



Luiz F. O. Chamon

robust
learning

Who am I?



Luiz F. O. Chamon

2022–: ELLIS-SimTech research group leader
IMPRS-IS faculty

2021–2022: Simons Institute, UC Berkeley (Postdoc)

2020: University of Pennsylvania (PhD)

< 2015: University of São Paulo, Brazil (BSc. & MSc.)

- I speak 4.5 languages (German = 0)

Disclaimer

What is about to follow...

- ...is **very** opinionated. Change my mind!
- ...contains ~~affiliated links~~ shameless plugs to my own work
- ...contains images and illustrations that I do not own: **do not share publicly**
- ...is **not** exhaustive

https://luizchamon.com/pdf/imprs_bootcamp.pdf

2

Agenda

What is “robust learning”?

Why are models brittle?

How to learn robustly?

What else are adversaries good for?

What is learning?

- Learning:** “model-free statistics” (data-driven)
- Inference:** “model-driven statistics” (data-free)

4

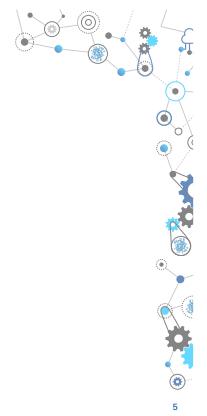
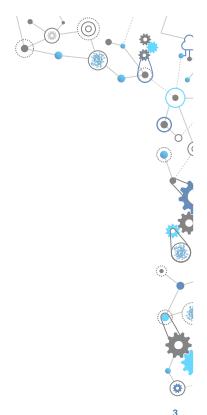
What is learning?

- Learning:** “model-free statistics” (data-driven)
- Inference:** “model-driven statistics” (data-free)

Examples

- How good is an estimator (sample mean)? \Rightarrow confidence interval vs. cross-validation

5



What is learning?

- **Learning:** “model-free statistics” (data-driven)
- **Inference:** “model-driven statistics” (data-free)

Examples

- How good is an estimator (sample mean)? \Rightarrow confidence interval vs. cross-validation
- Statistical optimization: $P^* = \min_{\theta} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\ell(f_{\theta}(\mathbf{x}), y)]$
 - ℓ is the bounded, Lipschitz continuous, possibly non-convex loss
 - f_{θ} is the (possibly nonlinear) parametrization [e.g., logistic classifier, (G)CNN]

[Rostamizadeh, Talwalkar, Mohri. Foundations of machine learning, 2012]; [Ben-David, Shalev-Shwartz. Understanding machine learning..., 2014]

5

What is learning?

- **Learning:** “model-free statistics” (data-driven)
- **Inference:** “model-driven statistics” (data-free)

Examples

- How good is an estimator (sample mean)? \Rightarrow confidence interval vs. cross-validation
- Statistical optimization: $\hat{P}^* = \min_{\theta} \frac{1}{N} \sum_{n=1}^N \ell(f_{\theta}(\mathbf{x}_n), y_n)$, $(\mathbf{x}_n, y_n) \sim \mathcal{D}$ (i.i.d.)
 - ℓ is the bounded, Lipschitz continuous, possibly non-convex loss
 - f_{θ} is the (possibly nonlinear) parametrization [e.g., logistic classifier, (G)CNN]

[Rostamizadeh, Talwalkar, Mohri. Foundations of machine learning, 2012]; [Ben-David, Shalev-Shwartz. Understanding machine learning..., 2014]

5

What is learning theory?

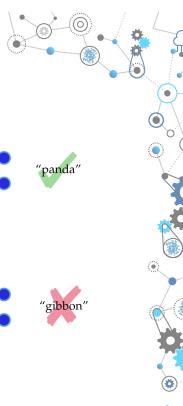
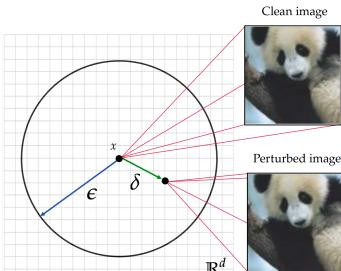
$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(\mathbf{x}_n), y_n) \xrightarrow{\text{"LLN"}} \min_{\theta} \mathbb{E} [\text{Loss}(f_{\theta}(\mathbf{x}), y)]$$

- Is this possible with a finite number of samples (learnability)? When?
e.g., linear functions, smooth functions (finite RKHS norm, bandlimited), NNs...
- How many samples do we need to get a good approximation (sample complexity)?
 $\approx 1/\epsilon^2$

[Rostamizadeh, Talwalkar, Mohri. Foundations of machine learning, 2012]; [Ben-David, Shalev-Shwartz. Understanding machine learning..., 2014]

6

Robustness in learning



8

Robustness in learning

“Since such external effects are rarely expected to help achieve desired system behavior [...] they are commonly referred as *disturbances*.

One of the main objectives of feedback is to render a system insensitive to such disturbances.”

[Dullerud, Paganini. A course in robust control theory, 2005.]

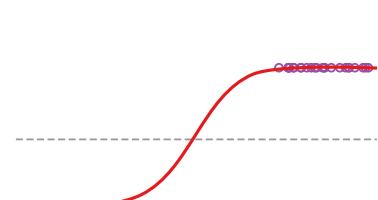
- Strong against input disturbances (\approx classical control: stability)
- Property of model/estimator/classifier



7

Smoothness vs. Robustness

Smoothness \Rightarrow Robustness

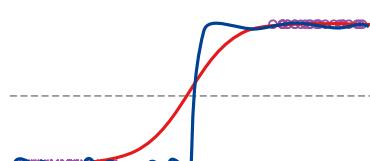


[Hein, Andriushchenko, NeurIPS'17]; [Pauli et al., IEEE Control Systems Letters'22]; [Bungert, Trillo, Murray, IMA Information and Inference'23]

9

Smoothness vs. Robustness

Smoothness \neq Robustness



9

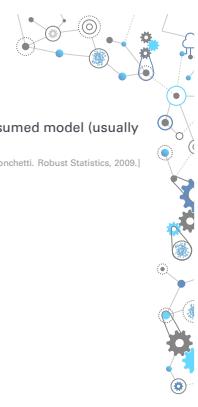
Robustness in statistics inference

“robustness signifies insensitivity to small deviations from the assumptions.”

[...] the shape of the true underlying distribution deviates slightly from the assumed model (usually the Gaussian law).”

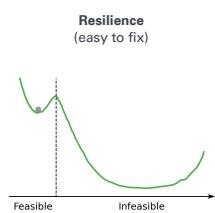
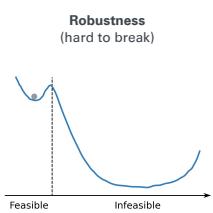
[Huber and Ronchetti. Robust Statistics, 2009.]

- Strong against model misspecification (\approx robust control)
strong against “model misspecification” \equiv (resistant) [Hampel’s theorem]
- Property of statistical problem



10

Resilience vs. Robustness



11

Risk vs. Robustness

Frequent	Rare
Relevant	Model Risk
Irrelevant	Noise Outlier

Robustness

- Value at risk (VaR, V@R), conditional VaR (CVaR, CV@R), expected shortfall...

[Shapiro et al. Lectures on Stochastic Programming, 2014]; [Kalogerias, C., Pappas, Ribeiro. Better Safe Than Sorry..., IEEE ICASSP'20 (best paper)]

Extreme, but rare events

12

Summary

...in learning

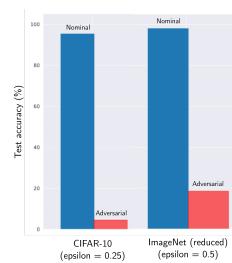
- Strong against input disturbances
- Property of estimator
- "Test-time" distribution shift

- ≠ ...in inference (property of statistical problem)
- ≠ smoothness
- ≠ resilience
- ≠ risk

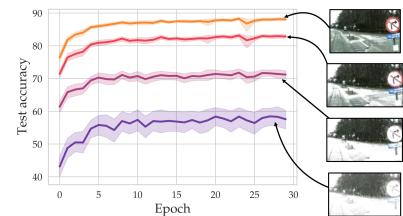
13



Is it actually an issue?



[Ilyas et al. Adversarial examples are not bugs..., NeurIPS'19]; [Robey, Hassani, Pappas. Model-based robust deep learning..., arXiv'20]



14

Agenda

What is "robust learning"?

Why are models brittle?

