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 $O(K^{3/2}d^{3}\epsilon^{-1})$, until the gradient magnitude drops to $O(\epsilon/K^{3/2}\sqrt{d})$. Here $d$ is the input dimension, $K = m_1 + n_1$ is the total number of neurons in the lowest layer of teacher and student. Note that we require a specific form of data augmentation and the sample complexity includes the additional data generated from augmentation. 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. The code is released in https://github.com/facebookresearch/luckmatters.

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 expressibility (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...
the number $m_t$ of teacher ($n_l \geq m_t$). 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 cor-
lates 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 be-
tween 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 thermodynam-
ics 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 Ker-
nel [Jacot et al., 2018] [Du et al., 2019] and mean-field ap-
proaches [Mei et al., 2018] requires infinite (or very large)
width, our setting applies to finite width as long as stu-
ent is slightly over-realized ($n_l \geq m_t$). (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 his-
tory [Engel & Van den Broeck, 2001; Gardner & Derrida,
1989; Saad & Solla, 1996; 1995; Freeman & Saad, 1997;
Mace & Coolen, 1998] and recently gains increasing inter-
est [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 mechan-
ics point of view in which the input dimension goes to in-
finiteness, or-called thermodynamics limits. They study sym-
metric solutions and locally analyze the symmetric break-
ning behavior and onset of specialization of the student

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 in-
duction 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, spuri-
ous 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 re-
covery properties in the student-teacher setting.

Learning very wide networks. Recent works on Neu-
ral Tangent Kernel [Jacot et al., 2018; Du et al., 2019;
Allen-Zhu et al., 2019a] 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 gen-
eralization, when the input data are assumed to be clus-
tered into classes. Both lines of work assume heavily
over-parameterized network, requiring polynomial growth
of number of nodes with respect to the number of sam-
ple. [Chizat & Bach, 2018] shows global convergence of
over-parameterized network with optimal transport. [Tian
et al., 2019] assumes mild over-realization and gives con-
vergence 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.
3. Mathematical Framework

**Notation.** Consider a student network and its associated teacher network (Fig. 1(a)). Denote the input as $x$. We focus on multi-layered networks with $\sigma(\cdot)$ as ReLU / Leaky ReLU [Maas et al. 2013] non-linearity. 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 $c_{\text{leaky}}$ is the leaky ReLU constant ($c_{\text{leaky}} = 0$ for ReLU). For node $j, f_j(x), z_j(x)$ and $g_j(x)$ 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 $n_l$ 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. $W_l = [w_{l,1}, w_{l,2}, \ldots, w_{l,n_l}]$ where each $w = [\tilde{w}, b] \in \mathbb{R}^{n_{l-1}+1}$ is the weight vector. Here $\tilde{w}$ is the weight and $b$ is the bias.

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

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

Let $f_l(x) = [f_{l,1}(x), \ldots, f_{l,n_l}(x)]^\top \in \mathbb{R}^{n_l+1}$ be the activation vector of layer $l$, $D_l(x) = \text{diag}(z_{l,1}(x), \ldots, z_{l,n_l}(x), 1) \in \mathbb{R}^{(n_l+1) \times (n_l+1)}$ be the diagonal matrix of gating function (for ReLU it is either 0 or 1), and $g_l(x) = [g_{l,1}(x), \ldots, g_{l,n_l}(x), 1]^\top \in \mathbb{R}^{n_l+1}$ be the backpropagated gradient vector. The last 1s are for bias. By definition, $f_0(x) = x \in \mathbb{R}^{n_0}$ is the input and $n_0 = n_0$. Note that $f_l(x), g_l(x)$ and $D_l(x)$ are all dependent on $W$. For brevity, we often use $f_l(x)$ rather than $f_l(x; W)$.

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

$$\hat{W}_l = \mathbb{E}_x [f_{l-1}(x)g_l^T(x)]$$ (1)

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

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

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

$$\min_W J(W) = \frac{1}{2} \mathbb{E}_x \|f^*_l(x) - f_L(x)\|^2$$ (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 $V_L = V_L^*: = I_{C \times C}$):

$$V_{l-1} := V_l D_l W_l^\top, \; \; V_{l-1}^* := V_l^* D_l^* W_l^{*\top}$$ (3)

We further define $A_l := V_l^T V_l^*$ in $\mathbb{R}^{n_l \times n_l}$ and $B_l = V_l^T V_l^*$ in $\mathbb{R}^{n_l \times m_l}$. Each element of $A_l$ is $a_{jj'} = v_j^T v_{j'}^*$. Similarly, each element of $B_l$ is $b_{jj'} = \tilde{v}_j^T \tilde{v}_{j'}$.

