On the statistical foundations of adversarially robust learning

Edgar Dobriban
Wharton, UPenn
joint work with Hamed Hassani, David Hong, and Alex Robey
slides at: github.com/dobriban/talks

October 27, 2020

Overview

Overview

Statistics in the present era

- Our age is characterized by a fast-growing interest in leveraging data
- Data Science, Analytics, AI, Machine Learning, Big Data ... Statistics?
- Dean Erika James:

  "Analytics and AI will be central to business going forward. Need to make sure we are world leaders. Business Analytics is the fastest growing concentation."

- Wharton initiatives: Analytics at Wharton ($5M), AI at Wharton ($5M)
- Penn initiatives: SEAS Data Science ($25M), AI Major
- National initiatives:
  1. National AI R&D Plan: "AI as an Administration Priority"
  2. Nat'l AI Institutes ($250M); NSF-Simons Collab ($20M); ...
- Statistics is key to all this

Statistics in the present era

- Our age is characterized by a fast-growing interest in leveraging data
- Data Science, Analytics, AI, Machine Learning, Big Data ... Statistics?
- Dean Erika James:

  "Analytics and AI will be central to business going forward. Need to make sure we are world leaders. Business Analytics is the fastest growing concentation."

- Wharton initiatives: Analytics at Wharton ($5M), AI at Wharton ($5M)
- Penn initiatives: SEAS Data Science ($25M), AI Major
- National initiatives:
  1. National AI R&D Plan: "AI as an Administration Priority"
  2. Nat'l AI Institutes ($250M); NSF-Simons Collab ($20M); ...
- Statistics is key to all this

Statistics in the present era

- Our age is characterized by a fast-growing interest in leveraging data
- Data Science, Analytics, AI, Machine Learning, Big Data ... Statistics?
- Dean Erika James:

  "Analytics and AI will be central to business going forward. Need to make sure we are world leaders. Business Analytics is the fastest growing concentation."

- Wharton initiatives: Analytics at Wharton ($5M), AI at Wharton ($5M)
- Penn initiatives: SEAS Data Science ($25M), AI Major
- National initiatives:
  1. National AI R&D Plan: "AI as an Administration Priority"
  2. Nat'l AI Institutes ($250M); NSF-Simons Collab ($20M); ...
- Statistics is key to all this

Statistics in the present era

- Our age is characterized by a fast-growing interest in leveraging data
- Data Science, Analytics, AI, Machine Learning, Big Data ... Statistics?
- Dean Erika James:

  "Analytics and AI will be central to business going forward. Need to make sure we are world leaders. Business Analytics is the fastest growing concentation."

- Wharton initiatives: Analytics at Wharton ($5M), AI at Wharton ($5M)
- Penn initiatives: SEAS Data Science ($25M), AI Major
- National initiatives:
  1. National AI R&D Plan: "AI as an Administration Priority"
  2. Nat'l AI Institutes ($250M); NSF-Simons Collab ($20M); ...
- Statistics is key to all this

Statistics in the present era

- Our age is characterized by a fast-growing interest in leveraging data
- Data Science, Analytics, AI, Machine Learning, Big Data ... Statistics?
- Dean Erika James:

  "Analytics and AI will be central to business going forward. Need to make sure we are world leaders. Business Analytics is the fastest growing concentation."

- Wharton initiatives: Analytics at Wharton ($5M), AI at Wharton ($5M)
- Penn initiatives: SEAS Data Science ($25M), AI Major
- National initiatives:
  1. National AI R&D Plan: "AI as an Administration Priority"
  2. Nat'l AI Institutes ($250M); NSF-Simons Collab ($20M); ...
- Statistics is key to all this

