Can we avoid robust overfitting in adversarial training? - An approximation viewpoint

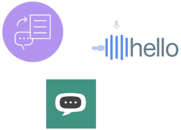
Fanghui Liu

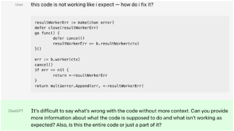
Department of Computer Science, University of Warwick, UK
Centre for Discrete Mathematics and its Applications (DIMAP), Warwick

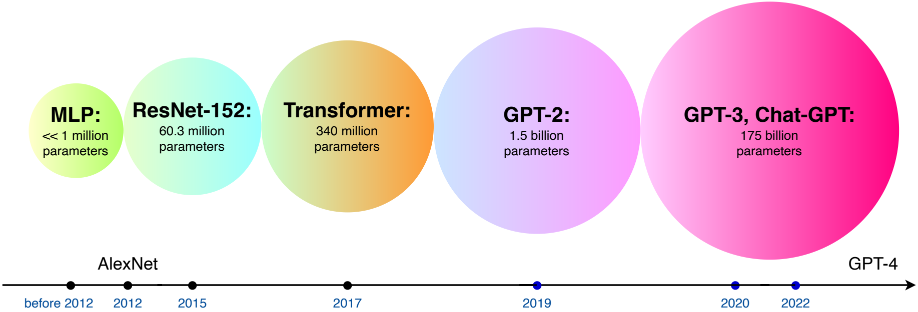
Based on joint work with
[Zhongjie Shi (HKU), Fanghui Liu, Yuan Cao (HKU), Johan A.K. Suykens (KU Leuven)]

at MLOPT Research Group Idea Seminar, UW-Madison

Over-parameterization: more parameters than training data







DNNs: the good in fitting...

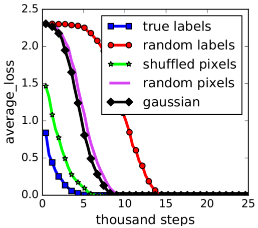


Figure: DNN Training curves on CIFAR10, from [1]

- Benign overfitting [2]
  - model complex enough to fit random labels
  - zero training error and low test error

DNNs: the good in fitting...

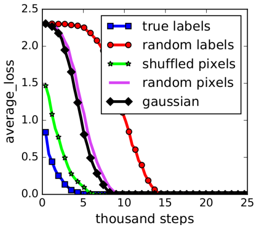


Figure: DNN Training curves on CIFAR10, from [1]

- Benign overfitting [2]
  - model complex enough to fit random labels
  - zero training error and low test error
- Empirical risk minimization

$$\hat{f} := \operatorname*{arg\,min}_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\Big(f_{\boldsymbol{w}}(\boldsymbol{x}_i), y_i\Big) \right\}$$

DNNs: the good in fitting...

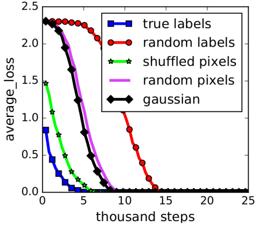
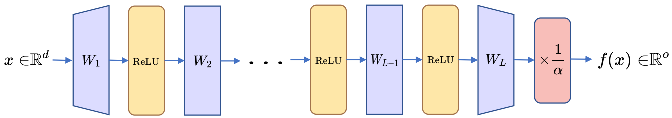


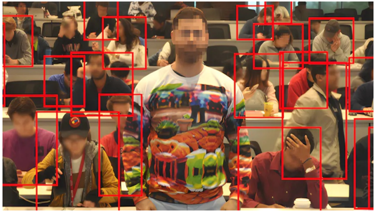
Figure: DNN Training curves on CIFAR10, from [1]

- Benign overfitting [2]
  - model complex enough to fit random labels
  - zero training error and low test error
- Empirical risk minimization

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\Big(f_{\boldsymbol{w}}(\boldsymbol{x}_i), y_i\Big) \right\}$$



DNNs: the bad in robustness...



