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

Yuandong Tian [1]

Abstract

We consider a deep ReLU / Leaky ReLU student network trained from the output of a fixed teacher network of the same depth, with Stochastic Gradient Descent (SGD). The student network is *over-realized*: at each layer l, the number n_l of student nodes is more than that (m_l) of teacher. Under mild conditions on dataset and teacher network, we prove that when the gradient is small at every data sample, each teacher node is *specialized* by at least one student node *at the lowest layer*. For two-layer network, such specialization can be achieved by training on any dataset of *polynomial* size $\mathcal{O}(K^{5/2}d^3\epsilon^{-1})$ (sample size including augmentation) until the gradient magnitude drops to $\mathcal{O}(\epsilon/K^{3/2}\sqrt{d})$, where d is the input dimension, $K = m_1 + n_1$ is the total number of neurons in the lowest layer of teacher and student. To our best knowledge, we are the first to give polynomial sample complexity for student specialization of training two-layer (Leaky) ReLU networks with finite depth and width in teacher-student setting, and finite complexity for the lowest layer specialization in multi-layer case, without parametric assumption of the input (like Gaussian). Our theory suggests that teacher nodes with large fan-out weights get specialized first when the gradient is still large, while others are specialized with small gradient, which suggests inductive bias in training. This shapes the stage of training as empirically observed in multiple previous works. Experiments on synthetic and CIFAR10 verify our findings.

1. Introduction

While Deep Learning has achieved great success in different empirical fields (Silver et al., 2016; He et al., 2016; Devlin et al., 2018), it remains an open question how such networks can generalize to new data. As shown by empirical studies (e.g., (Zhang et al., 2017)), for deep models, training on real or random labels might lead to very different generalization behaviors. Without any assumption on the dataset, the generalization bound can be vacuous, i.e., the same network with zero training error can either generalize well or perform randomly in the test set.

One way to impose that is via *teacher-student* setting: given N input samples, a fixed teacher network provides the label for a student to learn. The setting has a long history (Gardner & Derrida, 1989) and offers multiple benefits. First, while worst-case performance on arbitrary data distributions may not be a good model for real structured dataset and can be hard to analyze, using a teacher network implicitly enforces an inductive bias and could potentially lead to better generalization bound. Second, the existence of teacher is often guaranteed by expressiblility (e.g., even one-hidden layer can fit any function (Hornik et al., 1989)). Finally, a reference network could facilitate and deepen our understanding of the training procedure.

Specialization, i.e., a student node becomes increasingly correlated with a teacher node during training (Saad & Solla, 1996), is one important topic in this setup. If all student nodes are specialized to the teacher, then student tends to output the same as the teacher and generalization performance can be expected. Empirically, it has been observed in 2-layer networks (Saad & Solla, 1996; Goldt et al., 2019) and multi-layer networks (Tian et al., 2019; Li et al., 2016), in both synthetic and real dataset. In contrast, theoretical analysis is limited with strong assumptions (e.g., Gaussian inputs, infinite input dimension, local convergence, 2-layer setting, small number of hidden nodes).

In this paper, we analyze student specialization when both teacher and student are deep ReLU / Leaky ReLU (Maas et al., 2013) networks. Similar to (Goldt et al., 2019), the student is *over-realized* compared to the teacher: at each layer l, the number n_l of student nodes is larger than

[1] Facebook AI Research. Correspondence to: Yuandong Tian <yuandong@fb.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020. Copyright 2020 by the author(s).

the number m_l of teacher ($n_l \geq m_l$). Although over-realization is different from *over-parameterization*, i.e., the total number of parameters in the student model is larger than the training set size N, over-realization directly correlates with the width of networks and is a measure of over-parameterization. With finite input dimension, we show rigorously that when gradient at each training sample is small (i.e., the interpolation setting as suggested in (Ma et al., 2017; Liu & Belkin, 2018; Bassily et al., 2018)), student nodes *specialize* to teacher nodes *at the lowest layer*: **each teacher node is aligned with at least one student node**. This explains one-to-many mapping between teacher and student nodes and the existence of un-specialized student nodes, as observed empirically in (Saad & Solla, 1996).

Our setting is more relaxed than previous works. **(1)** While statistical mechanics approaches (Saad & Solla, 1996; Goldt et al., 2019; Gardner & Derrida, 1989; Aubin et al., 2018) assume both the training set size N and the input dimension d goes to infinite (i.e., the thermodynamics limits) and assume Gaussian inputs, our analysis allows finite d and N, and impose *no* parametric constraints on the input data distribution. **(2)** While Neural Tangent Kernel (Jacot et al., 2018; Du et al., 2019) and mean-field approaches (Mei et al., 2018) requires infinite (or very large) width, our setting applies to finite width as long as student is slightly over-realized ($n_l \geq m_l$). **(3)** While recent works (Hu et al., 2020) show convergence in classification for teacher-student setting when N grows exponentially w.r.t. number of teacher nodes (including 2-layer case), we address student specialization in regression problems and show polynomial sample complexity in 2-layer case.

We verify our findings with numerical experiments. For deep ReLU nodes, we show one-to-many specialization and existence of un-specialized nodes at each hidden layer, on both synthetic dataset and CIFAR10. We also perform ablation studies about the effect of student over-realization and how strong teacher nodes learn first compared to the weak ones, as suggested by our theory.

2. Related Works

Student-teacher setting. This setting has a long history (Engel & Van den Broeck, 2001; Gardner & Derrida, 1989; Saad & Solla, 1996; 1995; Freeman & Saad, 1997; Mace & Coolen, 1998) and recently gains increasing interest (Goldt et al., 2019; Aubin et al., 2018) in analyzing 2-layered network. The seminar works (Saad & Solla, 1996; 1995) studies 1-hidden layer case from statistical mechanics point of view in which the input dimension goes to infinity, or so-called *thermodynamics limits*. They study symmetric solutions and locally analyze the symmetric breaking behavior and onset of *specialization* of the student nodes towards the teacher. Recent follow-up works (Goldt et al., 2019) makes the analysis rigorous and empirically shows that random initialization and training with SGD indeed gives student specialization in 1-hidden layer case, which is consistent with our experiments. With the same assumption, (Aubin et al., 2018) studies phase transition property of specialization in 2-layer networks with small number of hidden nodes using replica formula. In these works, inputs are assumed to be Gaussian and step or Gauss error function is used as nonlinearity. Few works study teacher-student setting with more than two layers. (Allen-Zhu et al., 2019a) shows the recovery results for 2 and 3 layer networks, with modified SGD, batchsize 1 and heavy over-parameterization.

In comparison, our work shows that specialization happens around SGD critical points in the lowest layer for deep ReLU networks, without any parametric assumptions of input distribution.

Local minima is Global. While in deep linear network, all local minima are global (Laurent & Brecht, 2018; Kawaguchi, 2016), situations are quite complicated with nonlinear activations. While local minima is global when the network has invertible activation function and distinct training samples (Nguyen & Hein, 2017; Yun et al., 2018) or Leaky ReLU with linear separate input data (Laurent & von Brecht, 2017), multiple works (Du et al., 2018; Ge et al., 2017; Safran & Shamir, 2017; Yun et al., 2019) show that in GD case with population or empirical loss, spurious local minima can happen even in two-layered network. Many are specific to two-layer and hard to generalize to multi-layer setting. In contrast, our work brings about a generic formulation for deep ReLU network and gives recovery properties in the student-teacher setting.

Learning very wide networks. Recent works on Neural Tangent Kernel (Jacot et al., 2018; Du et al., 2019; Allen-Zhu et al., 2019b) show the global convergence of GD for multi-layer networks with infinite width. (Li & Liang, 2018) shows the convergence in one-hidden layer ReLU network using GD/SGD to solution with good generalization, when the input data are assumed to be clustered into classes. Both lines of work assume heavily over-parameterized network, requiring polynomial growth of number of nodes with respect to the number of samples. (Chizat & Bach, 2018) shows global convergence of over-parameterized network with optimal transport. (Tian et al., 2019) assumes mild over-realization and gives convergence results for 2-layer network when a subset of the student network is close to the teacher. Our work extends it to multilayer cases with much weaker assumptions.

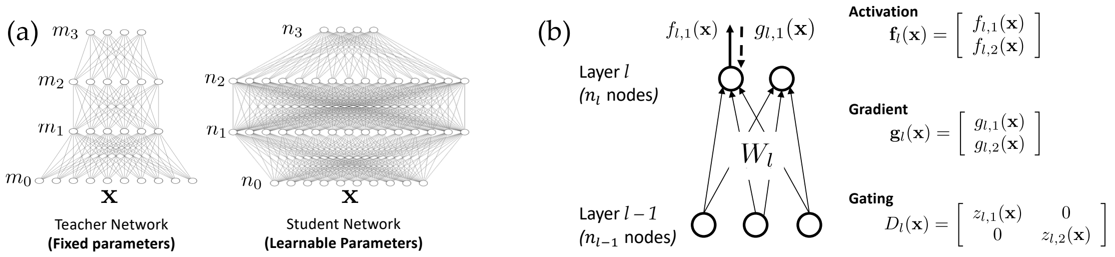

(a)
m_3
m_2
m_1
m_0
$\mathbf{x}$
Teacher Network
(Fixed parameters)
n_3
n_2
n_1
n_0
$\mathbf{x}$
Student Network
(Learnable Parameters)
(b)
$f_{l,1}(\mathbf{x})$
$g_{l,1}(\mathbf{x})$
Layer l
(n_l nodes)
W_l
Layer $l-1$
(n_{l-1} nodes)
Activation
$\mathbf{f}_l(\mathbf{x}) = \begin{bmatrix} f_{l,1}(\mathbf{x}) \\ f_{l,2}(\mathbf{x}) \end{bmatrix}$
Gradient
$\mathbf{g}_l(\mathbf{x}) = \begin{bmatrix} g_{l,1}(\mathbf{x}) \\ g_{l,2}(\mathbf{x}) \end{bmatrix}$
Gating
$D_l(\mathbf{x}) = \begin{bmatrix} z_{l,1}(\mathbf{x}) & 0 \\ 0 & z_{l,2}(\mathbf{x}) \end{bmatrix}$

Figure 1. Problem Setup. **(a)** Student-teacher setting. The student network learns from the output of a fixed teacher network via stochastic gradient descent (SGD). **(b)** Notations. All low cases are scalar, bold lower case are column vectors (row vectors are always with a transpose) and upper cases are matrices.

3. Mathematical Framework

Notation. Consider a student network and its associated teacher network (Fig. 1(a)). Denote the input as $\mathbf{x}$. We focus on multi-layered networks with $\sigma(\cdot)$ as ReLU / Leaky ReLU (Maas et al., 2013) nonlinearity. We use the following equality extensively: $\sigma(x) = \mathbb{I}[x \geq 0]x + \mathbb{I}[x < 0]c_{\text{leaky}}x$, where $\mathbb{I}[\cdot]$ is the indicator function and c_{leaky} is the leaky ReLU constant ($c_{\text{leaky}} = 0$ for ReLU). For node j, $f_j(\mathbf{x})$, $z_j(\mathbf{x})$ and $g_j(\mathbf{x})$ are its activation, gating function and backpropagated gradient *after the gating*.

Both teacher and student networks have L layers. The input layer is layer 0 and the topmost layer (layer that is closest to the output) is layer L. For layer l, let m_l be the number of teacher node while n_l be the number of student node. The weights $W_l \in \mathbb{R}^{(n_{l-1}+1)\times n_l}$ refers to the weight matrix that connects layer $l-1$ to layer l on the student side, with bias terms included. $W_l = [\mathbf{w}_{l,1}, \mathbf{w}_{l,2}, \ldots, \mathbf{w}_{l,n_l}]$ where each $\mathbf{w} = [\tilde{\mathbf{w}}; b] \in \mathbb{R}^{n_{l-1}+1}$ is the weight vector. Here $\tilde{\mathbf{w}}$ is the weight and b is the bias.

Without loss of generality, we assume teacher weights $\mathbf{w} = [\tilde{\mathbf{w}}, b] \in \mathbb{R}^{d+1}$ are *regular*, except for the topmost layer $l = L$.

Definition 1 (Regular Weight)**.** *A weight vector* $\mathbf{w} = [\tilde{\mathbf{w}}, b]$ *is* regular *if* $\|\tilde{\mathbf{w}}\|_2 = 1$.

Let $\mathbf{f}_l(\mathbf{x}) = [f_{l,1}(\mathbf{x}), \ldots, f_{l,n_l}(\mathbf{x}), 1]^\intercal \in \mathbb{R}^{n_l+1}$ be the activation vector of layer l, $D_l(\mathbf{x}) = \text{diag}[z_{l,1}(\mathbf{x}), \ldots, z_{l,n_l}(\mathbf{x}), 1] \in \mathbb{R}^{(n_l+1)\times(n_l+1)}$ be the diagonal matrix of gating function (for ReLU it is either 0 or 1), and $\mathbf{g}_l(\mathbf{x}) = [g_{l,1}(\mathbf{x}), \ldots, g_{l,n_l}(\mathbf{x}), 1]^\intercal \in \mathbb{R}^{n_l+1}$ be the backpropated gradient vector. The last 1s are for bias. By definition, $\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \in \mathbb{R}^{n_0}$ is the input and $m_0 = n_0$. Note that $\mathbf{f}_l(\mathbf{x})$, $\mathbf{g}_l(\mathbf{x})$ and $D_l(\mathbf{x})$ are all dependent on $\mathcal{W}$. For brevity, we often use $\mathbf{f}_l(\mathbf{x})$ rather than $\mathbf{f}_l(\mathbf{x}; \mathcal{W})$.

All notations with superscript * are from the teacher, only dependent on the teacher and remains the same throughout the training. For the topmost layer, $D_L^*(\mathbf{x}) = D_L(\mathbf{x}) \equiv I_{C\times C}$ since there is no ReLU gating, where C is the dimension of output for both teacher and student. With the notation, gradient descent is:

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

In SGD, the expectation $\mathbb{E}_{\mathbf{x}}[\cdot]$ is taken over a batch. In GD, it is over the entire dataset.

Let $E_j = \{\mathbf{x} : f_j(\mathbf{x}) > 0\}$ the activation region of node j and $\partial E_j = \{\mathbf{x} : f_j(\mathbf{x}) = 0\}$ its decision boundary.

