GENERALIZABILITY OF NEURAL NETWORKS MINIMIZING EMPIRICAL RISK BASED ON EXPRESSIVE ABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The primary objective of learning methods is generalization. Classic uniform generalization bounds, which rely on VC-dimension or Rademacher complexity, are applicable to all neural networks in the hypothesis space. On the other hand, algorithm-dependent generalization bounds, like stability bounds, address more practical scenarios and provide generalization conditions for neural networks trained using SGD. However, these bounds often rely on strict assumptions, such as the NTK hypothesis or convexity of the empirical loss, which are typically not satisfied by neural networks. Furthermore, these generalization bounds also fail to explain the significant attribute that over-parameterized models in deep learning exhibit nice generalizability. To establish generalizability under less stringent assumptions, which can also account for the effective generalizability of over-parameterized models, this paper investigates the generalizability of neural networks that minimize empirical risk. We establish a lower bound for population accuracy based on the expressiveness of these networks, which indicates that with an adequate large number of training samples and network sizes, these networks, including over-parameterized ones, can generalize effectively. Additionally, we provide a necessary condition for generalization, demonstrating that, for certain data distributions, the quantity of training data required to ensure generalization exceeds the network size needed to represent the corresponding data distribution. Finally, we provide theoretical insights into several phenomena in deep learning, including robust generalization, importance of over-parameterization networks, and effects of loss functions.

1 INTRODUCTION

Understanding the mechanisms behind the nice generalization ability of deep neural networks remains a fundamental challenge problem in deep learning theory. By generalization, it means that neural networks trained on finite data give high predict accuracy on unseen data. The generalization bound serves as a critical theoretical framework for evaluating the generalizability of learning algorithms. Let $\mathcal{F}$ be a network, $\mathcal{D}$ the data distribution, and $L(\mathcal{F}(x), y) = \mathbb{I}(\widehat{\mathcal{F}}(x) = y)$ where $\widehat{\mathcal{F}}(x)$ is the classification result of $\mathcal{F}(x)$. For a hypothesis space $\mathbf{H}$ and any $\mathcal{F} \in \mathbf{H}$, with probability $1 - \delta$ of dataset $\mathcal{D}_{tr}$ selected i.i.d. from $\mathcal{D}$, we have the classic generalization bound (Mohri et al., 2018)

$$|\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\mathcal{F}(x), y)] - \mathbb{E}_{(x,y)\in\mathcal{D}_{tr}}[L(\mathcal{F}(x), y)]| < \sqrt{(8d \ln \frac{2eN}{d} + 8 \ln \frac{4}{\delta})/N} \tag{1}$$

where d is the VC-dimension of $\mathbf{H}$ and $N = |\mathcal{D}_{tr}|$. There exist similar generalization bounds using Radermacher Complexity (Mohri et al., 2018).

The above-mentioned generalization bounds are valid for all networks in the hypothesis space. In practice, the generalizability of the networks trained by SGD is desirable. For that purpose, algorithmic-dependent generalization bounds are derived. It is shown that if the data satisfy the NTK condition, two-layer networks have a small generalization risk after training (Jacot et al., 2018; Ji & Telgarsky, 2019). Stability generalization bounds are also obtained by assuming the convexity and Lipschitz properties of the loss function (Hardt et al., 2016; Kuzborskij & Lampert, 2018), and more results of this type will be discussed in Section 2.

Unfortunately, uniform generalization bounds fail to explain the important phenomenon that over-parameterized models exhibit nice generalizability (Belkin et al., 2019), as pointed out by Nagarajan & Kolter (2019). For example, the VC-dimension is equal to the product of the number of parameters and the depth for ReLU networks (Bartlett et al., 2021), which renders the bound in equation 1 useless for over-parameterized models. Most of the algorithmic-dependent generalization bounds make strong and unrealistic assumptions about the training procedure. For example, the NTK condition is used to reduce the training to a convex optimization (Ji & Telgarsky, 2019) and the strong smoothness and convexity of the empirical loss are used to measure the effect in each training epoch (Hardt et al., 2016).

In order to give generalization conditions under more relaxed assumptions, and to provide more specific conditions for generalization to account for the nice generalizability of over-parameterized models, we will study the generalization of networks that minimize the empirical risk, that is, the networks $\mathcal{F} \in \mathbf{M} = \arg\min_{\mathcal{G}\in\mathbf{H}} \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{G}(x), y)$. The approach is reasonable because most practical training will lead to a very small empirical risk and the trained networks can be considered to minimize the empirical risk. So our main research objective is: **the generalization of the networks that minimize the empirical risk without requiring strong assumptions.**

In this paper, we consider two-layer networks, like many previous works (Ba et al., 2020; Luo & Yang, 2020; Ji & Telgarsky, 2019; Zeng & Lam, 2022). From the perspective of expressive ability, we show that when the number of training data and the size of the network are large enough, the network has generalizability. We further obtain that the sample complexity and the size of the network depend only on the cost required for the network to express such a distribution. As shown below.

Theorem 1.1 (Informal, Corollary 4.4)**.** *Let distribution $\mathcal{D}$ satisfy the condition that a two-layer network with width W_0 can reach accuracy 1 over $\mathcal{D}$. Then with high probability of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, if $N \geq \Omega(W_0^2)$ and width$(\mathcal{F}) \geq \Omega(W_0)$ for $\mathcal{F} \in \mathbf{M}$, then $\mathcal{F}$ has high population accuracy.*

From this result, we can determine the amount of training data and the size of the network that can ensure generalizability. Because the requirements for N and width$(\mathcal{F})$ are independent, our bounds can be used to explain the nice generalizability of over-parameterized models (Belkin et al., 2019).

We also give a lower bound for the sample complexity. For some data distribution, to ensure the generalizability of network which minimizes the empirical risk, the required number of data must be greater than the size of neural networks required to express such a distribution. As shown below.

Theorem 1.2 (Informal, Section 5)**.** *For some data distribution, if the width required for a two-layer network with* ReLU *activation function to express such distribution is at least W_0, then for a dataset with fewer than $O(W_0)$ elements, the network that minimizes the empirical risk for such dataset may have poor generalization.*

Finally, while networks that minimize the empirical risk exhibit good generalization, numerous classical experimental results indicate that these networks encounter several problems such as robustness, and so on. Therefore we provide some interpretability for these problems based on our theoretical results. Let $\mathcal{D}_{tr}$ be a dataset and $\mathcal{F} \in \mathbf{M}$. Then, three phenomena of deep learning are discussed with our theoretical results.

Robustness Generalization. (Section 6.1) It is known that robust memorization for a dataset $\mathcal{D}_{tr}$ is more difficult than memorization for $\mathcal{D}_{tr}$ (Park et al., 2021; Li et al., 2022; Yu et al., 2024a). We further show that when robust memorization of $\mathcal{D}_{tr}$ is much more difficult than memorization of $\mathcal{D}_{tr}$, then the robustness accuracy of $\mathcal{F}$ over $\mathcal{D}$ has an upper bound which may be low, or $\mathcal{F}$ has no robustness generalization over $\mathcal{D}$.

Importance of over-parameterization. (Section 6.2) It is recognized that over-parameterized networks have nice generalizaility (Belkin et al., 2019; Bartlett et al., 2021). In this regard, we show that when the network is large enough, a small empirical loss leads to high test accuracy. In contrast, when the network $\mathcal{F}$ is not large enough, there exist networks that achieve good generalization but cannot be found by minimizing the empirical risk.

Loss function. (Section 6.3) We show that for some loss function, generalization may not be achieved. If the loss function reached its minimum value or is a strictly decreasing concave function, $\mathcal{F}$ may have poor generalization.

2 Related Work

Generalization bound. Generalization bound is the central issue of learning theory and has been studied extensively (Valle-Pérez & A. Louis, 2022).

The algorithm-independent generalization bounds usually depend on the VC-dimension or the Rademacher complexity (Mohri et al., 2018). In (Harvey et al., 2017; Bartlett et al., 2019; Yang et al., 2023), the VC-dimension has been accurately calculated in terms of width, depth, and number of parameters. In (Wei & Ma, 2019; Arora et al., 2018; Li et al., 2018), some tighter generalization bound of networks was given based on Radermacher complexity. Generalization bounds were also studied for netwoks with special structures: Long & Sedghi (2019); Ledent et al. (2021); Li et al. (2018) gave the generalization bound of CNN, Vardi et al. (2022) gave the sample complexity of small networks, Brutzkus & Globerson (2021) studied the generalization bound of maxpooling networks, Trauger & Tewari (2024); Li et al. (2023) gave the generalization bound of transformers, and Ma et al. (2018); Luo & Yang (2020); Ba et al. (2020) studied the two-layer networks. Under some assumptions for the networks, Neyshabur et al. (2017); Barron & Klusowski (2018); Dziugaite & Roy (2017); Bartlett et al. (2017); Valle-Pérez & A. Louis (2022) gave the upper bounds of the generalization error. Generalization bounds based on information theory and Bayesian theory were also given (Alquier et al., 2024; Hellström et al., 2023). Nagarajan & Kolter (2019) shown that the uniform generalization bound cannot explain the generalizability for dep learning.

Algorithm-dependent generalization bounds were established in the algorithmic stability setting (Bousquet & Elisseeff, 2002; Elisseeff et al., 2005; Shalev-Shwartz et al., 2010). Under some assumptions on the empirical risk, Hardt et al. (2016); Wang & Ma (2022); Kuzborskij & Lampert (2018); Lei (2023); Bassily et al. (2020) gave the stability bounds under SGD. For small networks such as two-layer networks, Ji & Telgarsky (2019); Taheri & Thrampoulidis (2024); Li et al. (2020) proved the generalization of networks under some assumptions. Farnia & Ozdaglar (2021); Xing et al. (2021); Xiao et al. (2022); Wang et al. (2024); Allen-Zhu & Li (2022) gave stability generalization bounds for adversarial training under SGD. Regatti et al. (2019); Sun et al. (2023) gave stability generalization bounds under Asynchronous SGD. However, these algorithmic-dependent generalization bounds always impose strong assumptions on the training process or dataset. Generalization bounds for memorization networks were given in Yu et al. (2024b). However, minimizing empirical risk for cross-entropy loss does not necessarily lead to memorization, so our assumption is weaker than memorization.

Neural Network Interpretability. Interpretability is dedicated to providing reasonable explanations for phenomena that occur in neural networks. It was pointed in (Zhang et al., 2021) that interpretability is not always needed, but it is important for some prediction systems that are required to be highly reliable. For adversarial samples, it was shown that for certain data distributions and networks, there must be a trade-off between accuracy and adversarial accuracy (Shafahi et al., 2019; Bastounis et al., 2021). In (Yu et al., 2023), it was proven that a small perturbation of the network parameters will lead to low robustness. In (Allen-Zhu & Li, 2022), it was shown that the generation of adversarial samples after training is due to dense mixtures in the hidden weights. In (Yu et al., 2024a; Li et al., 2022), it was shown that ensuring generalization requires more parameters. For overfitting, long-term training has been shown to lead to a decrease in generalization (Xiao et al., 2022; Xing et al., 2021). In (Roelofs et al., 2019), comprehensive analysis of overfitting was given. In (Belkin et al., 2019; Bartlett et al., 2021), the importance of over-parameterized interpolation networks is mentioned, and in Arora et al. (2019); Cao & Gu (2019); Ji & Telgarsky (2020), the training and generalization of DNNs in the over-parameterized regime were studied. In this paper, we explain these phenomena from the perspective of the expressive ability of networks.

3 Notation

In this paper, for any $A \in \mathbb{R}$, $O(A)$ means a real number no more than cA for some $c > 0$ and $\Omega(A)$ means a real number not less than cA for some $c > 0$.

3.1 NEURAL NETWORK

In this paper, we consider two-layer neural networks $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}$ that can be written as:

$$F(x) = \sum_{i=1}^{W} a_i \sigma(W_i x + b_i) + c,$$

where σ is the activation function, $W_i \in \mathbb{R}^{1\times n}$ is the transition matrix, $b_i \in \mathbb{R}$ is the bias part, W is the width of the network, and $a_i, c \in \mathbb{R}$. Denote $\mathbf{H}_W^\sigma(n)$ as the set of all two-layer neural networks with input dimension n, width W, activation function σ, and all parameters are in $[-1, 1]$. To simplify the notation, we denote $\mathbf{H}_W^{\mathrm{ReLU}}(n)$ by $\mathbf{H}_W(n)$ when using the ReLU activation function.

3.2 DATA DISTRIBUTION

In this paper, we consider binary classification problems. To avoid extreme cases, we focus primarily on the distribution defined below.

Definition 3.1. For $n \in \mathbb{Z}_+$, $\mathcal{D}(n)$ is the set of distributions $\mathcal{D} \in [0, 1]^n \times \{-1, 1\}$ that have a *positive separation bound:* $\inf_{(x_1,y_1),(x_2,y_2)\sim\mathcal{D} \text{ and } y_1\neq y_2} ||x_1 - x_2||_2 > 0$.

The accuracy of a network $\mathcal{F}$ on a distribution $\mathcal{D}$ is defined as

$$A_{\mathcal{D}}(\mathcal{F}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(\mathrm{Sgn}(\mathcal{F}(x)) = y),$$

where Sgn is the sign function. We use $\mathcal{D}_{\mathrm{tr}} \sim \mathcal{D}^N$ to mean that $\mathcal{D}_{\mathrm{tr}}$ is a dataset of N samples drawn i.i.d. according to $\mathcal{D}$.

3.3 MINIMUM EMPIRICAL RISK

Consider the loss function $L(\mathcal{F}(x), y) = \ln(1 + e^{-\mathcal{F}(x)y})$, which is the cross-entropy loss for binary classification problems. For a dataset $\mathcal{D}_{tr} \subset [0, 1]^n \times \{-1, 1\}$ and a hypothesis space $\mathbf{H}_W^\sigma(n)$. To learn the features of the data in D_{tr}, a traditional method is empirical risk minimization (ERM), which minimizes empirical risk on the training dataset $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y)$ of the network $\mathcal{F}$.

Driven by this motivation, in this paper, we mainly consider networks $\mathcal{F} \in \mathbf{H}_W^\sigma(n)$ that can minimize empirical risk, that is, networks in

$$\mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n) = \arg\min_{\mathcal{G}\in\mathbf{H}_W^\sigma(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{G}(x), y). \tag{2}$$

It should be noted that such networks exist in most cases, as shown below.

Proposition 3.2. *Let $\mathcal{D}_{tr} \subset [0, 1]^n \times \{-1, 1\}$ and σ be a continuous function. Then for any $W \in \mathbb{Z}^+$, there exists an $\mathcal{F} \in \mathbf{H}_W^\sigma(n)$ such that $\mathcal{F} \in \mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n)$.*

Proof. Consider the empirical risk as a function of network parameters. Let a_i, b_i, c, W_i be the parameters of $\mathcal{F}$. Then empirical risk $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) = \sum_{(x,y)\in\mathcal{D}_{tr}} \ln e^{-y(\sum_{i=1}^{W} a_i\sigma(W_i x+b_i)+c)}$. Since σ is a continuous function, $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y)$ is a continuous function about a_i, b_i, c and W_i, and the domain of definition of parameters is $[-1, 1]^{W_g}$, where $W_g = W(n + 2) + 1$ is the number of parameters of $\mathcal{F}$. The proposition now comes from the fact that continuous functions have reachable upper and lower bounds on a closed domain. □

4 GENERALIZABILITY BASED ON NEURAL NETWORK EXPRESSIVE ABILITY

In this section, we demonstrate that, based on the expressive ability of neural networks, the generalization of the network that minimizes the empirical risk can be estimated. Specifically, in Section 4.1, we establish the relationship between expressive ability and generalizability. In Section 4.2, we extend our conclusion to local minima. In Section 4.3, we compare our generalization bounds with existing algorithm-independent and algorithm-dependent bounds, showcasing the superiority of our bound.

4.1 A Lower Bound for Accuracy Based on the Expressive Ability

We first define the expressive ability of neural networks to classify the data distribution.

Definition 4.1. We say that a distribution $\mathcal{D}$ over $[0,1]^n \times \{-1,1\}$ can be **expressed** by $\mathbf{H}_W^\sigma$ with confidence c, if there exists an $\mathcal{F} \in \mathbf{H}_W^\sigma$ such that

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}(y\mathcal{F}(x) \geq c) = 1.$$

For any distribution $\mathcal{D} \in \mathcal{D}(n)$, we can always find some activation function σ, such that $\mathcal{D}$ can be expressed by $\mathbf{H}_W^\sigma(n)$ with confidence c for some W and c. Therefore, this definition is reasonable. For example, if $\sigma = \mathrm{ReLU}$, according to the universal approximation theorem of neural networks (Cybenko, 1989), any $\mathcal{D} \in \mathcal{D}(n)$ can be represented by a network with ReLU as activation function, as shown by the following proposition. The proof is given in Appendix A.

Proposition 4.2. *For any distribution $\mathcal{D} \in \mathcal{D}(n)$, there exist $W \in \mathbb{N}_+$ and $c > 0$ such that $\mathcal{D}$ can be expressed by $\mathbf{H}_W(n)$ with confidence c.*

With such a definition, we have the following relationship between expressive ability and generalization ability. The proof is given in Appendix B.

Theorem 4.3. *Let σ be a continuous function with Lipschitz constant L_p, $W_0 \geq 2, n \in \mathbb{N}_+, c \in \mathbb{R}_+$. If $\mathcal{D} \in \mathcal{D}(n)$ can be expressed by $\mathbf{H}_{W_0}^\sigma$ with confidence c, then for any $W \geq W_0 + 1$, $N \in \mathbb{N}_+$, $\delta \in (0,1)$, with probability at least $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, the following bound stands for any $\mathcal{F} \in \mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n)$:*

$$A_{\mathcal{D}}(\mathcal{F}) \geq 1 - O\left(\frac{W_0}{cW} + \frac{nL_p(W_0+c)\sqrt{\log(4n)}}{c\sqrt{N}} + \sqrt{\frac{\ln(2/\delta)}{N}}\right).$$

Proof Idea. There are two main steps in the proof. The first step tries to estimate the minimum value of the empirical risk, which mainly uses the assumption: $\mathcal{D}$ can be expressed by $\mathbf{H}_{W_0}^\sigma$ with confidence c. The minimum value is based on W_0, c, W. Then, use such a minimum value to estimate the performance of the network on the dataset. In the second step, we can use the result in the first step and the classic generalization bound to estimate the performance of the network across the entire distribution and get the result. The core idea of this step is that the minimum value of empirical risk does not depend on N, but the Radermacher complexity will become smaller when increasing N. Then, when N is large enough, the performance of networks in distribution and datasets is similar.

Some experimental results used to verify Theorem 4.3 are included in Appendix L. Theorem 4.3 shows that increasing N and W leads to a better test accuracy. It is reasonable that more data make better generalization and a larger network makes better generalization, which also confirms the observation about the nice generalization ability of over-parameterized networks. This differs from classical algorithm-independent generalization bounds, which lack this advantageous property. Since the values of N and W to ensure generalization are only influenced by the size required for the network to express the data distribution, we can infer the following corollary.

Corollary 4.4. *With probability $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, it holds $A_{\mathcal{D}}(\mathcal{F}) \geq 1-\epsilon$ for any $\mathcal{F} \in \mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n)$, when $W \geq \Omega(W_0/(c\epsilon))$ and $N \geq \Omega(\frac{L_p(W_0+c)n\sqrt{\log(4n)}}{c\epsilon})^2 + \Omega(\frac{\ln(2/\delta)}{\epsilon^2})$.*

The above bounds of N and W depend only on constants about expressive ability W_0, c, Lipschitz constant L_p and ϵ, δ, which shows that as long as there are enough samples and enough large network size based on the expressive ability, the neural network that minimizes empirical risk will have generalization ability.