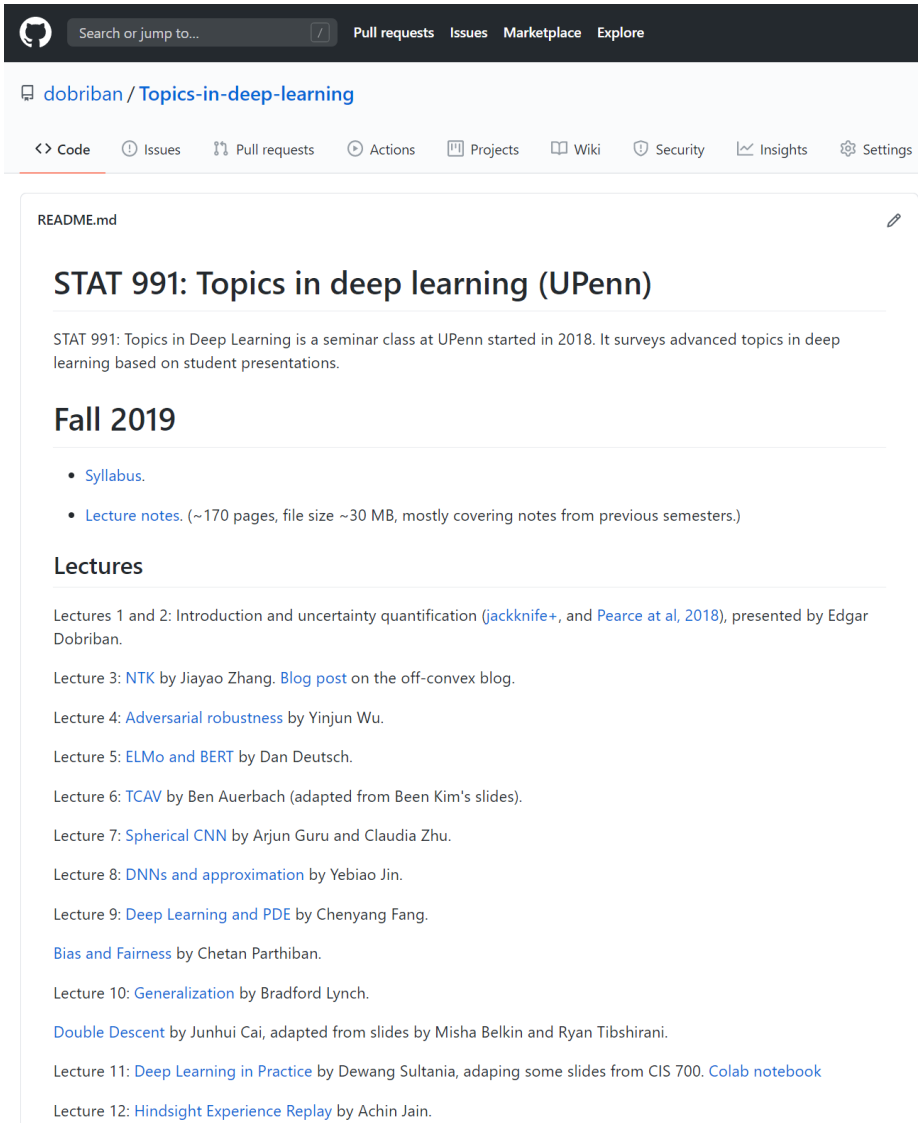
Statistics in the present era

- Our age is characterized by a fast-growing interest in leveraging data
- Data Science, Analytics, AI, Machine Learning, Big Data ... Statistics?
- Dean Erika James:

  "Analytics and AI will be central to business going forward. Need to make sure we are world leaders. Business Analytics is the fastest growing concentation."

- Wharton initiatives: Analytics at Wharton ($5M), AI at Wharton ($5M)
- Penn initiatives: SEAS Data Science ($25M), AI Major
- National initiatives:
  1. National AI R&D Plan: "AI as an Administration Priority"
  2. Nat'l AI Institutes ($250M); NSF-Simons Collab ($20M); ...
- Statistics is key to all this

Some activities



Notes for Stat 991: Topics in Deep Learning

Edgar Dobriban*

January 28, 2020

Abstract

Deep learning has achieved many empirical successes, and has attracted considerable attention. However, it continues to be poorly understood. This advanced seminar course will explore several topics in deep learning. We will discuss both theory and applications.

