Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization (NeurIPS 2023)

SeHyun Park

January 8, 2024

Seoul National University

Outline

Introduction

Introduction

- Existing study shows that common stochastic optimizers prefer flatter minimizers of the training loss, and thus flatness implies generalization.
- This paper critically examines this explanation.
- Through theoretical and empirical investigation, they 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.

(3) There exist non-generalizing flattest models, but sharpness minimization algorithms still generalize.

Setup

Setup

- **Data distribution**
  - x_i : sampled uniformly from the hypercube $\{-1, 1\}^d$ for $i = 1, \ldots, n$
  - $y_i = x_i[1]x_i[2]$,
    where $x_i[j]$ be the value of the j-th coordinate of vector x_i.
  - Let $(x_i, y_i) \overset{\text{iid}}{\sim} \mathcal{P}_{\text{xor}}$ for $i = 1, \ldots, n$
- **Architectures**
  - 2-MLP-No-Bias : $f_\theta^{nobias}(x) = W_2 \text{relu}(W_1 x)$ with $\theta = (W_1, W_2)$
  - 2-MLP-Bias : $f_\theta^{bias}(x) = W_2 \text{relu}(W_1 x + b_1)$ with $\theta = (W_1, b_1, W_2)$
  - 2-MLP-Sim-LN : $f_\theta^{sln}(x) = W_2 \frac{\text{relu}(W_1 x + b_1)}{\max\{||\text{relu}(W_1 x + b_1)||_2, \epsilon\}}$ where ϵ is a sufficiently small positive constant

Setup

- **Loss**
  - $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i \right)^2$
- **Sharpness**
  - Using $\mathrm{Tr}\left(\nabla^2 L(\theta)\right)$ to measure how sharp the loss is at θ
- **Interpolating Model**
  - A model f_θ interpolates the dataset $\{(x_i, y_i)\}_{i=1}^{n}$ if and only if $\forall i$, $f_\theta(x_i) = y_i$.

Three Scenario

Scenario I: All Flattest Models Generalize

- Scenario I

Theorem (1.1)

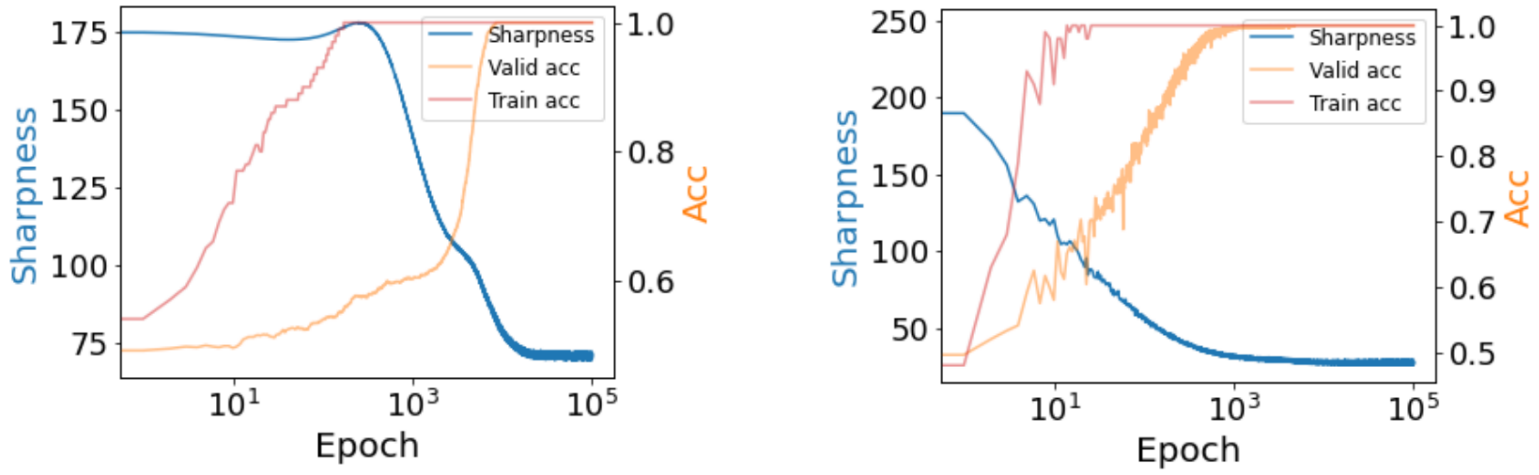
*For any $\delta \in (0,1)$ and input dimension d, for $n = \Omega\left(d \log\left(\frac{d}{\delta}\right)\right)$, with probability at least $1-\delta$ over the random draw of training set $\{(x_i, y_i)\}_{i=1}^{n}$ from $\mathcal{P}_{\mathrm{xor}}^{n}$, let $L(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(f_{\theta}^{nobias}(x_i)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

$$\mathrm{E}_{x,y\sim\mathcal{P}_{xor}}\left[\left(f_{\theta^*}^{nobias}(x) - y\right)^2\right] = O\left(\frac{d}{n} \cdot log\left(\frac{d}{n}\right)\right)$$