Remark 4.5. For deep networks, we can show that if the depth and width of the network and the number of data exceed a distribution-dependent threshold, then with high probability, the network minimizing the empirical risk can ensure generalization, as demonstrated in Appendix K. However, due to the complexity of deep networks, accurately determining the required depth, width, and data volume remains a challenge.

4.2 Generalization For Local Optimal Point

In practice, it is often challenging to accurately find the network parameters that minimize the empirical risk, but instead parameters are find which locally minimize the empirical risk. In this

section, we show that for networks with such parameters, if the value of the empirical risk is small, its generalization can also be guaranteed. We define such a set of networks.

Definition 4.6. For any $q \geq 1$ and dataset $\mathcal{D}_{tr}$, we say $\mathcal{F} \in \mathbf{H}_W^\sigma(n)$ is a q-approximation of minimize empirical risk if

$$\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq q \min_{f\in\mathbf{H}_W^\sigma(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y).$$

For all q-approximation networks, we have the following result. The proof is given in Appendix C.

Proposition 4.7. *Let σ be a continuous function with Lipschitz constant L_p, $W_0 \geq 2, n \in \mathbb{N}_+, c \in \mathbb{R}_+$. If $\mathcal{D} \in \mathcal{D}(n)$ can be expressed by $\mathbf{H}_{W_0}^\sigma$ with confidence c, then for any $W \geq W_0 + 1$, $N \in \mathbb{N}_+$, $q \geq 1$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, we have*

$$A_{\mathcal{D}}(\mathcal{F}) \geq 1 - O(\frac{qW_0}{cW} + \frac{nL_p(W_0 + c)\sqrt{\log(4n)}}{c\sqrt{N}} + \sqrt{\frac{\ln(2/\delta)}{N}}),$$

for any q-approximation $\mathcal{F} \in \mathbf{H}_W^\sigma(n)$ to minimize the empirical risk.

The theorem demonstrates that if a local optimal point is a q-approximation to minimize the empirical risk, then we can obtain conclusions similar to Theorem 4.3.

4.3 Comparison with classical conclusions

In this section, we compare our generalization bounds with previous ones. Compared to algorithm-independent generalization bounds, our bound performs better when the data size is not significantly larger than the network size. Compared to algorithm-dependent generalization bounds, our bound does not require overly strong assumptions as prerequisites.

Compare with the algorithm-independent generalization bound. When the scale of the network is bounded, a general generalization bound can be calculated by the VC dimension.

Theorem 4.8 (P.217 of (Mohri et al., 2018), Informal)**.** *Let $\mathcal{D}_{\text{tr}} \sim \mathcal{D}^N$ be the training set. For the hypothesis space $\mathbf{H} = \{\text{Sgn}(\mathcal{F}(x)) \,\|\, \mathcal{F}(x) : \mathbb{R}^n \to \mathbb{R}\}$ and $\delta \in \mathbb{R}_+$, with probability at least $1 - \delta$, for any $\text{Sgn}(\mathcal{F}(x)) \in \mathbf{H}$, we have*

$$|A_{\mathcal{D}}(\text{Sgn}(\mathcal{F})) - \mathbb{E}_{(x,y)\in\mathcal{D}_{\text{tr}}}[I(\text{Sgn}(\mathcal{F}(x)) = y)]| \leq O(\sqrt{\frac{\text{VC}(\mathbf{H})+\ln(1/\delta)}{N}}) \tag{3}$$

Theorem 4.8 demonstrates the relationship between the number of data and the size of the network to ensure generalization, which points out that when the number of data is much more than the VC-dimension of the network hypothesis space, generalization can be ensured. Since the VC-dimension is generally larger than the number of parameters of the network (Bartlett et al., 2019), Theorem 4.8 means that to ensure generalization, the number of data must be greater than the number of parameters of the network, which is contradictory to the fact that over-parameterized models have nice generalizability (Belkin et al., 2019; Bartlett et al., 2021). Similar results hold for the generalization bound based on Radermacher complexity, due to the observation that the Radermacher complexity for deep networks is close to one (Zhang et al., 2017). On the other hand, our generalization bounds in Theorem 4.3 can be used to explain the fact that over-parameterized models have nice generalizability.

Compare with the algorithm-dependent generalization bound. In the study of algorithm-dependent generalization bound, some works derive generalization bounds based on gradient descent under strong assumptions not met by neural networks.

Theorem 4.9 (Ji & Telgarsky (2019))**.** *Let $\epsilon \in (0, 1)$, $\delta \in (0, 1/4)$ and distribution $\mathcal{D}$ over $[0, 1]^n$ satisfy the NTK conditions with constant γ, and λ and M defined as*

$$\lambda = \frac{\sqrt{2\ln(4n/\delta)} + \ln(4/\epsilon)}{\gamma/4}, M = \frac{4096\lambda^2}{\gamma^6},$$

If the two-layer network with width $W > M$ and training step $\eta \leq 1$, with probability $1 - 4\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$ and training initiation point, after at most $\frac{2\lambda^2}{\eta\epsilon}$ times gradient descent on $\mathcal{D}_{tr}$, for the

trained network $\mathcal{F}$, it holds

$$A_{\mathcal{D}}(\mathcal{F}) \geq 1 - 2\epsilon - 16\frac{\sqrt{2\ln(4N/\delta)} + \ln(4/\epsilon)}{\gamma^2\sqrt{N}} - 6\sqrt{\frac{\ln(2/\delta)}{N}}.$$

Theorem 4.9 requires NTK conditions for distribution. These conditions can make the training approach to convex optimization, which is an overly strong condition. Theorem 4.3 only requires that a network interpolates the positive separation distribution, and it stands for any distribution $\mathcal{D} \in \mathcal{D}(n)$ as mentioned in Proposition 4.2.

Stability bounds represent another algorithm-dependent approach to generalization bound.

Theorem 4.10 (Theorem 3.7 in Hardt et al. (2016))**.** *Assume that for every sample (x, y), $L(\mathcal{F}_\theta(x), y)$ as a function based on θ is β-smooth, convex and L-Lipschitz. Let F^* be the network obtained by training on dataset $\mathcal{D}_{tr}$ by using SGD T times and each step size $\alpha_t < 2/\beta$. Then we have*

$$\mathbb{E}_{\mathcal{D}_{tr}\sim\mathcal{D}^N,\text{SGD}}\left|\frac{1}{N}\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}^*(x), y) - E_{(x,y)\sim\mathcal{D}}[L(F^*(x), y)]\right| \leq \frac{2L^2\sum_{i=1}^{T}\alpha_i}{N}.$$

Theorem 4.10 requires convex and smooth conditions for the risk function which are not satisfied by neural networks. Also, the Lipschitz constant is directly related to the network size, which cannot explain the over-parameterization phenomenon. In Theorem 4.3, there is no such problem because the network size W is in the denominator.

5 Lower bound for sample complexity based on expressive ability

In this section, on the other hand, we consider the lower bound of data complexity necessary for generalization.

5.1 Upper bound for Accuracy without Enough Data

This section illustrates that in the worst-case scenario, the minimum number of data points needed to guarantee accuracy is constrained by the VC-dimension of the smallest hypothesis space necessary to represent a distribution. We give a definition first.

Definition 5.1. For a hypothesis space $\mathbf{H} \subset \mathbb{R}^n \to \mathbb{R}$, $\text{VC}(\mathbf{H})$ is the maximum number of data in $[0,1]^n$ that $\mathbf{H}$ can shatter. Precisely, there exist $\text{VC}(\mathbf{H})$ samples $\{x_i\}_{i=1}^{\text{VC}(\mathbf{H})} \subset [0,1]^n$, such that for any $\{y_i\}_{i=1}^{\text{VC}(\mathbf{H})} \in \{-1, 1\}$, there is an $\mathcal{F} \in \mathbf{H}$ such that $\text{Sgn}(\mathcal{F}(x_i)) = y_i$ for all $i \in [\text{VC}(\mathbf{H})]$. But there do not exist $\text{VC}(\mathbf{H}) + 1$ such samples.

We have the following theorem. The proof is given in Appendix D.

Theorem 5.2. *For any $n, W, W_0 \in \mathbb{N}_+$ and activation function σ, there is a $\mathcal{D} \in \mathcal{D}(n)$ that satisfies the following properties.*

(1) There is an $\mathcal{F} \in \mathbf{H}^\sigma_{W_0}(n)$ such that $A_{\mathcal{D}}(\mathcal{F}) = 1$;

(2) For any given $\epsilon, \delta \in (0,1)$, if $N \leq \text{VC}(\mathbf{H}^\sigma_{W_0}(n))(1 - 4\epsilon - \delta)$, then with probability $1 - \delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, we have $A_{\mathcal{D}}(\mathcal{F}) < 1 - \epsilon$ for some $\mathcal{F} \in \mathbf{M}^\sigma_W(\mathcal{D}_{tr}, n)$.

This conclusion indicates that for distributions that require networks with width W_0 to express, some of them require at least $\Omega(\text{VC}(\mathbf{H}^\sigma_{W_0}(n)))$ data to ensure generalization. It is worth mentioning that this conclusion is true for any given W in the theorem. It is easy to see that a larger W_0 makes $\text{VC}(\mathbf{H}^\sigma_{W_0}(n))$ larger, so as the cost of expression increases, generalization becomes difficult. However, it is difficult to accurately calculate $\text{VC}(\mathbf{H}^\sigma_{W_0}(n))$ for some given σ. If we focus on ReLU networks, by the result in (Bartlett et al., 2019), we have

Corollary 5.3. *For any given $n, W, W_0 \in \mathbb{N}_+$, there is a $\mathcal{D} \in \mathcal{D}(n)$ that satisfies the following properties.*

(1) There is an $\mathcal{F} \in \mathbf{H}_{W_0}(n)$ such that $A_{\mathcal{D}}(\mathcal{F}) = 1$;

(2) For any given $\epsilon, \delta \in (0,1)$, if $N \leq O(nW_0(1-4\epsilon-\delta))$, then for all $\mathcal{D}_{tr} \sim \mathcal{D}^N$, it holds $A_{\mathcal{D}}(\mathcal{F}) < 1-\epsilon$ for some $\mathcal{F} \in \mathbf{M}_W^{\sigma}(\mathcal{D}_{tr}, n)$.

Besides, for any distribution, we can show that if the parameters required to express a distribution tend to infinity, the required number of data to ensure the generalization for such a distribution must also tend to infinity. As shown in the following theorem. The proof is given in Appendix E.

Theorem 5.4. *Suppose $\mathcal{D} \in \mathcal{D}(n)$, $W_0 \geq 2^{n+1}$, and $A_{\mathcal{D}}(\mathcal{F}) \leq 1-\epsilon$ for any ϵ and $\mathcal{F} \in \mathbf{H}_{W_0}(n)$. If $N \leq W_0^{\frac{1}{n+1}}(n+1)/e$, then for any $\mathcal{D}_{tr} \sim \mathcal{D}^N$ and $W \in \mathbb{N}_+$, there is an $\mathcal{F} \in \mathbf{M}_W(\mathcal{D}_{tr}, n)$ such that $A_{\mathcal{D}}(\mathcal{F}) \leq 1-\epsilon$.*

However, since Theorem 5.4 is correct for all distributions and datasets, it can only provide a relatively loose bound. If the distribution is given, we can calculate the relationship between the minimum number of data required and the minimum number of parameters required to fit it, as shown in the following section.

5.2 Appropriate network model helps with generalization

As mentioned in the previous sections, expressive ability and generalization ability are closely related. Section 4.1 demonstrates that simpler expressions facilitate generalization; Section 5.1 reveals that, in the worst-case scenario, the amount of data required to guarantee generalization approximates the VC-dim of the hypothesis space that can express the distribution.

Therefore, for a given distribution, selecting an appropriate network model that can fit the distribution easily may help facilitate better expression with fewer data and network size, ultimately leading to improved generalization. In this paper, focusing on two-layer networks, we illustrate that selecting an appropriate activation function for the neural network according to the target distribution enhances generalization.

To better explain this conclusion, let us examine the following distribution.

Definition 5.5. Let $\mathcal{D}_n \in [0,1]^n \times \{-1,1\}$ be a distribution defined on $\{(\frac{i}{n}\mathbf{1}, \mathbb{I}(i))\}_{i=1}^n$ where $\mathbf{1}$ is the vector with all one entries in $\mathbb{R}^n$, $\mathbb{I}(x) = 1$ if x is odd and $\mathbb{I}(x) = -1$ if x is even, and the probability of each point is the same.

ReLU networks need $\Omega(n)$ widths to express this distribution and require $\Omega(n)$ data to ensure generalization. The proof is given in Appendix F.

Proposition 5.6. *(1) For any n, $A_{\mathcal{D}_n}(\mathcal{F}) < 1$ for any $\mathcal{F} \in \mathbf{H}_W(n)$ when $W < n/2$;*

(2) If $N \leq \delta n$ where $\delta \in (0,1)$, then for all $\mathcal{D}_{tr} \sim \mathcal{D}_n^N$ and $W \in \mathbb{N}_+$, it holds $A_{\mathcal{D}}(\mathcal{F}) \leq 0.5 + 2\delta$ for some $\mathcal{F} \in \mathbf{M}_W(\mathcal{D}_{tr}, n)$.

But if we use the activation function $\sigma(x) = \sin(\pi x)$, the networks only need $O(1)$ width to express such a distribution and require fewer data to ensure generalization. The proof is given in Appendix G.

Proposition 5.7. *(1) For any n, $\mathcal{D}_n$ can be expressed by $\mathbf{H}_1^{\sigma}(n)$ with confidence 1;*

(2) For any $W \geq 2, n > 2$, $\delta \in (0,1)$ and $N \geq 4\frac{\ln(\delta/2)}{\ln(0.5+1/n)}$, with probability $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}_n^N$, it holds $A_{\mathcal{D}}(\mathcal{F}) = 1$ for all $\mathcal{F} \in \mathbf{M}_W^{\sigma}(\mathcal{D}_{tr}, n)$.

As shown in the above example, using $\sigma(x) = \sin(\pi x)$ as the activation function only requires $O(\ln(\delta/2))$ samples and $O(1)$ width to ensure generalization, but ReLU networks require at least $\Omega(n)$ samples and widths to ensure generalization. This demonstrates the crucial role of selecting the appropriate network model.

Remark 5.8. It is worth mentioning that for some very simple distributions like the Bernoulli distribution, the performance of various activation functions is similar, so we cannot provide a general conclusion for any distribution.

6 EXPLANATION OF SOME PHENOMENA IN DEEP NEURAL NETWORK

Although networks minimizing empirical risk are good for generalization, many classic experimental results have shown that networks still have problems. In this section, we will provide explanations for some classic experimental results based on our theoretical results.

6.1 WHY DO GENERAL NETWORKS LACK ROBUSTNESS?

Experiments show that using ERM to train a network can easily lead to low robustness accuracy (Szegedy, 2013). In this section, we provide some explanations for this fact.

The *robustness accuracy* of network $\mathcal{F}$ under distribution $\mathcal{D}$ and robust radius ϵ is defined as

$$\mathrm{Rob}_{\mathcal{D},\epsilon}(\mathcal{F}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(\mathbb{I}(\widehat{\mathcal{F}}(x') = y), \forall x' \in \mathbb{B}(x,\epsilon) \cap [0,1]^n).$$

The robustness accuracy requires not only correctness on the samples but also correctness within a neighborhood of the sample. We introduce a notation.

Definition 6.1. For a dataset $\mathcal{D}_{tr} = \{(x_i, y_i)\}_{i=1}^N$ and an $\epsilon > 0$, define

$$R(\mathcal{D}_{tr}, \epsilon) = \{\mathcal{D}_r \,\|\, \mathcal{D}_r = \mathcal{D}_{tr} \cup \{(x_i + \epsilon_i, y_i)\}_{i=1}^N, \text{ for some } ||\epsilon_i|| \leq \epsilon\}.$$

It is easy to see that $R(\mathcal{D}_{tr}, \epsilon)$ contains all the data formed by adding a perturbation with budget ϵ to $\mathcal{D}_{tr}$. In the above section, we mainly discussed the network expression ability in distribution. On the other hand, there are also some studies on the network expression ability on dataset such as memorization. Moreover, previous studies (Park et al., 2021; Li et al., 2022; Yu et al., 2024a) have shown that robustly memorizing a dataset may be much more difficult than memorizing a dataset. So, for a given hypothesis space $\mathbf{H}$ that can express a normal data set well, it may not be able to express the dataset after disturbance. In this case, in order to minimize the empirical risk, the network will prioritize simple features that are easy to fit, but will ignore the complex robust features, which leads to low robustness, as shown in the following theorem. The proof is given in Appendix H.

Theorem 6.2. *Let $\mathcal{D} \in \mathcal{D}(n)$ and L_p be the Lipschitz constant of activation function σ. If $N_0, W_0 \in \mathbb{N}_+$ and $\epsilon, \delta, c_0, c_1 > 0$ satisfy that with probability $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^{N_0}$, it holds*

(1) there exists an $\mathcal{F} \in \mathbf{H}_{W_0}^{\sigma}(n)$ such that $y\mathcal{F}(x) \geq c_0$ for all $(x,y) \in \mathcal{D}_{tr}$;

(2) there exists a $\mathcal{D}_r \in R(\mathcal{D}_{tr}, \epsilon)$, such that $\sum_{(x,y)\in\mathcal{D}_r} \frac{y\mathcal{F}(x)}{|\mathcal{D}_r|} \leq c_1$ for any $\mathcal{F} \in \mathbf{H}_{W_0}^{\sigma}(n)$.

Then, for any $W \geq W_0 + 1$, with probability $1 - O(\delta)$ of $\mathcal{D}_{tr} \sim \mathcal{D}^{N_0}$ and $\mathcal{F} \in \mathbf{M}_W^{\sigma}(\mathcal{D}_{tr}, n)$, we have $\mathrm{Rob}_{\mathcal{D},\epsilon}(\mathcal{F}) \leq 1 - \Omega(\frac{c_0 - 2c_1}{L_p W_0 n} - \frac{c_1}{L_p W_0 n}(\frac{W_0}{W} + \frac{1}{W_0}) - \sqrt{\frac{\ln(n/\delta)}{N_0}})$.

This theorem implies that if the dataset after adding perturbations becomes more difficult to fit, the network may have a low robustness generalization. Please note that although the conclusion is directly unrelated to ϵ, because c_1 is related to ϵ, ϵ also affects the conclusion implicitly.

Remark 6.3. Conditions (1) and (2) required in the theorem are reasonable. It is obvious that as ϵ increases, c_1 will decrease, and when ϵ is large enough, we have $c_0 >> c_1 \approx 0$. Hence, in some situation, a small ϵ is also enough to make $c_0 >> c_1$, such as the example given in the proof of Theorem 4.3 in (Li et al., 2022).

6.2 IMPORTANCE OF OVER-PARAMETERIZED NETWORKS

In the above section, we mainly consider $\mathcal{F} \in \mathbf{M}_W(\mathcal{D}_{tr}, n)$. But what we really need is $\mathcal{F} \in \arg\max_{\mathcal{G}\in\mathbf{H}_W(n)} A_{\mathcal{D}}(\mathcal{G})$. By Theorem 4.3, it is easy to show that when the number of data and the size of the network are large enough, the generalization of $\mathcal{F} \in \mathbf{M}_W(\mathcal{D}_{tr}, n)$ and $\mathcal{F} \in \arg\max_{f\in\mathbf{H}_W(n)} A_{\mathcal{D}}(f)$ are close, as shown below.