MSE Loss. We use the output of teacher as the supervision:

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

For convenience, we also define the following qualities:

Definition 2. *Define* $V_l \in \mathbb{R}^{C\times n_l}$ *and* $V_l^* \in \mathbb{R}^{C\times m_l}$ *in a top-down manner (for top-most layer* $V_L = V_L^* := I_{C\times C}$*):*

$$V_{l-1} := V_l D_l W_l^\intercal, \quad V_{l-1}^* := V_l^* D_l^* W_l^{*\intercal} \tag{3}$$

We further define $A_l := V_l^\intercal V_l^* \in \mathbb{R}^{n_l\times m_l}$ *and* $B_l = V_l^\intercal V_l \in \mathbb{R}^{n_l\times n_l}$*. Each element of* A_l *is* $\alpha_{jj'}^l = \mathbf{v}_j^\intercal \mathbf{v}_{j'}^*$*. Similarly, each element of* B_l *is* $\beta_{jj'}^l = \mathbf{v}_j^\intercal \mathbf{v}_{j'}$*.*

In this paper, we want to know whether *the student nodes specialize to teacher nodes at the same layers* during training? We define student node specialization as follows:

Definition 3 (ϵ-aligned)**.** *Two nodes* j *and* j' *are aligned if their weights* $[\tilde{\mathbf{w}}_j, b_j]$ *and* $[\tilde{\mathbf{w}}_{j'}, b_{j'}]$ *satisfy:*

$$\sin\tilde{\theta}_{jj'} \leq \epsilon, \qquad |b_j - b_{j'}| \leq \epsilon, \tag{4}$$

where $\tilde{\theta}_{jj'}$ *is the angle between* $\tilde{\mathbf{w}}_j$ *and* $\tilde{\mathbf{w}}_{j'}$*.*

One might wonder this is hard since the student's intermediate layer receives no *direct supervision* from the corresponding teacher layer, but relies only on backpropagated gradient. Surprisingly, Lemma 1 shows that the supervision is implicitly carried from layer to layer via gradient:

Lemma 1 (Recursive Gradient Rule)**.** *At layer* l*, the backpropagated* $\mathbf{g}_l(\mathbf{x})$ *satisfies*

$$\mathbf{g}_l(\mathbf{x}) = D_l(\mathbf{x})\left[A_l(\mathbf{x})\mathbf{f}_l^*(\mathbf{x}) - B_l(\mathbf{x})\mathbf{f}_l(\mathbf{x})\right] \tag{5}$$

Remark. Lemma 1 applies to arbitrarily deep ReLU networks and allows $n_l \neq m_l$. In particular, student can be over-realized. Note that A_l, B_l, V_l, V_l^* all depends on $\mathbf{x}$. Due to the property of (Leaky) ReLU, $A_l(\mathbf{x})$ and $B_l(\mathbf{x})$ are piecewise constant functions (Corollary 2 in Appendix).

Relationship to Distillation. Our setting is closely related but different from network distillation extensively used in practice. From our point of view, both the "pre-trained teacher" and "condensed student" in network distillation are large student networks, and the dataset are samples from an (inaccessible) and small teacher (ie., oracle) network. We will study connections to network distillation in our future works.

4. Simple Example: Two-layer Network, Zero Gradient and Infinite Samples

We first consider a two-layer ReLU network trained with infinite samples until the gradient is zero at every training sample. This ideal case reveals key structures of student specialization with intuitive proof. In 2-layer case, $A_1(\mathbf{x})$ and $B_1(\mathbf{x})$ are constant with respect to $\mathbf{x}$, since there is no ReLU gating at the top layer $l = 2$. The subscript l is omit for brevity. Sec. 5 proposes main theorems that consider finite sample, small gradient and multi-layer networks.

Obviously, some teacher networks cannot be reconstructed, e.g., a teacher network that outputs identically zero. Therefore, some assumptions on teacher network are needed.

Assumption 1. *(1) We have an infinite training set R_0, an* open *set around the origin. (2) Any two teacher weights are not co-linear; (3) The boundary of any teacher node j intersects with R_0: $R_0 \cap \partial E_j^* \neq \emptyset$. See Fig. 2.*

Intuitively, we want the teacher nodes to be well-separated, and all boundaries of teachers pass through a dataset, which reveals their nonlinear nature. "Open set" means R has interior and is *full rank.*

Definition 4 (Observer)**.** *Node k is an observer of node j if $R_0 \cap E_k \cap \partial E_j \neq \emptyset$. See Fig. 2(d).*

We assume the following *zero gradient condition*, which is feasible since student network is over-realized.

Assumption 2 (Zero Batch Gradient in SGD)**.** *For every batch $\mathcal{B} \subseteq R_0$, $\dot{W}_l = \mathbb{E}_{\mathbf{x}\in\mathcal{B}}\left[\mathbf{f}_{l-1}(\mathbf{x})\mathbf{g}_l^\intercal(\mathbf{x})\right] = 0$.*

Given these, we now arrive at the following theorem:

Theorem 1 (Student Specialization, 2-layers)**.** *If Assumption 1 and Assumption 2 hold, and **(1)** A teacher node j is observed by a student node k, and **(2)** $\alpha_{kj} \neq 0$ (defined in Def. 2), then there exists one student node k' with 0-alignment (exact alignment) with j.*

Proof sketch. Note that ReLU activations $\sigma(\mathbf{w}_k^\intercal\mathbf{x})$ are mutually linear independent, if their boundaries are within the training region R_0. On the other hand, the gradient of each student node k *when active*, is $\boldsymbol{\alpha}_k^\intercal\mathbf{f}_1(\mathbf{x}) - \boldsymbol{\beta}_k^\intercal\mathbf{f}_1(\mathbf{x}) = 0$, a linear combination of teacher and student nodes (note that $\boldsymbol{\alpha}_k^\intercal$ and $\boldsymbol{\beta}_k^\intercal$ are k-th rows of A_1 and B_1). Therefore, zero gradient means that the summation of coefficients of co-linear ReLU nodes is zero. Since teachers are not co-linear (Assumption 1), a non-zero coefficient $\alpha_{kj} \neq 0$ for teacher j means that it has to be co-linear with at least one student node, so that the summation of coefficients is zero. For details, please check detailed proofs in the Appendix.

Note that for one teacher node, multiple student nodes can specialize to it. For deep linear models, specialization does not happen since a linear subspace spanned by intermediate layer can be represented by different sets of bases.

Note that a necessary condition of a reconstructed teacher node is that its boundary is in the active region of student, or is *observed* (Definition 4). This is intuitive since a teacher node which behaves like a linear node is partly indistinguishable from a bias term.

For student nodes that are not aligned with any teacher node, if they are observed by other student nodes, then following a similar logic, we have the following:

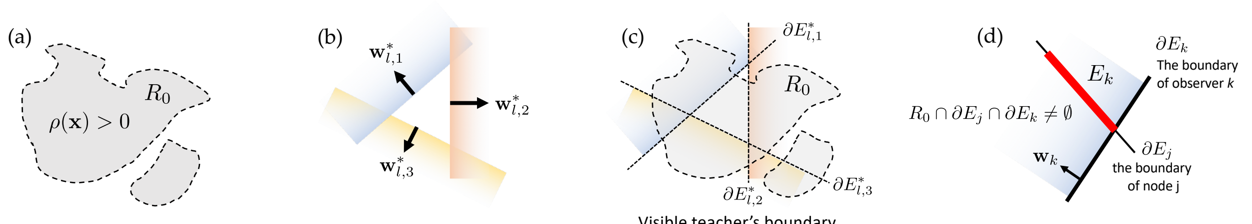
Theorem 2 (Un-specialized Student Nodes are Prunable)**.** *If an unaligned student k has C independent observers, i.e., the C-by-C matrix stacking the fan-out weights $\mathbf{v}$ of these observers is full rank, then $\sum_{k'\in\text{co-linear}(k)} \mathbf{v}_{k'}\|\mathbf{w}_{k'}\| = \mathbf{0}$. If node k is not co-linear with other students, then $\mathbf{v}_k = \mathbf{0}$.*

Corollary 1. *With sufficient observers, the contribution of all unspecialized student nodes is zero.*

Theorem 2 and Corollary 1 explain why network pruning is possible (LeCun et al., 1990; Hassibi et al., 1993; Hu et al., 2016). Note that a related theorem (Theorem 6) in (Laurent & von Brecht, 2017) studies 2-layer network with scalar output and linear separable input, and discusses characteristics of individual data point contributing loss in a local minima of GD. In our paper, no linear separable condition is imposed.

Network representations. To compare the intermediate representation of networks trained with different initialization, previous works use Canonical Correlation Analysis (CCA) (Hardoon et al., 2004) and its variants (e.g., SVCCA (Raghu et al., 2017)) that linearly transform activations of two networks into a common aligned space. This can be explained by Theorem 1 and Theorem 2: multiple student nodes who specialize to one teacher node can be aligned together after linear transformation and un-aligned students can be suppressed by a null transform.

Connectivity between two low-cost solutions. Previous works (Garipov et al., 2018; Draxler et al., 2018) discov-



Region as a training set

Distinct teacher nodes

Visible teacher's boundary in the training set

Node *j* is **observed** by node *k*

Figure 2. Simple Example (Sec. 4). **(a-c)** Assumption 1: Training set is an *open* region R_0 in the input space. Teacher nodes ($l = 1$) are distinct. Teacher boundaries are *visible* in R_0. Here $\partial E^*_{l,j} \cap R_0 \neq \emptyset$ for $j = 1, 2, 3$. **(d)** The definition of observer (Definition 4).

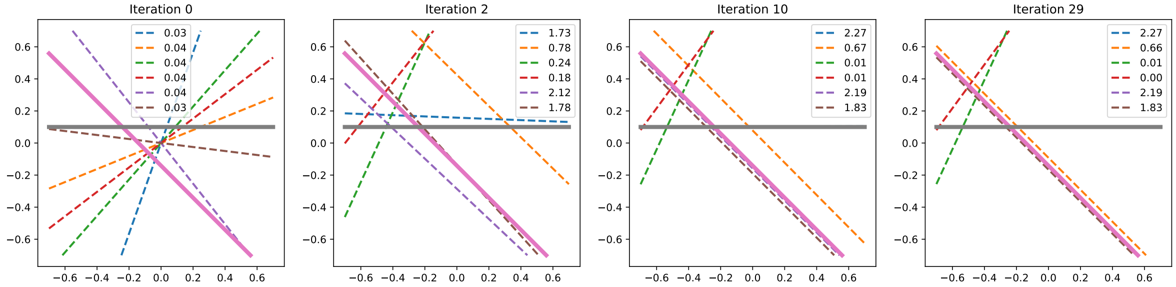


Figure 3. Convergence (2 dimension) for 2 teachers (solid line) and 6 students (dashed line). Legend shows $\|\mathbf{v}_k\|$ for student node k. $\|\mathbf{v}_k\| \to 0$ for nodes that are not aligned with teacher.

ered that low-cost solutions for neural networks can be connected via line segments, but not a single straight line. Our framework can explain this phenomenon using a construction similar to (Kuditipudi et al., 2019) but without the assumption of ϵ-dropout stableness of a trained network using Theorem 2. See Appendix E for the construction.

5. Main Theorems

In practice, stochastic gradient fluctuates around zero and only finite samples are available during training. In this case, we show a rough specialization still follows. For convenience, we define hyperplane band $I(\epsilon)$ as follows:

Definition 5 (Hyperplane band $I_{\mathbf{w}}(\epsilon)$)**.** $I_{\mathbf{w}}(\epsilon) = \{\mathbf{x} : |\mathbf{w}^\intercal\mathbf{x}| \leq \epsilon\}$. *We use* $I_j(\epsilon)$ *if* $\mathbf{w}$ *is from node* j*.*

Definition 6 ((η,μ)-Dataset)**.** *A dataset* D *is called* (η,μ)*-dataset, if there exists* $\eta,\mu > 0$ *so that for any regular weight* $\mathbf{w}$*, the number of samples in the hyperplane band* $N_D\left[I_{\mathbf{w}}(\epsilon)\right] \equiv N[D \cap I_{\mathbf{w}}(\epsilon)]$ *satisfies:*

$$N_D\left[I_{\mathbf{w}}(\epsilon)\right] \leq \eta\epsilon N_D + (d+1) \qquad (6)$$

and for any regular $\mathbf{w} = [\tilde{\mathbf{w}}, b]$ *with* $b = 0$ *(no bias):*

$$N_D\left[D\backslash I_{\mathbf{w}}\left(1/\epsilon\right)\right] \equiv N_D\left[|\tilde{\mathbf{w}}^\intercal\tilde{\mathbf{x}}| \geq \frac{1}{\epsilon}\right] \leq \mu\epsilon^2 N_D \qquad (7)$$

where N_D *is the size of the dataset.*

Intuitively, Eqn. 6 means that each point of the dataset is scattered around and any hyperplane band $|\mathbf{w}^\intercal\mathbf{x}| \leq \epsilon$ cannot cover them all. It is in some sense a high-rank condition for the dataset. The additional term $d+1$ exists because there always exists a hyperplane $\mathbf{w}_0$ that passes any $d+1$ points (excluding degenerating case). Therefore, $N_D[\mathbf{w}_0^\intercal\mathbf{x} = 0] = d+1$. Eqn. 7 can be satisfied with dataset sampled by any zero-mean distribution with finite variance, due to Chebyshev's inequality: $\mathbb{P}[|\tilde{\mathbf{w}}^\intercal\tilde{\mathbf{x}}| \geq 1/\epsilon] \leq \epsilon^2\mathrm{Var}(\tilde{\mathbf{w}}^\intercal\tilde{\mathbf{x}})$.

