Sharp Statistical Guarantees for Adversarially Robust Gaussian Classification

Chen Dan

Yuting Wei

Pradeep Ravikumar

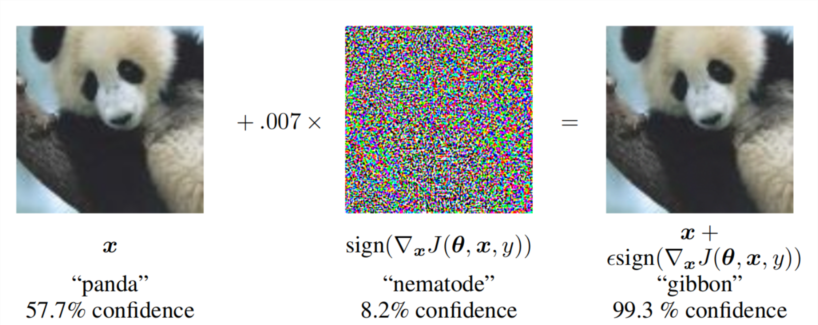
ICML 2020

Computer Science Department, Statistics Department, Machine Learning Department
Carnegie Mellon University

Outline

- Basics of Adversarial Robustness
- Prior works + Motivation of this work
- Main Results
- Proof Sketch

Adversarial Examples



(Goodfellow et al. 2014)

Deep Neural Networks are vulnerable to adversarial attacks.

Mathematical Formulation

Standard Classification:

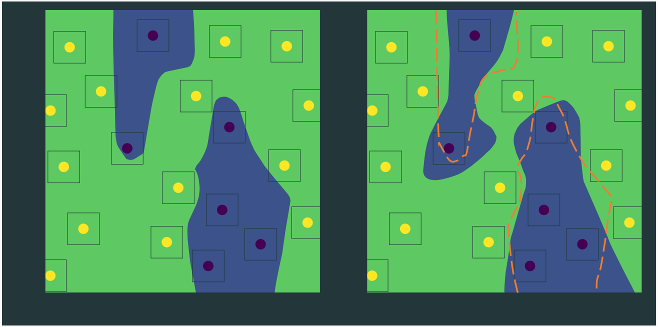
$$\min_{f} \mathbb{E} L\left(f(x), y\right)$$

Robust Classification: A 2-player game, defender and attacker:

$$\min_{f} \mathbb{E} \max_{\delta \in \Delta} L(f(x+\delta), y)$$

where Δ is a perturbation set, e.g. ℓ_∞ ball.

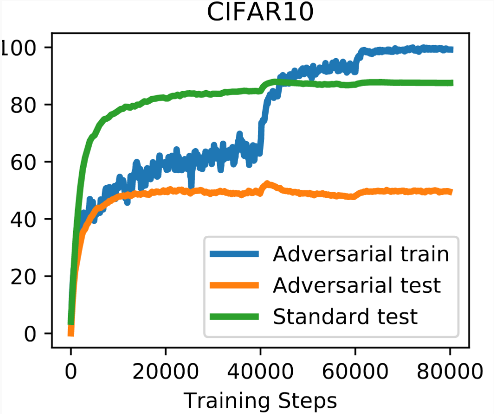
A 2D Example



Challenges

- Optimization
- Evaluation
- Generalization (This Work)

Generalization in Adversarial Setting



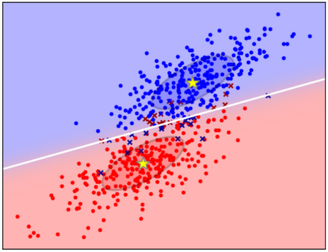
(Schmidt et al. NeurIPS'18) The generalization gap in Adv-Robust Classification is significantly larger than Standard Classification.

