Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity

Chulhee Yun, Suvrit Sra, Ali Jadbabaie

Laboratory for Information and Decision Systems, MIT

Introduction

Finite sample memorization is not well understood

Overparametried NNs with SGD memorize even random noise.

Given a network, can it memorize arbitrary datasets?

Results on function approximation are not very helpful.

Recent results on fully-connected, residual, convolutional networks require N hidden nodes to memorize N data points!

Can we exploit depth to memorize with less hidden nodes?

Main results

Tight memorization capacity of ReLU Networks

For 3-layer networks, $\Theta(\sqrt{N})$ hidden nodes are <u>necessary and sufficient</u> for memorizing N arbitrary data points.

ImageNet ($N =$ 1M, 1k classes) can be memorized with 4-layer ReLU networks with hidden layer size 2k-2k-4k.

L-layer network with W params: memorization capacity = $\Omega(W)$

If $L = 2,\ 3$: memorization capacity = $O(W)$ (tight)

If $L > 3$: memorization capacity = $O(WL \log W)$ (nearly tight)

Main results

Finite sample expressivity of residual networks

For ReLU ResNet with input/output dimension d_x/d_y,
$\Omega(N/d_x + d_y)$ hidden nodes are sufficient for memorizing N points

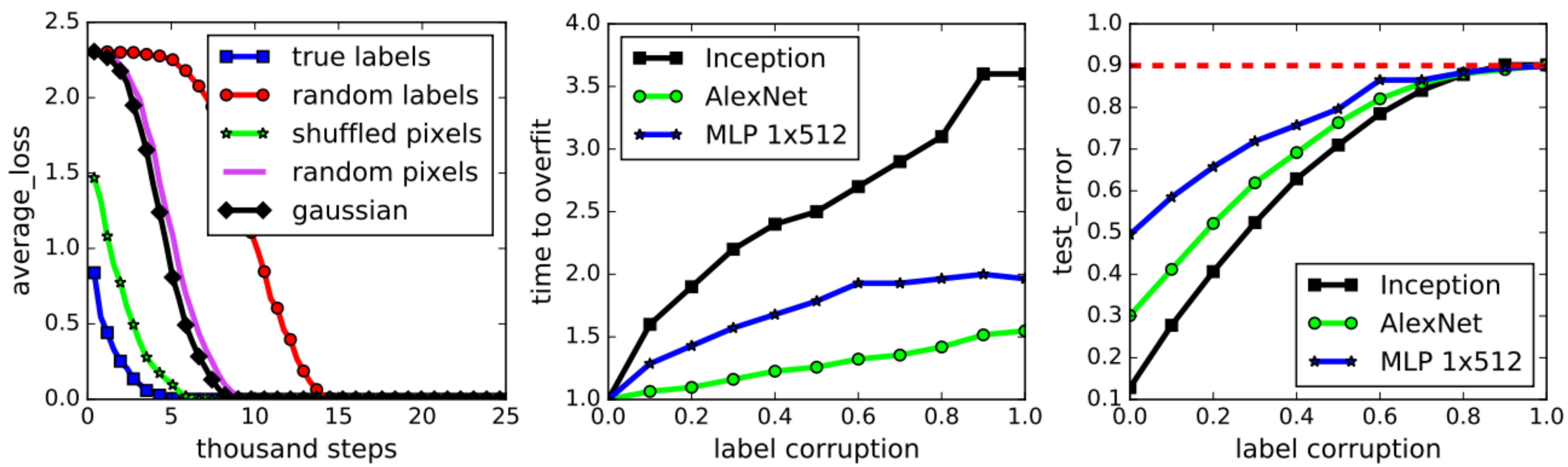
Trajectory of SGD near memorizing global minima

Without-replacement mini-batch SGD finds a point
with small risk when initialized close to global minima

(Long version starts here)

Memorization phenomenon in NNs

- Overparametrized NNs trained with SGD can memorize even random noise. [Zhang et al., 2017]