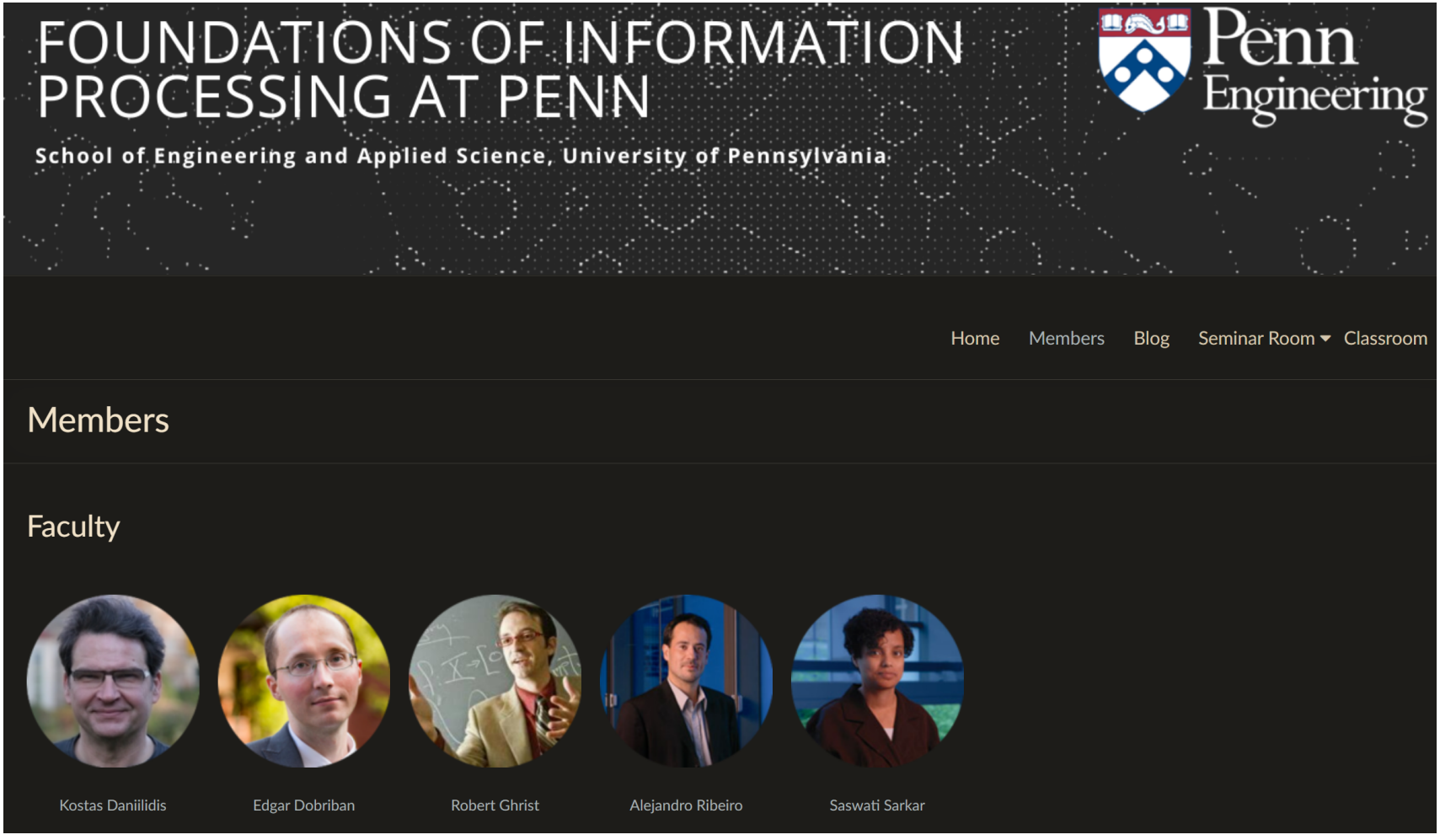
Contents

*Wharton Statistics Department, University of Pennsylvania. dobriban@wharton.upenn.edu. These notes draw inspiration from many sources, including David Donoho's course Stat 385 at Stanford, Andrew Ng's Deep Learning course on deeplearning.ai, David Silver's RL course, Tony Cai's reading group at Wharton. Disclaimer: the notes may contain factual and typographical errors. Thanks to several people who have provided parts of the notes, including Zongyu Dai, Georgios