How to learn robustly?

What else are adversaries good for?

Why are models brittle?

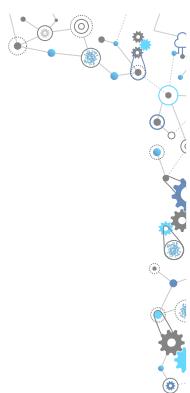
We don't really know
[long awkward pause]

• Overfitting

• Margin

• Features

15



Why are models brittle?

We don't really know
[long awkward pause]

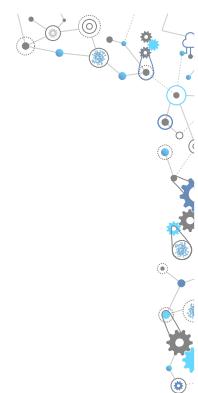
- Overfitting
- Margin
- Features

Why are models brittle?

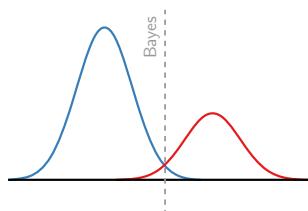
We don't really know
[long awkward pause]

- Overfitting → optimal estimators (and even the "true" one) can be brittle
- Margin
- Features

16



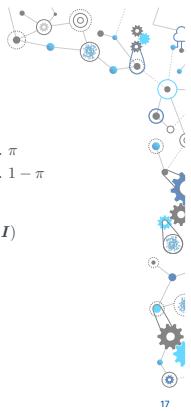
Binary classification



$$y = \begin{cases} +1, & \text{with prob. } \pi \\ -1, & \text{with prob. } 1 - \pi \end{cases}$$

$$x | y \sim \text{Normal}(y\mu, \sigma^2 I)$$

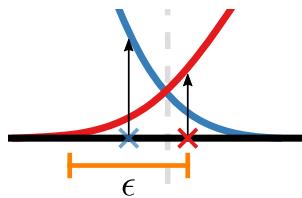
$$\Delta = \{\delta \mid \|\delta\|_2 \leq \epsilon\}$$



[Dobriban, Hassani, Hong, Robey. Provable tradeoffs in adversarially robust classification, IEEE TIT'22]

17

Binary classification



$$y = \begin{cases} +1, & \text{with prob. } \pi \\ -1, & \text{with prob. } 1 - \pi \end{cases}$$

$$x | y \sim \text{Normal}(y\mu, \sigma^2 I)$$

$$\Delta = \{\delta \mid \|\delta\|_2 \leq \epsilon\}$$

[Dobriban, Hassani, Hong, Robey. Provable tradeoffs in adversarially robust classification, IEEE TIT'22]

17

Why are models brittle?

We don't really know
[long awkward pause]

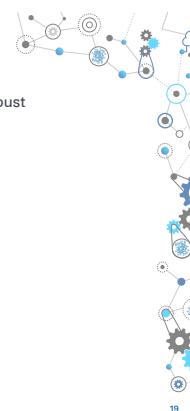
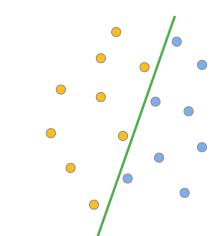


18

- Overfitting → optimal estimators (and even the “true” one) can be brittle
- Margin
- Features

The margin argument

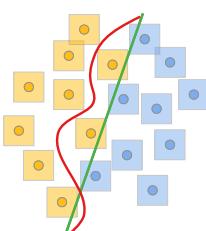
Data points close to the decision boundary are not robust



19

The margin argument

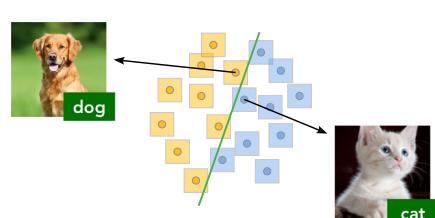
Data points close to the decision boundary are not robust



19

The margin argument

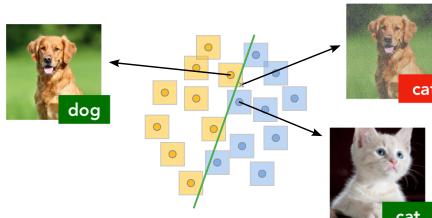
Data points close to the decision boundary are not robust



19

The margin argument

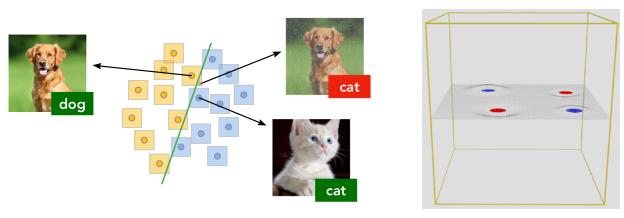
Data points close to the decision boundary are not robust



19

The margin argument

Data points close to the decision boundary are not robust



20

[Shamir, Melamed, BenShmuel. The dimpled manifold model of adversarial examples in machine learning, arXiv'21]

Why are models brittle?

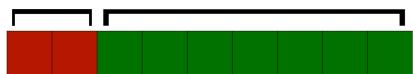
We don't really know
[long awkward pause]

- Overfitting → optimal estimators (and even the “true” one) can be brittle
- Margin
- Features

21

Some features are more robust than others

“Spurious” features “Truly” correlated features



- Adversarial examples arise from “spurious” features

22

[Ilyas, Santurkar,Tsipras, Engstrom,Tran, Mądry. Adversarial examples are not bugs, they are features, NeurIPS’19]

The features argument

Some features are more robust than others



- Adversarial examples arise from “spurious” features
- Useful, non-robust features

23

The features argument

Original training set

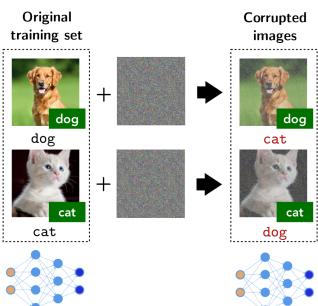


[Ilyas, Santurkar,Tsipras, Engstrom,Tran, Mądry. Adversarial examples are not bugs, they are features, NeurIPS’19]

24

[Ilyas, Santurkar,Tsipras, Engstrom,Tran, Mądry. Adversarial examples are not bugs, they are features, NeurIPS’19]

The features argument



[Ilyas, Santurkar,Tsipras, Engstrom,Tran, Mądry. Adversarial examples are not bugs, they are features, NeurIPS’19]

24

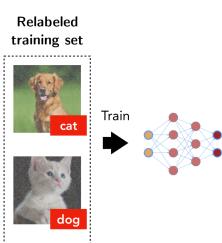
The features argument



[Ilyas, Santurkar,Tsipras, Engstrom,Tran, Mądry. Adversarial examples are not bugs, they are features, NeurIPS’19]

24

The features argument



[Ilyas, Santurkar,Tsipras, Engstrom,Tran, Mądry. Adversarial examples are not bugs, they are features, NeurIPS’19]

25

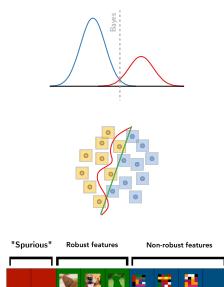
The features argument



[Ilyas, Santurkar,Tsipras, Engstrom,Tran, Mądry. Adversarial examples are not bugs, they are features, NeurIPS’19]

25

Summary



Overfitting
Margin
Features

26

Why are models brittle?



Learning is doing exactly what we asked for...
...and robustness is NOT what we asked for!

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)]$$

27

Agenda

What is “robust learning”?

Why are models brittle?

How to learn robustly?

What else are adversaries good for?



