Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization

Kaiyue Wen
Tsinghua University
wenky20@mails.tsinghua.edu.cn

Zhiyuan Li
Stanford University
zhiyuanli@stanford.edu

Tengyu Ma
Stanford University
tengyuma@stanford.edu

Abstract

Despite extensive studies, the underlying reason as to why overparameterized neural networks can generalize remains elusive. Existing theory shows that common stochastic optimizers prefer flatter minimizers of the training loss, and thus a natural potential explanation is that flatness implies generalization. This work critically examines this explanation. Through theoretical and empirical investigation, we identify the following three scenarios for two-layer ReLU networks: (1) flatness provably implies generalization; (2) there exist non-generalizing flattest models and sharpness minimization algorithms fail to generalize poorly, and (3) perhaps most strikingly, there exist non-generalizing flattest models, but sharpness minimization algorithms still generalize. Our results suggest that the relationship between sharpness and generalization subtly depends on the data distributions and the model architectures and sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. This calls for the search for other explanations for the generalization of over-parameterized neural networks.

1 Introduction

It remains mysterious why stochastic optimization methods such as stochastic gradient descent (SGD) can find generalizable models even when the architectures are overparameterized (Zhang et al., 2016; Gunasekar et al., 2017; Li et al., 2017; Soudry et al., 2018; Woodworth et al., 2020). Many empirical and theoretical studies suggest that generalization is correlated with or guaranteed by the flatness of the loss landscape at the learned model (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016; Dziugaite & Roy, 2017; Jastrzebski et al., 2017; Neyshabur et al., 2017; Wu et al., 2018; Jiang et al., 2019; Blanc et al., 2019; Wei & Ma, 2019a,b; HaoChen et al., 2020; Foret et al., 2021; Damian et al., 2021; Li et al., 2021; Ma & Ying, 2021; Ding et al., 2022; Nacson et al., 2022; Wei et al., 2022; Lyu et al., 2022; Norton & Royset, 2021; Wu & Su, 2023). Thus, a natural theoretical question is

Question 0. *Does the flatness of the minimizers always correlate with the generalization capability?*

The answer to the question turns out to be false. First, Dinh et al. (2017) theoretically construct *very sharp* networks with good generalization. Second, recent empirical results (Andriushchenko et al., 2023b) find that sharpness may not have a strong correlation with test accuracy for a collection of modern architectures and settings, partly due to the same reason—there exist sharp models with good generalization. We note that, technically speaking, Question 0 is ill-defined without specifying the collection of models on which the correlation is evaluated. However, those sharp but generalizable models appear to be the main cause for the non-correlation.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Architecture	All Flattest Minimizers Generalize Well.	Sharpness Minimization Algorithms Generalize.
2-layer w/o Bias	✓ (Theorem 3.1)	✓
2-layer w/ Bias	✗ (Theorem 4.1)	✗
2-layer w/ simplified BatchNorm	✓ (Theorem 3.2)	✓
2-layer w/ simplified LayerNorm	✗ (Theorem 5.1)	✓

Table 1: **Overview of Our Results.** Each row in the table corresponds to one architecture. The second column indicates whether all flattest minimizers of training loss generalize well. ✓ indicates that all (near) flattest minimizers of training loss provably generalize well and ✗ indicates that there provably exists flattest minimizers that generalize poorly. The third column indicates whether the sharpness minimization algorithms generalize well in our experiments. Results in row 2 and 4 deny Question 1 and Question 2 respectively.

Observing the existing theoretical and empirical evidence, it is natural to ask the one-side version of Question 0, where we are only interested in whether sharpness implies generalization but not vice versa.

Question 1. *Do all the flattest neural network minimizers generalize well?*

Though there are some theoretical works that answer Question 1 affirmatively for simplified linear models (Li et al., 2021; Ding et al., 2022; Nacson et al., 2022; Gatmiry et al., 2023), the answer to Question 1 for standard neural networks remains unclear. Those theoretical results linking generalization to sharpness for more general architectures typically also involve other terms in generalization bounds, such as parameter dimension or norm (Neyshabur et al., 2017; Foret et al., 2021; Wei & Ma, 2019a,b; Norton & Royset, 2021), thus do not answer Question 1 directly.

Our first contribution is a theoretical analysis showing that the answer to Question 1 can be **false**, even for simple architectures like 2-layer ReLU networks. Intriguingly, we also find that the answer to Question 1 subtly depends on the architectures of neural networks. For example, simply removing the bias in the first layer turns the aforementioned negative result into a positive result, as also shown in the Theorem 4.3 of Wu & Su (2023) (that the authors only came to be aware of after putting this work online).

More concretely, we show that for the 2 parity xor problem with mean square loss and with data sampled from hypercube $\{-1, 1\}^d$, all flattest 2-layer ReLU neural networks without bias provably generalize. However, when bias is added, for the same data distribution and loss function, there exists a flattest minimizer that fails to generalize for every unseen data. Since adding bias in the first layer can be interpreted as appending a constant input feature, this result suggests that the generalization of the flattest minimizer is sensitive to both network architectures and data distributions.

Recent theoretical studies (Wu et al., 2018; Blanc et al., 2019; Damian et al., 2021; Li et al., 2021; Arora et al., 2022; Wen et al., 2022; Nacson et al., 2022; Lyu et al., 2022; Bartlett et al., 2022; Li et al., 2022) also show that optimizers including SGD with large learning rates or label noise and Sharpness-Aware Minimization (SAM, Foret et al. (2021)) may implicitly regularize the sharpness of the training loss landscape. These optimizers are referred to as *sharpness minimization algorithms* in this paper. Because Question 1 is not always true, it is then natural to hypothesize that sharpness-minimization algorithms will fail for architectures and data distributions where Question 1 is not true.

Question 2. *Will sharpness minimization algorithm fail to generalize when there exist non-generalizing flattest minimizers?*

A priori, the authors were expecting that the answer to Question 2 is affirmative, which means that a possible explanation is that the sharpness minimization algorithm works if and only if for certain architecture and data distribution, Question 1 is true. However, surprisingly, we also answer this question negatively for some architectures and data distributions. In other words, we found that sharpness-minimization algorithms can still generalize well even when the answer to Question 1 is false. The result is consistent with our theoretical discovery that for many architectures, there exist both non-generalizing and generalizing flattest minimizers of the training loss. We show empirically that sharpness-minimization algorithms can find different types of minimizers for different architectures.

Our results are summarized in Table 1. We show through theoretical and empirical analysis that the relationship between sharpness and generalization can fall into three different regimes depending on the architectures and distributions. The three regimes include:

- **Scenario 1.** Flattest minimizers of training loss provably generalize and sharpness minimization algorithms find generalizable models. This regime (Theorems 3.1 and 3.2) includes 2-layer ReLU MLP without bias and 2-layer ReLU MLP with a simplified BatchNorm (without mean subtraction and bias). We answer both the Question 1 and Question 2 affirmatively in this scenario.[1]
- **Scenario 2.** There exists a flattest minimizer that has the worst generalization over all minimizers. Also, sharpness minimization algorithms fail to find generalizable models. This regime includes 2 layer ReLU MLP with bias. We deny Question 1 while affirm Question 2 in this scenario.
- **Scenario 3.** There exist flattest minimizers that do not generalize but the sharpness minimization algorithm still finds the generalizable flattest model empirically. This regime includes 2-layer ReLU MLP with a simplified LayerNorm (without mean subtraction and bias). In this scenario, the sharpness minimization algorithm relies other unknown mechanisms beyond minimizing sharpness to find a generalizable model. We deny both Question 1 and Question 2 in this scenario.

2 Setup

Rademacher Complexity. Given n data $S = \{x_i\}_{i=1}^n$, the *empirical Rademacher complexity* of function class $\mathcal{F}$ is defined as $\mathcal{R}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\epsilon\sim\{\pm1\}^n}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_i f(x_i)$. **Architectures.** As summarized in Table 1, we will consider multiple network architectures and discuss how architecture influences the relationship between sharpness and generalization. For each model f_θ parameterized by θ, we will use d to denote the input dimension and m to denote the network width. We will now describe the architectures in detail.

2-MLP-No-Bias. $f_\theta^{\text{nobias}}(x) = W_2\mathsf{relu}\left(W_1 x\right)$ with $\theta = (W_1, W_2)$.

2-MLP-Bias. $f_\theta^{\text{bias}}(x) = W_2\mathsf{relu}\left(W_1 x + b_1\right)$ with $\theta = (W_1, b_1, W_2)$. We additionally define MLP-Bias as $f_\theta^{\text{bias,D}}(x) = W_D\mathsf{relu}\cdots W_2\mathsf{relu}\left(W_1 x + b_1\right)$,

2-MLP-Sim-BN. $f_\theta^{\text{sbn}}(x, \{x_i\}_{i\in[n]}) = W_2\mathsf{SBN}_\gamma\left(\mathsf{relu}\left(W_1 x + b_1\right), \{\mathsf{relu}\left(W_1 x_i + b_1\right)\}\right)$, where the simplified BatchNorm SBN is defined as $\forall m, n \in N, \forall i \in [n], x, x_i \in \mathbb{R}^m, j \in [m], \mathsf{SBN}_\gamma(x, \{x_i\}_{i\in[n]})[j] = \gamma x[j]/\left(\sum_{i=1}^n (x_i[j])^2/n\right)^{1/2}$ and $\theta = (W_1, b_1, \gamma, W_2)$.

2-MLP-Sim-LN. $f_\theta^{\text{sln}}(x) = W_2\frac{\mathsf{relu}(W_1 x + b_1)}{\max\{\|\mathsf{relu}(W_1 x + b_1)\|_2, \epsilon\}}$ where ϵ is a sufficiently small positive constant.

Surprisingly, our results show that the relationships between sharpness and generalization are strikingly different among these simple yet similar architectures.

Data Distribution. We will consider a simple data distribution as our testbed. Data distribution $\mathcal{P}_{\text{xor}}$ is a joint distribution over data point x and label y. The data point is sampled uniformly from the hypercube $\{-1, 1\}^d$ and the label satisfies $y = x[1]x[2]$. Many of our results, including our generalization bound in Section 3 and experimental observations can be generalized to broader family of distributions (Appendix B).

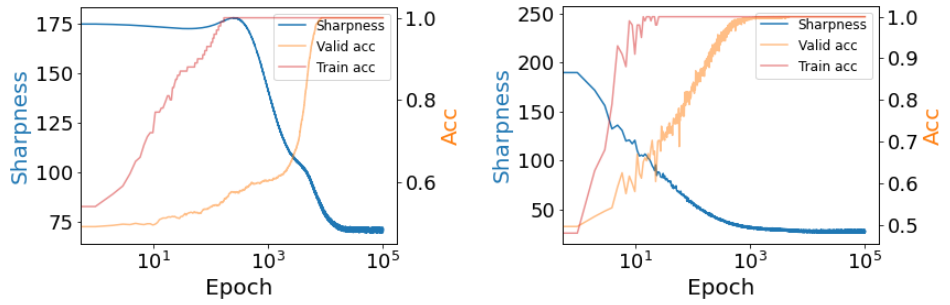
Loss. We will use mean squared error ℓ_{mse} for training and denote the training loss as L. In Appendix B, we will show that all our theoretical results and empirical observations hold for logistic loss with label smoothing probability $p > 0$. We will also consider zero one loss $\Pr(yf_\theta(x) > 0)$ for evaluating the model. We will use interpolating model to denote the model with parameter θ that minimizes L.

Definition 2.1 (Interpolating Model)**.** *A model f_θ interpolates the dataset $\{(x_i, y_i)\}_{i=1}^n$ if and only if $\forall i, f_\theta(x_i) = y_i$.*

Sharpness. Our theoretical analysis focuses on understanding the *sharpness* of the trained models. Precisely, for a model f_θ parameterized by θ, a dataset $\{(x_i, y_i)\}_{i=1}^n$ and loss function ℓ, we will use the trace of Hessian of loss function, $\mathrm{Tr}(\nabla^2 L(\theta))$ to measure how sharp the loss is at θ, which is a proxy for the sharpness along a random direction (Wen et al., 2022), or equivalently, the expected increment of loss under a random gaussian perturbation (Foret et al., 2021; Orvieto et al., 2022) .

$\mathrm{Tr}(\nabla^2 L(\theta))$ is not the only choice for defining sharpness, but theoretically many sharpness minimization algorithms have been shown to minimize this term over interpolating models. In particular,

[1]The condition for Question 2 is not satisfied and thus the answer to Question 2 is affirmative.



(a) Baseline (b) 1-SAM

Figure 1: **Scenario I.** We train a 2-layer MLP with ReLU activation without bias using gradient descent with weight decay and 1-SAM on $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and training set size $n = 100$. In both cases, the model reaches perfect generalization. Notice that although weight decay doesn't explicitly regularize model sharpness, the flatness of the model decreases through training, which is consistent with our Lemma 3.1 relating sharpness to the norm of the weight.

under the assumptions that the minimizer of the training loss form a smooth manifold Cooper (2018); Fehrman et al. (2020), Sharpness-Aware Minimization (SAM) (Foret et al., 2021) with batch size 1 and sufficiently small learning rate η and perturbation radius ρ (Wen et al., 2022; Bartlett et al., 2022), or Label Noise SGD with sufficiently small learning rate η (Blanc et al., 2019; Damian et al., 2021; Li et al., 2021), prefers interpolating models with small trace of Hessian of the loss. Hence, we choose to analyze trace of Hessian of the loss and will use SAM with batch size 1 (we denote it by 1-SAM) as our sharpness minimization algorithm in our experiments.

Notations. We use Tr to denote the trace of a matrix and $x[i]$ to denote the value of the i-th coordinate of vector x. We will use $\odot$ to represent element-wise product. We use $\mathbf{1}$ as the (coordinate-wise) indicator function, for example, $\mathbf{1}[x > 0]$ is a vector of the same length as x whose j-th entry is 1 if $x[j] > 0$ and 0 otherwise. We will use $\tilde{O}(x)$ to hide logarithmic multiplicative factors.

3 Scenario I: All Flattest Models Generalize

3.1 Flattest models provably generalize

When the architecture is 2-MLP-No-Bias, we will show that the flattest models can provably generalize, hence answering Question 1 affirmatively for this architecture and data distribution $\mathcal{P}_{\text{xor}}$.

Theorem 3.1. *For any $\delta \in (0,1)$ and input dimension d, for $n = \Omega\left(d \log\left(\frac{d}{\delta}\right)\right)$, with probability at least $1-\delta$ over the random draw of training set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{P}_{\text{xor}}^n$, let $L(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n \ell_{\text{mse}}(f_\theta^{\text{nobias}}(x_i), y_i)$ be the training loss for 2-MLP-No-Bias, it holds that for all $\theta^* \in \arg\min_{L(\theta)=0} \text{Tr}\left(\nabla^2 L(\theta)\right)$, we have that*

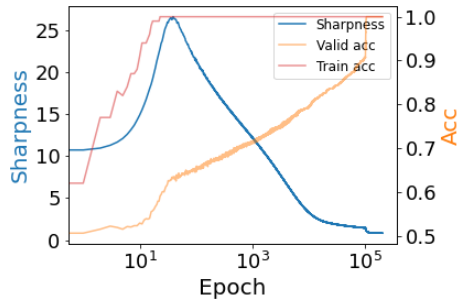
$$\mathrm{E}_{x,y\sim\mathcal{P}_{\text{xor}}}\left[\ell_{\text{mse}}\left(f_{\theta^*}^{\text{nobias}}(x), y\right)\right] \le \tilde{O}(d/n).$$

Theorem 3.1 shows that for $\mathcal{P}_{\text{xor}}$, flat models can generalize under almost linear sample complexity with respect to the input dimension. We note that Theorem 3.1 implies that $\Pr_{x,y\sim\mathcal{P}_{\text{xor}}}\left[f_{\theta^*}^{\text{nobias}}(x)y > 0\right] \le \tilde{O}(d/n)$. because if $f_{\theta^*}^{\text{nobias}}(x)y \le 0$, it holds that $\ell_{\text{mse}}\left(f_{\theta^*}^{\text{nobias}}(x), y\right) \ge 1$. This shows that the model can classify the input with high accuracy. The major proof step is relating sharpness to the norm of the weight itself.

