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

Zhongjie Shi

School of Computing and Data Science
The University of Hong Kong

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

DNNs: the good in fitting...

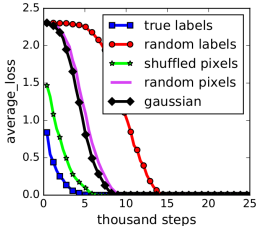
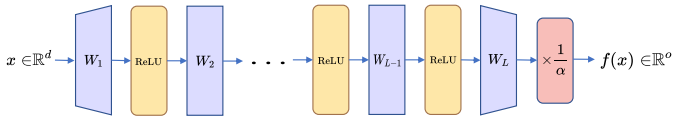


Figure: DNN Training curves on CIFAR10, from [1]

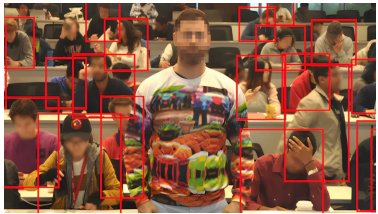
- Empirical risk minimization

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

- Benign overfitting [2]
  - model complex enough to fit random labels
  - zero training error and low test error
  - outside the scope of classical bias-variance tradeoff



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

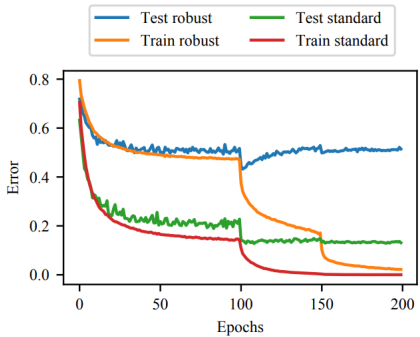


Figure: Results on CIFAR-10 with $\delta = 8/255$ [5].

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: classify separated sets

- ϵ-separated property for real datasets: input data points from different classes have at least 2ϵ distance in the pixel space.

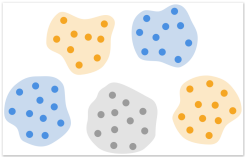


Figure: The class separation in image data. source from [9].

	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. [9]

Motivation: classify separated sets

Theorem (Curse of dimensionality [10])

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

Motivation: classify separated sets

Theorem (Curse of dimensionality [10])

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

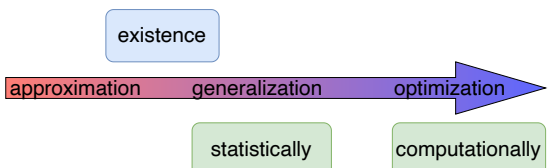
Question: Can overfitted DNNs in adversarial training generalize with a reasonable model complexity?

Motivation: classify separated sets

Theorem (Curse of dimensionality [10])

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

Question: Can overfitted DNNs in adversarial training generalize with a reasonable model complexity?



Preliminary: statistical learning theory (regression)

- Empirical risk minimization

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

- 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: $\|\hat{f} - f_\rho\|_\rho^2$

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: $\|\hat{f} - f_\rho\|_\rho^2$

- 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)

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

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

Assumptions

Assumption (source condition)

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

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

Assumption (non-irregularity of ρ_X)

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

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)

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

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

Assumption (non-irregularity of ρ_X)

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

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

$$q_X := \frac{1}{2} \min_{i\neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty \leq n^{-\frac{1}{d}}\,. \quad [11]$$

Standard genralization error under adversarial training

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

Assume $f_\rho \in W_\infty^\alpha(\mathcal{X})$ *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*

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

Standard genralization error under adversarial training

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

Assume $f_\rho \in W_\infty^\alpha(\mathcal{X})$ *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*

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

Textbook results (*optimal rates of convergence*) on Hölder space [12]

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

Standard genralization error under adversarial training

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

Assume $f_\rho \in W_\infty^\alpha(\mathcal{X})$ *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*

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

Textbook results (*optimal rates of convergence*) on Hölder space [12]

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

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

Robust overfitting: upper bound

- Robust generalization of a function f is bounded by the sum of its standard generalization and robustness

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

Robust overfitting: upper bound

- Robust generalization of a function f is bounded by the sum of its standard generalization and robustness