(a) Invisibility [3]

(b) Stop sign classified as 45 mph sign [4]

Adversarial training [6, 7, 8]

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\max_{\boldsymbol{x}_i' \in B_{\delta,\infty}(\boldsymbol{x}_i)} \ell\left(f_{\boldsymbol{w}}(\boldsymbol{x}_i'), y_i\right) \right] \right\}$$

with the perturbation ball $B_{\delta,\infty}(\boldsymbol{x}) = \{\boldsymbol{x}' : \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \delta\}$

Adversarial training [6, 7, 8]

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\max_{\boldsymbol{x}'_i \in B_{\delta,\infty}(\boldsymbol{x}_i)} \ell\left(f_{\boldsymbol{w}}(\boldsymbol{x}'_i), y_i\right) \right] \right\}$$

with the perturbation ball $B_{\delta,\infty}(\boldsymbol{x}) = \{\boldsymbol{x}' : \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \delta\}$

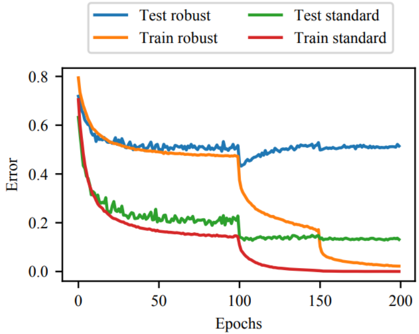


Figure: Results on CIFAR-10 with $\delta = 8/255$ [5].

Adversarial training [6, 7, 8]

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\max_{\boldsymbol{x}_i' \in B_{\delta,\infty}(\boldsymbol{x}_i)} \ell\left(f_{\boldsymbol{w}}(\boldsymbol{x}_i'), y_i \right) \right] \right\}$$

with the perturbation ball $B_{\delta,\infty}(\boldsymbol{x}) = \{\boldsymbol{x}' : \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \delta\}$

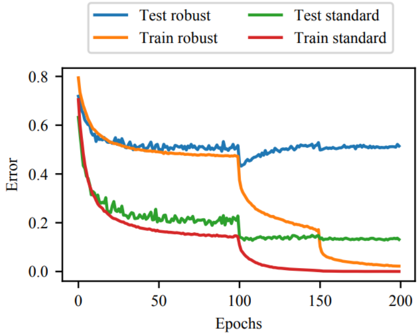


Figure: Results on CIFAR-10 with $\delta = 8/255$ [5].

Observations:

- **robust overfitting**: overfitting on adversarial training data harms the robust generalization

Adversarial training [6, 7, 8]

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\max_{\boldsymbol{x}'_i \in B_{\delta,\infty}(\boldsymbol{x}_i)} \ell\left(f_{\boldsymbol{w}}(\boldsymbol{x}'_i), y_i\right) \right] \right\}$$

with the perturbation ball $B_{\delta,\infty}(\boldsymbol{x}) = \{\boldsymbol{x}' : \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \delta\}$

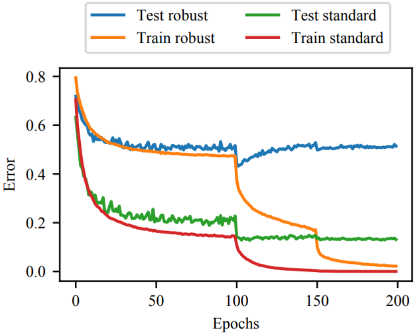


Figure: Results on CIFAR-10 with $\delta = 8/255$ [5].

Observations:

- **robust overfitting**: overfitting on adversarial training data harms the robust generalization
- **robust generalization gap**: gap between standard/robust generalization error
- robust-accuracy trade-off: adversarial training obtains a robust model but clean accuracy drops

Motivation: Can we avoid robust overfitting?

Theorem (Curse of dimensionality [9])

A ReLU DNN requires parameters $m = \Omega(\epsilon^{-d})$ to classify any two ϵ-separated sets A, $B \subseteq [0,1]^d$.

Motivation: Can we avoid robust overfitting?

