Max-Margin Works while Large Margin Fails: Generalization without Uniform Convergence

Kiarash Joolaei
Department of Computer Engineering

October 12, 2025

Introduction

- Many machine learning methods use **Uniform Convergence** to guarantee model generalization (Direct Implication)
- There are setups that UC doesn't hold
- **Main Question:** Is proving generalization possible for setups where UC fails?

Problem Setup

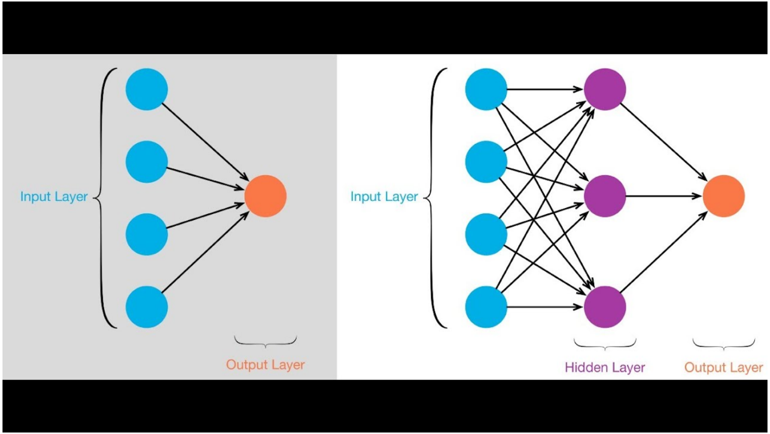


Figure: The paper proves generalization bounds for a **linear** and a **non-linear** setting

Linear Setting

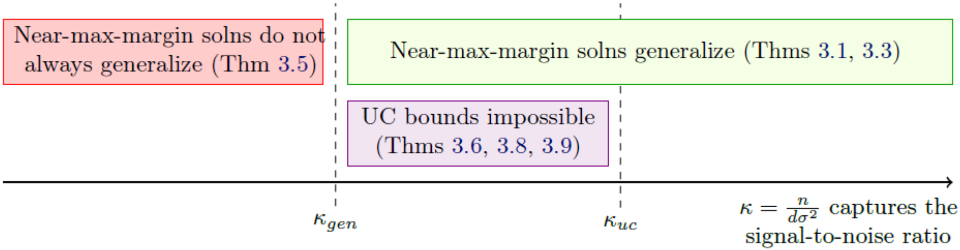
- **Data Distribution:** Fix some ground truth unit vector direction $\mu \in \mathbb{R}^d$. Let $x = z + \xi$, where $z \sim$ Uniform$(\{\mu, -\mu\})$, and ξ is uniform on the sphere of radius $\sqrt{d-1}\sigma$ in $d-1$ dimensions orthogonal to the direction of μ. Let $y = \mu^T x$ such that $y = 1$ with probability 0.5 and -1 with probability 0.5. Denote this distribution of (x, y) as $\mathcal{D}_{\mu,\sigma,d}$.
- **Model:** We learn a model $w \in \mathbb{R}^d$ that predicts $\hat{y} = \text{sign}(f_w(x))$ where $f_w(x) = w^T x$.

Non-linear Setting

- **Data Distribution:** Fix some ground truth unit vector directions $\mu_1, \mu_2 \in \mathbb{R}^d$. Let $x = z + \xi$, where $z \sim \text{Uniform}(\{\mu_1, -\mu_1, \mu_2, -\mu_2\})$, and ξ is uniform on the sphere of radius $\sqrt{d-2}\sigma$ in $d-2$ dimensions orthogonal to the direction of μ. Let $y = (\mu_1^T x)^2 - (\mu_2^T x)^2 = \text{XOR}((\mu_1 + \mu_2)^T x, (-\mu_1 + \mu_2)^T x)$. Denote this distribution of (x, y) as $\mathcal{D}_{\mu_1, \mu_2 \sigma, d}$.
- **Model:** Fix $a \in \{-1, 1\}^m$ so that $\sum_i a_i = 0$. The model is $f_W(x) = \sum_{i=1}^m a_i \phi(w_i^T x)$, where $W \in \mathbb{R}^{m \times d}$ and $\phi(z) = \max(0, z)^h$ for $h \in [1, 2)$. We also assume m is divisible by 4.