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

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 $\delta < \frac{q_X}{3} \leq \frac{1}{3}n^{-\frac{1}{d}}$, then there exist infinitely many $\widehat{f}^{over}$ 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 $\widehat{\mathcal{E}}^\delta(\widehat{f}_D^{over}) = 0$ and

$$\mathbb{E}\left[\mathcal{E}^\delta(\widehat{f}_D^{over}) - \mathcal{E}^\delta(f_\rho^\delta)\right] \lesssim \max\left\{\sqrt{d}\delta, (4\delta)^d n\right\}.$$

Robust overfitting: upper bound

- Robust generalization of a function f is bounded by the sum of its standard generalization and robustness

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

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 $\delta < \frac{q_X}{3} \leq \frac{1}{3}n^{-\frac{1}{d}}$, then there exist infinitely many $\widehat{f}^{over}$ 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 $\widehat{\mathcal{E}}^\delta(\widehat{f}_D^{over}) = 0$ and

$$\mathbb{E}\left[\mathcal{E}^\delta(\widehat{f}_D^{over}) - \mathcal{E}^\delta(f_\rho^\delta)\right] \lesssim \max\left\{\sqrt{d}\delta, (4\delta)^d n\right\}.$$

Remark:

- δ is sufficiently small: $\delta < n^{-\frac{1}{d-1}}$, we have robust excess risk $\lesssim \sqrt{d}\delta$
- 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 $\lesssim (4\delta)^d n$

Is construction optimal? - lower bound

Theorem (lower bound under the hinge loss (Shi, Liu, Cao, Suykens, 2024))

Under the same setting of the above theorem, denote σ^2 as the variance of the label noise. For any adversarial training global minimum $\widehat{f}_D^{over}$ of the empirical adversarial risk minimization algorithm over a DNN hypothesis space with $\mathcal{O}\left(\delta^{-\frac{d}{2\alpha-2}}\log\frac{1}{\delta}+nd\right)$ non-zero free parameters, we have $\widehat{\mathcal{E}^\delta}(\widehat{f}_D^{over})=0$, and

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}^\delta(\widehat{f}_D^{over})-\mathcal{E}^\delta(f_\rho^\delta)\right] &\gtrsim \sigma^2(4\delta)^d n-\left[\mathcal{E}^\delta(f_\rho^\delta)-\mathcal{E}(f_\rho)\right]\\ &\gtrsim \sigma^2(4\delta)^d n-\sqrt{d}\delta\,.\end{aligned}$$

- $\mathcal{E}^\delta(f_\rho^\delta)-\mathcal{E}(f_\rho)$ only depends on the distribution
- not optimal if $\delta < n^{-\frac{1}{d-1}}$
- optimal if $n^{-\frac{1}{d-1}}\le\delta<\frac{q_X}{3}\le\frac{1}{3}n^{-\frac{1}{d}}$

Summary: regression tasks

	#parameters	Upper bounds
standard generalization	$\mathcal{O}(nd)$	$\widetilde{\mathcal{O}}\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$
robust generalization	$\mathcal{O}\left(nd+\delta^{-\frac{d}{2\alpha-2}}\log\frac{1}{\delta}\right)$	$\mathcal{O}(\sqrt{d}\delta)$ (if $\delta < n^{-\frac{1}{d-1}}$) $\mathcal{O}((4\delta)^d n)$ (otherwise)

- more parameters for robust solutions
- more smooth, less #params
- smaller perturbation, less #params

target function is smooth enough + perturbation is small enough

$\Rightarrow$ **Avoid robust overfitting with a reasonable model complexity!**

Summary: regression tasks

	#parameters	Upper bounds
standard generalization	$\mathcal{O}(nd)$	$\widetilde{\mathcal{O}}\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$
robust generalization	$\mathcal{O}\left(nd+\delta^{-\frac{d}{2\alpha-2}}\log\frac{1}{\delta}\right)$	$\mathcal{O}(\sqrt{d}\delta)$ (if $\delta < n^{-\frac{1}{d-1}}$) $\mathcal{O}((4\delta)^d n)$ (otherwise)

- more parameters for robust solutions
- more smooth, less #params
- smaller perturbation, less #params

- **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	$\widetilde{\mathcal{O}}\left(nd+n^{\frac{\alpha d}{(2\alpha+d)(\alpha-1)}}\right)$	$\mathcal{O}\left(n^{-\frac{2\alpha}{2\alpha+d}}\right)$

- #parameters: $\widetilde{\mathcal{O}}(nd+n^{\frac{3}{2}})$ when taking $\alpha = 3$
- #parameters: $\widetilde{\mathcal{O}}(nd)$ when taking $\alpha \geq \Omega(\sqrt{d})$.