Conditional Gaussian Model

"Adversarially Robust Generalization Requires More Data", (Schmidt et al. NeurIPS'18)[1]

Binary Classification with Conditional Gaussian Model $P_{\mu,\Sigma}$:

$$p(y=1) = p(y=-1) = \frac{1}{2},$$
$$x|y=+1 \sim N(+\mu, \Sigma),$$
$$x|y=-1 \sim N(-\mu, \Sigma),$$
$$\mu \in \mathbb{R}^d, \quad \Sigma \in \mathbb{R}^{d\times d}.$$



Minimize Robust Classification Error:

$$R_{\text{robust}}(f) = \Pr[\exists x' : \|x' - x\|_B \le \varepsilon, f(x') \neq y]$$

where $\|\cdot\|_B$ is a norm, e.g. ℓ_p norm.

[1] arXiv:1804.11285

Sample Complexity

Theorem (Schmidt et al. NeurIPS'18)

When $\Sigma = \sigma^2 I, \|\mu\|_2 = \sqrt{d}, \sigma \leq \frac{1}{32} d^{1/4}$,

adversarial perturbation $\|x' - x\|_\infty \leq \frac{1}{4}$.

- $O(1)$ *samples sufficient for* 1% *standard classification error.*
- $\tilde{\Omega}(\sqrt{d})$ *samples necessary for* 49% *robust classification error.*

- What's the statistical rate of convergence?
- What happens in other regimes?
- Why do we need more data?

Prior works: Uniform Convergence

e.g. (Kim and Loh, 2018), (Yin, Ramchandran, Bartlett, 2019), (Awasthi, Frank, Mohri, 2020).

$$\sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \leq \tilde{O}(\sqrt{\frac{C(F)}{n}})$$

Assuming optimal classifier in $\mathcal{F}$: $O(\sqrt{\frac{1}{n}})$ convergence.

Independent work of (Dobriban et al. 2020)[2]: same setting with ours, $O(\sqrt{\frac{1}{n}})$ rate.

Can we get an $O(\frac{1}{n})$ fast rate?

[2] arXiv:2006.05161

Contributions

- Understanding the sample complexity through the lens of Statistical Minimax Theory.
- Introducing "Adversarial Signal-to-Noise Ratio", which helps explaining why robust classification requires more data.
- Near-optimal upper and lower bounds on minimax risk, with minimal assumptions.
- First $O(\frac{1}{n})$ "fast rate" in robust classification!

Minimax Theory

Goal: characterize the Statistical Minimax Error of robust Gaussian classification:

$$\min_{\widehat{f}} \max_{P_{\mu,\Sigma} \in D} [R_{\text{robust}}(\widehat{f}) - R^*_{\text{robust}}]$$

where:

- R^*_{robust} is the smallest classification error of any classifier.
- D is a class of distributions.
- $\hat{f}$ is any estimator based on n i.i.d samples $\{x_i, y_i\}_{i=1}^n \sim P_{\mu,\Sigma}$.

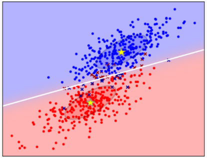
Fisher's LDA

Recall:

$$R_{\text{robust}}(f) = \Pr[\exists \|x' - x\|_B \le \varepsilon, f(x') \ne y]$$

When $\varepsilon = 0$, the problem reduces to Fisher's LDA.

The smallest classification error R^* is $\bar{\Phi}(\frac{1}{2}SNR)$, where:



- *SNR* is the *Signal-to-Noise Ratio* of the model:

$$SNR(P_{\mu,\Sigma}) = 2\sqrt{\mu^T \Sigma^{-1} \mu}.$$

- $\bar{\Phi}$: Gaussian tail probability $\bar{\Phi}(c) = \Pr_{X \sim N(0,1)}[X > c]$.

SNR measures the hardness of classification.

Minimax Rate of Fisher's LDA

Consider the family of distributions with a fixed SNR:

$$D_{\mathrm{std}}(r) := \{P_{\mu,\Sigma} | SNR(P_{\mu,\Sigma}) = 2\sqrt{\mu^T \Sigma^{-1} \mu} = r\}.$$

Theorem (Li et al. AISTATS'17[3])

$$\min_{\widehat{f}} \max_{P \in D_{\mathrm{std}}(r)} [R(\widehat{f}) - R^*] \geq \Omega\left(e^{-(\frac{1}{8}+o(1))r^2} \cdot \frac{d}{n}\right).$$

with a nearly-matching upper bound.