Assumption 3. *(a) Two teacher nodes* $j \neq j'$ *are not* ϵ_0*-aligned. (b) The boundary band* $I_j(\epsilon)$ *of each teacher* j *overlaps with the dataset:*

$$N_D[I_j(\epsilon)] \geq \tau\epsilon N_D \qquad (8)$$

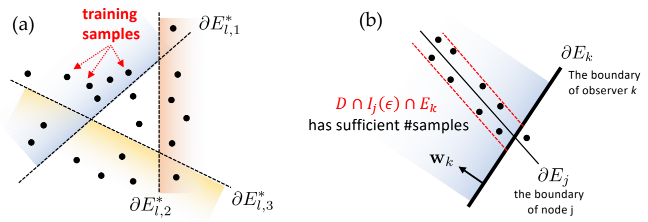
Intuitively, this means that we have sufficient samples near the boundary of each teacher nodes, in order to correctly identify the weights $\mathbf{w}^*_j$ of each teacher node j. Again, a teacher node that are not visible from the training set cannot be reconstructed. For this purpose, the reader might wonder why Eqn. 8 does not impose constraints that there need to be input samples on both sides of the teacher boundary ∂E^*_j. The answer is that we also assume proper data augmentation (see Definition 7).

5.1. Two Layer Case

The following two theorems show that with small gradient and polynomial number of samples, a rough specialization still follows in 2-layer network. Here let $K_1 = m_1 + n_1$.

Definition 7 (Augmentation)**.** *Given a dataset* D*, we construct* $\mathrm{aug}(D)$ *as* ***(a)*** *Teacher-agnostic:*

$$\mathrm{aug}(D) = \{\mathbf{x} \pm 2\epsilon\tilde{\mathbf{e}}_u/cK_1^{3/2},\ \mathbf{x} \in D,\ u = 1,\ldots,d\} \cup D$$



Visible teacher's boundary in the training set

Node j is **observed** by node k

Figure 4. Settings in Main Theorems (Sec. 5). **(a)** Assumption 3: Teacher boundaries are *visible* in dataset D. **(b)** The definition of observer now incorporates sample counts.

*where $\tilde{\mathbf{e}}_k$ is axis-aligned unit directions with $\|\tilde{\mathbf{e}}_k\| = 1$ or **(b)** Teacher-aware:*

$$\mathrm{aug}(D, \mathcal{W}^*) = \{\mathbf{x} \pm 2\epsilon\tilde{\mathbf{w}}^*_j/cK_1^{3/2},\ \ \mathbf{x} \in D \cap I_{\mathbf{w}^*_j}(\epsilon)\} \cup D$$

In both definitions, c is a constant related to (η, μ) of D (See proof of Theorem 6 in Appendix).

With the data augmentation, a polynomial number of samples suffice for student specialization to happen.

Theorem 3 (Two-layer Specialization with Polynomial Samples)**.** *For $0 < \epsilon \leq \epsilon_0$ and $0 < \kappa \leq 1$, for any finite dataset D with $N = \mathcal{O}(K_1^{5/2}d^2\epsilon^{-1}\kappa^{-1})$, for any teacher satisfying Assumption 3 and student trained on $D' = \mathrm{aug}(D)$ whose weight $\hat{\mathcal{W}}$ satisfies:*

(1) For $0 < \epsilon < \epsilon_0$, $I_j(\epsilon)$ of a teacher j is observed *by a student node k: $N_D[I_j(\epsilon) \cap E_k] \geq \kappa N_D[I_j(\epsilon)]$;*

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

then there exists a student k' so that (j, k') is ϵ-aligned.

Theorem 4 (Two-layer Specialization with Teacher-aware Dataset with Polynomial Samples)**.** *For $0 < \epsilon \leq \epsilon_0$ and $0 < \kappa \leq 1$, for any finite dataset D with $N = \mathcal{O}(K_1^{5/2}d\epsilon^{-1}\kappa^{-1})$, given a teacher network $\mathcal{W}^*$ satisfying Assumption 3 and student trained on $D' = \mathrm{aug}(D, \mathcal{W}^*)$ whose weight $\hat{\mathcal{W}}$ satisfies:*

(1) For $0 < \epsilon < \epsilon_0$, $I_j(\epsilon)$ of a teacher j is observed *by a student node k: $N_D[I_j(\epsilon) \cap E_k] \geq \kappa N_D[I_j(\epsilon)]$;*

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

then there exists a student k' so that (j, k') is ϵ-aligned.

Remark. Note that the definition of observation changes slightly in the finite sample setting: a sufficient number of samples needs to be observed in order to show convergence. Note that Theorem 3 contains an additional $\sqrt{d}$ factor in the gradient condition, due to teacher-agnostic augmentation. In fact, following the same proof idea, to remove the factor $\sqrt{d}$, exponential number of samples are needed, or knowledge of the teacher network (Theorem 4). This also suggests that for any teacher network $\mathbf{f}^*(\mathbf{x})$, there exists compatible input distribution so that the dataset $\{(\mathbf{x}, \mathbf{f}^*(\mathbf{x}))\}_{\mathbf{x} \in D}$ can be easy to learn.

5.2. Multi-layer Case

As in the 2-layer case, we can use similar intuition to analyze the behavior of the lowest layer for deep ReLU networks, thanks to Lemma 1 which holds for arbitrarily deep ReLU networks. In this case, $A_1(\mathbf{x})$ and $B_1(\mathbf{x})$ are no longer constant over $\mathbf{x}$, but are piece-wise constant. Note that $A_1(\mathbf{x})$ and $B_1(\mathbf{x})$ might contain exponential number of regions, $\mathcal{R} = \{R_0, R_1, \ldots, R_J\}$, where in each region R, $A_l(\mathbf{x})$ and $B_l(\mathbf{x})$ are constants.

As intersection of regions, the number boundaries are also exponential. The underlying intuition is that for each intermediate node, its boundary is "bended" to another direction, whenever the boundary meets with any boundary of its input nodes (Hanin & Rolnick, 2019b; Rolnick & Kording, 2019; Hanin & Rolnick, 2019a). All these boundaries will be reflected in the input region. Therefore, the number Q of hyper plane boundaries is exponential with respect to L, leading to exponential sample complexity.

Theorem 5 (Multi-layer alignment, Lowest Layer)**.** *Given $0 < \epsilon \leq \epsilon_0$, for any finite dataset D with $N = \mathcal{O}(Q^{5/2}d^2\epsilon^{-1}\kappa^{-1})$, if the first layer of the deep teacher network satisfies Assumption 3 and any student weight at the first layer $\hat{\mathcal{W}}_1$ satisfies:*

(1) For $0 < \epsilon < \epsilon_0$, $I_j(\epsilon)$ of a teacher j is observed by a student node k: $N_D[I_j(\epsilon) \cap E_k] \geq \kappa N_D[I_j(\epsilon)]$;

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

then there exists student node k' so that (j, k') is ϵ-aligned.

6. Discussions

The theorems suggest a few interesting consequences:

Strong and weak teacher nodes. Theorems show that the convergence is dependent on $\alpha_{kj} = \mathbf{v}_k^\intercal \mathbf{v}^*_j$, where $\mathbf{v}^*_j$ is the j-th column of V_1^* and $\mathbf{v}_k$ is the k-th column of V_1^*. Given the same magnitude of gradient norm, *strong* teacher j (large $\|\mathbf{v}^*_j\|$ and thus large α_{kj}) would achieve specialization, while weak teacher will not achieve specialization. This explains why early stopping (Caruana et al., 2001) could help and suggests how the inductive bias is created during training. We verify this behavior in our experiments.

Dataset matters. Theorem 3 and theorem 4 shows different sample complexity for datasets that are augmented with different augmentation methods, one is teacher-agnostic while the other is teacher-aware. This shows that if dataset is *compatible* with the underlying teacher network, the specialization would be much faster. We also verify this behavior in our experiments, showing that training with teacher-aware dataset yields much faster convergence as well as stronger generalization, given very few number of samples.

The role of over-realization. Theorem 1 suggests that over-realization (more student nodes in the hidden layer $l = 1$) is important. More student nodes mean more observers, and the existence argument in these theorems is more likely to happen and more teacher nodes can be covered by student, yielding better generalization.

Expected SGD conditions. In practice, the gradient conditions might hold in expectation (or high probability), e.g. $\mathbb{E}_t\left[\|\mathbf{g}_1(\mathbf{x})\|_\infty\right] \le \epsilon$. This means that $\|\mathbf{g}_1(\mathbf{x})\|_\infty \le \epsilon$ at least for some iteration t. All theorems still apply since they do not rely on past history of the weight or gradient.

7. Experiments

We first verify our theoretical finding on synthetic dataset generated by Gaussian distribution $\mathcal{N}(0, \sigma^2 I)$ with $\sigma = 10$. With other distributions (e.g., uniform $U[-1, 1]$), the result is similar. Appendix F gives details on how we construct the teacher network and training/eval dataset. Vanilla SGD is used with learning rate 0.01 and batchsize 16. To measure the degree of student specialization, we define ρ_{mean} as the mean of maximal normalized correlation ($\tilde{\mathbf{f}}_j$ is the normalized activation of node j over the evaluation set):

$$\rho_{\mathrm{mean}} = \underset{j \in \text{ teacher}}{\mathrm{mean}} \max_{j' \in \text{ student}} \rho_{jj'}, \quad \rho_{jj'} = \tilde{\mathbf{f}}_j^{*\intercal} \tilde{\mathbf{f}}_{j'}, \quad (9)$$

Strong/weak teacher nodes. To demonstrate the effect of strong and weak teacher nodes, we set up a diverse strength of teacher node by constructing the fan-out weights of teacher node j as follows:

$$\|\mathbf{v}_j^*\| \sim 1/j^p, \quad (10)$$

where p is the *teacher polarity factor* that controls how strong the energy decays across different teacher nodes. $p = 0$ means all teacher nodes have the same magnitude of fan-out weights, and large p means that the strength of teacher nodes are more polarized, i.e., some teacher nodes are very strong, some are very weak.

Two layer networks. First we verify Theorem 1 and Theorem 2 in the 2-layer setting. Fig. 5 shows for different degrees of over-realization (1×/2×/5×/10×), for nodes with weak specialization (i.e., its normalized correlation to the most correlated teacher is low, left side of the figure), their magnitudes of fan-out weights (y-axis) are small. The nodes with strong specialization have high fan-out weights.

Deep Networks. For deep ReLU networks (4-layer), we observe student specialization at *each* layer, shown in Fig. 6. We can also see the lowest layer converges better than the top layers at different sample sizes, in particular with MSE loss. Although our theory does not apply to Cross Entropy loss yet, empirically we still see specialization in multiple layers, in particular at the lowest layer.

7.1. Ablation studies

Strong/weak teacher node. We plot the average rate of a teacher node that is matched with at least one student node successfully (i.e., correlation > 0.95). Fig. 7 shows that stronger teacher nodes are more likely to be matched, while weaker ones may not be explained well, in particular when the strength of the teacher nodes are polarized (p is large). This is consistent with Theorem 3 which shows that a strong teacher node can be specialized even if the gradient magnitude is still relatively large, compared to a weak one. Fig. 8 further shows the dynamics of specialization for each teacher node: strong teacher node gets specialized fast, while it takes very long time to have students specialized to weak teacher nodes.

Effects of Over-realization. Over-realized student can explain more teacher nodes, while a student with 1× nodes has sufficient capacity to fit the teacher perfectly, it gets stuck despite long training.

Sample Complexity. Fig. 6 shows hows node correlation (and generalization) changes when sample complexity when we use MSE or Cross Entropy Loss. With more samples, generalization becomes better and ρ_{mean} also becomes better. Note that although our analysis doesn't include CE Loss, empirically we see ρ_{mean} grows, in particular at the lowest layer, when N becomes large.

Teacher-agnostic versus Teacher-aware. Theorem 4 and Theorem 3 shows that with different datasets (or same dataset but with different data augmentation), sample complexity can be very different. As shown in Fig. 9, if we construct a dataset with the knowledge of teacher, then a student trained on it will not overfit even with small number of samples. This shows the dependency of sample complexity with respect to the dataset.

CIFAR-10. We also experiment on CIFAR-10. We first pre-train a teacher network with 64-64-64-64 ConvNet (64 are channel sizes of the hidden layers, $L = 5$) on CIFAR-10 training set. Then the teacher network is pruned in a structured manner to keep strong teacher nodes. The student is over-realized based on teacher's remaining channels.

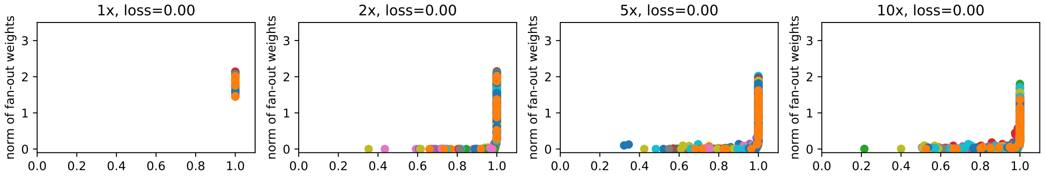


Figure 5. Student specialization of a 2-layered network with 10 teacher nodes and $1\times/2\times/5\times/10\times$ student nodes. For a student node k, we plot its degree of specialization (i.e., normalized correlation to its best correlated teacher) as the x coordinate and the fan-out weight norm $\|\mathbf{v}_k\|$ as the y coordinate. We plot results from 32 random seed. Student nodes of different seeds are in different color. An unspecialized student node has low correlations with teachers and low fan-out weight norm (Theorem 2). Higher p makes reconstruction of teacher node harder, in particular if the student network is not over-realized.

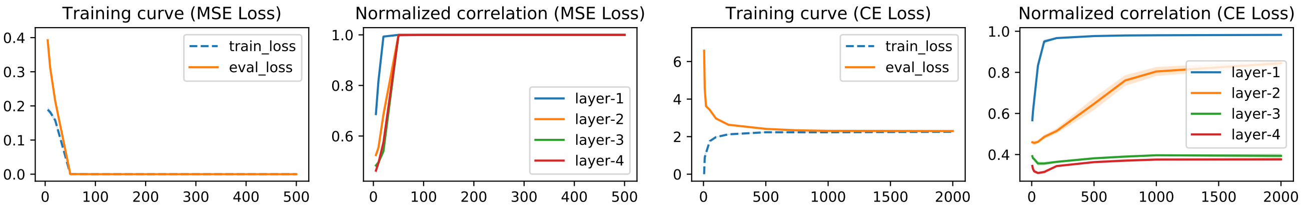


Figure 6. Relationship between evaluation loss and normalized correlation ρ_{mean} (y-axis), and sample complexity (x-axis, $\times 1000$) for MSE/Cross Entropy (CE) Loss function. Teacher is 4-layer with 50-75-100-125 hidden nodes and student is $2\times$ over-realization.

