组会分享

蒋佳锐

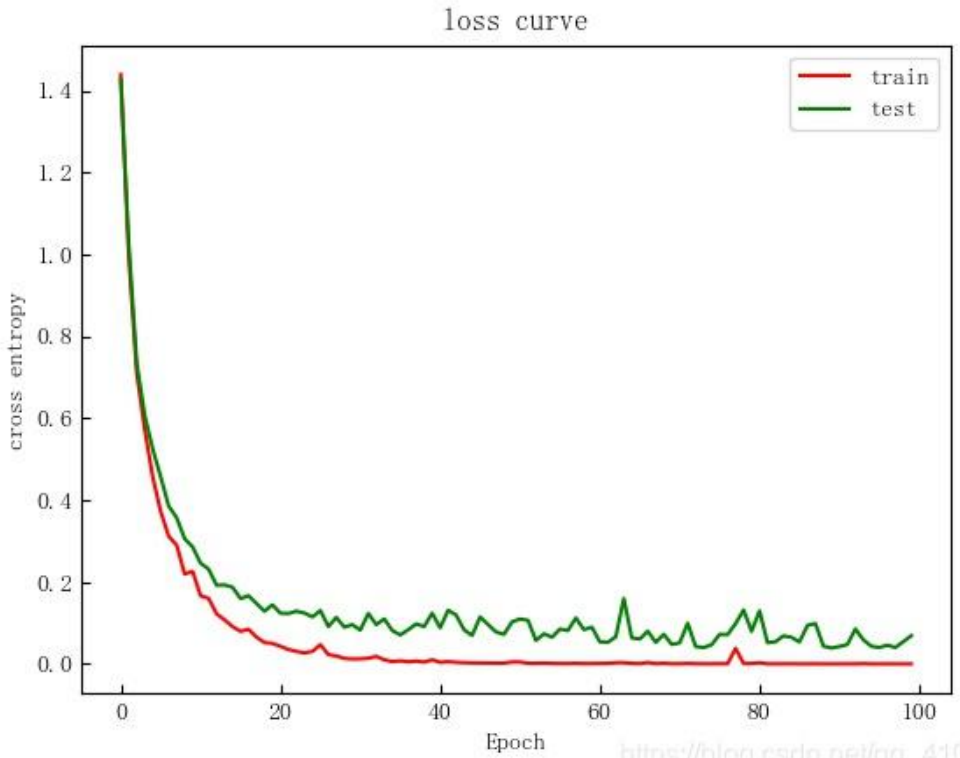
Benign Overfitting in Two-layer Convolutional Neural Networks

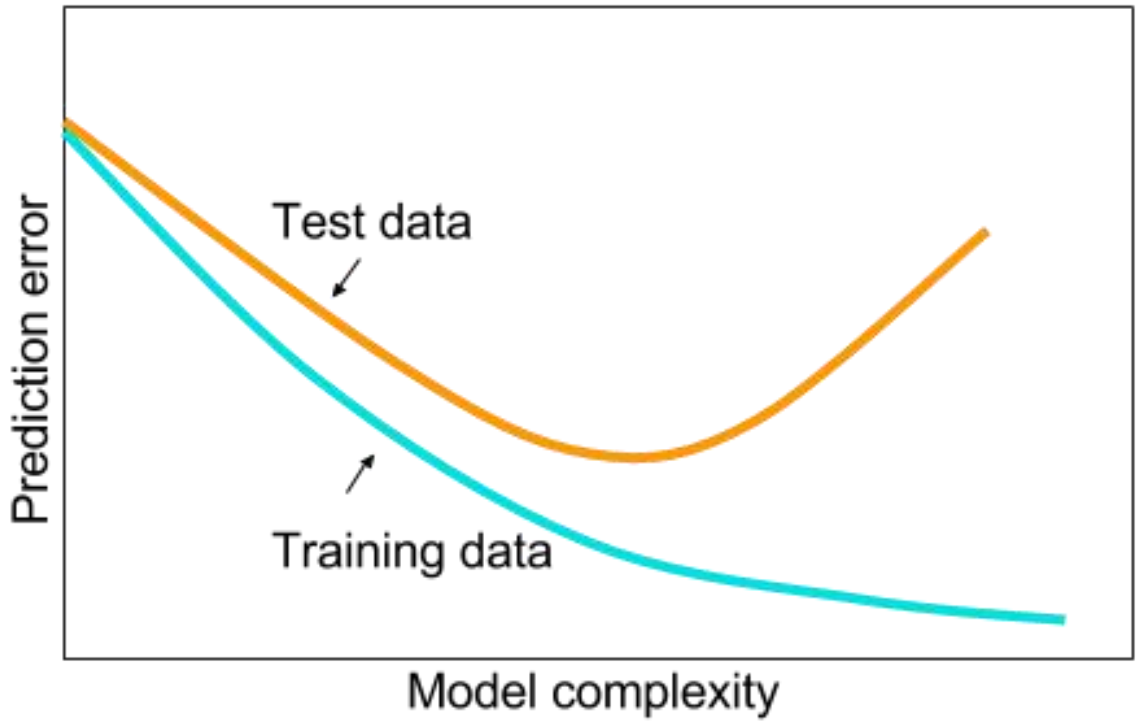
Yuan Cao*† and Zixiang Chen*‡ and Mikhail Belkin§ and Quanquan Gu¶

