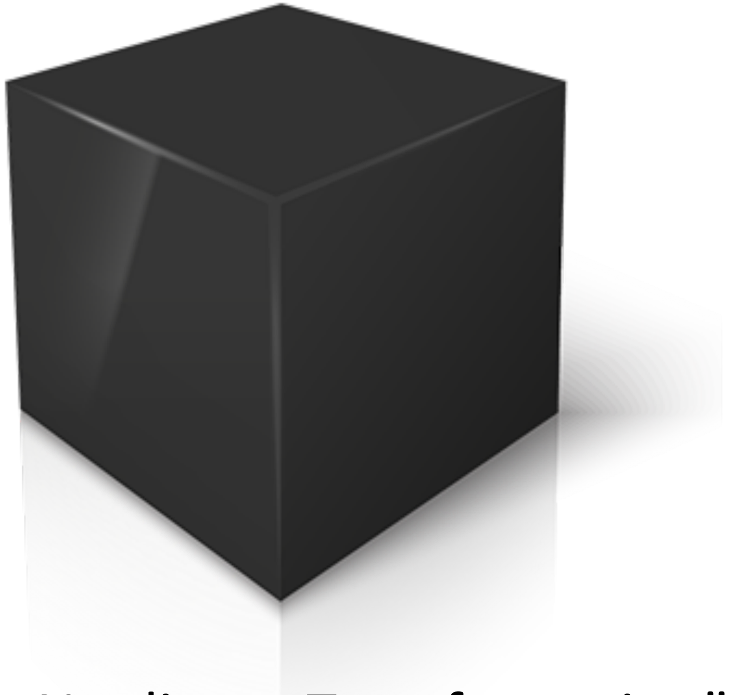
International Conference of Machine Learning (ICML), 2020

How do deep models work?

Input



This is an apple



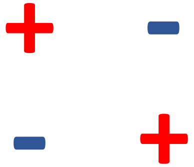
"Some Nonlinear Transformation"

Output

Three Major Problems

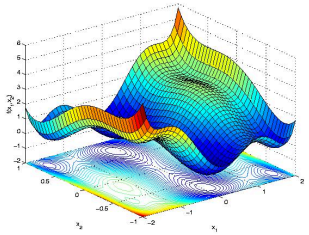
Understanding how Deep Models work

Expressibility



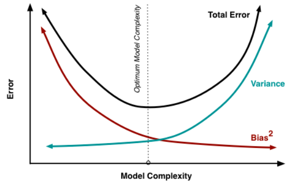
"Neural Network is a universal approximator"
"Deep Models can express functions more efficiently than shallow ones"

Optimization



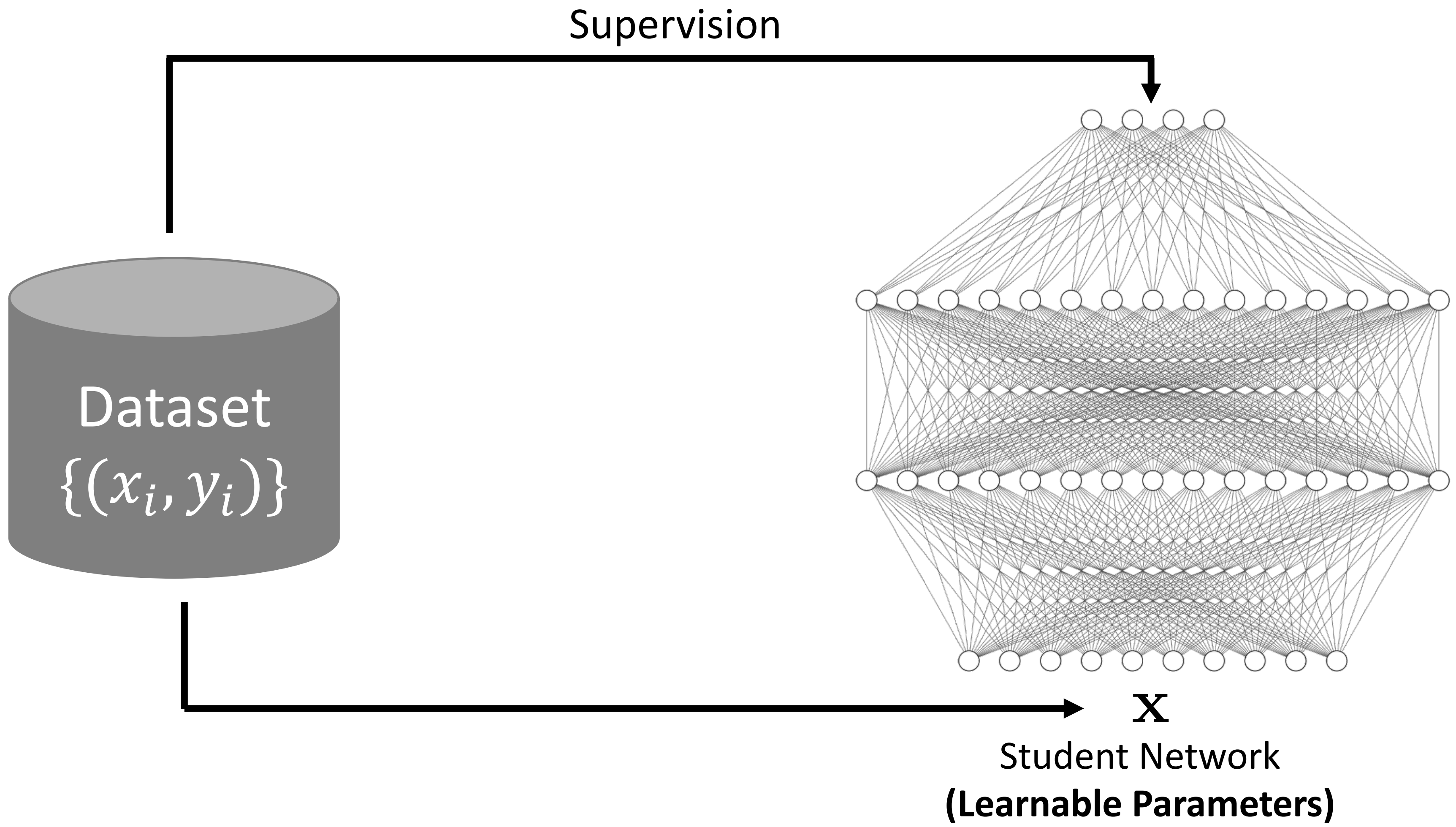
"Gradient vanishing/exploding"
"Gradient Descent might get stuck at saddle point / local minima"
"Can GD/SGD go to global optima? How fast?"

Generalization

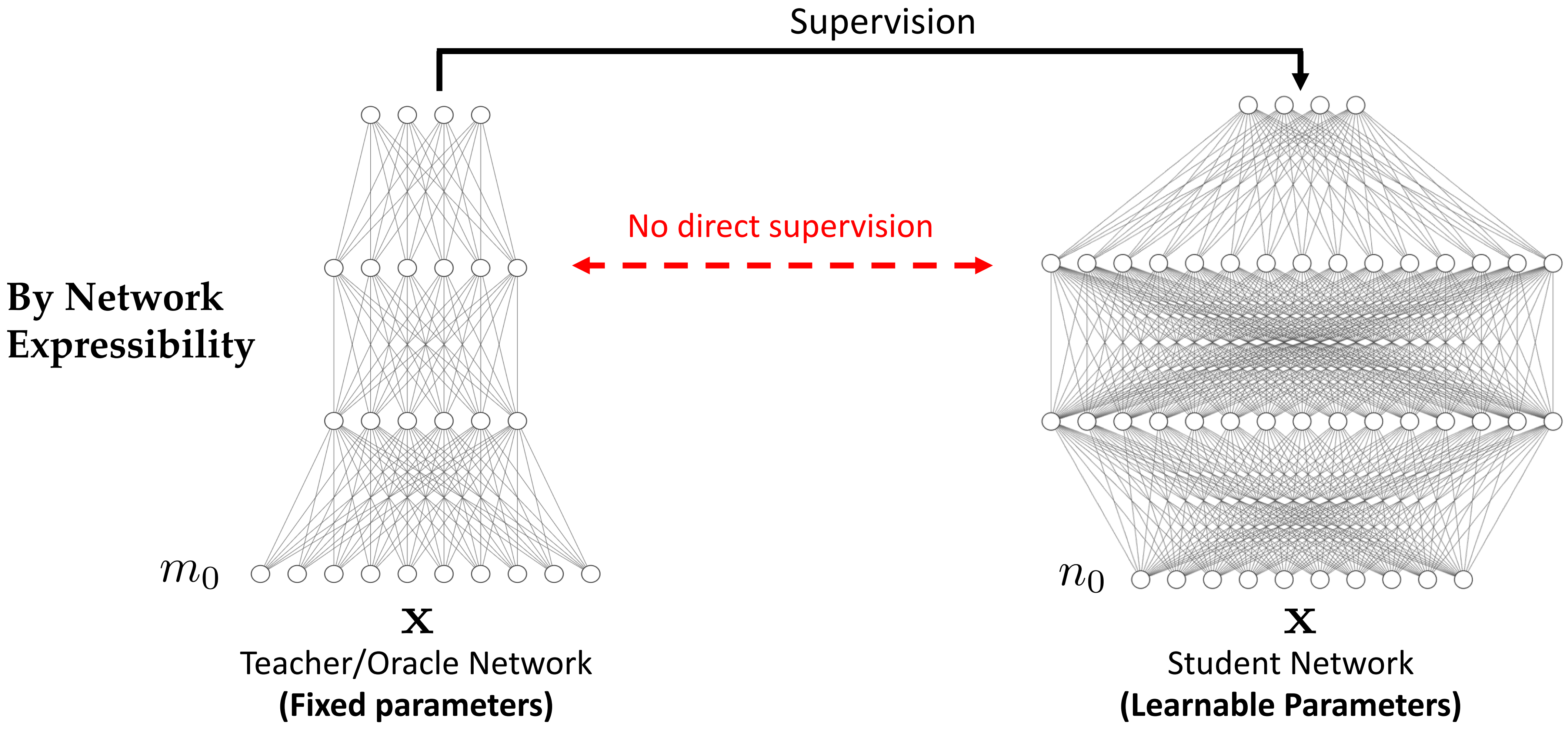


"Does zero training error often lead to overfitting?"
"More parameters might lead to overfitting."

Supervised Learning



Student-Teacher Setting

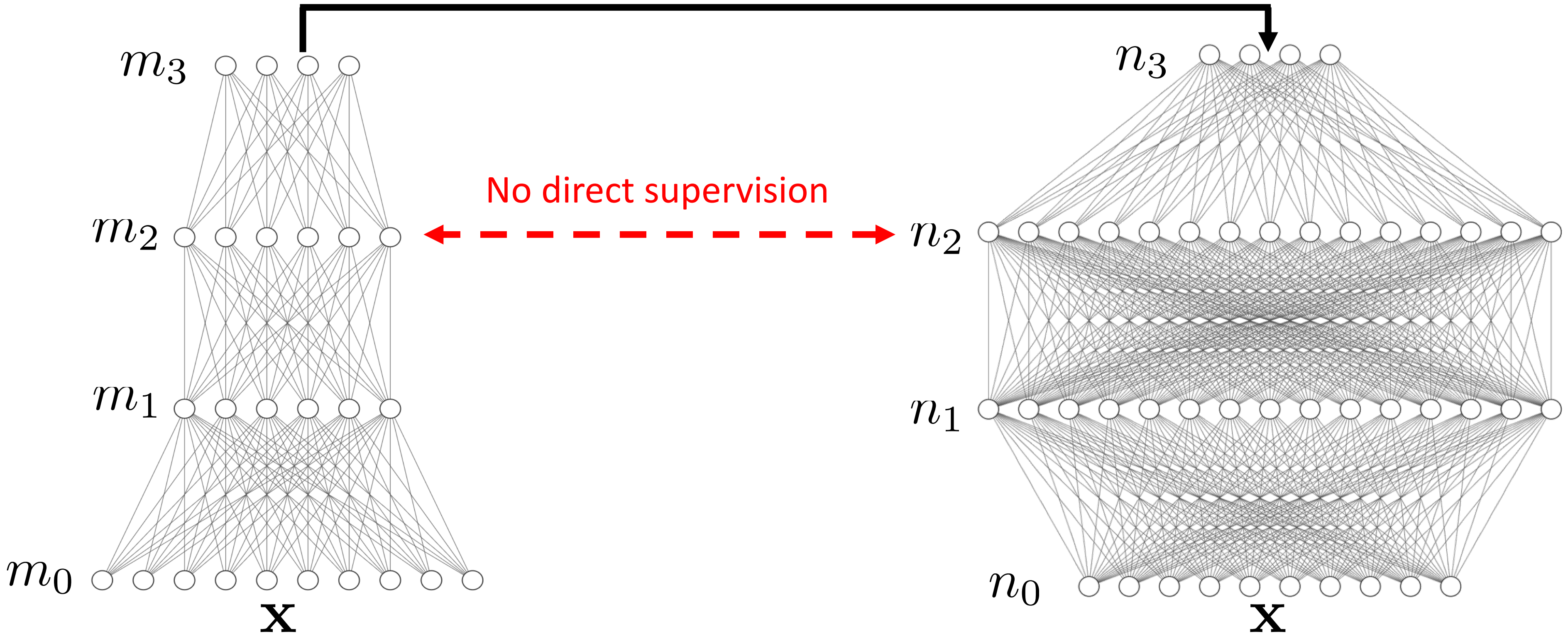


Setting in this paper

1. Finite m_0 and n_0
2. Works for $n_i \geq m_i$

(no crazy overparameterization)

$$\min_{\mathcal{W}} J(\mathcal{W}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[\|\mathbf{f}_L^*(\mathbf{x}) - \mathbf{f}_L(\mathbf{x})\|^2\right]$$



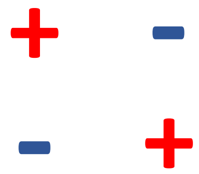
Teacher/Oracle Network
(Fixed parameters)

(Over-realized) Student Network
(Learnable Parameters)

Why Student-Teacher Setting?