To achieve $(R^* + \varepsilon)$ error, we need $\frac{d}{\varepsilon} e^{-(\frac{1}{8}+o(1))r^2}$ samples.

Large SNR $\rightarrow$ fewer samples.

[3]Link to the full paper

Proof Sketch: Upper Bound 1

Assume $\Sigma = I$ for simplicity.

For any linear classifier $f_w(x) = \mathrm{sign}(w^T x)$, the classification error is

$$R(f_w) = \bar{\Phi}(\frac{w^T \mu}{\|w\|_2})$$

Recall the Bayes Classifier: $f_{Bayes}(x) = \mathrm{sign}(\mu^T x)$, hence

$$R(f_w) - R^* = \bar{\Phi}(\frac{w^T \mu}{\|w\|_2}) - \bar{\Phi}(\|\mu\|_2)$$

Proof Sketch: Upper Bound 2

Taylor expansion of $\bar{\Phi}(\cdot)$:

$$\begin{aligned} R(f_w) - R^* &= \bar{\Phi}\left(\frac{w^T\mu}{\|w\|_2}\right) - \bar{\Phi}(\|\mu\|_2) \\ &\approx \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\|\mu\|_2^2\right)\left(\|\mu\|_2 - \frac{w^T\mu}{\|w\|_2}\right) \end{aligned}$$

Choose $w = \widehat{\mu}$, suffices to bound:

$$\delta_n = \|\mu\|_2 - \frac{\widehat{\mu}^T\mu}{\|\widehat{\mu}\|_2}$$

Proof Sketch: Upper Bound 3

$$
\begin{aligned}
\delta_n &= \|\mu\|_2 - \frac{\widehat{\mu}^T \mu}{\|\widehat{\mu}\|_2} \\
&= \frac{1}{\|\widehat{\mu}\|_2} \left(-\frac{1}{2} \|\widehat{\mu} - \mu\|_2^2 + (\|\widehat{\mu}\|_2 - \|\mu\|_2)^2 \right) \\
&= O(\frac{d}{n})
\end{aligned}
$$

Proof Sketch: Lower Bound

- Overall, similar to Linear Regression: Fano + Gilbert-Varshamov.
- Main difficulty: $R(f) - R^*$ does not satisfy triangle inequality.
- Approach 1 (Li et al. AISTATS 17): approximate version of triangle inequality using 2D Gaussian integrals.
- Approach 2 (Cai, Zhang JRSSB 19): Find another loss function L, which satisfies Triangle inequality and

$$R(f) - R^* \geq C \cdot L(f)^2$$

Then show that $L(f) \geq C' \sqrt{\frac{d}{n}}$.
Our proof technique is inspired by Approach 2.

What goes wrong in the robust setting?

Signal-to-Noise Ratio exactly characterizes the hardness of standard Gaussian classification problem.

Can we find a similar quantity for the robust setting?

SNR is not the correct answer!

Counterexample: exists two distributions P, P'

Distribution	SNR	R^*	R^*_{robust}
P	6	10^{-8}	10^{-8}
P'	6	10^{-8}	50%

Adversarial Signal-to-Noise Ratio

We define **Adversarial Signal-to-Noise Ratio(AdvSNR)** as:

$$AdvSNR(P_{\mu,\Sigma}) = \min_{\|z\|_B \le \varepsilon} SNR(P_{\mu-z,\Sigma}).$$

Using *AdvSNR*, we can re-formulate one of the main theorems in (Bhagoji et al. ,NeurIPS 2019)[4] as:

$$R^*_{\text{robust}} = \bar{\Phi}(\frac{1}{2} AdvSNR).$$

i.e. *AdvSNR* can be used to measure the hardness of robust classification.