Abstract

Modern neural networks often have great expressive power and can be trained to overfit the training data, while still achieving a good test performance. This phenomenon is referred to as "benign overfitting". Recently, there emerges a line of works studying "benign overfitting" from the theoretical perspective. However, they are limited to linear models or kernel/random feature models, and there is still a lack of theoretical understanding about when and how benign overfitting occurs in neural networks. In this paper, we study the benign overfitting phenomenon in training a two-layer convolutional neural network (CNN). We show that when the signal-to-noise ratio satisfies a certain condition, a two-layer CNN trained by gradient descent can achieve arbitrarily small training and test loss. On the other hand, when this condition does not hold, overfitting becomes harmful and the obtained CNN can only achieve constant level test loss. These together demonstrate a sharp phase transition between benign overfitting and harmful overfitting, driven by the signal-to-noise ratio. To the best of our knowledge, this is the first work that precisely characterizes the conditions under which benign overfitting can occur in training convolutional neural networks.

benign overfitting VS harmful overfitting





How to avoid harmful overfitting

- larger sample size **n**
- early stopping
- data augmentation
- weight decay
- greater **signal-noise ratio**

Problem Setup

Definition 3.1. *Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be a fixed vector representing the signal contained in each data point. Then each data point $(\mathbf{x}, y)$ with $\mathbf{x} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top]^\top \in \mathbb{R}^{2d}$ and $y \in \{-1, 1\}$ is generated from the following distribution $\mathcal{D}$:*

1. *The label y is generated as a Rademacher random variable.*
2. *A noise vector $\boldsymbol{\xi}$ is generated from the Gaussian distribution $N(\mathbf{0}, \sigma_p^2 \cdot (\mathbf{I} - \boldsymbol{\mu}\boldsymbol{\mu}^\top \cdot \|\boldsymbol{\mu}\|_2^{-2}))$.*
3. *One of $\mathbf{x}_1, \mathbf{x}_2$ is given as $y \cdot \boldsymbol{\mu}$, which represents the signal, the other is given by $\boldsymbol{\xi}$, which represents noises.*

$$\mathrm{SNR} = \frac{\|\mu\|_2}{\sigma_p \sqrt{d}} \approx \frac{\|\mu\|_2}{\|\xi\|_2}$$

Two-layer CNNs

Two-layer CNNs. We consider a two-layer convolutional neural network whose filters are applied to the two patches $\mathbf{x}_1$ and $\mathbf{x}_2$ separately, and the second layer parameters of the network are fixed as $+1/m$ and $-1/m$ respectively. Then the network can be written as $f(\mathbf{W}, \mathbf{x}) = F_{+1}(\mathbf{W}_{+1}, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$, where $F_{+1}(\mathbf{W}_{+1}, \mathbf{x})$, $F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$ are defined as:

$$F_j(\mathbf{W}_j, \mathbf{x}) = \frac{1}{m}\sum_{r=1}^{m}\big[\sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_1\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_2\rangle)\big] = \frac{1}{m}\sum_{r=1}^{m}\big[\sigma(\langle \mathbf{w}_{j,r}, y\cdot\boldsymbol{\mu}\rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}\rangle)\big],$$

$\sigma(z) = (\max\{0, z\})^q$

$$L_S(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n}\ell[y_i \cdot f(\mathbf{W}, \mathbf{x}_i)],$$

where $\ell(z) = \log(1 + \exp(-z))$, and $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is the training data set. We further define the true loss (test loss) $L_{\mathcal{D}}(\mathbf{W}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\ell[y \cdot f(\mathbf{W}, \mathbf{x})]$.

gradient descent update

$$\begin{aligned}\mathbf{w}_{j,r}^{(t+1)} &= \mathbf{w}_{j,r}^{(t)} - \eta \cdot \nabla_{\mathbf{w}_{j,r}} L_S(\boldsymbol{W}^{(t)}) \\ &= \mathbf{w}_{j,r}^{(t)} - \frac{\eta}{nm} \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot j y_i \boldsymbol{\xi}_i - \frac{\eta}{nm} \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \boldsymbol{\mu} \rangle) \cdot j \boldsymbol{\mu}\end{aligned}$$

for $j \in \{\pm 1\}$ and $r \in [m]$, where we introduce a shorthand notation $\ell_i'^{(t)} = \ell'[y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)]$.

