Supplementary Materials of Understanding Robustness in Teacher-Student Setting: A New Perspective

A Proofs

A.1 Lemma

Lemma 1. *If $\tilde{\mathbf{x}} = U\tilde{\mathbf{y}} + \tilde{\mathbf{x}}_0$, where $U \in \mathbb{R}^{d\times d'}$, then the inner product $\mathbf{w}^\intercal\mathbf{x}$ in the original space can be written as the inner product in the reduced space $\mathbf{w}_y^\intercal\mathbf{y}$ with*

$$\mathbf{w}^\intercal\mathbf{x} = \mathbf{w}_y^\intercal\mathbf{y}, \quad \text{for } \mathbf{w}_y := \begin{bmatrix} \tilde{\mathbf{w}}_y \\ b_y \end{bmatrix} = \begin{bmatrix} U^\intercal\tilde{\mathbf{w}} \\ \tilde{\mathbf{w}}^\intercal\mathbf{x}_0 + b \end{bmatrix} \tag{5}$$

Proof. Since the augmented vector $\mathbf{x} := [\tilde{\mathbf{x}}; 1] \in \mathbb{R}^{d+1}$, the inner product $\mathbf{w}^\intercal\mathbf{x}$ can be written as:

$$\mathbf{w}^\intercal\mathbf{x} := \tilde{\mathbf{w}}^\intercal\tilde{\mathbf{x}} + b = \underbrace{\tilde{\mathbf{w}}^\intercal U}_{\tilde{\mathbf{w}}_y^\intercal}\tilde{\mathbf{y}} + \underbrace{\tilde{\mathbf{w}}^\intercal\mathbf{x}_0 + b}_{b_y} \tag{6}$$

and the conclusion follows. □

A.2 Theorem 1

Proof. For low-dimensional input space $\mathcal{X}$, we could always find a set of orthonormal bases $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{d'}]$ so that for any point $\tilde{\mathbf{x}} \in X$, we have $\tilde{\mathbf{x}} = U\tilde{\mathbf{y}} + \tilde{\mathbf{x}}_0$. Therefore, by Lemma 1, the inner product $\mathbf{w}^\intercal\mathbf{x}$ can be written as

$$\mathbf{w}^\intercal\mathbf{x} = \mathbf{w}_y^\intercal\mathbf{y}, \quad \text{for } \mathbf{w}_y := \begin{bmatrix} \tilde{\mathbf{w}}_y \\ b_y \end{bmatrix} = \begin{bmatrix} U^\intercal\tilde{\mathbf{w}} \\ \tilde{\mathbf{w}}^\intercal\mathbf{x}_0 + b \end{bmatrix} \tag{7}$$

Then $\mathbf{y}$ is full-rank in X and we can apply Lemma 3 in Tian (2019) for the reduced space of $\mathbf{y}$ to draw the conclusion that for each teacher node j whose boundary is observed by a student node k with $\alpha_{jk} \neq 0$, there exists at least one student node k' so that $\mathbf{w}^*_{y,j} = \lambda\mathbf{w}_{y,k}$ with $\lambda > 0$. Taking its first d' components, we have $U^\intercal\tilde{\mathbf{w}}^*_j = \lambda U^\intercal\tilde{\mathbf{w}}_k$. Notice that $\mathrm{Proj}_{\mathcal{X}}[\tilde{\mathbf{w}}^*_j] = UU^\intercal\tilde{\mathbf{w}}^*_j$, we have $\mathrm{Proj}_{\mathcal{X}}[\tilde{\mathbf{w}}^*_j] = \lambda\mathrm{Proj}_{\mathcal{X}}[\tilde{\mathbf{w}}_k]$. □

A.3 Lemma 2

Lemma 2 (Relation between Hyperplanes (Lemma 5 in Tian (2019)))**.** *Let $\mathbf{w}_j$ and $\mathbf{w}_{j'}$ be two distinct hyperplanes with $\|\tilde{\mathbf{w}}_j\| = \|\tilde{\mathbf{w}}_{j'}\| = 1$. Denote $\theta_{jj'}$ as the angle between the two vectors $\mathbf{w}_j$ and $\mathbf{w}_{j'}$. Then there exists $\tilde{\mathbf{u}}_{j'} \perp \tilde{\mathbf{w}}_j$ and $\mathbf{w}_{j'}^\intercal\tilde{\mathbf{u}}_{j'} = \sin\theta_{jj'}$.*