Lemma 3.1. *Define $\Theta_C \triangleq \{\theta = (W_1, W_2) \mid \sum_{j=1}^m \|W_{1,j}\|_2 |W_{2,j}| \le C\}$. Under the setting of Theorem 3.1, there exists a absolute constant C independent of d and δ, such that with probability at least $1-\delta$, $\arg\min_{L(\theta)=0} \text{Tr}\left(\nabla^2 L(\theta)\right) \subseteq \Theta_C$ and $\mathcal{R}_S(\{f_\theta^{\text{nobias}} \mid \theta \in \Theta_C\}) \le \tilde{O}\left(\sqrt{d/n}\right)$.*

We would like to note that similar results of Theorem 3.1 and lemma 3.1 have also been shown in a prior work Wu & Su (2023) (that the authors were not aware of before the first version of this work was online).

The almost linear complexity in Theorem 3.1 is not trivial. For example, Wei et al. (2019) shows that learning the distribution will require $\Omega(d^2)$ samples for Neural Tangent Kernel (NTK) (Jacot et al.,



(a) 2-layer MLP with simplified BN

$W_{1,i}[1]$	$W_{1,i}[2]$	$\|W_{1,i}[3:d]\|_2$
18.581	-18.582	0.02
-14.363	-14.363	0.03
13.768	13.771	0.03
-12.601	12.601	0.01

(b) Weights of the four neurons with the largest norm in the first Layer

Figure 2: **Interpretable Flattest Solution** We train a 2-layer MLP with simplified BN using 1-SAM on $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and training set size $n = 100$. After training, we find that the model is indeed interpretable. In Figure 2b, we inspect the weight of the four neurons of the four largest neurons in the first layer and we observe that the four neurons approximately extract features $\pm x[1] \pm x[2]$.

2018). In contrast, our result shows that learning the distribution only requires $\tilde{O}(d)$ samples as long as the flatness of the model is controlled.

Beyond reducing model complexity, flatness may also encourage the model to find a more interpretable solution. We prove that under a stronger than i.i.d condition over the training set, the near flattest interpolating model with architecture 2-MLP-Sim-BN will provably generalize and the weight of the first layer will be centered on the first two coordinates of the input, i.e., $\|W_{1,i}[3:d]\|_2 \leq \epsilon \|W_{1,i}\|_2$.

Condition 1 (Complete Training Set Condition)**.** *There exists set $S \subset \{-1,1\}^{d-2}$, such that the linear space spanned by $S - S = \{s_1 - s_2 \mid s_1, s_2 \in S\}$ has rank $d-2$ and the training set is $\{(x,y) \mid x \in \mathbb{R}^d, x[3:d] \in S, x[1], x[2] \in \{-1,1\}, y = x[1] \times x[2]\}$.*

Theorem 3.2. *Given any training set $\{(x_i, y_i)\}_{i=1}^n$ satisfying Condition 1, for any width m and any $\epsilon > 0$, there exists constant $\kappa > 0$, such that for any width-m 2-MLP-Sim-BN, f^{sbn}, satisfying f_θ^{sbn} interpolates the training set and $\text{Tr}\left(\nabla^2 L(\theta)\right) \leq \kappa + \inf_{L(\theta')=0} \text{Tr}\left(\nabla^2 L(\theta')\right)$, it holds that $\forall x \in \{-1,1\}^d, \left|x[1]x[2] - f_\theta(x)\right| \leq \epsilon$ and that $\forall i \in [m], \|W_{1,i}[3:d]\|_2 \leq \epsilon \|W_{1,i}\|_2$.*

One may notice that in Theorem 3.2 we only consider the approximate minimizer of sharpness. This is because the gradient of output with respect to W_1, b_1, despite never being zero, will converge to zero as the norm of W_1, b_1 converges to ∞.

Condition 1 may seem stringent. In practice (Figure 2b), we find it not necessary for 1-SAM to find a generalizable solution. We hypothesize that this condition is mainly technical. Theorem 3.2 shows that sharpness minimization may guide the model to find an interpretable and low-rank representation. Similar implicit bias of SAM has also been discussed in Andriushchenko et al. (2023a) The proof is deferred to Appendix B.1

3.2 SAM empirically finds the flattest model that generalizes

We use 1-SAM to train 2-MLP-No-Bias on data distribution $\mathcal{P}_{\text{xor}}$ to verify our Theorem 3.1 (Figure 1). As expected, the model interpolates the training set and reaches a flat minimum that generalizes perfectly to the test set.

We then verify our Theorem 3.2 by training a 2-layer MLP with simplified BN on data distribution $\mathcal{P}_{\text{xor}}$ (Figure 2a). Here we do not enforce the strong theoretical Condition 1. However, we still observe that SAM finds a flat minimum that generalizes well. We then perform a detailed analysis of the model and find that the model is indeed interpretable. For example, the four largest neurons in the first layer approximately extract features $\{\text{relu}(c_1 x[1] + c_2 x[2]) \mid c_1, c_2 \in \{-1,1\}\}$ (Figure 2b). Also, the first 2 columns of the weight matrix of the first layer, corresponding to the useful features $\{\text{relu}(c_1 x[1] + c_2 x[2]) \mid c_1, c_2 \in \{-1,1\}\}$, have norms 42.47 and 42.48, while the largest column norm of the rest of the weight matrix is only 5.65.

4 Scenario II: Both Flattest Generalizing and Non-generalizing Models Exist, and SAM Finds the Former

4.1 Both generalizing and non-generalizing solutions can be flattest

In previous section, we show through Theorems 3.1 and 3.2 that sharpness benefits generalization under some assumptions. It is natural to ask whether it is possible to extend this bound to general architectures. However, in this section, we will show that the generalization benefit depends on model architectures. In fact, simply adding bias to the first layer of 2-MLP-No-Bias makes non-vacuous generalization bound impossible for distribution $\mathcal{P}_{\text{xor}}$. This then leads to a negative answer to Question 1.

Definition 4.1 (Set of extreme points)**.** *A finite set $S \subset \mathbb{R}^d$ is a set of extreme points if and only if for any $x \in S$, x is a vertex of the convex hull of S.*

Definition 4.2 (Memorizing Solutions)**.** *A D-layer network is a* memorizing solution *for a training dataset if (1) the network interpolates the training dataset, and (2) for any depth $k \in [D-1]$, there is an injection from the input data to the neurons on depth k, such that the activations in layer k for each input data is a one-hot vector with the non-zero entry being the corresponding neuron.*

Theorem 4.1. *For any $D \geq 2$, if the input data points $\{x_i\}$ of the training set form a set of extreme points (Definition 4.1), then there exists a width n layer D MLP-Bias that is a memorizing solution (Definition 4.2) for the training dataset and has minimal sharpness over all the interpolating solutions.*

As one may suspect, these memorizing solutions can have poor generalization performance.

Proposition 4.1. *For data distribution $\mathcal{P}_{\text{xor}}$, for any number of samples n, there exists a width-n 2-MLP-Bias that memorizes the training set as in Theorem 4.1, reaches minimal sharpness over all the interpolating models and has generalization error $\max\{1 - n/2^d, 0\}$ measured by zero one error.*

This corollary shows that a flat model can generalize poorly. Comparing Theorems 3.1 and 4.1, one may observe the perhaps surprising difference caused by slightly modifying the architectures (adding bias or removing the BatchNorm). To further show the complex relationship between sharpness and generalization, the following theorem suggests, despite the existence of memorizing solutions, there also exists a flattest model that *can* generalize well.

Proposition 4.2. *For data distribution $\mathcal{P}_{\text{xor}}$, for any number of samples n, there exists a width-n 2-MLP-Bias that interpolates the training dataset, reaches minimal sharpness over all the interpolating models, and has zero generalization error measured by zero one error.*

The flat solution constructed is highly simple. It contains four activated neurons, each corresponding to one feature in $\pm x[1] \pm x[2]$ (Equation (5)).

Proof sketch. For simplicity, we will consider 2-MLP-Bias here. The construction of the memorizing solution in Theorem 4.1 is as follows (visualized in Figure 3). As the input data points form a set of extreme points (Definition 4.1), for each input data point x_i, there exists a vector $\|w_i\| = 1, w_i \in \mathbb{R}^d$, such that $\forall j \neq i, w_i^\top x_i > w_i^\top x_j$. We can then choose

$$W_1 = [\sqrt{r_i|y_i|}w_i/\epsilon]_i^\top, b_1 = [\sqrt{r_i|y_i|}\left(-w_i^\top x_i + \epsilon\right)/\epsilon]^\top, W_2 = [\text{sign}(y_i)\sqrt{|y_i|/r_i}]_i.$$

Here $r_i = (\|x_i\|^2 + 1)^{1/2}$ and ϵ is a sufficiently small positive number. Then it holds that $\text{relu}(W_1 x_i + b_1) = \sqrt{r_i|y_i|}e_i$, where e_i is the $i-$th coordinate vector. This shows there is a one-to-one correspondence between the input data and the neurons. It is easy to verify that the model interpolates the training dataset. Furthermore, for $\mathcal{P}_{\text{xor}}$ and sufficiently small ϵ, for any input $x \notin \{x_i\}_{i\in[n]}$, it holds that $\text{relu}(W_1 x + b_1) = 0$. Hence the model will output the same label 0 for all the data points outside the training set. This indicates Proposition 4.1.

To show the memorization solution has minimal sharpness, we need the following lemma that relates the sharpness and the Jacobian of the model.

Lemma 4.1. *For mean squared error loss l_{mse}, if model f_θ is differentiable and interpolates dataset $\{(x_i, y_i)\}_{i\in[n]}$, then $\text{Tr}\left(\nabla^2 L(\theta)\right) = \frac{2}{n}\sum_{i=1}^n \|\nabla_\theta f_\theta(x_i)\|^2$.*

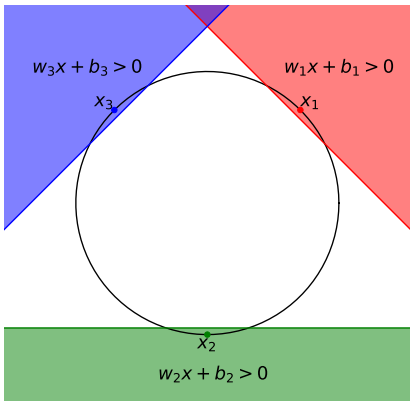


Figure 3: **Visualization of Memorization Solutions.** This is an illustration of the memorizing solutions constructed in Theorem 4.1. Here the input data points come from a unit circle and are marked as dots. The shady area with the corresponding color represents the region where the corresponding neuron is activated. One can see that the network can output the correct label for each input data point in the training set as long as the weight vector on the corresponding neuron is properly chosen. Further, the network will make the same prediction 0 for all the input data points outside the shady area and this volume can be made almost as large as the support of the training set by choosing ϵ sufficiently small. Hence the model can interpolate the training set while generalizing poorly.

Proof of Lemma 4.1. By standard calculus, it holds that,

$$\begin{aligned}
\mathrm{Tr}\left(\nabla^2 L(\theta)\right) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{Tr}\left(\nabla_\theta^2\left[(f_\theta(x_i)-y_i)^2\right]\right) \\
&= \frac{2}{n}\sum_{i=1}^{n} \mathrm{Tr}\left(\nabla_\theta^2 f_\theta(x_i)(f_\theta(x_i)-y_i) + \left(\nabla_\theta f_\theta(x_i)\right)\left(\nabla_\theta f_\theta(x_i)\right)^\top\right) \\
&= \frac{2}{n}\sum_{i=1}^{n} \mathrm{Tr}\left(\left(\nabla_\theta f_\theta(x_i)\right)\left(\nabla_\theta f_\theta(x_i)\right)^\top\right) = \frac{2}{n}\sum_{i=1}^{n} \|\nabla_\theta f_\theta(x_i)\|_2^2. \qquad (1)
\end{aligned}$$

The first equation in Equation (1) use $\forall i, f_\theta(x_i) = y_i$. The proof is then complete. □

After establishing Lemma 4.1, one can then explicitly calculate the lower bound of $\|\nabla_\theta f_\theta(x_i)\|^2$ condition on $f_\theta(x_i) = y_i$. For simplicity of writing, we will view the bias term as a part of the weight matrix by appending a 1 to the input data point. Precisely, we will use notation $x_i' \in \mathbb{R}^{d+1}$ to denote transformed input satisfying $\forall j \in [d], x_i'[j] = x_i[j], x_i'[d+1] = 1$ and $W_1' = [W_1, b_1] \in \mathbb{R}^{m\times(d+1)}$ to denote the transformed weight matrix.

By the chain rule, we have,

$$\begin{aligned}
\|\nabla_\theta f_\theta(x_i)\|^2 &= \|\nabla_{W_1'} f_\theta(x_i)\|_F^2 + \|\nabla_{W_2} f_\theta(x_i)\|_F^2 \\
&= \|(W_2 \odot \mathbf{1}\left[W_1' x_i' > 0\right])x_i'^\top\|_F^2 + \|\mathsf{relu}\left(W_1' x_i'\right)\|_2^2. \\
&= \|W_2 \odot \mathbf{1}\left[W_1' x_i' > 0\right]\|_2^2 \|x_i'\|^2 + \|\mathsf{relu}\left(W_1' x_i'\right)\|_2^2. \qquad (2)
\end{aligned}$$

Then by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\|\nabla_\theta f_\theta(x_i)\|^2 &= \|W_2 \odot \mathbf{1}\left[W_1' x_i' > 0\right]\|_2^2 \|x_i'\|^2 + \|\mathsf{relu}\left(W_1' x_i'\right)\|_2^2 \\
&\geq 2\|x_i'\| \left|\left(W_2 \odot \mathbf{1}\left[W_1 x_i > 0\right]\right)^\top \mathsf{relu}\left(W_1' x_i'\right)\right| = 2\|x_i'\||y_i|. \qquad (3)
\end{aligned}$$

In Equation (3), we use condition $f_\theta(x_i) = y_i$. Finally, notice that the lower bound is reached when

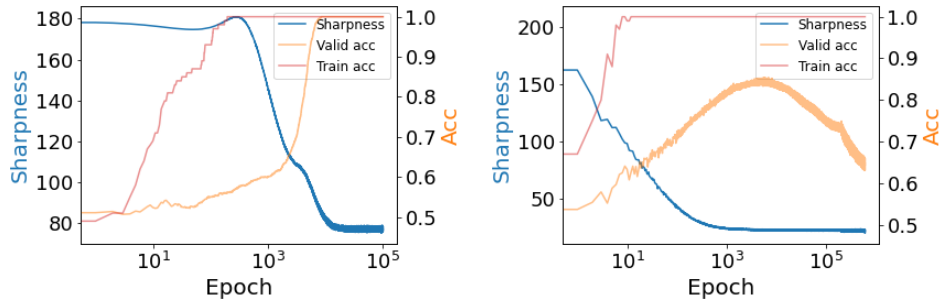
$$W_2 \odot \mathbf{1}\left[W_1' x_i' > 0\right] = \mathsf{relu}\left(W_1' x_i'\right)/\|x_i'\|. \qquad (4)$$

Condition Equation (4) is clearly reached for the memorization construction we constructed, where both sides of the equation are equal to $\sqrt{|y_i|/\|x_i'\|}e_i$. This completes the proof of Theorem 4.1.