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

$$\sin \tilde{\theta}_{jj'} \leq \epsilon, \; \; \|b_j - b_{j'}\| \leq \epsilon,$$ (4)

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

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 $g_l(x)$ satisfies

$$g_l(x) = D_l(x) [A_l(x)f_l^*(x) - B_l(x)f_l(x)]$$ (5)
Remark. Lemma[1] applies to arbitrarily deep ReLU networks and allows $n_1 \neq n_l$. In particular, student can be over-realized. Note that $A_1, B_1, V_1, V_1^\top$ all depends on $x$. Due to the property of (Leaky) ReLU, $A_1(x)$ and $B_1(x)$ are piecewise constant functions (Corollary[2] in Appendix).

Relationship to Network 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 network, i.e., the oracle. 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, $A_1(x)$ and $B_1(x)$ are constant with respect to $x$, 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. Therefor, 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 $B \subseteq R_0$, $W_i = E_{x \in B} [f_{l-1}(x)g_{l}(x)] = 0$.

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 $\sigma(w^j_1x)$ 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 $\alpha_k^1f_k(x) - \beta_k^1f_k(x) = 0$, a linear combination of teacher and student nodes (note that $\alpha_k^1$ and $\beta_k^1$ 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 $v$ of these observers is full rank, then $\sum_{k' \in C_{\text{linear node}}(k)} v_{k'}\|w_{k'}\| = 0$. If node $k$ is not co-linear with other students, then $v_k = 0$.

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-
Assumption 3. (a) Two teacher nodes \( j \neq j' \) are not \( \epsilon_0 \)-aligned. (b) The boundary band \( I_j(\epsilon) \) of each teacher \( j \) overlaps with the dataset:

\[
N_D[I_j(\epsilon)] \geq \tau \epsilon N_D
\]  

(8)

Intuitively, this means that we have sufficient samples near the boundary of each teacher nodes, in order to correctly identify the weights \( w_j \) 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 \( \partial E_j \). 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 \( \text{aug}(D) \) as (a) Teacher-agnostic:

\[
\text{aug}(D) = \{ x \pm 2\epsilon u / cK_1^{3/2} \mid x \in D, \ u = 1, \ldots, d \} \cup D
\]
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 $f^*(x)$, there exists compatible input distribution so that the dataset $\{(x, f^*(x))\}_{x \in D}$ 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, $A_1(x)$ and $B_1(x)$ are no longer constant over $x$, but are piece-wise constant. Note that $A_1(x)$ and $B_1(x)$ might contain exponential number of regions, $R = \{R_0, R_1, \ldots, R_L\}$, where in each region $R$, $A_1(x)$ and $B_1(x)$ 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 (Hann and Rolnick, 2019b, Rolnick & Kording, 2019, Hann & 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 $N = O(K^{5/2}d^2\epsilon^{-1}\kappa^{-1})$, if the first layer of the deep teacher network satisfies Assumption 3 and any student weight at the first layer $\hat{W}_1$ satisfies:

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

(2) Small gradient: $\|g_1(x, \hat{W})\|_{\infty} \leq \frac{\alpha_{\epsilon,j}}{5K_{\epsilon,j}^{5/2}\sqrt{d}} \epsilon$, for $x \in D'$,

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

6. Discussions

The theorems suggest a few interesting consequences:

Strong and weak teacher nodes. Theorems show that the convergence is dependent on $\alpha_{kj} = \|v_k^j\|_2^2$, where $v_k^j$ is the $j$-th column of $V_k^*$ and $v_k$ is the $k$-th column of $V_k^*$. Given the same magnitude of gradient norm, strong teacher $j$ (large $\|v_j^*\|$ and thus large $\alpha_{kj}$) 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. $E_t [\|g_t(x)\|_{\infty}] \leq \epsilon$. This means that $\|g_t(x)\|_{\infty} \leq \epsilon$ 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 $N(0, \sigma^2 I)$ with $\sigma = 10$. With other distributions (e.g., uniform $U[-1, 1]$), the result is similar. Appendix E 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 $\rho_{\text{mean}}$ as the mean of maximal normalized correlation ($f_j$ is the normalized activation of node $j$ over the evaluation set):

$$\rho_{\text{mean}} = \max_{j \in \text{teacher}} \max_{j' \in \text{student}} \rho_{jj'}, \quad \rho_{jj'} = \frac{f_j^* f_{j'}}{\|f_j^*\| \|f_{j'}\|}, \quad (9)$$

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:

$$\|v_j^*\| \sim 1/j^p, \quad (10)$$

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 ($1 \times 2 \times 5 \times 10 \times$), 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 $1 \times$ nodes has sufficient capacity to fit the teacher perfectly, it gets stuck despite long training.