28

Robust learning(ish)

Post-processing

- Filtering

$$f_{\theta^{\dagger}} \longrightarrow f_{\theta^{\dagger}} \circ \varphi$$

[Das, Shanbhogue, Chen, Hohman, Li, Chen, Kounavis, Chau. SHIELD: Fast, Practical... KDD'18]
[Shin, Dawn. JPEG-resistant adversarial images, arXiv'18]



29

Robust learning(ish)

Post-processing

- Filtering

$$f_{\theta^{\dagger}} \longrightarrow f_{\theta^{\dagger}} \circ \varphi$$

[Das, Shanbhogue, Chen, Hohman, Li, Chen, Kounavis, Chau. SHIELD: Fast, Practical... KDD'18]
[Shin, Dawn. JPEG-resistant adversarial images, arXiv'18]

- Smoothed classifiers

$$f_{\theta^{\dagger}}(\cdot) \longrightarrow \mathbb{E}_{z \sim m} [f_{\theta^{\dagger}}(\cdot + z)]$$

[Lecuyer, Attildakis, Geambas, Hsu, and Jana. Certified robustness to adversarial examples..., IEEE SSP'19]
[Cohen, Rosenfeld, Kolter. Certified adversarial robustness via randomized smoothing, ICML'19]
[Blum, Dick, Manoj, Zhang. Random smoothing might be unable to certify ℓ_{∞} robustness..., JMLR'20]
[Anderson, Sojoudi. Certified Robustness via Locally Biased Randomized Smoothing, L4DC'22]



29

Robust learning(ish)

Post-processing

- Filtering

$$f_{\theta^{\dagger}} \longrightarrow f_{\theta^{\dagger}} \circ \varphi$$

[Das, Shanbhogue, Chen, Hohman, Li, Chen, Kounavis, Chau. SHIELD: Fast, Practical... KDD'18]
[Shin, Dawn. JPEG-resistant adversarial images, arXiv'18]

- Smoothed classifiers

$$f_{\theta^{\dagger}}(\cdot) \longrightarrow \mathbb{E}_{z \sim m} [f_{\theta^{\dagger}}(\cdot + z)] = (f_{\theta^{\dagger}} * m)(\cdot)$$

[Lecuyer, Attildakis, Geambas, Hsu, and Jana. Certified robustness to adversarial examples..., IEEE SSP'19]
[Cohen, Rosenfeld, Kolter. Certified adversarial robustness via randomized smoothing, ICML'19]
[Blum, Dick, Manoj, Zhang. Random smoothing might be unable to certify ℓ_{∞} robustness..., JMLR'20]
[Anderson, Sojoudi. Certified Robustness via Locally Biased Randomized Smoothing, L4DC'22]



29

Robust learning

Data augmentation

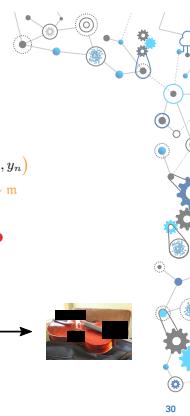
- Random augmentation

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n) \longrightarrow \min_{\theta} \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K \text{Loss}(f_{\theta}(x_n + \delta_k), y_n) \quad \delta_k \sim m$$

[Holmstrom, Koistinen. Using additive noise in back-propagation training, IEEE TNN'92]

[DeVries, Taylor. Improved regularization of convolutional neural networks with cutout, arXiv'17]

[Ford, Gilmer, Carlini, Cubuk. Adversarial examples are a natural consequence of test error in noise, ICML'19]



30

Robust learning

Data augmentation

- Random augmentation

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n) \longrightarrow \min_{\theta} \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K \text{Loss}(f_{\theta}(x_n + \delta_k), y_n) \quad \delta_k \sim m$$

[Holmstrom, Koistinen. Using additive noise in back-propagation training, IEEE TNN'92]

[DeVries, Taylor. Improved regularization of convolutional neural networks with cutout, arXiv'17]

[Ford, Gilmer, Carlini, Cubuk. Adversarial examples are a natural consequence of test error in noise, ICML'19]

- Pseudolabeling

$$\text{Pretrained network: } h^{\dagger} \quad \longrightarrow \min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n) + \frac{1}{K} \sum_{k=1}^K \text{Loss}(f_{\theta}(z_n), h^{\dagger}(z_n))$$

[Carmon, Raghunathan, Schmidt, Duchi, Liang. Unlabeled data improves adversarial robustness. NeurIPS'19]

[Gowal, Rebuffi, Wiles, Stimpert, Calan, Mann. Improving robustness using generated data. NeurIPS'21]



30



Robust learning

Robust optimization

- Adversarial training

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

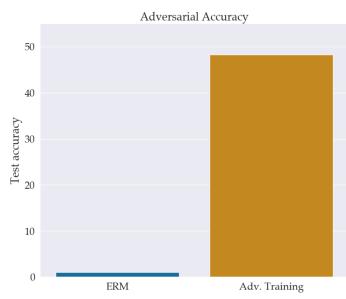
(Goodfellow, Shlens, Szegedy. Explaining and harnessing adversarial examples, ICLR'15)
 [Métry, Makelov, Schmidt, Tsipras, Vladu. Towards deep learning models resistant to adversarial attacks, ICLR'18]
 [HKSC'16, AJFR'17, PPAK'18, SBFZ'19, SCB'19, KCP'20, BT'21, KTM'22, JFEPF'22, WDJTM'23, NFPIY'23...](RobustBench)



31

Adversarial training

Robust optimization



[Robey*, Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

32

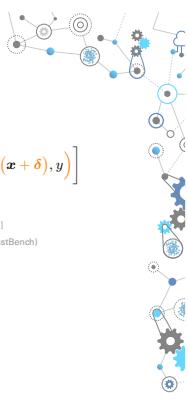
Robust learning

Robust optimization

- Adversarial training

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

(Goodfellow, Shlens, Szegedy. Explaining and harnessing adversarial examples, ICLR'15)
 [Métry, Makelov, Schmidt, Tsipras, Vladu. Towards deep learning models resistant to adversarial attacks, ICLR'18]
 [HKSC'16, AJFR'17, PPAK'18, SBFZ'19, SCB'19, KCP'20, BT'21, KTM'22, JFEPF'22, WDJTM'23, NFPIY'23...](RobustBench)



33

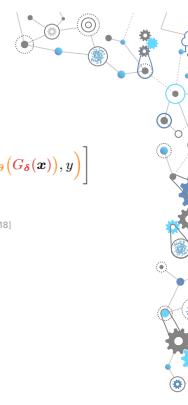
Robust learning

Robust optimization

- Adversarial training

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(G_{\delta}(x)), y) \right]$$

(Goodfellow, Shlens, Szegedy. Explaining and harnessing adversarial examples, ICLR'15)
 [Métry, Makelov, Schmidt, Tsipras, Vladu. Towards deep learning models resistant to adversarial attacks, ICLR'18]
 e.g., [Robey, Hassani, Pappas. Model-based robust deep learning: Generalizing to natural..., arXiv'20]



33

Robust learning

Robust optimization

- Adversarial training

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(G_{\delta}(x)), y) \right]$$

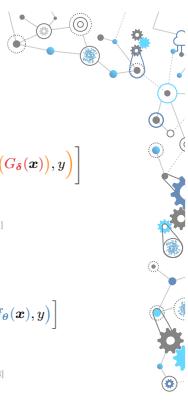
(Goodfellow, Shlens, Szegedy. Explaining and harnessing adversarial examples, ICLR'15)
 [Métry, Makelov, Schmidt, Tsipras, Vladu. Towards deep learning models resistant to adversarial attacks, ICLR'18]
 e.g., [Robey, Hassani, Pappas. Model-based robust deep learning: Generalizing to natural..., arXiv'20]

- Distributional robustness

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \sup_{\mathbf{P} \in \mathcal{P}} \mathbb{E}_{(x,y) \sim \mathbf{P}} [\text{Loss}(f_{\theta}(x), y)]$$

[Ben-Tal, Nemirovski, El Ghaoui. Robust Optimization, 2009]

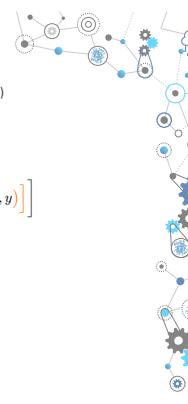
[Sinha, Namkoong, Volpi, Duchi. Certifying some distributional robustness with principled adversarial..., ICLR'18]



33

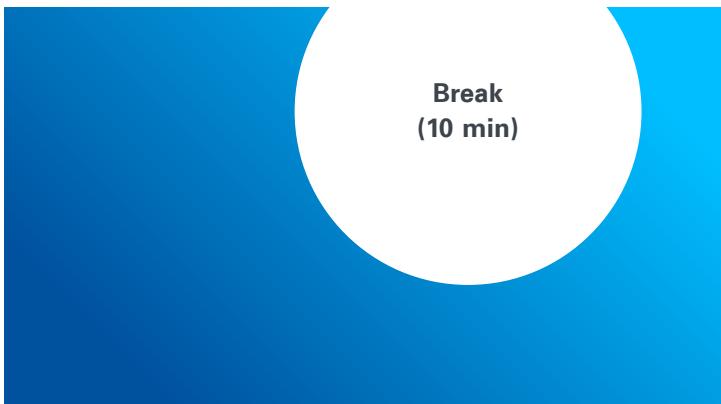
Summary

- Post-processing: filtering ($f_{\theta_1} \circ \varphi$), smoothed classifiers ($\mathbb{E}_{z \sim m}[f_{\theta_1}(\cdot + z)]$)
 - ✓ No need to retrain
 - ✗ Poor empirical performance, weak guarantees
- (Randomized) data augmentation: $\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x + \delta), y)]$
 - ✓ Low complexity
 - ✗ Poor empirical performance, no guarantees
- Adversarial training: $\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$
 - ✓ Great(!) empirically performance
 - ✗ ...



