KAIST AI616 DLT: Final Report **Spring 2022**

Minimum-norm solution in teacher-student setting: specialization of one-hidden-layer student network

Student name(s): Minhak Song, Sangheon Lee

1 Introduction

Over-parameterized neural networks have achieved remarkable performance in many machine learning tasks, but a theoretical understanding of how this performance is achieved is still a mystery. Two key questions to explain this mystery are: (1) how gradient-based optimization finds (almost) globally optimal solution, and (2) why over-parameterized networks do not over-fit but generalizes well. In this study, we tackle the second question – while the network with zero training error can either generalize well or perform randomly in the test set, which networks generalize well?

We try to answer this question in the *teacher-student* setting: given input samples, a fixed teacher network provides the label for a student to learn. This setting is widely used for theoretical works in deep learning. It should be highlighted that the existence of a teacher network is guaranteed by uniform approximation theorems, implying that the teacher-student setting is indeed not specific but general. Moreover, this setting makes the characterization of the training procedure easier compared to the arbitrary data distributions, which motivates our study.

This project analyzes the one-hidden-layer neural networks in a teacher-student setting. We study the interesting characteristic of training dynamics called *specialization*. Specialization refers to student neurons becoming increasingly correlated with teacher neurons during training [Saad and Solla, 1995], which leads the trained student to be identical to the teacher. This phenomenon was empirically observed in previous works – e.g., [Saad and Solla, 1995], [Goldt et al., 2019] – but the theoretical understanding is still limited by strong assumptions.

We show that in an over-parameterized setting, the *minimum-norm interpolated solution* is specialized, and every student neuron is aligned with one of the teacher neurons, resulting in the student representing the identical function as the teacher. We assume mild conditions on the training input's size and distribution. The minimum-norm interpolated solution refers to the student network with the smallest norm among the collection of student networks with zero training loss. Our result shows that the minimum-norm solution specializes among the infinitely many zero training loss solutions in an over-parameterized setting, resulting in low test loss. Indeed, the minimum-norm interpolated solution can be approximately obtained by minimizing the regularized empirical risk. We empirically observe that SGD with regularization results in a specialized solution that generalizes well, as we expected.

2 Related Works

The seminal work [Saad and Solla, 1995] analyzed student-teacher setting in one-hidden-layer case where the input dimension goes to infinity through statistical mechanics viewpoint. They empirically studied the specialization of the student neurons toward the teacher model.

Zhong et al. [2017] proved that the strong convexity holds in the neighborhood of the ground-truth parameters where the teacher and student are one-hidden-layer neural networks with the same width. However, it should be noted that the over-parameterized setting is not included in this study. Safran et al. [2021] further studied the effect of over-parameterization and proved that over-parameterization changes the non-global minima into the saddle points, demonstrating how mild over-parameterization helps eliminate spurious local minima.

The most relevant work is arguably Akiyama and Suzuki [2021], which showed that in sufficiently over-parameterized setting, the global minimizer of the regularized empirical risk is close to the teacher network. They also studied the global convergence of the special gradient method with norm-dependent step size. However, they assumed that the teacher network consists of the orthogonal neurons and required sufficiently large training set. Note that number the previous works, e.g., [Zhong et al., 2017], [Tian, 2017], [Safran and Shamir, 2018], [Safran et al., 2021], [Li et al., 2020], considered the orthogonality condition on the teacher neurons, which is quite strong, while our result does not assume such condition.

3 Problem Settings

In this section, we give the problem setting and the model that we consider in this project. We consider the regression task with training samples $S_n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ where the label is generated by the *teacher* model:

$$y_i = f^*_{m_0}(x_i), \tag{1}$$

where $f^*_{m_0}$ is the one-hidden-layer neural network with width m_0:

$$f^*_{m_0}(x; \theta^*) = \sum_{j=1}^{m_0} a_j^* \sigma(\langle b_j^*, x\rangle), \tag{2}$$

with fixed parameter $\theta^* = ((a_1^*, b_1^*), \ldots, (a_{m_0}^*, b_{m_0}^*)) \in (\mathbb{R}\times\mathbb{R}^d)^{m_0}$ and ReLU activation $\sigma(x) = \max\{x, 0\}$. We assume that the training inputs $(x_i)_{i=1}^n$ are independently and identically distributed from the uniform distribution on the unit ball $\mathbb{S}^{d-1}$.