A.4 Lemma 3

Lemma 3 (Evidence of Data points on Misalignment)**.** *Let $R \subset \mathbb{R}^d$ be an open set. Consider K ReLU nodes $f_j(\mathbf{x}) = \sigma(\mathbf{w}_j^\intercal\mathbf{x})$, $j = 1, \ldots, K$. $\|\tilde{\mathbf{w}}_j\| = 1$, $\mathbf{w}_j$ are not co-linear. Then for a node j with $\partial E_j \cap R \neq \emptyset$, either of the conditions holds:*

(1) There exists node $j' \neq j$ so that $\sin\theta_{jj'} \leq MK\epsilon/|c_j|$ and $|b_{j'} - b_j| \leq M_2\epsilon/|c_j|$.

(2) There exists $\mathbf{x}_j \in \partial E_j \cap R$ so that for any $j' \neq j$, $|\mathbf{w}_{j'}^\intercal\mathbf{x}_j| > 5\epsilon/|c_j|$.

where:

- *$\theta_{jj'}$ is the angle between $\tilde{\mathbf{w}}_j$ and $\tilde{\mathbf{w}}_{j'}$,*
- *r is the radius of a $d-1$ dimensional ball contained in $\partial E_j \cap R$,*
- *$M = \frac{10}{r}\sqrt{\frac{d}{2\pi}}$, $M_0 = \max_{\mathbf{x}\in\partial E_j \cap R} \|\mathbf{x}\|$ and $M_2 = 2M_0MK + 5$.*

Proof. Define $q_j = 5\epsilon/|c_j|$. For each $j' \neq j$, define $I_{j'} = \{\mathbf{x} : |\mathbf{w}_{j'}^{\intercal}\mathbf{x}| \leq q_j,\ \mathbf{x} \in \partial E_j\}$. We prove by contradiction. Suppose for any $j' \neq j$, $\sin\theta_{jj'} > KM\epsilon/|c_j|$ or $|b_{j'} - b_j| > M_2\epsilon/|c_j|$. Otherwise the theorem already holds.

Case 1. When $\sin\theta_{jj'} > KM\epsilon/|c_j|$ holds.

From Lemma 2, we know that for any $\mathbf{x} \in \partial E_j$, if $\mathbf{w}_{j'}^{\intercal}\mathbf{x} = -q_j$, with $a_{j'} \leq \frac{2q_j|c_j|}{MK\epsilon} = \frac{10}{MK}$, we have $\mathbf{x}' = \mathbf{x} + a_{j'}\mathbf{u}_{j'} \in \partial E_j$ and $\mathbf{w}_{j'}^{\intercal}\mathbf{x}' = +q_j$.

Consider a $d-1$-dimensional sphere $B \subseteq \Omega_j$ and its intersection of $I_{j'} \cap B$ for $j' \neq j$. Suppose the sphere has radius r. For each $I_{j'} \cap B$, its $d-1$-dimensional volume is upper bounded by:

$$V(I_{j'} \cap B) \leq a_{j'} V_{d-2}(r) \leq \frac{10}{MK} V_{d-2}(r) \tag{8}$$

where $V_{d-2}(r)$ is the $d-2$-dimensional volume of a sphere of radius r. Intuitively, the intersection between $\mathbf{w}_{j'}^{\intercal}\mathbf{x} = -q_j$ and B is at most a $d-2$-dimensional sphere of radius r, and the "height" is at most $a_{j'}$.

Case 2. When $\sin\theta_{jj'} \leq KM\epsilon/|c_j|$ but $|b_{j'} - b_j| > M_2\epsilon/|c_j|$ holds.

In this case, we want to show that for any $\mathbf{x} \in \Omega_j$, $|\mathbf{w}_{j'}^{\intercal}\mathbf{x}| > q_j$ and thus $I_{j'} \cap B = \emptyset$. If this is not the case, then there exists $\mathbf{x} \in \Omega_j$ so that $|\mathbf{w}_{j'}^{\intercal}\mathbf{x}| \leq q_j$. Then since $\mathbf{x} \in \partial E_j$, we have:

$$|\mathbf{w}_{j'}^{\intercal}\mathbf{x}| = |(\mathbf{w}_{j'} - \mathbf{w}_j)^{\intercal}\mathbf{x}| = |(\tilde{\mathbf{w}}_{j'} - \tilde{\mathbf{w}}_j)^{\intercal}\tilde{\mathbf{x}} + (b'_j - b_j)| \leq q_j \tag{9}$$