Preliminary: statistical learning theory (classification)

- Learn the Bayes rule

$$f_c(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \rho(y=1|\boldsymbol{x}) \geq \rho(y=-1|\boldsymbol{x})\,, \\ -1, & \text{if } \rho(y=1|\boldsymbol{x}) < \rho(y=-1|\boldsymbol{x})\,, \end{cases}$$

- The standard misclassification error

$$\mathcal{R}(f) := \int_{\mathcal{Z}} 1_{\{yf(\boldsymbol{x})=-1\}} d\rho$$

- Empirical ϕ-risk minimization with surrogate loss ϕ

$$\widehat{f}_{D,\phi} = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi\left(y_i f(\boldsymbol{x}_i)\right) \right\}$$

- Excess misclassification error

$$\mathcal{R}(\mathsf{sgn}(\widehat{f}_{D,\phi})) - \mathcal{R}(f_c)$$

Preliminary: statistical learning theory (classification)

- Learn the Bayes rule

$$f_c(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \rho(y=1|\boldsymbol{x}) \geq \rho(y=-1|\boldsymbol{x}), \\ -1, & \text{if } \rho(y=1|\boldsymbol{x}) < \rho(y=-1|\boldsymbol{x}), \end{cases}$$

- The standard misclassification error

$$\mathcal{R}(f) := \int_{\mathcal{Z}} 1_{\{yf(\boldsymbol{x})=-1\}} d\rho$$

- Empirical ϕ-risk minimization with surrogate loss ϕ

$$\widehat{f}_{D,\phi} = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \phi\left(y_i f(\boldsymbol{x}_i)\right) \right\}$$

- Excess misclassification error

$$\mathcal{R}(\mathsf{sgn}(\widehat{f}_{D,\phi})) - \mathcal{R}(f_c)$$

- Learn the robust target function

$$f_c^{\delta} = \arg\min_{f} \mathcal{R}^{\delta}(f)$$

- The robust misclassification error

$$\mathcal{R}^{\delta}(f) := \int_{\mathcal{Z}} \max_{\boldsymbol{x}' \in B_{\delta,\infty}(\boldsymbol{x})} 1_{\{yf(\boldsymbol{x}')=-1\}} d\rho$$

- Empirical robust ϕ-risk minimization

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

- Robust excess risk:

$$\mathcal{R}^{\delta}(\widehat{f}_{D,\phi}^{over}) - \mathcal{R}^{\delta}(f_c^{\delta})$$

Additional assumptions

Assumption (well separated data)

Denote $A = \{\boldsymbol{x} \in \mathcal{X} : f_c(\boldsymbol{x}) = 1\}$ *and* $B = \{\boldsymbol{x} \in \mathcal{X} : f_c(\boldsymbol{x}) = -1\}$, *clearly we have* $\mathcal{X} = A \cup B$. *The two classes are* 2δ*-separated if*

$$\|\boldsymbol{x}_A - \boldsymbol{x}_B\|_\infty \geq 2\delta, \quad \forall \boldsymbol{x}_A \in A,\ \boldsymbol{x}_B \in B\,.$$

Assumption (regularity assumption and high confidence of the Bayes rule)

Denote $\eta(\boldsymbol{x}) := \rho(y = 1|\boldsymbol{x})$. *We assume that* $\eta \in W_\infty^\alpha(\mathcal{X})$ *with* $\alpha \in \mathbb{N}$. *Besides, there exists some arbitrary small constant* $\zeta > 0$ *such that*

$$|\eta(\boldsymbol{x}) - 0.5| > \zeta, \quad \forall \boldsymbol{x} \in \mathcal{X}\,.$$

Robust overfitting: upper bound

- Robust misclassification of a classifier f is bounded by the sum of its standard misclassification and robustness

$$\mathcal{R}^{\delta}(f) - \mathcal{R}^{\delta}(f_c^{\delta}) \leq \mathcal{R}^{\delta}(f) - \mathcal{R}(f) + \mathcal{R}(f) - \mathcal{R}(f_c)$$