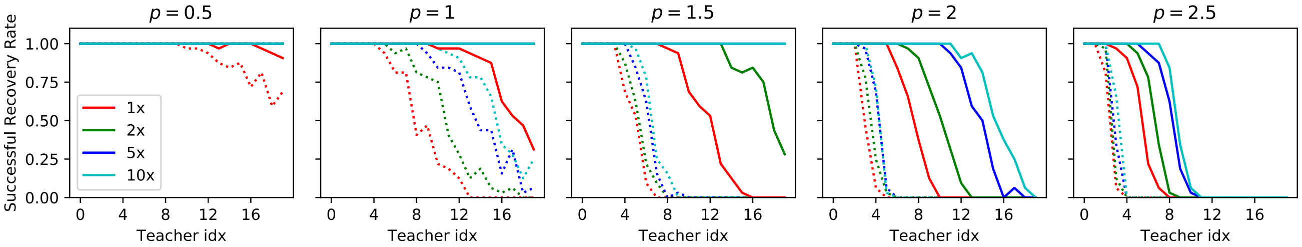


Figure 7. Success rate (over 32 random seeds) of recovery of 20 teacher nodes on 2-layer network at different teacher polarity p (Eqn. 10) under different over-realization. Dotted line: successful rate after 5 epochs. Solid line: successful rate after 100 epochs.

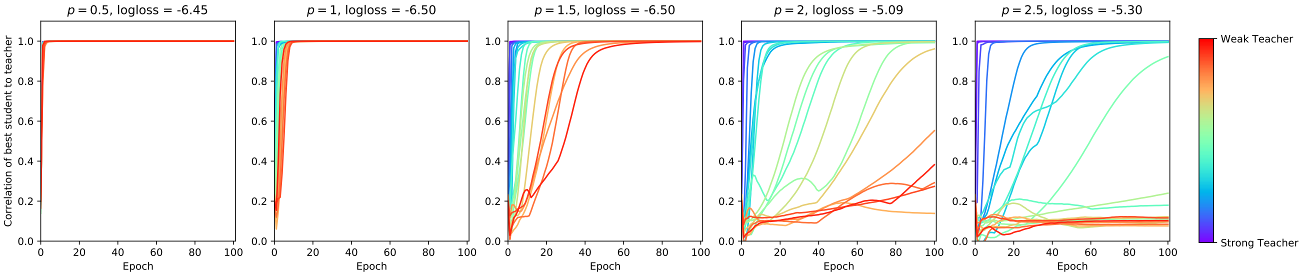


Figure 8. Dynamics of specialization (2-layer) over iterations. Each rainbow color represents one of the 20 teacher nodes (blue: strongest teacher, red: weakest). Strong teacher nodes (blue) can get specialized very quickly, while weak teacher nodes (red) is not specialized for a long time. The student is $5\times$ over-realization. Large p means strong polarity of teacher nodes (Eqn. 10).

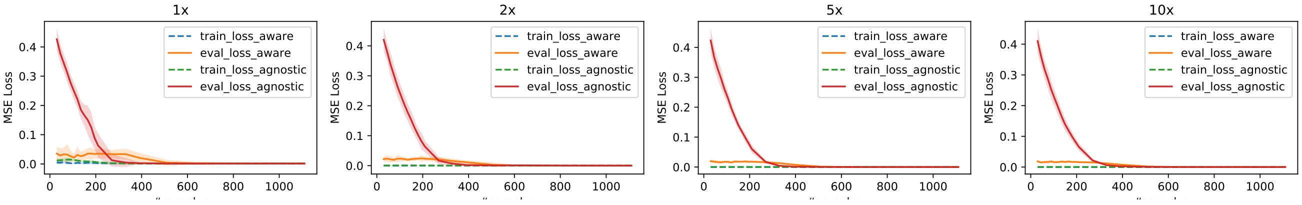


Figure 9. MSE Loss (y-axis) versus number of samples used (x-axis) using teacher-aware and teacher agnostic dataset in 2-layer network (10 random seeds). While training loss is low on both cases, teacher-aware dataset leads to substantial lower evaluation loss.

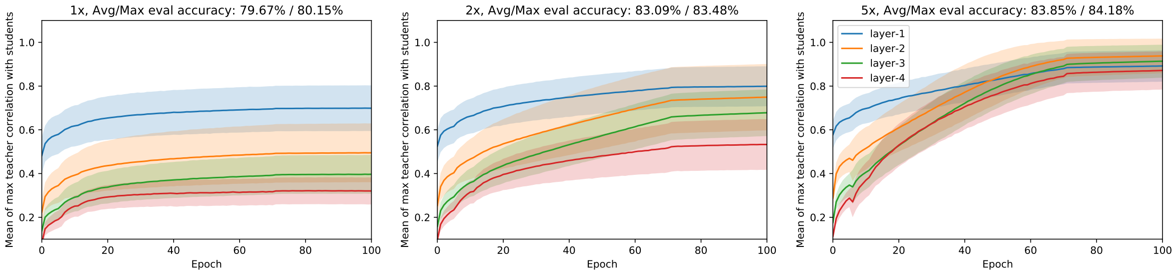


Figure 10. Mean of the max teacher correlation ρ_{mean} with student nodes over epochs in CIFAR10. More over-realization gives better student specialization across all layers and achieves strong generalization (higher evaluation accuracy on CIFAR-10 evaluation set).

Fig. 10 shows the convergence and specialization behaviors of student network. More over-realization leads to stronger specialization and improved generalization evaluated on CIFAR-10 evaluation set.

8. Conclusion and Future Work

In this paper, we use student-teacher setting to analyze how an (over-realized) deep ReLU student network trained with SGD learns from the output of a teacher. When the magnitude of gradient per sample is small, the teacher can be proven to be specialized by (possibly multiple) students and thus the teacher network is recovered at the lowest layer. We also provide finite sample analysis about when it happens. As future works, it is interesting to show specialization at *every* layer in deep networks and understand training dynamics in the teacher-student setting. On the empirical side, student-teacher setting can be useful when analyzing many practical phenomena (e.g., connectivity of low-cost solutions).

References

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *NeurIPS*, 2019a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252, Long Beach, California, USA, 09–15 Jun 2019b. PMLR. URL `http://proceedings.mlr.press/v97/allen-zhu19a.html`.

Arora, S., Cohen, N., Golowich, N., and Hu, W. A convergence analysis of gradient descent for deep linear neural networks. In *ICLR*, 2019. URL `https://openreview.net/forum?id=SkMQg3C5K7`.

Aubin, B., Maillard, A., Krzakala, F., Macris, N., Zdeborová, L., et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. In *Advances in Neural Information Processing Systems*, pp. 3223–3234, 2018.

Bassily, R., Belkin, M., and Ma, S. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

Bertsekas, D. P. and Tsitsiklis, J. N. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Caruana, R., Lawrence, S., and Giles, C. L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems*, pp. 402–408, 2001.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *ICLR*, 2017.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pp. 3036–3046, 2018.

Daneshmand, H., Kohler, J., Lucchi, A., and Hofmann, T. Escaping saddles with stochastic gradients. *arXiv preprint arXiv:1803.05999*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. Essentially no barriers in neural network energy landscape. *arXiv preprint arXiv:1803.00885*, 2018.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient descent can take exponential time to escape saddle points. In *Advances in neural information processing systems*, pp. 1067–1077, 2017.

Du, S. S., Lee, J. D., Tian, Y., Poczos, B., and Singh, A. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *ICML*, 2018.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *ICML*, 2019.

Engel, A. and Van den Broeck, C. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Freeman, J. A. and Saad, D. Online learning in radial basis function networks. *Neural Computation*, 9(7):1601–1622, 1997.

Gardner, E. and Derrida, B. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22 (12):1983–1994, jun 1989. doi: 10.1088/0305-4470/22/12/004. URL `https://doi.org/10.1088%2F0305-4470%2F22%2F12%2F004`.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, pp. 8789–8798, 2018.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Ge, R., Lee, J. D., and Ma, T. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Goldt, S., Advani, M. S., Saxe, A. M., Krzakala, F., and Zdeborová, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *NeurIPS*, 2019.

Hanin, B. and Rolnick, D. Complexity of linear regions in deep networks. *arXiv preprint arXiv:1901.09021*, 2019a.

Hanin, B. and Rolnick, D. Deep relu networks have surprisingly few activation patterns. In *Advances in Neural Information Processing Systems*, pp. 359–368, 2019b.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

Hassibi, B., Stork, D. G., and Wolff, G. J. Optimal brain surgeon and general network pruning. In *IEEE international conference on neural networks*, pp. 293–299. IEEE, 1993.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Hu, H., Peng, R., Tai, Y.-W., and Tang, C.-K. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.

Hu, T., Shang, Z., and Cheng, G. Optimal rate of convergence for deep neural network classifiers under the teacher-student setting. *arXiv preprint arXiv:2001.06892*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1724–1732. JMLR. org, 2017.

Kawaguchi, K. Deep learning without poor local minima. In *Advances in neural information processing systems*, pp. 586–594, 2016.

Kuditipudi, R., Wang, X., Lee, H., Zhang, Y., Li, Z., Hu, W., Arora, S., and Ge, R. Explaining landscape connectivity of low-cost solutions for multilayer nets. *CoRR*, abs/1906.06247, 2019. URL http://arxiv.org/abs/1906.06247.

Lampinen, A. K. and Ganguli, S. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *ICLR*, 2019.

Laurent, T. and Brecht, J. Deep linear networks with arbitrary loss: All local minima are global. In *International Conference on Machine Learning*, pp. 2908–2913, 2018.

Laurent, T. and von Brecht, J. The multilinear structure of relu networks. *arXiv preprint arXiv:1712.10132*, 2017.

LeCun, Y., Denker, J. S., and Solla, S. A. Optimal brain damage. In *Advances in neural information processing systems*, pp. 598–605, 1990.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

Li, Y., Yosinski, J., Clune, J., Lipson, H., and Hopcroft, J. E. Convergent learning: Do different neural networks learn the same representations? In *ICLR*, 2016.

Liu, C. and Belkin, M. Mass: an accelerated stochastic method for over-parametrized learning. *arXiv preprint arXiv:1810.13395*, 2018.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. *arXiv preprint arXiv:1712.06559*, 2017.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, pp. 3, 2013.

Mace, C. and Coolen, A. Statistical mechanical analysis of the dynamics of learning in perceptrons. *Statistics and Computing*, 8(1):55–88, 1998.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

Marceau-Caron, G. and Ollivier, Y. Natural langevin dynamics for neural networks. In *International Conference on Geometric Science of Information*, pp. 451–459. Springer, 2017.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Neyshabur, B., Salakhutdinov, R. R., and Srebro, N. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2015.

Nguyen, Q. and Hein, M. The loss surface of deep and wide neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2603–2612. JMLR. org, 2017.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pp. 6076–6085, 2017.

Rolnick, D. and Kording, K. P. Identifying weights and architectures of unknown relu networks. *arXiv preprint arXiv:1910.00744*, 2019.

Saad, D. and Solla, S. A. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225, 1995.

Saad, D. and Solla, S. A. Dynamics of on-line gradient descent learning for multilayer neural networks. In *Advances in neural information processing systems*, pp. 302–308, 1996.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., and Dahl, G. E. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research (JMLR)*, 2018.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Tian, Y., Jiang, T., Gong, Q., and Morcos, A. Luck matters: Understanding training dynamics of deep relu networks. *arXiv preprint arXiv:1905.13405*, 2019.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Wu, L., Ma, C., and Weinan, E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pp. 8279–8288, 2018.

Yun, C., Sra, S., and Jadbabaie, A. Global optimality conditions for deep neural networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BJk7Gf-CZ`.

Yun, C., Sra, S., and Jadbabaie, A. Small nonlinearities in activation functions create bad local minima in neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rke_YiRct7`.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

Supplementary Materials for Student Specialization in Deep Rectified Networks With Finite Width and Input Dimension

A. More related works

Deep Linear networks. For deep linear networks, multiple works (Lampinen & Ganguli, 2019; Saxe et al., 2013; Arora et al., 2019; Advani & Saxe, 2017) have shown interesting training dynamics. One common assumption is that the singular spaces of weights at nearby layers are aligned at initialization, which decouples the training dynamics. Such a nice property would not hold for nonlinear network. (Lampinen & Ganguli, 2019) shows that under this assumption, weight components with large singular value are learned first, while we analyze and observe empirically similar behaviors on the student node level. Generalization property of linear networks can also be analyzed in the limit of infinite input dimension with teacher-student setting (Lampinen & Ganguli, 2019). However, deep linear networks lack specialization which plays a crucial role in the nonlinear case. To our knowledge, we are the first to analyze specialization rigorously in deep ReLU networks.

SGD versus GD. Stochastic Gradient Descent (SGD) shows strong empirical performance than Gradient Descent (GD) (Shallue et al., 2018) in training deep models. SGD is often treated as an approximate, or a noisy version of GD (Bertsekas & Tsitsiklis, 2000; Hazan & Kale, 2014; Marceau-Caron & Ollivier, 2017; Goldt et al., 2019; Bottou, 2010). In contrast, many empirical evidences show that SGD achieves better generalization than GD when training neural networks, which is explained via implicit regularization (Zhang et al., 2017; Neyshabur et al., 2015), by converging to flat minima (Hochreiter & Schmidhuber, 1997; Chaudhari et al., 2017; Wu et al., 2018) , robust to saddle point (Jin et al., 2017; Daneshmand et al., 2018; Ge et al., 2015; Du et al., 2017) and perform Bayesian inference (Welling & Teh, 2011; Mandt et al., 2017; Chaudhari & Soatto, 2018).

Similar to this work, interpolation setting (Ma et al., 2017; Liu & Belkin, 2018; Bassily et al., 2018) assumes that gradient at each data point vanish at the critical point. While they mainly focus on convergence property of convex objective, we directly relate this condition to specific structure of deep ReLU networks.

B. A Mathematical Framework

B.1. Lemma 1

Proof. We prove by induction. When $l = L$ we know that $\mathbf{g}_L(\mathbf{x}) = \mathbf{f}^*_L(\mathbf{x}) - \mathbf{f}_L(\mathbf{x})$, by setting $V^*_L(\mathbf{x}) = V_L(\mathbf{x}) = I_{C\times C}$ and the fact that $D_L(\mathbf{x}) = I_{C\times C}$ (no ReLU gating in the last layer), the condition holds.