Therefore, from Cauchy inequality and triangle inequality, we have:

$$\|\tilde{\mathbf{w}}_{j'} - \tilde{\mathbf{w}}_j\|\|\tilde{\mathbf{x}}\| \geq |(\tilde{\mathbf{w}}_{j'} - \tilde{\mathbf{w}}_j)^{\intercal}\tilde{\mathbf{x}}| \geq |b'_j - b_j| - |\mathbf{w}_{j'}^{\intercal}\mathbf{x}| \tag{10}$$

From the condition, we have $\|\tilde{\mathbf{w}}_{j'} - \tilde{\mathbf{w}}_j\| = 2\sin\frac{\theta_{jj'}}{2} \leq 2\sin\theta_{jj'} \leq 2KM\epsilon/|c_j|$. Then

$$2M_0MK\epsilon/|c_j| \geq |(\tilde{\mathbf{w}}_{j'} - \tilde{\mathbf{w}}_j)^{\intercal}\tilde{\mathbf{x}}| \geq |b_{j'} - b_j| - q_j > M_2\epsilon/|c_j| - 5\epsilon/|c_j| \tag{11}$$

which is equivalent to:

$$2M_0MK > M_2 - 5 \tag{12}$$

which means that

$$M_2 < 2M_0MK + 5 \tag{13}$$

This is a contradiction. Therefore, $I_{j'} \cap B = \emptyset$ and thus $V(I_{j'} \cap B) = 0$.

Volume argument. Therefore, from the definition of M, we have $V(B) = V_{d-1}(r) \geq r\sqrt{\frac{2\pi}{d}}V_{d-2}(r) = \frac{10}{M}V_{d-2}(r)$, then we have:

$$V(B) \geq \frac{10}{M}V_{d-2}(r) > (K-1)\cdot\frac{10}{MK}V_{d-2}(r) \geq \sum_{j'\neq j, j' \text{ in case 1}} V(I_{j'} \cap B) \tag{14}$$

This means that there exists $\mathbf{x}_j \in B \subseteq \Omega_j$ so that $\mathbf{x}_j \notin I_{j'} \cap B$ for any $j' \neq j$ and j' in case 1. That is,

$$|\mathbf{w}_{j'}^{\intercal}\mathbf{x}_j| > q_j \tag{15}$$

On the other hand, for j' in case 2, the above condition holds for entire Ω_j, and thus hold for the chosen $\mathbf{x}_j$. $\square$

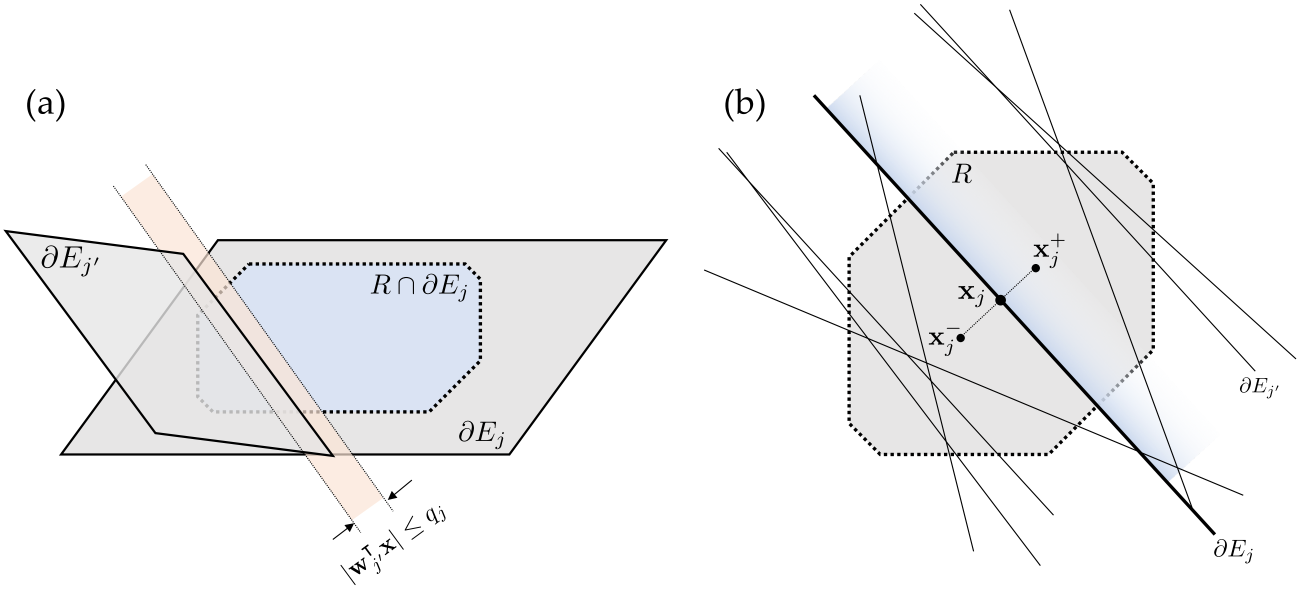

Figure 13: **(a)** Lemma 3. **(b)** Lemma 4.