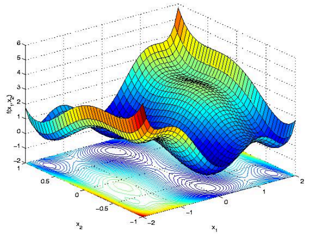
Understanding how Deep Models work

Expressibility



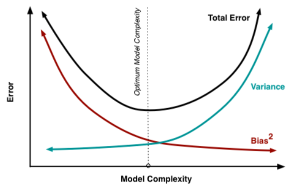
Provide a target function with bounded complexity

Optimization



Study fine dynamics behaviors by comparing with teacher

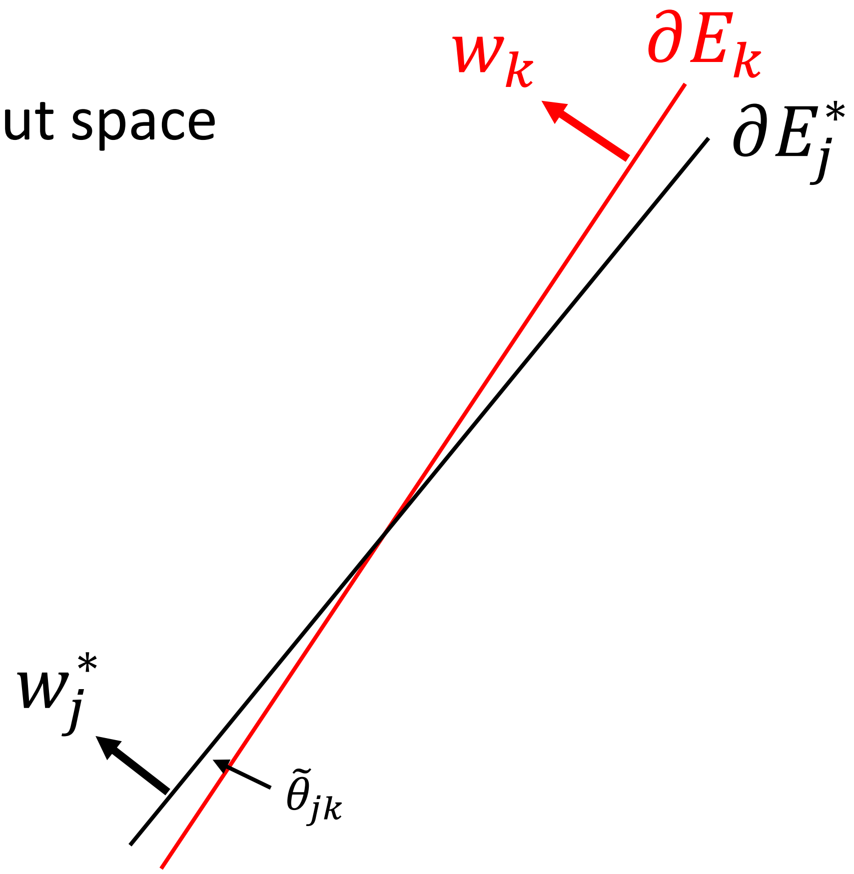
Generalization

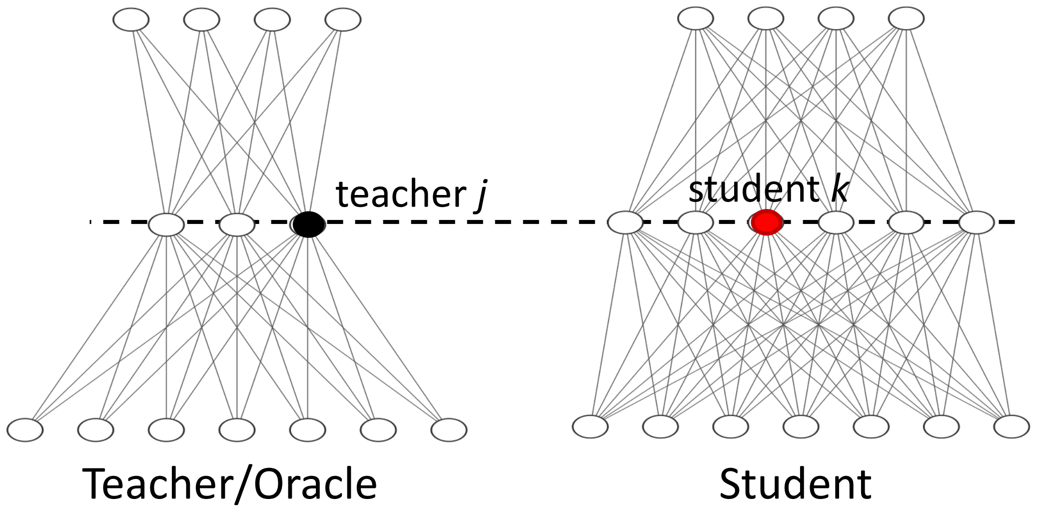


Our focus → ***Student Specialization*** yields generalization

Student Specialization

Input space





∂E_k: Boundary of node k

∂E_j^*: Boundary of teacher node j

ϵ-alignment: $\sin\tilde{\theta}_{jk} \leq \epsilon$ and $|b_j - b_k^*| \leq \epsilon$

Main Question

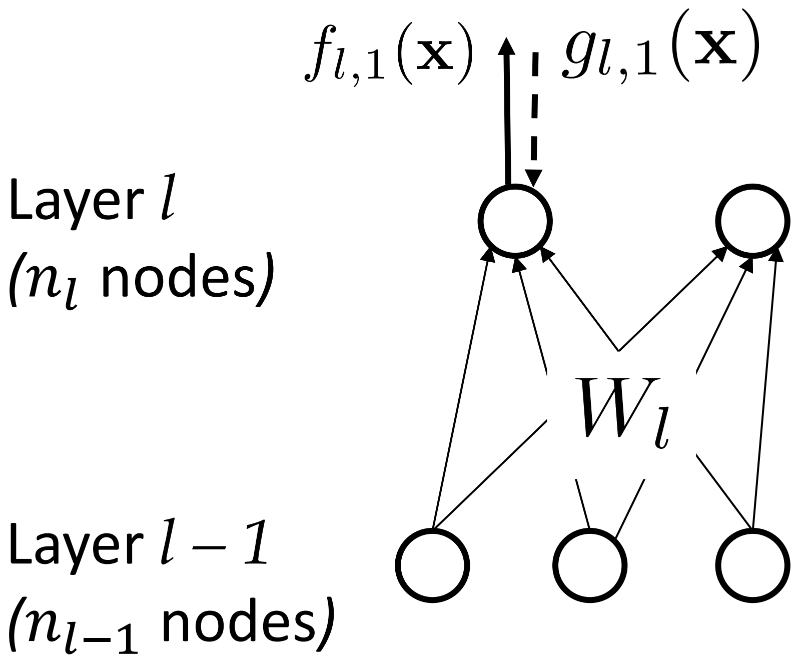
Small gradient
at every training sample
during training

? ? ?

Student aligns
with the teacher

➔ Small training error leads to good generalization

Notation



Activation

$$\mathbf{f}_l(\mathbf{x}) = \begin{bmatrix} f_{l,1}(\mathbf{x}) \\ f_{l,2}(\mathbf{x}) \end{bmatrix}$$

Gradient

$$\mathbf{g}_l(\mathbf{x}) = \begin{bmatrix} g_{l,1}(\mathbf{x}) \\ g_{l,2}(\mathbf{x}) \end{bmatrix}$$

Weight update rule: $\dot{W}_l = \mathbb{E}_{\mathbf{x}}\left[\mathbf{f}_{l-1}(\mathbf{x})\mathbf{g}_l^{\intercal}(\mathbf{x})\right]$

GD: expectation taken over the entire dataset

SGD: expectation taken over a batch

Lemma1: Recursive Gradient Rule

For layer l, there exists $A_l(x)$ and $B_l(x)$ so that:

$$\mathbf{g}_l(\mathbf{x}) = D_l(\mathbf{x})\left[A_l(\mathbf{x})\mathbf{f}_l^*(\mathbf{x}) - B_l(\mathbf{x})\mathbf{f}_l(\mathbf{x})\right]$$

Student gradient

Teacher mixture

Student mixture

Student gating

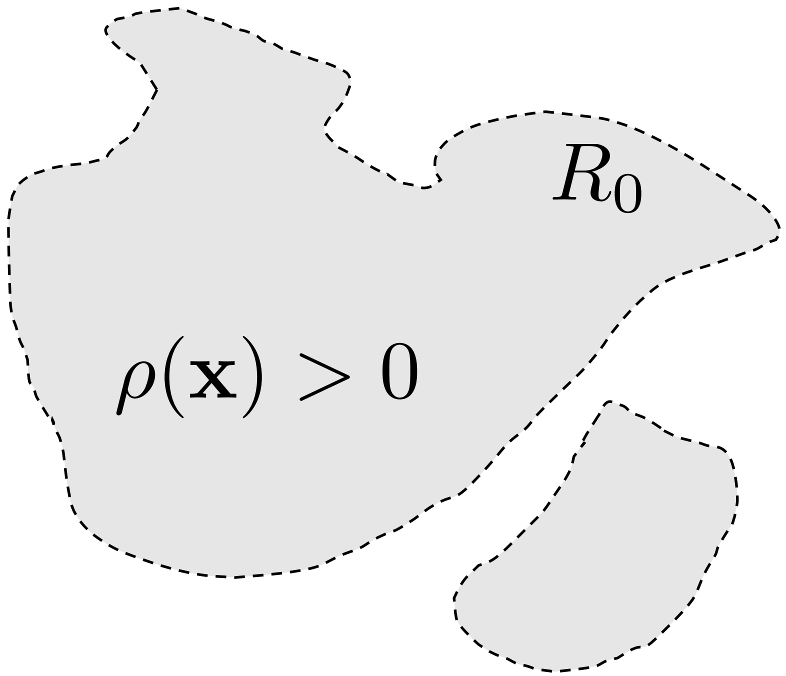
$A_l(x)$ and $B_l(x)$ are **piece-wise constant.**

Start with A Demonstrative Case:

Two-layer Network, **Zero Gradient** and **Infinite** Samples

Assumption of the dataset

No parametrized assumptions

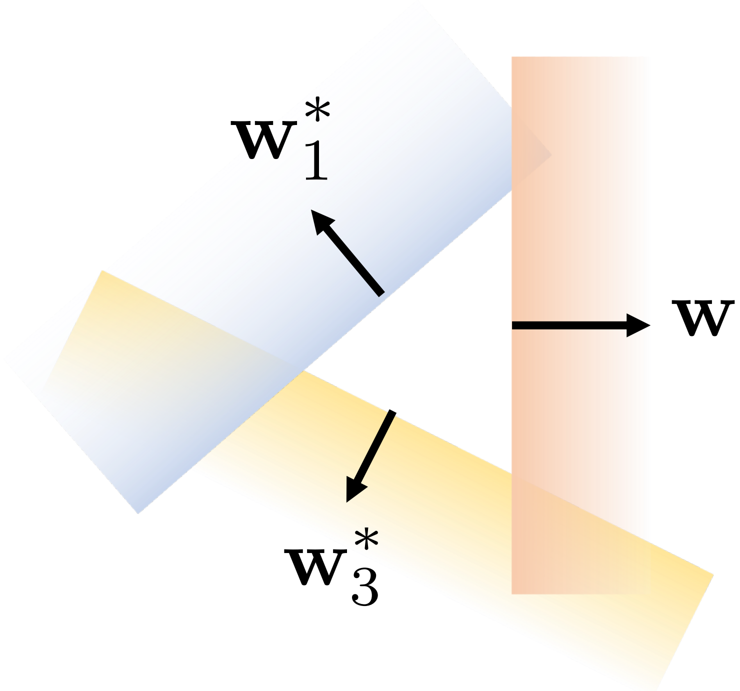


Infinite dataset!