34

**Break
(10 min)**



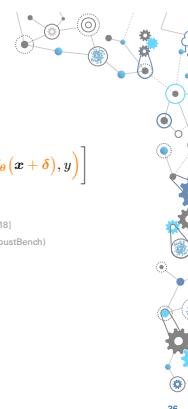
Robust learning

Robust optimization

- Adversarial training

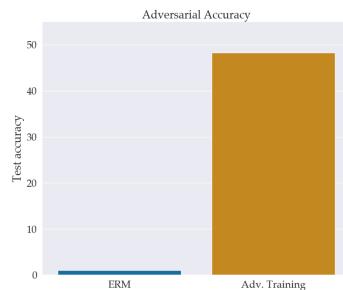
$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

(Goodfellow, Shlens, Szegedy. Explaining and harnessing adversarial examples, ICLR'15)
 [Métry, Makelov, Schmidt, Tsipras, Vladu. Towards deep learning models resistant to adversarial attacks, ICLR'18]
 [HKSC'16, AJFR'17, PPAK'18, SBFZ'19, SCB'19, KCP'20, BT'21, KTM'22, JFEPF'22, WDJTM'23, NFPIY'23...](RobustBench)



36

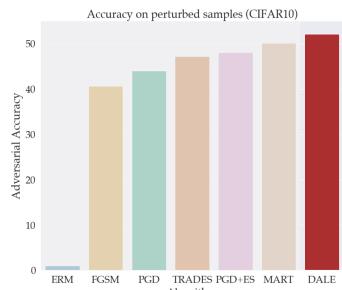
Adversarial training



[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

37

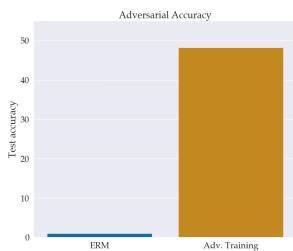
Adversarial training



[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

37

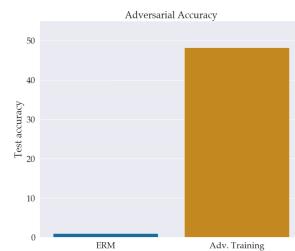
Adversarial training



- ✓ Great(!) empirical performance
- ✓ Grounded on a *robust* literature
- ✗ Statistical complexity
- ✗ Performance trade-offs
- ✗ Computational complexity

38

Adversarial training



- ✓ Great(!) empirical performance
- ✓ Grounded on a *robust* literature
- ✗ Statistical complexity
- ✗ Performance trade-offs
- ✗ Computational complexity

38

Statistical complexity

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x_n + \delta), y_n) \right] \xrightarrow{?} \min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- Is robust learning harder than non-robust learning? Do we need more samples?

Statistical complexity

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x_n + \delta), y_n) \right] \xrightarrow{?} \min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- Is robust learning harder than non-robust learning? Do we need more samples?

A: YES and NO

[Cullina, Bhagoji, Mittal. PAC-learning in the presence of evasion adversaries, NeurIPS'18] ✓

[Yin, Ramchandran, Bartlett. Rademacher Complexity for Adversarially Robust Generalization, ICML'19] ✓

[Montasser, Hanneke, Srebro. VC Classes are Adversarially Robust Learnable, but Only Improperly, COLT'19] ✓

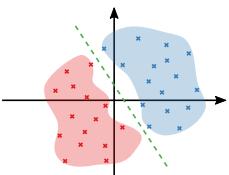
[Awasthi, Frank, Mohri. Adversarial Learning Guarantees for Linear Hypotheses and Neural Networks, ICML'20] ✓

[Montasser, Hanneke, Srebro. Adversarially robust learning: A generic minimax optimal learner & characterization, NeurIPS'22] ✓

Halfspace classifiers with 0-1 loss

$$P^* = \min_{\theta} \Pr_{(x,y)} [f_{\theta}(x) \neq y]$$

$$\hat{P}^* = \min_{\theta} \frac{1}{N} \sum_{n=1}^N \mathbb{I}[f_{\theta}(x_n) \neq y_n]$$



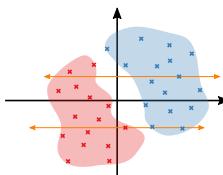
- 0-1 loss, i.e., probability of error
- Halfspace classifiers, i.e., $f_{\theta}(x) = \mathbb{I}[\theta^\top x + \bar{\theta} \geq 0]$
- $|P^* - \hat{P}^*| \leq \epsilon$ w.h.p. $\Leftrightarrow N \approx \frac{d+1}{\epsilon^2}$

40

Halfspace classifiers with 0-1 loss

$$P_r^* = \min_{\theta} \Pr_{(x,y)} [\exists \delta \in \Delta \text{ s.t. } f_{\theta}(x + \delta) \neq y]$$

$$\hat{P}_r^* = \min_{\theta} \frac{1}{N} \sum_{n=1}^N \mathbb{I}[\exists \delta \in \Delta \text{ s.t. } f_{\theta}(x_n + \delta) \neq y_n]$$



- Robust 0-1 loss
- $\Delta = \{\delta \mid \delta = cz, c \in \mathbb{R}\}$ for fixed z

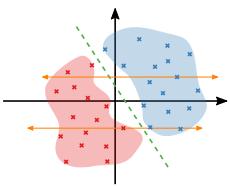
[Cullina, Bhagoji, Mittal. PAC-learning in the presence of evasion adversaries, NeurIPS'18]

41

Halfspace classifiers with 0-1 loss

$$P_r^* = \min_{\theta} \Pr_{(\mathbf{x}, y)} [\exists \delta \in \Delta \text{ s.t. } f_{\theta}(\mathbf{x} + \delta) \neq y]$$

$$\hat{P}_r^* = \min_{\theta} \frac{1}{N} \sum_{n=1}^N \mathbb{I} [\exists \delta \in \Delta \text{ s.t. } f_{\theta}(\mathbf{x}_n + \delta) \neq y_n]$$



- Robust 0-1 loss
- $\Delta = \{\delta \mid \delta = cz, c \in \mathbb{R}\}$ for fixed z

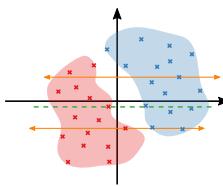
41

[Cullina, Bhagoji, Mittal. PAC-learning in the presence of evasion adversaries, NeurIPS'18]

Halfspace classifiers with 0-1 loss

$$P_r^* = \min_{\theta} \Pr_{(\mathbf{x}, y)} [\exists \delta \in \Delta \text{ s.t. } f_{\theta}(\mathbf{x} + \delta) \neq y]$$

$$\hat{P}_r^* = \min_{\theta} \frac{1}{N} \sum_{n=1}^N \mathbb{I} [\exists \delta \in \Delta \text{ s.t. } f_{\theta}(\mathbf{x}_n + \delta) \neq y_n]$$



[Cullina, Bhagoji, Mittal. PAC-learning in the presence of evasion adversaries, NeurIPS'18]

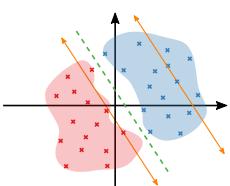
41

41

Halfspace classifiers with 0-1 loss

$$P_r^* = \min_{\theta} \Pr_{(\mathbf{x}, y)} [\exists \delta \in \Delta \text{ s.t. } f_{\theta}(\mathbf{x} + \delta) \neq y]$$

$$\hat{P}_r^* = \min_{\theta} \frac{1}{N} \sum_{n=1}^N \mathbb{I} [\exists \delta \in \Delta \text{ s.t. } f_{\theta}(\mathbf{x}_n + \delta) \neq y_n]$$

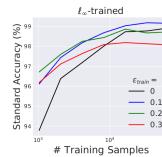


- Robust 0-1 loss
- $\Delta = \{\delta \mid \delta = cz, c \in \mathbb{R}\}$ for fixed z
- $|P_r^* - \hat{P}_r^*| \leq \epsilon \text{ w.h.p.} \Leftrightarrow N \approx \frac{d+1}{\epsilon^2}$
($P_r^* = P^*$!!)