- This shows that for $\mathcal{P}_{\mathrm{xor}}$, flat models can generalize under almost linear sample complexity with respect to the input dimension.

Scenario I: All Flattest Models Generalize

- SAM empirically finds the flattest model that generalizes.



(a) Baseline

(b) 1-SAM

Figure 1: Scenario I. Training a 2-layer MLP with ReLU activation without bias using gradient descent with weight decay and SAM on $\mathcal{P}_{\text{xor}}$ with batch size 1, dimension $d = 30$ and training set size $n = 100$.

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

- Scenario II

Definition (2.1)

(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 (2.2)

(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 training 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.

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

Theorem (2.1)

If the input data points $\{x_i\}$ of the training set form a set of extreme points (Definition 2.1), then there exists a width-n ***2-MLP-Bias*** *that is a memorizing solution (Definition 2.2) for the training dataset and has minimal sharpness over all the interpolating solutions.*

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

Proposition (2.1)

For data distribution $\mathcal{P}_{xor}$, for any number of samples n, there exists a width-n ***2-MLP-Bias*** *that memorizes the training set as in Theorem 2.1, reaches minimal sharpness over all the interpolating models and has generalization error $\max\{1 - n/2^d, 0\}$ measured by zero one error.*

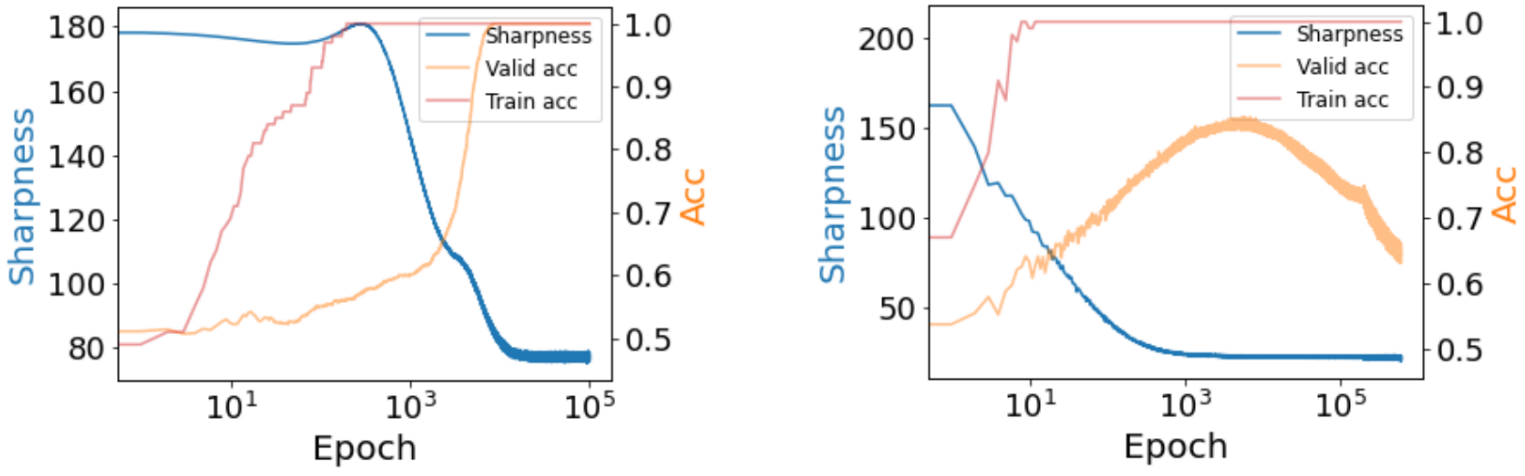
Proposition (2.2)

For data distribution $\mathcal{P}_{xor,}$, 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.*

- Prop. (2.1) and (2.2) show that both flattest generalizing and non-generalizing models with architecture **2-MLP-Bias** exist.

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

- SAM empirically finds the non-generalizing solutions.