Kissas, Jane Lee, Barry Plunkett, Matteo Sordello, Yibo Yang, Bo Zhang, Yi Zhang, Carolina Zheng. The images included in this note are subject to copyright by their rightful owners, and are included here for educational purposes.

Figure: 170pg. notes

Some activities: GPU Machines

Some activities



Some activities

Friday, September 4, 2020

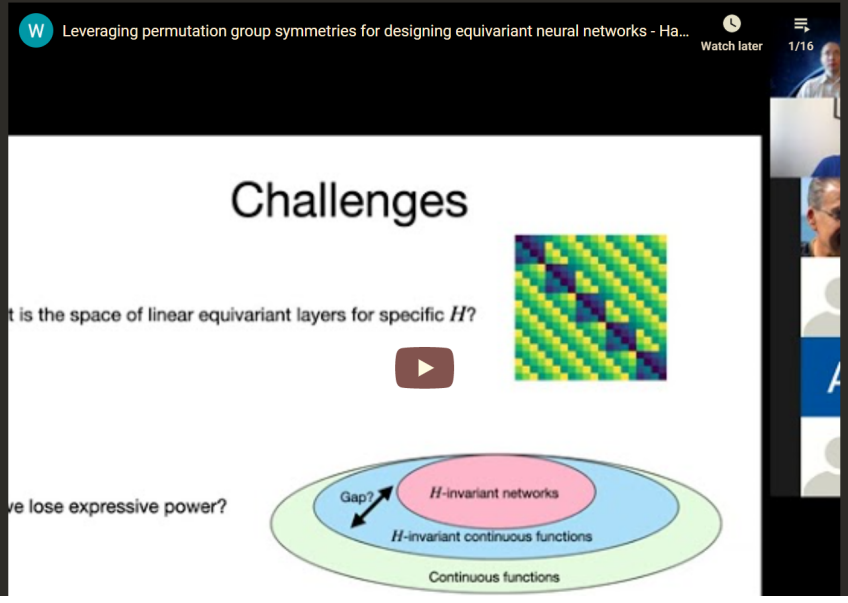
Workshop on Equivariance and Data Augmentation

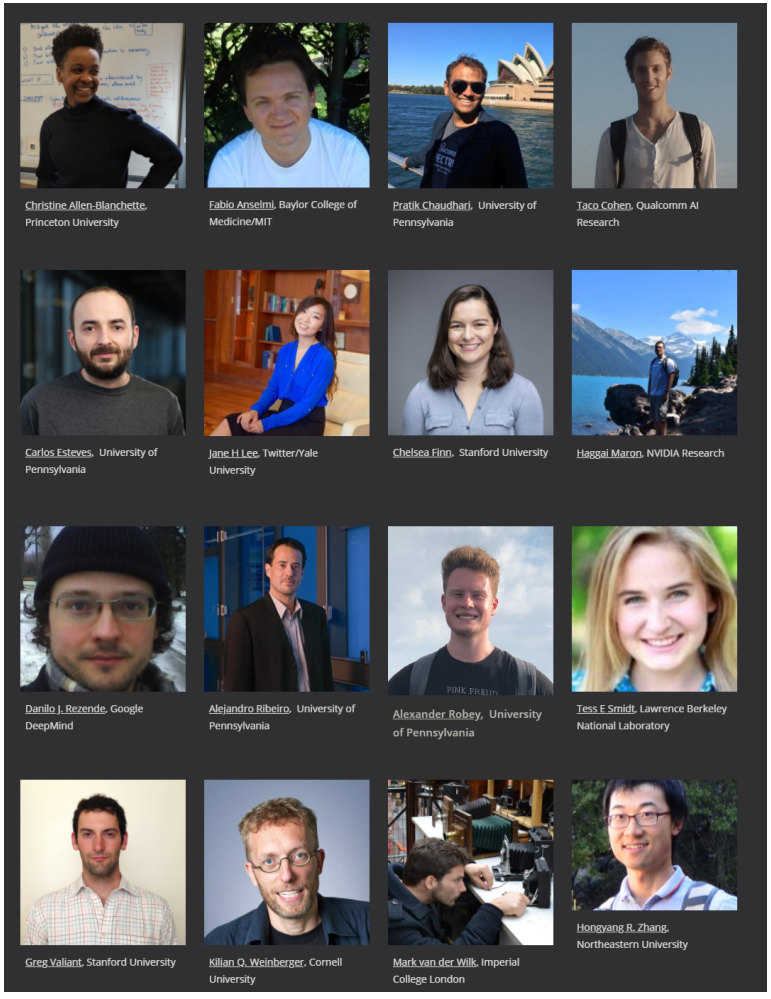
online, hosted by the University of Pennsylvania