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

Teacher-agnostic versus Teacher-aware. Theorem 4 and Theorem 5 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.
Student Specialization in Deep Rectified Networks With Finite Width and Input Dimension

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 un-specialized 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 \( \rho_{\text{mean}} \) (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.
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).

Acknowledgement


References


Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal


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-Carol & 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

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

Now suppose for layer \( l \), we have:

\[
g_l(x) = D_l(x) [ A_l(x) f^*_l(x) - B_l(x) f_l(x) ] \quad (11)
\]

\[
= D_l(x) V_l^T(x) [ V_l^*(x) f^*_l(x) - V_l(x) f_l(x) ] \quad (12)
\]

Using

\[
f_l(x) = D_l(x) W_l^T f_{l-1}(x) \quad (13)
\]

\[
f^*_l(x) = D_l^*(x) W_l^{*T} f^*_{l-1}(x) \quad (14)
\]

\[
g_{l-1}(x) = D_{l-1}(x) W_l g_l(x) \quad (15)
\]

we have:
\[ g_{l-1}(x) = D_{l-1}(x)W_lg_l(x) \]
\[ = \left( D_{l-1}(x) W_l D_l(x) V_l^T(x) \right) \left[ V_l^*(x) f_l^*(x) - V_l(x)f_l(x) \right] \]
\[ = D_{l-1}(x) V_l^{T*}(x) \left[ V_l^*(x) D_l(x) W_l^{T*} f_{l-1}^*(x) - V_l(x)D_l(x)W_l^T f_{l-1}(x) \right] \]
\[ = D_{l-1}(x) V_l^{T*}(x) \left[ V_l^*(x) f_{l-1}^*(x) - V_l(x)f_{l-1}(x) \right] \]
\[ = D_{l-1}(x) \left[ A_{l-1}(x) f_{l-1}^*(x) - B_{l-1}(x)f_{l-1}(x) \right] \]

\[ \Box \]

**B.2. Lemma**

**Lemma 2.** Denote \( D = \{ x_i \} \) as a dataset of \( N \) samples. If Assumption 2 holds, then either \( g_l(x_i; \hat{W}) = 0 \) or \( f_{l-1}(x_i; \hat{W}) = 0 \).

**Proof.** From Assumption 2, we know that for any batch \( B_j \), Eqn. 11 vanishes:

\[ \hat{W} = E_{\pi} \left[ g_l(x; \hat{W})f_{l-1}(x; \hat{W}) \right] = \sum_{i \in B_j} g_l(x_i; \hat{W})f_{l-1}^*(x_i; \hat{W}) = 0 \]  

Let \( U_i = g_l(x_i; \hat{W})f_{l-1}^*(x_i; \hat{W}) \). Note that \( B_j \) can be any subset of samples from the data distribution. Therefore, for a dataset of size \( N \), Eqn. 21 holds for all \( \binom{N}{B} \) 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 \( \| \hat{W} \|_\infty \leq \epsilon \), which is \( \| \sum_{i \in B_j} U_i \|_\infty \leq \epsilon \), then with simple Gaussian elimination for two batches \( B_1 \) and \( B_2 \) with only two sample difference, we will have for any \( i_1 \neq i_2 \), \( \| U_{i_1} - U_{i_2} \|_\infty = \| \sum_{i \in B_1} U_i - \sum_{i \in B_2} U_i \|_\infty \leq \| \sum_{i \in B_1} U_i \|_\infty + \| \sum_{i \in B_2} U_i \|_\infty = 2\epsilon \). Plug things back in and we have \( |B| U_i \|_\infty \leq |B| (|B| - 1) + 2\epsilon \), which is \( U_i \|_\infty \leq 2\epsilon \). If \( f_{l-1}(x; \hat{W}) \) has the bias term, then immediately we have \( \| g_l(x; \hat{W}) \|_\infty \leq \epsilon \).

**C. Two-layer, Infinite Samples and Zero Gradient**

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

**C.1. Corollary**

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

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

\[ V_{l-1}(x) = V_l(x)D_l(x)W_l^T, \quad V_{l-1}^*(x) = V_l^*(x)D_l^*(x)W_l^{T*} \]

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

Therefore, \( A_l(x) \) and \( B_l(x) \) are piece-wise constant with respect to input \( x \). They separate the region \( R_0 \) into constant regions with boundary points in a zero-measure set.
**C.2. Lemma 3**

**Lemma 3.** Consider $K$ ReLU activation functions $f_j(x) = \sigma(w_j^T x)$ for $j = 1 \ldots K$. If $w_j \neq 0$ and no two weights are co-linear, then $\sum_{j'} c_{j'} f_{j'}(x) = 0$ for all $x \in \mathbb{R}^{d+1}$ suggests that all $c_j = 0$.

