NEURAL INFORMATION
PROCESSING SYSTEMS

The Cost of Robustness: Tighter Bounds on Parameter Complexity for Robust Memorization in ReLU Nets

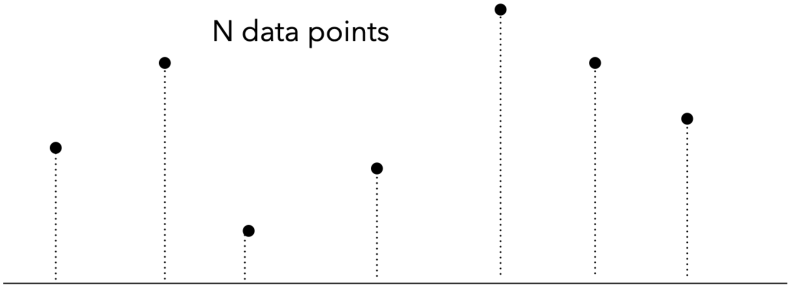
Yujun Kim*, Chaewon Moon*, Chulhee Yun

KAIST AI
Kim Jaechul Graduate School

OptiML Optimization & Machine Learning Laboratory

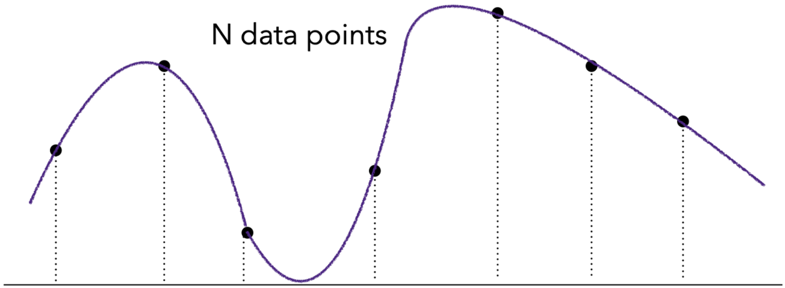
NeurIPS 2025

Memorization



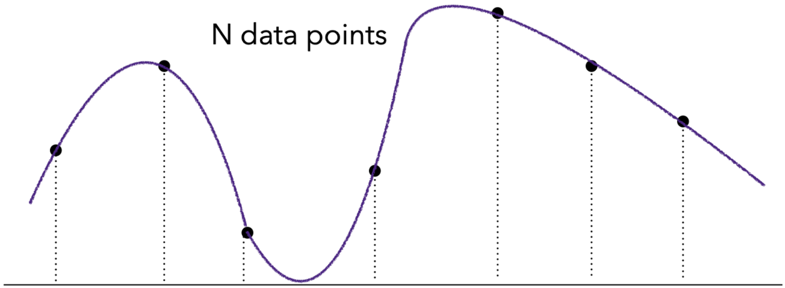
Find a network that maps all N data points to their corresponding labels.

Memorization



Find a network that maps all N data points to their corresponding labels.

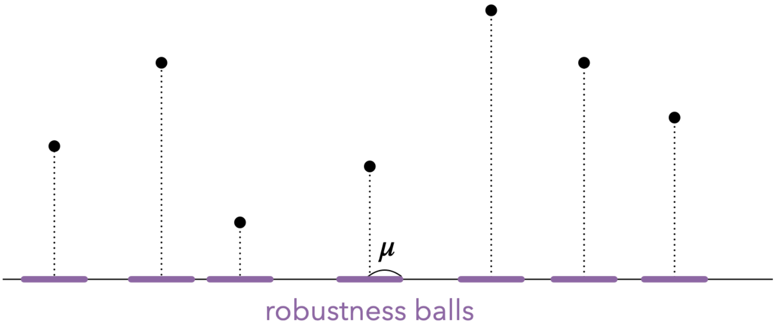
Memorization



Find a network that maps all N data points to their corresponding labels.

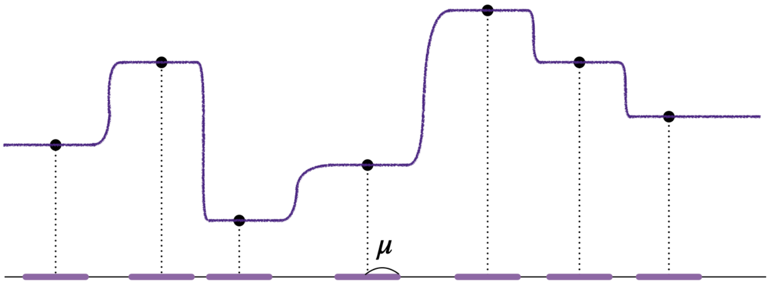
Prior work shows that memorizing N data points requires $\Theta(\sqrt{N})$ parameters.

Robust Memorization



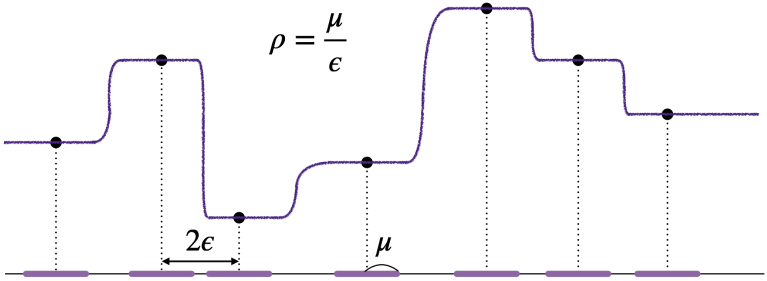
Find a network that maps all points within distance μ to their corresponding labels.