A.5 Lemma 4

Lemma 4 (Local ReLU Independence, Noisy case)**.** *Let R be an open set. Consider K ReLU nodes $f_j(\mathbf{x}) = \sigma(\mathbf{w}_j^\intercal \mathbf{x})$, $j = 1, \ldots, K$. $\|\tilde{\mathbf{w}}_j\| = 1$, $\mathbf{w}_j$ are not co-linear. If there exists $c_1, \ldots, c_K, c_\bullet$ and ϵ so that the following is true:*

$$\left| \sum_j c_j f_j(\mathbf{x}) + c_\bullet \mathbf{w}_\bullet^\intercal \mathbf{x} \right| \le \epsilon, \quad \forall \mathbf{x} \in R \tag{16}$$

and for a node j, $\partial E_j \cap R \neq \emptyset$. Then there exists node $j' \neq j$ so that $\sin\theta_{jj'} \le MK\epsilon/|c_j|$ and $|b_{j'} - b_j| \le M_2\epsilon/|c_j|$, where r, M, M_2 are defined in Lemma 3 but with $r' = r - 5\epsilon/|c_j|$.

Proof. Let $q_j = 5\epsilon/|c_j|$ and $\Omega_j = \{\mathbf{x} : \mathbf{x} \in \partial E_j \cap R, \ \ B(\mathbf{x}, q_j) \subseteq R\}$. If situation (1) in Lemma 3 happens then the theorem holds. Otherwise, applying Lemma 3 with $R' = \{\mathbf{x} : \mathbf{x} \in R, \ \ B(\mathbf{x}, q_j) \subseteq R\}$ and there exists $\mathbf{x}_j \in \Omega_j$ so that

$$|\mathbf{w}_{j'}^\intercal \mathbf{x}_j| \ge q_j = 5\epsilon/|c_j| \tag{17}$$

Let two points $\mathbf{x}_j^\pm = \mathbf{x}_j \pm q_j \tilde{\mathbf{w}}_j \in R$. In the following 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 $(\mathbf{w}_{j'}^\intercal \mathbf{x}_j)(\mathbf{w}_{j'}^\intercal \mathbf{x}_j^\pm) \ge 0$ (Figure 13):

$$\begin{aligned} (\mathbf{w}_{j'}^\intercal \mathbf{x}_j)(\mathbf{w}_{j'}^\intercal \mathbf{x}_j^\pm) &= (\mathbf{w}_{j'}^\intercal \mathbf{x}_j)\left[\mathbf{w}_{j'}^\intercal (\mathbf{x}_j \pm q_j \tilde{\mathbf{w}}_j)\right] && (18) \\ &= (\mathbf{w}_{j'}^\intercal \mathbf{x}_j)^2 \pm q_j (\mathbf{w}_{j'}^\intercal \mathbf{x}_j)\mathbf{w}_{j'}^\intercal \tilde{\mathbf{w}}_j && (19) \\ &= |\mathbf{w}_{j'}^\intercal \mathbf{x}_j|(|\mathbf{w}_{j'}^\intercal \mathbf{x}_j| \pm q_j \mathbf{w}_{j'}^\intercal \tilde{\mathbf{w}}_j) && (20) \end{aligned}$$

Since $|\mathbf{w}_{j'}^\intercal \tilde{\mathbf{w}}_j| \le 1$, it is clear that $(\mathbf{w}_{j'}^\intercal \mathbf{x}_j)(\mathbf{w}_{j'}^\intercal \mathbf{x}_j^\pm) \ge 0$. Therefore the three points $\mathbf{x}_j$ and $\mathbf{x}_j^\pm$ are on the same side of $\partial E_{j'}$ for any $j' \neq j$.