[4] arXiv:1909.12272

Main Result

Consider the family of distributions with a fixed AdvSNR:

$$D_{\text{robust}}(r) := \{P_{\mu,\Sigma} | AdvSNR(P_{\mu,\Sigma}) = r\}.$$

Theorem (Dan, Wei, Ravikumar, ICML'20)

$$\min_{\widehat{f}} \max_{P \in D_{robust}(r)} [R_{robust}(\widehat{f}) - R^*_{robust}] \geq \Omega\left(e^{-(\frac{1}{8}+o(1))r^2} \cdot \frac{d}{n}\right).$$

and there is a computationally efficient estimator which achieves this minimax rate!

Generalization of (Li et al. 2017) in adversarially robust setting - almost assumed nothing about covariance, norm of adversary, etc. !

Why does Adv-Robust Classification Require More Data?

The minimax rates for Standard vs. Adv-Robust classification:

$$\exp\{-\frac{1}{8}SNR^2\}\frac{d}{n} \quad \text{vs.} \quad \exp\{-\frac{1}{8}AdvSNR^2\}\frac{d}{n}$$

$AdvSNR \leq SNR \Rightarrow$ Adv-Robust Risk always converges slower.

Examples:

AdvSNR	SNR	#times slower
$\Theta(1)$	$\Theta(1)$	constant
$\Theta(1)$	$\Theta(\sqrt{logd})$	$poly(d)$
$\Theta(1)$	$\Omega(\sqrt{d})$	$exp(d)$

Upper Bound & Algorithm

- (Bhagoji et al. ,NeurIPS 2019)[5]: $f(x) = \mathrm{sign}(w_0^T x)$ has the minimal robust classification error, where

$$w_0 = \Sigma^{-1}(\mu - z_0),$$
$$z_0 = \underset{\|z\|_B \le \varepsilon}{\mathrm{argmin}} (\mu - z)^T \Sigma^{-1} (\mu - z).$$

- Replace (μ, Σ) by their empirical counterpart $(\widehat{\mu}, \widehat{\Sigma})$.
- Now we have an efficient algorithm that achieves the minimax rate!

[5] arXiv:1909.12272

Upper Bound & Algorithm

- Our Algorithm: $\widehat{f}(x) = \mathrm{sign}(\widehat{w_0}^T x)$ achieves minimax excess risk, where

$$\widehat{w_0} = \widehat{\Sigma}^{-1}(\widehat{\mu} - \widehat{z_0}),$$
$$\widehat{z_0} = \underset{\|z\|_B \leq \varepsilon}{\mathrm{argmin}}(\widehat{\mu} - z)^T \widehat{\Sigma}^{-1}(\widehat{\mu} - z).$$

- Replace (μ, Σ) by their empirical counterpart $(\widehat{\mu}, \widehat{\Sigma})$.
- Now we have an efficient algorithm that achieves the minimax rate!
- Proof is similar to Fisher LDA from high level, but requires a more careful decomposition of the loss (Lemma 6.3 in paper).

Lower Bound

- Main idea: Black-Box Reduction
    - Robust Classification is "harder" than Standard Classification.
    - For any distribution P with Signal-to-Noise Ratio r,
    - We can find a P' with $AdvSNR$ r, such that for any classifier f,

$$RobustExcessRisk_{P'}(f) \geq StdExcessRisk_P(f)$$

- Take $\min_f \max_{P \in D_{std}(r)}$,

$$\begin{aligned} &MinimaxRobustExcessRisk(D_{robust}(r)) \\ \geq &MinimaxStdExcessRisk(D_{std}(r)). \end{aligned}$$

- Apply (Li et al. 2017) and we get the minimax lower bound.

Summary

- We provided the first statistical minimax optimality result for Adversarially Robust Classification.
- We introduced AdvSNR, which characterizes the hardness of Adv-Robust Gaussian Classification.
- We proved matching upper and lower bounds for minimax excess risk, and proposed an efficient, minimax-optimal algorithm.
- Adversarially Robust Classification requires More Data, because adversarial perturbation decreases the Signal-to-Noise Ratio!

Acknowledgements

Kaizheng Wang
PhD @ Princeton $\rightarrow$
Asst. Prof @
Columbia

Tianle Cai
Undergrad @ PKU $\rightarrow$
PhD @ Princeton

Justin Khim
Postdoc @ CMU