Corollary 6.4. *Following Theorem 4.3, for all $\mathcal{F}_1 \in \mathbf{M}_W(\mathcal{D}_{tr}, n)$ and $\mathcal{F}_2 \in \arg\max_{f\in\mathbf{H}_W(n)} A_{\mathcal{D}}(f)$, we have* $|A_{\mathcal{D}}(\mathcal{F}_2) - A_{\mathcal{D}}(\mathcal{F}_1)| \leq O(\frac{W_0}{cW} + \frac{nL_p(W_0+c)\sqrt{\log(4n)}}{c\sqrt{N}} + \sqrt{\frac{\ln(2/\delta)}{N}})$.

Proof. Since $1 \geq A_{\mathcal{D}}(\mathcal{F}_2) \geq A_{\mathcal{D}}(\mathcal{F}_1)$, we have $|A_{\mathcal{D}}(\mathcal{F}_2) - A_{\mathcal{D}}(\mathcal{F}_1)| \leq 1 - A_{\mathcal{D}}(\mathcal{F}_1)$, and by Theorem 4.3, we obtain the result. □

The above corollary shows that if the size of the network is large enough, the gap will be small. In the following, we point out that for some distribution $\mathcal{D}$, if the size of network is too small, even with enough data, it may lead to a large gap of $A_{\mathcal{D}}(\mathcal{F}_2) - A_{\mathcal{D}}(\mathcal{F}_1)$. This emphasizes the importance of over-parameterization, as shown below. The proof is given in the Appendix I.

Proposition 6.5. *For some distribution $\mathcal{D} \in \mathcal{D}(n)$, there is a $W_0 > 0$, such that*

(1) There exists an $\mathcal{F} \in \mathbf{H}_{W_0}(n)$ such that $A_{\mathcal{D}}(\mathcal{F}) \geq 0.99$;

(2) For any $\delta > 0$, if $N \geq \Omega(n^2 \ln(n/\delta))$, with probability $1 - O(\delta)$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, we have $A_{\mathcal{D}}(\mathcal{F}) \leq 0.6$ for all $\mathcal{F} \in \mathbf{M}_{W_0}(\mathcal{D}_{tr}, n)$.

Remark 6.6. In Proposition 6.5, 0.99 can be changed to any real number in $(0, 1)$ and 0.6 can be changed to any real number in $(0.5, 1)$, and the result is still correct.

By Corollary 6.4, a large width does not make (2) in Proposition 6.5 true. So, the above conclusion indicates that for some distributions, when the network is not large enough, even if the network has generalizability, they cannot be found by minimizing the empirical risk. The distribution considered here contains some outliers. In order to fit these outliers, the small network must reduce generalization.

6.3 The Impact of Loss Function

In order to ensure generalizability of the network after empirical risk minimization, it is necessary to choose an appropriate loss function because minimizing some types of loss function is not good for generalization. In the previous sections, we mainly discussed the crossentropy loss function. In this section, we point out that not all loss functions can reach conclusions similar to Theorem 4.3.

Definition 6.7. We say that the loss function $L_b : \mathbb{R}^2 \to \mathbb{R}$ is bad if (1) or (2) is valid.

(1) There exist $x_{-1}, x_1 \in \mathbb{R}$ such that $L_b(x_{-1}, -1) = \min_{x \in \mathbb{R}} L_b(x, -1)$ and $L_b(x_1, 1) = \min_{x \in \mathbb{R}} L_b(x, 1)$.

(2) $L_b(\mathcal{F}(x), y) = \phi(y\mathcal{F}(x))$, where ϕ is a strictly decreasing concave function.

Condition (1) in the definition means that the loss function can reach its minimum value and condition (2) means that the loss function is a concave function. For example, some commonly used loss functions, such as the MSE loss function $L_{\text{MSE}}(\mathcal{F}(x), y) = ||\mathcal{F}(x) - y||_2$, or $L_q(\mathcal{F}(x), y) = -y\mathcal{F}(x)$, are all bad loss functions.

For such bad loss functions, we have

Theorem 6.8. *For any n and bad loss function L_b, there is a distribution $\mathcal{D} \in \mathcal{D}(n)$, such that for any $N \geq 0$, there is a $W_0 \geq 0$, such that if $W \geq W_0$, then with probability 0.99 of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, we have $A_{\mathcal{D}}(\mathcal{F}) \leq 0.5$ for some $\mathcal{F} \in \arg\min_{\mathcal{G} \in \mathbf{H}_W(n)} \sum_{(x,y) \in \mathcal{D}_{tr}} L_b(\mathcal{G}(x), y)$.*

This theorem means that to ensure generalizability, it is important to choose the appropriate loss function. The proof is given in the Appendix J.

7 Conclusion

In this paper, we give a lower bound for the population accuracy of the neural networks that minimize the empirical risk, which implies that as long as there exist enough training data and the network is large enough, generalization can be achieved. The data and network sizes required only depend on the size required for the network to represent the target data distribution. Furthermore, we show that if the scale required for the network to represent a data distribution increases, the amount of data required to achieve generalization on that distribution will also inevitably increase. Finally, the results are used to explain some phenomena in deep learning.

Limitation and future work. Although considering 2 layer networks is quite common in theoretical analysis of deep learning, it is still desirable to extend the result to deep neural networks. Preliminary results for deep neural networks are given in Appendix K, which need to be further studied. A more accurate estimate of the cost required to represent a given data distribution is needed.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.

Pierre Alquier et al. User-friendly introduction to pac-bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International conference on learning representations*, 2020.

Andrew R Barron and Jason M Klusowski. Approximation and estimation for high-dimensional deep learning networks. *arXiv preprint arXiv:1809.03090*, 2018.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33: 4381–4391, 2020.

Alexander Bastounis, Anders C Hansen, and Verner Vlačić. The mathematics of adversarial attacks in ai–why deep learning is unstable despite the existence of stable neural networks. *arXiv preprint arXiv:2109.06098*, 2021.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Alon Brutzkus and Amir Globerson. An optimization and generalization analysis for max-pooling networks. In *Uncertainty in Artificial Intelligence*, pp. 1650–1660. PMLR, 2021.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pp. 3174–3185. PMLR, 2021.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension bounds for piecewise linear neural networks. In *Conference on learning theory*, pp. 1064–1068. PMLR, 2017.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization bounds: Perspectives from information theory and pac-bayes. *arXiv preprint arXiv:2309.04381*, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *ICLR 2020*, 2020.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824. PMLR, 2018.

Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8279–8287, 2021.

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 191–227. PMLR, 2023.

Binghui Li, Jikai Jin, Han Zhong, John E Hopcroft, and Liwei Wang. Why robust generalization in deep learning is difficult: Perspective of expressive power. In *Advances in Neural Information Processing Systems*, 2022.

Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *arXiv preprint arXiv:1806.05159*, 2018.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023.

Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. Learning over-parametrized two-layer neural networks beyond ntk. In *Conference on learning theory*, pp. 2613–2682. PMLR, 2020.

Philip M Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. *arXiv preprint arXiv:1905.12600*, 2019.

Tao Luo and Haizhao Yang. Two-layer neural networks for partial differential equations: Optimization and generalization theory. *arXiv preprint arXiv:2006.15733*, 2020.

Chao Ma, Lei Wu, et al. A priori estimates of the population risk for two-layer neural networks. *arXiv preprint arXiv:1810.06397*, 2018.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Sejun Park, Jaeho Lee, Chulhee Yun, and Jinwoo Shin. Provable memorization via deep neural networks using sub-linear parameters. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3627–3661. PMLR, 15–19 Aug 2021.

Jayanth Regatti, Gaurav Tendolkar, Yi Zhou, Abhishek Gupta, and Yingbin Liang. Distributed sgd generalizes well under asynchrony. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 863–870. IEEE, 2019.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, 1972.

Ali Shafahi, Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *ICLR*, 2019.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Jianhui Sun, Ying Yang, Guangxu Xun, and Aidong Zhang. Scheduling hyperparameters to improve generalization: From centralized sgd to asynchronous sgd. *ACM Transactions on Knowledge Discovery from Data*, 17(2):1–37, 2023.

C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Hossein Taheri and Christos Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156):1–41, 2024.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pp. 1405–1413. PMLR, 2024.

Guillermo Valle-Pérez and Ard A. Louis. Generalization bounds for deep learning. *arXiv preprint arXiv:2203.14533*, 2022.

Gal Vardi, Ohad Shamir, and Nati Srebro. The sample complexity of one-hidden-layer neural networks. *Advances in Neural Information Processing Systems*, 35:9139–9150, 2022.

Mingze Wang and Chao Ma. Generalization error bounds for deep neural networks trained by sgd. *arXiv preprint arXiv:2206.03299*, 2022.

Yihan Wang, Shuang Liu, and Xiao-Shan Gao. Data-dependent stability analysis of adversarial training. *arXiv preprint arXiv:2401.03156*, 2024.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

Ming Wen, Yixi Xu, Yunling Zheng, Zhouwang Yang, and Xiao Wang. Sparse deep neural networks using l 1,infinite-weight normalization. *Statistica Sinica*, 31(3):1397–1414, 2021.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35:15446–15459, 2022.

Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. *Advances in neural information processing systems*, 34:26523–26535, 2021.

Yahong Yang, Haizhao Yang, and Yang Xiang. Nearly optimal vc-dimension and pseudo-dimension bounds for deep neural network derivatives. *Advances in Neural Information Processing Systems*, 36:21721–21756, 2023.

Lijia Yu, Yihan Wang, and Xiao-Shan Gao. Adversarial parameter attack on deep neural networks. In *International Conference on Machine Learning*, pp. 40354–40372. PMLR, 2023.

Lijia Yu, Xiao-Shan Gao, and Lijun Zhang. Optimal robust memorization with relu neural networks. In *International Conference on Learning Representations*, 2024a.

Lijia Yu, Xiao-Shan Gao, Lijun Zhang, and Yibo Miao. Generalizablity of memorization neural network. In *NeurIPS*, 2024b.

Yibo Zeng and Henry Lam. Generalization bounds with minimal dependency on hypothesis class via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 35: 27576–27590, 2022.

Chiyuan Zhang, Benjamin Recht, Samy Bengio, and Oriol Vinyals. Understanding deep learning requires re-thinking generalization. In *ICLR*, 2017.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021.

A PROOF OF PROPOSITION 4.2

A function σ is sigmoidal if $\text{limit}_{x\to-\infty}\sigma(x) = 0$ and $\text{limit}_{x\to\infty}\sigma(x) = 1$. Then, we have

Theorem A.1 (Theorem 1 in Cybenko (1989)). *For any continuous sigmoidal activation function σ, $\epsilon \in (0,1)$ and continuous function $f : [0,1]^n \to \mathbb{R}$, there exist $W \geq 0$ and $F \in \mathbf{H}^{\sigma}_{W}(n)$ such that $|f(x) - F(x)| \leq \epsilon$.*

We prove Proposition 4.2 by using the above Theorem.

Proof. It is easy to see that $\sigma(x) = \text{ReLU}(x+1) - \text{ReLU}(x)$ is a continuous sigmoidal activation function.

Denote $Z^{\sigma}_{W}(n)$ as the set of all two-layer neural networks with input dimension n, width W, and activation function σ. For simplicity, $Z_W(n)$ means $Z^{\text{ReLU}}_{W}(n)$.

Firstly, it is easy to see that $Z^{\sigma}_{W}(n) \subset Z_{2W}(n)$ for any $W \in \mathbb{N}_+$.

Then, because $\mathcal{D}$ has a positive separation distance with a different label, there is a continuous function f such that: $f(x) = 1$ if x has label 1 in distribution $\mathcal{D}$; $f(x) = -1$ if x has label -1 in distribution $\mathcal{D}$.

Finally, by Theorem A.1, there exist a W and a $\mathcal{F} \in Z^{\sigma}_{W}(n)$ such that $|\mathcal{F}(x) - f(x)| \leq 0.1$ for all $x \in [0,1]^n$. Thus, $\mathcal{F} \in Z^{\sigma}_{W}(n) \subset Z_{2W}(n)$ and $P_{(x,y)\sim\mathcal{D}}(y\mathcal{F}(x) \geq 0.9) = 1$.

Let the maximum of the absolute value of the parameters of $\mathcal{F}$ be A. If $A \leq 1$, then $\mathcal{F}$ is what we want. If $A > 1$, then let $\mathcal{F}_A$ be a network whose parameter is the corresponding parameter of $\mathcal{F}$ divided by A, so $\mathcal{F}_A = \mathcal{F}/A^2$. Hence, there are $\mathcal{F}_A \in \mathbf{H}_{2W}(n)$ and $P_{(x,y)\sim\mathcal{D}}(y\mathcal{F}(x) \geq 0.9/A^2) = 1$. The proposition is proved. □

B PROOF OF THEOREM 4.3

B.1 PREPARATORY WORK

We give some definitions of the hypothesis space.

Definition B.1. For a network $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}$ and an $a > 0$, let $\mathcal{F}_{-a,a}(x) = \min\{\max\{-a, \mathcal{F}(x)\}, a\}$, that is, clamp $\mathcal{F}$ in $[-a, a]$. Then for any hypothesis space $\mathbf{H}$, let $\mathbf{H}_{-a,a} = \{\mathcal{F}_{-a,a} \| \mathcal{F} \in H\}$.

We define the Radermacher complexity.

Definition B.2. For a hypothesis space $\mathbf{H}$ and dataset $\mathcal{D}$, the Radermacher complexity of $\mathbf{H}$ under dataset $\mathcal{D}$ is:

$$\text{Rad}_{\mathbf{H}}(\mathcal{D}) = \mathbb{E}_{(q_i)_{i=1}^{|\mathcal{D}|}}\left[\sup_{\mathcal{F}\in\mathbf{H}} \frac{\sum_{x_i\in\mathcal{D}} q_i\mathcal{F}(x_i)}{|\mathcal{D}|}\right]$$

where q_i satisfies that $P(q_i = 1) = P(q_i = -1) = 0.5$ and q_i are i.i.d.

Here are some results about the Radermacher complexity:

Lemma B.3. *For any hypothesis space $\mathbf{H}$, let $\mathbf{H}_{+a} = \{\mathcal{F} + a \| \mathcal{F} \in \mathbf{H}\}$, where $a \in \mathbb{R}$. Then for any hypothesis space $\mathbf{H}$, $a \in \mathbb{R}$ and dataset $\mathcal{D}$, there are* $\text{Rad}_{\mathbf{H}}(\mathcal{D}) = \text{Rad}_{\mathbf{H}_{+a}}(\mathcal{D})$.

Let the $L_{1,\infty}$ norm of a matrix W be the maximum value of the L_1 norm for each row of the matrix W.

Lemma B.4. *Let $\mathcal{F}_{n,d,(L_i),(c_i)} : \mathbb{R}^n \to \mathbb{R}$ be a network with d hidden layers, L_i Lipschitz-continuous activation function for i-th activation function, and the output layer does not contain an activation function. Let w_i be the i-th transition matrix and b_i be the i-th bias vector. Then the $L_{1,\infty}$ norm of w_i plus the $L_{1,\infty}$ norm of b_i is not more than c_i.*

Let $\mathbf{H}_{n,d,(L_i),(c_i)} = \{\mathcal{F}_{n,d,(L_i),(c_i)}\}$. Then when $L_i \geq 1$, $c_i \geq 1$, for any $\{x_i\}_{i=1}^N \subset [0,1]^n$, there are:

$$\text{Rad}_{H_{n,d,(L_i),(c_i)}}(\{x_i\}_{i=1}^N) \leq \frac{\Pi_{i=1}^{d} L_i \Pi_{i=1}^{d+1} c_i}{\sqrt{N}}\left(\sqrt{(d+3)\log(4)} + \sqrt{2\log(2n)}\right).$$

This lemma is an obvious corollary of Theorem 1 in (Wen et al., 2021). By the above two lemmas we can calculate the Radermacher complexity of $\mathbf{H}_W^\sigma(n)_{-a,a}$.

Lemma B.5. *Let σ be a L_p Lipschitz-continuous activation function and $L_p \geq 1$, and let $\mathbf{H} = \{F(x,y) : F(x,y) = y\mathcal{F}(x), \mathcal{F}(x) \in \mathbf{H}_W^\sigma(n)_{-a,a}\}$ where $a > 0$ is given in Definition B.1. Then for any $S = \{(x_i, y_i)\}_{i=1}^N \subset [0,1]^n \times \{-1,1\}$, there are*

$$\text{Rad}_{\mathbf{H}}(S) \leq \frac{2L_p(n+1)(W+1+a)}{\sqrt{N}}(\sqrt{5\log(4)} + \sqrt{2\log(2n)}).$$

Proof. First, there are $\text{Rad}_{\mathbf{H}}(S) = \text{Rad}_{\mathbf{H}}(\{(x_i,y_i)\}_{i=1}^N) = \mathbb{E}_{(q_i)_{i=1}^N}[\sup_{f\in\mathbf{H}_W^\sigma(n)_{-a,a}} \frac{\sum_{i=1}^N q_i y_i f(x_i)}{|\mathcal{D}|}]$. Taking into account the definition of q_i in definition B.2, there are $\text{Rad}_{\mathbf{H}}(\{(x_i,y_i)\}_{i=1}^N) = \mathbb{E}_{(q_i)_{i=1}^N}[\sup_{f\in\mathbf{H}_W^\sigma(n)_{-a,a}} \frac{\sum_{i=1}^N q_i f(x_i)}{|D|}] = \text{Rad}_{\mathbf{H}_W^\sigma(n)_{-a,a}}(\{x_i\}_{i=1}^N)$.

So, we just need to calculate $\text{Rad}_{\mathbf{H}_W^\sigma(n)_{-a,a}}(\{x_i\}_{i=1}^N)$.

First, for any function f and $a > 0$, $k \in N^+$, we have

$$\begin{aligned} & f_{-a,a}(x) \\ = \ & \text{ReLU}(f(x)+a) - \text{ReLU}(f(x)-a) - a \\ = \ & \textstyle\sum_{i=1}^k (\text{ReLU}(f(x)/k + a/k) - \text{ReLU}(f(x)/k - a/k)) - a \end{aligned}$$

On the other hand, let $H_{+a} = \{f + a \| f \in \mathbf{H}_W^\sigma(n)_{-a,a}\}$. Then for any $F \in \mathbf{H}_{+a}$, there are $F = f_{-a,a}(x) + a$ for some $f \in \mathbf{H}_W^\sigma(n)$. Then by the above form of expression, take $k = [W/2]$, F and write it as a network with:

(1): Depth 3. Because f has depth 2, after adding a ReLU activation function, it was depth 3.

(2): The first layer has an L_p Lipschitz-continuous activation function; the second layer has a 1 Lipschitz-continuous activation function, that is, ReLU.

(3): The $L_{1,\infty}$ norm of the three transition matrices plus bias vectors are $n+1$, $\frac{W+1+a}{[W/2]}$ and $2[W/2]$.

So, by Lemmas B.4 and B.3, there are $\text{Rad}_{\mathbf{H}_{+a}}(\{x_i\}_{i=1}^N) = \text{Rad}_{\mathbf{H}_W^\sigma(n)_{-a,a}}(\{x_i\}_{i=1}^N) = \frac{2L_p(n+1)(W+1+a)}{\sqrt{N}}(\sqrt{5\log(4)} + \sqrt{2\log(2n)})$. The theorem is proved. □

Another important Theorem is required.

Theorem B.6 (Theorem in Mohri et al. (2018))**.** *Let $H = \{F : \mathbb{R}^n \to [-a,a]\}$, and $\mathcal{D}$ be a distribution, then with probability $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, there are:*

$$|\mathbb{E}_{x\sim\mathcal{D}}[F(x)] - \sum_{x\in\mathcal{D}_{tr}} \frac{F(x)}{N}| \leq 2\text{Rad}_H(\mathcal{D}_{tr}) + 6a\sqrt{\frac{\ln(2/\delta)}{2N}},$$

for any $F \in H$.