Let $h(\mathbf{x}) = \sum_j c_j f_j(\mathbf{x}) + c_\bullet \mathbf{w}_\bullet^\intercal \mathbf{x}$, then $|h(\mathbf{x})| \le \epsilon$ for $\mathbf{x} \in R$. 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 (they are in the same side of the boundary $\partial E_{j'}$) and thus:

$$4\epsilon \ge |h(\mathbf{x}_j^+) + h(\mathbf{x}_j^-) - 2h(\mathbf{x}_j)| = |c_j q_j \mathbf{w}_j^\intercal \mathbf{w}_j| = |c_j| q_j = 5\epsilon \tag{21}$$

which is a contradiction. □

A.6 Theorem 2

Proof. Note that from Theorem 1, any input $\mathbf{x} \in \mathcal{U} \cap R$ can be written as $\tilde{\mathbf{x}} = U\tilde{\mathbf{y}} + \tilde{\mathbf{x}}_0$, where $U \in \mathbb{R}^{d\times d'}$ is a column-orthogonal matrix (i.e, $U^\intercal U = I_{d'\times d'}$). Also from Lemma 1, any inner-product $\mathbf{w}^\intercal \mathbf{x}$ can be written as $\mathbf{w}_y^\intercal \mathbf{y}$, with $\mathbf{w}_y := [\tilde{\mathbf{w}}_y; \tilde{\mathbf{w}}^\intercal \mathbf{x}_0 + b]$ and $\tilde{\mathbf{w}}_y := U^\intercal \tilde{\mathbf{w}}$, and the inner product of two projected weights is:

$$\tilde{\mathbf{p}}_j^\intercal \tilde{\mathbf{p}}_k := \mathrm{Proj}_{\mathcal{U}}[\tilde{\mathbf{w}}_j]^\intercal \mathrm{Proj}_{\mathcal{U}}[\tilde{\mathbf{w}}_k] = \tilde{\mathbf{w}}_j U U^\intercal U U^\intercal \tilde{\mathbf{w}}_k = \tilde{\mathbf{w}}_j U U^\intercal \tilde{\mathbf{w}}_k = \tilde{\mathbf{w}}_{y,j}^\intercal \tilde{\mathbf{w}}_{y,k} \tag{22}$$

Therefore, all the ReLU activations can be written in the reduced space, and the projected angle $\theta^{\mathcal{U}}_{jk} := \arccos \tilde{\mathbf{p}}_j^\intercal \tilde{\mathbf{p}}_k$ we are aiming for is also defined in the reduced space $\mathbf{y}$. Applying Lemma 4 on the reduced space $\mathbf{y}$ with $r = r(\mathcal{U} \cap R \cap \partial E_j)$, and the conclusion follows. □

A.7 Corollary 1

Proof. By Theorem 2, we know that for a node k_0, if it is observed by another student node k, then there exists a node j (can be either a teacher or another student node) so that their projected angle $\sin \theta^{\mathcal{U}}_{jk_0}$ has the following upper bound:

$$\sin \theta^{\mathcal{U}}_{jk_0} \le MK\epsilon / |\alpha_{kk_0}| \tag{23}$$

where $\alpha_{kk_0} := \mathbf{v}_k^\intercal \mathbf{v}_{k_0}$, $\mathbf{v}_k \in \mathbb{R}^C$ is the fan-out weights, and C is the number of output for the two-layer network. On the other hand, by the condition, we have $\sin \theta^{\mathcal{U}}_{jk_0} \ge c_0$ for any other teacher and student nodes, including j. Therefore, we have:

$$c_0 \le \sin \theta^{\mathcal{U}}_{jk_0} \le MK\epsilon / |\alpha_{kk_0}| \tag{24}$$

which leads to

$$|\mathbf{v}_k^\intercal \mathbf{v}_{k_0}| = |\alpha_{kk_0}| \le MK\epsilon / c_0 \tag{25}$$

If the student node k_0 is observed by C independent observers $k_1, k_2, \ldots, k_C$, then we have:

$$|\mathbf{v}_{k_m}^\intercal \mathbf{v}_{k_0}| = |\alpha_{k_0 k_m}| \le MK\epsilon / c_0, \quad m = 1, \ldots, C \tag{26}$$