We also define $\Omega_{\sigma,d}^{\text{linear}} := \{\mathcal{D}_{\mu,\sigma,d} : \mu \in \mathbb{R}^d, \|\mu\| = 1\}$ and $\Omega_{\sigma,d}^{\text{XOR}} := \{\mathcal{D}_{\mu_1,\mu_2,\sigma,d} : \mu_1 \perp \mu_2 \in \mathbb{R}^d, \|\mu_1\| = \|\mu_2\| = 1\}$

Generalization in SNR Regions



Large Dimension Assumption

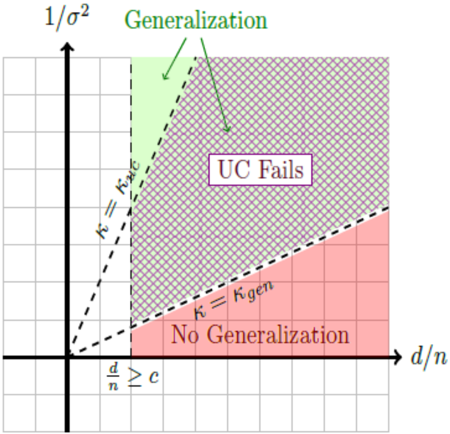


Figure: All the results in the paper require the assumption that $d \geqslant \Omega(n)$.

Preliminaries

In machine learning, the goal is to learn a hypothesis function h. One considers **global hypothesis class** $\mathcal{G}$, e.g., all two-layer neural networks. The learning is performed on a smaller subset $\mathcal{H} \subseteq \mathcal{G}$, meaning $h \in \mathcal{H}$, e.g., all two-layer neural networks with bounded norm.

Definition

For any loss function $\mathcal{L} : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$, and a hypothesis mapping $h : \mathcal{X} \longrightarrow \mathbb{R}$, the **test loss** on a distribution $\mathcal{D}$ is defined as $\mathcal{L}_{\mathcal{D}}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}(h(x), y)\right]$. For a set of examples $S = \{(x_i, y_i)\}_{i\in[n]}$, we define $\mathcal{L}_S(h) := \mathbb{E}_{i\sim[n]}\left[\mathcal{L}(h(x_i), y_i)\right]$

From now on, we assume $\mathcal{L}(y', y) = 1\{\text{sign}(y) \neq \text{sign}(y')\}$

Preliminaries (Uniform Convergence Bound)

Definition

A **two-sided** uniform convergence bound with parameter $\epsilon_{\textsf{unif}}$ for a problem class Ω, a set of hypotheses $\mathcal{H}$, and loss $\mathcal{L}$ is a bound that guarantees that for any $\mathcal{D} \in \Omega$, and for some $\delta \in (0,1)$

$$\mathsf{Pr}_{S \sim \mathcal{D}^n}\left(\sup_{h \in \mathcal{H}} |\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h)| \geqslant \epsilon_{\textsf{unif}}\right) \leqslant 1 - \delta$$

The **one-sided** version guarantees

$$\mathsf{Pr}_{S \sim \mathcal{D}^n}\left(\sup_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h) \geqslant \epsilon_{\textsf{unif}}\right) \leqslant 1 - \delta$$

In all of the future results, we consider $\delta = \frac{3}{4}$

Preliminaries (Useful Hypothesis Class)

Definition

A hypothesis class $\mathcal{H}$ is useful with respect to an algorithm $\mathcal{A}$ over a problem class Ω with confidence δ, if for any $\mathcal{D} \in \Omega$

$$\Pr_{S \sim \mathcal{D}^n} \left(\mathcal{A}(S) \in \mathcal{H} \right) \geqslant \delta$$

Here we also have $\delta = \frac{3}{4}$.
In this definition, alongside the UC Bound definition, instead of considering a single distribution $\mathcal{D}$, a **class of distributions** Ω is considered.