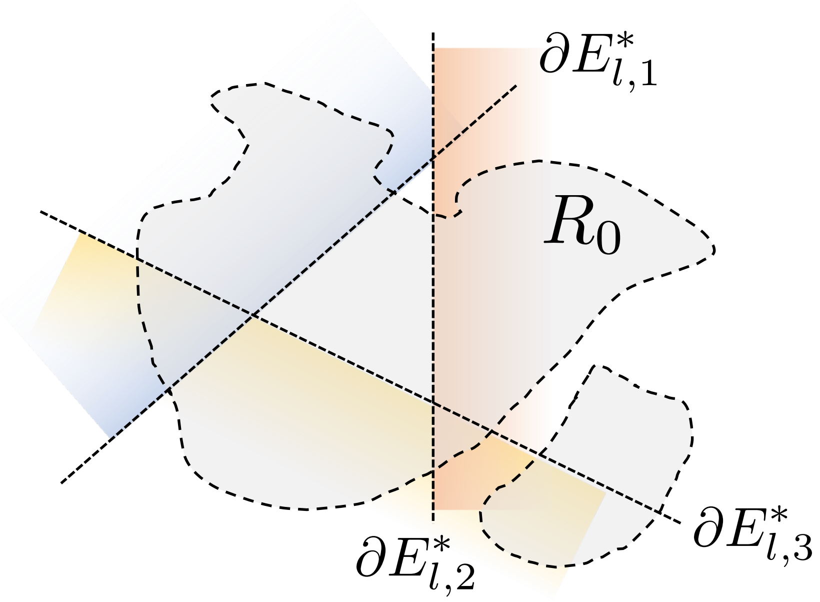
(Region needs to have interiors)

Assumptions on Teacher Network

- Cannot reconstruct arbitrary teachers
  - e.g., all ReLU nodes are dead

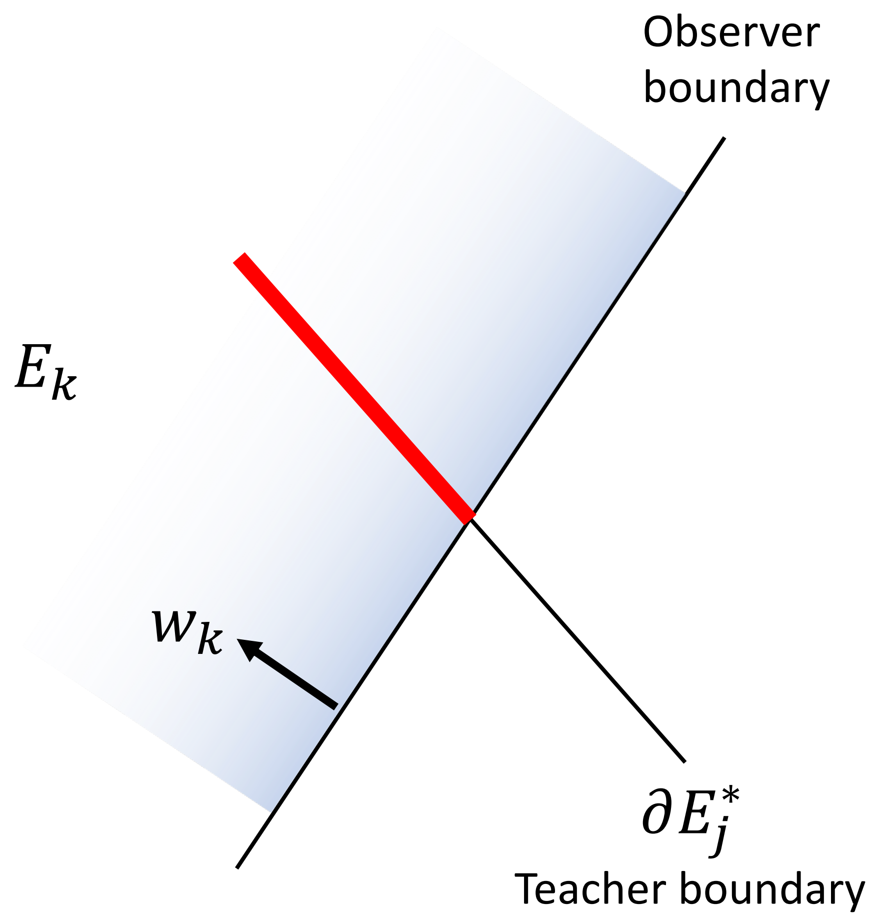


Distinct teacher nodes



Teacher's ReLU boundary are **visible** in the dataset

Definition of “Observation”



E_k: Activation region of node *k*

$$\partial E_j^* \cap E_k \neq \emptyset$$

Teacher *j* is **observed** by a student *k*

Main results: Alignment could happen!

$\boldsymbol{g}_1(x) = \mathbf{0}$ for all $x \in R_0$
(all input gradients at layer 1 is *zero* at all training samples)

Teacher node *j* is **observed** by a student node *k*

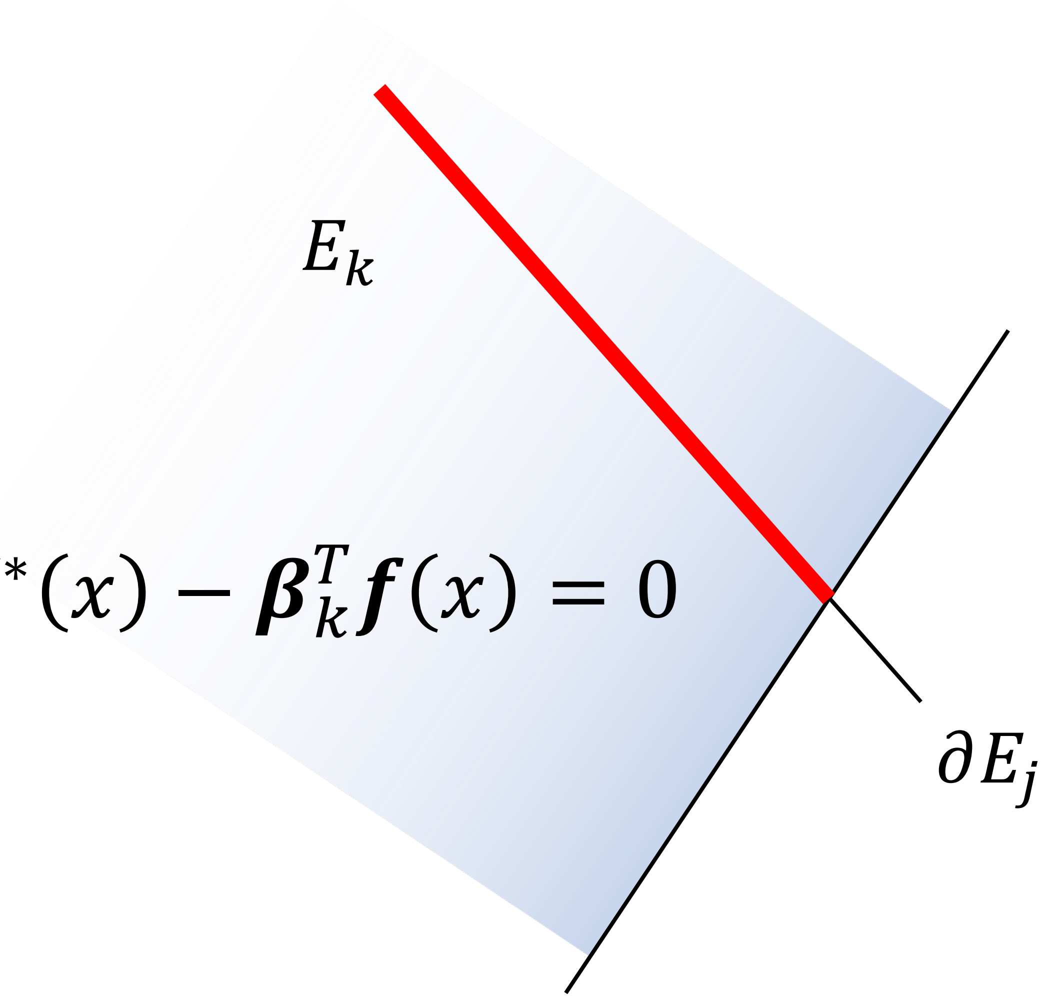
Teacher *j* is **aligned with** at least one student *k'*

Proof Sketch

The gradient of observer *k* is 0:

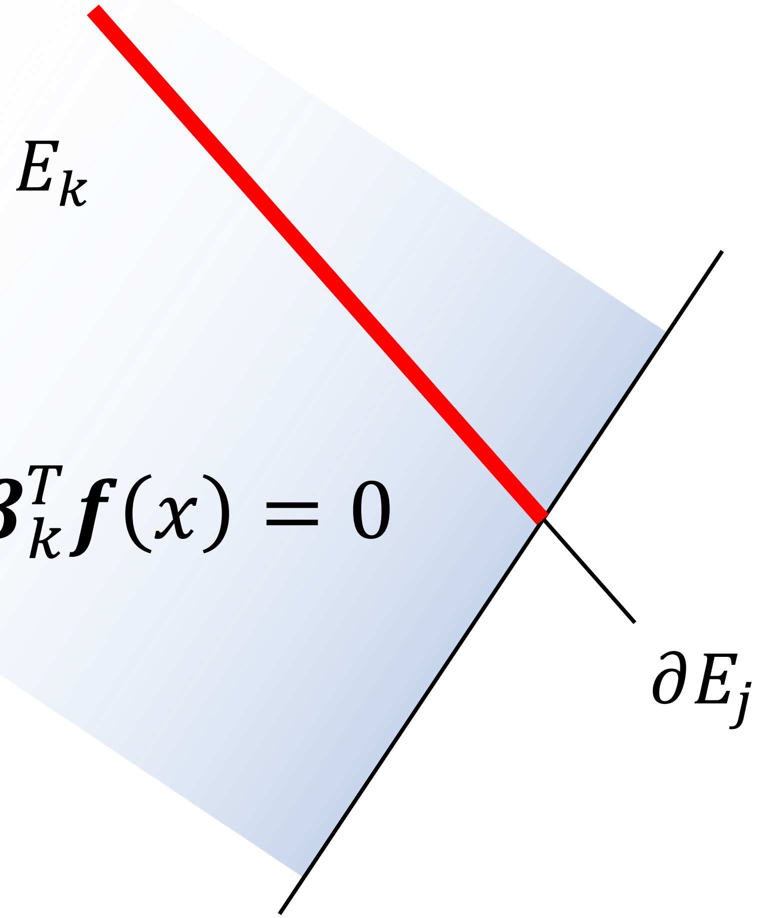
From Lemma 1, $g_k(x) = \boldsymbol{\alpha}_k^T \boldsymbol{f}^*(x) - \boldsymbol{\beta}_k^T \boldsymbol{f}(x) = 0$

If $x \in E_k$



Proof Sketch

The gradient of observer *k* is 0:



From Lemma 1, $g_k(x) = \boldsymbol{\alpha}_k^T \boldsymbol{f}^*(x) - \boldsymbol{\beta}_k^T \boldsymbol{f}(x) = 0$

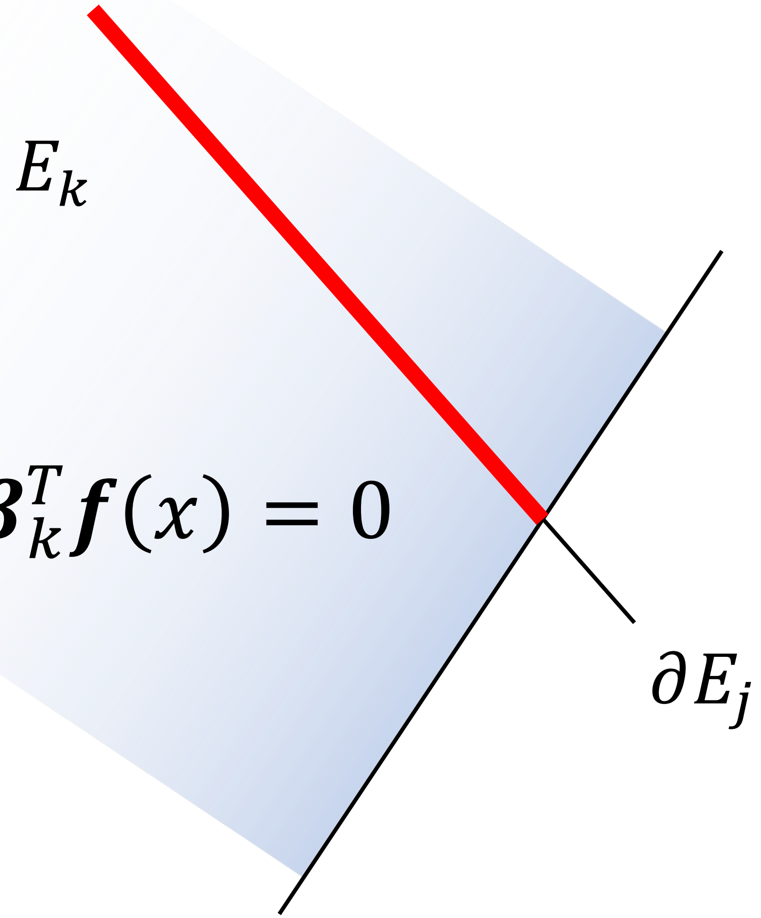
If $x \in E_k$

ReLUs are
linear independent!