However, the memorization network is not the only parameter that can reach the lower bound. For example, for distribution $\mathcal{P}_{\mathrm{xor}}$, if parameter θ satisfies,

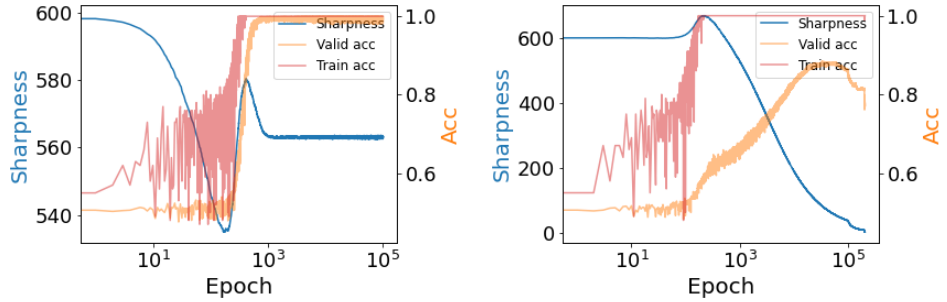
$$\forall i, j \in \{0,1\}, W_{1,2i+j+1} = r[(-1)^i, (-1)^j, ..., 0], b_1[2i+j+1] = -r, W_2[2i+j] = (-1)^{i+j}/r. \qquad (5)$$

$$\forall k > 4, W_{1,k} = [0, ..., 0], b_1[k] = 0, W_2[k] = 0,$$



(a) Baseline (b) 1-SAM

Figure 4: **Scenario II.** We train a 2-layer MLP with ReLU activation with Bias using gradient descent with weight decay and 1-SAM on $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and training set size $n = 100$. One can clearly observe a distinction between the two settings. The minimum reached by 1-SAM is flatter but the model fails to generalize and the generalization performance even starts to degenerate after 4000 epochs. The difference between Figures 1b and 4b indicates a small change in the architecture can lead to a large change in the generalization performance.



(a) Baseline (b) 1-SAM

Figure 5: **Scenario II with Softplus Activation.** We train a 2-layer MLP with Softplus activation ($\mathsf{SoftPlus}(x) = \log(1 + e^x)$) with bias using gradient descent with weight decay and 1-SAM on $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and training set size $n = 100$. We observe a similar phenomenon as Figure 4.

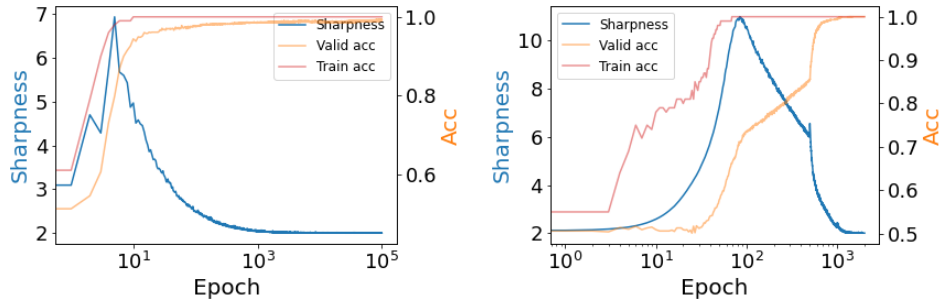
with $r = (d^2 + 1)^{1/4}$. then for any $x \in \{-1, 1\}^d$, it holds that $\mathsf{relu}(W_1 x + b_1) = re_{5/2 - x[1] - x[2]/2}$ and $f_\theta(x) = x[1] \times x[2]$. Hence it is possible for Equation (5) to hold while the model has perfect generalization performance.

4.2 SAM empirically finds the non-generalizing solutions

In this section, we will show that in multiple settings, SAM can find solutions that have low sharpness but fail to generalize compared to the baseline full batch gradient descent method with weight decay. This proves that flat minimization can hurt generalization performance. However, one should note that Question 2 is not denied for the current architectures.

Converged models found by SAM fail to generalize. We perform experiments on data distribution $\mathcal{P}_{\text{xor}}$ in Figure 4. We apply small learning rate gradient descent with weight decay as our baseline and observe that the converged model found by SAM has a much lower sharpness than the baseline. However, the generalization performance of SAM is much worse than the baseline. Moreover, the generalization performance even starts to degenerate after 4000 epochs. We conclude that in this scenario, sharpness minimization can empirically hurt generalization performance.

1-SAM may fail to generalize with other activation functions. A natural question is whether the phenomenon that 1-SAM fails to generalize is limited to ReLU activation. In Figure 5, we show empirically that 1-SAM fails to generalize for 2-layer networks with softplus activation trained on the same dataset, although there is no known guarantee for the existence of memorizing solutions.



(a) Standard Training

(b) Projected Training

Figure 6: **Scenario III.** We train two-layer ReLU networks with simplified LayerNorm on data distribution $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and sample complexity $n = 100$ using 1-SAM. In Figure 6a, we use standard training. In Figure 6b, we restricted the norm of the weight and the bias of the first layer as 10, to avoid minimizing the sharpness by simply increasing the norm. We can see that in both cases, the models almost perfectly generalize.

5 Scenario III: Both Flattest Generalizing and Non-generalizing Models Exist, and SAM Finds the Latter

5.1 Both generalizing and non-generalizing solutions can be flattest

Despite the surprising contrary between Theorems 3.1 and 4.1, experiments show that Question 2 consistently hold. However, we will provide a counterexample in this section. Specifically, we will consider data distribution $\mathcal{P}_{\text{xor}}$ and 2-layer ReLU MLP with simplified LayerNorm. One can first show both generalizing and non-generalizing solutions exist similar to Theorem 4.1 and propositions 4.1 and 4.2.

Theorem 5.1. *If the input data points $\{x_i\}$ of the training set form a set of extreme points (Definition 4.1), for sufficiently small ϵ, then there exists a width-n 2-MLP-Sim-LN with hyperparameter ϵ that is a memorizing solution (Definition 4.2) for the training dataset and has minimal sharpness over all the interpolating solutions.*

Proposition 5.1. *For data distribution $\mathcal{P}_{\text{xor}}$, for sufficiently small ϵ, for any number of samples n, there exists a width-n 2-MLP-Sim-LN with hyperparameter ϵ that memorizes the training set as in Theorem 4.1, reaches minimal sharpness over all the interpolating models and has generalization error $\max\{1 - n/2^d, 0\}$ measured by zero one error.*

Proposition 5.2. *For data distribution $\mathcal{P}_{\text{xor}}$, for sufficiently small ϵ, for any number of samples n, there exists a width-n 2-MLP-Sim-LN with hyperparameter ϵ that interpolates the training dataset, reaches minimal sharpness over all the interpolating models, and has zero generalization error measured by zero one error.*

The construction and intuition behind Theorem 5.1 and propositions 5.1 and 5.2 are similar to that of Theorem 4.1 and propositions 4.1 and 4.2. The proof is deferred to Appendix B.

5.2 SAM empirically finds generalizing models

Notice in Section 5.1 our theory makes the same prediction as in Section 4. However, strikingly, the experimental observation is reversed (Figure 6). Now running SAM can greatly improve the generalization performance till the model perfectly generalizes. This directly denies Question 2 as now we have a scenario in which sharpness minimization algorithms can improve generalization till perfect generalization while there exists a flattest minimizer that will generalize poorly.

6 Discussion and Conclusion

We present theoretical and empirical evidence for (1) whether sharpness minimization implies generalization subtly depends on the choice of architectures and data distributions, and (2) sharpness minimization algorithms including SAM may still improve generalization even when there exist flattest models that generalize poorly. Our results suggest that low sharpness may not be the only cause of the generalization benefit of sharpness minimization algorithms.

ACKNOWLEDGEMENTS

The authors would like to thank the support from NSF IIS 2045685.

References

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *arXiv preprint arXiv:2305.16292*, 2023a.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023b.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on edge of stability in deep learning. *arXiv preprint arXiv:2205.09745*, 2022.

Peter L Bartlett, Philip M Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *arXiv preprint arXiv:2210.01513*, 2022.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. *arXiv preprint arXiv:1904.09080*, 2019.

Yaim Cooper. The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*, 2018.

Alex Damian, Tengyu Ma, and Jason Lee. Label noise sgd provably prefers flat global minimizers, 2021.

Lijun Ding, Dmitriy Drusvyatskiy, and Maryam Fazel. Flat minima generalize for low-rank matrix recovery. *arXiv preprint arXiv:2203.03756*, 2022.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1019–1028. JMLR. org, 2017.

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.

Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21: 136, 2020.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Khashayar Gatmiry, Zhiyuan Li, Ching-Yao Chuang, Sashank Reddi, Tengyu Ma, and Stefanie Jegelka. The inductive bias of flatness regularization for deep matrix factorization. *arXiv preprint arXiv:2306.13239*, 2023.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

Jeff Z HaoChen, Colin Wei, Jason D Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*, 2020.

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

Jean Honorio and Tommi Jaakkola. Tight bounds for the expected risk of linear classifiers and pac-bayes finite-sample guarantees. In *Artificial Intelligence and Statistics*, pp. 384–392. PMLR, 2014.

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

Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *arXiv preprint arXiv:1712.09203*, pp. 2–47, 2017.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *International Conference on Learning Representations*, 2021.

Zhiyuan Li, Tianhao Wang, and Dingli Yu. Fast mixing of stochastic gradient descent with normalization and weight decay. *Advances in Neural Information Processing Systems*, 35:9233–9248, 2022.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *arXiv preprint arXiv:2206.07085*, 2022.

Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34:16805–16817, 2021.

Jiří Matoušek. On variants of the johnson–lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pp. 16270–16295. PMLR, 2022.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Matthew D Norton and Johannes O Royset. Diametrical risk minimization: Theory and computations. *Machine Learning*, pp. 1–19, 2021.

Antonio Orvieto, Anant Raj, Hans Kersting, and Francis Bach. Explicit regularization in over-parametrized models via noise injection. *arXiv preprint arXiv:2206.04613*, 2022.

Alessandro Rinaldo. 36-709: Advanced probability theory. https://www.stat.cmu.edu/~arinaldo/Teaching/36709/S19/Scribed_Lectures, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *Advances in Neural Information Processing Systems*, pp. 9722–9733, 2019a.

Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019b.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pp. 9709–9721, 2019.

Colin Wei, Yining Chen, and Tengyu Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? *arXiv preprint arXiv:2211.05729*, 2022.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*, 2020.

Lei Wu and Weijie J Su. The implicit regularization of dynamical stability in stochastic gradient descent. *arXiv preprint arXiv:2305.17490*, 2023.

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Contents

A Discussion on Limitation

Limitations and future work. Our results only cover a small subset of existing architectures. A natural extension for our work will be to examine whether flatness implies generalization for deep networks without the bias terms or if the flattest memorizing models exist for such architecture.

In our work, we assume 1-SAM always finds a valid global minimizer for sharpness. However, in previous works, only a local tendency of decreasing sharpness is proven. In all our experiments where the sharpness lower bound can be exactly characterized, we observe that the converged model found by 1-SAM always approximately reaches the lower bound. A possible future direction is characterizing under what condition can 1-SAM be used as a global optimizer for sharpness over minimizers.

Our work also suggests that there does not exist a universal generalization theory for neural networks only based on sharpness. A broader problem would be what other properties of the model can be used to explain the generalization of neural networks.

B Omitted Proofs

We will define $\arg\min_\theta(f(\theta))$ as the set of the minimizers f. When f has a unique minimizer, we also overload the definition to refer to that element. We will use $\mathbf{1}$ to denote the indicator function.

B.1 Formal results for 2-MLP-Sim-BN

B.1.1 Proof of Theorem 3.2

We will first prove a lemma showing that the model with a *sparse* second layer weight in L1 norm will satisfy the conclusion of Theorem 3.2.

Lemma B.1. *Given any training set $\{(x_i, y_i)\}_{i=1}^n$ satisfying Condition 1, for any width m and any $\epsilon > 0$, for any width-m 2-MLP-Sim-BN, f^{sbn} and parameter θ satisfying f_θ^{sbn} interpolates the training set and $\|\gamma \odot W_2\|_1 = \inf_{L(\theta')=0} \|\gamma' \odot W_2'\|_1$, then it holds that $\forall x \in \{-1,1\}^d, \left|x[1]x[2] - f_\theta(x)\right| = 0$ and $\forall i \in [m], \|W_{1,i}[3:d]\|_2 = 0$.*

Proof of Lemma B.1. As the training set satisfies Condition 1, there is an integer r such that $n = 4r$ and we assume WLOG for $k \in 0, 1, .., r-1$, $\{x_{4k+t}\}_{t\in\{0,1,2,3\}}$ has the same last $d-2$ coordinates and the first two coordinates are $(1,1), (1,-1), (-1,1), (-1,-1)$ respectively. Define $a_i \triangleq W_1 x_i + b_1$. Further define a as $a[j] = \sqrt{\sum_{i\in[n]} \frac{1}{n} \mathsf{relu}(a_i[j])^2}$. As the model interpolates the training set, we have

$$\sum_{j=1}^m (\gamma \odot W_2)[j] y_i \mathsf{relu}(a_i[j])/a[j] = 1, \forall i \in [n].$$

Summing the above n equalities, we have that

$$\sum_{j=1}^m (\gamma \odot W_2)[j] \sum_{i=1}^n y_i \mathsf{relu}(a_i[j])/a[j] = n.$$

This then implies

$$\sum_{j=1}^m |(\gamma \odot W_2)[j]| \geq \frac{n}{\max_j |\sum_{i=1}^n y_i \mathsf{relu}(a_i[j])/a[j]|}.$$

However, we have by Cauchy-Schwarz inequality that for each $1 \leq j \leq m$,

$$|\sum_{i=1}^n y_i \mathsf{relu}(a_i[j])| \leq \sqrt{r\left(\sum_{k=0}^{r-1}\left(\sum_{t=0}^3 y_{4k+t}\mathsf{relu}(a_{4k+t}[j])\right)^2\right)}.$$

Notice that $x_{4k+1} + x_{4k+4} = x_{4k+2} + x_{4k+3}$, it holds that $a_{4k+1} + a_{4k+4} = 2b_1 = a_{4k+2} + a_{4k+3}$ and hence by Lemma B.27, it holds that

$$\sqrt{r\left(\sum_{k=0}^{r-1}\left(\sum_{t=0}^3 y_{4k+t}\mathsf{relu}(a_{4k+t}[j])\right)^2\right)} \leq \sqrt{r\sum_{i=1}^n \mathsf{relu}(a_i[j])^2} = \frac{n}{2}a[j].$$

This then implies

$$\sum_{j=1}^m |(\gamma \odot W_2)[j]| \geq \frac{n}{n/2} = 2.$$

One can then show 2 is the minimum of $\|\gamma \odot W_2\|_1$ over minimizers by choosing the weights as

$\forall i, j \in \{0,1\}, W_{1,2i+j+1}' = [(-1)^i, (-1)^j, ..., 0], b_1'[2i+j+1] = -1, W_2'[2i+j] = \frac{1}{2}, \gamma[2i+j+1] = 1;$
$\forall k > 4, W_{1,k}' = [0, ..., 0], b_1'[k] = 0, W_2'[k] = 0, \gamma[k] = 1.$

The training loss is minimized and $\|\gamma \odot W_2\|_1 = 2$.

Hence $\|\gamma \odot W_2\|_1 = \inf_{L(\theta')=0} \|\gamma' \odot W_2'\|_1$ implies all the inequalities above must be equality. This implies $\forall j \in [m], |\sum_{i=1}^n y_i \mathsf{relu}(a_i[j])|/a[j] = \frac{n}{2}$, which by Lemma B.27 then implies the following condition: for any $j \in [m]$, there exists t_j such that $a_{4k+t}[j] \leq 0$ for $t \neq t_j$ and $a_{4k+t_j}[j]$ is a constant independent of k. [2]

This implies for any $s_1, s_2 \in S$ with S defined in Condition 1, it holds that $\forall i \in [m], W_{1,i}[3:d](s_1 - s_2) = 0$. As the linear space spanned by $S - S$ has rank $d-2$, this implies that $\forall i \in [m], \|W_{1,i}[3:d]\|_2 = 0$. As the model predict correctly over the training set and does not use the last $d-2$ coordinates, we have that $\forall x \in \{-1,1\}^d, \left|x[1]x[2] - f_\theta(x)\right| = 0$. The proof is then complete. □

We also have an approximate version of the above lemma.

