Adversarial Learning Bounds for Linear Classes and Neural Nets

Understanding Adversarial Learning through Rademacher Complexity

Pranjal Awasthi, Natalie Frank, Mehryar Mohri

Google Research & Courant Institute

August 14, 2020

Adversarial Attacks

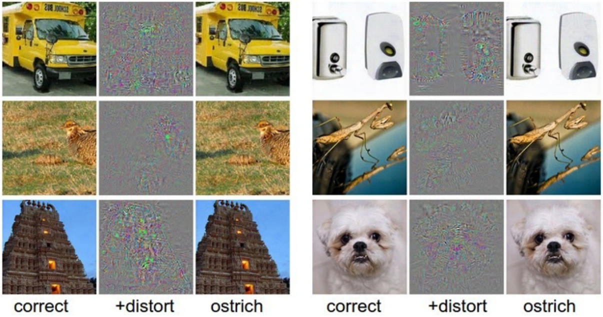


Figure: Imperceptible adversarial perturbations in ℓ_2. [5]

Adversarial Robustness

Figure: A sparse perturbation. [1]

Overarching Goal: Train classifiers robust to adversarial perturbations.

- Examples in many areas of applications.
- Different possible forms of perturbations: changing every pixel in an image vs. placing a sticker on a stop sign.
- Can we derive learning guarantees for adversarial robustness?

Outline of Talk

Goal of our paper: Understand what characterizes robust generalization and how it relates to non-robust generalization

1. Classification & Adversarial Classification setup
2. Rademacher complexity & Adversarial Rademacher Complexity
3. Better bounds for adversarial Rademacher complexity of linear classes
4. Better bounds for Rademacher complexity of linear classes
5. Adversarial Rademacher complexity of neural nets

Standard Classification Setting

Binary Classification: Data distributed over $\mathbb{R}^d \times \{-1, +1\}$ according to $\mathcal{D}$

Standard Setting:

- Given a predictor $h : \mathbb{R}^d \to \mathbb{R}$, a point $\mathbf{x}$ is classified as $\text{sign}(h(\mathbf{x}))$.
- There is an error if $yh(\mathbf{x}) < 0$, or $\mathbf{1}_{yh(\mathbf{x})<0} = 1$.
- The *classification error* is then

$$R(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{1}_{yh(\mathbf{x})<0}]$$

Defining Adversarial Perturbations

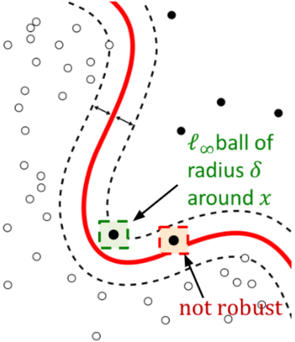
Adversarial Setting:

- The data is perturbed by ϵ in ℓ_p to "fool" the classifier into thinking there is an error, now an error occurs if

$$1 = \sup_{\|\mathbf{x}-\mathbf{x}'\|_r \leq \epsilon} \mathbf{1}_{yh(\mathbf{x}')<0} = \mathbf{1}_{\inf_{\|\mathbf{x}-\mathbf{x}'\|_r \leq \epsilon} yh(\mathbf{x}')<0}$$

- The *adversarial classification error* is then

$$\widetilde{R}(h) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{1}_{\inf_{\|\mathbf{x}-\mathbf{x}'\|_r \leq \epsilon} yh(\mathbf{x}')<0}]$$



Rademacher Complexity

The *empirical Rademacher complexity is*

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\mathbf{z}_i)\right]$$

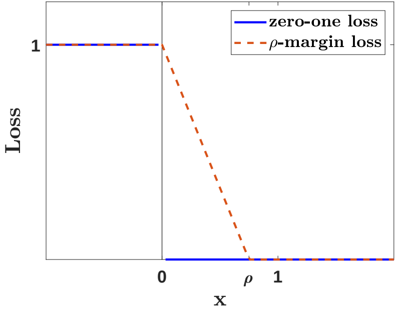
ρ-**Margin Loss:**

$$\Phi_\rho(u) = \min(1, \max(0, 1 - \frac{u}{\rho}))$$

Theorem (Margin Bounds [4])

$$R(h) \leq \widehat{R}_{\mathcal{S},\rho}(h) + \frac{2}{\rho}\mathfrak{R}_{\mathcal{S}}(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

holds with probability at least $1-\delta$ *for all* $h \in \mathcal{F}$.