Main Result

- Definition 4.1

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i.$$

signal learning

noise memorization

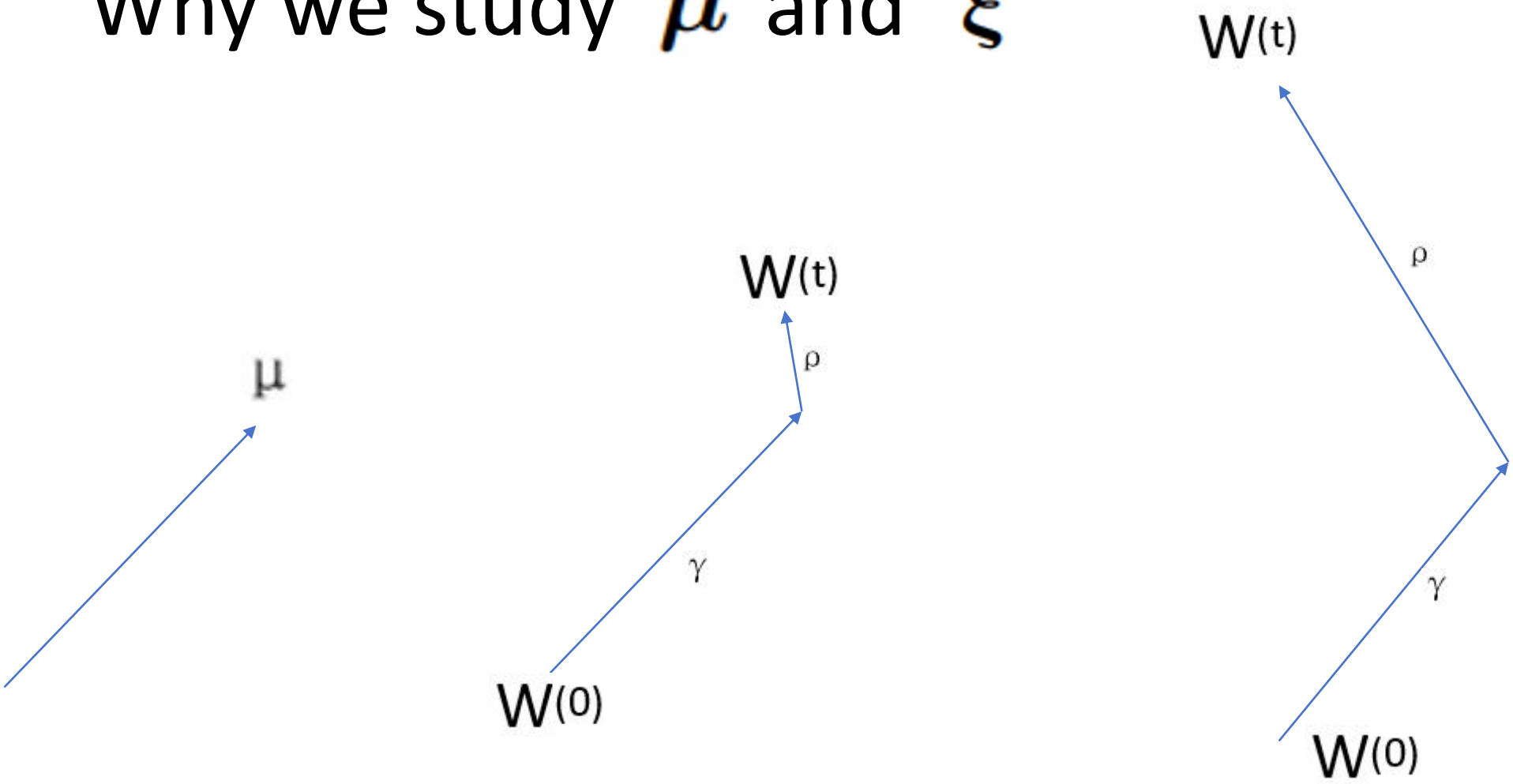
We further denote $\overline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \geq 0)$, $\underline{\rho}_{j,r,i}^{(t)} := \rho_{j,r,i}^{(t)} \mathbb{1}(\rho_{j,r,i}^{(t)} \leq 0)$. *Then we have that*