Lemma B.2. *Given any training set $\{(x_i, y_i)\}_{i\in[n]}$ satisfying Condition 1, for any $\epsilon > 0$ and width m, there exists $\kappa > 0$, such that for any width-m 2-MLP-Sim-BN f^{sbn} parameterized by θ satisfying f_θ^{sbn} interpolates the training set and $\|\gamma\|_2^2 + \|W_2\|_2^2 \leq \kappa + \inf_{L(\theta)=0} \left(\|\gamma\|_2^2 + \|W_2\|_2^2\right)$, it holds that $\forall x \in \{-1,1\}^d, \left|x[1]x[2] - f_\theta(x)\right| \leq \epsilon$ and $\forall i \in [m], \|W_{1,i}[3:d]\|_2 \leq \epsilon$.*

Proof of Lemma B.2. Suppose for any $\kappa > 0$, there exists θ_κ, such that $\sum_{x\in\{-1,1\}^d} \left|x[1]x[2] - f_{\theta_\kappa}(x)\right| > \epsilon$, $L(\theta_\kappa) = 0$ and $\|\gamma_\kappa\|_2^2 + \|W_{2,\kappa}\|_2^2 \leq \kappa + \inf_{L(\theta)=0} \|\gamma\|_2^2 + \|W_2\|_2^2$, We can normalize the first layer of θ_κ such that $\|W_{\kappa,1}\|_2^2 + \|b_{\kappa,1}\|^2 = 1$ without changing the function represented by the network. Then $(W_{\kappa,1}, b_{\kappa,1}, W_{2,\kappa}, \gamma_\kappa)$ falls in a bounded set. Therefore there exists an accumulation point $\theta^* = (W_1^*, b_1^*, W_2^*, \gamma^*)$ of $\{\theta_{1/i}\}_{i\in\mathbf{N}}$, however as $L(\theta)$ and $\|\gamma\|_2^2 + \|W_2\|_2^2$ are continuous functions of θ, this implies that $L(\theta^*) = 0$ and $\|\gamma^*\|_2^2 + \|W_2^*\|_2^2 = \inf_{L(\theta)=0} \left(\|\gamma\|_2^2 + \|W_2\|_2^2\right)$.

Notice by AM-GM inequality we have that $\|\gamma\|_2^2 + \|W_2\|_2^2 \geq 2\|\gamma \odot W_2\|_1$ and equality holds when $\gamma = W_2$. We then have $\inf_{L(\theta)=0} \left(\|\gamma\|_2^2 + \|W_2\|_2^2\right) = 2\inf_{L(\theta)=0} \|\gamma \odot W_2\|_1$ and $\gamma^* = W_2^*$. Then by Lemma B.1, we have that $\sum_{x\in\{-1,1\}^d} \left|x[1]x[2] - f_{\theta^*}(x)\right| = 0$. However θ^* is an accumulation point of θ_κ satisfying $\sum_{x\in\{-1,1\}^d} \left|x[1]x[2] - f_{\theta_{1/i}}(x)\right| > \epsilon$. This then leads to a contradiction. □

We will now lower bound the sharpness of the model,

Lemma B.3. *For any parameter θ for architecture 2-MLP-Sim-BN satisfying that $L(\theta) = 0$, it holds that $\mathrm{Tr}\left(\nabla^2 L(\theta)\right) \geq \|W_2\|_2^2 + \|\gamma\|_2^2$ and $\inf_{L(\theta)=0} \mathrm{Tr}\left(\nabla^2 L(\theta)\right) = \inf_{L(\theta)=0} \|W_2\|_2^2 + \|\gamma\|_2^2$.*

Proof of Lemma B.3. Define a_i and a as in proof of Lemma B.1. By Lemma 4.1,

$$\begin{aligned}
\mathrm{Tr}\left(\nabla^2 L(\theta)\right) &= \frac{1}{n}\sum_{i=1}^n \|\nabla_\theta f_\theta^{\mathrm{sbn}}(x_i)\|_2^2 \\
&\geq \frac{1}{n}\sum_{i=1}^n \|\nabla_\gamma f_\theta^{\mathrm{sbn}}(x_i)\|_2^2 + \|\nabla_{W_2} f_\theta^{\mathrm{sbn}}(x_i)\|_2^2 \\
&= \frac{1}{n}\sum_{j=1}^m \sum_{i=1}^n |W_2[j]a_i[j]/a[j]|_2^2 + |\gamma[j]a_i[j]/a[j]|_2^2 \\
&= \|W_2\|_2^2 + \|\gamma\|_2^2.
\end{aligned}$$

This inequality can approximately reach equality when $W_2 = \gamma$ and $\|W_1\|_2$ is sufficiently large. □

Now we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. This is a direct consequence of Lemma B.2 and Lemma B.3. □

[2] Notice that for any k, the order of $a_{4k+t}[j]$ is the same.

B.2 Formal results for 2-MLP-No-Bias

We will prove a more general result that holds for any data distribution satisfying the following condition.

Condition 2. *There exists constant C and σ independent of d and m such that the data distribution $\mathcal{D}$ over data points x and label y satisfies that x is symmetric[3] subgaussian random vector (Definition B.4) with parameter σ and variance I_d. Further, there exists parameter $\theta = (W_1, W_2)$ for width-m architecture 2-MLP-No-Bias such that $\Pr\left(f_\theta^{\text{nobias}}(x) = y\right) = 1$ and $\sum_{j=1}^m \|W_{1,j}\|_2 |W_{2,j}| \le C$.*

Theorem B.1. *Given any data distribution $\mathcal{D}$ satisfying Condition 2, for any $\delta \in (0,1)$ and input dimension d, for $n = \Omega\left(d \log\left(\frac{d}{\delta}\right)\right)$, with probability at least $1-\delta$ over the random draw of training set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, let $L(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n \ell_{\text{mse}}(f_\theta^{\text{nobias}}(x_i), y_i)$ be the training loss for width-m 2-MLP-No-Bias, it holds that for all $\theta^* \in \arg\min_{L(\theta)=0} \text{Tr}\left(\nabla^2 L(\theta)\right)$,*

$$\mathrm{E}_{x,y\sim\mathcal{D}}\left[\ell_{\text{mse}}\left(f_{\theta^*}^{\text{nobias}}(x), y\right)\right] \le \tilde{O}\left(d/n\right).$$

We will also prove a more general result than Lemma 3.1.

Lemma B.4. *Define $\Theta_C \triangleq \{\theta = (W_1, W_2) \mid \sum_{j=1}^m \|W_{1,j}\|_2 |W_{2,j}| \le C\}$. Under the setting of Theorem B.1, there exists a absolute constant C_1 independent of d and δ, such that with probability at least $1-\delta$ over the randomness of dataset $\{(x_i, y_i)\}_{i=1}^n$, $\arg\min_{L(\theta)=0} \text{Tr}\left(\nabla^2 L(\theta)\right) \subseteq \Theta_{C_1}$ and $\mathcal{R}_S(\{f_\theta^{\text{nobias}} \mid \theta \in \Theta_{C_1}\}) \le \tilde{O}\left(\sqrt{d/n}\right)$.*

B.2.1 Lemmas for uniform convergence

We will begin with two uniform convergence bounds that will be used in the proof of Lemma 3.1.

Lemma B.5. *Given any data distribution $\mathcal{D}$ satisfying Condition 2, there exists constant $C_2 > C_1 > 0$ depending on σ, for any $\delta \in (0,1)$, input dimension d, and number of samples $n = \Omega\left(d \log\left(\frac{d}{\delta}\right)\right)$, with probability at least $1-\delta$ over the random draw of set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, for any $w \in \mathbb{R}^d$, we have that,*

$$C_2 d \ge \frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2 \mathbf{1}\left(w^\top x_i > 0\right) \ge C_1 d. \tag{6}$$

Lemma B.6. *Given any data distribution $\mathcal{D}$ satisfying Condition 2, there exists constant $C_2 > C_1 > 0$ depending on σ, for any $\delta \in (0,1)$, input dimension d, and number of samples $n = \Omega\left(d \log\left(\frac{d}{\delta}\right)\right)$, with probability at least $1-\delta$ over the random draw of set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, for any $w \in \mathbb{R}^d$, $\|w\|_2 = 1$, we have that,*

$$C_2 \ge \frac{1}{n}\sum_{i=1}^n |w^\top x_i|_2^2 \mathbf{1}\left(w^\top x_i > 0\right) \ge C_1. \tag{7}$$

We will first prove Lemma B.5 by a combination of Concentration Inequalities and uniform convergence bound based on Rademacher complexity.

Proof of Lemma B.5. We will first prove the upper bound, by Lemma B.17, it holds that

$$\Pr(\|X_i\|_2^2 \ge 32\sigma^2 d + 8\sigma^2 \log(2n/\delta)) \le \frac{\delta}{2n}.$$

Hence when $\log(2n/\delta) \le 1024d$, the proof for the upper bound is complete. If $\log(2n/\delta) > 1024d$, then by Lemma B.17 and Chebyshev's inequality, we have that

$$\Pr(\frac{1}{n}\sum_{i=1}^n \|x_i\|_2^2 > 3d/2) \le \frac{4\text{Var}(\|x\|_2^2)}{d^2 n} \le \frac{2048d^2}{n} \le 2048d^2 \exp(-1024d)\delta \le \delta/2.$$

[3] x and $-x$ equal in distribution.

This concludes the proof of the upper bound.

We will then prove the lower bound. We will first show that there exists $\Omega(n)$ data points in $\{x_i\}$, such that $\|x_i\| \geq \Omega(\sqrt{d})$. By Lemma B.19, we have there exists constant ϵ, ζ such that $\Pr(\|x\|_2^2 > \epsilon d) > \zeta$.

Define indicator $b_i \triangleq \mathbf{1}\left(\|x_i\|_2^2 \geq \epsilon d\right)$, then b_i are i.i.d Bernoulli random variables. We then have $p = \Pr(b_i = 1) > \zeta$. By Chernoff's bound, it holds that

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} b_i \leq p/2\right) \leq \exp(-np/8) \leq \frac{\delta}{4},$$

for any $n > 8\frac{\log(1/\delta)}{\zeta}$. This shows that with probability at least $1 - \frac{\delta}{4}$, we have that,

$$\frac{1}{n}\sum_{i=1}^{n} b_i \geq p/2 \geq \zeta/2.$$

This shows that there exists $n' \geq \lfloor \zeta n/2 \rfloor$ data points in $\{x_i\}$, such that $\|x_i\|_2^2 \geq \epsilon d$. We can then relabel the data points as $z_1, ..., z_{n'}$. Then z_i are i.i.d random variables with $\|z_i\|_2^2 \geq 1/2$ conditioning on the value of n'. We can then have for any $w \in \mathbb{R}^d$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \|x_i\|_2^2 \mathbf{1}\left(w^\top x_i > 0\right) &\geq \frac{1}{n}\sum_{i=1}^{n'} \|z_i\|_2^2 \mathbf{1}\left(w^\top z_i > 0\right) \\
&\geq \frac{\zeta}{2}\frac{1}{n'}\sum_{i=1}^{n'} \|z_i\|_2^2 \mathbf{1}\left(w^\top z_i > 0\right) \\
&\geq \frac{\zeta\epsilon d}{2}\frac{1}{n'}\sum_{i=1}^{n'} \mathbf{1}\left(w^\top z_i > 0\right).
\end{aligned}$$

Finally, define $\mathcal{F} = \{z \to \mathbf{1}\left(w^\top z > 0\right)\}$, by Lemma B.22, we have that $\mathrm{VC}(\mathcal{F}) \leq d$. By Corollary B.1, we have that the empirical Rademacher complexity of $\mathcal{F}$ on $\{z_i\}_{i\in[n']}$ is upper bounded by $\sqrt{\frac{4d\log n'}{n'}} \leq 4\sqrt{\frac{d\log n}{\zeta n}}$.

By Lemma B.23, with probability at least $1 - \frac{\delta}{4}$, we have that

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n'}\sum_{i=1}^{n'} f(z_i) - \mathrm{E}[f(z)]\right| \leq 2\mathcal{R}_{n'}(\mathcal{F}) + 3\sqrt{\frac{\log\frac{4}{\delta}}{n'}}.$$

The symmetry of x_i implies that $\mathrm{E}[f(z_i) \mid n'] = 1/2$. This shows that with probability at least $1 - \frac{\delta}{4}$, we have that for any $w \in \mathbb{R}^d$,

$$\begin{aligned}
\frac{1}{n'}\sum_{i=1}^{n'} \mathbf{1}\left(w^\top z_i > 0\right) &\geq \frac{1}{2} - 2\mathcal{R}_{n'}(\mathcal{F}) - 3\sqrt{\frac{\log\frac{4}{\delta}}{n'}} \\
&\geq \frac{1}{2} - 8\sqrt{\frac{d\log n}{\zeta n}} - 3\sqrt{\frac{\log\frac{4}{\delta}}{\zeta n}}.
\end{aligned}$$

Hence when $n = \Omega(d\log(d/\delta))$, with probability at least $1 - \delta/2$, we have that for any $w \in \mathbb{R}^d$,

$$\frac{1}{n}\sum_{i=1}^{n} \|x_i\|_2^2 \mathbf{1}\left(w^\top x_i > 0\right) \geq \frac{\epsilon C_d}{2}\frac{1}{n'}\sum_{i=1}^{n'} \mathbf{1}\left(w^\top z > 0\right) \geq \frac{\epsilon\zeta d}{8}.$$

This concludes our proof. □

We will then prove Lemma B.6, and we will use the following lemma motivated by Matoušek (2008).

Lemma B.7. *Given any data distribution $\mathcal{D}$ satisfying Condition 2, there exists C depending on σ, such that for any $\delta \in (0,1)$, input dimension d, number of samples $n \geq C\log(2/\delta)$, and $w \in \mathbb{R}^d, \|w\|_2 = 1$, with probability at least $1-\delta$ over the random draw of set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, we have that,*

$$\frac{3}{2} \geq \frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2 \geq \frac{1}{2}. \tag{8}$$

Proof of Lemma B.7. Notice that $w^\top x_i$ is a subgaussian random variable with parameter σ^2. Hence by Lemma B.14, $w^\top x_i$ is a subexponential random variable with expectation 1. The rest follows from Lemma B.16. □

This lemma can be viewed as a variant of the Johnson-Lindenstrauss lemma. We will now proceed to show that we can prove a similar high probability bound when the indicator function $\mathbf{1}\left(w^\top x_i > 0\right)$ is taken into account.

Lemma B.8. *Given any data distribution $\mathcal{D}$ satisfying Condition 2, there exists C depending on σ, such that for any $\delta \in (0,1)$, input dimension d, sample complexity $n \geq C\log(4/\delta)$ and $w \in \mathbb{R}^d, \|w\|_2 = 1$, with probability at least $1-\delta$ over the random draw of set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, it holds that,*

$$\frac{3}{2} \geq \frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2 \mathbf{1}\left(w^\top x_i > 0\right) \geq \frac{1}{8}. \tag{9}$$

Proof of Lemma B.8. The upper bound is a direct consequence of Lemma B.7.

We will now prove the lower bound. We will first use a doubling trick using the symmetry of the data distribution. Randomly sample $\{b_i\}_{i\in n}$ uniformly from $\{-1,1\}^n$, and define $z_i = b_i x_i$. We have that z_i and x_i equals in distribution. Hence, we have that $|w^\top x_i|^2\mathbf{1}\left(w^\top x_i > 0\right)$ and $|w^\top b_i x_i|^2\mathbf{1}\left(b_i w^\top x_i > 0\right)$ equals in distribution. As b_i is independent with x_i, this shows that $|w^\top x_i|^2\mathbf{1}\left(w^\top x_i > 0\right)$ equals in distribution to $|w^\top x_i|^2 c_i$ where c_i is a Rademacher random variable independent of x_i.[4] Hence, we only need to prove that

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2 c_i < \frac{1}{8}\right) < \delta/2. \tag{10}$$

By Chernoff bound, we have that

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} c_i \leq \frac{1}{4}\right) \leq \exp\left(-\frac{n}{16}\right). \tag{11}$$

Hence when $n > 16\log(4/\delta)$, we have that $\Pr(\frac{1}{n}\sum_{i=1}^n c_i \leq \frac{1}{4}) < \delta/4$. This then implies that,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2 c_i < \frac{1}{8}\right) \leq \Pr\left(\frac{1}{n}\sum_{i=1}^{n} c_i \leq \frac{1}{4}\right) + \Pr\left(\frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2 c_i < \frac{1}{8} \mid \frac{1}{n}\sum_{i=1}^{n} c_i \geq \frac{1}{4}\right) \leq \frac{\delta}{2}, \tag{12}$$

for the last inequality, we apply Lemma B.7. This concludes our proof. □

Notice that Lemma B.8 is a point-wise bound, we will then use the technique of covering to prove a uniform bound. We will first prove a uniform bound for the case where the indicator function is not taken into account.