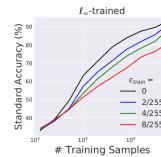
41

[Cullina, Bhagoji, Mittal. PAC-learning in the presence of evasion adversaries, NeurIPS'18]

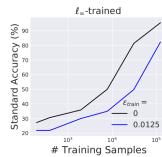
Nominal performance of robust models



(a) MNIST



(b) CIFAR-10



(c) Restricted ImageNet

[Tsipras, Santurkar, Engstrom, Turner, Mądry. Robustness May Be at Odds with Accuracy, ICLR'19]

42

"Softer" robustness

- Softmax or log-sum-exp

$$\min_{\theta} \mathbb{E}_{(\mathbf{x}, y)} \left[\frac{1}{\tau} \log \left(\mathbb{E}_{\delta \sim m} \left[e^{\tau \cdot \text{Loss}(f_{\theta}(\mathbf{x} + \delta), y)} \right] \right) \right]$$

- $\tau \rightarrow 0$: classical learning (with randomized data augmentation)
- $\tau \rightarrow \infty$: adversarial robustness (ess sup)

[Li, Beirami, Sanjabi, Smith. Tilted empirical risk minimization, ICLR'21]



43

"Softer" robustness

- Softmax or log-sum-exp

$$\min_{\theta} \mathbb{E}_{(\mathbf{x}, y)} \left[\frac{1}{\tau} \log \left(\mathbb{E}_{\delta \sim m} \left[e^{\tau \cdot \text{Loss}(f_{\theta}(\mathbf{x} + \delta), y)} \right] \right) \right]$$

- $\tau \rightarrow 0$: classical learning (with randomized data augmentation)
- $\tau \rightarrow \infty$: adversarial robustness (ess sup)

[Li, Beirami, Sanjabi, Smith. Tilted empirical risk minimization, ICLR'21]

- L_p norms

$$\min_{\theta} \mathbb{E}_{(\mathbf{x}, y)} \left[\mathbb{E}_{\delta \sim m} \left[|\text{Loss}(f_{\theta}(\mathbf{x} + \delta), y)|^{\tau} \right]^{1/\tau} \right]$$

- $\tau = 1$: classical learning (with randomized data augmentation)
- $\tau \rightarrow \infty$: adversarial robustness (ess sup)

[Nica, Bair, Zhang, Kolter. Robustness between the worst and average case, NeurIPS'21]



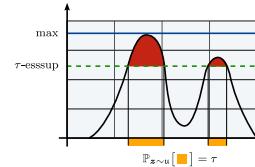
43

"Softer" robustness

- Probabilistic robustness

$$\min_{\theta} \mathbb{E}_{(\mathbf{x}, y)} \left[\underset{\delta \in \Delta}{\tau\text{-essup}} \text{Loss}(f_{\theta}(\mathbf{x} + \delta), y) \right]$$

- $\tau = 1/2$: classical learning (for symmetric m)
- $\tau = 0$: adversarial robustness (ess sup)



44

[Robey, C., Pappas, Hassani. Probabilistically robust learning: Balancing average- and worst-case..., ICML'22 (spotlight)]

"Softer" robustness

- Probabilistic robustness

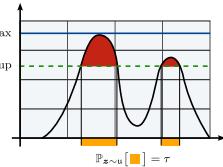
$$\min_{\theta} \mathbb{E}_{(\mathbf{x}, y)} \left[\underset{\delta \in \Delta}{\tau\text{-essup}} \text{Loss}(f_{\theta}(\mathbf{x} + \delta), y) \right]$$

- $\tau = 1/2$: classical learning (for symmetric m)
- $\tau = 0$: adversarial robustness (ess sup)

- Potentially better sample complexity

[Robey, C., Pappas, Hassani, ICML'22 (spotlight)]

[Raman, Tewari, Subedi, NeurIPS ML Safety Workshop'22]



44

[Robey, C., Pappas, Hassani. Probabilistically robust learning: Balancing average- and worst-case..., ICML'22 (spotlight)]

"Softer" robustness

- Probabilistic robustness

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\underset{\delta \in \Delta}{\tau\text{-essup}} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- $\tau = 1/2$: classical learning (for symmetric \mathbf{m})
- $\tau = 0$: adversarial robustness (ess-sup)

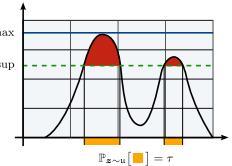
- Potentially better sample complexity

[Robey, C., Pappas, Hassani, ICML'22 (spotlight)]

[Raman, Tewari, Subedi, NeurIPS ML Safety Workshop'22]

- Better performance trade-off

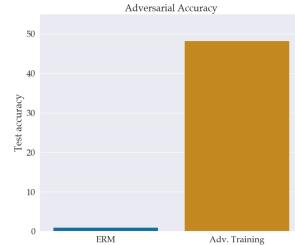
[Robey, C., Pappas, Hassani, ICML'22 (spotlight)]



44

[Robey, C., Pappas, Hassani. Probabilistically robust learning: Balancing average- and worst-case... ICML'22 (spotlight)]

Adversarial training



Great(!) empirical performance

Grounded on a robust literature

Statistical complexity

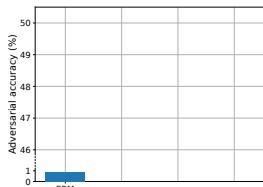
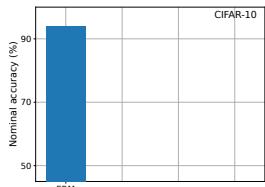
Performance trade-offs

Computational complexity

45

Nominal vs. adversarial performance

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)]$$

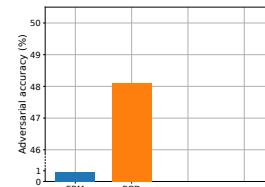
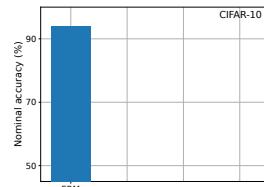


46

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

Nominal vs. adversarial performance

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

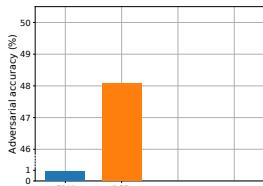
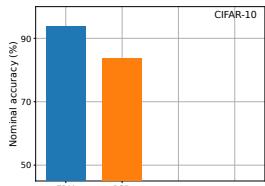


[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

46

Nominal vs. adversarial performance

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

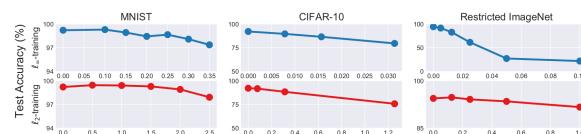


46

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

Nominal vs. adversarial performance

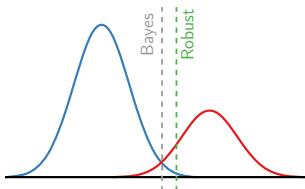
$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$



[Tsigaridas, Santurkar, Engstrom, Turner, Majdry. Robustness may be at odds with accuracy, ICLR'19]

46

Binary classification



$$y = \begin{cases} +1, & \text{with prob. } \pi \\ -1, & \text{with prob. } 1 - \pi \end{cases}$$

$$x | y \sim \text{Normal}(y\mu, \sigma^2 I)$$

$$\Delta = \{\delta \mid \|\delta\|_2 \leq \epsilon\}$$

47

[Dobriban, Hassani, Hong, Robey. Provable tradeoffs in adversarially robust classification, IEEE TIT'22]

Nominal vs. adversarial performance



Learning is doing exactly what we asked for!