Robust Memorization



Find a network that maps all points within distance μ to their corresponding labels.

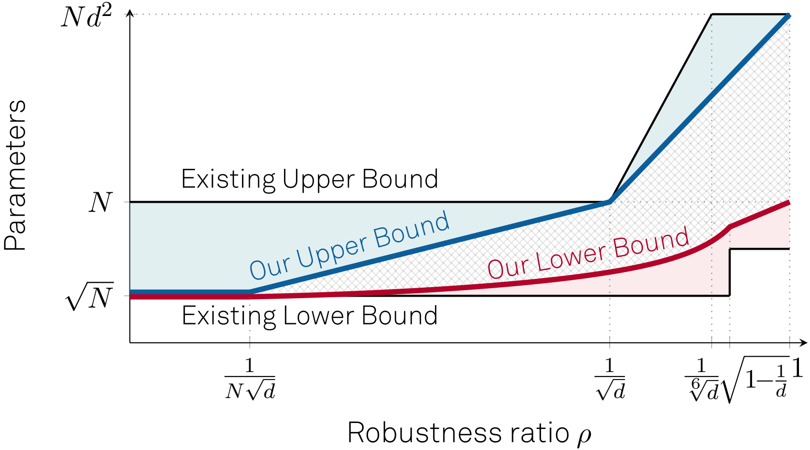
Robust Memorization



Find a network that maps all points within distance μ to their corresponding labels.

Its difficulty depends on the robustness ratio $\rho = \dfrac{\mu \quad \text{(robustness radius)}}{\epsilon \quad \text{(separation constant)}}$.

Contribution



N: Number of samples
d: Input dimension

- We provide tighter upper and lower bounds depending on ρ, covering the entire range $\rho \in (0, 1)$.

Lower Bounds

Theorem 3.1.

For given $\rho \in (0,1)$, if a trainable ReLU network can ρ-robustly memorize any N points in $\mathbb{R}^d$, it must have

$$P = \Omega\Big(\underbrace{\big(\rho^2 \min\{N, d\} + 1\big)\, d}_{①} + \underbrace{\min\{1/\sqrt{1-\rho^2}, \sqrt{d}\}\sqrt{N}}_{②}\Big)$$

trainable parameters.

① : Derived from the necessary condition on the width of first hidden layer.

② : Derived from the VC-dimension bound.

Upper Bounds

Theorem 4.2.

For any dataset $\mathcal{D} \in \boldsymbol{D}_{d,N,C}$ and $\eta \in (0,1)$, the following statements hold:

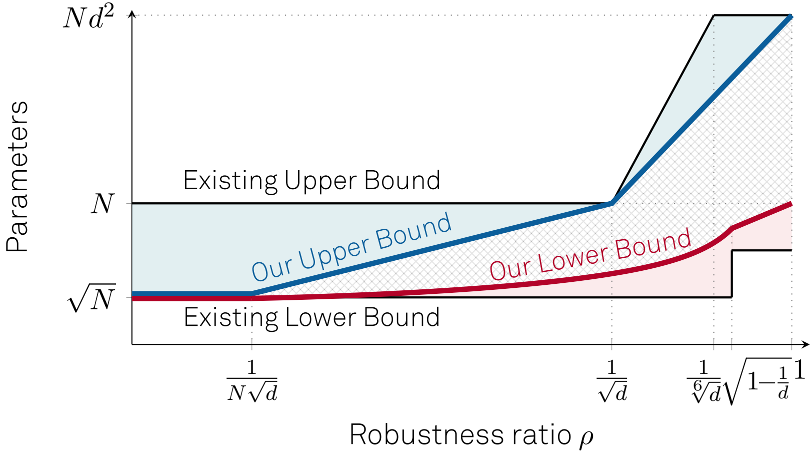
(i) If $\rho \in \left(0, \frac{1}{5N\sqrt{d}}\right]$, there exists $f \in \mathcal{F}_{d,P}$ with $P = \tilde{O}(\sqrt{N})$ that ρ-robustly memorizes $\mathcal{D}$.

(ii) If $\rho \in \left(\frac{1}{5N\sqrt{d}}, \frac{1}{5\sqrt{d}}\right]$, there exists $f \in \mathcal{F}_{d,P}$ with $P = \tilde{O}(Nd^{\frac{1}{4}}\rho^{\frac{1}{2}})$ that ρ-robustly memorizes $\mathcal{D}$ with error at most η.

(iii) If $\rho \in \left(\frac{1}{5\sqrt{d}}, 1\right)$, there exists $f \in \mathcal{F}_{d,P}$ with $P = \tilde{O}(Nd^2\rho^4)$ that ρ-robustly memorizes $\mathcal{D}$.

- (i)-(iii) are based on the Johnson-Lindenstrauss lemma.
- (i) and (ii) rely on the mapping from the grid to the lattice.

Conclusion



- For small ρ, robust memorization does not require higher cost than memorization.
- The cost of achieving robustness increases with larger ρ.

Poster Session 4

Thu 4 Dec

4:30 p.m. - 7:30 p.m.

https://arxiv.org/abs/2510.24643