Definition B.1 (ϵ-covering)**.** *A set $S \in \mathbb{R}^d$ is an ϵ-covering of a set $S' \in \mathbb{R}^d$, if and only if $\forall s' \in S', \exists s \in S, \|s - s'\|_2 \leq \epsilon$.*

[4] For special case where $w^\top x_i = 0$, this still holds as both sides are zero.

Lemma B.9. *For any $\epsilon > 0$, there exists an ϵ-covering S of the unit sphere in $\mathbb{R}^d$ with cardinality smaller than $\left(\frac{2}{\epsilon}+1\right)^d$.*

Proof. Consider a maximal subset T of the unit sphere in $\mathbb{R}^d$ satisfying that $\forall t \neq t' \in T, \|t-t'\|_2 \geq \epsilon$. As T is maximal, it is an ϵ-covering of the unit sphere.

Further consider set $T' = \{x \mid \exists t, \|x-t\|_2 \leq \frac{\epsilon}{2}\}$, which is the union of $|T|$ disjoint balls with radius $\epsilon/2$. Hence T' has volume $C|T|(\frac{\epsilon}{2})^d$ with C being the volume of the unit ball in $\mathbb{R}^d$. However, T' is contained in a ball with radius $1+\frac{\epsilon}{2}$ centered at origin. Hence it holds that $C|T|(\frac{\epsilon}{2})^d \leq C\left(1+\frac{\epsilon}{2}\right)^d$. This implies $|T| \leq \left(\frac{2}{\epsilon}+1\right)^d$. □

Lemma B.10. *Given any data distribution $\mathcal{D}$ satisfying Condition 2, there exists C depending on σ, such that for any $\delta \in (0,1)$, input dimension d and sample complexity $n \geq Cd\log(\frac{d}{\delta})$, with probability at least $1-\delta$ over the random draw of set $\{(x_i,y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, it holds that for any $w \in \mathbb{R}^d, \|w\|_2 = 1$,*

$$2 \geq \frac{1}{n}\sum_{i=1}^n |w^\top x_i|^2 \geq \frac{1}{4}. \tag{13}$$

Proof of Lemma B.10. Consider a $1/16$ covering of the unit sphere in $\mathbb{R}^d$, $w_1, ..., w_N$, we have that $N \leq 64^d$. By Lemma B.7 and Union Bound, we have with probability at least $1-\delta$ over the random draw of set $\{(x_i,y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, for any $k \in [N]$, we have that,

$$\frac{3}{2} \geq \frac{1}{n}\sum_{i=1}^n |w_k^\top x_i|^2 \geq \frac{1}{2}.$$

Now suppose the above event happens and

$$w^* = \mathop{\arg\max}_{w\in\mathbb{R}^d, \|w\|_2=1} \frac{1}{n}\sum_{i=1}^n |w^\top x_i|^2, \gamma = \mathop{\max}_{w\in\mathbb{R}^d, \|w\|_2=1} \frac{1}{n}\sum_{i=1}^n |w^\top x_i|^2. \tag{14}$$

Since $\{w_i\}_{i=1}^1$ is a 1/16 covering of unit sphere, there exists $k \in [N]$ such that $\|w^* - w_k\| \leq \frac{1}{16}$, then we have that by Cauchy-Schwarz inequality,

$$\begin{aligned}
\gamma - \frac{3}{2} &\leq \frac{1}{n}\sum_{i=1}^n |w^{*\top} x_i|^2 - \frac{1}{n}\sum_{i=1}^n |w_k^\top x_i|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n |(w^*-w_k)^\top x_i|_2 |(w^*+w_k)^\top x_i|_2 \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n |(w^*-w_k)^\top x_i|^2}\sqrt{\frac{1}{n}\sum_{i=1}^n |(w^*+w_k)^\top x_i|^2} \\
&\leq \gamma \|w^*-w_k\| \|w^*+w_k\| \\
&\leq \frac{\gamma}{8}.
\end{aligned}$$

This then implies that $\gamma \leq 2$. Hence, with probability $1-\delta$, we have that the upper bound holds.

Now for any $w \in \mathbb{R}^d$, $\|w\|_2 = 1$, suppose $\|w - w_k\|_2 \leq \frac{1}{16}$, we have that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2 &\geq \frac{1}{n}\sum_{i=1}^{n} |w_k^\top x_i|^2 + \frac{1}{n}\sum_{i=1}^{n} \left(|w^\top x_i|^2 - |w_k^\top x_i|^2\right) \\
&\geq \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n} |(w^* - w_k)^\top x_i||(w^* + w_k)^\top x_i| \\
&\geq \frac{1}{2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n} |(w^* - w_k)^\top x_i|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} |(w^* + w_k)^\top x_i|^2} \\
&\geq \frac{1}{2} - \gamma\frac{1}{8} \geq \frac{1}{4}.
\end{aligned}$$

This shows that with probability $1 - \delta$, the lower bound holds as well. The proof is complete. □

We can now prove Lemma B.6.

Proof of Lemma B.6. By Lemma B.10, we have that with probability at least $1 - \delta/2$ over the random draw of set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, for any $w \in \mathbb{R}^d$, $\|w\|_2 = 1$, we have that,

$$\frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2 \in [\frac{1}{4}, 2].$$

This directly implies the upper bound. For lower bound, consider a $1/64$ covering of the unit sphere in $\mathbb{R}^d$, $w_1, ..., w_N$, we have that $N \leq 128^d$. By Lemma B.8 and Union Bound, we have with probability at least $1 - \delta/2$ over the random draw of set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, for any $k \in [N]$, we have that,

$$\frac{1}{n}\sum_{i=1}^{n} |w_k^\top x_i|^2 \mathbf{1}\left(w_k^\top x_i > 0\right) \geq \frac{1}{8}.$$

Now suppose the above event happens and for any $w \in \mathbb{R}^d$, $\|w\|_2 = 1$, suppose $\|w - w_k\| \leq \frac{1}{32}$, by Lemma B.26, we have that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2 \mathbf{1}\left(w^\top x_i > 0\right) \\
\geq &\frac{1}{n}\sum_{i=1}^{n} |w_k^\top x_i|^2 \mathbf{1}\left(w_k^\top x_i > 0\right) + \frac{1}{n}\sum_{i=1}^{n} \left(|w^\top x_i|^2 \mathbf{1}\left(w^\top x_i > 0\right) - |w_k^\top x_i|^2 \mathbf{1}\left(w_k^\top x_i > 0\right)\right) \\
\geq &\frac{1}{8} - \frac{1}{n}\sum_{i=1}^{n} |(w - w_k)^\top x_i|_2 (|w^\top x_i| + |w_k^\top x_i|) \\
\geq &\frac{1}{8} - \sqrt{\frac{1}{n}\sum_{i=1}^{n} |w^\top x_i|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} |(w - w_k)^\top x_i|^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n} |w_k^\top x_i|^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} |(w - w_k)^\top x_i|^2} \\
\geq &\frac{1}{8} - 2 \times 2 \times \frac{1}{64} = \frac{1}{16}.
\end{aligned}$$

This completes the proof. □

B.2.2 Proof of Lemma B.4

Based on Appendix B.2.1, we are now ready to show that for 2-MLP-No-Bias, sharpness is within a constant factor of the norm of the parameters.

Lemma B.11. *Given any data distribution $\mathcal{D}$ satisfying Condition 2, there exists constant $C_2 > C_1 > 0$ depending on σ, for any $\delta \in (0,1)$, input dimension d and number of samples $n = \Omega\left(d \log\left(\frac{d}{\delta}\right)\right)$, with probability at least $1-\delta$ over the random draw of training set $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^n$, for any parameter $\theta = (W_1, W_2)$ of 2-MLP-No-Bias satisfying that $L(\theta) = 0$, it holds that,*

$$C_2\left(\|W_1\|_F^2 + d\|W_2\|^2\right) \geq \operatorname{Tr}\left(\nabla^2 L(\theta)\right) \geq C_1\left(\|W_1\|_F^2 + d\|W_2\|^2\right).$$

Proof of Lemma B.11. By Lemma 4.1, we have that,

$$\begin{aligned}
\operatorname{Tr}\left(\nabla^2 L(\theta)\right) &= \frac{2}{n}\sum_{i=1}^n \left\|\frac{\partial L}{\partial W_1}\right\|_F^2 + \left\|\frac{\partial L}{\partial W_2}\right\|_2^2 \\
&= \frac{2}{n}\sum_{i=1}^n \left(\|W_2 \odot \mathbf{1}\left[W_1 x_i > 0\right]\|_2^2 \|x_i\|^2 + \|\mathsf{relu}\left(W_1 x_i\right)\|_2^2\right) \\
&= \sum_{j=1}^m \left(\|W_{2,j}\|_2^2 \left(\frac{2}{n}\sum_{i=1}^n \mathbf{1}\left[W_{1,j} x_i > 0\right]\|x_i\|^2\right) + \sum_{i=1}^n |\mathsf{relu}\left(W_{1,j} x_i\right)|^2\right).
\end{aligned}$$

By Equations (6) and (7), there exists $C_2 > C_1$, such that for any $w \in \mathbb{R}^d$, it holds that

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\left[w^\top x_i > 0\right]\|x_i\|^2 \in [C_1 d/2, C_2 d/2].$$

$$\frac{1}{n}\sum_{i=1}^n |\mathsf{relu}\left(w^\top x_i\right)|^2 \in [C_1\|w\|^2/2, C_2\|w\|^2/2].$$

This then implies our result. □

We can now prove Lemma B.4.

Proof of Lemma B.4. By Condition 2, there exists parameter $\theta = (W_1, W_2)$, such that $L(\theta) = 0$ and $\sum_{j=1}^m \|W_{1,j}\|_2 |W_{2,j}| \leq C$. We can properly rescale $W_{1,j}$ and $W_{2,j}$ such that $\|W_1\|_F^2 + d\|W_2\|^2 = 2\sqrt{d}\sum_{j=1}^m \|W_{1,j}\|_2 |W_{2,j}| \leq 2C\sqrt{d}$.

Now by Lemma B.11, we have that there exists $C_2 > C_1 > 0$, such that for any $\theta^* = (W_1^*, W_2^*) \in \arg\min_{L(\theta)=0} \operatorname{Tr}\left(\nabla^2 L(\theta)\right)$, it holds that

$$\begin{aligned}
2C_2 C\sqrt{d} &\geq C_2\left(\|W_1\|_F^2 + d\|W_2\|^2\right) \\
&\geq \operatorname{Tr}\left(\nabla^2 L(\theta)\right) \geq \operatorname{Tr}\left(\nabla^2 L(\theta^*)\right) \\
&\geq C_1\left(\|W_1^*\|_F^2 + d\|W_2^*\|^2\right) \\
&= 2C_1\sqrt{d}\sum_{j=1}^m \|W_{1,j}^*\|_2 |W_{2,j}^*|.
\end{aligned}$$

This then implies that $\sum_{j=1}^m \|W_{1,j}^*\|_2 |W_{2,j}^*| \leq \frac{C_2 C}{C_1}$, completing the proof of the first claim.

By Lemma B.17, with probability at least $1-\delta$, we have that $\max_i \|x_i\|_2^2 = \tilde{O}(\sqrt{d})$. The second claim then follows from Lemma B.24. □

B.2.3 Proof of Theorem B.1

We are now ready to prove Theorem B.1 based on Lemma B.4.

Proof of Theorem B.1. Based on Lemma B.4, there exists constant $C_1 > C$ with C defined in Condition 2, with probability at least $1-\delta$, such that $\arg\min_{L(\theta)=0} \operatorname{Tr}\left(\nabla^2 L(\theta)\right) \subset \Theta_{C_1}$ and $\mathcal{R}_S(\Theta_{C_1}) = \tilde{O}(\sqrt{\frac{d}{n}})$. To get the faster rate $\tilde{O}(d/n)$, we would like to apply Theorem B.3. The main technical difficulty to apply Theorem B.3 here is that for distribution $\mathcal{D}$, the loss function L is not

necessarily bounded. To address this issue, we will consider a truncated version of the mean squared error (as in Gatmiry et al. (2023)).

$$l_c(x,y) = \ell_c(x-y) = \begin{cases} (x-y)^2, & \text{if } x-y \in [-c,c], \\ -(x-y)^2 + 4c|x-y| - 2c^2, & \text{if } x-y \in [-2c,-c] \cup [c,2c], \\ 2c^2, & \text{if } x-y \in (-\infty,-2c] \cup [2c,\infty). \end{cases} \tag{15}$$

We will choose $c = \tilde{O}(\sqrt{d})$ as in Lemma B.18 such that $\mathrm{E}_{x,y\sim\mathcal{D}}[\|x\|^2 \mathbf{1}\left[C_1\|x\| \geq c\right]] = \tilde{O}(\frac{d}{n})$ and $\mathrm{E}_{x,y\sim\mathcal{D}}[\|x\|^2 \mathbf{1}\left[C\|x\| \geq c\right]] = \tilde{O}(\frac{d}{n})$. By Condition 2, we have for $x, y \sim \mathcal{D}$, there exists $\theta_1^* \in \Theta_C$ such that $f_{\theta_1^*}^{\text{nobias}}(x) = y$, then

$$\begin{aligned} &\mathrm{E}_{x,y\sim\mathcal{D}}[\ell_{\text{mse}}(f_\theta^{\text{nobias}}(x), y)] - \mathrm{E}_{x,y\sim\mathcal{D}}[l_c(f_\theta^{\text{nobias}}(x), y)] \\ \leq &\mathrm{E}_{x,y\sim\mathcal{D}}[(f_\theta^{\text{nobias}}(x) - y)^2 \mathbf{1}\left[|f_\theta^{\text{nobias}}(x) - y| \geq c\right]] \\ \leq &2\mathrm{E}_{x,y\sim\mathcal{D}}[f_\theta^{\text{nobias}}(x)^2 \mathbf{1}\left[|f_\theta^{\text{nobias}}(x)| \geq c\right]] + 2\mathrm{E}_{x,y\sim\mathcal{D}}[f_{\theta*}^{\text{nobias}}(x)^2 \mathbf{1}\left[|f_{\theta*}^{\text{nobias}}(x)| \geq c\right]]. \end{aligned}$$

As we have $\theta \in \Theta_{C_1}$, it holds that

$$|f_\theta^{\text{nobias}}(x)| \leq \sum_{i=1}^m |W_{2,i}| \|W_{1,i}\|_2 \|x\| \leq C_1\|x\|.$$

This then implies that,

$$\mathrm{E}_{x,y\sim\mathcal{D}}[f_\theta^{\text{nobias}}(x)^2 \mathbf{1}\left[|f_\theta^{\text{nobias}}(x)| \geq c\right]] \leq C_1^2 \mathrm{E}_{x,y\sim\mathcal{D}}[\|x\|^2 \mathbf{1}\left[C_1\|x\| \geq c\right]] = \tilde{O}(\frac{d}{n}).$$

Similarly, $\mathrm{E}_{x,y\sim\mathcal{D}}[f_{\theta^*}^{\text{nobias}}(x)^2 \mathbf{1}\left[|f_{\theta^*}^{\text{nobias}}(x)| \geq c\right]] = \tilde{O}(\frac{d}{n})$. Hence, we have that

$$\mathrm{E}_{x,y\sim\mathcal{D}}[\ell_{\text{mse}}(f_\theta^{\text{nobias}}(x), y)] - \mathrm{E}_{x,y\sim\mathcal{D}}[l_c(f_\theta^{\text{nobias}}(x), y)] = \tilde{O}(\frac{d}{n}).$$

We then define the truncated version of L as $L_c(\theta) = \frac{1}{n}\sum_{i=1}^n l_c(W_2^\top \mathsf{relu}(W_1 x_i), y_i)$. Then we clearly have $L(\theta) = 0 \implies L_c(\theta) = 0$ Now by Theorem B.3, we have that for any $\theta \in \Theta_{C_1}$ and $L(\theta) = 0$, it holds that with probability at least $1 - \delta/2$,

$$\mathrm{E}_{x,y\sim\mathcal{D}}[l_c(f_\theta^{\text{nobias}}(x), y)] \leq \tilde{O}(\frac{d + c^2\log(1/\delta)}{n}) = \tilde{O}(\frac{d}{n}).$$

This completes the proof. □

B.2.4 Proof of Theorem 3.1 and lemma 3.1

One can easily construct width 4 2-MLP-No-Bias such that for $\Pr_{x,y\sim\mathcal{D}}\left(f_\theta^{\text{nobias}}(x) = y\right) = 1$. For example, one can have that

$$W_1 = \begin{bmatrix} 1+\epsilon & 1-\epsilon & 0 & \cdots \\ 1+\epsilon & -1+\epsilon & 0 & \cdots \\ -1-\epsilon & 1-\epsilon & 0 & \cdots \\ -1-\epsilon & -1+\epsilon & 0 & \cdots \end{bmatrix}, W_2 = \frac{1}{2-2\epsilon}\begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}.$$

Hence $\mathcal{P}_{\text{xor}}$ satisfies the condition in Condition 2 and this completes the proof of Theorem 3.1 and lemma 3.1.