Theorem (Curse of dimensionality [9])

A ReLU DNN requires parameters $m = \Omega(\epsilon^{-d})$ to classify any two ϵ-separated sets A, $B \subseteq [0,1]^d$.

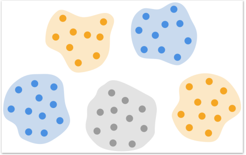


Figure: The class separation in image data. source from [10].

	perturbation ϵ	Train-Train	Test-Train
MNIST	0.1	0.737	0.812
CIFAR-10	0.031	0.212	0.220
SVHN	0.031	0.094	0.110
ResImageNet	0.005	0.180	0.224

Table: Separation of real data under typical perturbation radii. [10]

Motivation: Can we avoid robust overfitting?

Theorem (Curse of dimensionality [9])

A ReLU DNN requires parameters $m = \Omega(\epsilon^{-d})$ to classify any two ϵ-separated sets A, $B \subseteq [0,1]^d$.

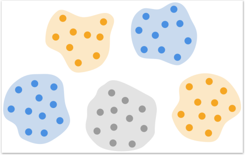
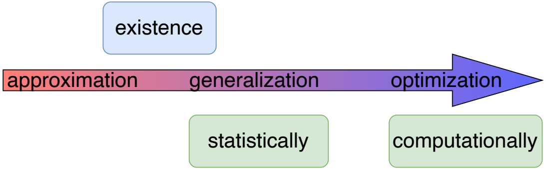


Figure: The class separation in image data. source from [10].

	perturbation ϵ	Train-Train	Test-Train
MNIST	0.1	0.737	0.812
CIFAR-10	0.031	0.212	0.220
SVHN	0.031	0.094	0.110
ResImageNet	0.005	0.180	0.224

Table: Separation of real data under typical perturbation radii. [10]



Preliminary: statistical learning theory (regression)

∘ Empirical risk minimization

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Big(f_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i \Big)^2 \right\}$$

∘ approximate the target function
$f_\rho := \arg\min_{f \in \mathcal{F}} \mathcal{E}(f)$
∘ the expected risk

$$\mathcal{E}(f) := \mathbb{E}_{(\boldsymbol{x},y)\sim\rho}(f_{\boldsymbol{w}}(\boldsymbol{x}) - y)^2$$

- excess risk $\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho)$
- using the squared loss: $\|\hat{f} - f_\rho\|_\rho^2$, where $\|f\|_\rho^2 = \int_X (f(\boldsymbol{x}))^2 \mathrm{d}\rho_X(\boldsymbol{x})$ [11]

Preliminary: statistical learning theory (regression)

◦ Empirical risk minimization

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Big(f_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i \Big)^2 \right\}$$

◦ approximate the target function
$f_\rho := \arg\min_{f \in \mathcal{F}} \mathcal{E}(f)$
◦ the expected risk

$$\mathcal{E}(f) := \mathbb{E}_{(\boldsymbol{x},y)\sim\rho} (f_{\boldsymbol{w}}(\boldsymbol{x}) - y)^2$$

- excess risk $\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho)$
- using the squared loss: $\|\hat{f} - f_\rho\|_\rho^2$, where $\|f\|_\rho^2 = \int_X (f(\boldsymbol{x}))^2 \mathrm{d}\rho_X(\boldsymbol{x})$ [11]

◦ Empirical adversarial risk minimization

$$\widehat{f}^{over} = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \max_{\boldsymbol{x}_i' \in B_{\delta,\infty}(\boldsymbol{x}_i)} \Big(f(\boldsymbol{x}_i') - y_i \Big)^2 \right\}$$

◦ approximate the robust target function
$f_\rho^\delta(\boldsymbol{x}) := \arg\min_{f \in \mathcal{F}} \mathcal{E}^\delta(f)$
◦ the robust expected risk

$$\mathcal{E}^\delta(f) := \mathbb{E}_{(\boldsymbol{x},y)\sim\rho} \max_{\boldsymbol{x}' \in B_{\delta,\infty}(\boldsymbol{x})} \Big(f_{\boldsymbol{w}}(\boldsymbol{x}') - y \Big)^2 .$$