Let $Q := [\mathbf{v}_{k_1}, \mathbf{v}_{k_2}, \ldots, \mathbf{v}_{k_C}] \in \mathbb{R}^{C\times C}$, then we have $\|Q^\intercal \mathbf{v}_{k_0}\|_\infty \le MK\epsilon/c_0$ and:

$$\|\mathbf{v}_{k_0}\|_\infty \le \|Q^{-\intercal}\|_\infty \|Q^\intercal \mathbf{v}_{k_0}\|_\infty \le \|Q^{-1}\|_1 MK\epsilon / c_0 \tag{27}$$

where $\|\cdot\|_1$ is the 1-norm of a matrix (or maximum absolute row sum). □

A.8 Theorem 3

Proof. Note that according to Lemma 1 in Tian (2019) (Appendix B.1), for any teacher f^{*m} and any student f of the same depth, we have at layer $l = 1$:

$$\begin{aligned} \mathbf{g}_1(\mathbf{x}) &= D_1(\mathbf{x}) V_1^\intercal(\mathbf{x}) \left[V_1^{*m}(\mathbf{x}) \mathbf{f}_1^{*m}(\mathbf{x}) - V_1(\mathbf{x}) \mathbf{f}_1(\mathbf{x})\right] && (28) \\ &= D_1(\mathbf{x}) V_1^\intercal(\mathbf{x}) (\mathbf{y}^{*m}(\mathbf{x}) - \mathbf{y}(\mathbf{x})) && (29) \end{aligned}$$

since for two-layer network, we have $\mathbf{y}(\mathbf{x}) = V_1(\mathbf{x})\mathbf{f}_1(\mathbf{x})$ is the output. Therefore, if the gradient computed between teacher f^* and student f has $\|\mathbf{g}_1\|_\infty \le \epsilon$, then

$$\begin{aligned} \|\mathbf{g}_1^m\|_\infty &= \|D_1 V_1^\intercal (\mathbf{y}^{*m}(\mathbf{x}) - \mathbf{y}(\mathbf{x}))\|_\infty && (30) \\ &\le \|D_1 V_1^\intercal (\mathbf{y}^{*m}(\mathbf{x}) - \mathbf{y}^*(\mathbf{x}))\|_\infty + \|D_1 V_1^\intercal (\mathbf{y}^*(\mathbf{x}) - \mathbf{y}(\mathbf{x}))\|_\infty && (31) \\ &\le \|D_1 V_1^\intercal (\mathbf{y}^{*m}(\mathbf{x}) - \mathbf{y}^*(\mathbf{x}))\|_\infty + \|\mathbf{g}_1\|_\infty && (32) \\ &\le \|V_1\|_1 \epsilon_0 + \epsilon && (33) \end{aligned}$$

where $\|V_1\|_1 = \max_j \|\mathbf{v}_j\|_1$ is the 1-norm (or the maximum absolute row sum) of matrix V_1. Then we apply Theorem 2 between the student f and teacher f^{*m} and the conclusion follows. □