We give a simple lemma below.

Lemma B.7. *(1): When $0 < x \leq e$, there are $\ln(1+x) \geq x/(e+1)$.*

(2): When $x > 0$, there are $xe^{-x} \leq 1/e$.

Proof. For (1): Consider $f(x) = \ln(1+x) - x/(e+1)$, there are $f'(x) = 1/(1+x) - 1/(1+e) \geq 0$, so $f(x) \geq f(0) = 0$, which means that $\ln(1+x) - x/(e+1) \geq 0$.

For (2): Consider $f(x) = xe^{-x}$, there are $f'(x) = e^{-x}(1-x)$, it is easy to see that $f'(x)$ become positive then negative when x from 0 to ∞, and $f'(1) = 0$, so $f(x) \leq f(1) = 1/e$. □

B.2 Proof of Theorem 4.3

Proof. Let $\mathcal{D}_{tr} \sim \mathcal{D}^N$ and $\mathcal{F}$ be a network in $\mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n)$. We prove Theorem 4.3 in four parts:

Part one: We have $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq N \ln(1 + e^{-c[\frac{W}{W_0+1}]})$.

Because $\mathcal{D}$ can be expressed by $\mathbf{H}_{W_0}^\sigma(n)$ with confidence c, so there is a network $\mathcal{F}_0 = \sum_{i=1}^{W_0} a_i\sigma(W_i x + b_i) + c_1$ such that $y\mathcal{F}(x) \geq c$ for all $(x, y) \sim \mathcal{D}$. Moreover, we can write such network as $\mathcal{F}_0 = \sum_{i=1}^{W_0+1} a_i\sigma(W_i x + b_i)$, where $a_{W_0+1} = \text{Sgn}(c_1)$, $W_{W_0+1} = 0$, $b_{W_0+1} = |c_1|$.

Now, we consider the following network in $\mathbf{H}_W^\sigma(n)$:

$$\mathcal{F}_W = \sum_{i=1}^{(W_0+1)[\frac{W}{W_0+1}]} a_{i\%(W_0+1)}\sigma(W_{i\%(W_0+1)}x + b_{i\%(W_0+1)}),$$

Here, we stipulate that $i\%(W_0 + 1) = W_0 + 1$ when $W_0 + 1|i$. Then we have $\mathcal{F}_W(x) = [\frac{W}{W_0+1}]\mathcal{F}_0(x)$ and $\mathcal{F}_W(x) \in \mathbf{H}_W^\sigma(n)$. Moreover, there are $y\mathcal{F}_W(x) = y[\frac{W}{W_0+1}]\mathcal{F}_0(x) \geq [\frac{W}{W_0+1}]c$ for all $(x, y) \sim \mathcal{D}$, so $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}_W(x), y) \leq N \ln(1 + e^{-c[\frac{W}{W_0+1}]})$. So for any $\mathcal{F} \in \arg\min_{f\in\mathbf{H}_W^\sigma(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$, there are $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}_W(x), y) \leq N \ln(1 + e^{-c[\frac{W}{W_0+1}]})$.

Part Two: Let $k = [\frac{W}{W_0+1}]$, by the assumption in Theorem, there is $k \geq 1$. We will show that $|\{(x, y) : (x, y) \in \mathcal{D}_{tr}, y\mathcal{F}(x) \leq kc/2\}| \leq Ne^{-kc/2+2}$.

Let $S = \{(x, y) : (x, y) \in \mathcal{D}_{tr}, y\mathcal{F}(x) \leq kc/2\}$, then according to part one, there are: $|S| \ln(1 + e^{-kc/2}) \leq \sum_{(x,y)\in S} L(\mathcal{F}(x), y) \leq \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq N \ln 1 + e^{-kc} \leq Ne^{-kc}$. So, there are $|S| \ln 1 + e^{-kc/2} \leq Ne^{-kc}$.

By Lemma B.7, there are $|S|e^{-kc/2}/(e + 1) \leq |S| \ln 1 + e^{-kc/2} \leq Ne^{-kc}$, so $|S| \leq Ne^{-kc/2}(e + 1) < Ne^{-kc/2+2}$.

Part Three: By Definition B.1, let network $g = \mathcal{F}_{-kc/2,kc/2}$, we show that, with high probability, $\mathbb{E}_{(x,y)\sim\mathcal{D}} yg(x)$ has a lower bound.

Firstly, by part two, there are $\sum_{(x,y)\in\mathcal{D}_{tr}} yg(x) \geq N(kc(1 - e^{-kc/2+2})/2 - kce^{-kc/2+2}/2) = Nkc(1 - 2e^{-kc/2+2})/2$.

Then, let $H = \{y\mathcal{F}(x) : \mathcal{F}(x) \in \mathbf{H}_W^\sigma(n)_{-kc/2,kc/2}\}$, by Lemma B.5, there are $Rad_H(\mathcal{D}_{tr}) \leq \frac{2(n+1)(W+1+kc/2)L_p}{\sqrt{N}}(\sqrt{5\log(4)} + \sqrt{2\log(2n)})$, $Rad_H(\mathcal{D}_{tr})$ is defined in definition B.2.

So, considering that $yg(x) \in H$ and by Theorem B.6, with probability $1 - \delta$ of $\mathcal{D}_{tr}$, there are

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}} yg(x) \\
\geq\ & \tfrac{1}{N}\textstyle\sum_{(x,y)\in\mathcal{D}_{tr}} yg(x) - 2\text{Rad}([\mathbf{H}_W^\sigma(n)]_{-kc/2,kc}) - 3kc\sqrt{\tfrac{\ln(2/\delta)}{2N}} \\
\geq\ & kc(1 - 2e^{-kc/2+2})/2 - \tfrac{2(n+1)L_p(W+1+kc/2)}{\sqrt{N}}(\sqrt{5\log(4)} + \sqrt{2\log(2n)}) - 3kc\sqrt{\tfrac{\ln(2/\delta)}{2N}}.
\end{aligned}$$

Part Four: Now, we prove Theorem 4.3.

Firstly, there are $A_{\mathcal{D}}(g) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(yg(x) > 0) \geq \mathbb{E}_{(x,y)\sim\mathcal{D}}[yg(x)]/(kc/2)$, we use $|g(x)| \leq kc/2$ in here. So, by part three, with probability of $\mathcal{D}_{tr}$, there are

$$A_{\mathcal{D}}(g) \geq 1 - 2e^{-kc/2+2} - \frac{4(n+1)L_p(W+1+kc/2)}{\sqrt{N}kc}(\sqrt{5\log(4)} + \sqrt{2\log(2n)}) - 6\sqrt{\frac{\ln(2/\delta)}{2N}}.$$

By Lemma B.7 and $k = [W/(W_0+1)] \geq \frac{W}{2W_0}$ which is because $[W/(W_0+1)] = k \geq 1$ and $W_0 \geq 2$, there are $2e^{-kc/2+2} \leq \frac{4e}{kc} = \frac{4e}{c[\frac{W}{W_0+1}]} \leq \frac{8eW_0}{Wc}$; and it is easy to see that $\frac{4(n+1)L_p(W+1+kc/2)}{\sqrt{N}kc} \leq$

$\frac{4(n+1)WL_p(2+kc/2W)}{\sqrt{N}kc} \leq \frac{8nWL_p(2+c/2W_0)}{\sqrt{N}[W/(W_0+1)]c} \leq \frac{8nL_p(4W_0+c)}{\sqrt{N}c}$, the last inequality uses $[W/(W_0+1)] \geq \frac{W}{2W_0}$.

The last step uses $k = [W/(W_0+1)] \geq \frac{W}{2W_0}$. And $\sqrt{5\log(4)}+\sqrt{2\log(2n)} \leq (\sqrt{5}+\sqrt{2})\sqrt{\log(4n)}$. So there are:

$$A_{\mathcal{D}}(g) \geq 1 - \frac{8eW_0}{Wc} - \frac{8nL_p(1+4\frac{W_0}{c})}{\sqrt{N}}(\sqrt{5}+\sqrt{2})\sqrt{\log(4n)} - 6\sqrt{\frac{\ln(2/\delta)}{2N}}.$$

Lastly, because $A_{\mathcal{D}}(g) = A_{\mathcal{D}}(\mathcal{F})$, we have $A_{\mathcal{D}}(\mathcal{F}) \geq 1 - O(\frac{W_0}{Wc} + \frac{nL_p(W_0+c)\sqrt{\log(4n)}}{\sqrt{N}c} + \sqrt{\frac{\ln(2/\delta)}{N}})$.

The theorem is proved. □

C Proof of Proposition 4.7

The proof is similar to the proof of Theorem 4.3, so we just follow the proof of Theorem 4.3.

Proof. Let $\mathcal{D}_{tr} \sim \mathcal{D}^N$, $\mathcal{F}$ be a network in $\mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n)$, and $\mathcal{F}_q$ be a network which is a q-approximation of minimization empirical risk.

We prove Theorem 4.7 in four parts below.

Part one: It holds $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq N\ln(1+e^{-c[\frac{W}{W_0+1}]})$. This is as same as in Part one in the proof of Theorem 4.3

Part Two: Let $k = [\frac{W}{W_0+1}] \geq 1$. Then, $|\{(x,y) : (x,y) \in \mathcal{D}_{tr}, y\mathcal{F}_q(x) \leq kc/2\}| \leq qNe^{-kc/2+2}$.

Let $S = \{(x,y) : (x,y) \in \mathcal{D}_{tr}, y\mathcal{F}_q(x) \leq kc/2\}$, then according to part one, there are: $|S|\ln(1+e^{-kc/2}) \leq \sum_{(x,y)\in S} L(\mathcal{F}_q(x), y) \leq q\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq qN\ln 1 + e^{-kc} \leq qNe^{-kc}$. So, there are $|S|\ln 1 + e^{-kc/2} \leq qNe^{-kc}$.

By Lemma B.7, there are $|S|e^{-kc/2}/(e+1) \leq |S|\ln 1 + e^{-kc/2} \leq qNe^{-kc}$, so $|S| \leq qNe^{-kc/2}(e+1) < qNe^{-kc/2+2}$.

Part Three: By Definition B.1, let network $g = (\mathcal{F}_q)_{-kc/2,kc/2}$. We will show that, with high probability, $\mathbb{E}_{(x,y)\sim\mathcal{D}} yg(x)$ has a lower bound.

Firstly, by part two, we have $\sum_{(x,y)\in\mathcal{D}_{tr}} yg(x) \geq N(kc(1-qe^{-kc/2+2})/2 - qkce^{-kc/2+2}/2) = Nkc(1-2qe^{-kc/2+2})/2$.

So, with probability $1-\delta$ of $\mathcal{D}_{tr}$, it holds

$$\begin{aligned}
& \mathbb{E}_{(x,y)\sim\mathcal{D}} yg(x) \\
\geq\ & \tfrac{1}{N}\textstyle\sum_{(x,y)\in\mathcal{D}_{tr}} yg(x) - 2\mathrm{Rad}([\mathbf{H}_W^\sigma(n)]_{-kc/2,kc}) - 3kc\sqrt{\tfrac{\ln(2/\delta)}{2N}} \\
\geq\ & kc(1-2qe^{-kc/2+2})/2 - \tfrac{2(n+1)L_p(W+1+kc/2)}{\sqrt{N}}(\sqrt{5\log(4)} + \sqrt{2\log(2n)}) - 3kc\sqrt{\tfrac{\ln(2/\delta)}{2N}}.
\end{aligned}$$

Part Four: Now, we prove Proposition 4.7.

Firstly, there are $A_{\mathcal{D}}(g) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(yg(x) > 0) \geq \mathbb{E}_{(x,y)\sim\mathcal{D}}[yg(x)]/(kc/2)$. So, by part three, with $1-\delta$ probability of $\mathcal{D}_{tr}$, there are

$$A_{\mathcal{D}}(g) \geq 1 - 2qe^{-kc/2+2} - \frac{4(n+1)L_p(W+1+kc/2)}{\sqrt{N}kc}(\sqrt{5\log(4)}+\sqrt{2\log(2n)}) - 6\sqrt{\frac{\ln(2/\delta)}{2N}}.$$

Then, similar as part four in proof of Theorem 4.3, there are

$$A_{\mathcal{D}}(g) \geq 1 - \frac{8qeW_0}{Wc} - \frac{8nL_p(1+4\frac{W_0}{c})}{\sqrt{N}}(\sqrt{5}+\sqrt{2})\sqrt{\log(4n)} - 6\sqrt{\frac{\ln(2/\delta)}{2N}},$$

which is what we want. □

D PROOF OF THEOREM 5.2

Proof. Assume that Theorem 5.2 is wrong, then there exist n, W and W_0 such that for given $\epsilon, \delta \in (0,1)$, if $\mathcal{D} \in \mathcal{D}(n)$ and $N \geq \text{VC}(\mathbf{H}^{\sigma}_{W_0}(n))(1-4\epsilon-\delta)$, with probability $1-\delta$ of $\mathcal{D}_{tr}$, we have $A_{\mathcal{D}}(\mathcal{F}) \geq 1-\epsilon$ for all $\mathcal{F} \in \arg\min_{f \in \mathbf{H}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$.

We will derive contradictions on the basis of this conclusion.

Part 1: Find some points and values.

For a simple expression, let $k = \text{VC}(\mathbf{H}^{\sigma}_{W_0}(n))$, and $\{u_i\}_{i=1}^{k}$ be k points that can be shattered by $\text{VC}(\mathbf{H}^{\sigma}_{W_0}(n))$. Let $q = \text{VC}(\mathbf{H}^{\sigma}_{W_0}(n))(1-4\epsilon-\delta)$.

Now, we consider the following types of distribution $\mathcal{D}$:

(c1): $\mathcal{D}$ is a distribution in $\mathcal{D}(n)$ and $\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in \{u_i\}_{i=1}^{k}) = 1$.

(c2): $\mathbb{P}_{(x,y)\sim\mathcal{D}}(x = u_i) = \mathbb{P}_{(x,y)\sim\mathcal{D}}(x = u_j) = 1/k$ for any $i, j \in [k]$.

Let S be the set that contains all such distributions, and it is easy to see that for any $\mathcal{D} \in S$, it can be expressed by $\mathbf{H}^{\sigma}_{W_0}(n)$.

Part 2: Some definition.

Moreover, for $\mathcal{D} \in S$, we define $S(\mathcal{D})$ as the following set:

$Z \in S(\mathcal{D})$ if and only if $Z \in [k]^q$ is a vector satisfying: Define $D(Z)$ as $D(Z) = \{(u_{z_i}, y_{z_i})\}_{i=1}^{q}$, then $A_{\mathcal{D}}(\mathcal{F}) \geq 1-\epsilon$ for all $\mathcal{F} \in \arg\min_{f \in \mathbf{H}_W(n)} \sum_{(x,y)\in D_Z} L(f(x), y)$, where z_i is the i-th weight of Z and y_{z_i} is the label of u_{z_i} in distribution $\mathcal{D}$.

It is easy to see that if we i.i.d. select q samples in distribution $\mathcal{D}$ to form a dataset $\mathcal{D}_{\text{tr}}$, then by $c2$, with probability 1, $\mathcal{D}_{\text{tr}}$ only contain the samples (u_j, y_j) where $j \in [k]$.

Now for any $\mathcal{D}_{tr}$ selected from $\mathcal{D}$, we construct a vector in $[k]^q$ as follows: the index of i-th selected samples as the i-th component of the vector. Then each selection situation corresponds to a vector in $[k]^q$ which is constructed as before. Then by the definition of $S(\mathcal{D})$, we have $A_{\mathcal{D}}(\mathcal{F}) \geq 1-\epsilon$ for all $\mathcal{F} \in \arg\min_{f \in \mathbf{H}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$ if and only if the corresponding vector of $\mathcal{D}_{\text{tr}}$ is in $S(\mathcal{D})$.

By the above result and by the assumption at the beginning of the proof, for any $\mathcal{D} \in S$ we have $\frac{|S(\mathcal{D})|}{q^k} \geq 1-\delta$.

Part 3: Prove the theorem.

Let S_s be a subset of S, and $S_s = \{\mathcal{D}_{i_1,i_2,\ldots,i_k}\}_{i_j\in\{-1,1\}, j\in[k]} \subset S$, where the distribution $\mathcal{D}_{i_1,i_2,\ldots,i_k}$ satisfies the label of u_j is i_j, where $j \in [k]$.

We will show that there exists at least one $\mathcal{D} \subset S_s$, such that $|S(\mathcal{D})| < (1-\delta)q^k$, which is contrary to the inequality $\frac{|S(\mathcal{D})|}{q^k} \geq 1-\delta$ as shown in the above. To prove that, we only need to prove that $\sum_{\mathcal{D}\in S_s} |S(\mathcal{D})| < (1-\delta)2^k q^k$, use $|S_s| = 2^k$ here.

To prove that, for any vector $Z \in [k]^q$, we estimate how many $\mathcal{D} \in S_s$ make Z included in $S(\mathcal{D})$.

Part 3.1, situation of a given vector Z and a given distribution $\mathcal{D}$.

For a $Z = (z_i)_{i=1}^{q}$ and $\mathcal{D}$ such that $Z \in S(\mathcal{D})$, let $\text{len}(Z) = \{c \in [k] : \exists i, c = z_i\}$. We consider the distributions in S_s that satisfy the following condition: for $i \in \text{len}(Z)$, the label of u_i is equal to the label of u_i in $\mathcal{D}$. Obviously, we have $2^{k-|\text{len}(Z)|}$ distributions that can satisfy the above condition in S_s. Let such distributions make up a set $S_{ss}(\mathcal{D}, Z)$. Now, we estimate how many distributions $\mathcal{D}_s$ in $S_{ss}(\mathcal{D}, Z)$ satisfy $Z \in S(\mathcal{D}_s)$.

It is easy to see that if $\mathcal{D}_s \in S_{ss}(\mathcal{D}, Z)$ such that there are more than $[2k\epsilon]$ of $i \in [k]$, $\mathcal{D}_s$ and D have different labels of u_i, then $\min\{A_{\mathcal{D}}(\mathcal{F}), A_{\mathcal{D}_s}(\mathcal{F})\} < 1-\epsilon$ for any $\mathcal{F}$. So considering $A_{\mathcal{D}}(\mathcal{F}) \geq 1-\epsilon$ for all $\mathcal{F} \in \arg\min_{f \in \mathbf{H}_W(n)} \sum_{(x,y)\in D_Z} L(f(x), y)$, by the above result, such kind

of $\mathcal{D}_s$ is at most $\sum_{i=0}^{[2k\epsilon]} C_{k-|\text{len}(Z)|}^i$. So, we have that: There are at most $\sum_{i=0}^{[2k\epsilon]} C_{k-|\text{len}(Z)|}^i$ numbers of distributions $\mathcal{D}_s$ in $S_{ss}(\mathcal{D}, Z)$ satisfy $Z \in S(\mathcal{D}_s)$.

Part 3.2, for any vector Z and distribution $\mathcal{D}$.

For any distribution $\mathcal{D} \in S_s$, let $y(\mathcal{D})_i$ be the label of u_i in distribution $\mathcal{D}$.

Firstly, for a given Z, we have at most $2^{|len(Z)|}$ different $\mathcal{S}_{ss}(\mathcal{D}, Z)$ for $\mathcal{D} \in \mathcal{D}_S$. Because when $\mathcal{D}_1$ and $\mathcal{D}_2$ satisfy $y(\mathcal{D}_1)_i = y(\mathcal{D}_2)_i$ for any $i \in \text{len}(Z)$, we have $\mathcal{D}_{ss}(\mathcal{D}_1, Z) = \mathcal{D}_{ss}(\mathcal{D}_2, Z)$, and $2^{|\text{len}(Z)|}$ situations of label of u_i where $i \in \text{len}(Z)$, so there exist at most $2^{|\text{len}(Z)|}$ different $\mathcal{S}_{ss}(\mathcal{D}, Z)$.