◦ robust excess risk: $\mathcal{E}^\delta(\widehat{f}^{over}) - \mathcal{E}^\delta(f_\rho^\delta)$

Assumptions

Assumption (source condition)

$f_\rho \in W^\alpha_\infty(X)$, *i.e., the* α*-Hölder continuous functions* $W^\alpha_\infty(X)$ *with* $\alpha > 0$.

$$\|f\|_{W^\alpha_\infty} = \|f\|_\infty + \|f\|_{W^\alpha_\infty} \quad \textit{with } \|f\|_{W^\alpha_\infty} = \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^\alpha} .$$

Assumptions

Assumption (source condition)

$f_\rho \in W^\alpha_\infty(X)$, *i.e., the* α*-Hölder continuous functions* $W^\alpha_\infty(X)$ *with* $\alpha > 0$.

$$\|f\|_{W^\alpha_\infty} = \|f\|_\infty + \|f\|_{W^\alpha_\infty} \quad \textit{with } \|f\|_{W^\alpha_\infty} = \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^\alpha} .$$

Assumption (non-irregularity of ρ_X)

$$\Phi_\rho := \{\rho_X : \rho_X \textit{ has bounded support}\}$$

Remark: consistency between $L^1(X)$ and $L^1_{\rho_X}(X)$ by introducing identity mapping J_ρ, $\bar{J}_\rho$

Assumptions

Assumption (source condition)

$f_\rho \in W^\alpha_\infty(X)$, *i.e., the* α*-Hölder continuous functions* $W^\alpha_\infty(X)$ *with* $\alpha > 0$.

$$\|f\|_{W^\alpha_\infty} = \|f\|_\infty + \|f\|_{W^\alpha_\infty} \quad \textit{with } \|f\|_{W^\alpha_\infty} = \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^\alpha} .$$

Assumption (non-irregularity of ρ_X)

$$\Phi_\rho := \{\rho_X : \rho_X \textit{ has bounded support}\}$$

Remark: consistency between $L^1(X)$ and $L^1_{\rho_X}(X)$ by introducing identity mapping J_ρ, $\bar{J}_\rho$

Separation distance

For separated data $X = \{\boldsymbol{x}_i\}_{i=1}^n$ in $[0,1]^d$, we have

$$q_X := \frac{1}{2} \min_{i \neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty \leq n^{-\frac{1}{d}} . \quad [12]$$

Standard genralization error under adversarial training

Theorem (standard generalization (Shi, Liu, Cao, Suykens, 2024))

Assume $f_\rho \in W^{\alpha}_{\infty}(\mathcal{X})$ *with* $\alpha > 0$, $\rho_X \in \Phi_\rho$ *is non-irregular. If* $\delta < \min\left\{\frac{q_X}{3}, n^{-\frac{2\alpha}{(2\alpha+d)d}-\frac{1}{d}}\right\}$, *then* $\exists \widehat{f}^{over}$ *with depth* $L = \mathcal{O}(\log n)$, *and width* $m_1 = \mathcal{O}(nd)$, $m_2, \ldots, m_L = \mathcal{O}(\log n)$, *such that*

$$\sup_{f_\rho \in W^{\alpha}_{\infty}(\mathcal{X}), \rho_X \in \Phi_\rho} \mathbb{E}\left[\mathcal{E}\left(\widehat{f}^{over}\right) - \mathcal{E}(f_\rho)\right] \lesssim \left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+d}}.$$

Standard genralization error under adversarial training

Theorem (standard generalization (Shi, Liu, Cao, Suykens, 2024))