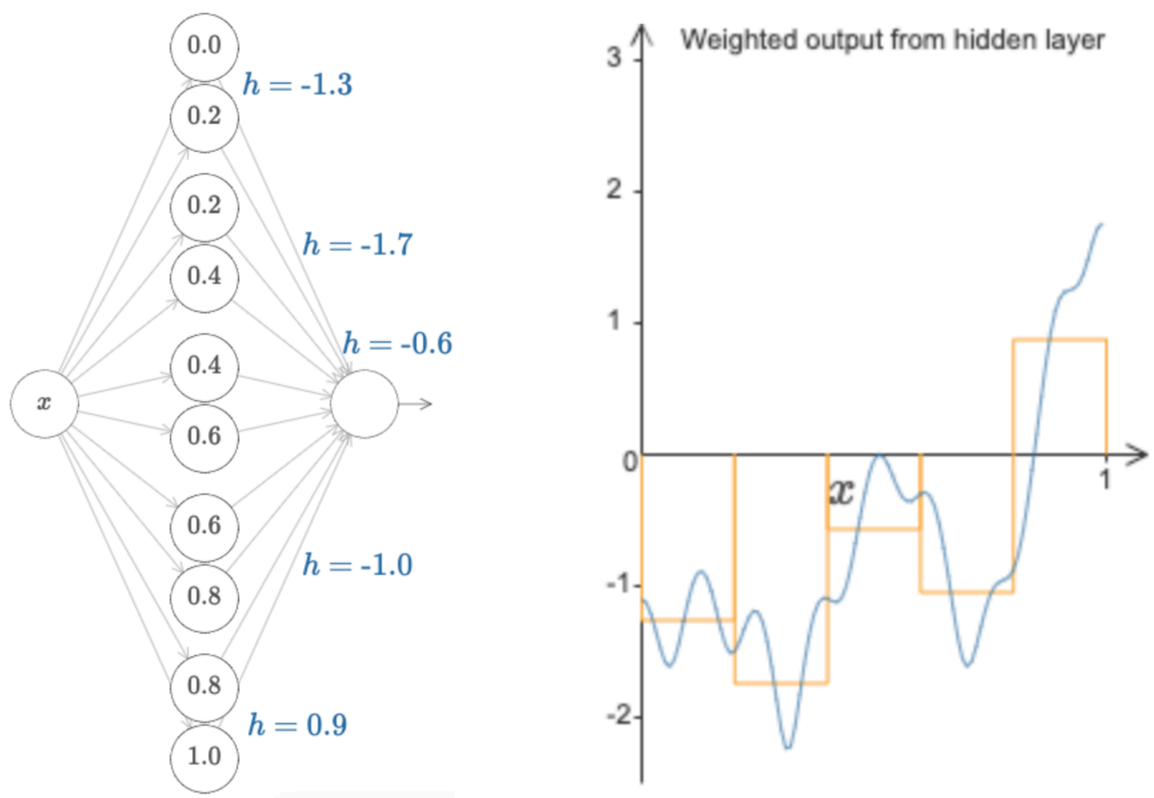


(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

Expressive power theory

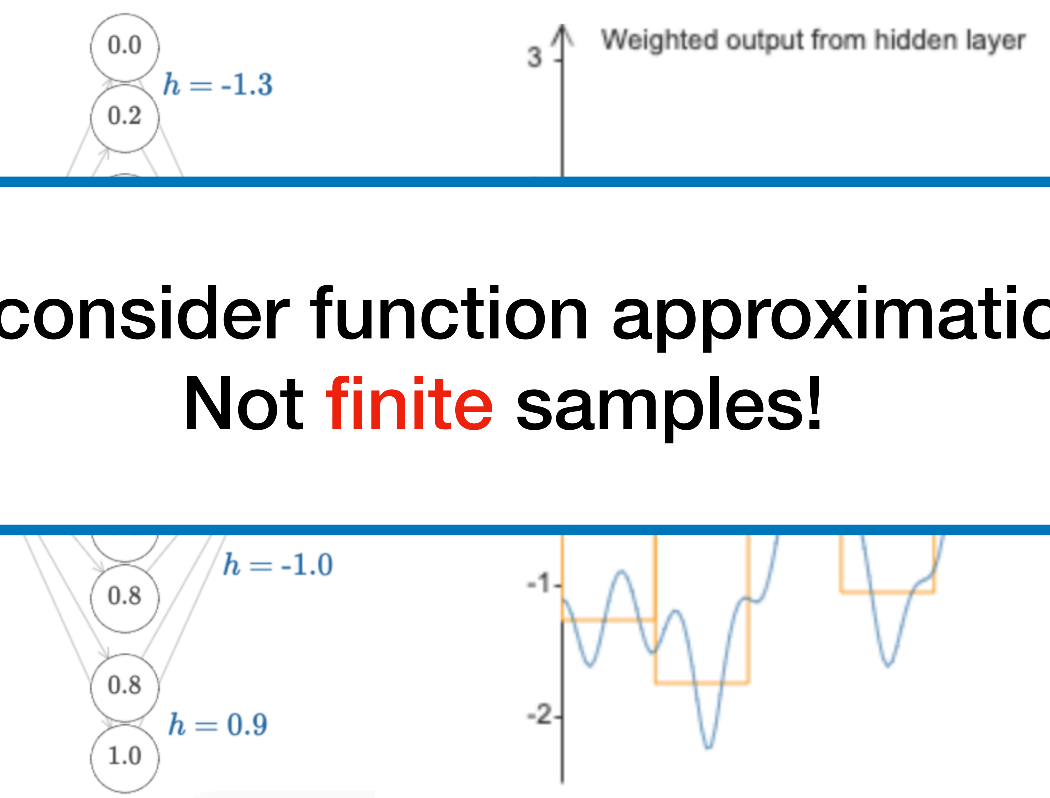
- To understand memorization phenomenon, it is important to understand NN's expressive power.
- Expressive power is a classic topic in NN theory, e.g., universal approximation theorem. [Cybenko, '89, Hornik, '91, …]



(http://neuralnetworksanddeeplearning.com/chap4.html)

Expressive power theory

- To understand memorization phenomenon, it is important to understand NN's expressive power.
- Expressive power is a classic topic in NN theory, e.g., universal approximation theorem. [Cybenko, '89, Hornik, '91, …]



Majority of results consider function approximation (infinite points), Not finite samples!

(http://neuralnetworksanddeeplearning.com/chap4.html)

Finite sample expressivity

Def. We define (universal) **finite sample expressivity** of a neural network $f_{\boldsymbol{\theta}}(\cdot)$ as the network's ability to satisfy the following:

For all $\{x_i\}_{i=1}^N \in \mathbb{R}^{d_x \times N}$ and for all $\{y_i\}_{i=1}^N \in \mathbb{R}^{d_y \times N}$, there exists a parameter $\boldsymbol{\theta}$ s.t. $f_{\boldsymbol{\theta}}(x_i) = y_i$ for all $1 \leq i \leq N$.

i.e., the net can memorize arbitrary dataset with N points.

Memorization capacity

Def. For $d_y = 1$, we define **memorization capacity** to be:

The maximum N such that for all $\{x_i\}_{i=1}^{N} \in \mathbb{R}^{d_x \times N}$ and for all $\{y_i\}_{i=1}^{N} \in \mathbb{R}^{N}$, there exists a parameter $\boldsymbol{\theta}$ s.t. $f_{\boldsymbol{\theta}}(x_i) = y_i$ for all i.

i.e., the **maximum value** of N for which the network has finite sample expressivity.

Comparison to VC dimension

- Definition of memorization capacity:

The maximum N such that **for all** $\{x_i\}_{i=1}^N \in \mathbb{R}^{d_x \times N}$ and for all $\{y_i\}_{i=1}^N \in \mathbb{R}^N$, there exists a parameter $\boldsymbol{\theta}$ s.t. $f_{\boldsymbol{\theta}}(x_i) = y_i$ for all i.