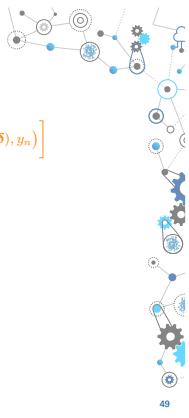
48

Nominal vs. adversarial performance

(Penalized) adversarial training

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n) \longrightarrow \min_{\theta} \frac{1}{N} \sum_{n=1}^N \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x_n + \delta), y_n) \right]$$

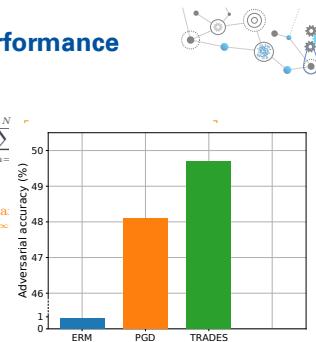
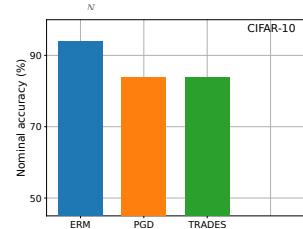
$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n) + \lambda \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x_n + \delta), y_n) \right]$$



49

Nominal vs. adversarial performance

(Penalized) adversarial training



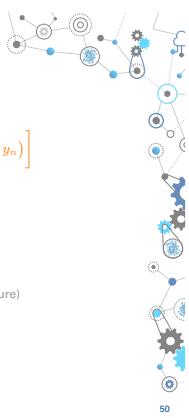
[Zhang, Yu, Jiao, Xing, El Ghaoui, Jordan. Theoretically principled trade-off between robustness and accuracy, ICML'19]

49

Nominal vs. adversarial performance

(Penalized) adversarial training

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n) + \lambda \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x_n + \delta), y_n) \right]$$



50

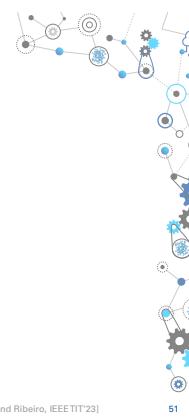
- No straightforward relation between λ and adversarial loss
- λ depends on the values of the losses (dataset, model, performance measure)
- Generalization

Nominal vs. adversarial performance

Constrained learning

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n)$$

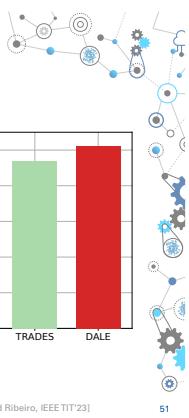
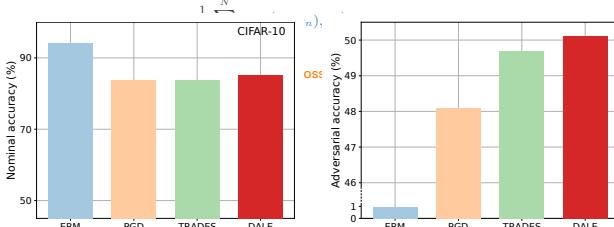
$$\text{subject to } \frac{1}{N} \sum_{n=1}^N \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x_n + \delta), y_n) \right] \leq c$$



51

Nominal vs. adversarial performance

Constrained learning

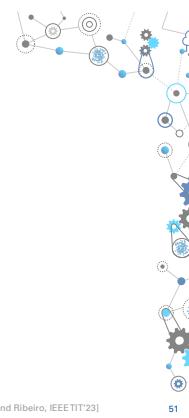
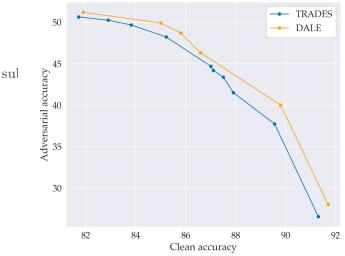


51

[C. and Ribeiro, NeurIPS'20]; [Robey*, C., Pappas, Hassani, and Ribeiro, NeurIPS'21]; [C., Paternain, Calvo-Fullana, and Ribeiro, IEEETIT'23]

Nominal vs. adversarial performance

Constrained learning



51

Nominal vs. adversarial performance

Constrained learning

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n)$$

$$\text{subject to } \frac{1}{N} \sum_{n=1}^N \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x_n + \delta), y_n) \right] \leq c$$



51

- Does constrained learning generalize?
- How many samples do we need?
- How do we solve constrained learning problems?

Nominal vs. adversarial performance

Constrained learning

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_{\theta}(x_n), y_n)$$

$$\text{subject to } \frac{1}{N} \sum_{n=1}^N \left[\max_{\|\delta\|_{\infty} \leq \epsilon} \text{Loss}(f_{\theta}(x_n + \delta), y_n) \right] \leq c$$



51

- Does constrained learning generalize? \Rightarrow yes, same as unconstrained learning
- How many samples do we need? \Rightarrow essentially the same as unconstrained learning
- How do we solve constrained learning problems? \Rightarrow using duality (despite non-convexity)

[C., Paternain, Calvo-Fullana, and Ribeiro. The empirical duality gap of constrained..., IEEE ICASSP'20 (best student paper)]
[C. and Ribeiro. Probably approximately correct constrained learning, NeurIPS'20]
[C., Paternain, Calvo-Fullana, and Ribeiro. Constrained learning with non-convex losses, IEEETIT'23]

Nominal vs. adversarial performance

Dual empirical constrained learning

$$\max_{\lambda \geq 0} \min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_\theta(x_n), y_n) + \lambda \left[\max_{\|\delta\|_\infty \leq \epsilon} \text{Loss}(f_\theta(x_n + \delta), y_n) - c \right]$$



52

Nominal vs. adversarial performance

Dual empirical constrained learning

$$\max_{\lambda \geq 0} \min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_\theta(x_n), y_n) + \lambda \left[\max_{\|\delta\|_\infty \leq \epsilon} \text{Loss}(f_\theta(x_n + \delta), y_n) - c \right]$$

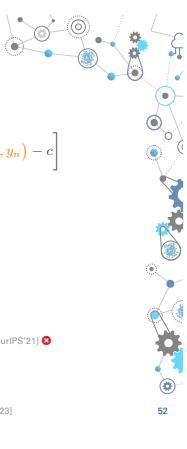
[C. et al., IEEE ICASSP'20 (best student paper)]; [C., Ribeiro, NeurIPS'20]; [C., Paternain, Calvo-Fullana, Ribeiro, IEEE TIT'23]

52

Nominal vs. adversarial performance

Dual empirical constrained learning

$$\max_{\lambda \geq 0} \min_{\theta} \frac{1}{N} \sum_{n=1}^N \text{Loss}(f_\theta(x_n), y_n) + \lambda \left[\max_{\|\delta\|_\infty \leq \epsilon} \text{Loss}(f_\theta(x_n + \delta), y_n) - c \right]$$



52

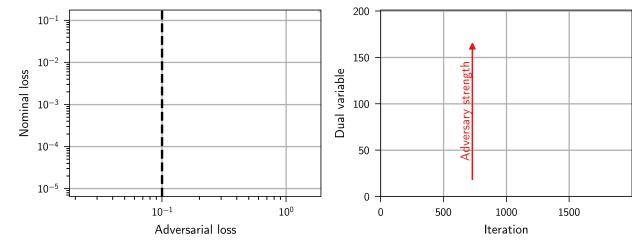
- c (static) is part of the problem, λ (dynamic) is part of the algorithm

- \neq penalty/regularization (static λ)

[Byrd, Lipton. What is the effect of importance weighting in deep learning?, ICML'19]
[Sagawa*, Koh*, Hashimoto, Liang. Distributionally robust neural networks, ICLR'20]
[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]
[C., Paternain, Calvo-Fullana, and Ribeiro. Constrained learning with non-convex losses, IEEE TIT'23]

[C. et al., IEEE ICASSP'20 (best student paper)]; [C., Ribeiro, NeurIPS'20]; [C., Paternain, Calvo-Fullana, Ribeiro, IEEE TIT'23]

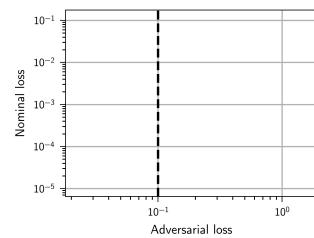
Robust image recognition



[C., Paternain, Calvo-Fullana, and Ribeiro, IEEE TIT'23]