Assume $f_\rho \in W^{\alpha}_{\infty}(\mathcal{X})$ *with* $\alpha > 0$, $\rho_X \in \Phi_\rho$ *is non-irregular. If* $\delta < \min\left\{ \frac{q_X}{3}, n^{-\frac{2\alpha}{(2\alpha+d)d} - \frac{1}{d}} \right\}$, *then* $\exists \widehat{f}^{over}$ *with depth* $L = \mathcal{O}(\log n)$, *and width* $m_1 = \mathcal{O}(nd)$, $m_2, \ldots, m_L = \mathcal{O}(\log n)$, *such that*

$$\sup_{f_\rho \in W^{\alpha}_{\infty}(\mathcal{X}), \rho_X \in \Phi_\rho} \mathbb{E}\left[\mathcal{E}\left(\widehat{f}^{over}\right) - \mathcal{E}(f_\rho)\right] \lesssim \left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+d}} .$$

Textbook results (*optimal rates of convergence*) on Hölder space [13]

$$\inf_{\hat{f} \in \mathcal{F}} \sup_{f_\rho \in W^{\alpha}_{\infty}(\mathcal{X}), \rho_X \in \Phi_\rho} \mathbb{E}\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho)\right] = \Theta\left(n^{-\frac{2\alpha}{2\alpha+d}}\right) .$$

Standard genralization error under adversarial training

Theorem (standard generalization (Shi, Liu, Cao, Suykens, 2024))

Assume $f_\rho \in W^\alpha_\infty(\mathcal{X})$ *with* $\alpha > 0$, $\rho_X \in \Phi_\rho$ *is non-irregular. If* $\delta < \min\left\{\frac{q_X}{3}, n^{-\frac{2\alpha}{(2\alpha+d)d} - \frac{1}{d}}\right\}$, *then* $\exists \widehat{f}^{over}$ *with depth* $L = \mathcal{O}(\log n)$, *and width* $m_1 = \mathcal{O}(nd)$, $m_2, \dots, m_L = \mathcal{O}(\log n)$, *such that*

$$\sup_{f_\rho \in W^\alpha_\infty(\mathcal{X}), \rho_X \in \Phi_\rho} \mathbb{E}\left[\mathcal{E}\left(\widehat{f}^{over}\right) - \mathcal{E}(f_\rho)\right] \lesssim \left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+d}} .$$

Textbook results (*optimal rates of convergence*) on Hölder space [13]

$$\inf_{\hat{f} \in \mathcal{F}} \sup_{f_\rho \in W^\alpha_\infty(\mathcal{X}), \rho_X \in \Phi_\rho} \mathbb{E}\left[\mathcal{E}(\hat{f}) - \mathcal{E}(f_\rho)\right] = \Theta\left(n^{-\frac{2\alpha}{2\alpha+d}}\right) .$$

- construction based on ρ and data
- linear over-parameterization is enough

Robust overfitting: upper bound

$$\mathcal{E}^\delta(f) - \mathcal{E}^\delta(f_\rho^\delta) \leq \mathcal{E}^\delta(f) - \mathcal{E}(f) + \mathcal{E}(f) - \mathcal{E}(f_\rho)$$

Robust overfitting: upper bound

$$\mathcal{E}^{\delta}(f) - \mathcal{E}^{\delta}(f_{\rho}^{\delta}) \leq \mathcal{E}^{\delta}(f) - \mathcal{E}(f) + \mathcal{E}(f) - \mathcal{E}(f_{\rho})$$

Theorem (robust generalization error (Shi, Liu, Cao, Suykens, 2024))

Assume $f_{\rho} \in W_{\infty}^{\alpha}([0,1]^d)$ *with* $\alpha \geq 2$, *and* $\rho_X \in \Phi_{\rho}$ *is non-irregular. If* $\delta < \frac{1}{3} \min \left\{ n^{-\frac{1}{d-1}}, q_X \right\}$, *then there exists* $\widehat{f}^{over}$ *with*