**Proof.** Suppose there exists some $c_j \neq 0$ so that $\sum_{j'} c_{j'} f_{j'}(x) = 0$ for all $x$. Pick a point $x_0 \in \partial E_j$ so that $w_j^T x_0 = 0$ but all $w_j'^T x_0 \neq 0$ for $j' \neq j$, which is possible due to the distinct weight conditions. Consider an $\epsilon$-ball $B_{x_0, \epsilon} = \{x : \|x - x_0\| \leq \epsilon\}$. We pick $\epsilon$ so that $\text{sign}(w_j^T x)$ for all $j' \neq j$ remains the same within $B_{x_0, \epsilon}$ (Fig. 11(a)). Denote $[j^+]$ as the indices of activated ReLU functions in $B_{x_0, \epsilon}$ except $j$.

Then for all $x \in B_{x_0, \epsilon} \cap E_j$, we have:

$$h(x) \equiv \sum_{j'} c_{j'} f_{j'}(x) = c_j w_j^T x + \sum_{j' \in [j^+]} c_{j'} w_{j'}^T x = 0 \quad (23)$$

Since $B_{x_0, \epsilon}$ is a $d$-dimensional object rather than a subspace, for $x_0$ and $x_0 + \epsilon e_k \in B(x_0, \epsilon)$, we have

$$h(x_0 + \epsilon e_k) - h(x_0) = \epsilon(c_j w_{j+k} + \sum_{j' \in [j^+]} c_{j'} w_{j'+k}) = 0 \quad (24)$$

where $e_k$ is axis-aligned unit vector ($1 \leq k \leq d$). This yields

$$c_j w_j + \sum_{j' \in [j^+]} c_{j'} w_{j'} = 0_d \quad (25)$$

Plug it back to Eqn. (23) yields

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

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

$$c_j w_j + \sum_{j' \in [j^+]} c_{j'} w_{j'} = 0_{d+1} \quad (27)$$

However, if we pick $x' = x_0 - \epsilon \frac{w_j}{\|w_j\|} \in B_{x_0, \epsilon} \cap E_j^c$, then $f_j(x') = 0$ but $\sum_{j' \in [j^+]} f_{j'}(x') = -c_j w_j^T x' = \epsilon c_j$ and thus

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

which is a contradiction. \qed
C.3. Lemma 4

**Lemma 4 (Local ReLU Independence).** Let $R$ be an open set. Consider $K$ ReLU nodes $f_j(x) = \sigma(w_j^T x)$, $j = 1, \ldots, K$.\; $w_j \neq 0$, $w_j \neq \gamma w_{j'}$ for $j \neq j'$ with any $\gamma > 0$.

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

$$\sum_j c_j f_j(x) + c, w_j^T x = 0, \quad \forall x \in R$$

(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 $\partial E_j$ is in $R$, we can find a similar $x_0$ so that $x_0 \in \partial E_j \cap R$ and $x_0 \notin \partial E_{j'}$ for any $j' \neq j$. We construct $B_{x_0, \epsilon}$. Since $R$ is an open set, we can always find $\epsilon > 0$ so that $B_{x_0, \epsilon} \subseteq R$ and no other boundary is in this $\epsilon$-ball. Following similar logic of Lemma 3, $c_j = 0$. \hfill \Box

C.4. Theorem 1

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

From Assumption 1 since $\rho(x) > 0$ in $R_0$, from Lemma 2 we know that either $g_1(x) = 0$ or $x = 0$. However, since $x = [x, 1]$ has bias term, $g_1(x) = D_1(x) [A_1 f_1^*(x) - B_1 t_1(x)] = 0$. Picking node $k$, the following holds for every node $k$ and every $x \in R_0 \cap E_k$:

$$\alpha_k^T f^*(x) - \beta_k^T f(x) = 0$$

(30)

Here $\alpha_k^T$ is the $k$-th row of $A_1$, $A_1 = [\alpha_1, \ldots, \alpha_n]^T$ and similarly for $\beta_k$. 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 an 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 $c_s w_j^*$ (with $\|w_j^*\| = 1$), we have:

$$c_s = \alpha_{kj} - \sum_{k' \in \text{co-linear}(j)} \|w_{k'}\| \beta_{kk'}$$

(31)