Adversarial Rademacher Complexity

Theorem (Robust margin bounds)

Define the class $\widetilde{\mathcal{F}}$ *by*

$$\widetilde{\mathcal{F}} = \left\{(\mathbf{x}, y) \mapsto \inf_{\|\mathbf{x}-\mathbf{x}'\|_r \leq \epsilon} yf(\mathbf{x}') \colon f \in \mathcal{F}\right\}.$$

The following holds with probability at least $1-\delta$ *for all* $h \in \mathcal{F}$*:*

$$\widetilde{R}(h) \leq \widetilde{R}_{\mathcal{S},\rho}(h) + \frac{2}{\rho}\mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Definition

We define the *adversarial Rademacher Complexity* as

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) := \mathfrak{R}_{\mathcal{S}}(\widetilde{\mathcal{F}})$$

Prior Work on Adversarial Rademacher Complexity of Linear Classes

$$\mathcal{F}_p = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle \colon \|\mathbf{w}\|_p \leq W\}$$

Yin et. al. [6]: For perturbations in the infinity norm, for some constant c

$$\max(\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p), c\epsilon W \frac{d^{\frac{1}{p^*}}}{\sqrt{m}}) \leq \widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_p) \leq \mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) + \epsilon W \frac{d^{\frac{1}{p^*}}}{\sqrt{m}}$$

Khim and Loh [3]: For perturbation in the r-norm, there exists a constant M_r for which

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_2) \leq \frac{W}{\sqrt{m}} \max_{(\mathbf{x}_i, y_i) \in \mathcal{S}} \|\mathbf{x}_i\|_2 + \epsilon \frac{M_{r^*}}{2\sqrt{m}}$$

Adversarial Rademacher Complexity of Linear Classes

$$\mathcal{F}_p = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle \colon \|\mathbf{w}\|_p \leq W\}$$

Theorem

Let $\epsilon > 0$, $r \geq 1$. Consider a sample $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$ and perturbations in the r-norm. Then

$$\begin{aligned}
\max\left(\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p), \epsilon \frac{W \max(d^{1-\frac{1}{r}-\frac{1}{p}}, 1)}{2\sqrt{2m}}\right) &\leq \widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}_p) \\
&\leq \mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) + \epsilon \frac{W}{2\sqrt{m}} \max(d^{1-\frac{1}{r}-\frac{1}{p}}, 1)
\end{aligned}$$

Rademacher Complexity of Linear Classes

$$\mathcal{F}_p = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_p \leq W\}$$

$$\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_m]$$

Group norms: $\|\mathbf{A}\|_{p,q} = \|(\|\mathbf{A}_1\|_p \cdots \|\mathbf{A}_m\|_p)\|_q$ where $\mathbf{A}_i$ is the ith row of $\mathbf{A}$.

Prior Work [2]:

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) \leq \begin{cases} W\sqrt{\frac{2\log(2d)}{m}}\|\mathbf{X}\|_{\max} & \text{if } p = 1 \\ \frac{W}{m}\sqrt{p^* - 1}\|\mathbf{X}\|_{p^*,2} & \text{if } 1 < p \leq 2 \end{cases}$$

Our new bounds:

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) \leq \begin{cases} \frac{W}{m}\sqrt{2\log(2d)}\|\mathbf{X}^T\|_{2,p^*} & \text{if } p = 1 \\ \frac{\sqrt{2}W}{m}\left[\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right]^{\frac{1}{p^*}}\|\mathbf{X}^T\|_{2,p^*} & \text{if } 1 < p \leq 2 \\ \frac{W}{m}\|\mathbf{X}^T\|_{2,p^*} & \text{if } p \geq 2 \end{cases}$$

Comparing the Bounds for $1 < p \leq 2$

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}_p) \leq \begin{cases} \frac{W}{m}\sqrt{p^* - 1}\|\mathbf{X}\|_{p^*,2} & \text{old bound} \\ \frac{\sqrt{2}W}{m}\left[\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right]^{\frac{1}{p^*}}\|\mathbf{X}^T\|_{2,p^*} & \text{new bound} \end{cases}$$