Robust overfitting: upper bound

- Robust misclassification of a classifier f is bounded by the sum of its standard misclassification and robustness

$$\mathcal{R}^{\delta}(f) - \mathcal{R}^{\delta}(f_c^{\delta}) \leq \mathcal{R}^{\delta}(f) - \mathcal{R}(f) + \mathcal{R}(f) - \mathcal{R}(f_c)$$

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

Assume $\eta \in W_{\infty}^{\alpha}([0,1]^d)$ *with* $\alpha \in \mathbb{N}$, $|\eta(\boldsymbol{x}) - 0.5| > \zeta, \quad \forall \boldsymbol{x} \in \mathcal{X}$ *for some artibtray small constant* ζ, $\rho_X \in \Phi_\rho$ *is non-irregular, and the two classes are* 2δ*-separtaed. If* $\delta < \frac{q_X}{3} \leq \frac{1}{3} n^{-\frac{1}{d}}$, *then there exist infinitely many* $\widehat{f}_D^{over}$ *with*

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

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

$$\mathbb{E}\left[\mathcal{R}\left(sgn\left(\widehat{f}_D^{over}\right)\right) - \mathcal{R}(f_c)\right] \lesssim (2\delta)^d n$$
$$\mathbb{E}\left[\mathcal{R}^{\delta}\left(sgn\left(\widehat{f}_D^{over}\right)\right) - \mathcal{R}^{\delta}(f_c^{\delta})\right] \lesssim (4\delta)^d n$$

Is construction optimal? - lower bound

Theorem (lower bound under the hinge loss (Shi, Liu, Cao, Suykens, 2024))

Under the same setting of the above theorem, for any adversarial training global minimum $\widehat{f}_D^{over}$ *of the empirical adversarial risk minimization algorithm over a DNN hypothesis space with* $\mathcal{O}\left(\zeta^{-\frac{d}{\alpha}}\log\frac{1}{\zeta}+nd\right)$ *non-zero free parameters, we have* $\widehat{\mathcal{E}_D^{\phi,\delta}}(\widehat{f}_D^{over})=0$, *and*

$$\mathbb{E}\left[\mathcal{R}^{\delta}\left(sgn\left(\widehat{f}_D^{over}\right)\right)-\mathcal{R}^{\delta}(f_c^{\delta})\right]\gtrsim\zeta\mathcal{R}(f_c)(4\delta)^d n\,. \tag{1}$$

- the robust misclassification error upper bound matches the lower bound

Summary: classification tasks

	Upper bound	Lower bound
standard misclassification error	$\mathcal{O}((2\delta)^d n)$	
robust misclassification error	$\mathcal{O}((4\delta)^d n)$	$\mathcal{O}((4\delta)^d n)$

- non-zero free parameters: $\mathcal{O}\left(\zeta^{-\frac{d}{\alpha}} \log \frac{1}{\zeta} + nd\right)$
- more smooth, better data quality, less #params
- a robust generalization gap exists

Summary: classification tasks

	Upper bound	Lower bound
standard misclassification error	$\mathcal{O}((2\delta)^d n)$	
robust misclassification error	$\mathcal{O}((4\delta)^d n)$	$\mathcal{O}((4\delta)^d n)$

- non-zero free parameters: $\mathcal{O}\left(\zeta^{-\frac{d}{\alpha}} \log \frac{1}{\zeta} + nd\right)$
- more smooth, better data quality, less #params
- a robust generalization gap exists

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

$\Rightarrow$ **Avoid robust overfitting with a reasonable model complexity!**

Summary: classification tasks

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

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

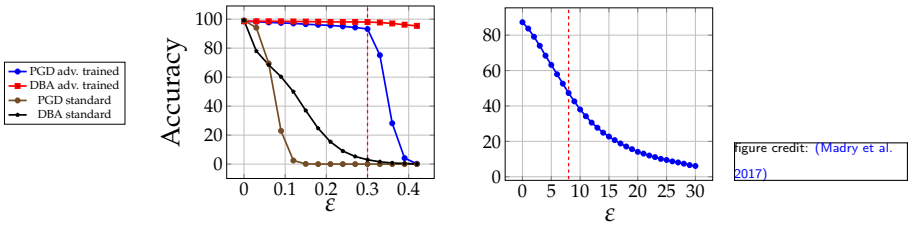


figure credit: (Madry et al. 2017)