Given the training data S_n, we consider training the *student* model f_m to estimate the teacher network by minimizing the empirical risk:

$$R(f_m) = \frac{1}{2n}\sum_{i=1}^{n}(f_m(x_i) - f^*_{m_0}(x_i))^2, \tag{3}$$

where f_m is the one-hidden-layer neural network with width m:

$$f_m(x; \theta) = \sum_{j=1}^{m} a_j \sigma(\langle b_j, x\rangle), . \tag{4}$$

with trainable parameter $\theta = ((a_1, b_1), \ldots, (a_m, b_m)) \in (\mathbb{R} \times \mathbb{R}^d)^m$.

We consider the over-parameterized setting in the sense that there are infinitely many student networks that represents the exact same function to the given teacher network, by considering the case $m_0 < m$. While there are multiple solutions with zero training loss, we analyze the following minimum l_2 norm interpolated solution:

$$\hat{\theta} \in \underset{\theta\in(\mathbb{R}\times\mathbb{R}^d)^m}{\arg\min} \|\theta\|_2 \quad \text{s.t. } f_m(x_i; \theta) = f^*_{m_0}(x; \theta^*), \quad i = 1, \ldots, n \tag{5}$$

We assume some mild conditions to make the analysis simple. For the teacher network, we use random weights normalized so that $\|a_j^* b_j^*\|_2 = 1$ for $j = 1, \ldots, m_0$. Furthermore, we assume that the training inputs are uniformly sampled on the unit sphere $\mathbb{S}^{d-1}$.

4 Warm up: minimum-norm solution with zero population risk

In this section, we consider an ideal case where the training set is infinitely large, so that the empirical risk can be considered as the population risk. We prove that the minimum-norm solution with zero population risk satisfies interesting property that the neuron of student network tends to align with the teacher network neurons. While specialization is a commonly used term for this phenomenon, we call this property as *condensation* in this paper, which is also widely used.

Definition 4.1. *Consider student network $f_m(\cdot; \theta)$. We call the network $f_m(\cdot; \theta)$ is* condensed *if the parameter can be grouped as $\theta = \bigcup_{k=1}^{m_0}\{(a_j^{(k)}, b_j^{(k)})\}_{j=1}^{m_k}$ with $\sum_{k=1}^{m_0} m_k = m$, so that $b_j^{(k)}/\|b_j^{(k)}\| \in \{\pm b_k^*/\|b_k^*\|\}$ for $k = 1, \ldots, m_0$, and $j = 1, \ldots, m_k$.*

Note that there are infinitely many student networks that has zero population risk, and among them, we prove that the minimum-norm solution is indeed condensed.

Proposition 4.2. *Let $\mathcal{F}$ be the collection of student networks with zero population risk, i.e.,*

$$\mathcal{F} = \{f_m(\cdot; \theta) : f_m(x; \theta) = f^*_{m_0}(x; \theta^*), \forall x \in \mathbb{S}^{d-1}\}. \tag{6}$$

Consider the minimum l_2 norm student network with zero population risk $f_m(\cdot; \hat{\theta}) \in \arg\min_{f_m(\cdot;\theta)\in\mathcal{F}}\|\theta\|_2$. Then,

1. *The network $f_m(\cdot;\hat{\theta})$ is condensed, and*
2. $\sum_{j=1}^{m_k}\|a_j^{(k)}b_j^{(k)}\|_2 = \|a_k^*b_k^*\|_2 = 1$ *for* $k = 1,\dots,m_0$.