Exploiting symmetry in structured data is a powerful way to improve the learning and generalization ability of AI systems, and extract higher quality information, in applications ranging from vision, imaging, and NLP to robotics. This is exemplified by convolutional neural nets, which are an ubiquitous architecture. Recently, there has been a great deal of progress to develop improved equivariant and invariant learning architectures, as well as improved data augmentation methods. There has also been progress on the theoretical foundations of the area, from the perspectives of statistics and optimization. The notion of adding data via data augmentation also arises in problems such as adversarial robustness. This workshop will bring together leading researchers in the area to discuss the state of the art of the field. The activity is part of the Center for Foundations of Information Processing at Penn, supported by NSF TRIPODS.

Due to Covid-19, the workshop will be held online as a Zoom webinar. While this means less in-person interaction, it also means that the talks are accessible to everyone for free, without the costs of travel.

Recorded talks are available as a Youtube playlist. Individual videos are also linked below. The link below takes you to the playlist in the "play all" mode.





Some activities

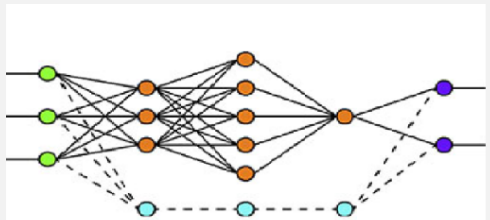
National Science Foundation
WHERE DISCOVERIES BEGIN

Search

Announcement

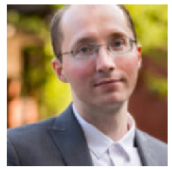
NSF and the Simons Foundation partner to uncover foundations of deep learning

Mathematicians, statisticians, engineers and theoretical computer scientists come together to work on deep learning and AI



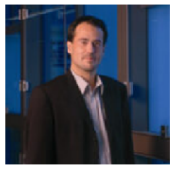

Emmanuel Candès (Stanford)


Ingrid Daubechies (Duke)


Edgar Dobriban (UPenn)


Rong Ge (Duke)


George Pappas (UPenn)


Alejandro Ribeiro (UPenn)


Guillermo Sapiro (Duke)


Shankar Sastry (UC Berkeley)

Some activities

C3.ai DTI

OVERVIEW ABOUT RESEARCH EVENTS NEWS CONTACT

Workshops

The Analytical Foundations of Deep Learning: Interpretability and Performance Guarantees

October 19-21 & 23, 2020

9 am to 2 pm PT (Noon to 5 pm ET) Daily

Attend: Zoom Meeting (You may need to authenticate through your institution's Zoom account before joining)

Watch: YouTube Channel (From our YouTube Channel select the "Live Now" stream)