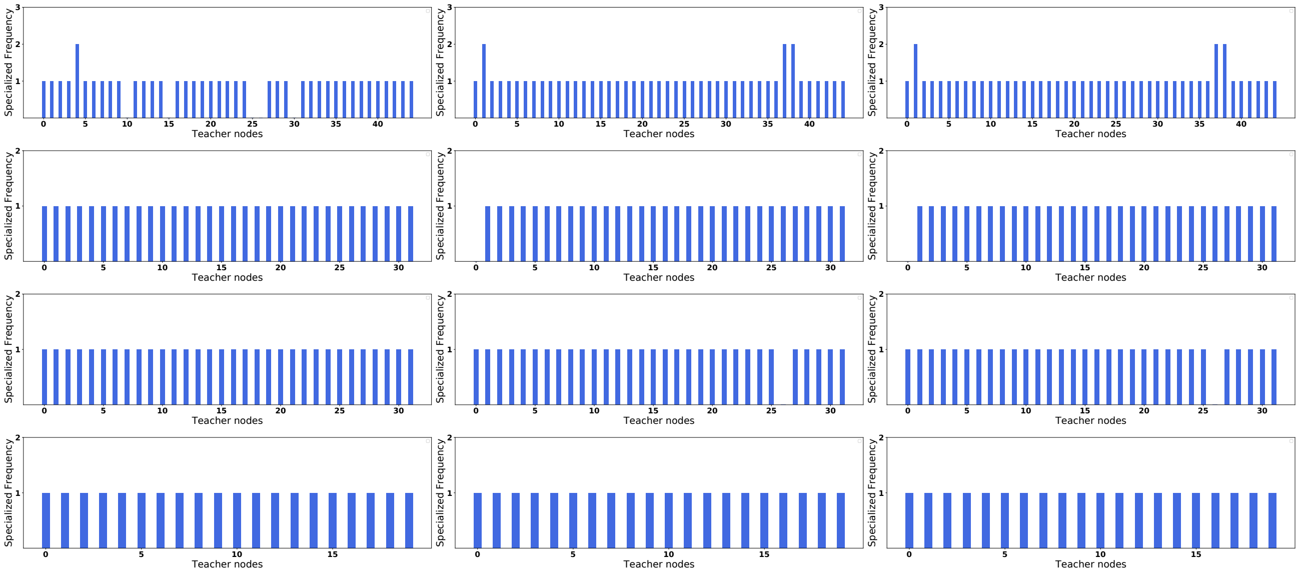

Figure 14: Specialized frequency of each teacher node by different student networks among different layers (We consider the node to be specialized if the NC is larger than 0.9). Figures in different columns refer to specialized evaluation with student network trained from different random initialization, while in different rows refer to evaluation on different layers. Here we can see 1) Teacher nodes are specialized uniformly by student nodes. 2) Different random initialization will lead to similar observations.

A.9 Unidentifiable teachers and Student Bias

We might wonder what would happen if there exist two teachers $f^{*1} \neq f^{*2}$ so that $\mathbf{y}_i = f^{*1}(\mathbf{x}_i) + \xi_i^1 = f^{*2}(\mathbf{x}_i) + \xi_i^2$ with different bias: $\|\xi_i^1\| \leq \epsilon_0$ and $\|\xi_i^2\| \leq \epsilon_0$. In this case, which teacher the student would converge into? We could use the same framework to analyze it:

Theorem 3. *For any two-layered network f^{*l} of the same architecture as f^* and $\|f^*(\mathbf{x}) - f^{*l}(\mathbf{x})\| \leq \epsilon_0$ for all $\mathbf{x} \in R$, when $\|\mathbf{g}_1\|_\infty \leq \epsilon$, for a teacher node j in f^{*l} observed by a student k, there exists a student k' so that $\sin \theta^{\mathcal{U}}_{jk'} \leq MK(\epsilon + \epsilon_0 \max_j \|\mathbf{v}_j\|_1)/\alpha^l_{jk}$.*

Note that this theorem can be applied to any teacher f^{*l} to yield a separate bound for the alignment. Some bounds are strong while others are loose. The larger α^l_{jk}, the tighter the bound. Therefore, there are two phases in the training: **(1)** at the early stage of training, ϵ is fairly large, the norm of the fan-out weights $\|\mathbf{v}_j\|_1$ is small, and many candidate teachers (as well as their hidden nodes) with reasonable ϵ_0 can stand out as long as their α^l_{jk} is large. Therefore, the student moves to salient (large α^l_{jk}) but potentially biased (large ϵ_0) explanation. **(2)** When the training converges and ϵ is small, some $\|\mathbf{v}_j\|_1$ becomes large, the "real" teacher with small bias ϵ_0 gives the tightest bound, and the student converges to it.

The case (1) is interesting since it shows that the student node doesn't go straight to the ground truth teacher node from the beginning, but has a bias towards simple models that could roughly explain data (with reasonable ϵ_0). This is a fixed bias for student nodes that only dependent on the dataset and regardless of the model initialization. This could be used to explain the adversarial transferability (Goodfellow et al., 2014). In this paper, we focus on the specialization of student nodes on a specific teacher network and leave the case of "one student multiple teachers" (i.e., Theorem 3) for future empirical study.

A.10 Ablation study on specialization distribution among teacher nodes

To investigate how well one teacher node could be specialized by student nodes and the existence of special teacher nodes which are easy to be specialized by student nodes, we conduct the ablation study by training three student networks with different random initialization and check the number of student nodes specialized to each teacher node as shown in Figure 14. We found that teacher nodes are specialized almost uniformly by different student nodes, showing that there may not be special "robust" teacher nodes, which could be an interesting finding.

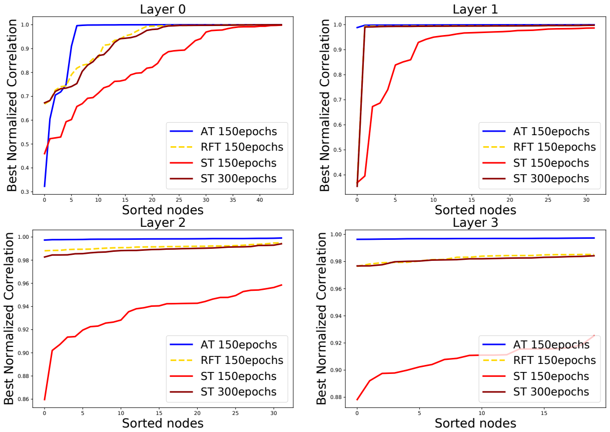

Figure 15: Sorted BNC curve of student models trained with Robust Feature Training (RFT), Standard Training (ST), and Adversarial Training (AT) trained for different epochs.