“At most” because from Assumption 1 all teacher weights are not co-linear. Note that 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 co-linear($j$) $\neq \emptyset$ and there exists at least one student $k'$ that is aligned with the teacher $j$. \hfill \Box

C.5. Theorem 2

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

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

$$\sum_{k' \in \text{co-linear}(k)} \beta_{k_o, k'} \|w_{k'}\| = 0$$

(32)

which is

$$v_{k_o}^T \sum_{k' \in \text{co-linear}(k)} v_{k'} \|w_{k'}\| = 0$$

(33)

Since $k$ is observed by $C$ students $k_1^o, k_2^o, \ldots, k_C^o$, then we have:

$$v_{k_o}^T \sum_{k' \in \text{co-linear}(k)} v_{k'} \|w_{k'}\| = 0$$

(34)
By the condition, all the $C$ vectors $\mathbf{v}_{k_\lambda}^T \in \mathbb{R}^C$ are linear independent, then we know that

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

C.6. Corollary [1]

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

$$\sum_{s} \sum_{k \in \text{co-linear}(s)} \mathbf{v}_k f_k(\mathbf{x}) = \sum_{s} \sum_{k \in \text{co-linear}(s)} \mathbf{v}_k \parallel \mathbf{w}_k \parallel \sigma(\mathbf{w}_s^T \mathbf{x}) \quad (36)$$

$$= \sum_{s} \sigma(\mathbf{w}_s^T \mathbf{x}) \sum_{k \in \text{co-linear}(s)} \mathbf{v}_k \parallel \mathbf{w}_k \parallel \quad (37)$$

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, $\sum_{k \in \text{co-linear}(s)} \mathbf{v}_k \parallel \mathbf{w}_k \parallel = 0$ and thus the net contribution is zero. \hfill \Box

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 $\mathbf{w}_j^T \tilde{\mathbf{u}}_{j'} = \sin \tilde{\theta}_{jj'}$.

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

$$\tilde{\mathbf{u}}_{j'} = \frac{1}{\sin \tilde{\theta}_{jj'}} P_{\mathbf{w}_j} \tilde{\mathbf{w}}_{j'} \quad (38)$$

It is easy to verify that $\|\tilde{\mathbf{u}}_{j'}\| = 1$ and $\mathbf{w}_j^T \tilde{\mathbf{u}}_{j'} = \sin \tilde{\theta}_{jj'}$. \hfill \Box

**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:

$$\sin \tilde{\theta}_{jj'} \leq M_1 \epsilon, \quad |b_j - b_{j'}| \leq M_2 \epsilon \quad (39)$$

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

$$N[D' \cap I_\mathbf{w}(\epsilon)] \leq \eta_{\mathbf{w}} \epsilon N_{D'} + (d + 1) \quad (40)$$

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

Note that $D'$ is always a constrained $\eta$-dataset for sufficiently large $\eta$.

**Lemma 6.** Consider $K$ hyperplanes, 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 $\eta$-dataset $D \subseteq I_j(\epsilon)$, if

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

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

Here

$$M_1 = (\gamma + 3)\eta K, \quad M_2 = 1 + \gamma + (\gamma + 3)\eta K \left(|b_j| + \sqrt{(3 + \gamma)\mu K \over 1 + \gamma}\right) \quad (41)$$
Case 2

Therefore, by the definition of $\eta$-dataset, we have:

$$N_D[I_j'((\gamma\epsilon)) \leq N_D \left[ |u_j^T x| \leq \frac{1 + \gamma}{M_1 \epsilon} \right] \leq \frac{1 + \gamma}{(3 + \gamma)\eta K}$$

(45)

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

$$\tilde{w}_j^T x = \cos \tilde{\theta}_{jj'} \tilde{w}_j^T \tilde{x} + \sin \tilde{\theta}_{jj'} \tilde{u}_j^T \tilde{x}$$

(46)
which means that
\[ w_j^T x - b_j = \cos \tilde{\theta}_{jj'} (w_j^T x - b_j) + \sin \tilde{\theta}_{jj'} \tilde{u}_j^T \tilde{x} \] (47)

Therefore, we have:
\[ |\tilde{u}_j^T \tilde{x}| = \frac{1}{\sin \tilde{\theta}_{jj'}} |w_j^T x - b_j - \cos \tilde{\theta}_{jj'} (w_j^T x - b_j)| \] (48)
\[ \geq \frac{1}{\sin \theta_{jj'}} \left[ |b_j - b_j| - (1 - \cos \tilde{\theta}_{jj'}) |b_j| - |w_j^T x| - |w_j^T x| \right] \] (49)