(a) MNIST, ℓ_∞-norm (b) CIFAR10, ℓ_∞-norm

Proof sketch: recall standard generalization error result

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

Assume $f_\rho \in W_\infty^\alpha(\mathcal{X})$ *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*

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

Proof sketch: trapezoid-shaped function

Let $\theta, a, b \in \mathbb{R}$ with $a < b$, denote the trapezoid-shaped function $T_{\theta,a,b}$ on $\mathbb{R}$ with a parameter $0 < \theta \leq 1$ as

$$T_{a,b,\theta}(t) := \frac{1}{\theta}\left\{\sigma(t-a+\theta) - \sigma(t-a) - \sigma(t-b) + \sigma(t-b-\theta)\right\}, \quad t \in \mathbb{R}.$$

Proof sketch: trapezoid-shaped function

Let $\theta, a, b \in \mathbb{R}$ with $a < b$, denote the trapezoid-shaped function $T_{\theta,a,b}$ on $\mathbb{R}$ with a parameter $0 < \theta \leq 1$ as

$$T_{a,b,\theta}(t) := \frac{1}{\theta}\left\{\sigma(t-a+\theta) - \sigma(t-a) - \sigma(t-b) + \sigma(t-b-\theta)\right\}, \quad t \in \mathbb{R}.$$

For $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, define (a two-layer ReLU NN with the width $4d$)

$$\Gamma_{a,b,\theta}(\boldsymbol{x}) := \sigma\left(\sum_{k=1}^{d} T_{a,b,\theta}\left(x_k\right) - (d-1)\right) = \begin{cases} 1, & \text{if } \boldsymbol{x} \in [a,b]^d. \\ 0, & \text{if } \boldsymbol{x} \notin [a-\theta, b+\theta]^d. \end{cases}$$

Proof sketch: trapezoid-shaped function

Let $\theta, a, b \in \mathbb{R}$ with $a < b$, denote the trapezoid-shaped function $T_{\theta,a,b}$ on $\mathbb{R}$ with a parameter $0 < \theta \leq 1$ as

$$T_{a,b,\theta}(t) := \frac{1}{\theta}\left\{\sigma(t-a+\theta) - \sigma(t-a) - \sigma(t-b) + \sigma(t-b-\theta)\right\}, \quad t \in \mathbb{R}.$$

For $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, define (a two-layer ReLU NN with the width $4d$)

$$\Gamma_{a,b,\theta}(\boldsymbol{x}) := \sigma\left(\sum_{k=1}^{d} T_{a,b,\theta}\left(x_k\right) - (d-1)\right) = \begin{cases} 1, & \text{if } \boldsymbol{x} \in [a,b]^d. \\ 0, & \text{if } \boldsymbol{x} \notin [a-\theta, b+\theta]^d. \end{cases}$$

$$\Rightarrow \quad \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{x} \in [\boldsymbol{x}_i-\delta, \boldsymbol{x}_i+\delta]^d \\ 0, & \text{if } \boldsymbol{x} \notin [\boldsymbol{x}_i-\delta-\tau, \boldsymbol{x}_i+\delta+\tau]^d. \end{cases}$$

- choosing $\tau \leq \delta < \frac{q_X}{3}$, we have $\Gamma_{\boldsymbol{x}_j-\delta,\boldsymbol{x}_j+\delta,\tau}(\boldsymbol{x}) = 0$ for all $j \neq i$.
- $1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) = 0$

Proof sketch: adversarial training error is zero

Lemma (product-gate property [13])

For any $\epsilon \in (0,1)$, there exists a deep ReLU FNN $\tilde{\times}_\epsilon : \mathbb{R}^2 \to \mathbb{R}$ with depth and free parameters $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ such that

$$\left|\tilde{\times}_\epsilon(x_1, x_2) - x_1 x_2\right| \leq \epsilon, \quad \forall x_1, x_2 \in [-1, 1].$$

Moreover, $\tilde{\times}_\epsilon(x_1, x_2) = 0$ if $x_1 = 0$ or $x_2 = 0$.