Coefficients for teacher *j*
direction must be 0

Proof Sketch

The gradient of observer *k* is 0:



From Lemma 1, $g_k(x) = \boldsymbol{\alpha}_k^T \boldsymbol{f}^*(x) - \boldsymbol{\beta}_k^T \boldsymbol{f}(x) = 0$

If $x \in E_k$

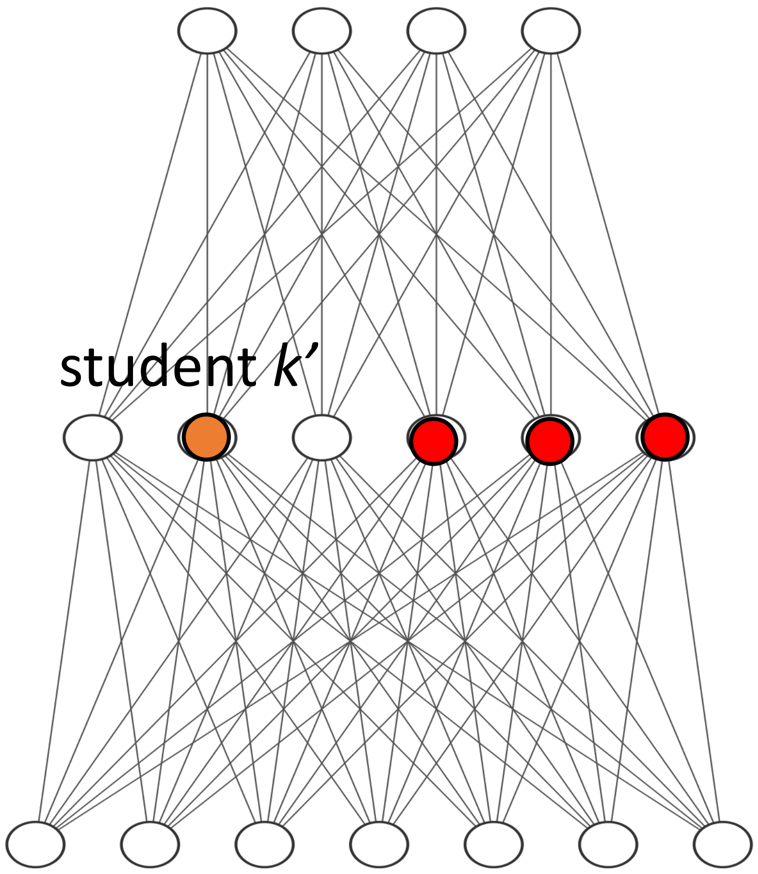
ReLUs are linear independent!

Coefficients for teacher *j* direction must be 0

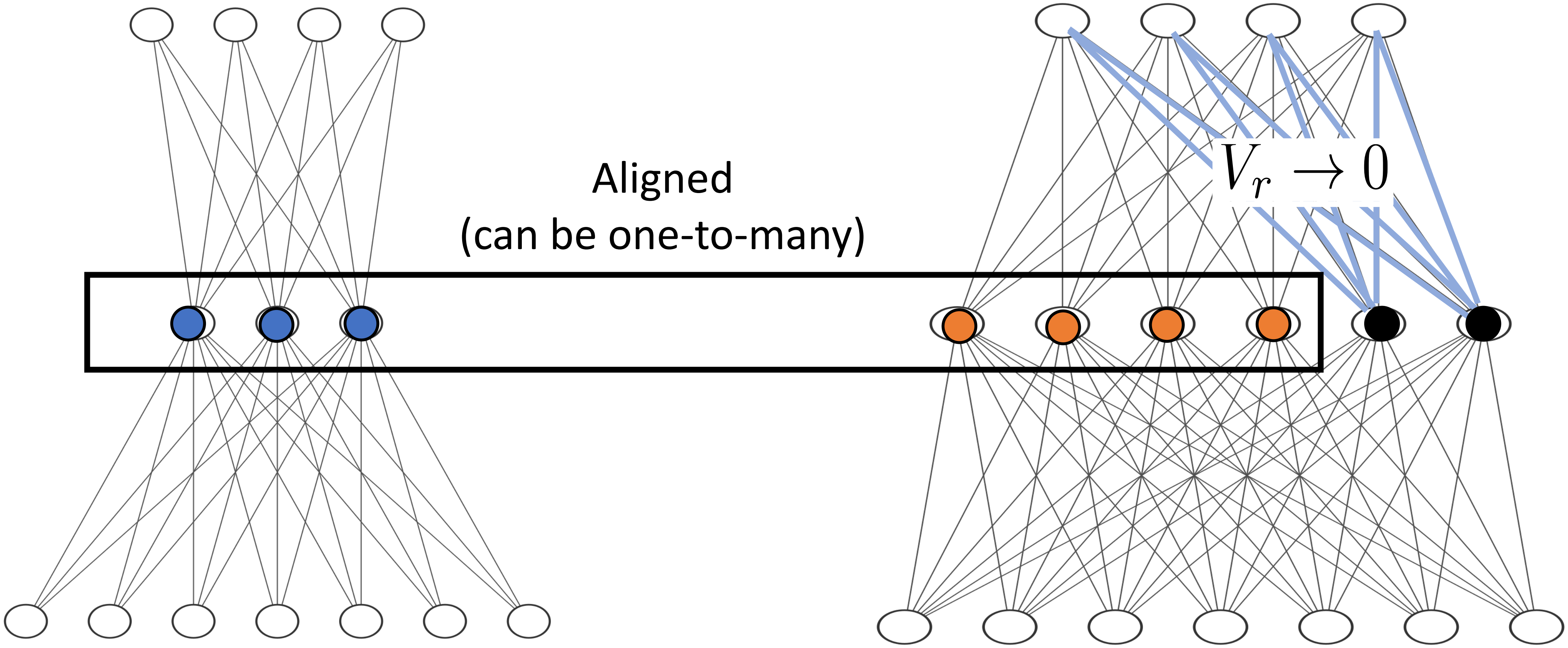
Teacher *j* is aligned with at least one student *k'* (sum of coefficients = 0)

Why Over-realization helps?

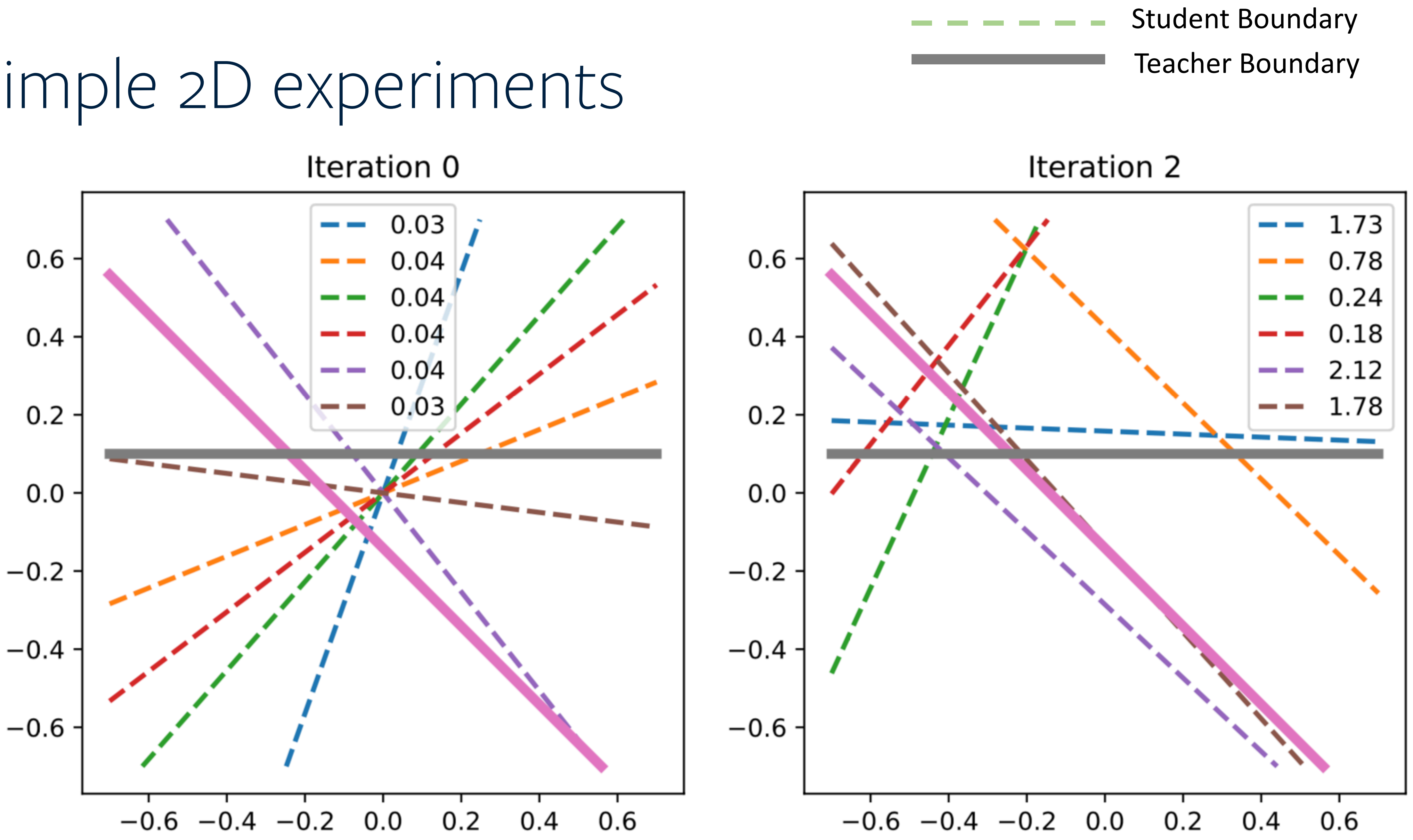
More observers!



What happens to unaligned students?