B.3 Formal results For MLP-Bias

We will prove Theorem 4.1 and a generalization of Proposition 4.2 in this section. We note that Proposition 4.1 is already proved in Section 4.

B.3.1 Proof of Theorem 4.1

We have demonstrated the proof for 2-MLP-Bias in Section 4, and the proof for layer-D MLP-Biasis conceptually similar.

Proof of Theorem 4.1. We will still use notation $x_i' \in R^{d+1}$ to denote transformed input satisfying $\forall j \in [d], x_i'[j] = x_i[j], x_i'[d+1] = 1$ and $W_1' = [W_1, b_1] \in \mathbb{R}^{m\times(d+1)}$ to denote the transformed weight matrix.

For the simplicity of writing, we will use the following notations,

$$a_{i,0} = x_i', a_{i,1} = \mathsf{relu}(W_1 x_i + b_1), a_{i,d} = \mathsf{relu}(W_d a_{i,d-1}), d > 1$$

We will also use $A_{i,d}$ to denote the diagonal matrix with $\mathbf{1}\left(a_{i,d} > 0\right)$ as the diagonal entries.

By Lemma 4.1 and the chain rule, we have that

$$\begin{aligned}\|\nabla_\theta f_\theta^{\mathrm{bias,D}}(x_i)\|_2^2 &= \sum_{j=2}^{D} \|\nabla_{W_j} f_\theta(x_i)\|_F^2 + \|\nabla_{W_1'} f_\theta(x_i)\|_F^2 \\ &= \sum_{j=1}^{D-1} \|W_D A_{i,D-1}\cdots W_{j+1}A_{i,j}\|_2^2 \|a_{i,j-1}\|_2^2 + \|a_{i,D-1}\|_2^2\end{aligned}$$

By AM-GM inequality and Cauchy-Schwarz inequality, we have that

$$\begin{aligned}\|\nabla_\theta f_\theta^{\mathrm{bias,D}}(x_i)\|_2^2 &= \sum_{j=1}^{D-1} \|W_D A_{i,D-1}\cdots W_{j+1}A_{i,j}\|_2^2 \|a_{i,j-1}\|_2^2 + \|a_{i,D-1}\|_2^2 \\ &\geq D\left(\left(\Pi_{j=1}^{D-1}\|W_D A_{i,D-1}\cdots W_{j+1}A_{i,j}\|_2^2 \|a_{i,j-1}\|_2^2\right)\|a_{i,D-1}\|_2^2\right)^{\frac{1}{D}} \\ &\geq D\left(\left(\Pi_{j=1}^{D-1}\|W_D A_{i,D-1}\cdots W_{j+1}A_{i,j}\|_2^2 \|a_{i,j}\|_2^2\right)\|x_i'\|_2^2\right)^{\frac{1}{D}} \\ &\geq D|y_i|^{2(D-1)/D}\|x_i'\|_2^{2/D}.\end{aligned}$$

As every training data point is an extreme point of the convex hull of $\{x_i\}$, for each input data point x_i, there exists a vector $\|w_i\| = 1, w_i \in \mathbb{R}^d$, such that $\forall j \notin i, w_i^\top x_i > w_i^\top x_j$. Finally, the above inequality can be reached by a memorizing solution when we choose,

$$\begin{aligned}&W_1 = [u_i w_i/\epsilon]_i^\top, b_1 = [u_i\left(-w_i^\top x_i + \epsilon\right)/\epsilon]_i^\top, \\ &W_j = \mathrm{diag}([1/r_i]_{i\in[n]}), \forall 2 \leq j \leq D-1, \\ &W_D = [\mathrm{sign}(y_i)/r_i]_{i\in[n]},\end{aligned}$$

with r_i, u_i satisfyng $r_i = (\|x_i'\|/|y_i|)^{1/D}, u_i = |y_i| r_i^{D-1}$ when $y_i \neq 0$, $r_i = u_i = 1$ when $y_i = 0$. The proof is then completed. □

B.3.2 Generalization of Proposition 4.2

We will directly prove a more general version of Proposition 4.2, which is Proposition B.1.

Proposition B.1. *Given any constant s, for any data distribution $\mathcal{D}$ over input x and label y satisfying that (1) the label y depends only on the first s coordinates $\mathcal{I}$ of the input, (2) $x_\mathcal{I}$ are sampled from a set of extreme points in $\mathbb{R}^{|\mathcal{I}|}$ and (3) $\Pr(\|x\|_2 = R) = 1$, for sufficiently large width m depending on $\mathcal{D}$, there exists a flattest minimizer θ^* for width-m 2-MLP-Bias with generalization error 0.*

Proof of Proposition B.1. The proof is similar to the proof of Proposition 4.2. Suppose the set of extreme points in $\mathbb{R}^{|\mathcal{I}|}$ contains k elements $v_1, ..., v_k$ satisfying $\|v_k\| = v$ and corresponds to label $y_1, ..., y_k$. Then there exist vectors $\|w_i\| = 1, w_i \in \mathbb{R}^{|\mathcal{I}|}$, such that $\forall j \neq i, w_i^\top v_i > w_i^\top v_j$. We will then choose $m = k$ and let,

$$\forall j \in [k], W_{1,j} = r[v_j, ..., 0]/\epsilon, b_1[j] = r(-w_j^\top v_j + \epsilon)/\epsilon, W_2[j] = y_j/r, \tag{16}$$

with $r^2 = |y_j|(R^2+1)$ and ϵ sufficiently small. It is easy to verify that the construction will reach the smallest sharpness for any training set. □

The construction above critically relies on the fact that there exists a set of extreme points in $\mathbb{R}^{|\mathcal{I}|}$ containing the input data points. We will show that this is not necessary by the following example.

Proposition B.2. *Given any constant L, for any data distribution $\mathcal{D}$ over input x and label $y = f(x)$ satisfying that (1) the label function f depends only on the first 2 coordinates $\mathcal{I}$ of the input and is L-lipschiz, (2) the input data points satisfy $x_{\mathcal{I}}$ are sampled uniformly from the unit circle in $\mathbb{R}^2$, and (3) $\Pr(\|x\|_2 = R) = 1$, for any $\delta \in (0, 1/20)$ and $n = \Omega(\log(1/\delta)/\delta)$, with probability $1 - \delta$ over the random draw of training set $\{(x_i, y_i)\}_{i\in[n]}$, there exists a flattest minimizer θ^* for width-n 2-MLP-Bias with generalization error $O(\delta^2)$.*

Proof. Suppose the largest value of label y is Y. Suppose for dataset $\{(x_i, y_i)\}$, the first two coordinates of $\{x_i\}$ forms a set $\{v_i\}$ that lies on the unit circle in $\mathbb{R}^2$ and corresponds to label $\{y_i\}$. Suppose WLOG v_i is sorted by the angle it forms with the x-axis. We will then define z_i as the midpoint of the arc $v_{i-1}v_i$ and w_i as the unit vector perpendicular to $z_i z_{i+1}$. Here $z_{n+1} = z_1$ and $w_0 = w_n$. The flattest minimizers θ^* is then defined as,

$$\forall j \in [n], W_{1,j} = r[w_j, ..., 0]/w_j^\top z_j, b_1[j] = r(-w_j^\top v_j + w_j^\top z_j)/w_j^\top z_j, W_2[j] = y_j/r, \quad (17)$$

with $r^2 = |y_j|\sqrt{R^2 + 1}$. Verifying that the construction will reach the smallest sharpness for the training set is easy. Now splitting the sphere into $N = \lceil 2\pi/\delta \rceil > 1/\delta$ arcs with length no longer than δ. Then by the standard coupon collector problem, with probability at least $1 - \delta$, when $n \geq N \log \delta$, there is at least one point in each arc. Under such case, we have that $z_j z_{j+1}$ has length no greater than 2δ and $w_j^\top z_j > 1 - 10\delta$ for any j.

Therefore, for any $v \in \mathbb{R}^2, \|v\| = 1$, suppose WLOG v fails in arc $z_1 z_2$ and corresponds to label y, then $f_{\theta^*}^{\text{bias}}(x) = y_1 w_1^\top (v - v_1 + z_1)/w_1^\top z_1$ for $x[1:2] = v$. Therefore, we have that

$$\begin{aligned}
\|f_{\theta^*}^{\text{bias}}(x) - y\|_2^2 &\leq \|f_{\theta^*}^{\text{bias}}(v) - y_1\|_2^2 + \|y_1 - y\|_2^2 \\
&\leq \|y_1 w_1^\top (v - v_1)/w_1^\top z_1\|_2^2 + L^2\|v - v_1\|_2^2 \\
&\leq 4Y^2\delta^2/(1 - 10\delta)^2 + L^2\delta^2.
\end{aligned}$$

This shows that the expected generalization error is bounded by $O(\delta^2)$. The proof is completed. □

B.4 Formal results for 2-MLP-Sim-LN

B.4.1 Proof of Theorem 5.1

We will first lower bound the sharpness of all minimizers of 2-MLP-Sim-LN by the following lemma.

Lemma B.12. *Given any number of samples n and $\epsilon > 0$, for any training set $\{(x_i, y_i)\}_{i\in[n]}$ satisfying that the input data points $\{x_i\}$ of the training set form a set of extreme points, for width-n 2-MLP-Sim-LN with hyperparameter ϵ, it holds that*

$$\inf_{L(\theta)=0} \operatorname{Tr}\left(\nabla^2 L(\theta)\right) \geq \frac{2}{n}\sum_{i=1}^{n} \min(1, \frac{2}{\epsilon}\sqrt{\|x_i\|_2^2 + 1}|y_i|).$$

Proof. By Lemma 4.1, we have that

$$\operatorname{Tr}\left(\nabla^2 L(\theta)\right) = \frac{2}{n}\sum_{i=1}^{n} \|\nabla_\theta f_\theta(x_i)\|_2^2.$$

We will then discuss by cases to show the lower bound of $\|\nabla_\theta f_\theta(x_i)\|_2^2$ for each $i \in [n]$ when $f_\theta(x_i) = y_i$, we will continue to use notation $x_i' \in R^{d+1}$ to denote transformed input satisfying $\forall j \in [d], x_i'[j] = x_i[j], x_i'[d+1] = 1$ and $W_1' = [W_1, b_1] \in \mathbb{R}^{m\times(d+1)}$ to denote the transformed weight matrix.

1. If $\|\mathsf{relu}(W_1 x_i + b_i)\|_2 > \epsilon$, then it holds that

$$\begin{aligned}
\|\nabla_\theta f_\theta(x_i)\|_2^2 &\geq \|\nabla_{W_2} f_\theta(x_i)\|_2^2 \\
&= \|\frac{\mathsf{relu}(W_1 x_i + b_i)}{\|\mathsf{relu}(W_1 x_i + b_i)\|_2}\|_2^2 = 1.
\end{aligned}$$

2. If $\|\mathsf{relu}(W_1 x_i + b_i)\|_2 \leq \epsilon$, then it holds that

$$\begin{aligned}\|\nabla_\theta f_\theta(x_i)\|_2^2 &\geq \|\nabla_{W_1'} f_\theta(x_i)\|_2^2 + \|\nabla_{W_2} f_\theta(x_i)\|_2^2 \\ &= \frac{1}{\epsilon^2}(\|W_2^\top \mathbf{1}\left(\mathsf{relu}(W_1 x_i + b_i) > 0\right)\|_2^2 \|x_i'\|_2^2 + \|\mathsf{relu}(W_1 x_i + b_i)\|_2^2) \\ &\geq \frac{2}{\epsilon^2}\|x_i'\|_2 |W_2^\top \mathsf{relu}(W_1 x_i + b_i)| \\ &\geq \frac{2}{\epsilon}\|x_i'\|_2 |y_i|.\end{aligned}$$

This concludes the proof. □

Proof of Theorem 5.1. By Lemma B.12, we only need to construct a memorizing solution that has sharpness $\frac{2}{n}\sum_{i=1}^n \min(1, \frac{2}{\epsilon}\sqrt{\|x_i\|_2^2+1}|y_i|)$.

As the input data points form a set of extreme points, for each input data point x_i, there exists a vector $\|w_i\| = 1, w_i \in \mathbb{R}^d$, such that $\forall j \notin i, w_i^\top x_i > w_i^\top x_j$. We can then construct the minimal sharpness solution by choosing for sufficiently small δ,

$$W_1 = [u_i w_i/\delta]_i^\top, b_1 = [u_i\left(-w_i^\top x_i + \delta\right)/\delta]_i^\top, W_2 = [r_i y_i]_{i\in[n]},$$

with r_i, u_i satisfying

1. $r_i = 1, u_i = 2\epsilon$ when $\sqrt{\|x_i\|_2^2+1}|y_i| > \epsilon$.
2. $r_i = (\frac{\epsilon}{\sqrt{\|x_i\|^2+1}|y_i|})^{1/2}, u_i = \epsilon(\frac{\sqrt{\|x_i\|^2+1}|y_i|}{\epsilon})^{1/2}$ when $0 < \sqrt{\|x_i\|_2^2+1}|y_i| \leq \epsilon$.
3. $r_i = 0, u_i = 2\epsilon$ when $y_i = 0$.

It is easy to check this is a memorizing solution that minimizes sharpness.[5] The proof is then completed. □

B.4.2 Proof of Propositions 5.1 and 5.2

Proof of Proposition 5.1. We will suppose $\epsilon < \sqrt{d+1}$, then for $\mathcal{P}_{\text{xor}}$, the minimal sharpness is always 2 by Lemma B.12. Consider the following construction for sufficiently small δ,

$$W_1 = [2\epsilon x_i/\delta]_i^\top, b_1 = [2\epsilon\left(-d + \delta\right)/\delta]_i^\top, W_2 = [y_i]_{i\in[n]},$$

Then first this is a memorizing solution that minimizes sharpness. Second, the generalization error is $1 - n/2^d$ because for any $x \notin \{x_i\}_{i\in[n]}$, it holds that $\mathsf{relu}(W_1 x + b_1) = 0$ and hence $f_\theta(x) = 0$, The proof is then completed. □

Proof of Proposition 5.2. We will suppose $\epsilon < \sqrt{d+1}$, then for $\mathcal{P}_{\text{xor}}$, the minimal sharpness is always 2 by Lemma B.12. Consider the following construction for sufficiently small δ,

$$\begin{aligned}&\forall i, j \in [2], W_{1,2i+j} = 2\epsilon[(-1)^i, (-1)^j, ..., 0], b_1[2i+j] = -2\epsilon, W_2[2i+j] = (-1)^{i+j}. \\ &\forall k > 4, W_{1,k} = [0, ..., 0], b_1[k] = 0, W_2[k] = 0,\end{aligned} \tag{18}$$

This is an interpolating parameter that minimizes sharpness that can perfectly generalize. □

[5] When $\sqrt{\|x_i\|_2^2+1}|y_i| > \epsilon$, one can notice that $\nabla_{W_1'} f_\theta(x_i) = 0$ as the activation in layer 1 is nonzero only in one dimension.