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i.$$

signal learning

noise memorization

Why we study μ and ξ



Main Result

Condition 4.2. *Suppose that*

1. *Dimension d is sufficiently large:* $d = \widetilde{\Omega}(m^{2\vee[4/(q-2)]}n^{4\vee[(2q-2)/(q-2)]})$.
2. *Training sample size n and neural network width m satisfy* $n, m = \Omega(\mathrm{polylog}(d))$.
3. *The learning rate η satisfies* $\eta \leq \widetilde{O}(\min\{\|\boldsymbol{\mu}\|_2^{-2}, \sigma_p^{-2}d^{-1}\})$.
4. *The standard deviation of Gaussian initialization σ_0 is appropriately chosen such that* $\widetilde{O}(nd^{-1/2})\cdot \min\{(\sigma_p\sqrt{d})^{-1}, \|\boldsymbol{\mu}\|_2^{-1}\} \leq \sigma_0 \leq \widetilde{O}(m^{-2/(q-2)}n^{-[1/(q-2)]\vee 1}) \cdot \min\{(\sigma_p\sqrt{d})^{-1}, \|\boldsymbol{\mu}\|_2^{-1}\}$.

Main Result

Theorem 4.3. *For any $\epsilon > 0$, let $T = \widetilde{\Theta}(\eta^{-1} m \sigma_0^{-(q-2)} \|\boldsymbol{\mu}\|_2^{-q} + \eta^{-1}\epsilon^{-1} m^3 \|\boldsymbol{\mu}\|_2^{-2})$. Under Condition 4.2, if $n \cdot \mathrm{SNR}^q = \widetilde{\Omega}(1)$[1], then with probability at least $1 - d^{-1}$, there exists $0 \le t \le T$ such that:*

1. *The CNN learns the signal:* $\max_r \gamma_{j,r}^{(t)} = \Omega(1)$ *for* $j \in \{\pm 1\}$.
2. *The CNN does not memorize the noises in the training data:* $\max_{j,r,i} |\rho_{j,r,i}^{(T)}| = \widetilde{O}(\sigma_0 \sigma_p \sqrt{d})$.
3. *The training loss converges to ϵ, i.e.,* $L_S(\mathbf{W}^{(t)}) \le \epsilon$.
4. *The trained CNN achieves a small test loss:* $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \le 6\epsilon + \exp(-n^2)$

benign overfitting

Main Result

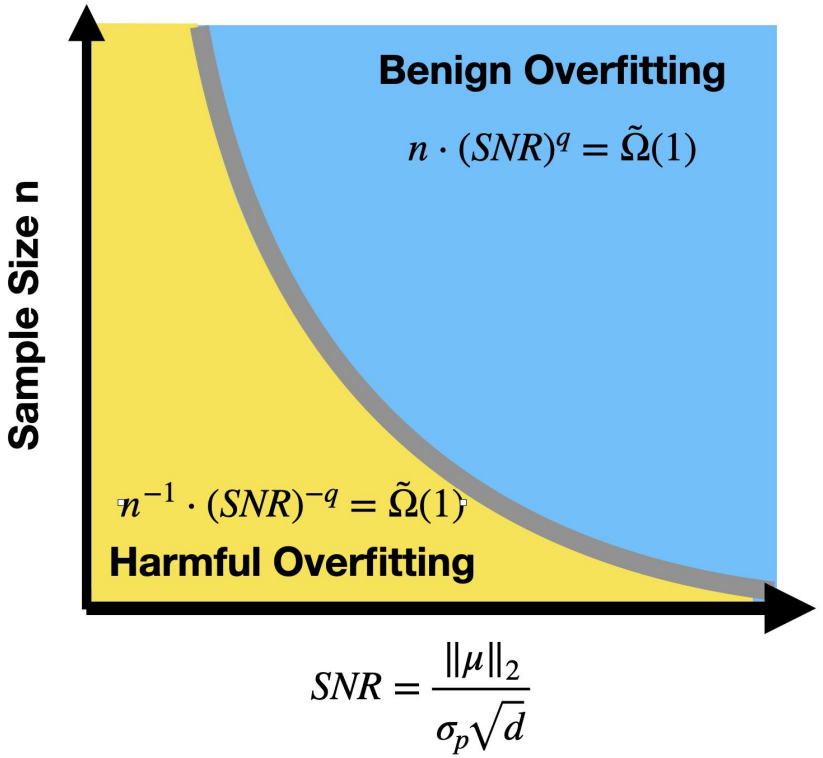
Theorem 4.4. *For any $\epsilon > 0$, let $T = \widetilde{\Theta}(\eta^{-1} m \cdot n(\sigma_p \sqrt{d})^{-q} \cdot \sigma_0^{-(q-2)} + \eta^{-1}\epsilon^{-1} n m^3 d^{-1} \sigma_p^{-2})$. Under Condition 4.2, if $n^{-1} \cdot \mathrm{SNR}^{-q} = \widetilde{\Omega}(1)$, then with probability at least $1 - d^{-1}$, there exists $0 \le t \le T$ such that:*