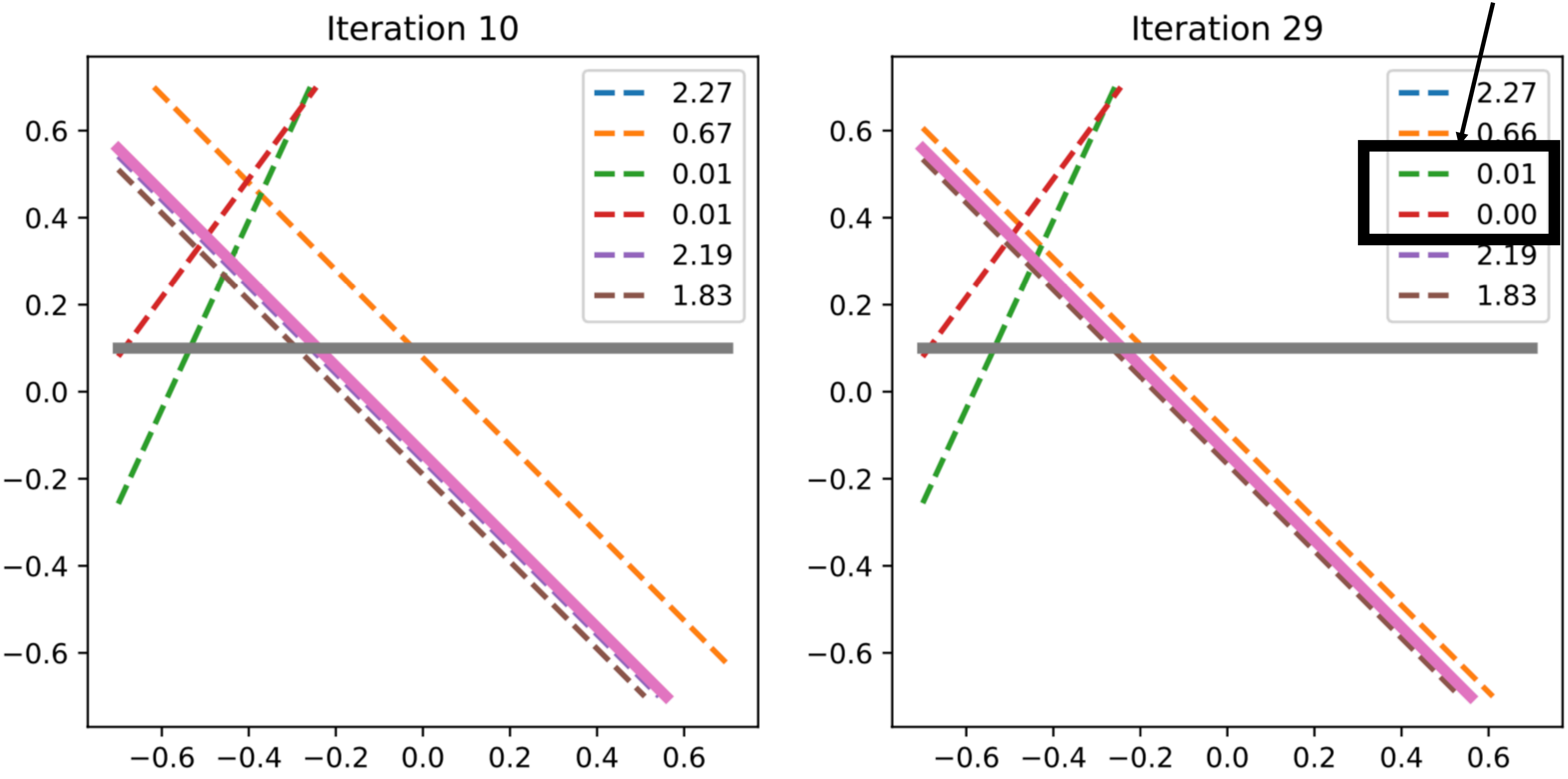


Simple 2D experiments



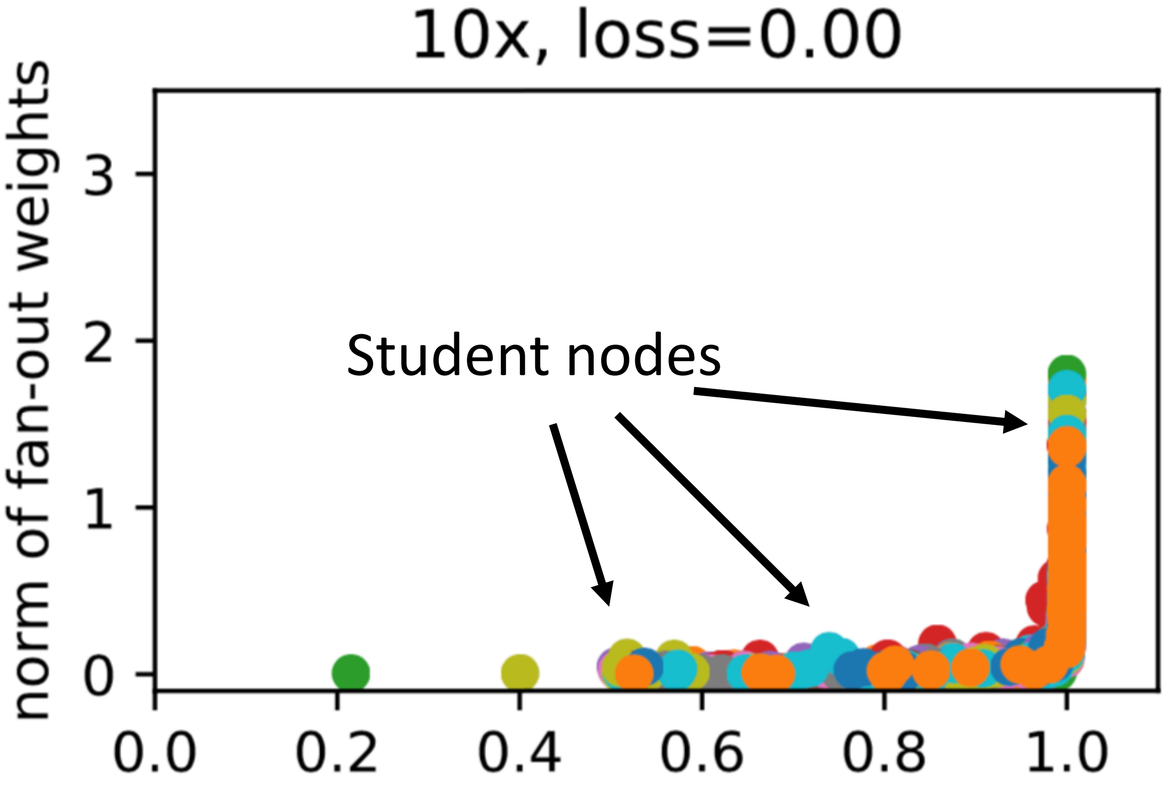
Simple 2D experiments

Norm of the fan-out weights.



L-shape curve at convergence

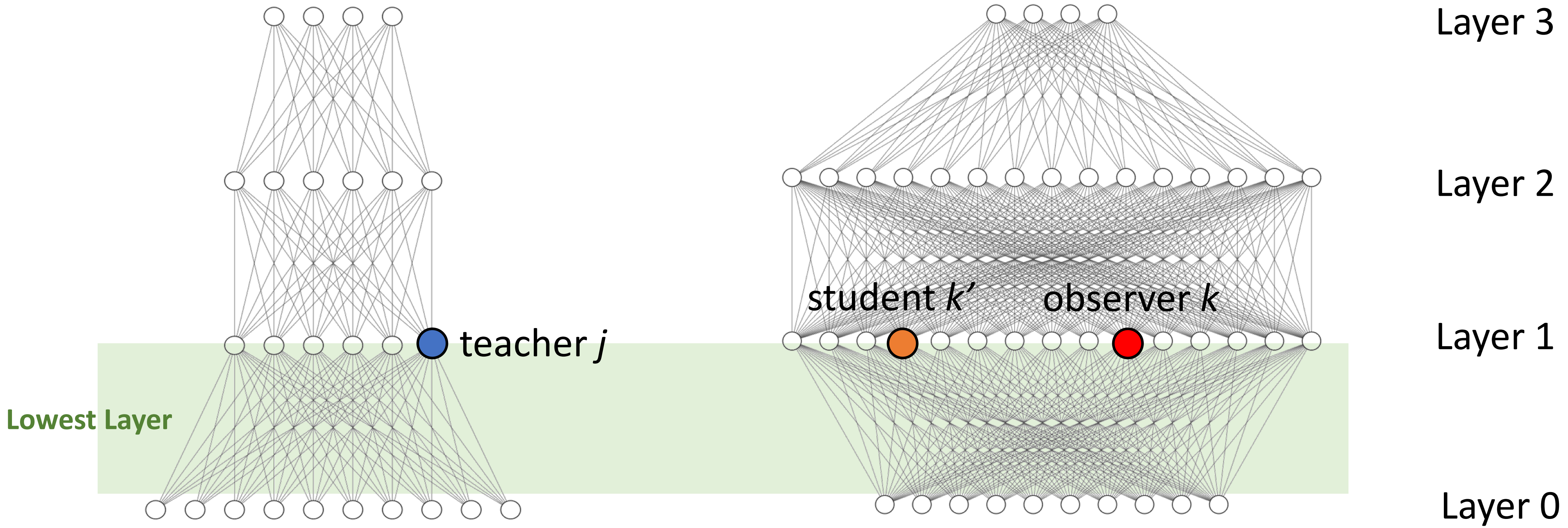
10x over-realization



$\|\mathbf{v}_k\|$

Normalized correlation of a student node
to its best correlated teacher (degree of specialization)

Multi-Layer case: Alignment could happen!



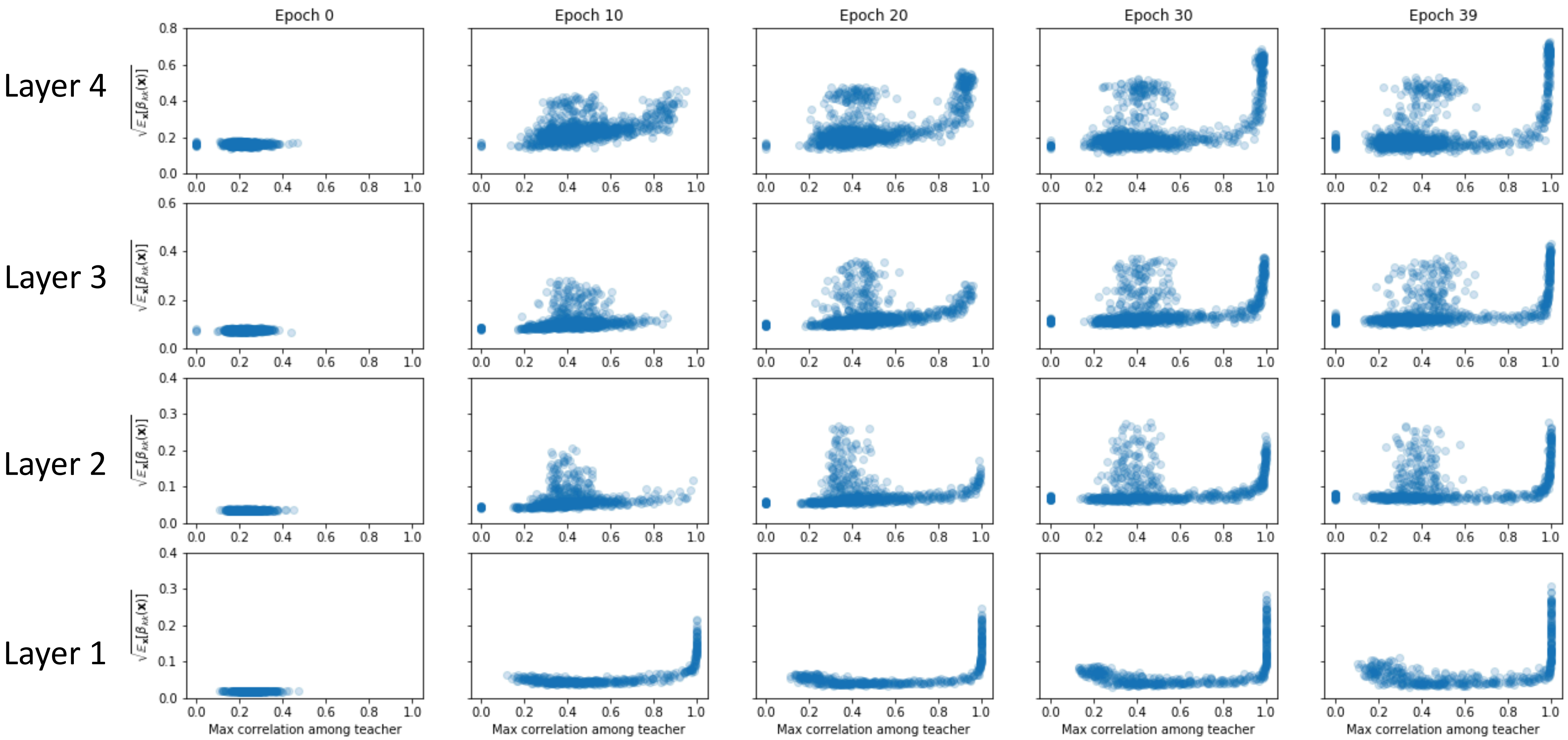
$$\boldsymbol{\alpha}_k^{\intercal}(\mathbf{x})\mathbf{f}^*(\mathbf{x}) - \boldsymbol{\beta}_k^{\intercal}(\mathbf{x})\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

Piece wise constant, apply the same logic **per region!**

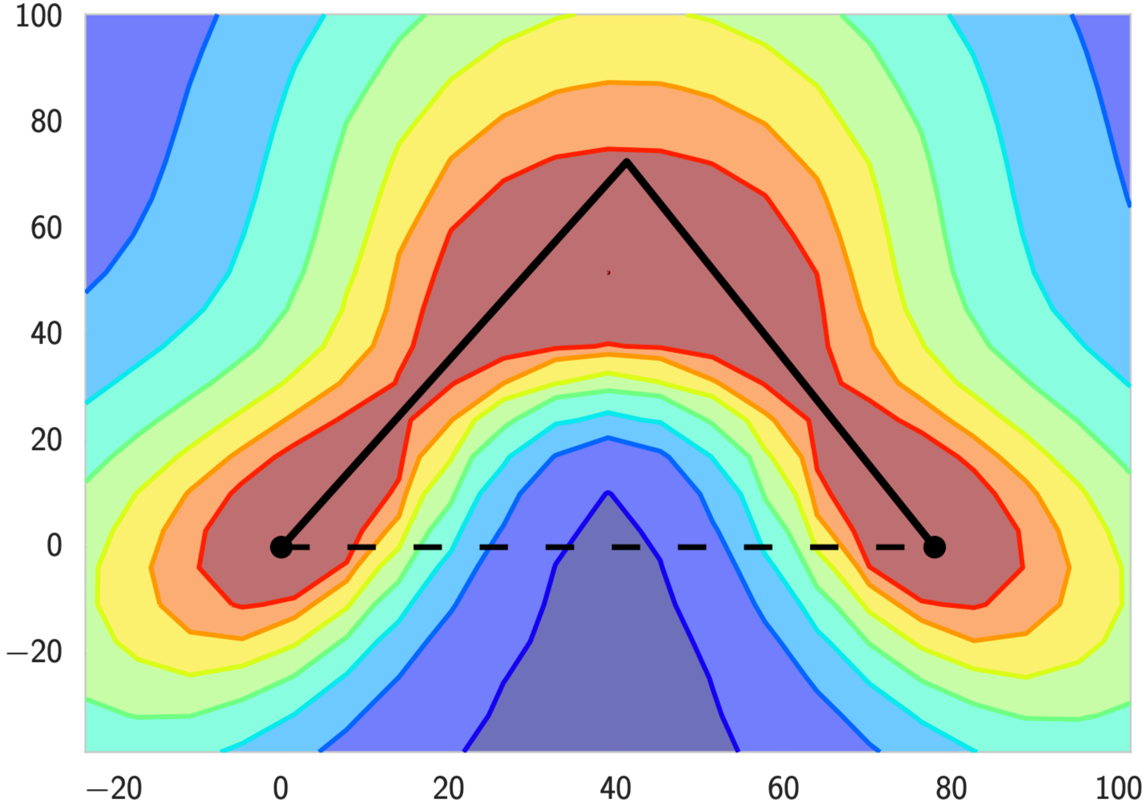
For 2-layer:

$$\sqrt{\mathbb{E}_{\mathbf{x}}\left[\beta_{kk}(\mathbf{x})\right]} = \|\mathbf{v}_k\|$$