B.5 Discussion on the choice of loss function

In this section, we will show why our theoretical results hold for logistic loss with label smoothing by showing that using the logistic loss with label smoothing yields the same set of minimizers and flattest minimizers as a corresponding problem using mean squared error.

Definition B.2 (Logistic Loss with Label Smoothing)**.** *Logistic loss with label smoothing probability p is defined as, $\ell : \ell_{\text{logistic,p}}(a, b) = -p \log\left(\frac{e^{ba}}{1+e^{a}}\right) - (1-p)\log\left(\frac{e^{(1-b)a}}{1+e^{a}}\right), b \in \{0, 1\}$. We will denote the training loss yield as $\ell_{\text{logistic,p}}$ as $L^{\log}$.*

Theorem B.2. *For any probability $p \in (0, 1)$, and for any training set $\{(x_i, y_i)\}_{i\in[n]}$ satisfying that $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$, let $\gamma_p = \ln(\frac{1-p}{p})$, if the minimum of the mean squared error L^{mse} over set $\{x_i, \gamma_p(2y_i - 1)\}$ is 0, then the minimizers of L^{mse} over set $\{x_i, \gamma_p(2y_i - 1)\}$ and the minimizers of $L^{\log}$ over set $\{(x_i, y_i)\}$ are the same.*

Proof. This theorem is a direct consequence of the following inequality,

$$\ell_{\text{logistic,p}}(a, b) = -p \log\left(\frac{e^{ba}}{1+e^{a}}\right) - (1-p)\log\left(\frac{e^{(1-b)a}}{1+e^{a}}\right) \geq -p\log p - (1-p)\log(1-p).$$

The minimal is reached when $a = (2b-1)\gamma_p$ where $\gamma_p = \ln(\frac{1-p}{p})$. □

Lemma B.13. *For any probability $p \in (0, 1)$, and for any training set $\{(x_i, y_i)\}_{i\in[n]}$ satisfying that $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$, let $\gamma_p = \ln(\frac{1-p}{p})$, for any model f_θ that is differentiable and interpolates dataset $\{x_i, \gamma_p(2y_i - 1)\}_{i\in[n]}$, it holds that $\operatorname{Tr}\left(\nabla^2 L^{\log}(\theta)\right) = \frac{1}{p(1-p)}\frac{1}{n}\sum_{i=1}^{n}\|\nabla_\theta f_\theta(x_i)\|^2$.*

Proof. By standard calculus, it holds that,

$$\begin{aligned}
\operatorname{Tr}\left(\nabla^2 L(\theta)\right) =& \frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\left(\nabla_\theta^2\left[\ell_{\text{logistic,p}}(f_\theta(x_i), y_i)\right]\right) \\
=& \frac{1}{n}\sum_{i=1}^{n}\operatorname{Tr}\left(\partial_\theta\left[\frac{d\ell_{\text{logistic,p}}(f_\theta(x_i), y_i)}{df_\theta(x_i)}\nabla_\theta f_\theta(x_i)\right]\right) \\
=& \frac{1}{n}\sum_{i=1}^{n}\frac{d\ell_{\text{logistic,p}}(f_\theta(x_i), y_i)}{df_\theta(x_i)}\operatorname{Tr}\left(\nabla_\theta^2 f_\theta(x_i)\right) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{d^2\ell_{\text{logistic,p}}(a, y_i)}{da^2}\,|_{a=f_\theta(x_i)}\operatorname{Tr}\left(\nabla_\theta^2 f_\theta(x_i)\right)\operatorname{Tr}\left(\left(\nabla_\theta f_\theta(x_i)\right)\left(\nabla_\theta f_\theta(x_i)\right)^\top\right) \\
=& \frac{1}{n}\sum_{i=1}^{n}\frac{d^2\ell_{\text{logistic,p}}(a, y_i)}{da^2}\,|_{a=(2y-1)\gamma_p}\operatorname{Tr}\left(\left(\nabla_\theta f_\theta(x_i)\right)\left(\nabla_\theta f_\theta(x_i)\right)^\top\right) \\
=& \frac{1}{n}\frac{1}{p(1-p)}\sum_{i=1}^{n}\|\nabla_\theta f_\theta(x_i)\|_2^2. \qquad (19)
\end{aligned}$$

The proof is then complete. □

By Lemmas 4.1 and B.13, we have that the sharpness yields by both loss functions are the same up to a constant factor. Therefore, the flattest minimizers of both loss functions are the same.

B.6 Technical Lemmas

B.6.1 Concentration inequalities

Subgaussian random variables are defined as follows.

Definition B.3 (Subgaussian random variable)**.** *A random variable X is called σ-subgaussian if $E[X] = 0$ and $\mathrm{E}\left[\exp\left(\lambda X\right)\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.*

Subgaussian random vectors are defined as,

Definition B.4 (Subgaussian random vector)**.** *A random vector $x \in \mathbb{R}^d$ is called σ-subgaussian if* $\mathrm{E}[x] = 0$ *and* $\mathrm{E}\left[\exp\left(\lambda^T x\right)\right] \leq \exp\left(\frac{\sigma^2\|\lambda\|_2^2}{2}\right)$ *for all* $\lambda \in \mathbb{R}^d$.

We will further define subexponential random variables.

Definition B.5 (Subexponential random variable)**.** *A random variable X is (σ, α)-subexponential if* $\mathrm{E}\left[\exp\left(\lambda(X - \mathrm{E}(X))\right)\right] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right)$ *for all* $|\lambda| \leq \frac{1}{\alpha}$.

Lemma B.14 (Honorio & Jaakkola (2014))**.** *If random variable X is σ-subgaussian, then X^2 is $(4\sqrt{2}\sigma^2, 4\sigma^2)$-subexponential.*

Lemma B.15 (Hoeffding's Bound)**.** *If $\{X_i\}_{i\in[n]}$ are σ-subgaussian and independent, then there exists $C_\sigma > 0$, for all $t \geq 0$,* $\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-nt^2 C_\sigma\right)$.

Lemma B.16 (Rinaldo (2019))**.** *If $\{X_i\}_{i\in[n]}$ are (σ, α)-subexponential and independent, then there exists $C_{\alpha,\sigma} > 0$, for all $t \geq 0$,* $\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mathrm{E}[X_i])\right| \geq t\right) \leq 2\exp\left(-n\min\left(tC_{\alpha,\sigma}, t^2 C_{\alpha,\sigma}\right)\right)$.

Lemma B.17 (Rinaldo (2019))**.** *If $x \in \mathbb{R}^d$ is a σ-Subgaussian random vector then for any $t \geq 0$,*

$$\Pr\left(\|x\|_2^2 \geq 32\sigma^2 d + 8\sigma^2 t\right) \leq \exp(-t). \tag{20}$$

It also holds that $\|x\|_2^2$ has bounded second moment $\mathrm{E}[\|x\|_2^4] \leq 2048\sigma^4 d^2$.

We will also need the following lemma bounding the truncated second-order moment of a subgaussian random variable.

Lemma B.18. *For any $n > 0$ and dimension d, for any d-dimension σ-subgaussian random vector x, there exists $c = O(\sqrt{d\log(dn)}\sigma)$, such that* $\mathrm{E}\left[\|x\|^2 \mathbf{1}\left(\|x\| > c\right)\right] \leq \frac{d}{n}\sigma^2$.

Proof. We have that by Equation (20),

$$\begin{aligned}
&\mathrm{E}\left[\|x\|^2 \mathbf{1}\left(\|x\| > c\right)\right] \\
=&c^2 \Pr\left(\|x\|^2 > c^2\right) + \int_c^\infty \Pr\left(\|x\|^2 > t^2\right) dt^2 \\
\leq&c^2 \exp(-\frac{c^2 - 32\sigma^2 d}{8\sigma^2}) + \int_c^\infty 2t\exp(-\frac{t^2 - 32\sigma^2 d}{8\sigma^2})dt \\
=&c^2 \exp(-\frac{c^2 - 32\sigma^2 d}{8\sigma^2}) + 8\sigma^2 \exp(-\frac{t^2 - 32\sigma^2 d}{8\sigma^2})\,|_\infty^c \\
\leq&c^2 \exp(-\frac{c^2 - 32\sigma^2 d}{8\sigma^2}) + 8\sigma^2 \exp(-\frac{c^2 - 32\sigma^2 d}{8\sigma^2})
\end{aligned}$$

Hence there exists $c = O(\sqrt{d\log(dn)}\sigma)$ such that $\mathrm{E}\left[\|x\|^2 \mathbf{1}\left(\|x\| > c\right)\right] \leq \frac{d}{n}$. □

We will finally show a constant probability lower bound on the norm of a subgaussian random vector with unit variance.

Lemma B.19. *Given any $\sigma > 0$, there exists constant ϵ, ζ, for any dimension d, for any σ-subgaussian random vector x with connvariance I_d, it holds that* $\Pr(\|x\|_2^2 > \epsilon d) > \zeta$.

Proof. As x is σ-subgaussian, it holds that for any $\lambda \in \mathbb{R}$,

$$\mathrm{E}_{\|v\|=1}[\exp(\lambda v^\top x)] \leq \exp(\lambda^2\sigma^2/2). \tag{21}$$

Here the expectation over v in Equation (21) is taken over a uniform distribution over a unit ball and v is independent of x. Hence $v^\top x$ equals in distribution to $v[1]\|x\|_2$. Hence it holds that,

$$\mathrm{E}_{\|v\|=1}[\exp(v[1]\|x\|_2)] \leq \exp(\sigma^2/2). \tag{22}$$

Note that $\exp(x) \geq 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120}$ and $\forall t \in \mathbb{N}, \mathrm{E}[(v[1])^{2t+1}] = 0$, it holds that

$$1 + \frac{1}{2}\mathrm{E}[\|x\|_2^2]\mathrm{E}_{\|v\|=1}[(v[1])^2] + \frac{1}{24}\mathrm{E}[\|x\|_2^4]\mathrm{E}_{\|v\|=1}[(v[1])^4] \leq \exp(\sigma^2/2).$$

It is well known that $\mathrm{E}_{\|v\|=1}[(v[1])^2] = \frac{1}{d}$ and $\mathrm{E}_{\|v\|=1}[(v[1])^4] = \frac{3}{(d+2)(d+4)}$. Also it holds that $\mathrm{E}[\|x\|_2^2] = d$. Hence,

$$\mathrm{E}[\|x\|_2^4] \leq \left(\exp(\sigma^2/2) - 3/2\right) \frac{(d+2)(d+4)}{3}.$$

This implies that

$$\begin{aligned}
&\left(\exp(\sigma^2/2) - 3/2\right) \frac{(d+2)(d+4)}{3} \\
\geq &\mathrm{E}[\|x\|_2^4 I(\|x\|_2^2 > \frac{1}{2}d)] \\
\geq &\frac{\left(\mathrm{E}[\|x\|_2^2 I(\|x\|_2^2 > \frac{1}{2}d)]\right)^2}{\Pr(\|x\|_2^2 > \frac{1}{2}d)} \\
= &\frac{\left(\mathrm{E}[\|x\|_2^2] - \mathrm{E}[\|x\|_2^2 I(\|x\|_2^2 \leq \frac{1}{2}d)]\right)^2}{\Pr(\|x\|_2^2 > \frac{1}{2}d)} \\
\geq &\frac{d^2}{4\Pr(\|x\|_2^2 > \frac{1}{2}d)}
\end{aligned}$$

Hence, we can conclude that

$$\Pr(\|x\|_2^2 > \frac{1}{2}d) \geq \frac{3d^2}{4(d+2)(d+4)\left(\exp(\sigma^2/2) - 3/2\right)} \geq \frac{1}{20\left(\exp(\sigma^2/2) - 3/2\right)}.$$

This concludes the proof. □

B.6.2 Rademacher Complexity

Recall the definition of Rademacher complexity,

Definition B.6 (Rademacher complexity)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$. Let $S = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ be a set of points. The empirical Rademacher complexity of $\mathcal{F}$ with respect to S is defined as $\mathcal{R}_S(\mathcal{F}) = \frac{1}{n}\,\mathbb{E}_{\epsilon \sim \{\pm 1\}^n} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(x_i)$.*

We will also define the following notion of the shattered set and VC dimension.

Definition B.7 (Shattered set)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y} = \{0, 1\}$. A set $S = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ is said to be shattered by $\mathcal{F}$ if for every $T \subset S$, there exists $f \in \mathcal{F}$ such that $f(x) = 1$ for all $x \in T$ and $f(x) = 0$ for all $x \in S \setminus T$.*

Definition B.8 (VC dimension)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y} = \{0, 1\}$. The VC dimension of $\mathcal{F}$ is defined as $\mathrm{VC}(\mathcal{F}) = \sup\{n \in \mathbb{N} \mid$ there exists a set of size n shattered by $\mathcal{F}\}$.*

We will use the following well-known lemmas.

Lemma B.20 (Massart's Lemma)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y} = \{0, 1\}$. Further, suppose $\mathcal{A} = \{(f(x_i))_{i \in n} \mid f \in F\}$, then, $\mathcal{R}_S(\mathcal{F}) \leq \sqrt{\frac{2\log|\mathcal{A}|}{n}}$.*

Lemma B.21 (Sauer's Lemma)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y} = \{0, 1\}$. Further, suppose $\mathcal{A} = \{(f(x_i))_{i \in [n]} \mid f \in F\}$, then $|\mathcal{A}| \leq \sum_{i=0}^{\mathrm{VC}(\mathcal{F})} \binom{n}{i}$.*

Combining the above two lemmas, we get the following corollary.

Corollary B.1. *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y} = \{0, 1\}$, then $\mathcal{R}_S(\mathcal{F}) \leq \sqrt{\frac{4\mathrm{VC}(\mathcal{F})\log n}{n}}$.*

Further, we also have the following lemma controlling the VC dimension.

Lemma B.22. *Suppose* $\mathcal{F} = \{x \in \mathbb{R}^d \to \mathbf{1}\left(w^\top x > 0\right) \mid w \in \mathbb{R}^d\}$, *then* $\mathrm{VC}(\mathcal{F}) = d$.

The following uniform convergence bound based on Rademacher complexity is also well known.

Lemma B.23 (Shalev-Shwartz & Ben-David (2014))**.** *Suppose for all* $f \in \mathcal{F}$, $0 \le f(x) \le 1$, *then with probability at least* $1-\delta$ *over the randomness of i.i.d. sampled* $S = \{x_1, \ldots, x_n\} \subset \mathcal{X}$, *it holds that*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathrm{E}[f(x)] \right| \le 2\mathcal{R}_S(\mathcal{F}) + 3\sqrt{\frac{\log \frac{4}{\delta}}{n}}. \tag{23}$$

To prove our main results, we will also need the following theorem due to Srebro et al. (2010).

Definition B.9. *A loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *is* $H-$*smooth, if and only* $\frac{d\ell(x,y)}{dx}$ *is* $H-$*lipschitz.*

Theorem B.3 (Theorem 1 of Srebro et al. (2010))**.** *For an* H*-smooth non-negative loss* ℓ *s.t.* $\forall_{x,y,f} |\ell(f(x), y)| \le b$, *for any* $\delta > 0$ *we have that with probability at least* $1-\delta$ *over a random sample of size* n, *for any* $f \in \mathcal{F}$ *with zero training loss* $\hat{L}(h) = 0$,

$$L(h) \le O\left(H \log^3 n \mathcal{R}_n^2(\mathcal{F}) + \frac{b \log(1/\delta)}{n} \right).$$

Finally, we also need lemmas bounding the Rademacher complexity of norm-bounded linear hypothesis and 2-MLP-No-Bias.

Lemma B.24. *For any constant* $C > 0$ *and number of samples* n, *for the set of parameters for 2-MLP-No-Bias* $\Theta_C \triangleq \{\theta \mid \sum_{j=1}^{m} \|W_{1,j}\|_2 |W_{2,j}| \le C, \theta = (W_1, W_2)\}$ *and any training set* $\{x_i\}_{i \in [n]}$ *satisfying that* $\|x_i\|_2 \le B$, *it holds that* $\mathcal{R}_S(\{f_\theta^{\text{nobias}} \mid \theta \in \Theta_C\}) \le \frac{2CB}{\sqrt{n}}$.