Proof Strategy: Gradient Difference Proposition 4.2 can be shown through characterizing the difference of gradients between neighboring sectors. For simplicity, we consider the case $d = 2$. For a given function $g : \mathbb{S}^1 \to \mathbb{R}$, we define the set

$$\mathcal{A}(g) := \{\theta : 0 \le \theta < 2\pi, g \text{ has discontinuity at } x = (\cos\theta, \sin\theta)\}. \tag{7}$$

For a given two-layer ReLU network $f : \mathbb{S}^1 \to \mathbb{R}$, $\mathcal{A}(\nabla f)$ is a finite set. Note that here we consider the gradient with respect to the data input, which is different from the gradient with respect to the parameter used in SGD. Now we define a sum of the size of the jumps at jump discontinuities of ∇f as

$$\kappa(f) := \sum_{\theta\in\mathcal{A}(\nabla f)} \left\| \lim_{\alpha\to\theta+} \nabla f((\cos\alpha,\sin\alpha)) - \lim_{\alpha\to\theta-} \nabla f((\cos\alpha,\sin\alpha)) \right\|_2. \tag{8}$$

Remark 4.3. *A few remarks on the function κ are in order.*

1. *Consider a one-hidden-neuron ReLU network $f(x;\theta) = a\cdot\sigma(\langle b, x\rangle)$. Then $\kappa(f) = 2\|ab\|_2$.*
2. *Consider two-layer ReLU networks f_1, f_2 defined on $\mathbb{S}^1$. Then $\kappa(f_1+f_2) \le \kappa(f_1)+\kappa(f_2)$.*
3. *Consider a two-layer ReLU network $f_m(x;\theta) = \sum_{j=1}^m a_j\sigma(\langle b_j, x\rangle)$. Then $\kappa(f_m) \le 2\sum_{j=1}^m\|a_jb_j\|_2$.*

Proof: We can check that the second remark is straightforward from the definition. For the first remark, consider the polar coordinate $b = (r\cos\phi, r\sin\phi) \in \mathbb{R}^2$. Then, $\mathcal{A}(\nabla f) = \{\phi+\pi/2, \phi-\pi/2\}$ modulo 2π. Note that $\|\nabla f\|_2$ has value $\|ab\|_2$ when ReLU is activated and has value 0 if ReLU is not activated. Hence, it holds $\|\lim_{\alpha\to\theta+}\nabla f(\alpha) - \lim_{\alpha\to\theta-}\nabla f(\alpha)\|_2 = \|ab\|_2$ for $\theta = \phi+\pi/2, \phi-\pi/2$. The third remark directly follows from the first and second remarks. □

Following the Remark 4.3, the teacher network $f_{m_0}^*$ has $\kappa(f_{m_0}^*) = 2\sum_{j=1}^{m_0}\|a_j^*b_j^*\|_2 = 2m_0$ and for every student network $f_m \in \mathcal{F}$, we have $\kappa(f_m) = \kappa(f_{m_0}^*) = 2m_0$. Then the student network $f_m(x;\theta) = \sum_{j=1}^m a_j\sigma(\langle b_j, x\rangle)$ in $\mathcal{F}$ satisfies $\kappa(f_m) = 2m_0 \le 2\sum_{j=1}^m\|a_jb_j\|_2$. Moreover, it holds $\mathcal{A}(\nabla f_m) = \mathcal{A}(\nabla f_{m_0}^*)$ and for each $\theta \in \mathcal{A}(\nabla f_m)$, the size of the discontinuity jump of at θ is equal for both ∇f_m and $\nabla f_{m_0}^*$. Then l_2 norm of the parameter θ is bounded by

$$\|\theta\|_2 = \sum_{j=1}^m(a_j^2 + \|b_j\|^2) \ge 2\sum_{j=1}^m\|a_jb_j\|_2 \ge 2\sum_{j=1}^{m_0}\|a_j^*b_j^*\|_2 = 2m_0. \tag{9}$$

If we set the parameters $a_j = a_j^*/|a_j^*|$ and $b_j = b_j^*/\|b_j^*\|_2$ for $j = 1,\dots,m_0$ and zero for remaining parameters, the student network is minimum-norm solution with l_2 norm $\|\theta\|_2 = 2m_0$. Therefore, for every minimum-norm solution, the equality $\sum_{j=1}^m\|a_jb_j\|_2 = \sum_{j=1}^{m_0}\|a_j^*b_j^*\|_2$ should be satisfied.