Then, by part 3.1, for an $\mathcal{S}_{ss}(\mathcal{D}, Z)$, at most $\sum_{i=0}^{[2k\epsilon]} C_{k-|\text{len}(Z)|}^i$ of $\mathcal{D}_s \in S_{ss}(\mathcal{D}, Z)$ satisfies $Z \in S(\mathcal{D}_s)$. So by the above result and consider that $\mathcal{D}_s = \cup_{\mathcal{D} \in \mathcal{D}_s} \mathcal{S}_{ss}(\mathcal{D}, Z)$, at most $2^{|\text{len}(Z)|} \sum_{i=0}^{[2k\epsilon]} C_{k-|\text{len}(Z)|}^i$ number of $\mathcal{D}_s \in S_s$ such that $Z \in S(\mathcal{D}_s)$.

And there exist q^k different Z, so $\sum_{\mathcal{D} \in S_s} |S(\mathcal{D})| = \sum_Z \sum_{\mathcal{D} \in S_s} I(Z \in S(\mathcal{D})) \leq \sum_Z 2^{|\text{len}(Z)|} \sum_{i=0}^{[2k\epsilon]} C_{k-|\text{len}(Z)|}^i \leq \sum_Z 2^k(1-\delta) = q^k 2^k(1-\delta)$. For the last inequality, we use $\sum_{i=0}^{[2k\epsilon]} C_{k-|\text{len}(Z)|}^i < 2^{k-|\text{len}(Z)|}(1-\delta)$, which can be shown by $|\text{len}(Z)| \leq q \leq k(1-4\epsilon-\delta)$ and Lemma D.1.

This is what we want. We proved the theorem. □

A required lemma is given.

Lemma D.1. *If $\epsilon, \delta \in (0,1)$ and $k, x \in \mathbb{Z}_+$ satisfy that: $x \leq k(1-2\epsilon-\delta)$, then $2^x(\sum_{j=0}^{[k\epsilon]} C_{k-x}^j) < 2^k(1-\delta)$.*

Proof. We have

$$2^x(\textstyle\sum_{j=0}^{[k\epsilon]} C_{k-x}^j) \leq 2^x 2^{k-x} \frac{[k\epsilon]}{k-x} \leq 2^k \frac{k\epsilon}{k-x} < 2^k(1-\delta).$$

The first inequality sign uses $\sum_{j=0}^m C_n^m \leq m2^n/n$ where $m \leq n/2$, and by $x \leq k(1-2\epsilon-\delta)$, so $[k\epsilon] \leq (k-x)/2$. The third inequality sign uses the fact $x \leq k(1-2\epsilon-\delta)$. □

E Proof of Theorem 5.4

We give the proof of Theorem 5.4.

Proof. Let $\mathcal{D}_{tr} \sim \mathcal{D}^N$ and $\mathcal{D}_{tr} = \{(x_i, y_i)\}_{i=1}^N$. For any given W, let $\mathcal{F}$ be a network in $\mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n)$ and $\mathcal{F} = \sum_{i=1}^W a_i \text{ReLU}(W_i x + b_i) + c$.

Then, we consider another network F_f that is constructed in the following way:

(1): For a $v \in \{-1,1\}^N$, we say $i \in S_v$ if: $\text{ReLU}(W_i x_j + b_i) \geq 0$ for all j such that $v_j = 1$; $\text{ReLU}(W_i x_j + b_i) < 0$ for all j such that $v_j = -1$.

(2): For any $v \in \{-1,1\}^N$, if $S_v \neq \phi$, let $\mathbb{P}_v = \sum_{i \in S_v} a_i W_i / |S_v|$ and $Q_v = \sum_{i \in S_v} a_i b_i / |S_v|$.

(3): Define F_f as: $F_f(x) = \sum_{v \in \{-1,1\}^N, S_v \neq \phi} \sum_{i=1}^{|S_v|} \text{ReLU}(\mathbb{P}_v x + Q_v) + c$.

Then we have the following result:

(r1): $F_f \in \arg\min_{f \in \mathbf{H}_W^\sigma(n)} \sum_{(x,y) \in \mathcal{D}_{tr}} L(f(x), y)$.

Firstly, it is easy to see that each parameter of F_f is in $[-1,1]$, because for any v, $||\mathbb{P}_v||_\infty = ||\sum_{i \in S_v} \frac{a_i W_i}{|S_v|}||_\infty \leq \sum_{i \in S_v} \frac{||a_i W_i||_\infty}{|S_v|} \leq |S_v| \frac{1}{|S_v|} = 1$, and $||Q_v||_\infty = ||\sum_{i \in S_v} \frac{a_i b_i}{|S_v|}||_\infty \leq \sum_{i \in S_v} \frac{||a_i b_i||_\infty}{|S_v|} \leq |S_v| \frac{1}{|S_v|} = 1$.

Then, F_f has width W, because for each i, there is only one v such that $i \in S_v$, so $\sum_{v\in\{-1,1\}^N, S_v\neq\phi}\sum_{i=1}^{|S_v|} 1 = W$, which implies that F_f has width W.

Finally, there are $\mathcal{F}_f(x_i) = \mathcal{F}(x_i)$ for all $(x_i, y_i) \in \mathcal{D}_{tr}$. We just need to show that for x_1, others are similar.

There are $\mathcal{F}(x_1) = \sum_{i=1}^{W} a_i \mathrm{ReLU}(W_i x_1 + b_i) + c = \sum_{i\in[W], W_i x_1 + b_i \geq 0} a_i(W_i x_1 + b_i) + c$. Hence, letting $V1 = \{v : v \in \{-1,1\}^N, v_1 = 1\}$, then there is $F_f(x_1) = \sum_{v\in\{-1,1\}^N, S_v\neq\phi}\sum_{i=1}^{|S_v|} \mathrm{ReLU}(\mathbb{P}_v x_1 + Q_v) + c = \sum_{v\in V1, S_v\neq\phi}\sum_{i=1}^{|S_v|}(\mathbb{P}_v x_1 + Q_v) + c$.

Consider that $\{i \in [W], W_i x_1 + b_i \geq 0\} = \{i : i \in S_v, v \in V1\}$, so:

$$\begin{array}{rl} & \mathcal{F}(x_1) \\ = & \sum_{i\in[W], W_i x_1 + b_i > 0} a_i(W_i x_1 + b_i) + c \\ = & \sum_{i: i\in S_v, v\in V1} a_i(W_i x_1 + b_i) + c \\ = & \sum_{v\in V1, S_v\neq\phi}\sum_{i\in S_v} a_i(W_i x_1 + b_i) + c \\ = & \sum_{v\in V1, S_v\neq\phi} |S_v|(\mathbb{P}_v x_1 + b_v) + c \\ = & \mathcal{F}_f(x_1). \end{array}$$

By such three points and considering $\mathcal{F} \in \arg\min_{f\in\mathbf{H}^{\sigma}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$, so there are $F_f \in \arg\min_{f\in\mathbf{H}^{\sigma}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$.

(r2): $A_{\mathcal{D}}(F_f) \leq 1 - \delta$ when $N \leq W_0^{\frac{1}{n+1}}(n+1)/e$, where W_0 is defined in Theorem. This is what we want.

Firstly, we show that $|\{v : S_v \neq \phi\}| \leq max\{2^{n+1}, \frac{eN}{n+1}^{n+1}\}$, just by Lemma E.1.

Secondly, consider the network $F_{f1} = \sum_{v\in\{-1,1\}^N, S_v\neq\phi} \mathrm{ReLU}(|S_v|\mathbb{P}_v x_1 + |S_v|Q_v) + c$. By the assumption of $\mathcal{D}$ and $|\{v : S_v \neq \phi\}| \leq max\{2^{n+1}, \frac{eN}{n+1}^{n+1}\}$, then we know that, when $N \leq W_0^{\frac{1}{n+1}}(n+1)/e$, there are $A_{\mathcal{D}}(F_{f1}) \leq 1 - \delta$.

Moreover, there are $F_{f1}(x) = \sum_{v\in\{-1,1\}^N, S_v\neq\phi} \mathrm{ReLU}(|S_v|\mathbb{P}_v x + |S_v|Q_v) + c = \sum_{v\in\{1,1\}^N, S_v\neq\phi}\sum_{i=1}^{|S_v|} \mathrm{ReLU}(\mathbb{P}_v x + Q_v) + c = F_f(x)$, so $A_{\mathcal{D}}(\mathcal{F}_f) = A_{\mathcal{D}}(F_{f1}) \leq 1 - \delta$, this is what we want. □

A required lemma is given:

Lemma E.1. *For any $S = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$, let $\Pi(S) = \{(\mathrm{Sgn}(Wx_i + b))_{i=1}^n : W \in \mathbb{R}^n, b \in \mathbb{R}\}$. Then $|\Pi(S)| \leq max\{2^{n+1}, \frac{eN}{n+1}^{n+1}\}$.*

Proof. It is easy to see that $|\Pi(S)| \leq 2^N$ because $\mathrm{Sgn}(Wx_i + b) \in \{-1, 1\}$. So, when $N \leq n + 1$, it is obviously correct.

When $N > n + 1$. Consider that the VC-dim of the linear space is $n + 1$, and $\Pi(S) = \{(\mathrm{Sgn}(Wx_i + b))_{i=1}^n : W \in \mathbb{R}^n, b \in \mathbb{R}\}$ is the growth function of linear space under dataset S. So by Theorem 1 of (Sauer, 1972), we have $|\Pi(S)| \leq \sum_{i=0}^{n+1} C_N^i$.

Moreover, there are $\sum_{i=0}^{n+1} C_N^i \leq \frac{eN}{n+1}^{n+1}$ as shown in (Sauer, 1972), this is what we want. □

F Proof of Proposition 5.6

We give the proof of Proposition 5.6.

Proof. Firstly, it is easy to show that $\mathcal{D}_n$ cannot be expressed by $\mathbf{H}_W(n)$ when $W < n/2$ by Lemma F.1, so we have proved (1) of Proposition 5.6.

Let $\mathcal{D}_{tr} \sim \mathcal{D}_n^N$ and $N \leq n\delta$, for any given W, let $\mathcal{F}$ be a network in $\mathbf{M}^{\sigma}_W(\mathcal{D}_{tr}, n)$, and $\mathcal{F} = \sum_{i=1}^{W} a_i \mathrm{ReLU}(W_i x + b_i) + c$.

Now we prove (2) of Proposition 5.6. Let $\mathcal{D}_{tr} = \{(\frac{x_i}{n}\mathbf{1}, \mathbb{I}(x_i))\}_{i=1}^N$ where $x_i \in [n]$ be selected from the distribution, without loss of generality, let $x_i < x_{i+1}$ for any $i \in [N]$.

We will divide $[W]$ into several subsets based on the intersection of the plane $W_j x + b$ and the line $-\infty\mathbf{1} \to \infty\mathbf{1}$, let $[W] = \cup_{i=1}^{2N} s_i$, and:

1. For any $i \in [N-1]$: if $j \in [W]$ such that $\frac{x_i}{n} W_j\mathbf{1} + b_i < 0$ and $\frac{x_{i+1}}{n} W_j\mathbf{1} + b_j \geq 0$, then $j \in s_i$;

2. If $j \in [W]$ such that $\frac{x_i}{n} W_j\mathbf{1} + b_i < 0$ for any $i \in [N]$, then $j \in s_N$;

3. For any $i \in \{N+1, N+2, \ldots, 2N-1\}$: if $j \in [W]$ such that $\frac{x_{i-N}}{n} W_j\mathbf{1} + b_i \geq 0$ and $\frac{x_{i-N+1}}{n} W_j\mathbf{1} + b_j < 0$, then $j \in s_i$;

4. If $j \in [W]$ such that $\frac{x_i}{n} W_j\mathbf{1} + b_i \geq 0$ for any $i \in [N]$, then $j \in s_{2N}$.

Now, by such $2N$ subset, we consider another network $\mathcal{F}_f$ that is defined as:

For any $i \in [2N]$, if $S_i \neq \phi$, define $P_i = \sum_{j \in S_i} a_i W_i / |S_i|$ and $Q_i = \sum_{j \in S_i} a_i b_i / |S_i|$. Then $F_f = \sum_{i \in [2N], S_i \neq \phi} \sum_{j=1}^{|S_i|} \mathrm{ReLU}(P_i x + Q_i) + c = \sum_{i \in [2N], S_i \neq \phi} |S_i| \mathrm{ReLU}(P_i x + Q_i) + c$.

Because there is only one intersection point between a straight line and a plane, each $j \in [W]$ is only in one subset s_i. So, $\mathcal{F}_f \in \mathbf{H}_W^\sigma(n)$. Moreover, we show that $\mathcal{F}_f(x) = \mathcal{F}(x)$ for any $(x, y) \in \mathcal{D}_{tr}$, which implies $\mathcal{F}_f \in \arg\min_{f \in \mathbf{H}_W^\sigma(n)} \sum_{(x,y) \in \mathcal{D}_{tr}} L(f(x), y)$.

For any $j \in [N]$, by the definition of s_i, we know that $\frac{x_j}{n} W_i\mathbf{1} + b_i \geq 0$ if and only if $i \in \{1, 2, \ldots, j-1\} \cup \{N+j, N+j+1, \ldots, 2N\}$, so:

$$
\begin{aligned}
& \mathcal{F}_f(x_j) \\
= \; & \textstyle\sum_{i \in [2N], S_i \neq \phi} \sum_{j=1}^{|S_i|} \mathrm{ReLU}(P_i x_j + Q_i) + c \\
= \; & \textstyle\sum_{i \in \{1,2,\ldots,j-1\} \cup \{N+j, N+j+1, \ldots, 2N\}, S_i \neq \phi} \sum_{j=1}^{|S_i|} (P_i x_j + Q_i) + c \\
= \; & \textstyle\sum_{k \in \cup_{i \in \{1,2,\ldots,j-1, N+j, N+j+1, \ldots, 2N\}} s_i} (W_k x_j + b_k) + c \\
= \; & \textstyle\sum_{k \in [W]} \mathrm{ReLU}(W_k x_j + b_k) + c \\
= \; & \mathcal{F}(x_j)
\end{aligned}
$$

This is what we want. At last, by $\mathcal{F}_f = \sum_{i \in [2N], S_i \neq \phi} |S_i| \mathrm{ReLU}(P_i x + Q_i) + c$ has width at most $2N$ and Lemma F.1, and consider that $N \leq n\delta$, we have that: $A_{\mathcal{D}}(\mathcal{F}_f) \leq 0.5 + 2\delta$, this is what we want.

□

A required lemma is given below.

Lemma F.1. *If $x_1 < x_2 < x_3 < \cdots < x_N$, and x_i has label $y_i = 1$ when i is odd, or x_i has label $y_i = -1$. We consider dataset $S = \{(x_i\mathbf{1}(n), y_i)\}$, where $\mathbf{1}$ is all-one vector in $\mathbb{R}^n$. Then: For any two-layer network width M, this network can correctly classify at most to $M + \frac{N}{2}$ samples in S.*

Proof. Let $\mathcal{F} = \sum_{i=1}^M a_i \mathrm{ReLU}(W_i x + b_i) + c$. Let $W_i x + b_i$ and the line $-\infty\mathbf{1}(n) \to \infty\mathbf{1}(n)$ intersect at one point $P_i\mathbf{1}(n)$. Let $P_i \leq P_j$ if $i \leq j$. Let $P_{M+1} = \infty$.

Then it is easy to see that in the line segment $P_i\mathbf{1}(n) \to P_{i+1}\mathbf{1}(n)$, $\mathcal{F}(x)$ is a linear function. So, there is $P_{i+0.5} \in (P_i, \mathbb{P}_{i+1})$ such that $\mathcal{F}$ maintains the positive and negative polarity unchanged in $P_i\mathbf{1}(n) \to P_{i+0.5}\mathbf{1}(n)$ and $P_{i+0.5}\mathbf{1}(n) \to P_{i+1}\mathbf{1}(n)$.

So if $P_i \leq x_u < x_{u+1} < \cdots < x_{u+k} < P_{i+0.5}$, F gives the same label to $x_u\mathbf{1}(n), x_{u+1}\mathbf{1}(n), \ldots, x_{u+k}\mathbf{1}(n)$, which means that $\mathcal{F}$ can classify at most $[\frac{(k+1)+1}{2}]$ samples in them. Similar to when $P_{i+0.5} \leq x_u < x_{u+1} < \cdots < x_{u+k} < P_{i+1}$.

Let $q_i = |\{j : P_{i/2} \leq x_j < P_{i/2+0.5}\}|$ where $i \in [2M]$. Consider that each sample in S is appeared in a $P_i\mathbf{1}(n) \to P_{i+0.5}\mathbf{1}(n)$ or $P_{i+0.5}\mathbf{1}(n) \to P_{i+1}\mathbf{1}(n)$, so $\sum_{i=1}^{2M} q_i = N$.

So, the whole network can classify at most $\sum_{i=1}^{2M} [\frac{1+q_i}{2}] \leq \sum_{i=1}^{2M} \frac{1+q_i}{2} = M + \frac{N}{2}$. □

G PROOF OF PROPOSITION 5.7

Proof. Proof of (1): Let $\mathbf{1}$ be the all one vector, $\sum x = \sum_{i=1}^{n} x_i$ where x_i is the i-th weight of x. We show that $\mathcal{F} = \sigma(\mathbf{1}x - 0.5) \in \mathbf{H}_1^{\sigma}(n)$ is what we want. Because if $\sum x$ is odd, then $\sigma(\mathbf{1}x - 0.5) = \sigma(\sum x - 0.5) = sin(\pi(\sum x - 0.5)) = 1$; if $\sum x$ is even, then $\sigma(\mathbf{1}x - 0.5) = \sigma(\sum x - 0.5) = sin(\pi(\sum x - 0.5)) = -1$.

Proof of (2): we will prove it into three parts:

Part one: For any W and $\mathcal{D}_{tr} \sim \mathcal{D}_n^N$, let $\mathcal{F} \in \mathbf{M}_W^{\sigma}(\mathcal{D}_{tr}, n)$ and $\mathcal{F} = \sum_{i=1}^{W} \sigma(W_i x + b_i) + c$. Then there are: for any $(x, y) \in \mathcal{D}_{tr}$, there are $y\sigma(W_i x + b_i) = 1$ for any $i \in [W]$.

If not, without loss of generality, assume that $y\sigma(W_1 x + b_1) < 1$ for some $(x, y) \in \mathcal{D}_{tr}$. According to the proof of (1), there are W_0 and b_0 such that $y\sigma(W_0 x + b_0) = 1$ for any $(x, y) \in \mathcal{D}_{tr}$. Now we consider the network $\mathcal{F}_c(x) = \sum_{i=2}^{W} \sigma(W_i x + b_i) + \sigma(W_0 x + b_0) + c$, then we have that:

Firstly, it is easy to see that $\mathcal{F}_c \in \mathbf{H}_W^{\sigma}(n)$.

Secondly, we show that $y\mathcal{F}(x) \leq y\mathcal{F}_c(x)$ for any $(x, y) \in \mathcal{D}_{tr}$ and $y\mathcal{F}(x) < y\mathcal{F}_c(x)$ for some $(x, y) \in \mathcal{D}_{tr}$.