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,
\[ |\tilde{u}_j^T \tilde{x}| \geq \frac{M_2 - 1 - \gamma}{M_1} - |b_j| = \sqrt{\frac{(3 + \gamma)\mu K}{1 + \gamma}} \] (50)

Therefore, we have:
\[ N_D[I_{j'}(\gamma\epsilon)] \leq N_D \left[ |\tilde{u}_j^T \tilde{x}| \geq \sqrt{\frac{(3 + \gamma)\mu K}{1 + \gamma}} \right] \leq \frac{1 + \gamma}{(3 + \gamma)K} N_D \] (51)

Combining the two cases, since \( N_D \geq (3 + \gamma)K(d + 1) \), we know that
\[ \sum_{j' \neq j} N_D[I_{j'}(\gamma\epsilon)] \leq \frac{1 + \gamma K - 1}{3 + \gamma} K N_D + (K - 1)(d + 1) \] (52)
\[ < \frac{1 + \gamma}{3 + \gamma} N_D + \frac{1}{3 + \gamma} N_D \] (53)
\[ = \left( 1 - \frac{1}{3 + \gamma} \right) N_D < N_D \] (54)

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

**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 \( |b_j| \). For small \( \epsilon \), this term is negligible.

**Lemma 7.** Define
\[ h(x) = \sum_{j'=1}^{K} c_{j'} \sigma(w_{j'}^T x) + c.w_j^T x \] (55)

Suppose there exists \( x_j \in I_j(\epsilon_l) \setminus \cup_{j' \neq j} I_{j'}(\epsilon_h) \) with \( \epsilon_l < \epsilon_h \) and there exists a vector \( \tilde{e} \) so that \( \epsilon_l < \epsilon_0 \leq w_j^T \tilde{e} \leq \epsilon_h \). Construct 2 datapoints \( x_j^\pm = x_j \pm \tilde{e} \) and set \( D = \{ x_j^+, x_j^- \} \). Then there exists \( x \in D \) so that \( |h(x)| > \frac{|x_j^\pm|}{2}(\epsilon_0 - \epsilon_l) \).

**Proof.** We show that the three points \( x_j \) and \( x_j^\pm \) are on the same side of \( \partial E_{j'} \) for any \( j' \neq j \). This can be achieved by checking whether \( (w_j^T x_j)(w_j^T x_j^\pm) \geq 0 \) (Fig. 12(b) and (c)):
\[ (w_j^T x_j)(w_j^T x_j^\pm) = (w_j^T x_j)^2 \] (56)
\[ = (w_j^T x_j^\pm)^2 \] (57)
\[ = |w_j^T x_j|^2 |w_j^T x_j| \] (58)

Since \( x_j \notin I_{j'}(\epsilon_h) \) and \( w_j^T \tilde{e} \leq \epsilon_h \), we have:
\[ |w_j^T x_j| \pm w_j^T \tilde{e} > \epsilon_h \pm \epsilon_h \geq 0 \] (59)
Therefore, \((w_j^T x_j)(w_j^T x_j^+) \geq 0\) and the three points \(x_j\) and \(x_j^+\) are on the same side of \(\partial E_{j'}\) for any \(j' \neq j\).

If the conclusion is not true, then consider \(h(x_j^+) + h(x_j^-) - 2h(x_j)\). Since \(x_j^+ + x_j^- = 2x_j\), we know that all terms related to \(w\) and \(w_{j'}\) 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,

\[
h(x_j^+) + h(x_j^-) - 2h(x_j) = c_j \left[ \sigma(w_j^T x_j^+) + \sigma(w_j^T x_j^-) - 2\sigma(w_j^T x_j) \right]
\]

Since \(|w_j^T x_j| \leq \epsilon_t\) and \(|w_j^T \hat{e} \geq \epsilon_t > \epsilon_t|\), it is always the case that:

(a) \(w_j^T x_j^+ = w_j^T x_j + w_j^T \hat{e} > 0\) and by ReLU properties \(\sigma(w_j^T x_j^+) = w_j^T x_j^+\).

(b) \(w_j^T x_j^- = w_j^T x_j - w_j^T \hat{e} < 0\) so by ReLU properties \(\sigma(w_j^T x_j^-) = 0\).