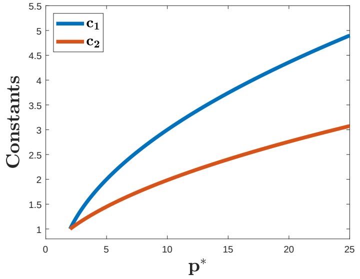
Comparing the Norms: If $p \leq 2$, then

$$\min(m,d)^{\frac{1}{2}-\frac{1}{p^*}}\|\mathbf{X}^T\|_{2,p^*} \geq \|\mathbf{X}\|_{p^*,2} \geq \|\mathbf{X}^T\|_{2,p^*}$$

Comparing the Constants:

$$c_1(p) = \sqrt{p^* - 1}$$

$$c_2(p) = \sqrt{2}\left[\frac{\Gamma(\frac{p^*+1}{2})}{\sqrt{\pi}}\right]^{\frac{1}{p^*}}$$



Adversarial Rademacher Complexity of the ReLU

$$\mathcal{G}_p = \{(\mathbf{x}, y) \mapsto (y\langle \mathbf{w}, \mathbf{x}\rangle)_+ \colon \|\mathbf{w}\|_p \leq W, y \in \{-1, 1\}\}$$
$$\mathcal{F}_p = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle \colon \|\mathbf{w}\|_p \leq W\}$$

Theorem

The adversarial Rademacher complexity of $\mathcal{G}_p$ can be bounded as follows:

$$\frac{W\delta\epsilon}{2\sqrt{2}m}|T^{\delta}_{\epsilon,\mathbf{s}^*}| \max(d^{1-\frac{1}{p}-\frac{1}{r}}, 1) \leq \widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p)$$
$$\leq \mathfrak{R}_{T_\epsilon}(\mathcal{F}_p) + \epsilon\frac{W}{2\sqrt{m}} \max(1, d^{1-\frac{1}{r}-\frac{1}{p}}),$$

where

$$T_\epsilon = \{i \colon y_i = -1 \text{ or }, y_i = 1 \text{ and } \|\mathbf{x}_i\|_r > \epsilon\}$$
$$T^{\delta}_{\epsilon,\mathbf{s}} = \{i \colon \langle \mathbf{s}, \mathbf{x}_i\rangle - (1 + \delta y_i) y_i \epsilon \|\mathbf{s}\|_{r^*} > 0\}$$

and $\mathbf{s}^*$ is the adversarial perturbation.

Adversarial Rademacher Complexity of Neural Nets

$$\mathcal{G}_p^n = \left\{(\mathbf{x}, y) \mapsto y \sum_{j=1}^{n} u_j \rho(\mathbf{w}_j \cdot \mathbf{x})\colon \|\mathbf{u}\|_1 \le \Lambda, \|\mathbf{w}_j\|_p \le W\right\}.$$

Theorem

Let ρ be a function with Lipschitz constant L_ρ with $\rho(0) = 0$. Then, the following upper bound holds for the adversarial Rademacher complexity of $\mathcal{G}_p^n$:

$$\widetilde{\mathfrak{R}}_{\mathcal{S}}(\mathcal{G}_p^n) \le L_\rho \left[\frac{W\Lambda \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|\mathbf{X}\|_{r,\infty} + \epsilon)}{\sqrt{m}}\right] \times \left(1 + \sqrt{d(n+1)\log(36)}\right).$$

Towards Dimension Independent Bounds

- Studying the structure of adversarial perturbations leads to equations qualitatively similar to γ-fat shattering.
- Under appropriate assumptions, this can lead to dimension independent bounds.

Conclusion

We covered

- New bounds for Rademacher complexity of linear classes.
- New bounds for adversarial Rademacher complexity of linear classes.
- New bounds for adversarial Rademacher complexity of Neural nets.

Open problems

- Generalize to arbitrary norms: in general is the dual norm a good regularizer?
- Improve the adversarial neural nets generalization bound or find a matching lower bound.

Bibliography

[1] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, 2017.

[2] Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of NIPS*, pages 793–800, 2008.

[3] Justin Khim and Po-Ling Loh. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

[4] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, second edition, 2018.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of ICLR*, 2014.

[6] Dong Yin, Kannan Ramchandran, and Peter L. Bartlett. Rademacher complexity for adversarially robust generalization. In *Proceedings of ICML*, pages 7085–7094, 2019.