- Recall the definition of VC dimension:

The maximum N such that **there exists** $\{x_i\}_{i=1}^N \in \mathbb{R}^{d_x \times N}$ s.t. for all $\{y_i\}_{i=1}^N \in \{\pm 1\}^N$, there exists a parameter $\boldsymbol{\theta}$ s.t. $f_{\boldsymbol{\theta}}(x_i) = y_i$ for all i.

memorization capacity ≤ VC dimension

Previous works

- Classical works focus on memorization capacity of NNs with activations such as **linear threshold** or **sigmoid** [Cover, 1965; Baum, 1988; Huang & Huang, 1991; Huang & Babri, 1998; Huang, 2003; etc…]

- Recent results on **modern** architectures, for:

  - ReLU fully-connected NNs (FNNs) [Zhang et al., 2017]

  - Residual networks (ResNets) [Hardt & Ma, 2017]

  - Convolutional neural networks (CNNs) [Nguyen & Hein, 2017]

Previous works

- However, recent results impose strong assumptions on the number of hidden nodes!

- A 1-hidden-layer ReLU network with N **hidden nodes** can memorize any arbitrary dataset with N points.
  [Zhang et al., 2017]

- Results on ResNets and CNNs require N **hidden nodes**.
  [Hardt & Ma, 2017, Nguyen & Hein, 2017]

Previous works

- However, recent results impose strong assumptions on the number of hidden nodes!

- A 1-hidden-layer ReLU network with N **hidden nodes** can memorize any arbitrary dataset with N points.
  [Zhang et al., 2017]

Can we use depth to memorize with less hidden nodes?

Agenda

Tight memorization capacity of fully-connected NNs

Number of hidden nodes necessary and sufficient for memorization

Memorization capacity of residual networks

Number of hidden nodes sufficient for memorization

Trajectory of SGD near memorizing global minima

Analysis of without-replacement SGD near global minima

Finite sample expressivity of FNNs

- Training data $\{(x_i, y_i)\}_{i=1}^N, \;\; x_i \in \mathbb{R}^{d_x}, \;\; y_i \in \mathbb{R}^{d_y}$

- **Assumption**: all x_i's are distinct and all $y_i \in [-1,1]^{d_y}$.

- Fully-connected neural networks

$$a^0(x) = x, \;\; a^l(x) = \underbrace{\sigma}_{\substack{\text{Activation}\\\text{function}}}(\underbrace{W^l}_{\substack{\text{Weight}\\\text{matrix}}} a^{l-1}(x) + \underbrace{b^l}_{\substack{\text{Bias}\\\text{vector}}}), \;\; l \in \{1,\ldots,L-1\}, \;\; f_{\boldsymbol{\theta}}(x) = W^L a^{L-1}(x) + b^L$$

- Activation $\sigma(t) = \max\{s_+ t, s_- t\}, \; s_+ > s_- \geq 0.$
  (includes ReLU and Leaky ReLU)

Sufficiency results

Theorem 1.

A 2-hidden-layer ReLU network with hidden layer dimensions $d_1 d_2 \geq 4Nd_y$ can memorize any arbitrary dataset with N distinct points.

$d_1 = d_2 = 2\sqrt{Nd_y}$ **suffices!**

Proposition 2 (classification).

A 3-hidden-layer ReLU network with hidden layer dimensions $d_1 d_2 \geq 4N$ and $d_3 \geq 4d_y$ can memorize any arbitrary classification data dataset with N distinct points.

Necessity results

Theorem 3. ($d_y = 1$)