By the definition of $\mathcal{F}$ and $\mathcal{F}_c$, for any $(x, y) \in \mathcal{D}_{tr}$, there are $y\mathcal{F}_c(x) - y\mathcal{F}(x) = y(\sigma(W_0 x + b_0) - \sigma(W_1 x + b_1)) = 1 - y\sigma(W_1 x + b_1) \geq 0$, and by the assumption, there is a $(x, y) \in \mathcal{D}_{tr}$ such that $1 > y\sigma(W_1 x + b_1))$, then $y\mathcal{F}_c(x) - y\mathcal{F}(x) > 0$ for such $(x, y) \in \mathcal{D}_{tr}$, this is what we want.

By the above two results, and considering that $L(\mathcal{F}(x), y)$ is a strictly decreasing function about $y\mathcal{F}(x)$, there are $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) > \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}_c(x), y)$, which is contradictory to $\mathcal{F} \in \arg\min_{f\in\mathbf{H}_W^{\sigma}(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$. So we prove part one.

Part Two. For any $j \in Z$, let $x_j = \frac{j}{n}\mathbf{1}$ and $y_j = \mathbb{I}(j)$, where $\mathbb{I}(x)$ is defined in the definition of distribution $\mathcal{D}_n$. If $i_j \in \mathbb{Z}$ where $j \in [4]$ such that $i_1 - i_2$ and $i_3 - i_4$ are co-prime, then there are: if $W_0 \in [-1, 1]^n$ and $b_0 \in [-1, 1]$ such that $y_{i_j}\sigma(W_0 x_{i_j} + b_0) = 1$ for any $j \in [4]$, then $y_p\sigma(W_0 x_p + b_0) = 1$ for all $p \in Z$.

When there is $y_{i_j}\sigma(W_0 x_{i_j} + b_0) = y_{i_j} sin(\pi(W_0 x_{i_j} + b_0)) = y_{i_j} sin(\pi(< W_0, \mathbf{1} > i_j/n + b_0)) = 1$, consider that $y_{i_j} \in \{-1, 1\}$, then there is $< W_0, \mathbf{1} > i_j/n + b_0 = m_{i_j} - 0.5$ for $m_{i_j} \in Z$, moreover, m_{i_j} and i_j are same parity.

Now consider $(W_0 x_{i_1} + b_0) - (W_0 x_{i_2} + b_0)$ and $(W_0 x_{i_3} + b_0) - (W_0 x_{i_4} + b_0)$, there are $< W_0, \mathbf{1} > (i_1 - i_2)/n = m_{i_1} - m_{i_2}$ and $< W_0, \mathbf{1} > (i_3 - i_4)/n = m_{i_3} - m_{i_4}$. So, there are $\frac{i_1 - i_2}{i_3 - i_4} = \frac{m_{i_1} - m_{i_2}}{m_{i_3} - m_{i_4}}$.

By $i_1 - i_2$ and $i_3 - i_4$ are co-prime, and $|m_{i_1} - m_{i_2}| = | < W_0, \mathbf{1} > (i_1 - i_2)/n| \leq |i_1 - i_2|$, $|m_{i_3} - m_{i_5}| = | < W_0, \mathbf{1} > (i_3 - i_4)/n| \leq |i_3 - i_4|$, there are $< W_0, \mathbf{1} > /n = 1$ or $< W_0, \mathbf{1} > /n = -1$.

Hence, by $m_{i_j} - i_j =< W_0, \mathbf{1} > i_j/n + b_0 + 0.5 - i_j$ and $< W_0, \mathbf{1} > /n = 1$ or $< W_0, \mathbf{1} > /n = -1$, consider that m_{i_j} and i_j are the same parity, so $b = -0.5$.

So for any $p \in \mathbb{Z}$, there are $y_p\sigma(W_0 x_p + b_0) = y_p sin(\pi(< W_0, \mathbf{1} > p/n + b_0)) = y_p sin(\pi(p - 0.5)) = 1$, this is what we want.

Part Three, if $\mathcal{D}_{tr} \sim \mathcal{D}_n^N$ and $N \geq 4\frac{\ln(\delta/2)}{\ln(0.5+1/n)}$, with probability $1 - \delta$, there are four samples (x_i, y_i) where $i \in [4]$ in $\mathcal{D}_{tr}$, such that $x_i = \frac{m_i}{n}\mathbf{1}$, $m_1 - m_2$ and $m_3 - m_4$ are co-prime.

By the definition of $\mathcal{D}_n$, it is equivalent to: repeatable randomly select $N \geq 4\frac{\ln(\delta/2)}{\ln(0.5+1/n)}$ points from $[n]$, with probability $1 - \delta$, there are four samples m_i such that $m_1 - m_2$ and $m_3 - m_4$ are co-prime.

By Lemma G.1, when $N \geq 4\frac{\ln(\delta/2)}{\ln(0.5+1/n)}$, with probability at least $1 - (0.5 + 1/n)^{\frac{\ln(\delta/2)}{\ln(0.5+1/n)}}/(0.5 + 1/n) = 1 - \delta/(1 + 2/n) \geq 1 - \delta$. This is what we want.

Part Four, we prove the result.

Let $\mathcal{D}_{tr} \sim \mathcal{D}_n^N$. For any W, let $\mathcal{F} \in \arg\min_{f\in\mathbf{H}_W^\sigma(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x),y)$ and $\mathcal{F} = \sum_{i=1}^W \sigma(W_i x + b_i) + c$.

Firstly, with probability $1-\delta$, there are four samples in $\mathcal{D}_{tr}$ satisfying part three. Then, according to part one, there are $y\sigma(W_i x + b_i) = 1$ for such four samples. Finally, in part two, there are $y\sigma(W_i x + b_i) = y\sigma(\sum x) = 1$ for any $(x,y) \sim \mathcal{D}_n$. So, $y\mathcal{F}(x) \geq W - 1 > 0$ for any $(x,y) \sim \mathcal{D}_n$, we prove the result. □

A required lemma is given.

Lemma G.1. *Randomly select N points from $[n]$, where $n \geq 3$ and $N \geq 4$. With probability $1-(0.5+1/n)^{N/4-1}$, there are four samples m_i such that $m_1 - m_2$ and $m_3 - m_4$ are co-prime.*

Proof. Firstly, we consider the situation that $N = 4$, let $\{m_i\}_{i=1}^4$ are the selected number. Then we have

$$\begin{aligned}
& P((m_1 - m_2, m_3 - m_4) = 1) \\
= \; & P(m_1 - m_2 \neq 0, m_3 - m_4 \neq 0) - P((m_1 - m_2, m_3 - m_4) \neq 1, m_1 - m_2 \neq 0, m_3 - m_4 \neq 0) \\
= \; & (1-1/n)^2(1 - P((|m_1 - m_2|, |m_3 - m_4|) \neq 1 | m_1 - m_2 \neq 0, m_3 - m_4 \neq 0)) \\
\geq \; & (1-1/n)^2(1 - \textstyle\sum_{q\in Prime} P(q|(|m_1 - m_2|, |m_3 - m_4|) | m_1 - m_2 \neq 0, m_3 - m_4 \neq 0)) \\
\geq \; & (1-1/n)^2(1 - \textstyle\sum_{q\in Prime} \frac{1}{q^2}) \\
\geq \; & 0.5(1-1/n)^2 \geq 0.5 - 1/n
\end{aligned}$$

where $Prime$ is the set of all primes. For the second inequality sign, we use

$$\begin{aligned}
& P(q | m_1 - m_2 \, | m_1 - m_2 \neq 0) \\
= \; & \textstyle\sum_{i=1}^{n-1} P(q | i, i = |m_1 - m_2| \, | m_1 - m_2 \neq 0) \\
= \; & [(n-1)/q] * \frac{1}{n-1} \\
\leq \; & 1/q.
\end{aligned}$$

Similar for $m_3 - m_4$. For the last inequality sign, we use $P(2) = \sum_{i\in Prime} \frac{1}{i^2} < 0.5$, where P is Riemann function.

So, when we select N samples, it contains $[N/4] > N/4 - 1$ pairs of four independent samples randomly selected. So, with probability $1-(0.5+1/n)^{N/4-1}$, there are four samples m_i such that $m_1 - m_2$ and $m_3 - m_4$ are co-prime. □

H Proof of Theorem 6.2

Now, we prove Theorem 6.2.

Proof. we prove the proposition into three parts.

Part One, with probability $1-2\delta$ of $\mathcal{D}_{tr} \sim D^{N_0}$, there are $\mathbb{E}_{x\sim\mathcal{D}}[y\mathcal{F}(x)] \geq c_0 N_0 [\frac{W}{W_0+1}] - 2\frac{L_p(W+1)(n+1)(\sqrt{4\log(4)}+\sqrt{2\log(2n)})}{\sqrt{N_0}} - 6\mathcal{F}_{\max}\sqrt{\frac{\ln(2/\delta)}{2N_0}}$ for all $\mathcal{F} \in \mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n)$, where $\mathcal{F}_{\max} = \max_{x+\delta\in[0,1]^n} |\mathcal{F}(x+\delta)|$.

Firstly, we show that there are $\sum_{(x,y)\in\mathcal{D}_{tr}} y\mathcal{F}(x) \geq N_0[\frac{W}{W_0+1}]c_0$ for all $\mathcal{F} \in \mathbf{M}_W^\sigma(\mathcal{D}_{tr}, n)$ when $\mathcal{D}_{tr} \sim \mathcal{D}^{N_0}$ satisfies the conditions of the proposition.

Because $\mathcal{D}_{tr}$ can be expressed in the network space $\mathbf{H}_{W_0}^\sigma(n)$ with confidence c_0, there is a network $\mathcal{F}_0 = \sum_{i=1}^{W_0} a_i\sigma(W_i x + b_i) + c$ such that $y\mathcal{F}(x) \geq c_0$ for all $(x,y) \in \mathcal{D}_{tr}$. Moreover, we can write such networks as: $\mathcal{F}_0 = \sum_{i=1}^{W_0+1} a_i\sigma(W_i x + b_i)$, where $a_{W_0+1} = \text{Sgn}(c)$, $W_{W_0+1} = 0$, $b_{W_0+1} = |c|$.

Now, we consider the following network in $\mathbf{H}_W^\sigma(n)$:

$$\mathcal{F}_W = \sum_{i=1}^{(W_0+1)[\frac{W}{W_0+1}]} a_{i\%(W_0+1)}\sigma(W_{i\%(W_0+1)}x + b_{i\%(W_0+1)}),$$

Here, we stipulate that $i\%(W_0+1) = W_0+1$ when $W_0+1|i$. Then we have $\mathcal{F}_W(x) = [\frac{W}{W_0+1}]\mathcal{F}_0(x)$ and $\mathcal{F}_W(x) \in \mathbf{H}_W^{\sigma}(n)$. Moreover, there are $y\mathcal{F}_W(x) = y[\frac{W}{W_0+1}]\mathcal{F}_0(x) \geq [\frac{W}{W_0+1}]c_0$ for all $(x,y) \in \mathcal{D}_{tr}$, so $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}_W(x), y) \leq N_0 \ln(1 + e^{-c_0[\frac{W}{W_0+1}]})$.

Then, because $\ln 1 + e^x$ is a convex function, so that:

$$\begin{aligned} & N_0 \ln 1 + e^{-\frac{\sum_{(x,y)\in\mathcal{D}_{tr}} y\mathcal{F}(x)}{N}} \\ \leq & \sum_{(x,y)\in\mathcal{D}_{tr}} \ln 1 + e^{-y\mathcal{F}(x)} \\ = & \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \\ \leq & \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}_W(x), y) \\ \leq & N_0 \ln(1 + e^{-c_0[\frac{W}{W_0+1}]}) \end{aligned}$$

So $\sum_{(x,y)\in\mathcal{D}_{tr}} y\mathcal{F}(x) \geq c_0 N[\frac{W}{W_0+1}]$.

Hence, by Lemma B.4 and Theorem B.6, with probability $1-\delta$ of $\mathcal{D}_{tr}$, there are:

$$|\mathbb{E}_{x\sim\mathcal{D}}[\mathcal{F}(x)] - \sum_{x\in\mathcal{D}_{tr}} \frac{\mathcal{F}(x)}{N_0}| \leq 2\frac{L_p(W+1)(n+1)(\sqrt{4\log(4)} + \sqrt{2\log(2n)})}{\sqrt{N_0}} + 6\mathcal{F}_{\max}\sqrt{\frac{\ln(2/\delta)}{2N_0}},$$

for all $\mathcal{F} \in \mathbf{H}_W(n)$.

Finally, combining the above two results, with probability $1-2\delta$, there is $\mathbb{E}_{x\sim\mathcal{D}}[\mathcal{F}(x)] \geq c_0 N_0[\frac{W}{W_0+1}] - 2\frac{L_p(W+1)(n+1)(\sqrt{4\log(4)}+\sqrt{2\log(2n)})}{\sqrt{N_0}} - 6\mathcal{F}_{\max}\sqrt{\frac{\ln(2/\delta)}{2N_0}}$.

Part Two, there is an upper bound of $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\min_{||\delta||\leq\epsilon} y\mathcal{F}(x+\delta)]$, if $\mathcal{D}_{tr}$ satisfies Part One.

For any $\mathcal{F} \in \mathbf{H}_W(n)$, we can write $\mathcal{F} = \sum_{i=0}^{\lceil\frac{W}{W_0}\rceil-1}\sum_{j=1}^{W_0} \mathrm{ReLU}(W_{iW_0+j}x + b_{iW_0+j}) + c$, which is a representation of the sum of $\lceil\frac{W}{W_0}\rceil$ small networks with width of W_0. So by part one and by the assumption in the theorem, with probability $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, there is a $\mathcal{D}_r \in R(\mathcal{D}_{tr}, \epsilon)$ such that $\sum_{(x,y)\in\mathcal{D}_r} y\mathcal{F}_1(x) \leq 2N_0c_1$ for all $\mathcal{F}_1 \in \mathbf{H}_{W_0}(n)$. Then we have $\sum_{(x,y)\in\mathcal{D}_r} y\mathcal{F}(x) \leq 2N_0c_1\lceil\frac{W}{W_0}\rceil$, by the definition of $\mathcal{D}_r$, which implies that $\sum_{(x,y)\in\mathcal{D}_{tr}} \min_{||\delta||\leq\epsilon} y\mathcal{F}(x+\delta) + y\mathcal{F}(x) \leq 2N_0c_1\lceil\frac{W}{W_0}\rceil$.

And then, by McDiarmid inequality, with probability $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^{N_0}$, there are $|\mathbb{E}_{(x,y)\sim\mathcal{D}}[\min_{||\delta||\leq\epsilon} y\mathcal{F}(x+\delta) + y\mathcal{F}(x)] - \frac{1}{N_0}\sum_{(x,y)\in\mathcal{D}_{tr}} \min_{||\delta||\leq\epsilon} y\mathcal{F}(x+\delta) + y\mathcal{F}(x)| \leq 2\mathcal{F}_{\max}\sqrt{\frac{\ln 1/\delta}{2N_0}}$. So if there are $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\min_{||\delta||\leq\epsilon} y\mathcal{F}(x+\delta) + y\mathcal{F}(x)] > 2c_1\lceil\frac{W}{W_0}\rceil + 2\mathcal{F}_{\max}\sqrt{\frac{\ln 1/\delta}{2N_0}}$, according to McDiarmid inequality, with probability $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^{N_0}$, $\sum_{(x,y)\in\mathcal{D}_{tr}} \min_{||\delta||\leq\epsilon} y\mathcal{F}(x+\delta) + y\mathcal{F}(x) > 2N_0c_1\lceil\frac{W}{W_0}\rceil$ stand, which is a contradiction with the above result.

So there must be $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\min_{||\delta||\leq\epsilon} y\mathcal{F}(x+\delta) + y\mathcal{F}(x)] \leq 2c_1\lceil\frac{W}{W_0}\rceil + 2\mathcal{F}_{\max}\sqrt{\frac{\ln 1/\delta}{2N_0}}$. Finally, considering the result in Part one, we have that:

$$\begin{aligned} & \mathbb{E}_{(x,y)\sim\mathcal{D}}[\min_{||\delta||\leq\epsilon} y\mathcal{F}(x+\delta)] \\ \leq & 2c_1\lceil\frac{W}{W_0}\rceil - c_0[\frac{W}{W_0+1}] + 2\frac{L_p(W+1)(n+1)(\sqrt{4\log(4)}+\sqrt{2\log(2n)})}{\sqrt{N_0}} + 8\mathcal{F}_{\max}\sqrt{\frac{\ln(2/\delta)}{2N_0}} \end{aligned}$$

Part Three, Now we can get the result.

By Lemma H.1 and part two, there are $Rob_{\mathcal{D},\epsilon}(\mathcal{F}) \leq 1 - \frac{c_0[\frac{W}{W_0+1}]-2c_1\lceil\frac{W}{W_0}\rceil}{\mathcal{F}_{\max}} + 8\sqrt{\frac{\ln 2/\delta}{2N_0}} + 2\frac{L_p(W+1)(n+1)(\sqrt{4\log(4)}+\sqrt{2\log(2n)})}{\sqrt{N_0}\mathcal{F}_{\max}}$, and we consider that each parameter of $\mathcal{F}$ is not greater than 1 and Lipschitz constant of σ is not more than L_p, so $\mathcal{F}_{\max} = \max_{x+\delta\in[0,1]^n} |\mathcal{F}(x+\delta)| = \max_{x\in[0,1]^n} |\mathcal{F}(x)| \leq L_pW(n+1)+1$.

Let $T = [\frac{W}{W_0+1}]$, by $L_p, n, W_0 \geq 1$ and $W \geq W_0+1$, there are:

$$\begin{aligned}
& \frac{c_0[\frac{W}{W_0+1}]-2c_1\lceil\frac{W}{W_0}\rceil}{L_pW(n+1)+1} \\
\geq\ & \frac{c_0T-2c_1(\frac{(T+1)(W_0+1)}{W_0}+1)}{L_p(T+1)(W_0+1)(n+1)+1} \\
=\ & \frac{c_0T-2c_1T}{L_p(T+1)(W_0+1)(n+1)+1} - \frac{4c_1}{L_p(T+1)(W_0+1)(n+1)+1} - \frac{2c_1}{L_pW_0(W_0+1)(n+1)+1} \\
\geq\ & \frac{c_0-2c_1}{8L_pW_0n} - \frac{4c_1}{L_pW_0n}(\frac{1}{W/W_0}+\frac{1}{W_0})
\end{aligned}$$

and

$$\begin{aligned}
& \frac{L_p(W+1)(n+1)(\sqrt{4\log(4)}+\sqrt{2\log(2n)})}{\sqrt{N_0}\mathcal{F}_{\max}} \\
\geq\ & \frac{L_p(W+1)(n+1)(\sqrt{4\log(4)}+\sqrt{2\log(2n)})}{2\sqrt{N_0}L_p(W+1)(n+1)} \\
=\ & 2\frac{\sqrt{4\log(4)}+\sqrt{2\log(2n)}}{\sqrt{N_0}}
\end{aligned}$$

So, there are $\mathrm{Rob}_{\mathcal{D},\epsilon}(\mathcal{F}) \leq 1-\frac{c_0-2c_1}{8L_pW_0n}+\frac{4c_1}{L_pW_0n}(\frac{1}{W/W_0}+\frac{1}{W_0})+2\sqrt{\frac{\ln 2/\delta}{2N_0}}+4\frac{\sqrt{4\log(4)}+\sqrt{2\log(2n)}}{\sqrt{N_0}}$. Merge some items and ignore constants, this is what we want. □

A required lemma is given below.