Now suppose for layer l, we have:

$$\begin{aligned} \mathbf{g}_l(\mathbf{x}) &= D_l(\mathbf{x})\left[A_l(\mathbf{x})\mathbf{f}^*_l(\mathbf{x}) - B_l(\mathbf{x})\mathbf{f}_l(\mathbf{x})\right] && (11) \\ &= D_l(\mathbf{x})V_l^\intercal(\mathbf{x})\left[V^*_l(\mathbf{x})\mathbf{f}^*_l(\mathbf{x}) - V_l(\mathbf{x})\mathbf{f}_l(\mathbf{x})\right] && (12) \end{aligned}$$

Using

$$\begin{aligned} \mathbf{f}_l(\mathbf{x}) &= D_l(\mathbf{x})W_l^\intercal \mathbf{f}_{l-1}(\mathbf{x}) && (13) \\ \mathbf{f}^*_l(\mathbf{x}) &= D^*_l(\mathbf{x})W_l^{*\intercal}\mathbf{f}^*_{l-1}(\mathbf{x}) && (14) \\ \mathbf{g}_{l-1}(\mathbf{x}) &= D_{l-1}(\mathbf{x})W_l\mathbf{g}_l(\mathbf{x}) && (15) \end{aligned}$$

we have:

$$\begin{aligned}
\mathbf{g}_{l-1}(\mathbf{x}) &= D_{l-1}(\mathbf{x})W_l\mathbf{g}_l(\mathbf{x}) && (16)\\
&= D_{l-1}(\mathbf{x})\underbrace{W_lD_l(\mathbf{x})V_l^{\intercal}(\mathbf{x})}_{V_{l-1}^{\intercal}(\mathbf{x})}\left[V_l^*(\mathbf{x})\mathbf{f}_l^*(\mathbf{x}) - V_l(\mathbf{x})\mathbf{f}_l(\mathbf{x})\right] && (17)\\
&= D_{l-1}(\mathbf{x})V_{l-1}^{\intercal}(\mathbf{x})\left[\underbrace{V_l^*(\mathbf{x})D_l^*(\mathbf{x})W_l^{*\intercal}}_{V_{l-1}^*(\mathbf{x})}\mathbf{f}_{l-1}^*(\mathbf{x}) - \underbrace{V_l(\mathbf{x})D_l(\mathbf{x})W_l^{\intercal}}_{V_{l-1}(\mathbf{x})}\mathbf{f}_{l-1}(\mathbf{x})\right] && (18)\\
&= D_{l-1}(\mathbf{x})V_{l-1}^{\intercal}(\mathbf{x})\left[V_{l-1}^*(\mathbf{x})\mathbf{f}_{l-1}^*(\mathbf{x}) - V_{l-1}(\mathbf{x})\mathbf{f}_{l-1}(\mathbf{x})\right] && (19)\\
&= D_{l-1}(\mathbf{x})\left[A_{l-1}(\mathbf{x})\mathbf{f}_{l-1}^*(\mathbf{x}) - B_{l-1}(\mathbf{x})\mathbf{f}_{l-1}(\mathbf{x})\right] && (20)
\end{aligned}$$

□

B.2. Lemma 2

Lemma 2. *Denote $\mathcal{D} = \{\mathbf{x}_i\}$ as a dataset of N samples. If Assumption 2 holds, then either $\mathbf{g}_l(\mathbf{x}_i;\hat{\mathcal{W}}) = \mathbf{0}$ or $\mathbf{f}_{l-1}(\mathbf{x}_i;\hat{\mathcal{W}}) = \mathbf{0}$.*

Proof. From Assumption 2, we know that for any batch $\mathcal{B}_j$, Eqn. 1 vanishes:

$$\dot{W}_l = \mathbb{E}_{\mathbf{x}}\left[\mathbf{g}_l(\mathbf{x};\hat{\mathcal{W}})\mathbf{f}_{l-1}^{\intercal}(\mathbf{x};\hat{\mathcal{W}})\right] = \sum_{i\in\mathcal{B}_j}\mathbf{g}_l(\mathbf{x}_i;\hat{\mathcal{W}})\mathbf{f}_{l-1}^{\intercal}(\mathbf{x}_i;\hat{\mathcal{W}}) = 0 \qquad (21)$$

Let $U_i = \mathbf{g}_l(\mathbf{x}_i;\hat{\mathcal{W}})\mathbf{f}_{l-1}^{\intercal}(\mathbf{x}_i;\hat{\mathcal{W}})$. Note that $\mathcal{B}_j$ can be any subset of samples from the data distribution. Therefore, for a dataset of size N, Eqn. 21 holds for all $\binom{N}{|\mathcal{B}|}$ batches, but there are only N data samples. With simple Gaussian elimination we know that for any $i_1 \neq i_2$, $U_{i_1} = U_{i_2} = U$. Plug that into Eqn. 21 we know $U = 0$ and thus for any i, $U_i = 0$. Since U_i is an outer product, the theorem follows.

Note that if $\|\dot{W}_l\|_\infty \leq \epsilon$, which is $\|\sum_{i\in\mathcal{B}_j}U_i\|_\infty \leq \epsilon$, then with simple Gaussian elimination for two batches $\mathcal{B}_1$ and $\mathcal{B}_2$ with only two sample difference, we will have for any $i_1 \neq i_2$, $\|U_{i_1} - U_{i_2}\|_\infty = \|\sum_{i\in\mathcal{B}_1}U_i - \sum_{i\in\mathcal{B}_2}U_i\|_\infty \leq \|\sum_{i\in\mathcal{B}_1}U_i\|_\infty + \|\sum_{i\in\mathcal{B}_2}U_i\|_\infty = 2\epsilon$. Plug things back in and we have $|\mathcal{B}|\|U_i\|_\infty \leq [2(|\mathcal{B}| - 1) + 1]\epsilon$, which is $\|U_i\|_\infty \leq 2\epsilon$. If $\mathbf{f}_{l-1}(\mathbf{x};\hat{\mathcal{W}})$ has the bias term, then immediately we have $\|\mathbf{g}_l(\mathbf{x};\hat{\mathcal{W}})\|_\infty \leq \epsilon$. □

C. Two-layer, Infinite Samples and Zero Gradient

Definition 8 (Regular weight vector)**.** *A weight vector $\mathbf{w} = [\tilde{\mathbf{w}}, b] \in \mathbb{R}^{d+1}$ is called* regular*, if $\|\tilde{\mathbf{w}}\|_2 = 1$.*

C.1. Corollary 2

Corollary 2 (Piecewise constant)**.** *R_0 can be decomposed into a finite (but potentially exponential) set of regions $\mathcal{R}_{l-1} = \{R_{l-1}^1, R_{l-1}^2, \ldots, R_{l-1}^J\}$ plus a zero-measure set, so that $A_l(\mathbf{x})$ and $B_l(\mathbf{x})$ are constant within each region R_{l-1}^j with respect to $\mathbf{x}$.*

Proof. The base case is that $V_L(\mathbf{x}) = V_L^*(\mathbf{x}) = I_{C\times C}$, which is constant (and thus piece-wise constant) over the entire input space. If for layer l, $V_l(\mathbf{x})$ and $V_l^*(\mathbf{x})$ are piece-wise constant, then by Eqn. 3 (rewrite it here):

$$V_{l-1}(\mathbf{x}) = V_l(\mathbf{x})D_l(\mathbf{x})W_l^{\intercal}, \quad V_{l-1}^*(\mathbf{x}) = V_l^*(\mathbf{x})D_l^*(\mathbf{x})W_l^{*\intercal} \qquad (22)$$

since $D_l(\mathbf{x})$ and $D_l^*(\mathbf{x})$ are piece-wise constant and $W_l^{\intercal}$ and $W_l^{*\intercal}$ are constant, we know that for layer $l-1$, $V_{l-1}(\mathbf{x})$ and $V_{l-1}^*(\mathbf{x})$ are piece-wise constant. Therefore, for all $l = 1, \ldots L$, $V_l(\mathbf{x})$ and $V_l^*(\mathbf{x})$ are piece-wise constant.

Therefore, $A_l(\mathbf{x})$ and $B_l(\mathbf{x})$ are piece-wise constant with respect to input $\mathbf{x}$. They separate the region R_0 into constant regions with boundary points in a zero-measured set. □

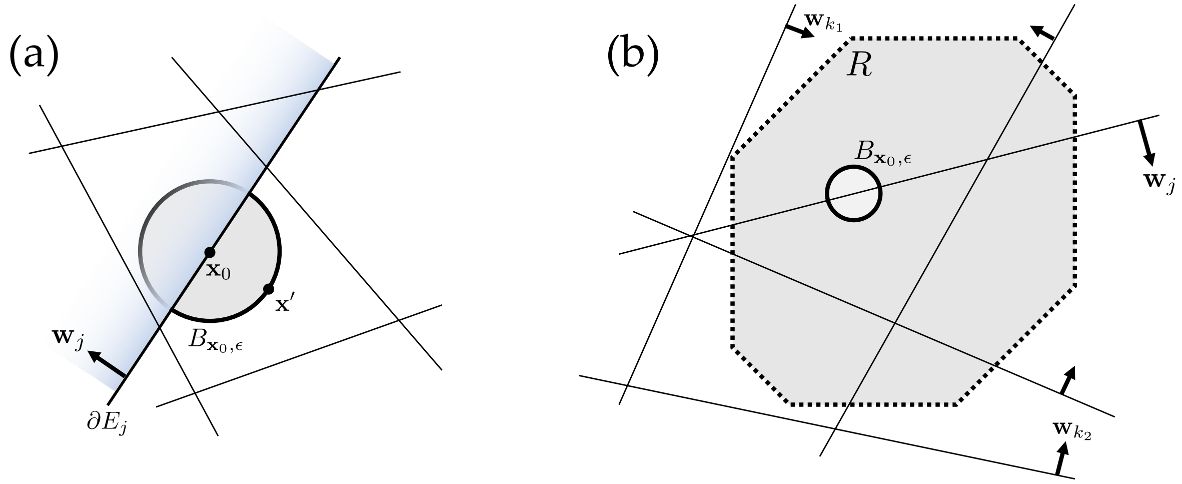


Figure 11. Proof illustration for **(a)** Lemma 3, **(b)** Lemma 4.

C.2. Lemma 3

Lemma 3. *Consider K ReLU activation functions $f_j(\mathbf{x}) = \sigma(\mathbf{w}_j^\intercal \mathbf{x})$ for $j = 1 \dots K$. If $\mathbf{w}_j \neq 0$ and no two weights are co-linear, then $\sum_{j'} c_{j'} f_{j'}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{R}^{d+1}$ suggests that all $c_j = 0$.*

Proof. Suppose there exists some $c_j \neq 0$ so that $\sum_j c_j f_j(\mathbf{x}) = 0$ for all $\mathbf{x}$. Pick a point $\mathbf{x}_0 \in \partial E_j$ so that $\mathbf{w}_j^\intercal \mathbf{x}_0 = 0$ but all $\mathbf{w}_{j'}^\intercal \mathbf{x}_0 \neq 0$ for $j' \neq j$, which is possible due to the distinct weight conditions. Consider an ϵ-ball $B_{\mathbf{x}_0,\epsilon} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| \leq \epsilon\}$. We pick ϵ so that $\mathrm{sign}(\mathbf{w}_{j'}^\intercal \mathbf{x})$ for all $j' \neq j$ remains the same within $B_{\mathbf{x}_0,\epsilon}$ (Fig. 11(a)). Denote $[j^+]$ as the indices of activated ReLU functions in $B_{\mathbf{x}_0,\epsilon}$ except j.

Then for all $\mathbf{x} \in B_{\mathbf{x}_0,\epsilon} \cap E_j$, we have:

$$h(\mathbf{x}) \equiv \sum_{j'} c_{j'} f_{j'}(\mathbf{x}) = c_j \mathbf{w}_j^\intercal \mathbf{x} + \sum_{j' \in [j^+]} c_{j'} \mathbf{w}_{j'}^\intercal \mathbf{x} = 0 \tag{23}$$

Since $B_{\mathbf{x}_0,\epsilon}$ is a d-dimensional object rather than a subspace, for $\mathbf{x}_0$ and $\mathbf{x}_0 + \epsilon \mathbf{e}_k \in B(\mathbf{x}_0, \epsilon)$, we have

$$h(\mathbf{x}_0 + \epsilon \mathbf{e}_k) - h(\mathbf{x}_0) = \epsilon (c_j w_{jk} + \sum_{j' \in [j^+]} c_{j'} w_{j'k}) = 0 \tag{24}$$

where $\mathbf{e}_k$ is axis-aligned unit vector ($1 \leq k \leq d$). This yields

$$c_j \tilde{\mathbf{w}}_j + \sum_{j' \in [j^+]} c_{j'} \tilde{\mathbf{w}}_{j'} = \mathbf{0}_d \tag{25}$$

Plug it back to Eqn. 23 yields

$$c_j b_j + \sum_{j' \in [j^+]} c_{j'} b_{j'} = 0 \tag{26}$$

where means that for the (augmented) $d + 1$ dimensional weight:

$$c_j \mathbf{w}_j + \sum_{j' \in [j^+]} c_{j'} \mathbf{w}_{j'} = \mathbf{0}_{d+1} \tag{27}$$

However, if we pick $\mathbf{x}' = \mathbf{x}_0 - \epsilon \frac{\tilde{\mathbf{w}}_j}{\|\tilde{\mathbf{w}}_j\|^2} \in B_{\mathbf{x}_0,\epsilon} \cap E_j^\complement$, then $f_j(\mathbf{x}') = 0$ but $\sum_{j' \in [j^+]} f_j'(\mathbf{x}') = -c_j \mathbf{w}_j^\intercal \mathbf{x}' = \epsilon c_j$ and thus

$$\sum_{j'} c_{j'} f_{j'}(\mathbf{x}') = \epsilon c_j \neq 0 \tag{28}$$

which is a contradiction. $\square$

C.3. Lemma 4

Lemma 4 (Local ReLU Independence)**.** *Let R be an open set. Consider K ReLU nodes $f_j(\mathbf{x}) = \sigma(\mathbf{w}_j^\intercal \mathbf{x})$, $j = 1, \ldots, K$. $\mathbf{w}_j \neq 0$, $\mathbf{w}_j \neq \gamma \mathbf{w}_{j'}$ for $j \neq j'$ with any $\gamma > 0$.*

If there exists $c_1, \ldots, c_K, c_\bullet$ so that the following is true:

$$\sum_j c_j f_j(\mathbf{x}) + c_\bullet \mathbf{w}_\bullet^\intercal \mathbf{x} = \mathbf{0}, \quad \forall \mathbf{x} \in R \tag{29}$$

and for node j, $\partial E_j \cap R \neq \emptyset$, then $c_j = 0$.