- *depth* $L = \mathcal{O}\left(\log \frac{1}{\delta}\right)$
- *width* $m_1 = \mathcal{O}\left(\delta^{-\frac{d}{2\alpha-2}} \log \frac{1}{\delta} + nd\right)$, $m_2, \ldots, m_L = \mathcal{O}\left(\delta^{-\frac{d}{2\alpha-2}} \log \frac{1}{\delta}\right)$
- *non-zero free parameters* $\mathcal{O}\left(\delta^{-\frac{d}{2\alpha-2}} \log \frac{1}{\delta} + nd\right)$

such that $\widehat{\mathcal{E}}^{\delta}(\widehat{f}^{over}) = 0$ *and*

$$\mathbb{E}\left[\mathcal{E}^{\delta}(\widehat{f}^{over}) - \mathcal{E}^{\delta}(f_{\rho}^{\delta})\right] \lesssim \sqrt{d}\delta\,.$$

Robust overfitting: upper bound

$$\mathcal{E}^{\delta}(f) - \mathcal{E}^{\delta}(f_{\rho}^{\delta}) \leq \mathcal{E}^{\delta}(f) - \mathcal{E}(f) + \mathcal{E}(f) - \mathcal{E}(f_{\rho})$$

Theorem (robust generalization error (Shi, Liu, Cao, Suykens, 2024))

Assume $f_{\rho} \in W_{\infty}^{\alpha}([0,1]^d)$ *with* $\alpha \geq 2$*, and* $\rho_X \in \Phi_{\rho}$ *is non-irregular. If* $\delta < \frac{1}{3} \min \left\{ n^{-\frac{1}{d-1}}, q_X \right\}$*, then there exists* $\widehat{f}^{over}$ *with*

- *depth* $L = \mathcal{O}\left(\log \frac{1}{\delta}\right)$
- *width* $m_1 = \mathcal{O}\left(\delta^{-\frac{d}{2\alpha-2}} \log \frac{1}{\delta} + nd\right)$, $m_2, \ldots, m_L = \mathcal{O}\left(\delta^{-\frac{d}{2\alpha-2}} \log \frac{1}{\delta}\right)$
- *non-zero free parameters* $\mathcal{O}\left(\delta^{-\frac{d}{2\alpha-2}} \log \frac{1}{\delta} + nd\right)$

such that $\widehat{\mathcal{E}}^{\delta}(\widehat{f}^{over}) = 0$ *and*

$$\mathbb{E}\left[\mathcal{E}^{\delta}(\widehat{f}^{over}) - \mathcal{E}^{\delta}(f_{\rho}^{\delta})\right] \lesssim \sqrt{d}\delta\,.$$

Remark:

- If $\frac{1}{3}n^{-\frac{1}{d-1}} \leq \delta < \frac{q_X}{3} \leq \frac{1}{3}n^{-\frac{1}{d}}$, we have robust excess risk $\lesssim \sqrt{d}\left((4+2C_0)\delta\right)^d n, \quad \forall C_0 \in (0,1]$.

Summary: take-away messages

	#parameters	Upper bound
standard generalization	$\mathcal{O}(nd)$	$\widetilde{\mathcal{O}}\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$
robust generalization	$\mathcal{O}\left(nd + \delta^{-\frac{d}{2\alpha-2}} \log \frac{1}{\delta}\right)$	$\mathcal{O}(\sqrt{d}\delta)$

- more smooth, less #params
- Examples: $\delta < n^{-\frac{1}{d}}$
  - $\delta = \frac{1}{n}$: robust overfitting?

Summary: take-away messages

	#parameters	Upper bound
standard generalization	$\mathcal{O}(nd)$	$\widetilde{\mathcal{O}}\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$
robust generalization	$\mathcal{O}\left(nd+\delta^{-\frac{d}{2\alpha-2}}\log\frac{1}{\delta}\right)$	$\mathcal{O}(\sqrt{d}\delta)$

- more smooth, less #params
- Examples: $\delta < n^{-\frac{1}{d}}$
  - $\delta = \frac{1}{n}$: robust overfitting?

well-separated data + target function is smooth enough + perturbation is small enough

$\Rightarrow$ **Avoid robust overfitting!**

Summary: take-away messages

	#parameters	Upper bound
standard generalization	$\mathcal{O}(nd)$	$\widetilde{\mathcal{O}}\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$
robust generalization	$\mathcal{O}\left(nd + \delta^{-\frac{d}{2\alpha-2}} \log \frac{1}{\delta}\right)$	$\mathcal{O}(\sqrt{d}\delta)$

- more smooth, less #params
- Examples: $\delta < n^{-\frac{1}{d}}$
  - $\delta = \frac{1}{n}$: robust overfitting?