Preliminaries (Margin)

Definition

The **margin** $\gamma(h, S)$ of a classifier h on a sample S equals $\min_{(x,y)\in S} yh(x)$.
The **normalized margin** for a scalar c and an h-homogeneous function f_W ($f_{cW}(x) = c^h f_W(x)$) is defined as:

$$\bar{\gamma}(f_W, S) := \frac{\gamma(f_W, S)}{\|W\|^h} = \gamma(f_{\frac{W}{\|W\|}}, S)$$

where $\|W\| := \sqrt{\mathbb{E}_{i\sim[m]}\left[\|w_i\|^2\right]}$. The **maximum normalized margin** is defined as: $\gamma^*(S) := \sup_{W:\|W\|\leqslant 1} \gamma(f_W, S)$

Preliminaries (Near-Max-Margin Solution)

Definition

Let $\epsilon > 0$. A classifier h is a $(1-\epsilon)$-**max-margin solution** for S if

$$\gamma(h, S) \geqslant (1-\epsilon)\gamma^*(S)$$

We refer to a bound that holds for $(1-\epsilon)$-max-margin solutions as an **extremal margin bound**.

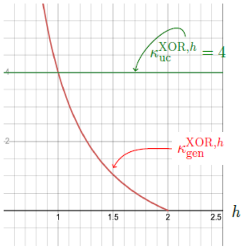
Main Results

- For the linear problem:

$$\kappa_{\text{gen}}^{\text{linear}} := 0, \quad \kappa_{\text{uc}}^{\text{linear}} := 1$$

- For the XOR problem with activation ReLU^h, for $h \in [1,2)$:

$$\kappa_{\text{gen}}^{\text{XOR},h} := \text{the solution to } 2^{\frac{1}{h}}\sqrt{\frac{2}{\kappa}} = \sqrt{\frac{\kappa}{4+\kappa}} + \sqrt{\frac{16}{\kappa(4+\kappa)}}, \quad \kappa_{\text{uc}}^{\text{XOR},h} := 4$$



ReLU^h Intuition

- If $h = 1$, then $\kappa_{\text{gen}}^{\text{XOR},h} = \kappa_{\text{uc}}^{\text{XOR},h}$. Thus we do not a regime with UC failure, but max-margin solutions generalize.

Extremal-Margin Generalization for Linear Problem

This theorem states that when $\kappa > \kappa_{\text{gen}}$, any near-max-margin solution generalizes.

Theorem

Let $\delta > 0$. There exist constants $\epsilon = \epsilon(\delta)$ and $c = c(\delta)$ such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega_{\sigma,d}^{\text{linear}}$ satisfying $\kappa_{\text{gen}}^{\text{linear}} + \delta \leqslant \kappa \leqslant \frac{1}{\delta}$, and $d \geqslant cn$, then with probability $1 - 3e^{-n}$ over the randomness of a training set $S \sim \mathcal{D}^n$, for any $w \in \mathbb{R}^d$ that is a $(1-\epsilon)$-max-margin solution, we have $\mathcal{L}_{\mathcal{D}}(f_w) \leqslant e^{-\frac{n}{64d\sigma^4}} + e^{-\frac{n}{8}}$

Extremal-Margin Generalization for XOR on Neural Network

A similar generalization result holds for XOR problem learned on two-layer neural networks.