1. *The CNN memorizes noises in the training data: $\max_r \overline{\rho}_{y_i,r,i}^{(t)} = \Omega(1)$.*
2. *The CNN does not sufficiently learn the signal: $\max_{j,r} \gamma_{j,r}^{(t)} \le \widetilde{O}(\sigma_0 \|\boldsymbol{\mu}\|_2)$.*
3. *The training loss converges to ϵ, i.e., $L_S(\mathbf{W}^{(t)}) \le \epsilon$.*
4. *The trained CNN has a constant order test loss: $L_{\mathcal{D}}(\mathbf{W}^{(t)}) = \Theta(1)$.*

harmful overfitting

Main Result

- If $n \cdot \mathrm{SNR}^q = \widetilde{\Omega}(1)$, then the CNN learns the signal and achieves a $O(\epsilon + \exp(-n^2))$ test loss. This is the regime of benign overfitting.
- If $n^{-1} \cdot \mathrm{SNR}^{-q} = \widetilde{\Omega}(1)$ then the CNN can only memorize noises and will have a $\Theta(1)$ test loss. This is the regime of harmful overfitting.



Proof Technique

- Iterative Analysis of the Signal-Noise Decomposition

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j \cdot \gamma_{j,r}^{(t)} \cdot \|\boldsymbol{\mu}\|_2^{-2} \cdot \boldsymbol{\mu} + \sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \cdot \|\boldsymbol{\xi}_i\|_2^{-2} \cdot \boldsymbol{\xi}_i.$$

$$\gamma_{j,r}^{(0)}, \overline{\rho}_{j,r,i}^{(0)}, \underline{\rho}_{j,r,i}^{(0)} = 0,$$

$$\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \cdot \boldsymbol{\mu} \rangle) \cdot \|\boldsymbol{\mu}\|_2^2,$$

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_i = j),$$

$$\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \cdot \ell_i'^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \|\boldsymbol{\xi}_i\|_2^2 \cdot \mathbb{1}(y_i = -j).$$

Iterative Analysis of the Signal-Noise Decomposition

$$\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'(y_i \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle + y_i \cdot j \cdot \gamma_{j,r}^{(t)}) \cdot \|\boldsymbol{\mu}\|_2^2,$$

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \cdot \ell_i'^{(t)} \sigma' \left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \sum_{i'=1}^{n} \overline{\rho}_{j,r,i'}^{(t)} \frac{\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2} + \sum_{i'=1}^{n} \underline{\rho}_{j,r,i'}^{(t)} \frac{\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2} \right) \cdot \|\boldsymbol{\xi}_i\|_2^2 \quad (y_i = j)$$

$$\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm} \cdot \ell_i'^{(t)} \sigma' \left(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \sum_{i'=1}^{n} \overline{\rho}_{j,r,i'}^{(t)} \frac{\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2} + \sum_{i'=1}^{n} \underline{\rho}_{j,r,i'}^{(t)} \frac{\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2} \right) \cdot \|\boldsymbol{\xi}_i\|_2^2 \quad (y_i = -j)$$

Bound γ

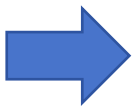
$$\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \cdot \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'(y_i \cdot \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle + y_i \cdot j \cdot \gamma_{j,r}^{(t)}) \cdot \|\boldsymbol{\mu}\|_2^2,$$

bound $l_i^{'(t)}$ $\quad \ell_i'^{(t)} = \ell'[y_i \cdot f(\mathbf{W}^{(t)}, \mathbf{x}_i)]$

$$\left|\langle w_{j,r}^{(0)}, \mu \rangle\right|, \left|\langle w_{j,r}^{(0)}, \xi_i \rangle\right|, \|\xi\|_2^2, \left|\langle \xi_i, \xi_{i'} \rangle\right|, \left|\langle w_{j,r}^{(t)}, \mu \rangle\right|, \left|\langle w_{j,r}^{(t)}, \xi_i \rangle\right|$$

$$F_j\left(W_j^{(t)}, x_i\right) = \frac{1}{m} \sum_{r=1}^{m} \left[\sigma\left(\langle w_{j,r}^{(t)}, y_i \mu \rangle\right) + \sigma\left(\langle w_{j,r}^{(t)}, \xi_i \rangle\right)\right] = O(1)$$

$$-l_i^{'(t)} \geq C_1$$



$$\begin{aligned}\gamma_{1,r}^{(t+1)} &= \gamma_{1,r}^{(t)} - \frac{\eta}{nm} \cdot \sum_{i=1}^{n} \ell_i'^{(t)} \cdot \sigma'(y_i \cdot \langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\mu} \rangle + y_i \cdot \gamma_{1,r}^{(t)}) \cdot \|\boldsymbol{\mu}\|_2^2 \\ &\geq \gamma_{1,r}^{(t)} + \frac{C_1 \eta}{nm} \cdot \sum_{y_i=1} \sigma'(\langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\mu} \rangle + \gamma_{1,r}^{(t)}) \cdot \|\boldsymbol{\mu}\|_2^2.\end{aligned}$$