Now consider the student network $f_m(x;\theta) = \sum_{j=1}^m a_j\sigma(\langle b_j, x\rangle)$ in $\mathcal{F}$ where there exists a second-layer weight b_j distinct from the second-layer weights of the teacher networks, i.e., $b_j/\|b_j\| \notin \bigcup_{k=1}^{m_0}\{\pm b_k^*/\|b_k^*\|\}$ for some j. Then, this neuron does not contributes to the jump discontinuity of ∇f_m at points in $\mathcal{A}(\nabla f_m)$. Hence, the equality $\sum_{j=1}^m\|a_jb_j\|_2 = 2m_0$ cannot be achieved in this case. Therefore, the minimum-norm solution should have condensed second-layer weights, i.e., the parameter can be grouped as $\hat{\theta} = \bigcup_{k=1}^{m_0}\{(a_j^{(k)}, b_j^{(k)})\}_{j=1}^{m_k}$ with $\sum_{k=1}^{m_0} m_k = m$ satisfying

$$b_j^{(k)}/\|b_j^{(k)}\| \in \{\pm b_k^*/\|b_k^*\|\} \text{ for } k = 1,\dots,m_0, \text{ and } j = 1,\dots,m_k. \tag{10}$$

Moreover, for each condensed group of parameters, to obtain the same output as the teacher network, the student network satisfies $\sum_{j=1}^{m_k}\|a_j^{(k)}b_j^{(k)}\|_2 = \|a_k^*b_k^*\|_2 = 1$ for $k = 1,\dots,m_0$ as desired.

Remark 4.4. *A few remarks on Proposition 4.2 are in order.*

1. *The condensation property in this setting does not imply anything about the generalization performance of the student model since we assume that the student network has zero test loss. However, analyzing the gradient difference is also used as a critical idea in proving the main result in Section 5.*
2. *The condensed student network can have a neuron a_jb_j aligned with the neuron of the teacher network in the exact opposite direction. We observed that these opposite-direction neurons can exist if and only if the vector sum of such neurons is exactly zero. This can be considered an independently interesting result, but we will not discuss it in detail in this paper.*

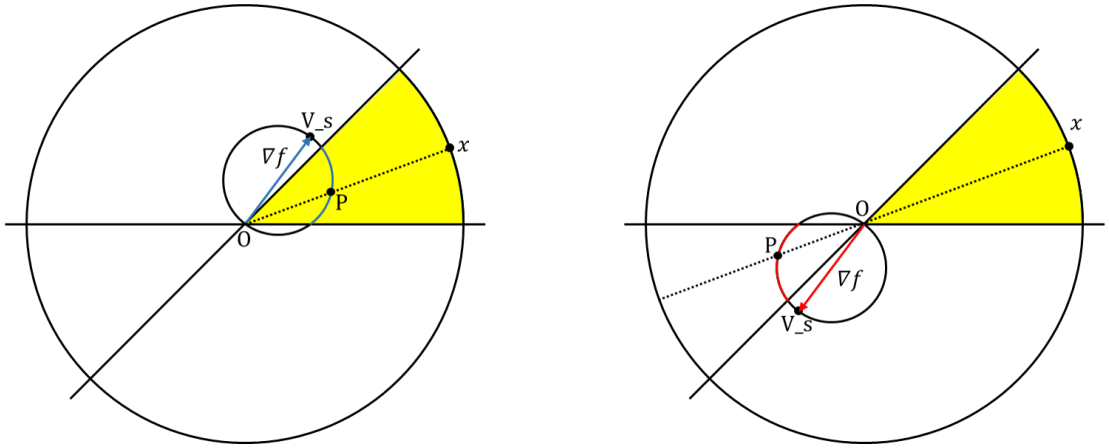

Figure 1: Looking at the right picture, given a sector (yellow region), the gradient of the network is constant (blue arrow). Then, the function value can be drawn as blue arc in the figure. Given a point x on the circle, $f(x) = |\overline{OP}|$. In the case of right picture, where gradient is on the opposite direction to the sector, $f(x) = -|\overline{OP}|$,