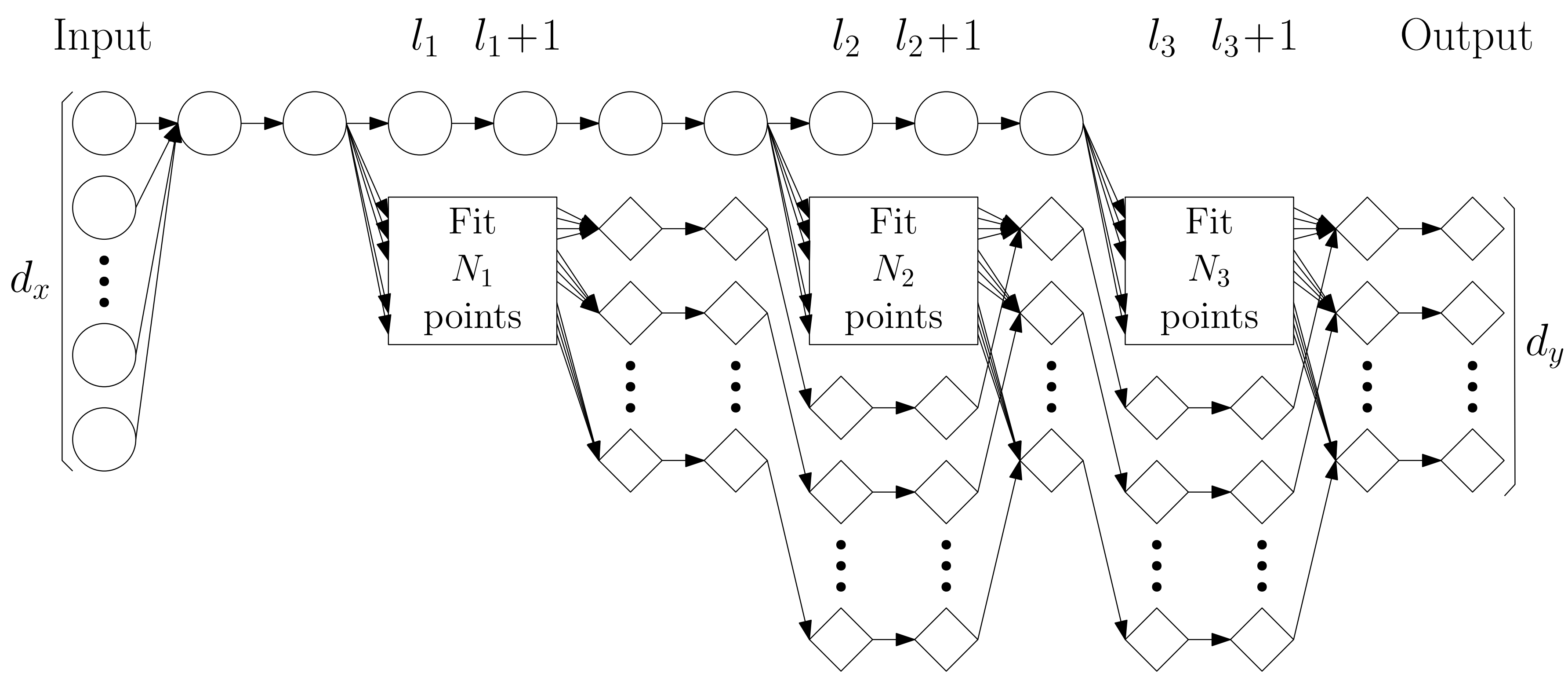
A 1-hidden-layer ReLU network with $d_1 + 2 < N$, or
a 2-hidden-layer ReLU network with $2d_1d_2 + d_2 + 2 < N$
can **not** memorize any arbitrary dataset with N points.
(i.e., there exist datasets that they fail to memorize)

Discussion

- Depth-width tradeoff in finite sample expressivity
  - **Necessary** and **sufficient** width for memorizing ($d_y = 1$): 1-hidden-layer $\Theta(N)$ vs 2-hidden-layer $\Theta(\sqrt{N})$
  - For d_y-class classification, $\Omega(\sqrt{Nd_y})$ requirement of 2-hidden-layer improves to $\Omega(\sqrt{N} + d_y)$ by one more layer
- ImageNet ($N \approx 10^6, d_y = 10^3$) memorized with a 2k-2k-4k FNN.
- Surprisingly small network size is required to memorize + achieve zero training loss at global minimum.

Extension to deeper networks

- Extension to deeper networks possible:
  if there are $\Omega(Nd_y)$ parameters between hidden layers,
  the network can memorize N points.



Tight bounds on capacity

- L-layer network with W params, $d_y = 1$
- $\Omega(N)$ parameters sufficient to memorize N points
  $\Longrightarrow$ a lower bound $\Omega(W)$ on memorization capacity
- Theorem 3 $\Longrightarrow$ For $L = 2,\ 3$, capacity = $O(W)$

Tight!