Proof. We can apply the same logic as Lemma 3 to the region R (Fig. 11(b)). For any node j, since its boundary ∂E_j is in R, we can find a similar $\mathbf{x}_0$ so that $\mathbf{x}_0 \in \partial E_j \cap R$ and $\mathbf{x}_0 \notin \partial E_{j'}$ for any $j' \neq j$. We construct $B_{\mathbf{x}_0,\epsilon}$. Since R is an open set, we can always find $\epsilon > 0$ so that $B_{\mathbf{x}_0,\epsilon} \subseteq R$ and no other boundary is in this ϵ-ball. Following similar logic of Lemma 3, $c_j = 0$. □

C.4. Theorem 1

Proof. In this situation, because $D_2(\mathbf{x}) = D_2^*(\mathbf{x}) = I$, according to Eqn. 3, $V_1(\mathbf{x}) = W_1^\intercal$ and $V_1^*(\mathbf{x}) = W_1^{*\intercal}$ are independent of input $\mathbf{x}$. Therefore, both A_1 and B_1 are independent of input $\mathbf{x}$.

From Assumption 1, since $\rho(\mathbf{x}) > 0$ in R_0, from Lemma 2, we know that either $\mathbf{g}_1(\mathbf{x}) = \mathbf{0}$ or $\mathbf{x} = \mathbf{0}$. However, since $\mathbf{x} = [\tilde{\mathbf{x}}, 1]$ has bias term, $\mathbf{g}_1(\mathbf{x}) = D_1(\mathbf{x})\left[A_1 \mathbf{f}_1^*(\mathbf{x}) - B_1 \mathbf{f}_1(\mathbf{x})\right] = \mathbf{0}$. Picking node k, the following holds for every node k and every $\mathbf{x} \in R_0 \cap E_k$:

$$\boldsymbol{\alpha}_k^\intercal \mathbf{f}^*(\mathbf{x}) - \boldsymbol{\beta}_k^\intercal \mathbf{f}(\mathbf{x}) = \mathbf{0} \tag{30}$$

Here $\boldsymbol{\alpha}_k^\intercal$ is the k-th row of A_1, $A_1 = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{n_1}]^\intercal$ and similarly for $\boldsymbol{\beta}_k^\intercal$. Note here layer index $l = 1$ is omitted for brevity.

For teacher j, suppose it is observed by student k, i.e., $\partial E_j^* \cap E_k \neq \emptyset$. Given all teacher and student nodes, note that co-linearity is a equivalent relation, we could partition these nodes into disjoint groups. Suppose node j is in group s. In Eqn. 30, if we combine all coefficients in group s together into one term $c_s \mathbf{w}_j^*$ (with $\|\mathbf{w}_j^*\| = 1$), we have:

$$c_s = \alpha_{kj} - \sum_{k' \in \text{co-linear}(j)} \|\mathbf{w}_{k'}\| \beta_{kk'} \tag{31}$$

"At most" because from Assumption 1, all teacher weights are not co-linear. Note that $\text{co-linear}(j)$ might be an empty set.

By Assumption 1, $\partial E_j^* \cap R_0 \neq \emptyset$ and by observation property, $\partial E_j^* \cap E_k \neq \emptyset$, we know that for $R = R_0 \cap E_k$, $\partial E_j^* \cap R \neq \emptyset$. Applying Lemma 4, we know that $c_s = 0$. Since $\alpha_{kj} \neq 0$, we know $\text{co-linear}(j) \neq \emptyset$ and there exists at least one student k' that is aligned with the teacher j. □

C.5. Theorem 2

Proof. We basically apply the same logic as in Theorem 1. Consider the colinear group $\text{co-linear}(k)$. If for all $k' \in \text{co-linear}(k)$, $\beta_{k'k'} \equiv \|\mathbf{v}_{k'}\|^2 = 0$, then $\mathbf{v}_{k'} = \mathbf{0}$ and the proof is complete.

Otherwise, if there exists some student k so that $\mathbf{v}_k \neq \mathbf{0}$. By the condition, it is observed by some student node k_o, then with the same logic we will have

$$\sum_{k' \in \text{co-linear}(k)} \beta_{k_o,k'} \|\mathbf{w}_{k'}\| = 0 \tag{32}$$

which is

$$\mathbf{v}_{k_o}^\intercal \sum_{k' \in \text{co-linear}(k)} \mathbf{v}_{k'} \|\mathbf{w}_{k'}\| = 0 \tag{33}$$

Since k is observed by C students $k_o^1, k_o^2, \ldots, k_o^J$, then we have:

$$\mathbf{v}_{k_o^j}^\intercal \sum_{k' \in \text{co-linear}(k)} \mathbf{v}_{k'} \|\mathbf{w}_{k'}\| = 0 \tag{34}$$

By the condition, all the C vectors $\mathbf{v}^{\intercal}_{k^j_o} \in \mathbb{R}^C$ are linear independent, then we know that

$$\sum_{k' \in \text{co-linear}(k)} \mathbf{v}_{k'} \|\mathbf{w}_{k'}\| = \mathbf{0} \tag{35}$$

□

C.6. Corollary 1

Proof. We can write the contribution of all student nodes which are not aligned with any teacher nodes as follows:

$$\begin{aligned} \sum_s \sum_{k \in \text{co-linear}(s)} \mathbf{v}_k f_k(\mathbf{x}) &= \sum_s \sum_{k \in \text{co-linear}(s)} \mathbf{v}_k \|\mathbf{w}_k\| \sigma(\mathbf{w}'^{\intercal}_s \mathbf{x}) && (36) \\ &= \sum_s \sigma(\mathbf{w}'^{\intercal}_s \mathbf{x}) \sum_{k \in \text{co-linear}(s)} \mathbf{v}_k \|\mathbf{w}_k\| && (37) \end{aligned}$$

where $\mathbf{w}'_s$ is the unit vector that represents the common direction of the co-linear group s. From Theorem 2, for group s that is not aligned with any teacher, $\sum_{k \in \text{co-linear}(s)} \mathbf{v}_k \|\mathbf{w}_k\| = \mathbf{0}$ and thus the net contribution is zero. □

D. Main Theorems

D.1. Lemma 5

Lemma 5 (Relation between Hyperplanes)**.** *Let $\mathbf{w}_j$ and $\mathbf{w}_{j'}$ two distinct hyperplanes with $\|\tilde{\mathbf{w}}_j\| = \|\tilde{\mathbf{w}}_{j'}\| = 1$. Denote $\tilde{\theta}_{jj'}$ as the angle between the two vectors $\tilde{\mathbf{w}}_j$ and $\tilde{\mathbf{w}}_{j'}$. Then there exists $\tilde{\mathbf{u}}_{j'} \perp \tilde{\mathbf{w}}_j$ and $\mathbf{w}^{\intercal}_{j'} \tilde{\mathbf{u}}_{j'} = \sin \tilde{\theta}_{jj'}$.*

Proof. Note that the projection of $\tilde{\mathbf{w}}_{j'}$ onto $\tilde{\mathbf{w}}_j$ is:

$$\tilde{\mathbf{u}}_{j'} = \frac{1}{\sin \tilde{\theta}_{jj'}} P^{\perp}_{\tilde{\mathbf{w}}_j} \tilde{\mathbf{w}}_{j'} \tag{38}$$

It is easy to verify that $\|\tilde{\mathbf{u}}_{j'}\| = 1$ and $\mathbf{w}^{\intercal}_{j'} \tilde{\mathbf{u}}_{j'} = \sin \tilde{\theta}_{jj'}$. □

Definition 9 (Alignment of (j, j') by error $(M_1\epsilon, M_2\epsilon)$)**.** *Two nodes j and j' are called aligned, if their weights $\mathbf{w}_j = [\tilde{\mathbf{w}}_j, b_j]$ and $\mathbf{w}_{j'} = [\tilde{\mathbf{w}}_{j'}, b_{j'}]$ satisfy the following:*

$$\sin \tilde{\theta}_{jj'} \le M_1 \epsilon, \qquad |b_j - b_{j'}| \le M_2 \epsilon \tag{39}$$

Definition 10 (Constrained η-Dataset)**.** *For a weight vector $\mathbf{w}$ and $\epsilon \le \epsilon_0$, a dataset $D' = D \cap I_{\mathbf{w}}(\epsilon)$ is called a constrained η-Dataset, if for any regular $\mathbf{w}'$ with $\tilde{\mathbf{w}}^T \tilde{\mathbf{w}}' = 0$, we have:*

$$N\left[D' \cap I_{\mathbf{w}'}(\epsilon)\right] \le \eta_{\mathbf{w}} \epsilon N_{D'} + (d+1) \tag{40}$$

Note $\eta_{\mathbf{w}}$ is independent of ϵ.

Note that D' is always a constrained η-dataset for sufficiently large η.

Lemma 6. *Consider K hyper planes, each with regular weight $\{\mathbf{w}_j\}_{j=1}^K$. $\gamma > 0$ and $\epsilon > 0$ are constants. Consider one hyper-plane j. For a constrained η-dataset $D \subseteq I_j(\epsilon)$, if*

(a) $N_D \ge (\gamma + 3)K(d + 1)$.

(b) For $j' \ne j$, j' is not aligned with j by error $(M_1\epsilon, M_2\epsilon)$. Here

$$M_1 = (\gamma + 3)\eta K, \quad M_2 = 1 + \gamma + (\gamma + 3)\eta K \left(|b_j| + \sqrt{\frac{(3+\gamma)\mu K}{1+\gamma}} \right) \tag{41}$$

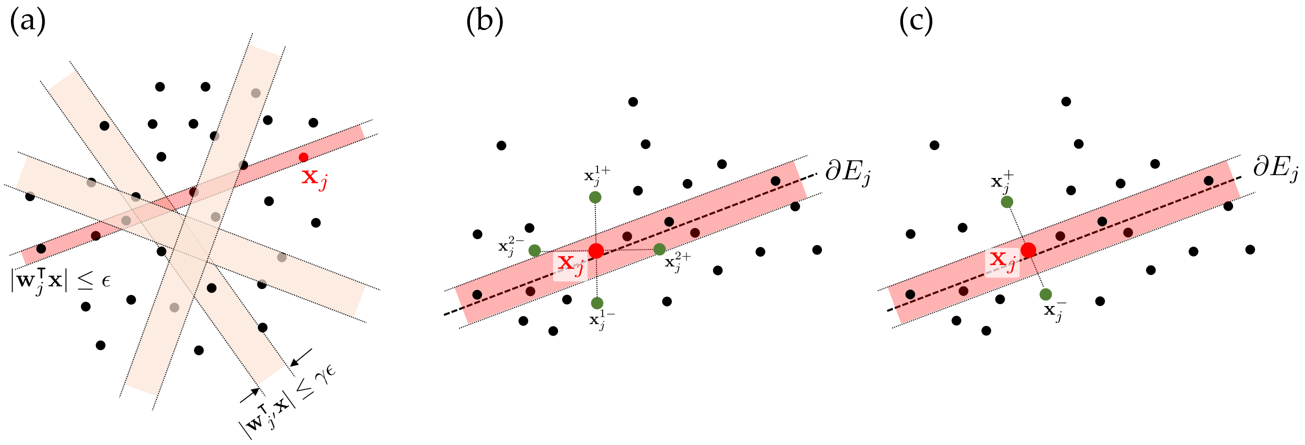


Figure 12. **(a)** Lemma 6: either there exists a node j' aligned with node j, or there exists $\mathbf{x}_j$ so that $|\mathbf{w}_j^\intercal \mathbf{x}| \le \epsilon$ but for all $j' \neq j$, $|\mathbf{w}_{j'}^\intercal \mathbf{x}| > \gamma\epsilon$, i.e., $\mathbf{x}_j \in I_j(\epsilon) \backslash \cup_{j' \neq j} I_{j'}(\gamma\epsilon)$. **(c)** Lemma 7. For such $\mathbf{x}_j$, since each data point is augmented according to teacher-agnostic augmentation (Definition 7), $\mathbf{x}_j$ also has its own augmentation $\mathbf{x}_j^{k\pm}$. Due to the property of $\mathbf{x}_j$, we know $\mathbf{x}_j$ and its augmentation $\mathbf{x}_j^{k\pm}$ are on the same side of gradient function $h(\mathbf{x})$ and thus each node $j' \neq j$ are all linear. By checking the gradient h evaluated on $\mathbf{x}_j$ and its augmentation $\mathbf{x}_j^{k\pm}$, we could show at least one gradient has large magnitude by contradiction. **(c)** In teacher-aware case, contradiction could follow with only 2 augmented samples, if they are constructed alone teacher j's weight $\tilde{\mathbf{w}}_j^*$.

Then there exists $\mathbf{x} \in D$ so that for any $j' \neq j$, $\mathbf{x} \notin I_{j'}(\gamma\epsilon)$.

Proof. Without loss of generality, we assume any angle $\tilde{\theta}_{jj'} \in [0, \pi/2]$. If not, we can always flip the hyper plane by sending $\mathbf{w} = [\tilde{\mathbf{w}}, b]$ to $\mathrm{flip}(\mathbf{w}) = [-\tilde{\mathbf{w}}, b]$. This gives $I_{\mathbf{w}}(\epsilon) = I_{\mathrm{flip}(\mathbf{w})}(\epsilon)$ and keep the definition of ϵ-alignment: $(\mathbf{w}_1, \mathbf{w}_2)$ is ϵ-aligned if and only if $(\mathbf{w}_1, \mathrm{flip}(\mathbf{w}_2))$ is ϵ-aligned, due to the fact that $\sin\theta = \sin(\pi - \theta)$.

For any $j' \neq j$, since j' is not aligned with j by error $(M_1\epsilon, M_2\epsilon)$, we know that either of the two cases hold.