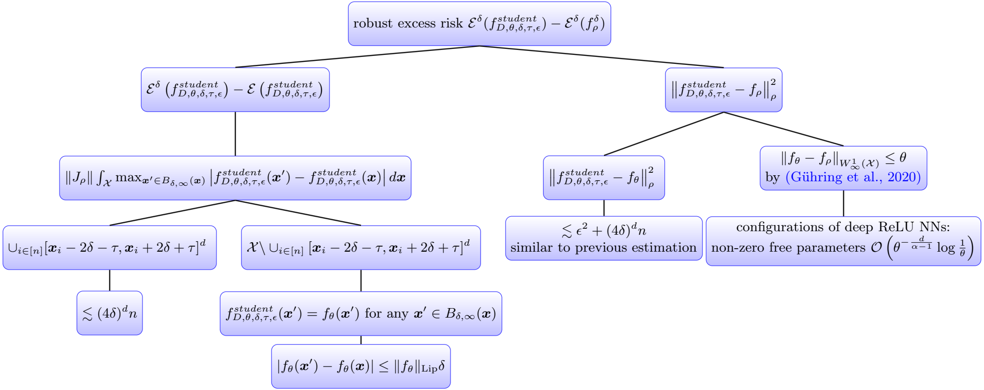
well-separated data + target function is smooth enough + perturbation is small enough

$\Rightarrow$ **Avoid robust overfitting!**

- **robust generalization gap** by taking $\delta := n^{-\frac{2\alpha}{2\alpha+d}} < n^{-\frac{1}{d}}$

$\alpha > \frac{d}{2(d-1)}$ and $\alpha > 2$	#parameters	Upper bound
robust generalization	$\widetilde{\mathcal{O}}\left(nd + n^{\frac{\alpha d}{(2\alpha+d)(\alpha-1)}}\right)$	$\mathcal{O}\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$

Proof roadmap



$$f^{student}_{D,\theta,\delta,\tau,\epsilon}(\boldsymbol{x}) := \sum_{i=1}^n y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + c_3 \tilde{\times}_\epsilon \left(\frac{f_\theta(\boldsymbol{x})}{c_3}, 1 - \sum_{i=1}^n \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right).$$

Is construction optimal? - robust generalization

Theorem (Robust generalization error: lower bound)

Under the same setting of results from [Theorem robust generalization error (upper bound)], we have

$$\mathbb{E}\left[\mathcal{E}^{\delta}(\widehat{f}_D^{over}) - \mathcal{E}^{\delta}(f_\rho^{\delta})\right] \geq \|\bar{J}_\rho\|\sigma^2(4\delta)^d n - \left[\mathcal{E}^{\delta}(f_\rho^{\delta}) - \mathcal{E}\left(f_\rho\right)\right]$$
$$\geq \|\bar{J}_\rho\|\sigma^2(4\delta)^d n - \bar{C}_1\|J_\rho\|\sqrt{d}\delta\,,$$

where $\bar{C}_1$ is a constant independent of d, n and δ.

Is construction optimal? - robust generalization

Theorem (Robust generalization error: lower bound)

Under the same setting of results from [Theorem robust generalization error (upper bound)], we have

$$\mathbb{E}\left[\mathcal{E}^{\delta}(\widehat{f}_D^{over}) - \mathcal{E}^{\delta}(f_\rho^{\delta})\right] \geq \|\bar{J}_\rho\|\sigma^2(4\delta)^d n - \left[\mathcal{E}^{\delta}(f_\rho^{\delta}) - \mathcal{E}\left(f_\rho\right)\right]$$
$$\geq \|\bar{J}_\rho\|\sigma^2(4\delta)^d n - \bar{C}_1\|J_\rho\|\sqrt{d}\delta\,,$$

where $\bar{C}_1$ is a constant independent of d, n and δ.