Training Progresses



Solutions can be connected by line segments

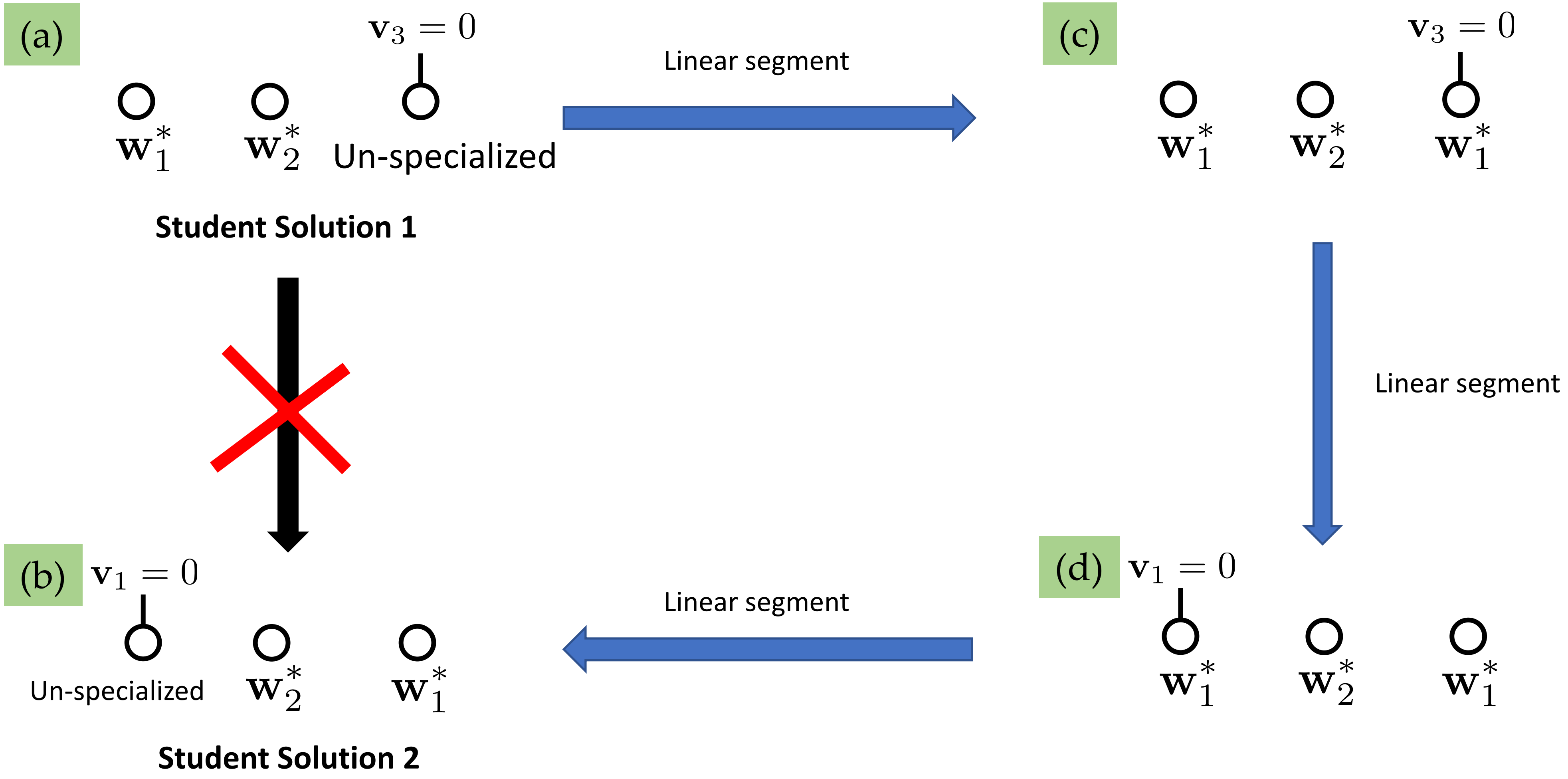


[Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs, Garipov et al. NeurIPS 2018]

[Essentially No Barriers in Neural Network Energy Landscape, Draxler et al, 2018]

[Explaining Landscape Connectivity of Low-cost Solutions for Multilayer Nets, Kuditipudi et al, 2019]

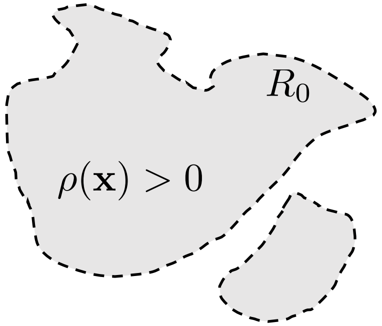
Our Explanation



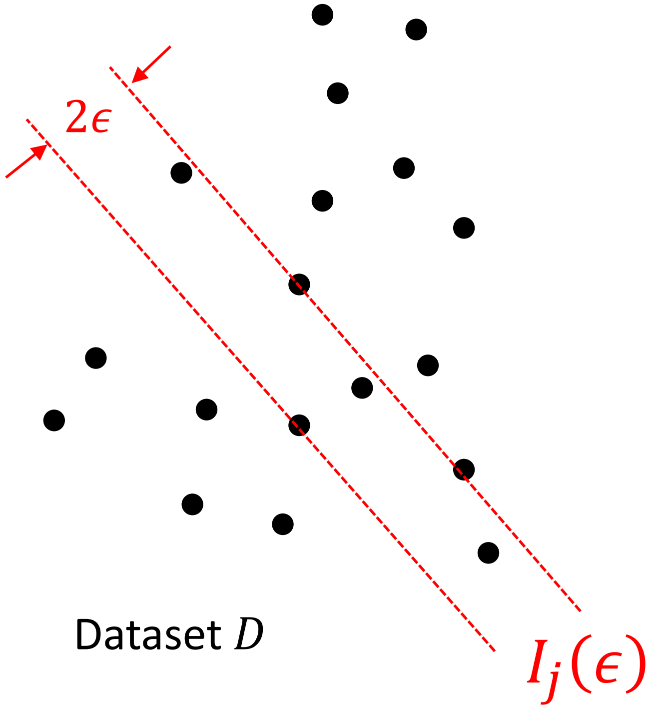
More Realistic Case:

Student Specialization with **2-layers** ReLU nets, **Small** Gradient and **Finite** Samples

Dataset Assumption



Infinite case



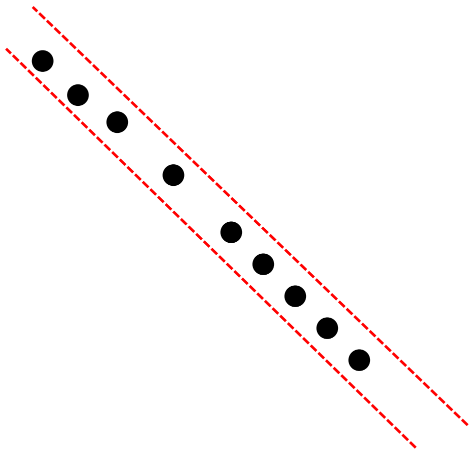
For any hyper-plane band $I_j(\epsilon)$:

$$\left|D \cap I_j(\epsilon)\right| \leq \eta\epsilon|D| + (d+1)$$

Intuition: Data should be full-rank

$\left|D \cap I_j(\epsilon)\right| \approx |D|$

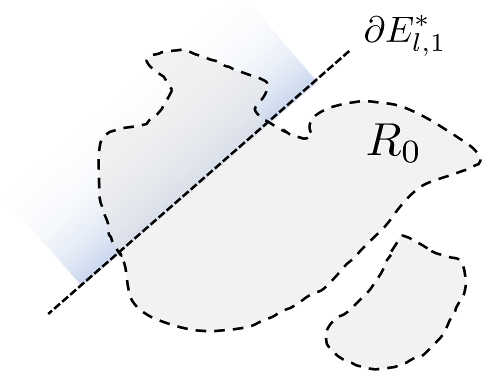
But ϵ is small



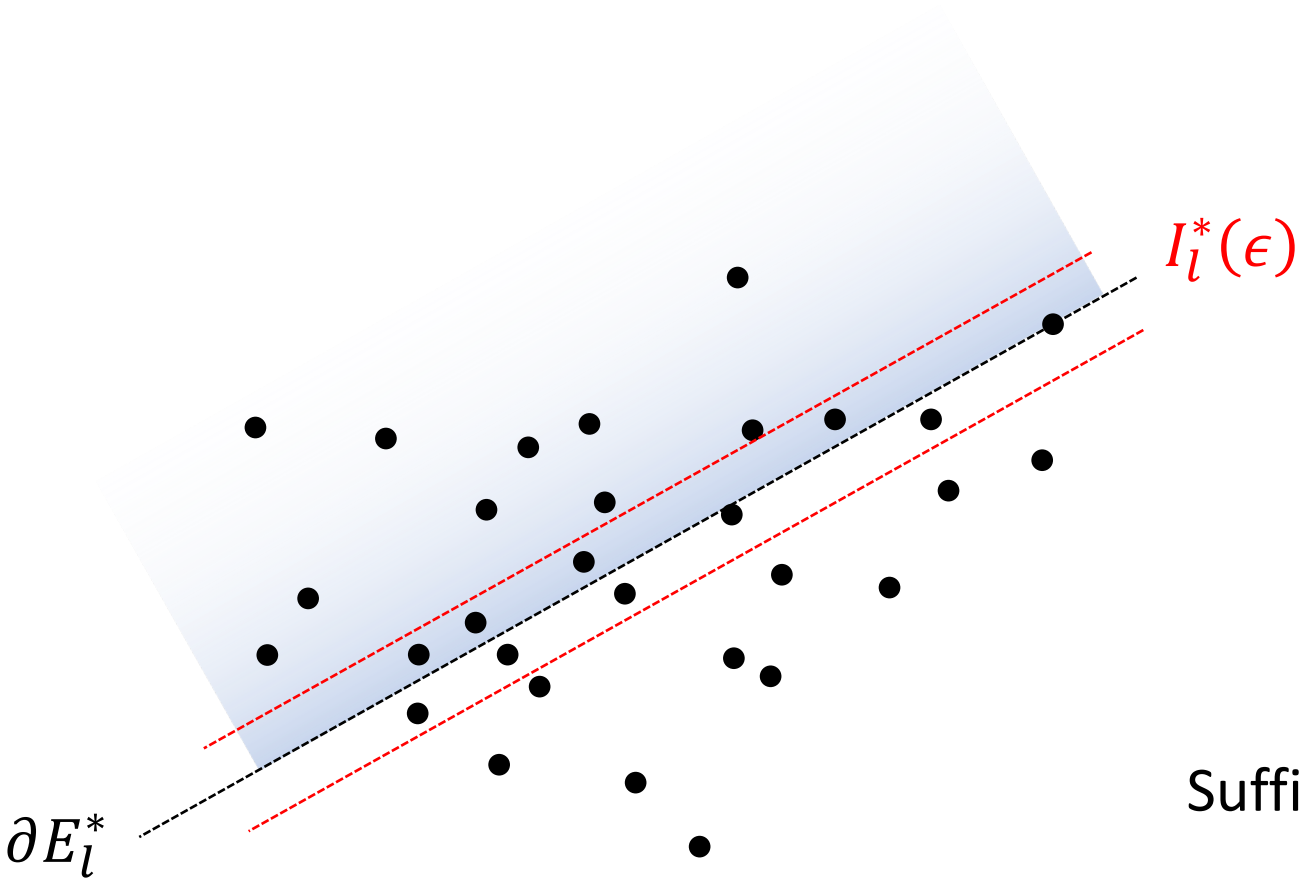
Failure case

(low-rank)