Proof. Let $\overline{u}$ denotes $u/\|u\|_2$ for $u \neq 0$ and 0 when $u = 0$,

$$\begin{aligned}
R_S(\{f_\theta^{\text{nobias}} \mid \theta \in \Theta_C\}) &= \frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\theta} \sum_{i=1}^{n} \sigma_i f_\theta^{\text{nobias}}(x_i) \right] \\
&= \frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\theta} \sum_{i=1}^{n} \sigma_i \left[\sum_{j=1}^{m} W_{2,j} \mathsf{relu}\left(W_{1,j} x_i\right) \right] \right] \\
&= \frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\theta} \sum_{i=1}^{n} \sigma_i \left[\sum_{j=1}^{m} W_{2,j} \|W_{1,j}\|_2 \mathsf{relu}\left(\overline{W_{1,j}}^T x_i\right) \right] \right] \\
&= \frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\theta} \sum_{j=1}^{m} W_{2,j} \|W_{1,j}\|_2 \left[\sum_{i=1}^{n} \sigma_i \mathsf{relu}\left(\overline{W_{1,j}}^T x_i\right) \right] \right] \\
&\le \frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\theta} \sum_{j=1}^{m} |W_{2,j}| \|W_{1,j}\|_2 \max_{k \in [n]} \left| \sum_{i=1}^{n} \sigma_i \mathsf{relu}\left(\overline{W_{1,k}}^T x_i\right) \right| \right] \\
&\le= \frac{C}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\bar{u}: \|\bar{u}\|_2 = 1} \left| \sum_{i=1}^{n} \sigma_i \mathsf{relu}\left(\bar{u}^T x_i\right) \right| \right] \\
&\le \frac{C}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\bar{u}: \|\bar{u}\|_2 \le 1} \left| \sum_{i=1}^{n} \sigma_i \mathsf{relu}\left(\bar{u}^T x_i\right) \right| \right] \\
&\le \frac{2C}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{\bar{u}: \|\bar{u}\|_2 \le 1} \sum_{i=1}^{n} \sigma_i \mathsf{relu}\left(\bar{u}^T x_i\right) \right] \\
&= 2C R_S\left(\mathcal{H}'\right),
\end{aligned}$$

where $\mathcal{H}' = \left\{x \mapsto \mathsf{relu}\left(\bar{u}^\top x\right) : \bar{u} \in \mathbb{R}^d, \|\bar{u}\|_2 \leq 1\right\}$. By Talagrand's lemma, since relu is 1-Lipschitz, $R_S\left(\mathcal{H}'\right) \leq R_S\left(\mathcal{H}''\right)$ where $\mathcal{H}'' = \left\{x \mapsto \bar{u}^\top x : \bar{u} \in \mathbb{R}^d, \|\bar{u}\|_2 \leq 1\right\}$ is a linear hypothesis space. Using $R_S\left(\mathcal{H}''\right) \leq \frac{B}{\sqrt{n}}$ by Lemma B.25 concludes the proof. □

Lemma B.25. *For any constant $C > 0$ and number of samples n, for any set $S = \{x_i\}_{i\in[n]}$ satisfying that $\forall i, x_i \in \mathbb{R}^d, \|x_i\|_2^2 \leq C^2$ and function class $\mathcal{F} = \left\{x \mapsto \langle w, x\rangle \mid w \in \mathbb{R}^d, \|w\|_2 \leq 1\right\}$, it holds that,*

$$\mathcal{R}_S(\mathcal{F}) \leq \frac{C}{\sqrt{n}}.$$

Proof.

$$\begin{aligned}
\mathcal{R}_S(\mathcal{F}) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|w\|_2\leq 1} \frac{1}{n}\sum_{i=1}^n \sigma_i \langle w, x_i\rangle\right] \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|w\|_2\leq 1} \left\langle w, \sum_{i=1}^n \sigma_i x_i\right\rangle\right] \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma}\left[\left\|\sum_{i=1}^n \sigma_i x_i\right\|_2\right] \\
&= \frac{1}{n}\sqrt{\sum_{i=1}^n \|x_i\|_2^2} \leq \frac{C}{\sqrt{n}}.
\end{aligned}$$

□

B.6.3 Elementary inequalities

We will prove some elementary inequalities that will be useful in the proof of our main results.

Lemma B.26. *For any $x, y \in \mathbb{R}$, $|\mathsf{relu}(x)^2 - \mathsf{relu}(y)^2| \leq (|x| + |y|)|x - y|$.*

Proof. We will assume WLOG that $x > y$. We then discuss the following three cases.

1. $0 \geq x > y$, then the result is trivial.
2. $x > 0 \geq y$, then $|\mathsf{relu}(x)^2 - \mathsf{relu}(y)^2| = x^2 \leq (|x| + |y|)|x - y|$.
3. $x > y > 0$, then $|\mathsf{relu}(x)^2 - \mathsf{relu}(y)^2| = x^2 - y^2 = (|x| + |y|)|x - y|$.

This completes the proof. □

Lemma B.27. *For any $a, b, c, d \in \mathbb{R}$, if $a + d = b + c$, then*

$$|\mathsf{relu}(a) + \mathsf{relu}(d) - \mathsf{relu}(b) - \mathsf{relu}(c)|^2 \leq (\mathsf{relu}(a))^2 + (\mathsf{relu}(d))^2 + (\mathsf{relu}(b))^2 + (\mathsf{relu}(c))^2.$$

The equality holds if and only if three of the four values are not positive.

Proof. WLOG we assume $a \geq b \geq c \geq d$. As ReLU is convex, we have that $\mathsf{relu}(a) + \mathsf{relu}(d) - \mathsf{relu}(b) - \mathsf{relu}(c) \geq 0$. Further, we have that $\mathsf{relu}(a) + \mathsf{relu}(d) - \mathsf{relu}(b) - \mathsf{relu}(c) \leq \mathsf{relu}(a) - \mathsf{relu}(b) \leq \mathsf{relu}(a)$. Thus, we have the desired result. □

	Learning Rate	Perturbation Radius	Batch size	Weight Decay	Epochs
Figure 1a	0.01	0	100	0.05	1e5
Figure 1b	0.01	0.05	1	0	1e5
Figure 2a	0.005	0.1	1	0	1e5
	0.003	1	1	0	1e5
Figure 4a	0.01	0	100	0.05	1e5
Figure 4b	0.01	0.05	1	0	2e5
	0.01	0.1	1	0	4e5
Figure 5a	0.001	0	1	0.05	1e5
Figure 5b	0.0005	0.05	1	0	1e5
	0.001	0.1	1	0	1e5
	0.005	1	1	0	5e3
Figure 6a	0.1	0.1	1	0	1e5
Figure 6b	0.01	0.1	1	0	5e2
	0.01	0.5	1	0	5e2
	0.01	1	1	0	5e2
Figure 7a	0.01	0.2	1	0	1e5
Figure 7b	0.01	0.2	1	0	1e5
Figure 8a	0.1	0	10	0.01	1e5
Figure 8b	0.1	0.2	1	0	1e5
Figure 9a	0.01	0	1	0.05	1e5
Figure 9b	0.1	0.2	1	0	1e5
Figure 10a	0.1	0.2	1	0	4e4
Figure 10b	1	0.5	1	0	1e3
	1	1	1	0	1e5
Figure 11a	0.01	0	1	0.05	1e5
Figure 11b	0.01	0.05	1	0	1e5

Table 2: **Training details for Experiments.** For Figures 6a and 10b, we scale down the initialization of the first layer by a factor of 100 to avoid minimizing the sharpness by simply increasing the norm at the beginning.

C Experiments

C.1 Training Details

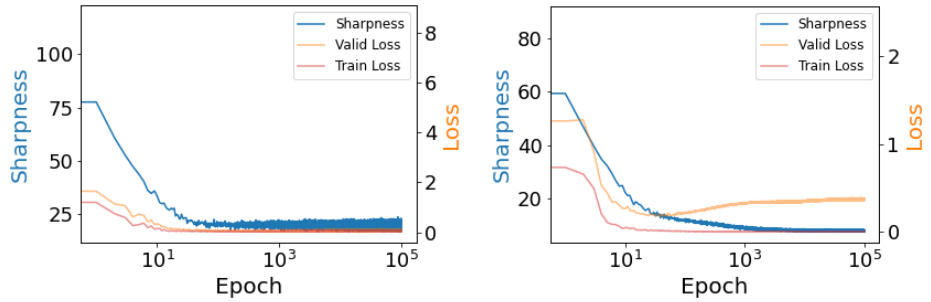
For all the experiments, we use networks with width 500. The learning rates, perturbation radius, and training epochs are summarized in Table 2. For those experiments where there are adjustments in hyperparameters through the training process, we report all the hyperparameters in multiple rows. We use 8 NVIDIA 2080 GPUs to train the models. The training time for each experiment is around 12 hours per 1e5 epochs

C.2 Extension To Uniform Ball Distribution

As our Theorems 4.1 and B.1 suggests, the generalization and memorization results should hold for data distribution other than boolean hypercube. We perform experiments on uniform ball distribution to verify this. Specifically, we sample data points uniformly from the ball with radius $\sqrt{d}$ with dimension $d = 10$ and the label is defined as $y = |x[1]| - |x[2]|$. The results are shown in Figure 7. We can see that the flattest minimizers of the two architectures have very different generalization behavior. The flattest minimizer of the MLP without bias has a much better generalization performance than the one with bias. This is consistent with our theoretical results.

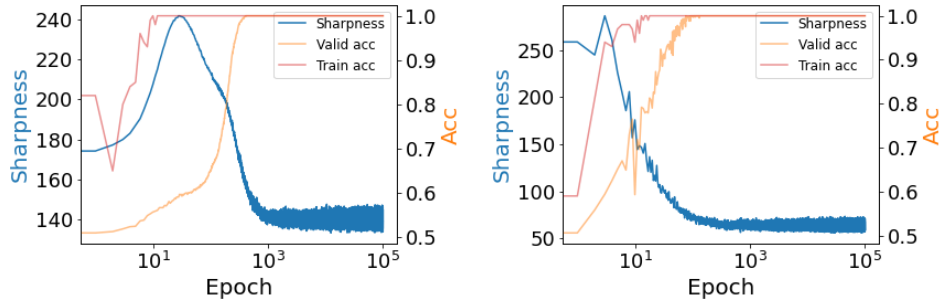
C.3 Extension To Logistic Loss

As Theorem B.2 and lemma B.13 suggests, our results can be extended to logistic loss with label smoothing. We perform all our experiments mentioned in the main text on the same distribution $\mathcal{P}_{\text{xor}}$, with the mean squared error loss replaced by logistic loss with label smoothing $p = 0.2$ to verify this. The results are shown in Figures 8 to 10.



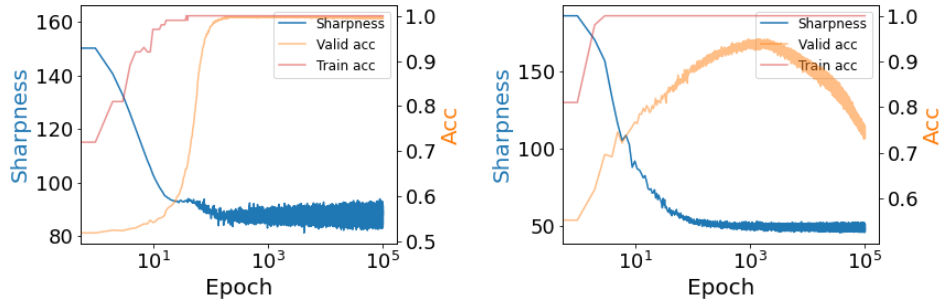
(a) 2-MLP-No-Bias (b) 2-MLP-Bias

Figure 7: **Uniform Ball Distribution.** We train a 2-layer MLP with ReLU activation with and without Bias using 1-SAM on uniform ball distribution with dimension $d = 10$ and training set size $n = 100$. One can again see the striking difference between the generalization behavior of the flattest minimizers of the two architectures.



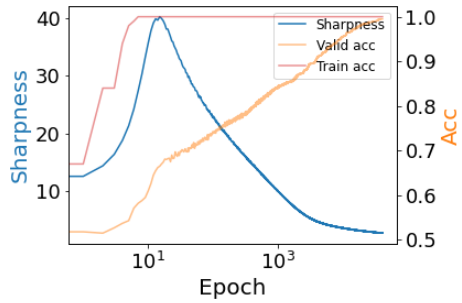
(a) Baseline (b) 1-SAM

Figure 8: **Scenario I with Logistic Loss.** We train a 2-layer MLP with ReLU activation without Bias using gradient descent with weight decay and 1-SAM on $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and training set size $n = 100$. In both cases, the model reaches perfect generalization. Notice that although weight decay doesn't explicitly regularize model sharpness, the flatness of the model decreases through training, which is consistent with our Lemma 3.1 relating sharpness to the norm of the weight.

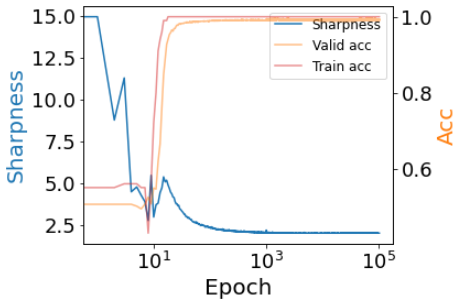


(a) Baseline (b) 1-SAM

Figure 9: **Scenario II with Logistic Loss.** We train a 2-layer MLP with ReLU activation with Bias using gradient descent with weight decay and 1-SAM on $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and training set size $n = 100$. One can observe a distinction between the two settings. The minimum reached by 1-SAM is flatter but the model generalizes much worse and even starts to degenerate after 2000 epochs. The difference between Figures 8b and 9b is similar to the difference between Figures 1b and 4b

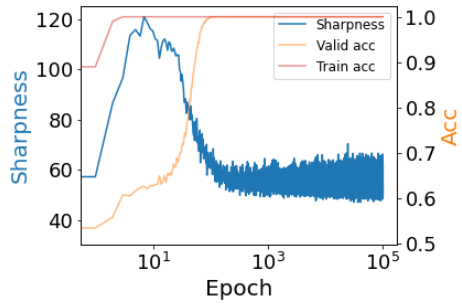


(a) Simplified BatchNorm

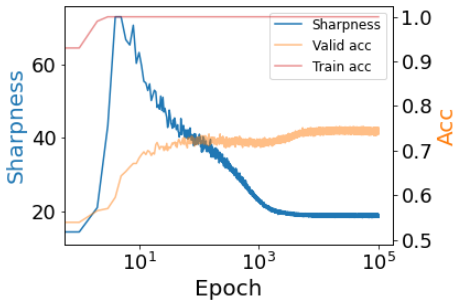


(b) Simplified LayerNorm

Figure 10: **Models with Normalization and Logistic Loss.** We train two-layer ReLU networks with simplified BatchNorm and LayerNorm on data distribution $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and sample complexity $n = 100$ using 1-SAM. We can see that in both cases, the models nearly perfectly generalize.



(a) Baseline



(b) 1-SAM

Figure 11: **Scenario II with Deeper Networks.** We train a 3-layer MLP with ReLU activation with Bias using gradient descent with weight decay and 1-SAM on $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and training set size $n = 100$. One can observe a distinction between the two settings. The minimum reached by 1-SAM is flatter, but the model generalizes much worse.

C.4 Extension To Deeper Networks

Our Theorem 4.1 suggests that memorization solutions can exist for deeper networks with biased terms in the first layer. We perform experiments on deeper networks to verify this. Specifically, we train a 3-layer MLP with ReLU activation with bias term in the first layer on $\mathcal{P}_{\text{xor}}$ with dimension $d = 30$ and training set size $n = 100$. The results are shown in Figure 11. We can see that the flattest minimizer of the 3-layer MLP with bias term in the first layer has a much worse generalization performance than the baseline. This is consistent with our theoretical results.