5 Main result: minimum-norm solution with zero empirical risk

In this section we extend Proposition 4.2 to zero training loss solutions, and prove that the student neurons align with the teacher neurons given some mild conditions on the training input distribution.

Definition 5.1. *Let $f^*_{m_0}(x;\theta^*) = \sum_{j=1}^{m_0} a_j^* \sigma(\langle b_j^*, x\rangle)$ be the teacher neural network. For $1 \le j \le m_0$, let l_j be the line passing through the origin that is perpendicular to b_j. Then l_j's divide $\mathbb{R}^2$ by $2m_0$ regions. We define* sector *to designate each region divided by l_j's.*

Note that the gradient of $f^*_{m_0}(x;\theta^*)$ with respect to the input x is constant in each sector. The following lemma shows the condensation of the student network when there are more than two training inputs in each sector, assuming the input dimension $d = 2$.

Theorem 5.2. *let $X = \{x_i : 1 \le i \le N\} \subset \mathbb{S}^1$ be the set of training inputs, where for each sector made by $f^*_{m_0}$, there are at least three $x \in X$ such that the sector contains x. Let $\mathcal{F}$ be the collection of student networks with zero training loss, i.e.,*

$$\mathcal{F} = \{f_m(\cdot;\theta) : f_m(x;\theta) = f^*_{m_0}(x;\theta^*), \forall x \in X\}. \tag{11}$$

Consider the minimum l_2 norm student network with zero training loss $f_m(\cdot;\hat{\theta}) \in \arg\min_{f_m(\cdot;\theta)\in\mathcal{F}} \|\theta\|_2$. Then,

1. *The network $f_m(\cdot;\hat{\theta})$ is condensed, and*
2. *$\sum_{j=1}^{m_k} \|a_j^{(k)} b_j^{(k)}\|_2 = \|a_k^* b_k^*\|_2 = 1$ for $k = 1, \dots, m_0$.*

Proof: According to the proof of Proposition 4.2, it suffices to to prove that $\kappa(f) > 2m_0$ for any non-condensed solution f.

For each sector of the network f, the gradient (with respect to input data) is constant over it. Then, the function value at point x can be drawn as figure 1. For sector s, let's call $V(f,s)$ to be point at ∇f, and call $S(f)$ to be set of all sectors of f. Then, $\kappa(f)$ is a sum of the length of all lines connecting $V(f,s)$ and $V(f,s')$ where $s, s' \in S(f)$ are adjacent. $\kappa(f)$ can be interpreted as the perimeter of the polygon constructed by connecting $V(f,s)$ of adjacent sectors.

If f is not condensed, there exists a sector s^* of the parent network $f^*_{m_0}$ such that the gradient of f is distinct from the gradient of $f^*_{m_0}$ on s^*. Let P_1, P_2, P_3 be training points on s^*, and let $s_1, \dots, s_k$ be the sectors of f that has intersection with s^* in clockwise order.

If $V(f,s_l) \neq V(f^*_{m_0}, s^*)$ for all $1 \le l \le k$, then all points P_1, P_2, P_3 belong to different sectors. (If two points are in the same sector, gradient of the that sector is same with the gradient of s^*) So, $k \ge 3$. For simplicity, assume s^* is divided into three parts in student network, each containing one of P_i. Let s_i be sector of f containing point P_i, and $V_i = V(f,s_i)$. Since each circle corresponding to each sector should