- Upper bound on VC dim $O(WL \log W)$ [Bartlett et al., 2019]
  $\Longrightarrow$ For $L > 3$, capacity = $O(WL \log W)$

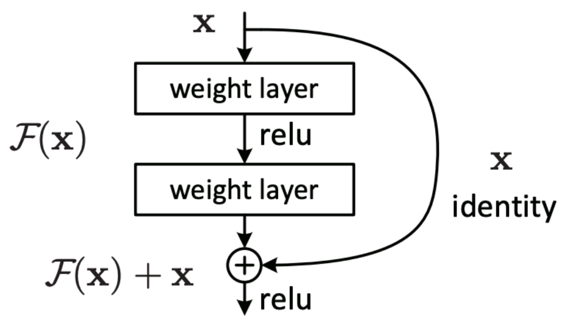
Almost tight!

Finite sample expressivity of ResNets

- Assumption: data points x_i's are in general position, i.e., no $d_x + 1$ data points lie on the same affine hyperplane.

- Assumption: $y_i \in \{0,1\}^{d_y}$ is a one-hot encoding.

- Residual network (ResNet)



$$h^0(x) = x,$$

$$h^l(x) = h^{l-1}(x) + V^l \sigma(U^l h^{l-1}(x) + b^l) + c^l, \quad l \in \{1,\ldots,L-1\}$$

$$g_\theta(x) = V^L \sigma(U^L h^{L-1}(x) + b^L) + c^L$$

- d_l is the # hidden nodes in l-th residual layer

Sufficiency result for ResNets

Theorem 4.

A ResNet with hidden layer dimensions $\sum_{l=1}^{L-1} d_l \geq \frac{4N}{d_x} + 4d_y$ and $d_L \geq 2d_y$ can memorize any arbitrary classification dataset with N points.

- Under a different assumption, improve requirement $N + d_y$ [Hardt & Ma, 2017] to $\frac{4N}{d_x} + 6d_y$.
- For CIFAR-10 ($N =$ 50k, $d_x =$ 3,072, $d_y =$ 10): 50,010 nodes vs 126 nodes

SGD near memorizers

- We want to solve the empirical risk minimization problem:

$$\text{minimize}_{\boldsymbol{\theta}} \quad \mathfrak{R}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} \ell(f_{\boldsymbol{\theta}}(x_i); y_i)$$

- **Assumption.** The loss $\ell(z; y)$ is strictly convex and three times differentiable in z. For any y, there exists a global minimizer z of $\ell(z; y)$.

- **Def.** A point $\boldsymbol{\theta}^*$ is a **memorizing global minimum** of $\mathfrak{R}(\boldsymbol{\theta})$ if $\ell'(f_{\boldsymbol{\theta}^*}(x_i); y_i) = 0$ for all $1 \leq i \leq N$.

SGD near memorizers

- We analyze without-replacement mini-batch SGD, with mini-batch size B.
- At every $E = N/B$ steps, dataset reshuffled and partitioned into $S^{(kE)}, S^{(kE+1)}, \ldots, S^{(kE+E-1)}$
- SGD update

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \frac{\eta}{B} \sum_{i \in S^{(t)}} \nabla_{\boldsymbol{\theta}} \ell(f_{\boldsymbol{\theta}^{(t)}}(x_i); y_i)$$

SGD near memorizers

Theorem 5 (informal).

If the initialization $\boldsymbol{\theta}^{(0)}$ satisfies $\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\| \leq \rho$ for some memorizing global minimum $\boldsymbol{\theta}^*$ and small constant ρ, initialization satisfies $\mathfrak{R}(\boldsymbol{\theta}^{(0)}) - \mathfrak{R}(\boldsymbol{\theta}^*) = O(\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\|^2)$.
If we run SGD with small enough η, it finds a point $\boldsymbol{\theta}$ that satisfies

$$\mathfrak{R}(\boldsymbol{\theta}) - \mathfrak{R}(\boldsymbol{\theta}^*) = O(\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\|^4), \text{ and}$$
$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 2\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\|.$$

SGD near memorizers

- Theorem restricted to initialization **very close** to memorizing global minima
- However, holds **without** any *width/depth requirement* on the network or *distributional assumption* on data— the only requirement: $\boldsymbol{\theta}^*$ **memorizes** the data.
- Completely **deterministic**, independent of the partition of dataset taken by SGD
- The behavior of SGD after finding $\boldsymbol{\theta}$ is not well understood

Thank you for your attention!

Reference:

Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity
https://arxiv.org/abs/1810.07770