(a) Baseline

(b) 1-SAM

Figure 2: Scenario II. Training a 2-layer MLP with ReLU activation with Bias using gradient descent with weight decay and SAM on $\mathcal{P}_{\text{xor}}$ with batch size 1, dimension $d = 30$ and training set size $n = 100$.

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

- Scenario III

Theorem (3.1)

If the input data points $\{x_i\}$ of the training set form a set of extreme points (Definition 2.1), for sufficiently small ϵ, then there exists a width-n ***2-MLP-Sim-LN*** *with hyperparameter ϵ that is a memorizing solution (Definition 2.2) for the training dataset and has minimal sharpness over all the interpolating solutions.*

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

Proposition (3.1)

*For data distribution $\mathcal{P}_{xor}$, 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 3.1, reaches minimal sharpness over all the interpolating models and has generalization error $\max\left\{1 - n/2^d, 0\right\}$ measured by zero one error.*

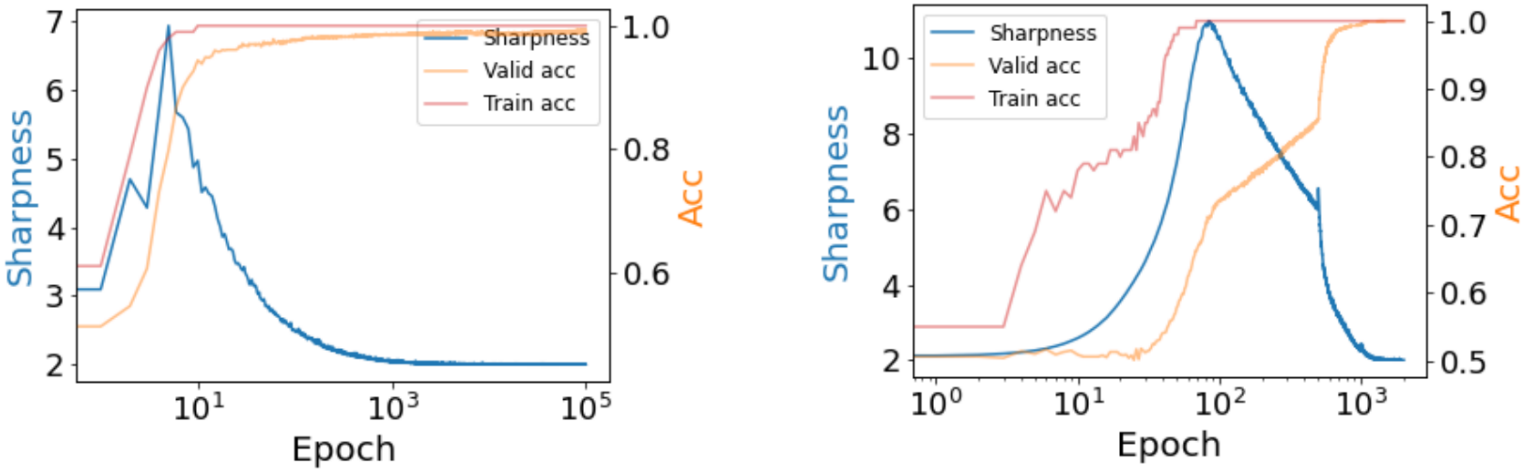
Proposition (3.2)

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

- Prop. (3.1) and (3.2) show that both flattest generalizing and non-generalizing models with architecture **2-MLP-Sim-LN** exist.

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

- SAM empirically finds generalizing models.



(a) Baseline

(b) 1-SAM

Figure 3: : Scenario III. Training two-layer ReLU networks with simplified LayerNorm on data distribution $\mathcal{P}_{\mathrm{xor}}$ with dimension $d = 30$ and sample complexity $n = 100$ using SAM with batch size 1.

Conclusion

Conclusion

▶ Authors present theoretical and empirical evidence for whether sharpness minimization implies generalization subtly depends on the choice of architectures and data distributions.

Architecture	All Flattest Minimizers Generalize Well.	Sharpness Minimization Algorithms Generalize.
2-layer w/o Bias	✓ (Theorem 1.1)	✓
2-layer w/ Bias	✗ (Theorem 2.1)	✗
2-layer w/ simplified LayerNorm	✗ (Theorem 3.1)	✓

Figure 4: Summary of results

▶ Limitations

- The setup is too simplistic. No noise is considered in the label $y_i = x_i[1]x_i[2]$.
- Results only cover a small subset of existing architectures.

End