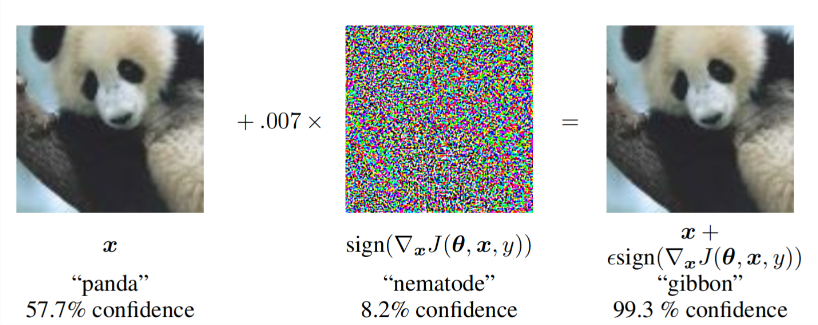
Optimal Statistical Guarantees for Adversarially Robust Gaussian Classification

Chen Dan, Yuting Wei, Pradeep Ravikumar

ICML 2020

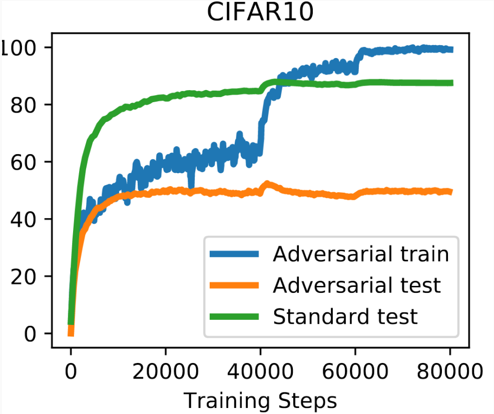
Computer Science Department, Statistics Department, Machine Learning Department
Carnegie Mellon University

Adversarial Example



Deep Neural Networks are vulnerable to adversarial attacks.

Statistical Challenges



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

Conditional Gaussian Model

(Mixture of two gaussians picture here)

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

$$p(y=1) = p(y=-1) = \frac{1}{2},$$
$$x|y=+1 \sim N(+\mu, \Sigma),$$
$$x|y=-1 \sim N(-\mu, \Sigma).$$

Minimize Robust Classification Error:

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

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

Sample Complexity

"Adversarially Robust Generalization Requires More Data":

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* 99% *standard accuracy.*
- $\tilde{\Omega}(\sqrt{d})$ *samples necessary for* 51% *robust accuracy.*

- Why do we need more data?
- What happens in other regimes?

Contributions

- Understanding the sample complexity through the lens of Statistical Minimax Theory.
- Introducing "Adversarial Signal-to-Noise Ratio", which explains why robust classification requires more data.
- Near-optimal upper and lower bounds on minimax risk.
- ** Computationally efficient minimax-optimal estimator.
- ** Minimal assumptions.

Minimax Theory

Our goal is to 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:

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

Fisher's LDA: Bayes Risk

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

The smallest possible 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 characterizes the hardness of classification problem.

Minimax Rate of Fisher LDA

Consider the family of distributions with a fixed SNR:

$$D_{\mathrm{std}}(r) := \{P_{\mu,\Sigma} | SNR(P_{\mu,\Sigma}) = r\}.$$

The following minimax rate is proved by prior works:

Theorem (Li et al. AISTATS'17)

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

with a nearly-matching upper bound.

Signal-to-Noise Ratio

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!
- Two distributions with same *SNR* can have very different optimal robust classification error (e.g. 0.1% vs 50%)!

Adversarial Signal-to-Noise Ratio

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

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

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

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

which recovers the results in Fisher LDA when $\varepsilon = 0$!

Main Result

Consider the family of distributions with a fixed AdvSNR:

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

Our Main Theorem:

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!

Why does Adv-Robust Classification Require More Data?

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

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

- $AdvSNR \leq SNR$, so Adv-Robust Risk always converges slower.
- Sometimes $AdvSNR = \Theta(1)$ and $SNR = \Theta(1)$, the convergence is only a constant factor slower.
- Sometimes $AdvSNR = \Theta(1)$ and $SNR = \Theta(d)$, the convergence is $\exp(\Omega(d))$ times slower!

Upper Bound & Algorithm

- (Bhagoji et al. ,NeurIPS 2019) showed that a linear classifier $f(x) = \mathrm{sign}(w_0^T x)$ 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 you have an efficient algorithm that achieves the minimax rate!

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

- In this paper, we provide 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 an efficient, minimax-optimal algorithm.
- Adversarially Robust Classification requires More Data, because adversarial perturbation decreases the Signal-to-Noise Ratio!