Bound γ

$$\gamma_{1,r}^{(t+1)} \geq \gamma_{1,r}^{(t)} + \frac{C_1\eta}{nm} \cdot \sum_{y_i=1} \sigma'(\langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\mu} \rangle + \gamma_{1,r}^{(t)}) \cdot \|\boldsymbol{\mu}\|_2^2.$$

- bound $A^{(t)} = max_r \left(\langle w_{1,r}^{(0)}, \mu \rangle + \gamma_{1,r}^{(t)} \right)$

$$A^{(t+1)} \geq \left(1 + \frac{C_1\eta q \sigma_0^{q-2} \|\boldsymbol{\mu}\|_2^q}{2^q m}\right) A^{(t)}$$

$$A^{(t)} \geq \left(1 + \frac{C_1\eta q \sigma_0^{q-2} \|\boldsymbol{\mu}\|_2^q}{2^q m}\right)^t A^{(0)} \geq \exp\left(\frac{C_1\eta q \sigma_0^{q-2} \|\boldsymbol{\mu}\|_2^q}{2^{q+1} m} t\right) A^{(0)} \geq \exp\left(\frac{C_1\eta q \sigma_0^{q-2} \|\boldsymbol{\mu}\|_2^q}{2^{q+1} m} t\right) \frac{\sigma_0 \|\boldsymbol{\mu}\|_2}{2},$$

if t is large enough, then:

$$\max_r \gamma_{j,r}^{(t)} = \Omega(1)$$

Bound ρ

$$\Psi^{(t)} = \max_{j,r,i} |\rho_{j,r,i}^{(t)}| = \max_{j,r,i} \{\overline{\rho}_{j,r,i}^{(t)}, -\underline{\rho}_{j,r,i}^{(t)}\}$$

$$\psi^{(t+1)} = \psi^{(t)} + \Delta\psi \qquad \text{upper bound } \Delta\psi$$

$$\psi^{(t)} \leq t * \Delta\psi$$

$$\max_{j,r,i} |\rho_{j,r,i}^{(T)}| = \widetilde{O}(\sigma_0 \sigma_p \sqrt{d})$$

Proof Technique

- Decoupling with a Two-Stage Analysis

  Stage 1: $\left|l_i^{'(t)}\right| = \Theta(1)$-before loss function significantly decreases

  Stage 2: Training loss converges to ε

Bound $L_s\left(W^{(t)}\right)$

$$\widehat{\mathcal{R}}(W) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i f_i(W)\right)$$

Ziwei Ji Matus Telgarsky

$$\left\|W_{t+1} - \overline{W}\right\|_F^2 = \left\|W_t - \overline{W}\right\|_F^2 - 2\eta_t \left\langle \nabla\widehat{\mathcal{R}}(W_t), W_t - \overline{W}\right\rangle + \eta_t^2 \left\|\nabla\widehat{\mathcal{R}}(W_t)\right\|_F^2$$

$$\begin{aligned}
\left\langle \nabla\widehat{\mathcal{R}}(W_t), W_t - \overline{W}\right\rangle &= \frac{1}{n}\sum_{i=1}^{n} \ell'\left(y_i f_i(W_t)\right) y_i \left\langle \nabla f_i(W_t), W_t - \overline{W}\right\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n} \ell'\left(y_i f_i(W_t)\right)\left(y_i f_i(W_t) - y_i f_i^{(t)}\left(\overline{W}\right)\right) \\
&\geq \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(y_i f_i(W_t)\right) - \ell\left(y_i f_i^{(t)}\left(\overline{W}\right)\right)\right) = \widehat{\mathcal{R}}(W_t) - \widehat{\mathcal{R}}^{(t)}\left(\overline{W}\right)
\end{aligned}$$

Homogeneity

- ReLU - 1 homogeneous

$$\langle \nabla f_i(W), W \rangle = f_i(W).$$

- ReLU^q - q homogeneous

$$\langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}), \mathbf{W}^{(t)} \rangle = q \cdot f(\mathbf{W}^{(t)}, \mathbf{x})$$