Theorem

Let $h \in (1, 2)$, and let $\delta > 0$. There exist constants $\epsilon = \epsilon(\delta)$ and $c = c(\delta)$ such that the following holds. For any n, d, σ and $\mathcal{D} \in \Omega_{\sigma,d}^{XOR}$ satisfying $\kappa = \frac{n}{d\sigma^2} \geqslant \kappa_{gen}^{XOR,h} + \delta$ and $d \geqslant cn$, then with probability $1 - 3e^{-\frac{n}{c}}$ over the randomness of a training set $S \sim \mathcal{D}^n$, for any two-layer neural network with activation function ReLU^h and weight matrix W that is a $(1-\epsilon)$-max-margin solution, we have $\mathcal{L}_{\mathcal{D}}(f_W) \leqslant e^{-\frac{1}{c\sigma^2}}$

This result is meaningful whenever σ is small enough (in terms of δ), because we assumed that $\frac{d}{n} \in \left[c, \frac{1}{\sigma^2(\kappa_{\text{gen}}^{\text{XOR},h}+\delta)}\right]$ and this interval needs to be non-empty. Also, test loss tends to zero as σ approaches 0.

Region where Max-Margin Generalization not Guaranteed

If $\kappa < \kappa_{\text{gen}}$, it is possible that a near-max margin solution does not generalize at all. Since $\kappa_{\text{gen}} = 0$ in the linear setting, we only state this result for the XOR problem.

Theorem

Suppose $\kappa < \kappa_{\text{gen}}^{\text{XOR},h}$. For any $\epsilon > 0$, there exists a constant $c = c(\kappa, \epsilon)$ such that if $d \geqslant cn$, then for any $\mathcal{D} \in \Omega_{\sigma,d}^{\text{XOR}}$, with probability $1 - 3e^{-\frac{n}{c}}$ over $S \sim \mathcal{D}^n$, there exists some W with $\|W\| = 1$ and $\gamma(f_W, S) \geqslant (1-\epsilon)\gamma^*(S)$ such that $\mathcal{L}_{\mathcal{D}}(f_W) = \frac{1}{2}$.

The last two theorems demonstrate that in the XOR problem, there is a threshold in κ (κ_{gen}) above which generalization occurs. As long as κ is above this threshold, we achieve generalization when $\sigma^2 \ll 1$.

One-sided UC Bounds are Vacuous (XOR)

Theorem

Fix $h \in (1, 2)$, and suppose $\kappa_{\textit{gen}}^{\textit{XOR},h} < \kappa < \kappa_{\textit{uc}}^{\textit{XOR},h}$. For any $\delta > 0$ there exist strictly positive constants $\epsilon = \epsilon(\kappa, \delta)$ and $c = c(\kappa, \delta)$ such that the following holds. Let $\mathcal{A}$ be any algorithm that outputs a $(1 - \epsilon)$-max-margin two-layer neural network f_W for any $S \in (\mathbb{R}^d \times \{-1, 1\})^n$. Let $\mathcal{H}$ be any concept class that is **useful** *for $\mathcal{A}$ on $\Omega_{\sigma,d}^{h,\textit{XOR}}$. Suppose that $\epsilon_{\textit{unif}}$ is a uniform convergence bound for the XOR problem $\Omega_{\sigma,d}^{h,\textit{XOR}}$: that is, for any $\mathcal{D} \in \Omega_{\sigma,d}^{h,\textit{XOR}}$, $\epsilon_{\textit{unif}}$ satisfies*

$$Pr_{S \sim \mathcal{D}^n} \left(\sup_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h) \geqslant \epsilon_{\textit{unif}} \right) \leqslant \frac{1}{4}$$

Then if $d \geqslant cn$ and $n > c$ we must have $\epsilon_{\textit{unif}} \geqslant 1 - \delta$

One-sided UC Bounds are Vacuous (Linear)

Theorem

Consider the problem class $\Omega_{\sigma,d}^{\textit{linear}}$ *and suppose that* $\kappa_{\textit{gen}}^{\textit{linear}} < \kappa < \kappa_{\textit{uc}}^{\textit{linear}}$*. Then the following result holds for a universal constant (independent of* κ*): If* $\epsilon \leqslant \frac{\kappa(\kappa_{\textit{uc}}^{\textit{linear}} - \kappa)^2}{c}$*, and* $\frac{d}{n} \geqslant \frac{c}{\kappa^2(\kappa_{\textit{uc}}^{\textit{linear}} - \kappa)^4}$*, then we achieve the guarantee that* $\epsilon_{\textit{unif}} \geqslant 1 - e^{-\frac{n}{36 d \sigma^2}} - e^{-\frac{n}{8}}$