Therefore if \(w_j^T x_j \geq 0\) then:

\[
|h(x_j^+) + h(x_j^-) - 2h(x_j)| = |c_j w_j^T(x_j^+ - 2x_j)| = |c_j w_j^T(x_j + \hat{e} - 2x_j)| \\
= |c_j w_j^T(\hat{e} - x_j)| \geq |c_j(\epsilon_0 - w_j^T x_j)| \\
\geq |c_j|(|\epsilon_0 - \epsilon_t|)
\]

if \(w_j^T x_j < 0\) then:

\[
|h(x_j^+) + h(x_j^-) - 2h(x_j)| = |c_j w_j^T x_j^+| = |c_j w_j^T(x_j + \hat{e})| \geq |c_j|(|\epsilon_0 - \epsilon_t|)
\]

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

\[
|h(x_j^+) + h(x_j^-) - 2h(x_j)| \leq \frac{4}{5}|c_j|(|\epsilon_0 - \epsilon_t|)
\]

which is a contradiction.

For Leaky ReLU, the proof is similar except that the final condition has an additional \(1 - \epsilon_{\text{leaky}}\) factor.

**Assumption 4.** (a) Two teacher nodes \(j \neq j'\) are not \(\epsilon_0\)-aligned. (b) The boundary band \(I_j(\epsilon)\) of each teacher \(j\) overlaps with the dataset:

\[N_D[I_j(\epsilon)] \geq \tau \epsilon N_D\]

**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 \(N = \Theta(\epsilon K^{5/2} d^2 \tau^{-1} \epsilon^{-1} \kappa^{-1})\), for any teacher satisfying Assumption 3 and student trained on \(D' = \text{aug}(D)\) whose weight \(W\) 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:

\[
\|g_1(x, W)\|_{\infty} \leq \frac{\alpha_{\text{leaky}}}{5cK^{5/2} \sqrt{d}} \epsilon, \ x \in D',
\]

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

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

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

\[\text{aug}(D) = \{x_k^+ = x \pm 2\epsilon \hat{e}_u/cK^{3/2}, \ x \in D, \ u = 1, \ldots, d\} \cup D\]

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

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

\[N_D[I_j(\epsilon')] = N[D \cap I_j(\epsilon')] \geq \tau \epsilon' N_D = O(K d^{3/2} \kappa^{-1})\]
and thus for $N_D[I_j(\epsilon') \cap E_k]$, we know that
\[
N_D[I_j(\epsilon') \cap E_k] \geq \kappa N_D[I_j(\epsilon')] = \mathcal{O}(Kd^{3/2})
\] (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:
\[
\sin \theta_{j,k'} \leq M_1'_{j} \epsilon = M_2'_{j} \epsilon \leq M_2'_{j} \epsilon = M_2'_{j} \epsilon
\] (70)
Where $M_1'_{j} = \Theta(K\gamma)$ and $M_2'_{j} = \Theta(K^{3/2} \gamma)$. Therefore, $M_1'_{j} = \Theta(K\gamma/cK^{3/2} \sqrt{d}) = o(1)$ and $M_2'_{j} = \Theta(1)$. Choosing the constant $c > 0$ so that we have $M_1'_{j} \leq 1$ and $M_2'_{j} \leq 1$ and thus
\[
\sin \theta_{j,k'} \leq \epsilon, \quad |b_j^* - b_{k'}| \leq \epsilon,
\] (71)
which means that $(j, k')$ are $\epsilon$-aligned. Note that no other teacher is $\epsilon$-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 $x_j$ so that (note that $\gamma = 2\sqrt{d}$):
\[
x_j \in I_j(\epsilon') \setminus \cup_{j' \neq j} I_j(2\sqrt{d} \epsilon').
\] (72)
Here all $j'$ includes all teacher and student nodes (a total of $K$ nodes), excluding the current node $j$ under consideration. Since $\{\hat{e}_u\}^d_{u=1}$ forms orthonormal bases, there exists at least one $u$ so that $1 \geq w_j^T \hat{e}_u \geq 1/\sqrt{d}$ (with proper sign flipping of $\hat{e}_u$). Let $\hat{e} = 2\epsilon \hat{e}_u / cK^{3/2} = 2\epsilon \sqrt{d} \hat{e}_u$, we have $2\epsilon' \leq w_j^T \hat{e} \leq 2\epsilon \sqrt{d} \epsilon'$. Applying Lemma 7 and we know the additional samples required in the lemma is already in $\text{aug}(D)$. Therefore, there exists $x \in \text{aug}(D) \subset D$ so that $|h(x)| > \frac{|x_{j,k}|}{5}(2\epsilon' - \epsilon') = \frac{|x_{j,k}|}{5cK^{3/2} \sqrt{d} \epsilon'}$, 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 $N = \Theta(cK^{5/2} d\gamma^{-1} \epsilon^{-1})$, given a teacher network $W^*$ satisfying Assumption D and student trained on $D' = \text{aug}(D, W^*)$ whose weight $W$ satisfies
\begin{enumerate}
  \item For $\epsilon \in [0, \epsilon_0]$, the hyperplane band $I_j(\epsilon)$ of a teacher is observed by a student node $k$: $N_D[I_j(\epsilon) \cap E_k] \geq \kappa N_D[I_j(\epsilon')]$;
  \item Small gradient: $\|g_1(x, \hat{W})\|_\infty \leq \frac{\alpha_k}{5cK^{3/2} \sqrt{d} \epsilon}$, for $x \in D'$,
\end{enumerate}
then there exists a student $k'$ so that $(j, k')$ is $\epsilon$-aligned.