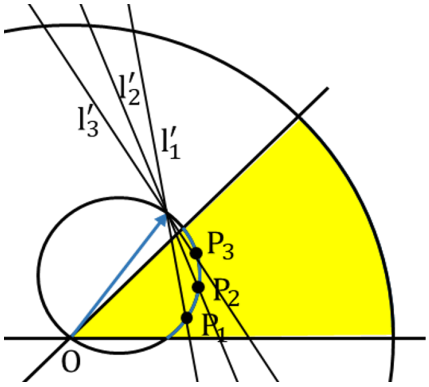

Figure 2: Pictoral illustration that each V_i should lie on l_i'

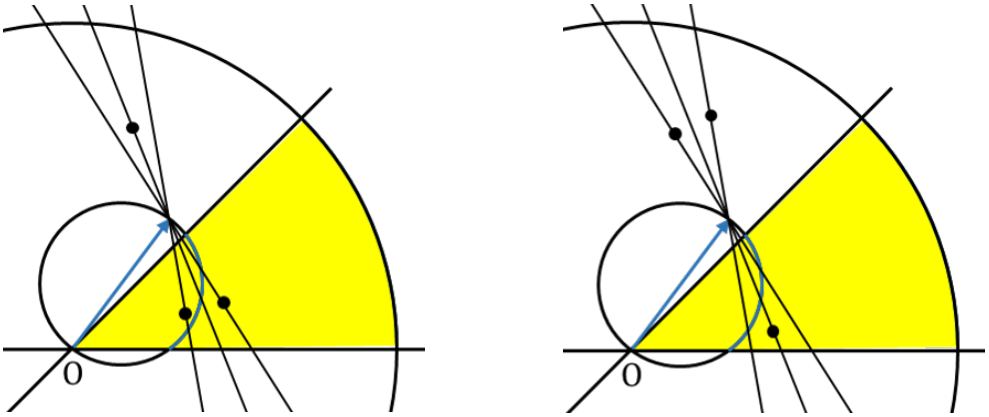

Figure 3: Two possibilities of alignment of V_i's. the black points are showing example location of V_i's.

contain O and P_i, each V_i lie on the line l_i' of Figure 2. Since P_1 and P_2 belong to different sectors, the slope of the direction of the change of the gradient is between the slope of l_1' and the slope l_2'. Hence, V_1 and V_2 should be located at the opposite side with respect to the line $\overline{OV(f_{m_0}^*, s^*)}$. So there is only two possibility of alignment of V_i's, which is drawn in Figure 3. In both of the case, the triangle with vertexes V_1, V_2, V_3 have $V(f_{m_0}^*, s^*)$ inside. The similar argument can be applied when s^* is divided into more then three parts.

If there exists l such that $V(f, s_l) = V(f_{m_0}^*, s^*)$, there is $l' \neq l$ such that $V(f, s_{l'}) \neq V(f_{m_0}^*, s^*)$. As a result, polygon made by connecting $V(f, s)$ of adjacent sectors has bigger perimeter than the polygon made by connecting $V(f_{m_0}^*, s^*)$ of adjacent sectors. So $\kappa(f)$ is larger than $2m_0$. If $\kappa(f) > 2m_0$, f can't be minimum norm solution. Therefore, $f_m(\cdot; \hat{\theta})$ is condensed.

□

Remark 5.3. *Theorem 5.2 implies that given the mild conditions on the training inputs, the minimum-norm solution with zero training loss has zero test loss, i.e., the student network represents exactly the same function as the teacher network. This result is surprising since given only finite number of training samples, the generalization error is guaranteed to be exactly zero.*

6 Experiments

In this section, we present the empirical findings given the same teacher-student setting where we minimize the regularized empirical risk. We observed that SGD with regularization can approximate the minimum-norm interpolated solution. We set up an experiment to check whether the minimum-norm solution found by SGD with regularization condenses. The teacher network used in the experiment has width $m_0 = 3$, where b_i's have $\frac{2}{3}\pi$ degree pairwise. The student network used in the experiment has a width $m = 100$. With this fixed parent network, we can get a lower bound of the probability that the student network condenses. The probability that the student network condenses is larger than the