Proof sketch: adversarial training error is zero

Lemma (product-gate property [13])

For any $\epsilon \in (0,1)$, there exists a deep ReLU FNN $\tilde{\times}_\epsilon : \mathbb{R}^2 \to \mathbb{R}$ with depth and free parameters $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ such that

$$\left|\tilde{\times}_\epsilon(x_1, x_2) - x_1 x_2\right| \leq \epsilon, \quad \forall x_1, x_2 \in [-1, 1].$$

Moreover, $\tilde{\times}_\epsilon(x_1, x_2) = 0$ if $x_1 = 0$ or $x_2 = 0$.

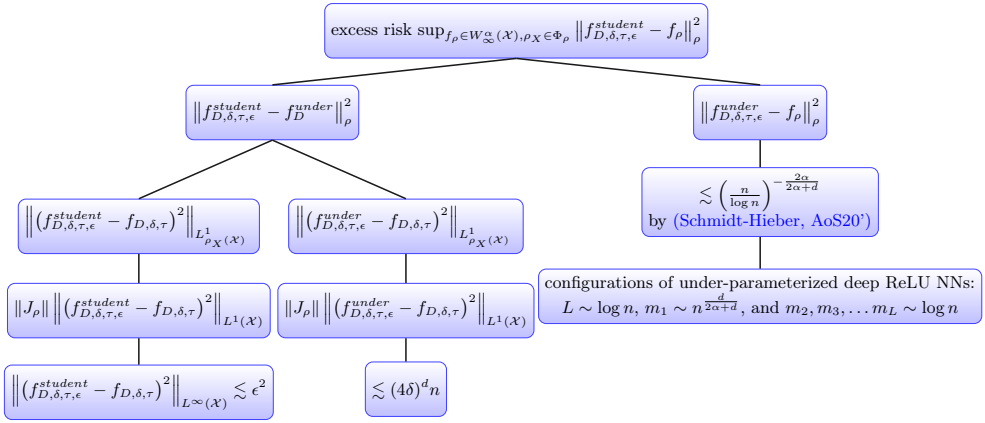
- construct student network

$$f_{D,\delta,\tau,\epsilon}^{student}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + c_1 \tilde{\times}_\epsilon \left(\frac{f_D^{under}(\boldsymbol{x})}{c_1}, 1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right).$$

  based on the teacher network f_D^{under} that achieves near-optimal convergence rates through an under-parameterized ERM algorithm

- $\tilde{\times}_\epsilon \left(\frac{f_D^{under}(\boldsymbol{x})}{c_1}, 1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right) = 0$
- $f_{D,\delta,\tau,\epsilon}^{student}(\boldsymbol{x}) = y_i, \quad$ when $\boldsymbol{x} \in [\boldsymbol{x}_i - \delta, \boldsymbol{x}_i + \delta]^d$

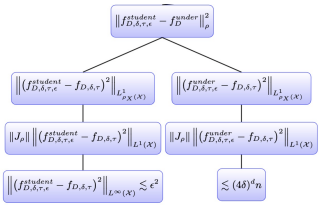
Proof sketch: standard generalization error roadmap



- right side: textbook results for the teacher network f^{under}_D

Proof sketch: generalization error is small

- left side: introducing an intermediate term $f_{D,\delta,\tau}(\boldsymbol{x})$

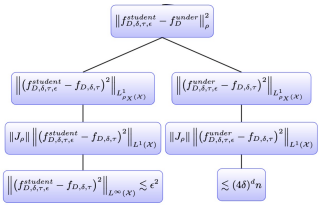


$$f^{student}_{D,\delta,\tau,\epsilon}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + c_1 \tilde{\times}_\epsilon \left(\frac{f^{under}_D(\boldsymbol{x})}{c_1}, 1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right)$$

$$f_{D,\delta,\tau}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + f^{under}_D(\boldsymbol{x}) \left(1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right)$$

Proof sketch: generalization error is small

- left side: introducing an intermediate term $f_{D,\delta,\tau}(\boldsymbol{x})$



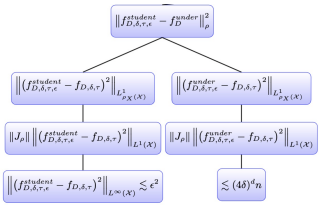
$$f^{student}_{D,\delta,\tau,\epsilon}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + c_1 \tilde{\times}_\epsilon \left(\frac{f^{under}_D(\boldsymbol{x})}{c_1}, 1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right)$$

$$f_{D,\delta,\tau}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + f^{under}_D(\boldsymbol{x}) \left(1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right)$$