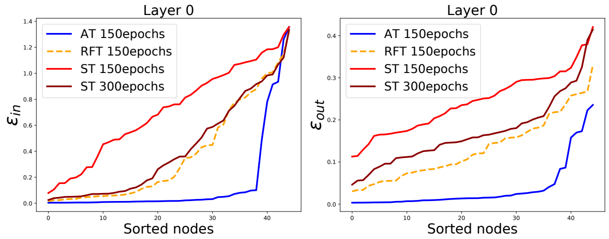

Figure 16: $(\epsilon_{\text{in}}, \epsilon_{\text{out}})$ curve of student models trained with Robust Feature Training (RFT), Standard Training (ST), and Adversarial Training (AT) trained for different epochs.

A.11 Analysis on Robust feature dataset

Robust feature disentanglement, proposed by Ilyas et al. (2019b), is a general method to generate a robust feature dataset from a robustly trained model. Specifically, the robust feature dataset $\mathcal{D} = \{x_r\}$ is generated by minimizing the feature representation distance as below:

$$x_r = \arg\min_{x_r} ||f_M(x) - f_M(x_r)||_2 \tag{34}$$

while f_M represents the representation output of model f and x is drawn from the raw dataset. For every x as the target image, the robust feature image x_r is optimized from a randomly selected image or random noise.

In teacher-student setting, we define f to be a robust student model if its prediction can be consistent with the teacher's prediction against oracle-adversarial or data-adversarial. Different from the standard setting, the generated x_r may lie in different categories with x from the teacher's perspective. In order to avoid the inconsistency, we add another term into robust feature generation's goal to minimize the logit difference between robust feature image x_r and target image x given by the teacher model:

$$x_r = \arg\min_{x_r} \alpha||L_t(x_r) - L_t(x)||_2 + ||f_M(x) - f_M(x_r)||_2 \tag{35}$$

where α is the balancing hyperparameter. We choose $\alpha = 0.5$ for the default setting.

We choose the AT model trained with 150 epochs and generate the corresponding robust feature dataset $\mathcal{D}$. Based on $\mathcal{D}$, we train the robust feature model for 150 epochs via fine-tuning on top of a 150 epochs trained ST model. In order to make a fair comparison, we compare the 150 epochs Robust Feature Training (RFT) model to $150, 300$ epochs trained ST models and 150 epochs AT model. All models are trained with the teacher's logit feedback.

Table 5: Robustness of student models trained with Robust Feature Training (RFT), Standard Training (ST), and Adversarial Training (AT) for different epochs.

Model	AT (150 epochs)	ST (150 epochs)	ST (300 epochs)	RFT (150 epochs)
Robust Accuracy	83.27%	35.88%	61.73%	45.39%

Table 5 and Figure 15 show the robustness and neuron specialization of student models with RFT, ST, and AT. We can see **1)** AT model achieves the best robustness as well as the best neuron specialization; **2)** RFT model (150 epochs) fine-tuned from ST model (150 epochs) achieves better model robustness and specialization than ST model (150 epochs); **3)** The neuron specialization of RFT model (150 epochs) and ST model (300 epochs) is close but ST model (150 epochs) achieves better robustness. Figure 16 shows the $\epsilon_{\text{in}}, \epsilon_{\text{out}}$ curve and we can see the 150 epochs RFT model shows the similar ϵ_{in} curve but slightly better ϵ_{out} curve to 300 epochs ST model. We analyze this phenomenon by considering the robust feature dataset mainly captures the out-plane vulnerability. As we discussed in Section 5.3, the in-plane vulnerability could be more severe to the model's robustness and that could be the reason why the 150 epochs RFT model achieves slightly worse robustness than the 300 epochs ST model.

Remarks. Based on the comparison between RFT, AT, ST models, we can conclude *again* that the neuron specialization of student models highly indicates their robustness. On the other hand, when the neuron specialization is close, the robustness comparison between them is less informative since other factors such as data distribution may have an impact on it. In addition, the teacher-student provides an in-depth explanation of why the robust feature dataset exists from the neuron specialization perspective. The robust feature dataset can help model capture the in-plane data projection and out-plane vulnerability therefore improve the correlation between student and teacher, which leads to better model robustness.