- $\mathcal{E}^{\delta}(f_\rho^{\delta}) - \mathcal{E}(f_\rho)$ only depends on the distribution
- C_0 can be sufficiently small to match the upper bound (if $n^{-\frac{1}{d-1}} \leq \delta < \frac{q_X}{3}$)

Is construction optimal? - robust generalization

Theorem (Robust generalization error: lower bound)

Under the same setting of results from [Theorem robust generalization error (upper bound)], we have

$$\mathbb{E}\left[\mathcal{E}^{\delta}(\widehat{f}_D^{over}) - \mathcal{E}^{\delta}(f_\rho^{\delta})\right] \geq \|\bar{J}_\rho\|\sigma^2(4\delta)^d n - \left[\mathcal{E}^{\delta}(f_\rho^{\delta}) - \mathcal{E}\left(f_\rho\right)\right]$$
$$\geq \|\bar{J}_\rho\|\sigma^2(4\delta)^d n - \bar{C}_1\|J_\rho\|\sqrt{d}\delta\,,$$

where $\bar{C}_1$ is a constant independent of d, n and δ.

- $\mathcal{E}^{\delta}(f_\rho^{\delta}) - \mathcal{E}(f_\rho)$ only depends on the distribution
- C_0 can be sufficiently small to match the upper bound (if $n^{-\frac{1}{d-1}} \leq \delta < \frac{q_X}{3}$)
- optimal for classification (not included in this talk)

Refer to more results arxiv:2401.13624

Thanks for your attention!

Q & A

my homepage www.lfhsgre.org for more information!

References I

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals.
Understanding deep learning (still) requires rethinking generalization.
Communications of the ACM, 64(3):107–115, 2021.
(Cited on pages 3, 4, and 5.)

[2] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler.
Benign overfitting in linear regression.
the National Academy of Sciences, 2020.
(Cited on pages 3, 4, and 5.)

[3] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein.
Making an invisibility cloak: Real world adversarial attacks on object detectors.
In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.
(Cited on page 6.)

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song.
Robust physical-world attacks on deep learning visual classification.
In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.
(Cited on page 6.)

References II

[5] Leslie Rice, Eric Wong, and Zico Kolter.
Overfitting in adversarially robust deep learning.
In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.
(Cited on pages 7, 8, 9, and 10.)

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.
Explaining and harnessing adversarial examples.
In *International Conference on Learning Representations*, 2015.
(Cited on pages 7, 8, 9, and 10.)

[7] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein.
Adversarial training for free!
In *Advances in Neural Information Processing Systems*, 2019.
(Cited on pages 7, 8, 9, and 10.)

[8] Eric Wong, Leslie Rice, and J Zico Kolter.
Fast is better than free: Revisiting adversarial training.
In *International Conference on Learning Representations*, 2019.
(Cited on pages 7, 8, 9, and 10.)

References III

[9] Binghui Li, Jikai Jin, Han Zhong, John Hopcroft, and Liwei Wang.
Why robust generalization in deep learning is difficult: Perspective of expressive power.
In *Advances in Neural Information Processing Systems*, pages 4370–4384, 2022.
(Cited on pages 11, 12, and 13.)

[10] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri.
A closer look at accuracy vs. robustness.
In *Advances in Neural Information Processing Systems*, pages 8588–8601, 2020.
(Cited on pages 11, 12, and 13.)

[11] Felipe Cucker and Dingxuan Zhou.
Learning theory: an approximation theory viewpoint, volume 24.
Cambridge University Press, 2007.
(Cited on pages 14 and 15.)

[12] Holger Wendland.
Scattered data approximation, volume 17.
Cambridge university press, 2004.
(Cited on pages 16, 17, and 18.)

References IV

[13] László Györfi, Michael Köhler, Adam Krzyżak, and Harro Walk.
A Distribution-Free Theory of Nonparametric Regression, volume 1.
Springer, 2002.
(Cited on pages 19, 20, and 21.)