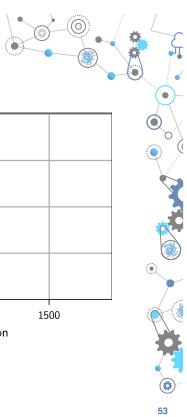
53

Robust image recognition

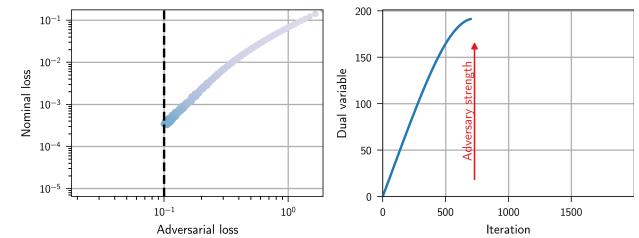


[C., Paternain, Calvo-Fullana, and Ribeiro, IEEE TIT'23]

53



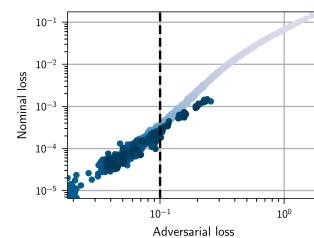
Robust image recognition



[C., Paternain, Calvo-Fullana, and Ribeiro, IEEE TIT'23]

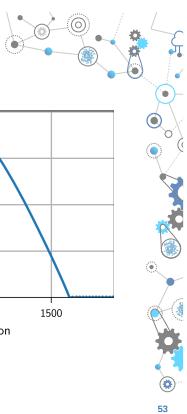
53

Robust image recognition

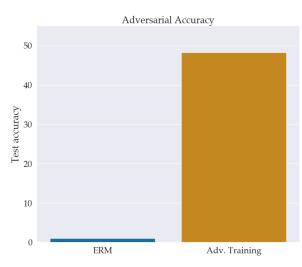


Empirical observations: [Zhang et al., ICML'20; Sitawarin, ArXiv'20]

53



Adversarial training



- Great(!) empirical performance
- Grounded on a robust literature
- Statistical complexity
- Performance trade-offs
- Computational complexity

54

Computational complexity

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- Hard to evaluate/optimize



Computational complexity

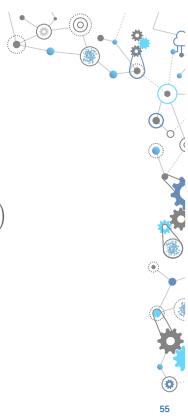
$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- Hard to evaluate/optimize

$$\begin{aligned} \text{“}\nabla_{\theta}\text{”} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right] &= \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta^*), y) \\ \delta^* &\in \operatorname{argmax}_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \end{aligned}$$

▪ Typically non-convex, e.g., $f_{\theta}(\cdot) = \sum_{m=1}^M \theta_m \kappa(\cdot, c_m)$

▪ Often underparametrized ($\theta \in \mathbb{R}^p$, $x \in \mathbb{R}^n$, and $p \gg n$)



Computational complexity

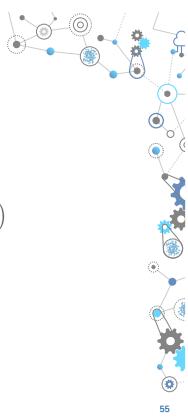
$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- Hard to evaluate/optimize

$$\begin{aligned} \partial_{\theta} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right] &\ni \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta^*), y) \\ \delta^* &\in \operatorname{argmax}_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \end{aligned}$$

▪ Typically non-convex, e.g., $f_{\theta}(\cdot) = \sum_{m=1}^M \theta_m \kappa(\cdot, c_m)$

▪ Often underparametrized ($\theta \in \mathbb{R}^p$, $x \in \mathbb{R}^n$, and $p \gg n$)



“PGD”

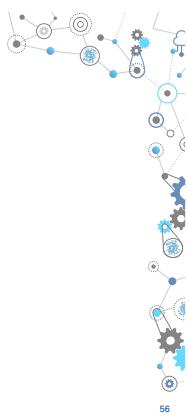
$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- “PGD” [Médré et al., ICLR’18] (alternating optimization, gradient ascent-descent)

```

1:  $\delta^1 \leftarrow \delta_{t-1}$ 
2: for  $k = 1, \dots, K$ 
3:    $\delta^{k+1} \leftarrow \operatorname{proj}_{\Delta} \left[ \delta^k + \eta \operatorname{sign} (\nabla_{\delta} \text{Loss}(f_{\theta^k}(x + \delta^k), y)) \right]$ 
4: end
5:  $\delta_t \leftarrow \delta^{K+1}$ 
6:  $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta_t), y)$ 

```



“PGD”

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- “PGD” [Médré et al., ICLR’18] (alternating optimization, gradient ascent-descent)

```

1:  $\delta^1 \leftarrow \delta_{t-1}$ 
2: for  $k = 1, \dots, K$ 
3:    $\delta^{k+1} \leftarrow \operatorname{proj}_{\Delta} \left[ \delta^k + \eta \operatorname{sign} (\nabla_{\delta} \text{Loss}(f_{\theta^k}(x + \delta^k), y)) \right]$ 
4: end
5:  $\delta_t \leftarrow \delta^{K+1}$ 
6:  $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta_t), y)$ 

```

- Random initialization
- Restarts
- Pruning
- Adaptive step size



“PGD”

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- “PGD” [Médré et al., ICLR’18] (alternating optimization, gradient ascent-descent)

```

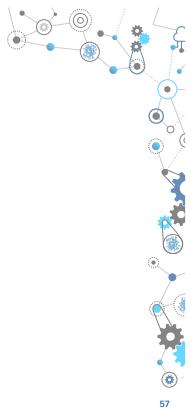
1:  $\delta^1 \leftarrow \delta_{t-1}$ 
2: for  $k = 1, \dots, K$ 
3:    $\delta^{k+1} \leftarrow \operatorname{proj}_{\Delta} \left[ \delta^k + \eta \operatorname{sign} (\nabla_{\delta} \text{Loss}(f_{\theta^k}(x + \delta^k), y)) \right]$ 
4: end
5:  $\delta_t \leftarrow \delta^{K+1}$ 
6:  $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta_t), y)$ 

```

≈ data augmentation

Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$



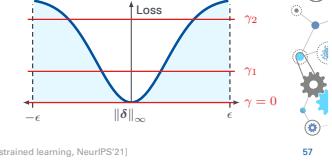
Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\mathbb{E}_{\delta \sim \mu(\delta|x,y)} [\text{Loss}(f_{\theta}(x + \delta), y)] \right]$$

- For any approximation error, $\exists \gamma(x, y)$ such that

$$\mu(\delta|x, y) \propto [\text{Loss}(f_{\theta}(x + \delta), y) - \gamma(x, y)]_+$$



[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

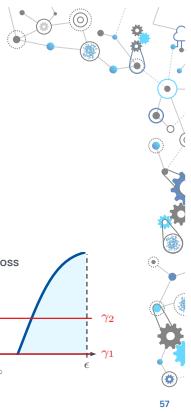
57

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

57

Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$



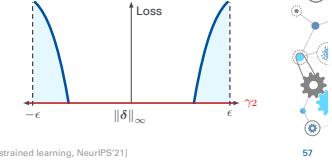
Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\mathbb{E}_{\delta \sim \mu(\delta|x,y)} [\text{Loss}(f_{\theta}(x + \delta), y)] \right]$$

- For any approximation error, $\exists \gamma(x, y)$ such that

$$\mu(\delta|x, y) \propto [\text{Loss}(f_{\theta}(x + \delta), y) - \gamma(x, y)]_+$$



[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

57

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

57

Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$



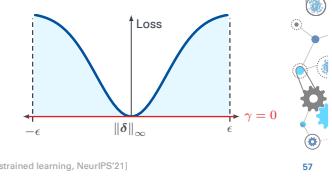
Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\mathbb{E}_{\delta \sim \mu(\delta|x,y)} [\text{Loss}(f_{\theta}(x + \delta), y)] \right]$$

- For any approximation error, $\exists \gamma(x, y)$ such that

$$\mu(\delta|x, y) \propto [\text{Loss}(f_{\theta}(x + \delta), y) - \gamma(x, y)]_+$$



[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

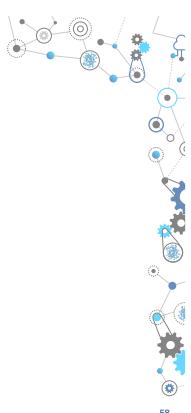
57

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

57

Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$



Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

$$1: \delta_t \sim \text{Loss}(f_{\theta}(x + \delta), y) \\ 2: \theta_{t+1} = \theta_t - \eta \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta_t), y)$$