Day 4 (Friday, Oct. 23): Brainstorm and Discussion

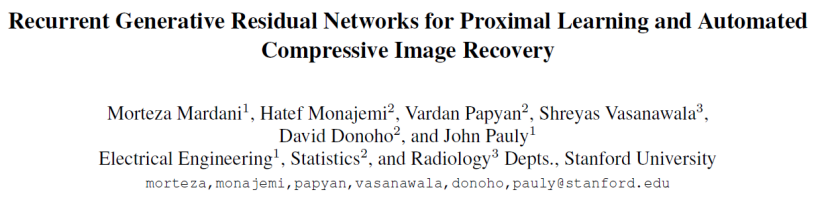
9 am – 10:30 am: Robustness

Lead: Edgar Dobriban (University of Pennsylvania)

Participants: Sebastien Bubeck (Microsoft Research), Jinghui Chen (University of California, Los Angeles), Soheil Feizi (University of Maryland), Micah Goldblum (University of Maryland), Zico Kolter (Carnegie Mellon University), Omar Montasser (Toyota Technological Institute at Chicago), Cyrus Rashtchian (University of California, San Diego), Aditi Raghunathan (Stanford University), Alex Robey (University of Pennsylvania), Chong You (University of California, Berkeley), Hongyang Zhang (Toyota Technological Institute at Chicago)

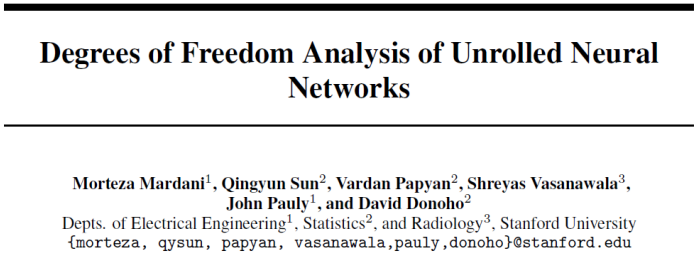
Some considerations

- Criticism of deep learning (resource inefficiency etc) by prominent statisticians (e.g., Donoho)... followed by work from the same people



Recurrent Generative Residual Networks for Proximal Learning and Automated Compressive Image Recovery

Morteza Mardani[1], Hatef Monajemi[2], Vardan Papyan[2], Shreyas Vasanawala[3], David Donoho[2], and John Pauly[1]
Electrical Engineering[1], Statistics[2], and Radiology[3] Depts., Stanford University
morteza,monajemi,papyan,vasanawala,donoho,pauly@stanford.edu



Degrees of Freedom Analysis of Unrolled Neural Networks

Morteza Mardani[1], Qingyun Sun[2], Vardan Papyan[2], Shreyas Vasanawala[3], John Pauly[1], and David Donoho[2]
Depts. of Electrical Engineering[1], Statistics[2], and Radiology[3], Stanford University
{morteza, qysun, papyan, vasanawala,pauly,donoho}@stanford.edu

Neural Proximal Gradient Descent for Compressive Imaging

Morteza Mardani[1], Qingyun Sun[4], Shreyas Vasawanala[2], Vardan Papyan[3], Hatef Monajemi[3], John Pauly[1], and David Donoho[3]
Depts. of [1]Electrical Eng., [2]Radiology, [3]Statistics, and [4]Mathematics; Stanford University
morteza,qysun,vasanawala,papyan,monajemi,pauly,donoho@stanford.edu

Prevalence of neural collapse during the terminal phase of deep learning training

Vardan Papyan[a,1], X. Y. Han[b,1], and David L. Donoho[a,2]

[a]Department of Statistics, Stanford University, Stanford, CA 94305-4065; and [b]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850

Some takeaways

- To ensure that our voice is heard, and that proper methods are used, statisticians need to be involved
- Our contributions are encouraged and valued
- There is a learning curve, and not everyone is welcoming (but worth it)

Overview

This talk

- Emerging area of adversarial robustness in AI/Machine Learning
- Answers to some fundamental statistical questions

Robustness

- Want our results to not be affected by accidental perturbations of the data
- Long history: median (middle ages), L_1 regression (Boscovich, Laplace), M-estimation (Huber, Rousseeuw, ...), ...
- This talk: test-time (adversarial) robustness of prediction methods (?2004–2013–...)