Polynomial Margin Bounds Fail for Linear Problem

Theorem

Suppose $\kappa_{\text{gen}}^{\text{linear}} < \kappa < \kappa_{\text{uc}}^{\text{linear}}$. There exists a universal constant c such that the following holds. Let $\epsilon = \frac{\kappa(\kappa_{uc}-\kappa)^2}{c}$, and let $\mathcal{A}$ be any algorithm so that $\mathcal{A}(S)$ outputs a $(1-\epsilon)$-max-margin solution f_w for any $S \in (\mathbb{R}^d \times \{-1,1\})^n$. Let $\mathcal{H}$ be any concept class that is **useful** *for $\mathcal{A}$. Suppose that there exists a polynomial margin bound of integer degree p for the linear problem $\Omega_{\sigma,d}^{\text{linear}}$ that is, for any $\mathcal{D} \in \Omega_{\sigma,d}^{\text{linear}}$, there is some G that satisfies*

$$Pr_{S\sim\mathcal{D}^n}\left(\sup_{h\in\mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h) \geqslant \frac{G}{\gamma(h,S)^p}\right) \leqslant \frac{1}{4}$$

Then for any $\mathcal{D} \in \Omega_{\sigma,d}^{\text{linear}}$, if $\frac{d}{n} \geqslant \frac{c}{\kappa^2(\kappa_{uc}-\kappa)^4}$, with probability $\frac{1}{2} - 3e^{-n}$ over $S \sim \mathcal{D}^n$, the margin bound is weak even on the max-margin solution, that is, $\frac{G}{\gamma^*(S)^p} \geqslant \max(\frac{1}{c}, 1 - e^{-\frac{\kappa}{36\sigma^2}} - e^{-\frac{n}{8}} - \frac{3\kappa}{c})^p$, which is more than an absolute constant.

Polynomial Margin Bounds Fail for XOR on Neural Network

Theorem

Fix an integer $p \geqslant 1$, and suppose $\kappa_{\textit{gen}}^{\textit{XOR},h} < \kappa < \kappa_{\textit{uc}}^{\textit{XOR},h}$. For any $\epsilon > 0$, there exists $c = c(\kappa, p, \epsilon)$ such that the following holds. Let $\mathcal{H}$ be any hypothesis class such that for all $\mathcal{D} \in \Omega_{\sigma,d}^{\textit{XOR}}$,

$$\textit{Pr}_{S\sim\mathcal{D}^n}\left(\textit{all } (1-\epsilon)\textit{max-margin two-layer neural networks } f_W \textit{ for } S \textit{ lie in } \mathcal{H}\right) \geqslant \frac{3}{4}$$

Suppose that there exists an polynomial margin bound of degree p for the XOR problem $\Omega_{\sigma,d}^{\textit{XOR}}$: that is, for any $\mathcal{D} \in \Omega_{\sigma,d}^{\textit{XOR}}$, there exists some G that satisfies

$$\textit{Pr}_{S\sim\mathcal{D}^n}\left(\sup_{h\in\mathcal{H}} \mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h) \geqslant \frac{G}{\gamma(h,S)^p}\right) \leqslant \frac{1}{4}$$

Then for any $\mathcal{D} \in \Omega_{\sigma,d}^{\textit{XOR}}$if $d \geqslant cn$ and $n \geqslant c$, with probability $\frac{1}{2} - 3e^{-\frac{n}{c}}$ over $S \sim \mathcal{D}^n$, on the max-margin solution, the generalization guarantee is no better than $\frac{1}{c}$, ie. $\frac{G}{\gamma^*(S)^p} \geqslant \frac{1}{c}$

The End