Bound $L_s(W^{(t)})$

$$y_i\langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^* \rangle \geq 2q\log(2q/\epsilon) - \widetilde{O}(\sigma_0\|\boldsymbol{\mu}\|_2) - \widetilde{O}(\sigma_0\sigma_p\sqrt{d}) \geq q\log(2q/\epsilon)$$

$$\begin{aligned}
&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\
&= 2\eta\langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\
&= \frac{2\eta}{n}\sum_{i=1}^{n} \ell_i'^{(t)}[qy_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) - \langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^* \rangle] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\
&\geq \frac{2\eta}{n}\sum_{i=1}^{n} \ell_i'^{(t)}[qy_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) - q\log(2q/\epsilon)] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\
&\geq \frac{2q\eta}{n}\sum_{i=1}^{n}[\ell\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big) - \epsilon/(2q)] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 \\
&\geq (2q-1)\eta L_S(\mathbf{W}^{(t)}) - \eta\epsilon,
\end{aligned}$$

$$\frac{1}{T - T_1 + 1}\sum_{s=T_1}^{T} L_S(\mathbf{W}^{(s)}) \leq \frac{\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2}{(2q-1)\eta(T - T_1 + 1)} + \frac{\epsilon}{2q-1} \leq \frac{3\epsilon}{2q-1} < \epsilon,$$

Bound $L_D(W^{(t)})$

event $\mathcal{E}$

Lemma D.7. *Under the same conditions as Theorem 4.3, with probability at least* $1 - 4mT \cdot \exp(-C_2^{-1}\sigma_0^{-2}\sigma_p^{-2}d^{-1})$, *we have that* $\max_{j,r} |\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi} \rangle| \leq 1/2$ *for all* $0 \leq t \leq T$, *where* $C_2 = \widetilde{O}(1)$.

$$\mathbb{E}\big[\ell\big(yf(\mathbf{W}^{(t)}, \mathbf{x})\big)\big] = \underbrace{\mathbb{E}[\mathbb{1}(\mathcal{E})\ell\big(yf(\mathbf{W}^{(t)}, \mathbf{x})\big)]}_{I_1} + \underbrace{\mathbb{E}[\mathbb{1}(\mathcal{E}^c)\ell\big(yf(\mathbf{W}^{(t)}, \mathbf{x})\big)]}_{I_2}.$$

Bound I1

$$\exp(-y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)) \overset{(i)}{\leq} 2\log\left(1+\exp(-y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i))\right) = 2\ell\big(y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big) \leq 2L_S(\mathbf{W}^{(t)})$$

$$\begin{aligned}
|yf(\mathbf{W}^{(t)}, \mathbf{x}) - y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)| &\leq \frac{1}{m}\sum_{j,r}\sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i\rangle) + \frac{1}{m}\sum_{j,r}\sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}\rangle) \\
&\leq \frac{1}{m}\sum_{j,r}\sigma(1/2) + \frac{1}{m}\sum_{j,r}\sigma(1/2) \\
&\leq 1,
\end{aligned}$$

$$\begin{aligned}
I_1 &\leq \mathbb{E}[\mathbb{1}(\mathcal{E})\exp(-yf(\mathbf{W}^{(t)}, \mathbf{x}))] \\
&\leq e\cdot\mathbb{E}[\mathbb{1}(\mathcal{E})\exp(-y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i))] \\
&\leq 2e\cdot\mathbb{E}[\mathbb{1}(\mathcal{E})L_S(\mathbf{W}^{(t)})],
\end{aligned}$$

Bound I2

$$
\begin{aligned}
\ell\big(yf(\mathbf{W}^{(t)},\mathbf{x})\big) &\le \log(1+\exp(F_{-y}(\mathbf{W}^{(t)},\mathbf{x})))\\
&\le 1+F_{-y}(\mathbf{W}^{(t)},\mathbf{x})\\
&= 1+\frac{1}{m}\sum_{j=-y,r\in[m]}\sigma(\langle\mathbf{w}_{j,r}^{(t)},y\boldsymbol{\mu}\rangle)+\frac{1}{m}\sum_{j=-y,r\in[m]}\sigma(\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}\rangle)\\
&\le 1+F_{-y_i}(\mathbf{W}_{-y_{i'}},\mathbf{x}_{i'})+\frac{1}{m}\sum_{j=-y,r\in[m]}\sigma(\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}\rangle)\\
&\le 2+\frac{1}{m}\sum_{j=-y,r\in[m]}\sigma(\langle\mathbf{w}_{j,r}^{(t)},\boldsymbol{\xi}\rangle)\\
&\le 2+\widetilde{O}((\sigma_0\sqrt{d})^q)\|\boldsymbol{\xi}\|^q,
\end{aligned}
$$

$$
\begin{aligned}
I_2 &\le \sqrt{\mathbb{E}[\mathbb{1}(\mathcal{E}^c)]}\cdot\sqrt{\mathbb{E}\Big[\ell\big(yf(\mathbf{W}^{(t)},\mathbf{x})\big)^2\Big]}\\
&\le \sqrt{\mathbb{P}(\mathcal{E}^c)}\cdot\sqrt{4+\widetilde{O}((\sigma_0\sqrt{d})^{2q})\mathbb{E}[\|\boldsymbol{\xi}\|_2^{2q}]}\\
&\le \exp[-\widetilde{\Omega}(\sigma_0^{-2}\sigma_p^{-2}d^{-1})+\mathrm{polylog}(d)]\\
&\le \exp(-n^2),
\end{aligned}
$$