1) $\sin\tilde{\theta}_{jj'} > M_1\epsilon$.

2) $\sin\tilde{\theta}_{jj'} \le M_1\epsilon$ but $|b_j - b_{j'}| > M_2\epsilon$.

Case 1: By Lemma 5 we know that there exists $\tilde{\mathbf{u}}_{j'} \perp \mathbf{w}_j$ so that

$$\tilde{\mathbf{w}}_{j'} = \cos\tilde{\theta}_{jj'}\tilde{\mathbf{w}}_j + \sin\tilde{\theta}_{jj'}\tilde{\mathbf{u}}_{j'} \tag{42}$$

Let $b_u = (b_{j'} - b_j \cos\tilde{\theta}_{jj'})/\sin\tilde{\theta}_{jj'}$ and $\mathbf{u}_{j'} = [\tilde{\mathbf{u}}_{j'}, b_u]$. Then we have:

$$\mathbf{w}_{j'} = \cos\tilde{\theta}_{jj'}\mathbf{w}_j + \sin\tilde{\theta}_{jj'}\mathbf{u}_{j'} \tag{43}$$

Notice that we have the following fact: if $\mathbf{x} \in D \cap I_{j'}(\gamma\epsilon) \subseteq I_j(\epsilon) \cap I_{j'}(\gamma\epsilon)$, then

$$|\mathbf{u}_{j'}^\intercal \mathbf{x}| \le \frac{1}{\sin\tilde{\theta}_{jj'}}\left[|\mathbf{w}_{j'}^\intercal \mathbf{x}| + |\cos\tilde{\theta}_{jj'}||\mathbf{w}_j^\intercal \mathbf{x}|\right] \le \frac{(1+\gamma)\epsilon}{M_1\epsilon} = \frac{1+\gamma}{(3+\gamma)\eta K} \tag{44}$$

Therefore, by the definition of η-dataset, we have:

$$N_D\left[I_{j'}(\gamma\epsilon)\right] \le N_D\left[|\mathbf{u}_{j'}^\intercal \mathbf{x}| \le \frac{1+\gamma}{(3+\gamma)\eta K}\right] \le \frac{1+\gamma}{(3+\gamma)K}N_D + (d+1) \tag{45}$$

Case 2: Notice the following fact: if $\mathbf{x} \in D \cap I_{j'}(\gamma\epsilon) \subseteq I_j(\epsilon) \cap I_{j'}(\gamma\epsilon)$, then from Eqn. 43 we know:

$$\tilde{\mathbf{w}}_{j'}^\intercal \tilde{\mathbf{x}} = \cos\tilde{\theta}_{jj'}\tilde{\mathbf{w}}_j^\intercal \tilde{\mathbf{x}} + \sin\tilde{\theta}_{jj'}\tilde{\mathbf{u}}_{j'}^\intercal \tilde{\mathbf{x}} \tag{46}$$

which means that

$$\mathbf{w}_{j'}^\intercal \mathbf{x} - b_{j'} = \cos\tilde{\theta}_{jj'}(\mathbf{w}_j^\intercal \mathbf{x} - b_j) + \sin\tilde{\theta}_{jj'}\tilde{\mathbf{u}}_{j'}^\intercal \tilde{\mathbf{x}} \tag{47}$$

Therefore, we have:

$$\begin{aligned} |\tilde{\mathbf{u}}_{j'}^\intercal \tilde{\mathbf{x}}| &= \frac{1}{\sin\tilde{\theta}_{jj'}}\left|\mathbf{w}_{j'}^\intercal \mathbf{x} - b_{j'} - \cos\tilde{\theta}_{jj'}(\mathbf{w}_j^\intercal \mathbf{x} - b_j)\right| \qquad (48) \\ &\geq \frac{1}{\sin\tilde{\theta}_{jj'}}\left[|b_j - b_{j'}| - (1-\cos\tilde{\theta}_{jj'})|b_j| - |\mathbf{w}_{j'}^\intercal \mathbf{x}| - |\mathbf{w}_j^\intercal \mathbf{x}|\right] \qquad (49) \end{aligned}$$

Note that since $\tilde{\theta}_{jj'} \in [0, \pi/2]$, we have $1 - \cos\tilde{\theta}_{jj'} \leq 1 - \cos^2\tilde{\theta}_{jj'} = \sin^2\tilde{\theta}_{jj'} \leq \sin\tilde{\theta}_{jj'} \leq M_1\epsilon$. Therefore,

$$|\tilde{\mathbf{u}}_{j'}^\intercal \tilde{\mathbf{x}}| \geq \frac{M_2 - 1 - \gamma}{M_1} - |b_j| = \sqrt{\frac{(3+\gamma)\mu K}{1+\gamma}} \tag{50}$$

Therefore, we have:

$$N_D[I_{j'}(\gamma\epsilon)] \leq N_D\left[|\tilde{\mathbf{u}}_{j'}^\intercal \tilde{\mathbf{x}}| \geq \sqrt{\frac{(3+\gamma)\mu K}{1+\gamma}}\right] \leq \frac{1+\gamma}{(3+\gamma)K}N_{D_j} \tag{51}$$

Combining the two cases, since $N_{D_j} \geq (3+\gamma)K(d+1)$, we know that

$$\begin{aligned} \sum_{j' \neq j} N_D\left[I_{j'}(\gamma\epsilon)\right] &\leq \frac{1+\gamma}{3+\gamma}\frac{K-1}{K}N_D + (K-1)(d+1) \qquad (52) \\ &< \frac{1+\gamma}{3+\gamma}N_{D_j} + \frac{1}{3+\gamma}N_{D_j} \qquad (53) \\ &= \left(1 - \frac{1}{3+\gamma}\right)N_D < N_D \qquad (54) \end{aligned}$$

Moreover, since $N_D/(3+\gamma) \geq K(d+1) \geq 1$, so there exists at least one $\mathbf{x} \in D$ so that $\mathbf{x}$ doesn't fall into any bucket $I_{j'}(\gamma\epsilon)$. This means that for any $j' \neq j$, $\mathbf{x} \notin I_{j'}(\gamma\epsilon)$ and the proof is complete. □

Remark. Note that the constant $|b_j|$ in M_2 can be smaller since we could always use a stronger bound $1 - \cos\theta \leq 1 - \cos^2\theta = \sin^2\theta$ and as a result, M_2 would contain $|b_j|\epsilon$. For small ϵ, this term is negligible.

Lemma 7. *Define*

$$h(\mathbf{x}) = \sum_{j'=1}^{K} c_{j'}\sigma(\mathbf{w}_{j'}^\intercal \mathbf{x}) + c_{\bullet}\mathbf{w}_{\bullet}^\intercal \mathbf{x} \tag{55}$$

Suppose there exists $\mathbf{x}_j \in I_j(\epsilon_l)\backslash \cup_{j'\neq j} I_{j'}(\epsilon_h)$ *with* $\epsilon_l < \epsilon_h$ *and there exists a vector* $\tilde{\mathbf{e}}$ *so that* $\epsilon_l < \epsilon_0 \leq \mathbf{w}_j^\intercal \tilde{\mathbf{e}} \leq \epsilon_h$. *Construct* 2 *datapoints* $\mathbf{x}_j^{\pm} = \mathbf{x}_j \pm \tilde{\mathbf{e}}$ *and set* $D = \{\mathbf{x}_j, \mathbf{x}_j^{\pm}\}$. *Then there exists* $\mathbf{x} \in D$ *so that* $|h(\mathbf{x})| > \frac{|c_j|}{5}(\epsilon_0 - \epsilon_l)$.

Proof. We show that the three points $\mathbf{x}_j$ and $\mathbf{x}_j^{\pm}$ are on the same side of $\partial E_{j'}$ for any $j' \neq j$. This can be achieved by checking whether $(\mathbf{w}_{j'}^\intercal \mathbf{x}_j)(\mathbf{w}_{j'}^\intercal \mathbf{x}_j^{\pm}) \geq 0$ (Fig. 12(b) and (c)):

$$\begin{aligned} (\mathbf{w}_{j'}^\intercal \mathbf{x}_j)(\mathbf{w}_{j'}^\intercal \mathbf{x}_j^{\pm}) &= (\mathbf{w}_{j'}^\intercal \mathbf{x}_j)\left[\mathbf{w}_{j'}^\intercal(\mathbf{x}_j \pm \tilde{\mathbf{e}})\right] \qquad (56) \\ &= (\mathbf{w}_{j'}^\intercal \mathbf{x}_j)^2 \pm (\mathbf{w}_{j'}^\intercal \mathbf{x}_j)\mathbf{w}_{j'}^\intercal \tilde{\mathbf{e}} \qquad (57) \\ &= |\mathbf{w}_{j'}^\intercal \mathbf{x}_j|(|\mathbf{w}_{j'}^\intercal \mathbf{x}_j| \pm \mathbf{w}_{j'}^\intercal \tilde{\mathbf{e}}) \qquad (58) \end{aligned}$$

Since $\mathbf{x}_j \notin I_{j'}(\epsilon_h)$ and $\mathbf{w}_{j'}^\intercal \tilde{\mathbf{e}} \leq \epsilon_h$, we have:

$$|\mathbf{w}_{j'}^\intercal \mathbf{x}_j| \pm \mathbf{w}_{j'}^\intercal \tilde{\mathbf{e}} > \epsilon_h \pm \epsilon_h \geq 0 \tag{59}$$

Therefore, $(\mathbf{w}_{j'}^\intercal \mathbf{x}_j)(\mathbf{w}_{j'}^\intercal \mathbf{x}_j^\pm) \geq 0$ and the three points $\mathbf{x}_j$ and $\mathbf{x}_j^\pm$ are on the same side of $\partial E_{j'}$ for any $j' \neq j$.

If the conclusion is not true, then consider $h(\mathbf{x}_j^+) + h(\mathbf{x}_j^-) - 2h(\mathbf{x}_j)$. Since $\mathbf{x}_j^+ + \mathbf{x}_j^- = 2\mathbf{x}_j$, we know that all terms related to $\mathbf{w}_\bullet$ and $\mathbf{w}_{j'}$ with $j \neq j$ will cancel out, due to the fact that they are in the same side of the boundary $\partial E_{j'}$ and thus behave linearly. Therefore,

$$h(\mathbf{x}_j^+) + h(\mathbf{x}_j^-) - 2h(\mathbf{x}_j) = c_j \left[\sigma(\mathbf{w}_j^\intercal \mathbf{x}_j^+) + \sigma(\mathbf{w}_j^\intercal \mathbf{x}_j^-) - 2\sigma(\mathbf{w}_j^\intercal \mathbf{x}_j)\right] \tag{60}$$

Since $|\mathbf{w}_j^\intercal \mathbf{x}_j| \leq \epsilon_l$ and $\mathbf{w}_j^\intercal \tilde{\mathbf{e}} \geq \epsilon_0 > \epsilon_l$, it is always the case that:

(a) $\mathbf{w}_j^\intercal \mathbf{x}_j^+ = \mathbf{w}_j^\intercal \mathbf{x}_j + \mathbf{w}_j^\intercal \tilde{\mathbf{e}} > 0$ and by ReLU properties $\sigma(\mathbf{w}_j^\intercal \mathbf{x}_j^+) = \mathbf{w}_j^\intercal \mathbf{x}_j^+$.

(b) $\mathbf{w}_j^\intercal \mathbf{x}_j^- = \mathbf{w}_j^\intercal \mathbf{x}_j - \mathbf{w}_j^\intercal \tilde{\mathbf{e}} < 0$ so by ReLU properties $\sigma(\mathbf{w}_j^\intercal \mathbf{x}_j^-) = 0$.

Therefore if $\mathbf{w}_j^\intercal \mathbf{x}_j \geq 0$ then:

$$\begin{aligned}
|h(\mathbf{x}_j^+) + h(\mathbf{x}_j^-) - 2h(\mathbf{x}_j)| &= |c_j \mathbf{w}_j^\intercal (\mathbf{x}_j^+ - 2\mathbf{x}_j)| = |c_j \mathbf{w}_j^\intercal (\mathbf{x}_j + \tilde{\mathbf{e}} - 2\mathbf{x}_j)| && (61)\\
&= |c_j \mathbf{w}_j^\intercal (\tilde{\mathbf{e}} - \mathbf{x}_j)| \geq |c_j(\epsilon_0 - \mathbf{w}_j^\intercal \mathbf{x}_j)| && (62)\\
&\geq |c_j|(\epsilon_0 - \epsilon_l) && (63)
\end{aligned}$$

if $\mathbf{w}_j^\intercal \mathbf{x}_j < 0$ then:

$$|h(\mathbf{x}_j^+) + h(\mathbf{x}_j^-) - 2h(\mathbf{x}_j)| = |c_j \mathbf{w}_j^\intercal \mathbf{x}_j^+| = |c_j \mathbf{w}_j^\intercal (\mathbf{x}_j + \tilde{\mathbf{e}})| \geq |c_j|(\epsilon_0 - \epsilon_l) \tag{64}$$

On the other hand, from gradient condition, we have:

$$|h(\mathbf{x}_j^+) + h(\mathbf{x}_j^-) - 2h(\mathbf{x}_j)| \leq \frac{4}{5}|c_j|(\epsilon_0 - \epsilon_l) \tag{65}$$

which is a contradiction. □

For Leaky ReLU, the proof is similar except that the final condition has an additional $1 - c_{\mathrm{leaky}}$ factor.

Assumption 4. *(a) Two teacher nodes $j \neq j'$ are not ϵ_0-aligned. (b) The boundary band $I_j(\epsilon)$ of each teacher j overlaps with the dataset:*

$$N_D[I_j(\epsilon)] \geq \tau\epsilon N_D \tag{66}$$

Theorem 6 (Two-layer Specialization with Polynomial Samples)**.** *Let $K = m_1 + n_1$. For $0 < \epsilon \leq \epsilon_0$, for any finite dataset D with $N = \Theta(cK^{5/2}d^2\tau^{-1}\epsilon^{-1}\kappa^{-1})$, for any teacher satisfying Assumption 3 and student trained on $D' = \mathrm{aug}(D)$ whose weight $\hat{\mathcal{W}}$ satisfies:*

(1) For $\epsilon \in [0, \epsilon_0]$, the hyperplane band $I_j(\epsilon)$ of a teacher is observed by a student node k: $N_D[I_j(\epsilon) \cap E_k] \geq \kappa N_D[I_j(\epsilon)]$;

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

where $c > 0$ is a constant related to dataset properties. Then there exists a student k' so that (j, k') is ϵ-aligned.