Dataset-Teacher Compatibility



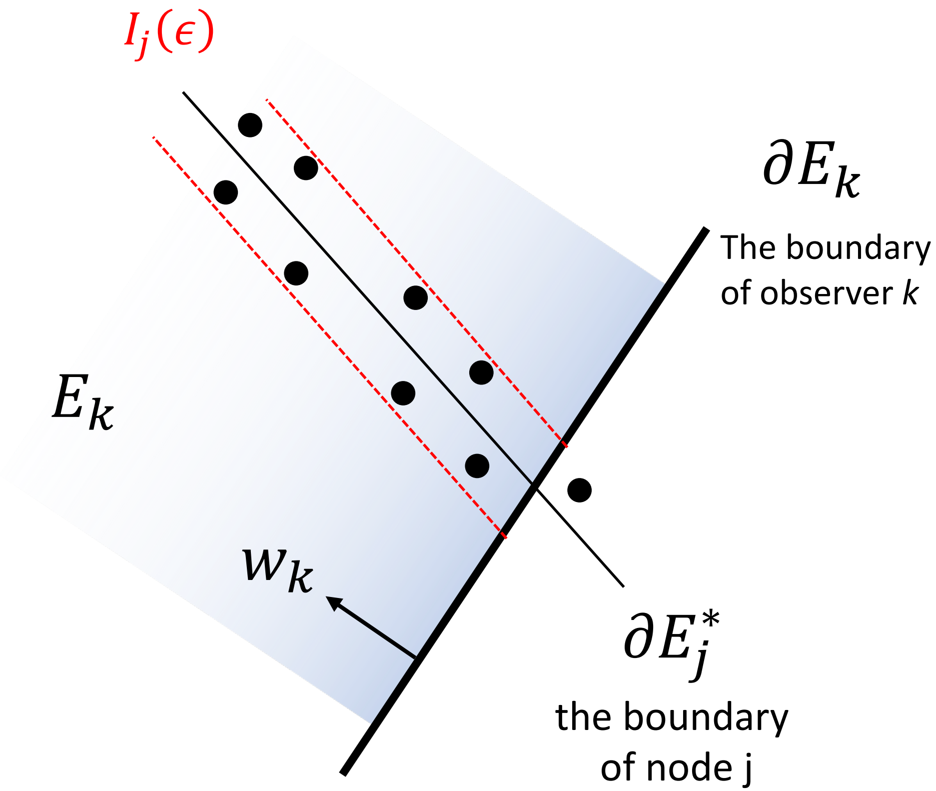
Infinite case

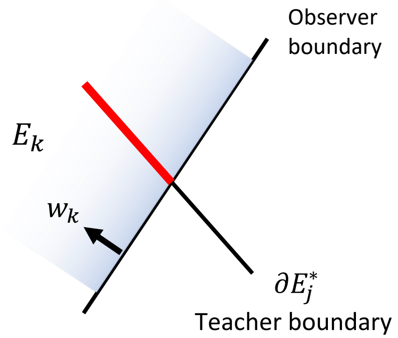


$$|D \cap I_l^*(\epsilon)| \geq \tau\epsilon|D|$$

Sufficient Data around teacher boundary

Observation in Finite Sample Case





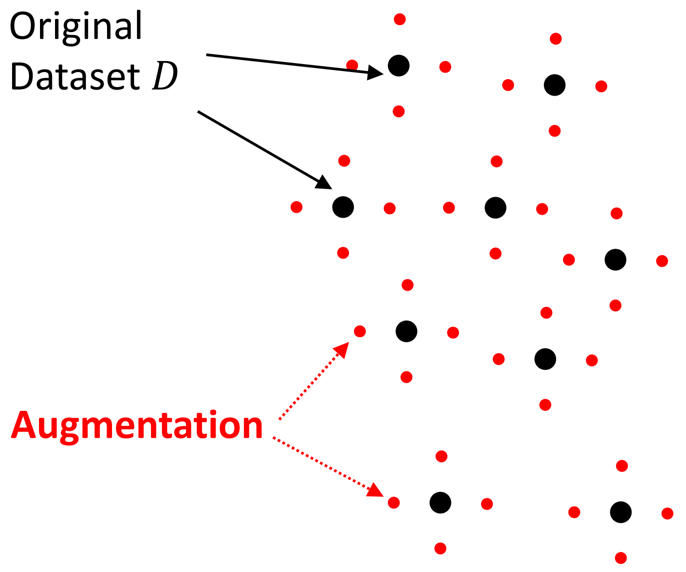
Infinite case

For a teacher node j, there exists a student k:

$$|D \cap I_j(\epsilon) \cap E_k| \geq \kappa \, |D \cap I_j(\epsilon)|$$

A sufficient portion of boundary samples lie in E_k

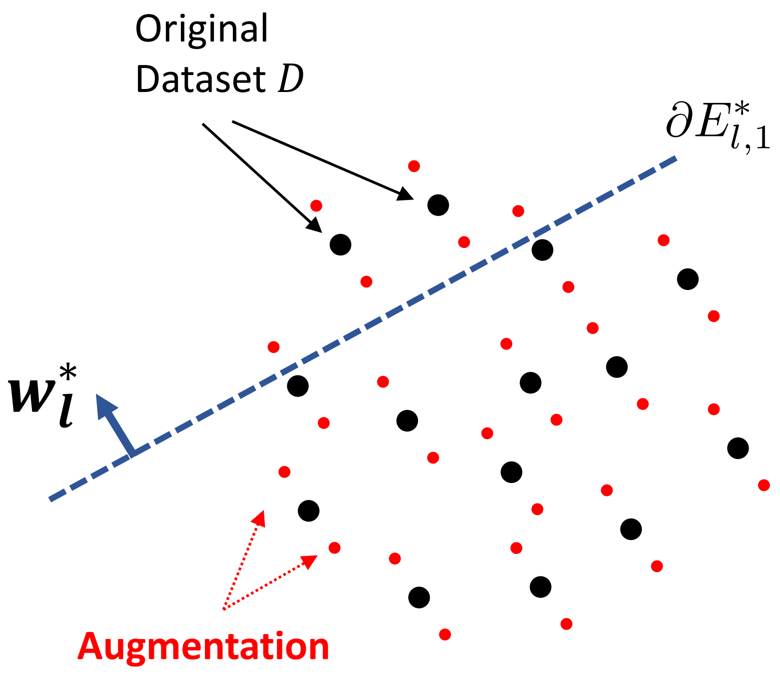
Data Augmentation



Teacher-agnostic augmentation

D' = Aug(D)

$|D'| = (2d+1)|D|$



Teacher-aware augmentation

D' = Aug(D)

$|D'| = 2m|D|$

Polynomial Complexity for 2-layered Network

To achieve ϵ-alignment between a teacher *j* and student *k*

$$K_1 = m_1 + n_1$$

Small gradient

$$\|\mathbf{g}_1(\mathbf{x}, \hat{\mathcal{W}})\|_\infty \le \frac{\alpha_{kj}}{5K_1^{3/2}\sqrt{d}}\epsilon,\ \mathbf{x} \in D'$$

$$\|\mathbf{g}_1(\mathbf{x}, \hat{\mathcal{W}})\|_\infty \le \frac{\alpha_{kj}}{5K_1^{3/2}}\epsilon$$

Sample Complexity of original Dataset D

$$N = \mathcal{O}(K_1^{5/2} d^2 \epsilon^{-1} \kappa^{-1})$$

$$\mathcal{O}(K_1^{5/2} d \epsilon^{-1} \kappa^{-1})$$

Teacher-agnostic augmentation

D' = Aug(D)

|D'| = (2d+1)|D|

Teacher-aware augmentation

D' = Aug(D)

|D'| = m|D|

Polynomial Complexity for 2-layered Network

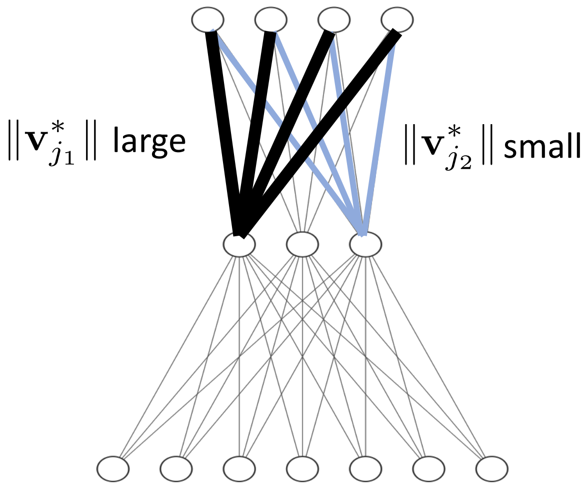
To achieve ϵ-alignment between a teacher *j* and student *k*

$$K_1 = m_1 + n_1$$

Small gradient

$$\|\mathbf{g}_1(\mathbf{x}, \hat{\mathcal{W}})\|_\infty \leq \frac{\boxed{\alpha_{kj}}}{5K_1^{3/2}\sqrt{d}}\epsilon,\ \mathbf{x} \in D'$$

$$\|\mathbf{g}_1(\mathbf{x}, \hat{\mathcal{W}})\|_\infty \leq \frac{\boxed{\alpha_{kj}}}{5K_1^{3/2}}\epsilon$$



$$\alpha_{kj} := \boldsymbol{v}_k^{\mathrm{T}} \boldsymbol{v}_j^*$$

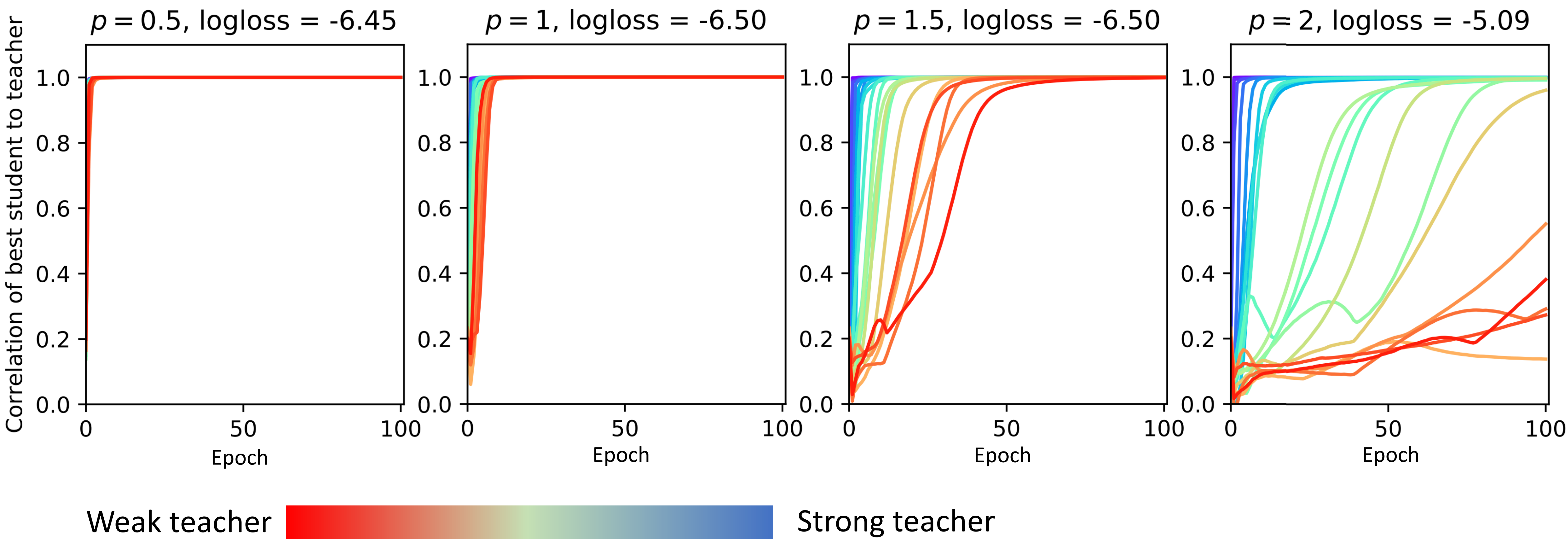
Strong teacher nodes are learned faster

1. Robust to Noise! 😃
2. Hard to learn weak teacher nodes 😢

Weak teacher nodes are slow to learn

Teacher j:

$$\|\mathbf{v}_j^*\| \propto 1/j^p$$



Polynomial Complexity for 2-layered Network

To achieve ϵ-alignment between a teacher j and student k

$$K_1 = m_1 + n_1$$

Small gradient

$$\|\mathbf{g}_1(\mathbf{x}, \hat{\mathcal{W}})\|_\infty \le \frac{\alpha_{kj}}{5K_1^{3/2}\sqrt{d}}\epsilon, \ \mathbf{x} \in D'$$

$$\|\mathbf{g}_1(\mathbf{x}, \hat{\mathcal{W}})\|_\infty \le \frac{\alpha_{kj}}{5K_1^{3/2}}\epsilon$$