- ✓ Use MCMC techniques (e.g., Langevin Monte Carlo)

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

58

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

58

Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

```

1:  $\delta^t \leftarrow \delta_{t-1}$ 
2: for  $k = 1, \dots, K$ 
3:    $\delta^{k+1} = \text{proj}_{\Delta} \left[ \delta^k + \eta \text{sign} \left[ \nabla_{\delta} \log (\text{Loss}(f_{\theta}(x + \delta^k), y)) \right] + \sqrt{2\eta T} \zeta \right], \quad \zeta \sim \text{Laplace}(0, I)$ 
4: end
5:  $\delta_t \leftarrow \delta^{K+1}$ 
6:  $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta_t), y)$ 

```

- ✓ Use MCMC techniques (e.g., Langevin Monte Carlo) [DALE → "PGD" as $T \rightarrow 0, \dots$]

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

58

Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

```

1:  $\delta^t \leftarrow \delta_{t-1}$ 
2: for  $k = 1, \dots, K$ 
3:    $\delta^{k+1} = \text{proj}_{\Delta} \left[ \delta^k + \eta \text{sign} \left[ \nabla_{\delta} \log (\text{Loss}(f_{\theta}(x + \delta^k), y)) \right] + \sqrt{2\eta T} \zeta \right], \quad \zeta \sim \text{Laplace}(0, I)$ 
4: end
5:  $\delta_t \leftarrow \delta^{K+1}$ 
6:  $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta_t), y)$ 

```

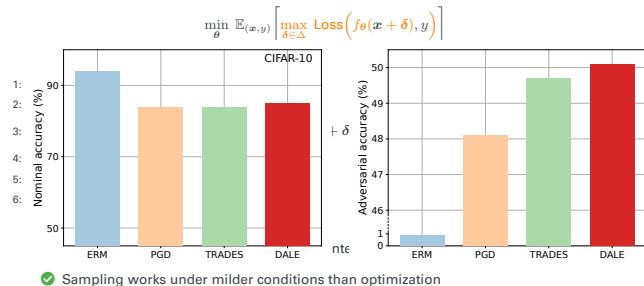
- ✓ Use MCMC techniques (e.g., Langevin Monte Carlo) [DALE → "PGD" as $T \rightarrow 0, \dots$]

- ✓ Sampling works under milder conditions than optimization

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

58

Dual Adversarial Learning



- ✓ Sampling works under milder conditions than optimization

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

58

Dual Adversarial Learning

$$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

```

1:  $\delta^t \leftarrow \delta_{t-1}$ 
2: for  $k = 1, \dots, K$ 
3:    $\delta^{k+1} = \text{proj}_{\Delta} \left[ \delta^k + \eta \text{sign} \left[ \nabla_{\delta} \log (\text{Loss}(f_{\theta}(x + \delta^k), y)) \right] + \sqrt{2\eta T} \zeta \right], \quad \zeta \sim \text{Laplace}(0, I)$ 
4: end
5:  $\delta_t \leftarrow \delta^{K+1}$ 
6:  $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \text{Loss}(f_{\theta}(x + \delta_t), y)$ 

```

- ✓ Sampling works under milder conditions than optimization

[Robey*, C., Pappas, Ribeiro, Hassani. Adversarial robustness with semi-infinite constrained learning, NeurIPS'21]

58

Summary

- Adversarial training

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\text{Loss}(f_{\theta}(x), y)] \longrightarrow \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \Delta} \text{Loss}(f_{\theta}(x + \delta), y) \right]$$

- ✓ Great(!) empirically performance

- ✓ Statistical complexity
⇒ change what robustness means (e.g., probabilistic robustness)

- ✓ Performance trade-offs
⇒ constrained learning

- ✓ Computational complexity
⇒ sampling and DALE

[Houle, C., Ribeiro. Automatic data augmentation via invariance-constrained learning, ICML'23]

59

Agenda

What is "robust learning"?

Why are models brittle?

How to learn robustly?

What else are adversaries good for?

Friendly adversaries

- Non-robustness robust properties

- Invariance [Houle, C., Ribeiro. Automatic data augmentation via invariance-constrained learning, ICML'23]

- Smoothness [Cervino, C., Haefele, Vidal, Ribeiro. Learning Globally Smooth Functions on Manifolds, ICML'23]

- ...

[Cervino, C., Haefele, Vidal, Ribeiro. Learning Globally Smooth Functions on Manifolds, ICML'23]

61

Friendly adversaries

- Non-robustness robust properties

- Invariance [Houle, C., Ribeiro. Automatic data augmentation via invariance-constrained learning, ICML'23]

$$\text{Loss}(f_{\theta}(gx), y) = \text{Loss}(f_{\theta}(x), y), \forall g \in \mathcal{G} \Rightarrow \max_{g \in \mathcal{G}} \text{Loss}(f_{\theta}(gx), y) - \text{Loss}(f_{\theta}(x), y) \leq \epsilon$$

⋮

- Smoothness [Cervino, C., Haefele, Vidal, Ribeiro. Learning Globally Smooth Functions on Manifolds, ICML'23]

- ...

61

Friendly adversaries

- Non-robustness robust properties

- Invariance** [Houle, C., Ribeiro. Automatic data augmentation via invariance-constrained learning, ICML'23]

$$\text{Loss}(f_\theta(gx), y) = \text{Loss}(f_\theta(x), y), \forall g \in \mathcal{G} \Rightarrow \max_{y \in \mathcal{Y}} \text{Loss}(f_\theta(gx), y) - \text{Loss}(f_\theta(x), y) \leq \epsilon$$

⋮

- Smoothness** [Cervino, C., Haeffele, Vidal, Ribeiro. Learning Globally Smooth Functions on Manifolds, ICML'23]

$$\begin{aligned} \min_{\theta} \quad & \max_{z \in \mathcal{M}} \|\nabla_{\mathcal{M}} f_\theta(z)\|^2 \quad (\text{Lipschitz constant}) \\ \text{subject to} \quad & \mathbb{E}_{(x,y \sim \mathcal{D})} [\text{Loss}(f_\theta(x), y)] \leq c \end{aligned}$$

- ...



61

Friendly adversaries

- Density estimation (GANs) [Goodfellow et al. Generative adversarial nets, NeurIPS'14]

$$\min_{\theta} \max_d \mathbb{E}_{x \sim \mathcal{D}} [\log(d(x))] + E_{z \sim \mathcal{M}} [\log(1 - d(f_\theta(z)))]$$

[$d : x \leftarrow 1$ and $f_\theta(z) \leftarrow 0$]



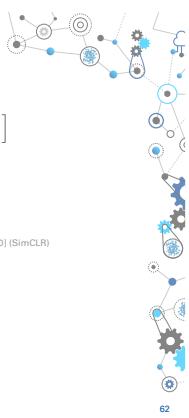
62

Friendly adversaries

- Density estimation (GANs) [Goodfellow et al. Generative adversarial nets, NeurIPS'14]

$$\min_{\theta} \max_d \mathbb{E}_{x \sim \mathcal{D}} [\log(d(x))] + E_{z \sim \mathcal{M}} [\log(1 - d(f_\theta(z)))]$$

[$d : x \leftarrow 1$ and $f_\theta(z) \leftarrow 0$]



62

Concluding remarks

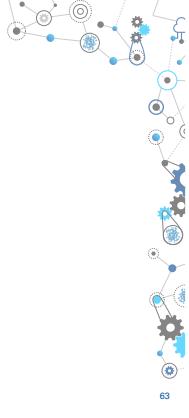
- Semi-supervised and unsupervised learning
- ...
- ...



63

Concluding remarks

- robust learning = strong against input disturbances
Important, yes... but essential?



63

Concluding remarks

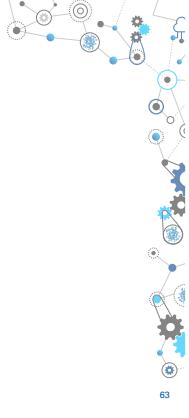
- robust learning = strong against input disturbances
Important, yes... but essential?
- Is adversarial training THE solution?
Not when it matters...



63

Concluding remarks

- robust learning = strong against input disturbances
Important, yes... but essential?
- Is adversarial training THE solution?
Not when it matters...
- So are we done with adversarial training?
Absolutely not!



63