Lemma H.1. *If $\mathcal{F}:\mathbb{R}^n\to\mathbb{R}$ and the distribution $\mathcal{D}\in[0,1]^n\times\{-1,1\}$ satisfy $\mathbb{E}_{(x,y)\sim\mathcal{D}}[y\mathcal{F}(x)]\leq A$ and $\max_{x\in[0,1]^n}|\mathcal{F}(x)|\leq B$, then $A_{\mathcal{D}}(\mathcal{F})\leq 1+\frac{A}{B}$.*

Proof. There are $\mathbb{E}_{(x,y)\sim\mathcal{D}}[y\mathcal{F}(x)] \geq -(\max_{x\in[0,1]^n}|\mathcal{F}(x)|)\mathbb{P}_{(x,y)\sim\mathcal{D}}(y\neq \mathrm{Sgn}(\mathcal{F}(x))) = -B(1-A_{\mathcal{D}}(\mathcal{F}))$, so $A\geq -B+BA_{\mathcal{D}}(\mathcal{F})$, that is, $A_{\mathcal{D}}(\mathcal{F})\leq 1-\frac{A}{B}$. □

I Proof of Proposition 6.5

Proof. We take a $c>0$ such that $\ln(1+e^{-c})\geq \ln 2-\ln 2/800$, $1-(1/e)^{4c}<0.1$. Then take an n such that $\ln(1+e^{-n/2+2c})<\ln 2/2$. Let N satisfy $(\frac{4(n+1)(\sqrt{5\log(4)}+\sqrt{2log2n})}{\sqrt{98N/200}}+6(n+2)\sqrt{\frac{\ln(2/\delta)}{2N}})<\ln 2/800$.

We consider the following distribution $\mathcal{D}$:

(c1): Let $s_1=\{(x,1):x\in[0,1],\sum x=n/2+c,||x||_{-\infty}\geq 2c/n\}$, $||x||_{-\infty}$ mean the minimum of the weight of $|x|$; $s_2=\{(x,-1):x\in[0,1],\sum x=n/2-c,||x||_\infty\leq 1-2c/n\}$; $s_3=\{(x,-1):x\in[0,1],\sum x=n-c\}$;

(c2): $\mathbb{P}_{(x,y)\sim\mathcal{D}}(\sum x=n-c)=1/100$, and $\mathcal{D}$ is a uniform distribution in s_3;

(c3): $\mathbb{P}_{(x,y)\sim\mathcal{D}}(\sum x=n/2+c)=\mathbb{P}_{(x,y)\sim\mathcal{D}}(\sum x=n/2-c)=99/200$, and $\mathcal{D}$ is a uniform distribution in $s_1\cup s_2$.

Let $W_0=1$, then we show this distribution and W_0 are what we want.

(1) in Theorem: Let $\mathcal{F}_1=Relu(\mathbf{1}x)-c/2\in\mathbf{H}_1(n)$. Then $\mathcal{F}_1(x)>0$ for all x such that $\sum x=c$, and $\mathcal{F}_1(x)<0$ for all x such that $\sum x=-c$, so $A_{\mathcal{D}}(\mathcal{F}_1)\geq 0.99$.

(2) in Theorem: We use the following parts to show the (2) in the Theorem.

Part One. With probability at least $1-3e^{-2N/200^2}$ of $\mathcal{D}_{tr}\sim\mathcal{D}^N$, there are at least $N/200$ points in $\mathcal{D}_{tr}\cap s_3$, and at least $98/200N$ points with label 1 in $\mathcal{D}_{tr}$, at least $98/200N$ points with label -1 in $\mathcal{D}_{tr}$.

Using the Hoeffding inequality and $\mathbb{P}_{(x,y)\sim\mathcal{D}}(\sum x=n-c)=1/100$, we know that with probability at least $1-e^{-2N/200^2}$ of $\mathcal{D}_{tr}$, there are at least $N/200$ points in s_3. Using also the Hoeffding inequality and $\mathbb{P}_{(x,y)\sim\mathcal{D}}(y=1)=99/200$, we know that with probability at least $1-e^{-2N(99/200-98/200)^2}$ of $\mathcal{D}_{tr}$, there are at least $98/200N$ points with label 1 in $\mathcal{D}_{tr}$; similar, with probability at least

$1 - e^{-2N(101/200-98/200)^2}$ of $\mathcal{D}_{tr}$, there are at least $98/200N$ points with label -1 in $\mathcal{D}_{tr}$. Adding them, we get the result.

Part Two. For a $\mathcal{D}_{tr}$ that satisfies Part One, if $\mathcal{F} \in \arg\min_{f \in \mathbf{H}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$, then there is $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq \frac{199\ln 2 + \ln(1+e^{-n/2+2c})}{200} N$.

We just consider the following network $\mathcal{F}_1 \in \mathbf{H}_1(n)$: $\mathcal{F}_1 = -\mathrm{ReLU}(\mathbf{1}x - (n/2 + c))$, then $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}_1(x), y) = \ln 2|\mathcal{D}_{tr}/s_3| + \ln 1 + e^{-n/2+2c}|\mathcal{D}_{tr} \cap s_3| \leq \frac{199\ln 2 + \ln(1+e^{-n/2+2c})}{200}$. Hence, for any $\mathcal{F} \in \arg\min_{f\in\mathbf{H}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$, there must be $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}_1(x), y) \leq \frac{199\ln 2 + \ln(1+e^{-n/2+2c})}{200} N$, which is what we want.

Part Three. If $\mathcal{F} \in \mathbf{H}_1(n)$ such that $\mathcal{F}(x) \geq 0$ for all $(x, -1) \in s_3$. Then $\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\mathcal{F}(x), y)] \geq 99/100 \ln 1 + e^{-c} + 1/100 \ln 2$.

Consider that for any $(x_1, 1) \in s_1$, there must be $(x_1 - 2c\mathbf{1}/n, -1) \in s_2$; on the other hand, if $(x_2, -1) \in s_2$, there must be $(x_2 + 2c\mathbf{1}/n, 1) \in s_1$. So we can match the points in s_1 and s_2 one by one by adding or subtracting a vector $2c\mathbf{1}/n$.

Moreover, for any $x \in [0, 1]$ and $x \in \mathbf{H}_1(n)$, there are $|\mathcal{F}(x) - \mathcal{F}(x - 2c\mathbf{1}/n)| \leq 2c$, which implies $L(\mathcal{F}(x), 1) + L(\mathcal{F}(x - 2c\mathbf{1}/n), -1) = \ln(1 + e^{-\mathcal{F}(x)}) + \ln(1 + e^{\mathcal{F}(x-2c\mathbf{1}/n)}) \geq 2\ln 1 + e^{-c}$. So for a $(x_1, 1) \in s_1$ and $(x_2, -1) \in s_2$ where $x_2 = x_1 - 2c\mathbf{1}/n$, there must be $L(\mathcal{F}(x_1), 1) + L(\mathcal{F}(x_2), -1) \geq 2\ln 1 + e^{-c}$.

Hence, by $\mathcal{F}(x) > 0$ for all $(x, -1) \in s_3$, $\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\mathcal{F}(x), y)] \geq 99/200 \ln(1 + e^{-c}) + \ln 2/100$.

Part Four. For any network $\mathcal{F} \in \mathbf{H}_1(n)$ such that $\mathcal{F}(x) < 0$ for a $x \in s_3$, then $A_{\mathcal{D}}(\mathcal{F}) < 60\%$.

Firstly, we show that if z_1, z_2, z_3 are collinear, without loss of generality, assuming z_2 is between z_1 and z_3, then $\mathcal{F}(z_1) \geq \mathcal{F}(z_2) \geq \mathcal{F}(z_3)$ or $\mathcal{F}(z_1) \leq \mathcal{F}(z_2) \leq \mathcal{F}(z_3)$. Consider that z_1, z_2, z_3 are collinear, so $z_2 = \lambda z_1 + (1 - \lambda)z_3$ for some $\lambda \in (0, 1)$. So let $f(k) = \mathrm{ReLU}(k(Wz_1 + b) + (1 - k)(Wz_3 + b))$, there are $f(0) = \mathrm{ReLU}(Wz_3 + b)$, $f(1) = \mathrm{ReLU}(Wz_1 + b)$ and $f(\lambda) = \mathrm{ReLU}(\lambda(Wz_1 + b) + (1 - \lambda)(Wz_3 + b)) = \mathrm{ReLU}(Wz_2 + b)$. Consider that $\mathrm{ReLU}(\cdot)$ is a monotonic function, so that $f(k)$ is also an monotonic function about $k \in \mathbb{R}$, so we get the result.

Secondly, for any $(z, -1) \in s_2$, let x_z satisfy: $(x_z, 1) \in s_1$ and x, x_z, z are collinear. Then we have that:

(1): For any $(z, -1) \in s_2$, $\mathcal{F}$ must give the wrong label to x_z or z. If not, there are $\mathcal{F}(x) < 0$, $\mathcal{F}(x_z) > 0$ and $\mathcal{F}(z) < 0$. By the above result, it is not possible.

(2): Let $S = \{x_z : (z, 1) \in s_2\} \subset s_1$, then $\mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in S | x \in s_1) \geq (1 - 4c/n)^{n-1}$. Because for any $(z, 1) \in s_2$, $\frac{||x-x_z||_2}{||x-z||_2} = \frac{\sum(x-x_z)}{\sum(x-z)} = \frac{n/2-2c}{n/2}$, which is a constant value, where $\sum x$ means the sum of the weights of x, so S is a proportional scaling of s_1 with the ratio $\frac{n-4c}{n}$, we get the result.

So, there are: $A_{\mathcal{D}}(\mathcal{F}) \leq max\{\mathbb{P}_{(x,y)\sim\mathcal{D}}((x, y) \in s_2), \mathbb{P}_{(x,y)\sim\mathcal{D}}((x, y) \in S)\} + \mathbb{P}_{(x,y)\sim\mathcal{D}}((x, y) \in s_3) + \mathbb{P}_{(x,y)\sim\mathcal{D}}(s_1/S) \leq \frac{101+99(1-(1-4c/n)^{n-1})}{200} \leq 101/200 + 99/200 * (1 - (1/e)^{4c}) \leq 0.6$, use the definition of c.

Part Five. Prove the Theorem.

We show that with probability $1 - 3e^{-2N/200^2} - \delta$ of $\mathcal{D}_{tr}$, for any $\mathcal{F} \in \arg\min_{f\in\mathbf{H}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$, $\mathcal{F}$ must give the correct label to some points in s_3. Then by part four, we can get the result.

By part one, with probability at least $1 - 3e^{-2N/200^2}$ of $\mathcal{D}_{tr}$, there are at least $N/200$ points in $\mathcal{D}_{tr} \cap s_3$, and at least $98N/200(98N/200)$ points has label 1(-1). Hence, by Lemma I.1 and Theorem 4.8, we know that, with probability $1 - \delta$ of $\mathcal{D}_{tr}$, there are $|\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y)/N - \mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\mathcal{F}(x), y)]| \geq \frac{4(n+1)(\sqrt{5\log(4)}+\sqrt{2log2n})}{\sqrt{98N/200}} + 6(n + 2)\sqrt{\frac{\ln(2/\delta)}{2N}})$. So, with probability $1 - 3e^{-2N/200^2} - \delta$, $\mathcal{D}_{tr}$ satisfies the above two conditions.

For such a $\mathcal{D}_{tr}$, assume that $\mathcal{F} \in \arg\min_{f \in \mathbf{H}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(f(x), y)$, and $\mathcal{F}$ must give the correct label to some points in s_3.

If not, by part two, we know that $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \leq \frac{199\ln 2+\ln(1+e^{-n/2+2c})}{200} N$.

Then, by part three, $\mathbb{E}_{(x,y)\sim\mathcal{D}} L(\mathcal{F}(x), y) \geq 99/100 \ln 1 + e^{-c} + 1/100 \ln 2$. Hence, by the definition of $\mathcal{D}_{tr}$, there are $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y) \geq N(99/100 \ln 1 + e^{-c} + 1/100 \ln 2) - N(\frac{4(n+1)(\sqrt{5\log(4)}+\sqrt{2log2n})}{\sqrt{98N/200}} - 6(n+2)\sqrt{\frac{\ln(2/\delta)}{2N}})$.

By the definition of c, n and N, there are $\sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y)/N \geq (99/100 \ln 1 + e^{-c} + 1/100 \ln 2) - (\frac{4(n+1)(\sqrt{5\log(4)}+\sqrt{2log2n})}{\sqrt{98N/200}} + 6(n+2)\sqrt{\frac{\ln(2/\delta)}{2N}}) \geq \frac{199.5\ln 2}{200} > \frac{199\ln 2+\ln(1+e^{-n/2+2c})}{200} \geq \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y)/N$, which leads to contradiction. And we prove the result. □

A required lemma is given below.

Lemma I.1. *For any given $D = \{(x_i, y_i)\}_{i=1}^N$, if there are at least K samples have label 1 in it and there are at least K samples have label -1 in it, then there are:*

$$\mathbb{E}_{\sigma_i}[\max_{\mathcal{F}\in\mathbf{H}_1(n)} \frac{1}{N}\sum_{i=1}^{N} \sigma_i L(\mathcal{F}(x_i), y_i)] \leq \frac{4(n+1)(\sqrt{5\log(4)} + \sqrt{2\log(2n)})}{\sqrt{K}},$$

where σ_i are i.i.d and $P(\sigma_i = 1) = P(\sigma_i = -1) = 0.5$.

Proof. We have

$$\begin{aligned} & \mathbb{E}_{\sigma_i}[\max_{\mathcal{F}\in\mathbf{H}_1(n)} \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \sigma_i L(\mathcal{F}(x_i), y_i)] \\ = \; & \mathbb{E}_{\sigma_i}[\max_{\mathcal{F}\in\mathbf{H}_1(n)} \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \sigma_i \ln 1 + e^{y_i\mathcal{F}(x_i)}] \\ \leq \; & \mathbb{E}_{\sigma_i}[\max_{\mathcal{F}\in\mathbf{H}_1(n)} \tfrac{1}{|D_1|}\textstyle\sum_{x\in D_1} \sigma_i \ln 1 + e^{\mathcal{F}(x)}] + \mathbb{E}_{\sigma_i}[\max_{\mathcal{F}\in\mathbf{H}_1(n)} \tfrac{1}{|D_2|}\textstyle\sum_{x\in D_2} \sigma_i \ln 1 + e^{\mathcal{F}(x)}] \end{aligned}$$

Hence, see $2\ln(1+e^x)$ as an activation of the second layer, and the output layer is $\mathcal{F}_2(x) = x/2$. By Lemma B.4, we have $\mathbb{E}_{\sigma_i}[\max_{\mathcal{F}\in\mathbf{H}_1(n)} \frac{1}{|D_1|}\sum_{x\in D_1} \sigma_i \ln 1 + e^{\mathcal{F}(x)}] \leq \frac{2(n+1)(\sqrt{5\log(4)}+\sqrt{2\log(2n)})}{\sqrt{|D_1|}}$. Similar for an other part, so we get the result. □

J Proof of Theorem 6.8

We give the proof of when loss function L_b satisfies (1) in definition 6.7 at first.

Proof. We first define some symbols.

Let the loss function L_b be a bad loss function that satisfies (1) in Definition 6.7. Let $L_b(z_1, 1) = \min_{x\in\mathbb{R}} L_b(x, 1)$ and $L_b(z_{-1}, -1) = \min_{x\in\mathbb{R}} L_b(x, -1)$, assume $|z_1| + |z_{-1}| = z$. For any given $x \in \mathbb{R}^n$, let $x_t = (x_2, x_3, \ldots, x_n) \in \mathbb{R}^{n-1}$, where x_i is the i-the weight of x; let $x^t = (0, x_1, x_2, x_3, \ldots, x_n) \in \mathbb{R}^{n+1}$.

Then we prove the Theorem in three parts:

Part One: We construct the following distribution $\mathcal{D}_b \in [0,1]^n \times \{-1, 1\}$:

(1): $\mathcal{D}_b$ is defined on $\{x : x \in [0,1]^n, 0.6 \leq x_1 \; or \; x_1 \leq 0.4\} \times \{-1, 1\}$, where x_1 is the first weight of x.

(2): x has label 1 if and only if $x_1 \geq 0.6$, or x has label -1.

(3): The marginal distribution about x of $\mathcal{D}_b$ is an uniform distribution.

Part Two: For any $\mathcal{D}_{tr} \sim \mathcal{D}_b^N$, we consider the following network $\mathcal{F}_{\mathcal{D}_{tr}}$.

Let $\mathcal{D}_{tr-t} = \{(x_t, y) \| (x, y) \in \mathcal{D}_{tr}\}$. By Lemma J.2, with probability 0.99, there is a $\mathcal{F}_t$ with width W not greater than $O(zN^5n^2)$ such that: if $(x_t, 1) \in \mathcal{D}_{tr-t}$, there are $\mathcal{F}_t(x_t) = z_1$; if $(x_t, -1) \in \mathcal{D}_{tr-t}$, there are $\mathcal{F}_t(x_t) = z_{-1}$. Let $\mathcal{F}_t(x) = \sum_{i=1}^{W} a_i \mathrm{ReLU}(W_i x + b_i) + c$.

Then, we construct $\mathcal{F}_{\mathcal{D}_{tr}} : \mathbb{R}^n \to \mathbb{R}$ as $\mathcal{F} = \sum_{i=1}^{W} a_i \mathrm{ReLU}(W_i^t x + b_i) + c$.

Part Three: We prove the Theorem.

For any $\mathcal{D}_{tr} \sim \mathcal{D}_b^N$, we consider the network $\mathcal{F}_{\mathcal{D}_{tr}}$ mentioned in part two. Firstly, we show that $\mathcal{F}_{\mathcal{D}_{tr}}(x) \in \arg\min_{\mathcal{F} \in \mathcal{H}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L(\mathcal{F}(x), y)$. Because $\mathcal{F}_{\mathcal{D}_{tr}}(x) = \mathcal{F}_t(x_t) = z_1$ when $(x, 1) \in \mathcal{D}_{tr}$ and $\mathcal{F}_{\mathcal{D}_{tr}}(x) = \mathcal{F}_t(x_t) = z_{-1}$ when $(x, -1) \in \mathcal{D}_{tr}$. So $L(\mathcal{F}_{\mathcal{D}_{tr}}(x), y)$ reaches the minimum value for any $(x, y) \in \mathcal{D}_{tr}$, which implies $\mathcal{F}_{\mathcal{D}_{tr}} \in \arg\min_{\mathcal{F} \in \mathcal{H}_W(n)}$.

Secondly, there are $A_{\mathcal{D}}(\mathcal{F}_{\mathcal{D}_{tr}}(x)) = 0.5$. If $A_{\mathcal{D}}(\mathcal{F}_{\mathcal{D}_{tr}}(x)) > 0.5$, then there must be a pair of $(x_1, 1)$ and $(x_2, -1)$ in distribution $\mathcal{D}_b$ such that $(x_1)_t = (x_2)_t$ and $\mathcal{F}_{\mathcal{D}_{tr}}(x)$ give the correct label to x_1 and x_2. But it is easy to see that $\mathcal{F}_{\mathcal{D}_{tr}}(x) = \mathcal{F}_t(x_t)$ where $\mathcal{F}_t$ is mentioned in part two, so, $\mathcal{F}_{\mathcal{D}_{tr}}(x_1) = \mathcal{F}_t(x)((x_1)_t) = \mathcal{F}_t(x)((x_2)_t) = \mathcal{F}_{\mathcal{D}_{tr}}(x_2)$, which is in contradiction to $\mathcal{F}_{\mathcal{D}_{tr}}(x)$ gives the correct label to x_1 and x_2. This is what we want.

□

Some required lemmas are given.

Lemma J.1. *For any $v \in \mathbb{R}^n$ and $T \geq 1$, let $u \in \mathbb{R}^n$ be uniformly randomly sampled from the hypersphere S^{n-1}. Then we have $\mathbb{P}(|\langle u, v\rangle| < \frac{||v||_2}{T}\sqrt{\frac{8}{n\pi}}) < \frac{2}{T}$.*

This is Lemma 13 in (Park et al., 2021).