No $\sqrt{d}$

Sample Complexity of original Dataset D

$$N = \mathcal{O}(K_1^{5/2} d^2 \epsilon^{-1} \kappa^{-1})$$

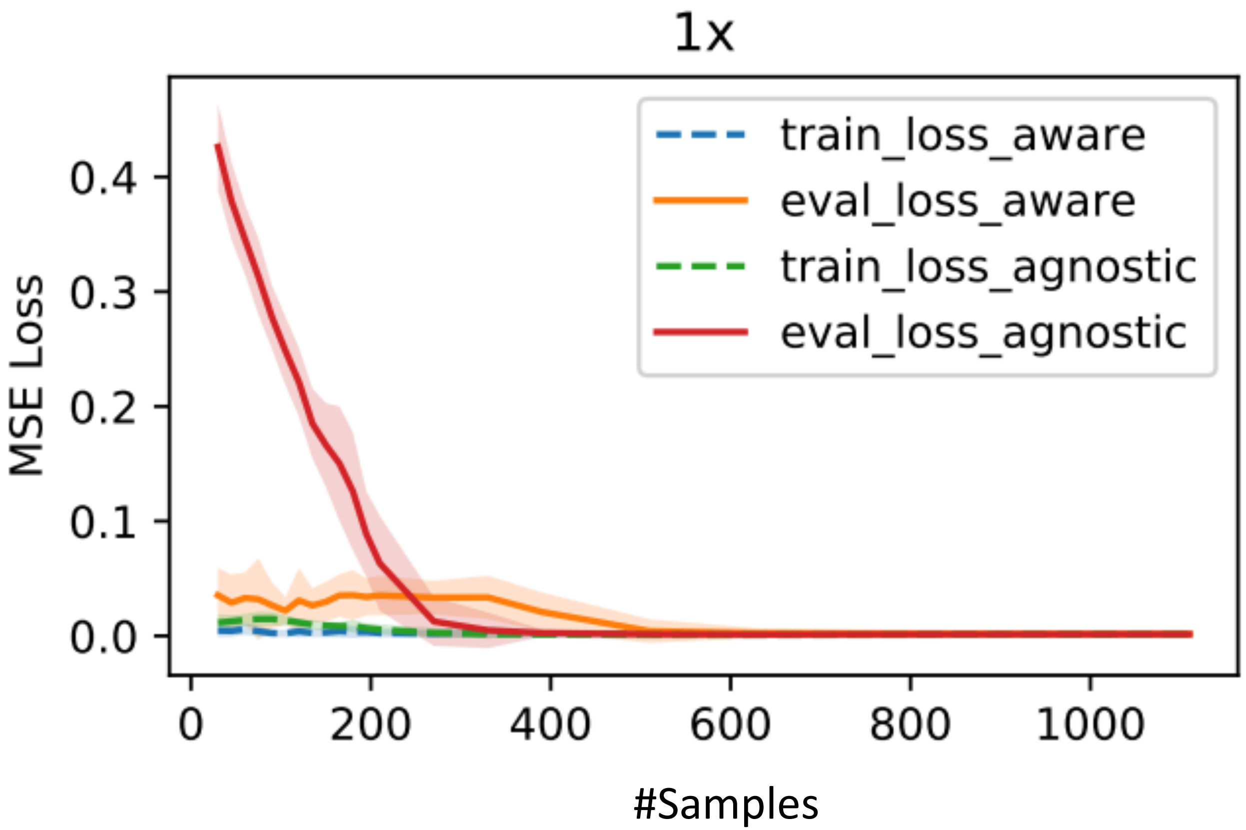
$$\mathcal{O}(K_1^{5/2} d \epsilon^{-1} \kappa^{-1})$$

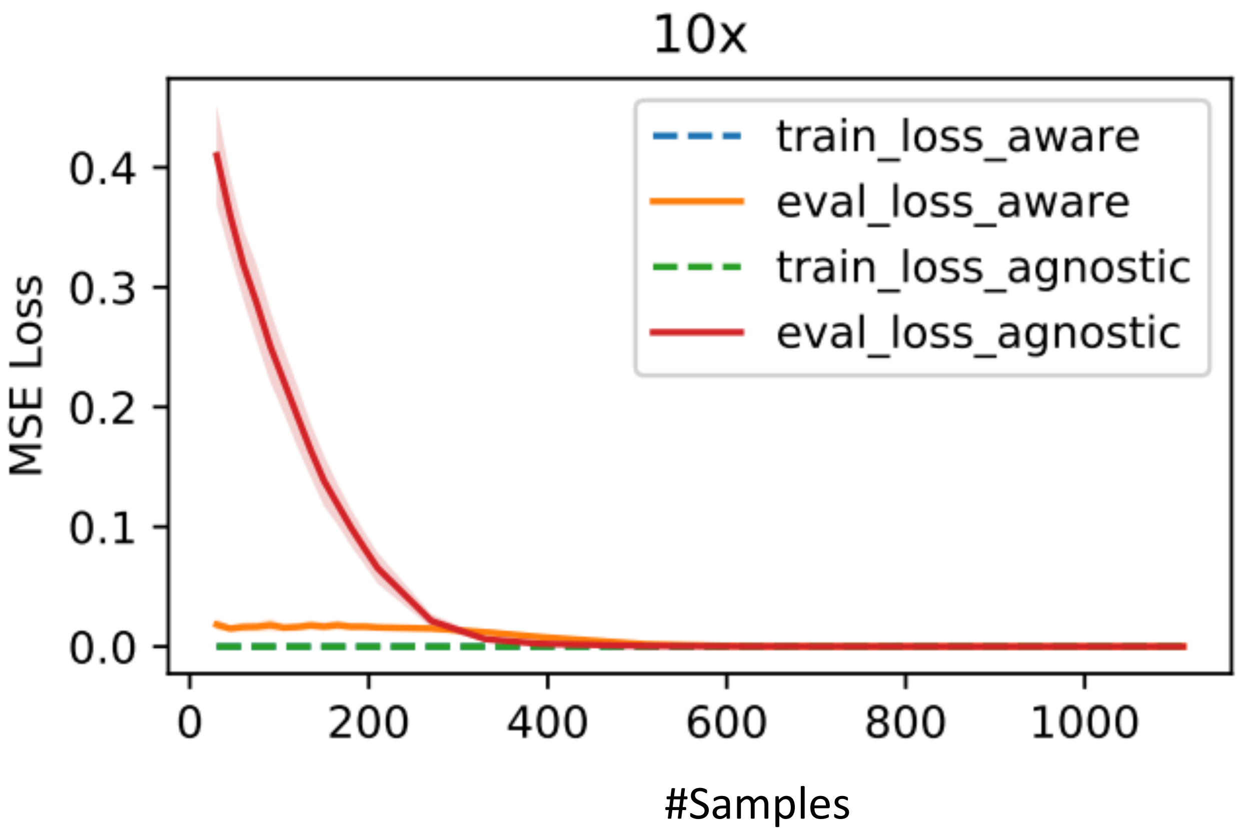
Linear w.r.t d

Teacher-agnostic augmentation
D' = Aug(D)
|D'| = (2d+1)|D|

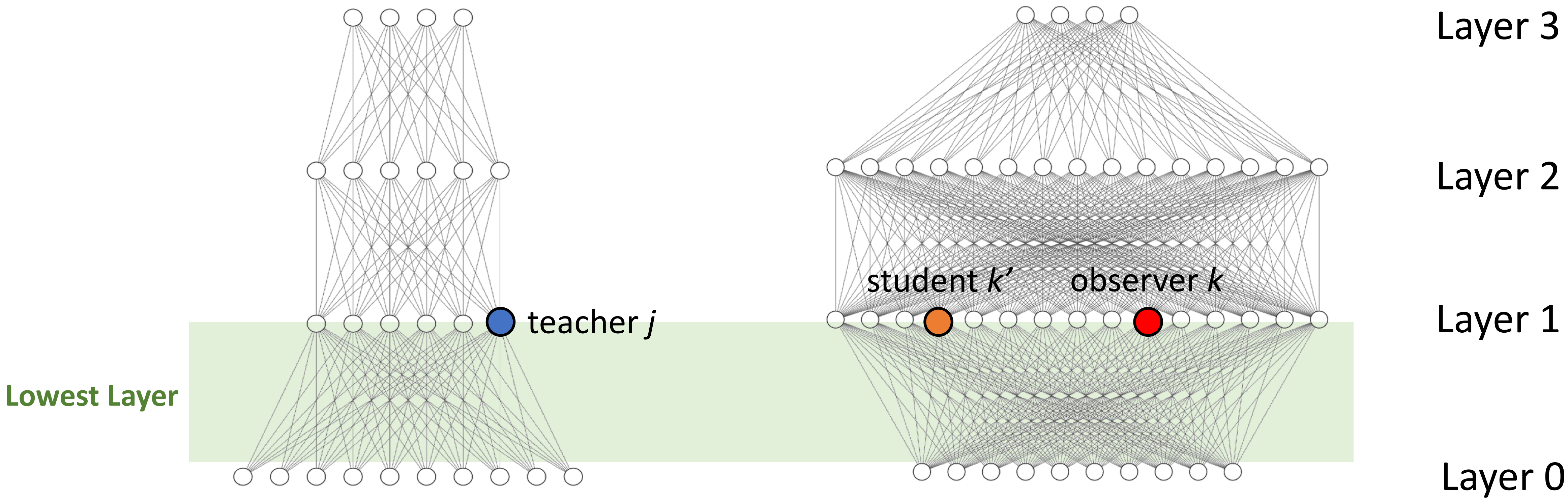
Teacher-aware augmentation
D' = Aug(D)
|D'| = m|D|

Teacher-Agnostic versus Teacher-aware





Multi-layer case



Small gradient

$$\|\mathbf{g}_1(\mathbf{x}, \hat{\mathcal{W}})\|_\infty \le \frac{\min_{R\in\mathcal{R}} \alpha_{kj}(R)}{5Q^{3/2}\sqrt{d}}\epsilon, \textit{ for } \mathbf{x} \in D'$$

Sample Complexity of original Dataset D

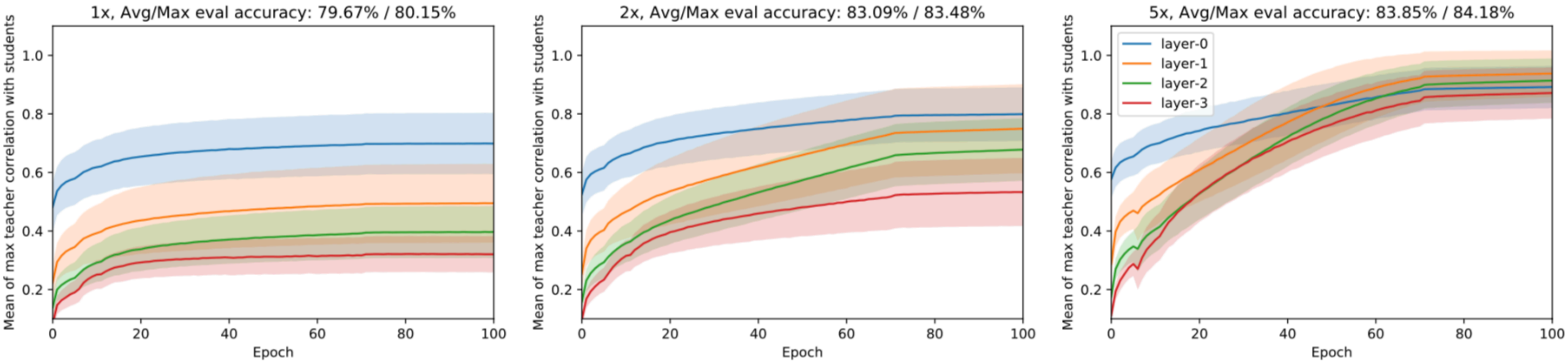
$$\mathcal{O}(Q^{5/2}d^2\epsilon^{-1}\kappa^{-1})$$

Q: #boundaries of hyperplanes (w.r.t network depth)

CIFAR 10

1. Train a conv teacher network of size 64-64-64-64.
2. **[Construct Oracle]** Prune the teacher network with [0.3,0.5,0.5,0.7] rate.
3. Then train a student network to mimic teacher‘s output (before softmax)

The student network has more parameters



Summary and Future Works

- Student Specialization in finite width and finite input dimension
  - Polynomial sample complexity in 2-layer ReLU networks.
  - Specialization in the lowest layer of deep ReLU networks
  - Experiments verify the claims.

- Future Works
  - Specialization at intermediate layers
  - Generalization Bound
  - Training Dynamics
  - Connect with empirical practices (e.g., network distillations).

Thanks!