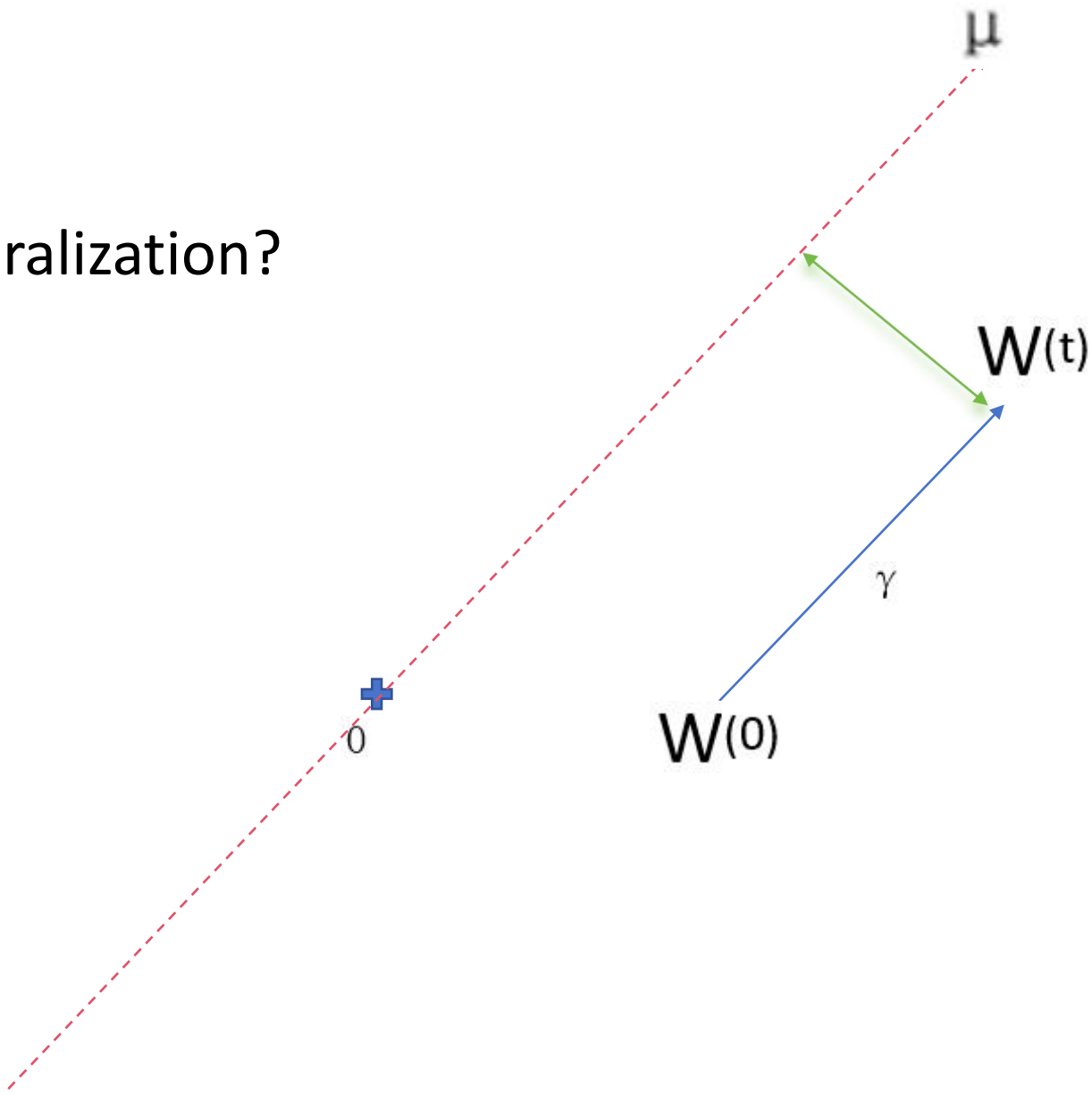
$$
L_{\mathcal{D}}(\mathbf{W}^{(t)}) \le 6\epsilon+\exp(-n^2)
$$

Inequalities and Bounds

- Markov’s inequality
- Hoeffding’s inequality
- Bernstein’s inequality
- Jensen’s inequality
- Triangle inequality
- Cauchy-Schwartz inequality
- Gaussian tail bound
- Union bound

discussion

- Can noise help generalization?
- How to initialize W



Thanks!