Proof. Let $K = m_1 + n_1$. Let $\epsilon' = \epsilon / cK^{3/2}\sqrt{d}$ with some constant $c = \Theta(\max_j \mu_{\mathbf{w}_j^*} \max_j \eta_{\mathbf{w}_j^*}) > 0$.

We construct a basic dataset D with $N = \Theta(cK^{5/2}d^2\epsilon^{-1}\tau^{-1}\kappa^{-1})$ samples and use the augmentation operator $\mathrm{aug}(D)$:

$$\mathrm{aug}(D) = \{\mathbf{x}_k^\pm = \mathbf{x} \pm 2\epsilon\tilde{\mathbf{e}}_u / cK^{3/2},\ \ \mathbf{x} \in D,\ \ u = 1, \ldots, d\} \cup D \tag{67}$$

where $\tilde{\mathbf{e}}_k$ is axis-aligned unit directions with $\|\tilde{\mathbf{e}}_k\| = 1$. It is clear that $|\mathrm{aug}(D)| = (2d+1)|D|$. Let $D' = \mathrm{aug}(D)$.

Let j be some teacher node. Consider a slice of basic dataset $D \cap I_j(\epsilon')$. By Assumption 4(b), we know that

$$N_D[I_j(\epsilon')] = N[D \cap I_j(\epsilon')] \geq \tau\epsilon' N_D = \mathcal{O}(Kd^{3/2}\kappa^{-1}) \tag{68}$$

and thus for $N_D[I_j(\epsilon') \cap E_k]$, we know that

$$N_D[I_j(\epsilon') \cap E_k] \geq \kappa N_D[I_j(\epsilon')] = \mathcal{O}(Kd^{3/2}) \tag{69}$$

With a sufficiently large constant in N, we have $N_D[I_j(\epsilon') \cap E_k] \geq (\gamma + 3)K(d+1)$ with $\gamma = 2\sqrt{d}$. We then apply Lemma 6, which leads the two following cases:

Case 1. There exists weight k' so that the following alignment condition holds:

$$\sin \tilde{\theta}_{jk'} \leq M'_{1j}\epsilon' = M_{1j}\epsilon, \quad |b^*_j - b_{k'}| \leq M'_{2j}\epsilon' = M_{2j}\epsilon \tag{70}$$

Where $M'_{1j} = \Theta(K\gamma)$ and $M'_{2j} = \Theta(K^{3/2}\gamma)$. Therefore, $M_{1j} = \Theta(K\gamma/cK^{3/2}\sqrt{d}) = o(1)$ and $M_{2j} = \Theta(1)$. Choosing the constant $c > 0$ so that we have $M_{1j} \leq 1$ and $M_{2j} \leq 1$ and thus

$$\sin \tilde{\theta}_{jk'} \leq \epsilon, \quad |b^*_j - b_{k'}| \leq \epsilon, \tag{71}$$

which means that (j, k') are ϵ-aligned. Note that no other teacher is ϵ-aligned with j. So k' has to be a student node and the proof is complete.

Case 2. If the alignment condition doesn't hold, then according to Lemma 6, there exists $\mathbf{x}_j$ so that (note that $\gamma = 2\sqrt{d}$):

$$\mathbf{x}_j \in I_j(\epsilon') \backslash \cup_{j' \neq j} I_{j'}(2\sqrt{d}\epsilon'). \tag{72}$$

Here all j' includes all teacher and student nodes (a total of K nodes), excluding the current node j under consideration. Since $\{\tilde{\mathbf{e}}_\mathbf{u}\}_{u=1}^d$ forms orthonormal bases, there exists at least one u so that $1 \geq \mathbf{w}_j^{*\intercal}\tilde{\mathbf{e}}_u \geq 1/\sqrt{d}$ (with proper sign flipping of $\tilde{\mathbf{e}}_u$). Let $\tilde{\mathbf{e}} = 2\epsilon\tilde{\mathbf{e}}_u/cK^{3/2} = 2\epsilon'\sqrt{d}\tilde{\mathbf{e}}_u$, we have $2\epsilon' \leq \mathbf{w}_j^{*\intercal}\tilde{\mathbf{e}} \leq 2\sqrt{d}\epsilon'$. Applying Lemma 7 and we know the additional samples required in the lemma is already in $\mathrm{aug}(D)$. Therefore, that there exists $\mathbf{x} \in \mathrm{aug}(D_j) \subset D$ so that $|h(\mathbf{x})| > \frac{|\alpha_{kj}|}{5}(2\epsilon' - \epsilon') = \frac{|\alpha_{kj}|}{5cK^{3/2}\sqrt{d}}\epsilon$, which is a contradiction. □

Theorem 7 (Two-layer Specialization with Teacher-aware Dataset with Polynomial Samples)**.** *For $0 < \epsilon \leq \epsilon_0$, for any finite dataset D with $N = \Theta(cK^{5/2}d\tau^{-1}\epsilon^{-1})$, given a teacher network $\mathcal{W}^*$ satisfying Assumption 3 and student trained on $D' = \mathrm{aug}(D, \mathcal{W}^*)$ whose weight $\hat{\mathcal{W}}$ satisfies*

(1) For $\epsilon \in [0, \epsilon_0]$, the hyperplane band $I_j(\epsilon)$ of a teacher is observed by a student node k: $N_D[I_j(\epsilon) \cap E_k] \geq \kappa N_D[I_j(\epsilon)]$;

(2) Small gradient: $\|\mathbf{g}_1(\mathbf{x}, \hat{\mathcal{W}})\|_\infty \leq \frac{\alpha_{kj}}{5cK^{3/2}}\epsilon$, for $\mathbf{x} \in D'$,

then there exists a student k' so that (j, k') is ϵ-aligned.

Proof. Pick $\epsilon' = \epsilon/cK^{3/2}$ (c defined as in Theorem 6). For dataset D, by Assumption 3, we know that for $D \cap I_j(\epsilon') \cap E_k$:

$$N[D \cap I_j(\epsilon') \cap E_k] = N_D[I_j(\epsilon') \cap E_k] \geq \kappa N_D[I_j(\epsilon')] \geq \kappa\tau\epsilon' N_D = \mathcal{O}(Kd) \tag{73}$$

Apply Lemma 6 with $\gamma = 2$ and similarly we know that either there exists a student k' so that (j, k') is ϵ-align (with $M_{1j} \leq 1$ and $M_{2j} \leq 1$), or there exists $\mathbf{x}_j$ such that

$$\mathbf{x}_j \in I_j(\epsilon') \backslash \cup_{j' \neq j} I_{j'}(2\epsilon'). \tag{74}$$

Setting $\tilde{\mathbf{e}} = 2\epsilon'\mathbf{w}^*_j$ and we know that $\mathbf{w}_j^{*\intercal}\tilde{\mathbf{e}} = 2\epsilon'$. Since we have used teacher-aware augmentation, applying Lemma 7 with $\epsilon_h = \epsilon_0 = 2\epsilon'$ and $\epsilon_l = \epsilon'$, the conclusion follows. □

D.2. Theorem 5

Proof. The proof is similar to 2-layer case (Theorem 3 in the main text or Theorem 6 in Appendix). The only difference is that instead of thinking about $K_1 = m_1 + n_1$ boundaries, we need to think about all the Q boundaries introduced by the top-level and obtain a data point $\mathbf{x}_j$ so that it is within the boundary of node j, but far away from all other possible boundaries:

$$\mathbf{x}_j \in I_j(\epsilon') \backslash \cup_{j' \neq j} I_{j'}(2\sqrt{d}\epsilon'). \tag{75}$$

Where j' includes all the boundaries induced. This could be exponential. Note that for each intermediate node j', its boundary $\mathbf{w}_{j'}\mathbf{f} = 0$ will be "bent" whenever the underlying feature $\mathbf{f}$, which is the output of a set of ReLU / Leaky ReLU nodes, has shifted their activation patterns. □

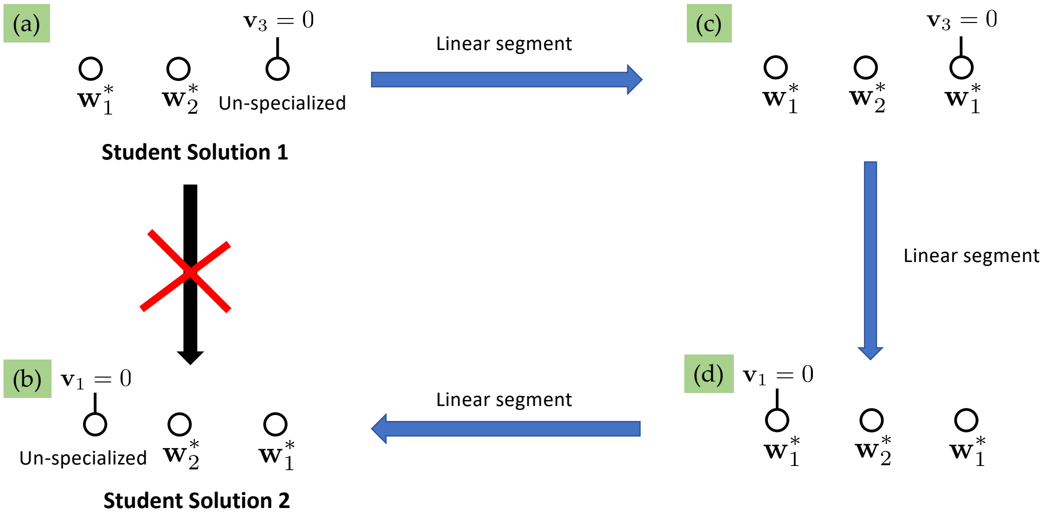


Figure 13. A piece-wise linear curve between two lost-cost student solutions.

E. Connectivity

We construct a piece-wise linear curve from two low-cost student solutions as in Fig. 13. Consider two student networks with 3 hidden nodes trained with the same teacher with 2 nodes. Once they converge, assuming their gradients are zero, there could be many different ways of specializations that satisfy Theorem 1.

Fig. 13(a) and (b) show two such specializations $\mathcal{W}^{(1)}$ and $\mathcal{W}^{(2)}$. For $\mathcal{W}^{(1)}$, $\mathbf{w}_1^{(1)} = \mathbf{w}_1^*$, $\mathbf{w}_2^{(1)} = \mathbf{w}_2^*$ and $\mathbf{w}_3^{(1)}$ is an un-specialized node whose fan-out weights are zero ($\mathbf{v}_3^{(1)} = \mathbf{0}$). For $\mathcal{W}^{(2)}$, $\mathbf{w}_3^{(2)} = \mathbf{w}_1^*$, $\mathbf{w}_2^{(2)} = \mathbf{w}_2^*$ and $\mathbf{w}_1^{(2)}$ is an un-specialized node whose fan-out weights are zero ($\mathbf{v}_1^{(2)} = \mathbf{0}$).

If we directly connect these two solutions using a straight line, the intermediate solution will be no longer low-cost since a linear combination $\lambda\mathbf{w}_1^{(1)} + (1-\lambda)\mathbf{w}_1^{(2)} = \lambda\mathbf{w}_1^* + (1-\lambda)\mathbf{w}_1^{(2)}$ can be a random (un-specialized) vector, and its corresponding fan-out weights $\lambda\mathbf{v}_1^{(1)} + (1-\lambda)\mathbf{v}_1^{(2)} = \lambda\mathbf{v}_1^{(1)}$ is also non-zero. This yields a high-cost solution.

On the other hand, if we take a piece-wise linear path (a)-(c)-(d)-(b), then each line segment will have low-cost and we move $\mathbf{w}_1^*$ from node 1 to node 3. We list the line segment construction as follows:

- Start from $\mathcal{W}^{(1)}$.
- **(a)-(c)**. Blend $\mathbf{w}_1^*$ into an un-specialized weight: $\mathbf{w}_3(t) = (1-t)\mathbf{w}_3^{(1)} + t\mathbf{w}_1^*$. This won't change the output since $\mathbf{v}_3^{(1)} = \mathbf{0}$.
- **(c)-(d)**. Move $\mathbf{v}_1^{(1)}$ from node 1 to node 3:

$$\mathbf{v}_3(t) = (1-t)\mathbf{v}_3^{(1)} + t\mathbf{v}_1^{(1)} = t\mathbf{v}_1^{(1)} \quad (76)$$
$$\mathbf{v}_1(t) = (1-t)\mathbf{v}_1^{(1)} + t\mathbf{v}_3^{(1)} = (1-t)\mathbf{v}_1^{(1)} \quad (77)$$

  This won't change the output since $\mathbf{v}_1(t) + \mathbf{v}_3(t) = \mathbf{v}_1^{(1)}$ and their weights are both $\mathbf{w}_1^*$.
- **(d)-(b)**. Change $\mathbf{w}_1$ to be the unspecified weight in $\mathcal{W}^{(2)}$: $\mathbf{w}_1(t) = (1-t)\mathbf{w}_1^* + t\mathbf{w}_2^{(1)}$. This won't change the output since now $\mathbf{v}_1 = \mathbf{0}$.
- Arrive at $\mathcal{W}^{(2)}$.

F. Empirical results

We construct teacher networks in the following manner. For two-layered network, the output dimension $C = 50$ and input dimension $d = m_0 = n_0 = 100$. For multi-layered network, we use 50-75-100-125 (i.e, $m_1 = 50, m_2 = 75, m_3 =$

$100, m_4 = 125$, $L = 5$, $d = m_0 = n_0 = 100$ and $C = m_5 = n_5 = 50$). The teacher network is constructed to satisfy Assumption 3: at each layer, teacher filters are distinct from each other and their bias is set so that $\sim 50\%$ of the input data activate the nodes, maximizing the number of samples near the boundary.

We generate the input distribution using $\mathcal{N}(0, \sigma^2 I)$ with $\sigma = 10$. For all 2-layer experiments, we sample 10000 as training and another 10000 as evaluation. We also tried other distribution (e.g., uniform distribution $U[-1, 1]$), and the results are similar.

G. Code Release

We have attached all Python codes in our supplementary material submission.