- left part: by the product-gate property

Proof sketch: generalization error is small

- left side: introducing an intermediate term $f_{D,\delta,\tau}(\boldsymbol{x})$



$$f^{student}_{D,\delta,\tau,\epsilon}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + c_1 \tilde{\times}_\epsilon \left(\frac{f^{under}_D(\boldsymbol{x})}{c_1}, 1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right)$$

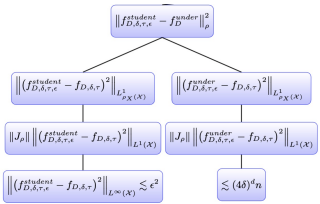
$$f_{D,\delta,\tau}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + f^{under}_D(\boldsymbol{x}) \left(1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right)$$

- left part: by the product-gate property
- right part: divide $\mathcal{X}$ into two parts
  - when $\boldsymbol{x} \in \mathcal{X} \backslash \left(\cup_{i\in\{1,\ldots,n\}} [\boldsymbol{x}_i - \delta - \tau, \boldsymbol{x}_i + \delta + \tau]^d \right)$

    $\Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) = 0$ for all $i \quad \Rightarrow \quad f_{D,\delta,\tau}(\boldsymbol{x}) = f^{under}_D(\boldsymbol{x})$

Proof sketch: generalization error is small

- left side: introducing an intermediate term $f_{D,\delta,\tau}(\boldsymbol{x})$



$$f^{student}_{D,\delta,\tau,\epsilon}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + c_1 \tilde{\times}_\epsilon \left(\frac{f^{under}_D(\boldsymbol{x})}{c_1}, 1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right)$$

$$f_{D,\delta,\tau}(\boldsymbol{x}) := \sum_{i=1}^{n} y_i \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) + f^{under}_D(\boldsymbol{x}) \left(1 - \sum_{i=1}^{n} \Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) \right)$$

- left part: by the product-gate property
- right part: divide $\mathcal{X}$ into two parts
  - when $\boldsymbol{x} \in \mathcal{X} \setminus \left(\cup_{i\in\{1,\ldots,n\}} [\boldsymbol{x}_i - \delta - \tau, \boldsymbol{x}_i + \delta + \tau]^d \right)$

    $\Gamma_{\boldsymbol{x}_i-\delta,\boldsymbol{x}_i+\delta,\tau}(\boldsymbol{x}) = 0$ for all $i \quad \Rightarrow \quad f_{D,\delta,\tau}(\boldsymbol{x}) = f^{under}_D(\boldsymbol{x})$
  - when $\boldsymbol{x} \in \cup_{i\in\{1,\ldots,n\}} [\boldsymbol{x}_i - \delta - \tau, \boldsymbol{x}_i + \delta + \tau]^d$

$$\left\| \left(f_{D,\delta,\tau} - f^{under}_D \right)^2 \right\|_{L^1(\mathcal{X})} = \sum_{i=1}^{n} \int_{[\boldsymbol{x}_i-\delta-\tau,\boldsymbol{x}_i+\delta+\tau]^d} \left(f_{D,\delta,\tau}(\boldsymbol{x}) - f^{under}_D(\boldsymbol{x}) \right)^2 d\boldsymbol{x} \leq (c_1 + M)^2 (4\delta)^d n$$

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 page 2.)

[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 page 2.)

[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 3.)

[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 3.)

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 4 and 5.)

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.
Explaining and harnessing adversarial examples.
In *International Conference on Learning Representations*, 2015.
(Cited on pages 4 and 5.)

[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 4 and 5.)

[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 4 and 5.)

References III

[9] 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 page 6.)

[10] 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 7, 8, and 9.)

[11] Holger Wendland.
Scattered data approximation, volume 17.
Cambridge university press, 2004.
(Cited on pages 12, 13, and 14.)

[12] László Györfi, Michael Köhler, Adam Krzyżak, and Harro Walk.
A Distribution-Free Theory of Nonparametric Regression, volume 1.
Springer, 2002.
(Cited on pages 15, 16, and 17.)

References IV

[13] Dmitry Yarotsky.
Error bounds for approximations with deep ReLU networks.
Neural Networks, 94:103–114, 2017.
(Cited on pages 37 and 38.)