Lemma J.2. *For any N points $\{x_i\}_{i=1}^N$ randomly selected in $[0, 1]^n$, and any N given point $\{y_i\}_{i=1}^N$ in $[-a, a]$. With probability 0.99 of $\{x_i\}_{i=1}^N$, there is a network $\mathcal{F}$ with width not more than $O(aN^5n^2)$ and $\mathcal{F}(x_i) = y_i$.*

Proof. **Part One:** First, we show that with probability 0.99, there is $||x_i - x_j||_2 \geq \frac{0.01}{2N^2\sqrt{n}}$ for all pairs i, j.

For any $i, j \in \mathbb{N}$ and $\epsilon > 0$, there are:

$$\begin{aligned} & P(||x_i - x_j||_2 \geq \epsilon) \\ = \; & P(\sum_{k=1}^n ((x_i)_k - (x_j)_k)^2 \geq \epsilon^2) \\ \geq \; & \Pi_{k=1}^n P(((x_i)_k - (x_j)_k)^2 \geq \epsilon^2/n) \\ \geq \; & \Pi_{k=1}^n (1 - \frac{2\epsilon}{\sqrt{n}}) \\ \geq \; & 1 - 2\epsilon\sqrt{n} \end{aligned}$$

So $\mathbb{P}(||x_i - x_j||_2 \geq \epsilon, \forall(i, j)) \geq 1 - \sum_{i\neq j} P(||x_i - x_j||_2 < \epsilon) \geq 1 - 2\epsilon\sqrt{n}N^2$. Take $\epsilon = \frac{0.01}{2\sqrt{n}N^2}$, we get the result.

Part Two: There is a $w \in \mathbb{R}^n$ such that $||w||_2 = 1$ and $|w(x_i - x_j)| \geq \frac{0.01}{4N^4n}\sqrt{\frac{8}{\pi}}$

By Lemma J.1, for any pair i, j, $\mathbb{P}_u(|u(x_i - x_j)| < \frac{||x_i - x_j||_2}{2N^2}\sqrt{\frac{8}{n\pi}}) < \frac{1}{N^2}$. So, $\mathbb{P}_u(|u(x_i - x_j)| \geq \frac{||x_i - x_j||_2}{2N^2}\sqrt{\frac{8}{n\pi}}, \forall(i, j)) \geq 1 - \sum_{i\neq j} \mathbb{P}_u(|u(x_i - x_j)| < \frac{||x_i - x_j||_2}{2N^2}\sqrt{\frac{8}{n\pi}}) > 1 - 1 = 0$, which implies that there is a w such that $||w||_2 = 1$ and for any pair (i, j), there are $|w(x_i - x_j)| \geq \frac{||x_i - x_j||_2}{2N^2}\sqrt{\frac{8}{n\pi}} \geq \frac{0.01}{4N^4n}\sqrt{\frac{8}{\pi}}$, use the result of part one.

Part Three: Prove the result.

Let w be the vector mentioned in part two, and $wx_i < wx_j$ when $i \neq j$. Let $\delta = \frac{0.01}{4N^4n}\sqrt{\frac{8}{\pi}}$. Now, we consider the following network:

$$\mathcal{F}(x) = \sum_{i=1}^{N} \frac{y_i}{\delta}(\mathrm{ReLU}(wx - (wx_i + \delta)) + \mathrm{ReLU}(wx - (wx_i - \delta)) - 2\mathrm{ReLU}(wx - wx_i)).$$

Easy to verify $\mathcal{F}(x_i) = y_i$. Consider $|wx_i| \leq n$ and $|\frac{y_i}{\delta}| < 400aN^4n$, so $\mathcal{F} \in \mathbf{H}_{O(aN^5n^2)}(n)$. This is what we want. □

We now give the proof of when the loss function L_b satisfies (2) in definition 6.7.

Proof. In this proof, we only consider a very simple distribution $\mathcal{D}$: $\mathbb{P}_{(x,y)\sim\mathcal{D}}((x,y) = (0,-1)) = \mathbb{P}_{(x,y)\sim\mathcal{D}}((x,y) = (\mathbf{1},1)) = 0.5$, where $\mathbf{1}$ is a all one vector.

We show that for any $\mathcal{D}_{tr}$ and W, let $\mathcal{F} \in \arg\min_{f\in\mathbf{H}_W(n)} \sum_{(x,y)\in\mathcal{D}_{tr}} L_b(f(x),y)$, there are $A_{\mathcal{D}}(\mathcal{F}) = 0.5$.

Part one: When $\mathcal{D}_{tr}$ contains only $(0,-1)$, then there must be $\mathcal{F} = \sum_{i=1}^{W} -\mathrm{ReLU}(w_i x + 1) - 1$ for some x_i, which implies $\mathcal{F}(\mathbf{1}) < 0$, so $A_{\mathcal{D}}(\mathcal{F}) = 0.5$.

Part two: When $\mathcal{D}_{tr}$ contains only $(\mathbf{1},1)$, then there must be $\mathcal{F} = \sum_{i=1}^{W} \mathrm{ReLU}(\mathbf{1}x + 1) + 1$, which implies $\mathcal{F}(0) > 0$, so $A_{\mathcal{D}}(\mathcal{F}) = 0.5$.

Part Three: When $\mathcal{D}_{tr}$ contains $(\mathbf{1},1)$ and $(0,-1)$, we will show that $\mathcal{F} = \sum_{i=1}^{W} \mathrm{ReLU}(\mathbf{1}x + 1) + 1 \in \arg\min_{f\in\mathbf{H}_W(n)} L_b(f(0),-1) + L_b(f(\mathbf{1}),1)$. Consider that $A_{\mathcal{D}}(\mathcal{F}) = 0.5$ for such $\mathcal{F}$, we can prove the Theorem.

If $\mathcal{F} = \sum_{i=1}^{W} \mathrm{ReLU}(\mathbf{1}x + 1) + 1 \notin \arg\min_{f\in\mathbf{H}_W(n)} L_b(f(0),-1) + L_b(f(\mathbf{1}),1)$. Let $\mathcal{F}_0(x) = \sum_{i=1}^{W} a_i \mathrm{ReLU}(W_i x + b_i) + c \in \arg\min_{f\in\mathbf{H}_W(n)} L_b(f(0),-1) + L_b(f(\mathbf{1}),1)$. Then, let $\mathcal{F}_0(0) = b$ and $\mathcal{F}_0(\mathbf{1}) = a$.

By $\phi(a)+\phi(-b) = L_b(\mathcal{F}_0(0),-1)+L_b(\mathcal{F}_0(\mathbf{1}),1) < L_b(\mathcal{F}(0),-1)+L_b(\mathcal{F}(\mathbf{1}),1) = \phi(W(n+1)+1)+\phi(-W-1)$, and ϕ is a decreasing concave function, there must be $W(n+1)+1-a < -b+W+1$, which implies $|a-b| > Wn$.

Consider $|a-b| = |\sum_{i=1}^{W} a_i \mathrm{ReLU}(b_i) - \sum_{i=1}^{W} a_i \mathrm{ReLU}(W_i\mathbf{1} + b_i)| \leq |\sum_{i=1}^{W} a_i \mathbf{1} W_i| \leq Wn$. This is a contradiction to $|a-b| > Wn$ which was shown above. So, assumption is wrong, so $\mathcal{F} = \sum_{i=1}^{W} \mathrm{ReLU}(\mathbf{1}x + 1) + 1 \in \arg\min_{f\in\mathbf{H}_W(n)} L_b(f(0),-1) + L_b(f(\mathbf{1}),1)$, this is what we want. □

K EXTEND THE RESULT TO GENERAL NEURAL NETWORK

For multi-layer neural networks, we can show that if there is enough data and the network is large enough, then generalization can also be ensured for the network which can minimum the empirical risk. Unfortunately, due to the complexity of depth networks, we are unable to provide a good generalization bound of such network.

Denote $\mathbf{H}_{W,D}(n)$ to be the set of all neural networks of layers D with input dimension n, width W for each hidden layer, activation function ReLU, and all parameters of the transition matrix are in $[-1,1]$. Then, there are:

Theorem K.1. *For any given $n \in \mathbb{N}_+$, if $\mathcal{D} \in \mathcal{D}(n)$ satisfies: there is a network $\mathcal{F} \in \mathbf{H}_{W_0,D_0}(n)$ such that $\mathbb{P}_{(x,y)\sim\mathcal{D}}(y\mathcal{F}(x) > c) = 1$ for a $W_0, D_0 \in \mathbb{N}_+, c > 0$. Then we have that for any $W \geq \Omega(W_0)$, $D \geq \Omega(D_0)$ and $\delta > 0$, with probability at least $1-\delta$ of $\mathcal{D}_{tr} \sim \mathcal{D}^N$, it holds $A_{\mathcal{D}}(\mathcal{F}) \geq 1 - O(e^{-W^D/K} + K^n\sqrt{\frac{\ln(K/\delta)}{N}})$ for all $\mathcal{F} \in \mathbf{M}_{W,D}(\mathcal{D}_{tr}, n)$, where $K = (\frac{c}{2^{D_0+2}W_0^{D_0-1}n})^{-1}$.*

However, this bound is relatively loose, and how to obtain a bound that is polynomial in W_0, D_0, c is an important question.

Proof. **Part One**. For any given $\mathcal{D}_{tr} \sim \mathcal{D}^N$, we show that there is a network $\mathcal{F} \in \mathbf{H}_{W,D}$ such that $y\mathcal{F}(x) \geq [\frac{W}{W_0}]^{D_0-1}\frac{cW^{D-D_0}}{2}$ for any $(x,y) \in \mathcal{D}_{tr}$.

By the assumption of $\mathcal{D}$ in the theorem, let $\mathcal{F}_1 \in \mathbf{H}_{W_0,D_0}(n)$ satisfy $\mathbb{P}_{(x,y)\sim\mathcal{D}}(y\mathcal{F}_1(x) \geq c) = 1$. And W_i is the i-th transition matrix of $\mathcal{F}_1$, b_i is the i-th bias vector of $\mathcal{F}_1$.

We will construct $\mathcal{F}$ as $\mathcal{F} = \mathcal{F}_{p2} \circ \mathcal{F}_{p1}$, and we construct the two networks $\mathcal{F}_{p1}$ and $\mathcal{F}_{p2}$ as following:

$\mathcal{F}_{p1} : \mathbb{R}^n \to \mathbb{R}^W$ which has width W and depth D_0, and the output layer of $\mathcal{F}_{p1}$ also uses the ReLU activation function.

Let W be a matrix in $\mathbb{R}^{a,b}$ where $a, b \in \mathbb{N}^+$, and $T(W, a_1, b_1)$ is a matrix in $\mathbb{R}^{a_1,b_1}$ defined as: for any $i \in [a], j \in [b], k_1, k_2 \in \mathbb{Z}$, there are $T(W, a_1, b_1)_{k_1[\frac{a_1}{a}]+i, k_2[\frac{b_1}{b}]+j} = W_{i,j}$; other weights of $T(W, a_1, b_1)$ are 0. Then $\mathcal{F}_{p1}$ is defined as:

(1): The first transition matrix is $T(W_1, W, n)$, and the first bias vector is $T(b_1, W, 1)$;

(2): When $i > 2$, the i-th transition matrix is $T(W_i, W, W)$, and the i-th bias vector is $[\frac{W}{W_0}]^{i-1} T(b_i, W, 1)$.

Then, we have $\mathcal{F}_{p1}(x) = [\frac{W}{W_0}]^{D_0-1} \mathrm{ReLU}(\mathcal{F}_1(x))$.

For $\mathcal{F}_{p2} : \mathbb{R}^W \to \mathbb{R}$, which has width W and depth $D - D_0$, we define it as:

(1): When $i < D - D_0$, the i-th transition matrix is $\mathbb{I}_{W,W}$, and the i-th bias vector is 0, where $\mathbb{I}$ means all one matrix;

(2): The last transition matrix is $\mathbb{I}(1, W)$, and the last bias vector is $-[\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{2}$.

Then, $\mathcal{F} = \mathcal{F}_2 \circ \mathcal{F}_1$ is what we want.

Part two. Similar to the proof of 4.3, there are at most $Ne^{-[\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4}+2}$ points in $\mathcal{D}_{tr}$ such that $y\mathcal{F}(x) \leq [\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4}$.

Part three. If $y\mathcal{F}(x) \geq [\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4}$, then $y\mathcal{F}(x') > 0$ for all $||x' - x||_\infty \leq \frac{c}{2^{D_0+1} W_0^{D_0-1} n}$.

As shown in Lemma K.2, there are $y\mathcal{F}(x') \geq [\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4} - W^{L-1} n ||x - x'||_\infty$. So when $||x - x'||_\infty \leq \frac{c[\frac{W}{W_0}]^{L_0-1}}{4nW^{L_0-1}} \leq \frac{c}{2^{D_0+1} W_0^{D_0-1} n}$, there are $y\mathcal{F}(x') > 0$.

Part four. Let $r = \frac{c}{2^{D_0+1} W_0^{D_0-1} n}$. we can divide $[0, 1]^n$ into $\frac{1}{(r/2)^n}$ disjoint cubes that have side length $r/2$. Then by part three, we know that in a cube, $\mathcal{F}$ gives the same label to every point in such a cube when $|\mathcal{F}(x)| \geq [\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{2}$ for at least one x in such cube.

Part Five. Prove the result.

By part four, name such m cubes as $c_1, c_2, \cdots, c_m$, and let $\mathbb{P}_i = \mathbb{P}_{(x,y)\sim\mathcal{D}}(x \in c_i)$ and $\mathbb{P}_i \geq \mathbb{P}_j$ when $i \geq j$.

As shown in part four, let $S = \{i \in [N], \exists (x, y) \in \mathcal{D}_{tr} \cap c_i, y\mathcal{F}(x) \geq [\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4}\}$, then we have $A_{\mathcal{D}}(\mathcal{F}) \geq \sum_{i \in S} \mathbb{P}_i$.

For any i, by Hoeffding inequality, with probability $1 - e^{-N\mathbb{P}_i^2/2}$, there are at least $N\mathbb{P}_i/2$ points in cube c_i. So for any given $\epsilon_0 > 0$, let $\mathbb{P}_{k_0} \geq \epsilon_0$, then, with probability at least $1 - \sum_{i=k_0}^m e^{-N\epsilon_0^2/2}$ of $\mathcal{D}_{tr}$, there are at least $N\mathbb{P}_i/2$ points in C_i for any $i \geq k_0$.

As shown in part two, there are at most $Ne^{-[\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4}+2}$ points in $\mathcal{D}_{tr}$ such that $y\mathcal{F}(x) \leq [\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4}$. So, by the above result, let $T = \{k_0, k_0+1, \ldots, N\}/S$ and $N(C_i)$ is the number of points in C_i, with probability at least $1 - \sum_{i=k_0}^m e^{-N\epsilon_0^2/2}$ of $\mathcal{D}_{tr}$, there are $\sum_{i \in T} N\mathbb{P}_i/2 \leq \sum_{i \in T} N(C_i) \leq Ne^{-[\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4}+2}$.

Hence,

$$\begin{aligned} & \mathbb{P}_{\mathcal{D}}(\mathcal{F}) \\ \geq\ & \sum_{i \in S} \mathbb{P}_i \\ \geq\ & 1 - \sum_{i \in [k_0]} \mathbb{P}_i - \sum_{i \in T} \mathbb{P}_i \\ \geq\ & 1 - m\epsilon_0 - 2e^{-[\frac{W}{W_0}]^{D_0-1} \frac{cW^{D-D_0}}{4}+2}. \end{aligned}$$

Now, we take $\epsilon_0 = \sqrt{\frac{2\ln(m/\delta)}{N}}$. We get the result. □

A required lemma is given below.

Lemma K.2. *If a network with depth L and width W, the L_∞ norm of each transition matrix does not exceed 1. Then $|\mathcal{F}(x) - \mathcal{F}(z)| \leq nW^{L-1}||x - z||_\infty$.*

Proof. It is easy to see that $||Relu(Wx+b) - \text{ReLU}(Wz+b)||_\infty \leq ||W(x-z)||_\infty \leq ||W||_{1,\infty}||x-z||_\infty$. Let $\mathcal{F}_i$ is the output of i-th layer of $\mathcal{F}$, then

$$\begin{array}{rl} & |\mathcal{F}(x) - \mathcal{F}(z)| \\ \leq & W||\mathcal{F}_{D-1}(x) - \mathcal{F}_{D-1}(z)||_\infty \\ \leq & W^2||\mathcal{F}_{D-2}(x) - \mathcal{F}_{D-2}(z)||_\infty \\ \cdots & \\ \leq & W^{D-1}||\mathcal{F}_1(x) - \mathcal{F}_1(z)||_\infty \\ \leq & nW^{D-1}||x - z||_\infty \end{array}$$

which proves the lemma. □

L EXPERIMENTS

In this section, we give some simple experiments to validate our theoretical conclusions. Our experimental setup is as follows. We used MNIST data set and two-layer networks with ReLU activation function. When training the network, we ensure that the absolute value of each parameter is smaller than 1 by weight-clipping after each gradient descent. Two experiments are considered:

About size and accuracy: For networks with widths 100,200,...,900,1000, we observe their accuracy on the test set after training. The results are shown in Figure 1.

About data and precision: Using training sets with 10%, 20%, ..., 90%, 100% of the original training set to train a network with widths 200, 400 and 600. The results are shown in Figure 2.

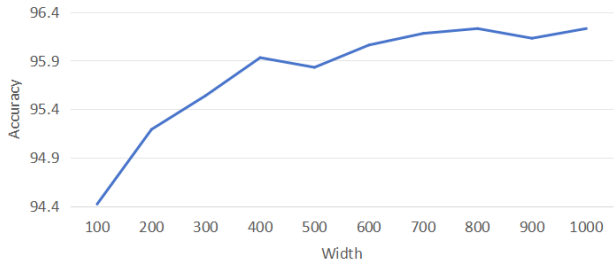

Figure 1: The accuracy on the different width networks.

Based on the experimental results, we have the following conclusions which confirm the correctness of Theorem 4.3.

(1) Increasing the amount of data or enlarging the network leads to greater accuracy. Specifically, when there are fewer data (smaller networks), increasing the number of data (width of the network) leads to a greater improvement in accuracy, which is consistent with Theorem 4.3 where the number of data (network size) is located on the denominator.

(2) When the number of data is fixed, increasing the network size has a limitation effect on improving accuracy, as shown in Figure 1, which is consistent with Theorem 4.3, because the number of data cannot affect the item in the generalization bound about network size.

(3) When the network is small, increasing the number of data can only have a limited effect on improving accuracy, as shown in Figure 2. Accuracy on training 200-width network with the entire dataset is almost equivalent to accuracy on training 400-width network with 40% data in the entire

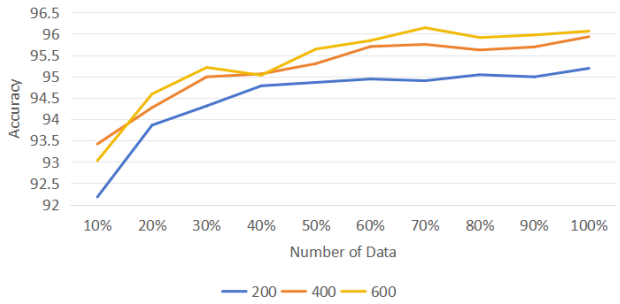

Figure 2: The accuracy on the 200,400,600 width networks with different number of data.

dataset. This is consistent with Theorem 4.3, because the size of the network cannot affect the item in the generalization bound about the number of data.