**Proof.** Pick $\epsilon' = \epsilon / cK^{3/2}$ ($c$ defined as in Theorem 6). For dataset $D$, by Assumption 3 we know that for $D \cap I_j(\epsilon') \cap E_k$:
\[
N[D \cap I_j(\epsilon') \cap E_k] = N_D[I_j(\epsilon') \cap E_k] \geq \kappa N_D[I_j(\epsilon')] \geq \kappa \epsilon' \kappa N_D = \mathcal{O}(Kd)
\] (73)
Apply Lemma 6 with $\gamma = 2$ and similarly we know that either there exists a student $k'$ so that $(j, k')$ is $\epsilon$-align (with $M_1'_{j} \leq 1$ and $M_2'_{j} \leq 1$), or there exists $x_j$ such that
\[
x_j \in I_j(\epsilon') \setminus \cup_{j' \neq j} I_j(2\epsilon').
\] (74)
Setting $\hat{e} = 2\epsilon \epsilon^* w_j^*$ and we know that $w_j^T \hat{e} = 2\epsilon'$. 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 6**

**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 $K_j = m_j + n_j$ boundaries, we need to think about all the $Q$ boundaries introduced by the top-level and obtain a data point $x_j$ so that it is within the boundary of node $j$, but far away from all other possible boundaries:
\[
x_j \in I_j(\epsilon') \setminus \cup_{j' \neq j} I_j(2\sqrt{d} \epsilon').
\] (75)
Where $j'$ includes all the boundaries induced. This could be exponential. Note that for each intermediate node $j'$, its boundary $w_j f = 0$ will be “bent” whenever the underlying feature $f$, which is the output of a set of ReLU / Leaky ReLU nodes, has shifted their activation patterns.
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 \( W^{(1)} \) and \( W^{(2)} \). For \( W^{(1)} \), \( w_1^{(1)} = w_1^* \), \( w_2^{(1)} = w_2^* \) and \( w_3^{(1)} \) is an un-specialized node whose fan-out weights are zero (\( v_1^{(1)} = 0 \)). For \( W^{(2)} \), \( w_3^{(2)} = w_1^* \), \( w_2^{(2)} = w_2^* \) and \( w_1^{(2)} \) is an un-specialized node whose fan-out weights are zero (\( v_1^{(2)} = 0 \)).

If we directly connect these two solutions using a straight line, the intermediate solution will be no longer low-cost since a linear combination

\[
\lambda w_1^{(1)} + (1 - \lambda) w_1^{(2)} = \lambda w_1^* + (1 - \lambda) w_1^*
\]

can be a random (un-specialized) vector, and its corresponding fan-out weights \( \lambda v_1^{(1)} + (1 - \lambda) v_1^{(2)} = \lambda v_1^{(1)} \) 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 \( w_1^* \) from node 1 to node 3. We list the line segment construction as follows:

- **Start from** \( W^{(1)} \).
- **(a)-(c)**. Blend \( w_1^* \) into an un-specialized weight: \( w_3(t) = (1 - t)w_3^{(1)} + tw_1^* \). This won’t change the output since \( v_3(t) = 0 \).
- **(c)-(d)**. Move \( v_1^{(1)} \) from node 1 to node 3:

\[
\begin{align*}
    v_3(t) & = (1 - t)v_3^{(1)} + tv_1^{(1)} = tv_1^{(1)} \\
    v_1(t) & = (1 - t)v_1^{(1)} + tv_3^{(1)} = (1 - t)v_1^{(1)}
\end{align*}
\]

This won’t change the output since \( v_1(t) + v_3(t) = v_1^{(1)} \) and their weights are both \( w_1^* \).
- **(d)-(b)**. Change \( w_1 \) to be the unspecified weight in \( W^{(2)} \): \( w_1(t) = (1 - t)w_1^* + tw_2^{(1)} \). This won’t change the output since now \( v_1 = 0 \).
- **Arrive at** \( W^{(2)} \).

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.