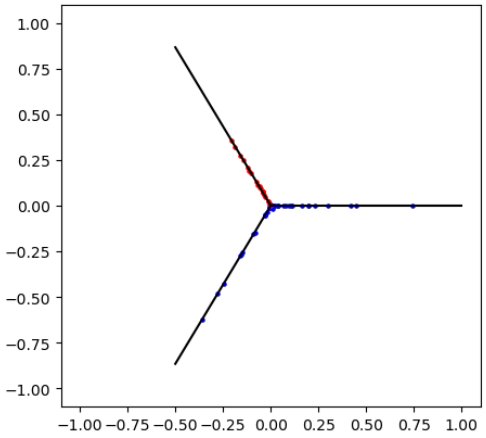

Figure 4: Experiment result: $|a|b_i$'s after SGD with regularization. The color is red if $a > 0$ and blue otherwise. The black lines represents $|a|b_i$'s of the teacher network.

probability that there are at least three training data in each sector, which is lower bounded by

$$1 - \sum_{i=1}^{2m_0} \left((1-p_i)^N + Np_i(1-p_i)^{N-1} + \frac{N(N-1)}{2} \cdot p_i^2(1-p_i)^{N-2} \right) \tag{12}$$

where N is the number of training data, and p_i is probability that the random data point is placed inside sector i. In this experiment setting, $m_0 = 3, |S(f^*_{m_0})| = 6$, and $p_i = \frac{1}{6}$. With $N = 40$, probability lower bound is 0.83. When we run an experiment, as one can observe in in Figure 4, the resulting student network condenses, and squared L2 norm of the trained weight was 5.99, which is close to $6 = 2m_0$, the theoretical minimum that student network with zero population risk can have. (An error might have occurred since objective with regularization is biased.) We repeated the same experiment, and observed that the network condensed in all cases, and the squared L2 norm of the trained weight was always close to 6.

Furthermore, we observed the interesting phenomenon when we conduct the experiment on even smaller N. When $N = 20$, the probability lower bound obtained by Equation 12 is zero, but the network successfully condenses on almost all of the cases. This suggests that our sufficient condition, that at least three train data should exist in each sector, can be further improved.

7 Conclusion

In this project, we have studied the generalization aspects of the minimum-norm interpolated solution for one-hidden-layer ReLU neural networks in teacher-student settings. Surprisingly, we have shown that if each sector given by the teacher network contains more than two training inputs, the minimum l_2 norm student network with zero training loss represents the same function as the teacher, i.e., the test loss is zero. Furthermore, we have empirically observed that SGD with regularization can approximate the minimum-norm solution, showing that the student network's neurons align with those of the teacher network, resulting in good generalization performance.

Our result is novel in the sense that we considered arbitrary given teacher neurons, while most works on the teacher-student setting assumed that the neurons of the teacher network are orthogonal to each other. We restricted our analysis to 2-dimensional training input, but we conjecture that our results can be extended to general input dimensions, which we leave as future work. Moreover, further analysis of how gradient methods with regularization can approximate the minimum-norm interpolated solution will be an interesting topic. We believe that our study gives a new insight into how minimizing the norm contributes to the learnability in the teacher-student setting.

References

S. Akiyama and T. Suzuki. On learnability via gradient method for two-layer relu neural networks in teacher-student setting. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 152–162. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/akiyama21a.html.

S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/cab070d53bd0d200746fb852a922064a-Paper.pdf.

Y. Li, T. Ma, and H. R. Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2613–2682. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/li20a.html.

D. Saad and S. Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper/1995/file/a1519de5b5d44b31a01de013b9b51a80-Paper.pdf.

I. Safran and O. Shamir. Spurious local minima are common in two-layer ReLU neural networks. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4433–4441. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/safran18a.html.

I. M. Safran, G. Yehudai, and O. Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. In M. Belkin and S. Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3889–3934. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/safran21a.html.

Y. Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3404–3413. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/tian17a.html.

K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4140–4149. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/zhong17a.html.