Robustness

- Want our results to not be affected by accidental perturbations of the data
- Long history: median (middle ages), L_1 regression (Boscovich, Laplace), M-estimation (Huber, Rousseeuw, ...), ...
- This talk: test-time (adversarial) robustness of prediction methods (?2004–2013–...)

Robustness

- Want our results to not be affected by accidental perturbations of the data
- Long history: median (middle ages), L_1 regression (Boscovich, Laplace), M-estimation (Huber, Rousseeuw, ...), ...
- This talk: test-time (adversarial) robustness of prediction methods (?2004–2013–...)

Adversarial robustness

- Let $f(x, \hat{\theta})$ be a predictor (classifier, regression function) estimated/learned from data
- For a new test datapoint x, we predict $\hat{y}(x) = f(x, \hat{\theta})$
- Suppose an adversary perturbs $x \to x' = x + \delta$ for some "small" adversarially chosen δ
- Want prediction $\hat{y}(x')$ to be robust/stable (not change much)

Adversarial robustness

- Let $f(x, \hat{\theta})$ be a predictor (classifier, regression function) estimated/learned from data
- For a new test datapoint x, we predict $\hat{y}(x) = f(x, \hat{\theta})$
- Suppose an adversary perturbs $x \to x' = x + \delta$ for some "small" adversarially chosen δ
- Want prediction $\hat{y}(x')$ to be robust/stable (not change much)

Modern methods are not robust

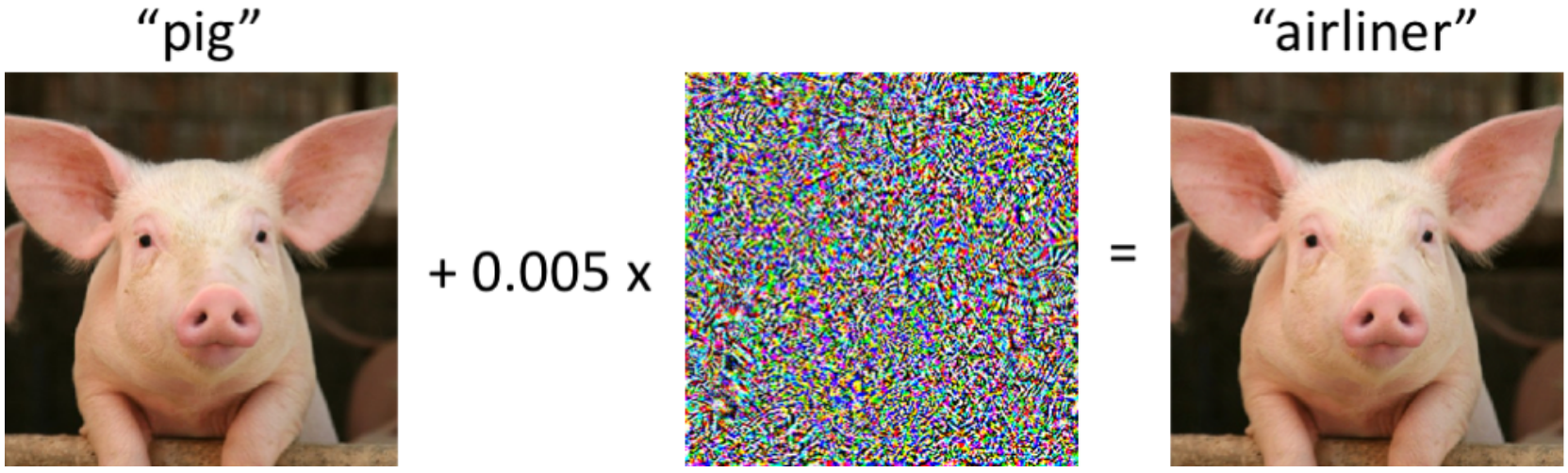


Figure: `gradientscience.org/intro_adversarial/`, by Madry & Schmidt.

- Key challenge in deploying learning algorithms for real problems

Modern methods are not robust: Adversarial patch

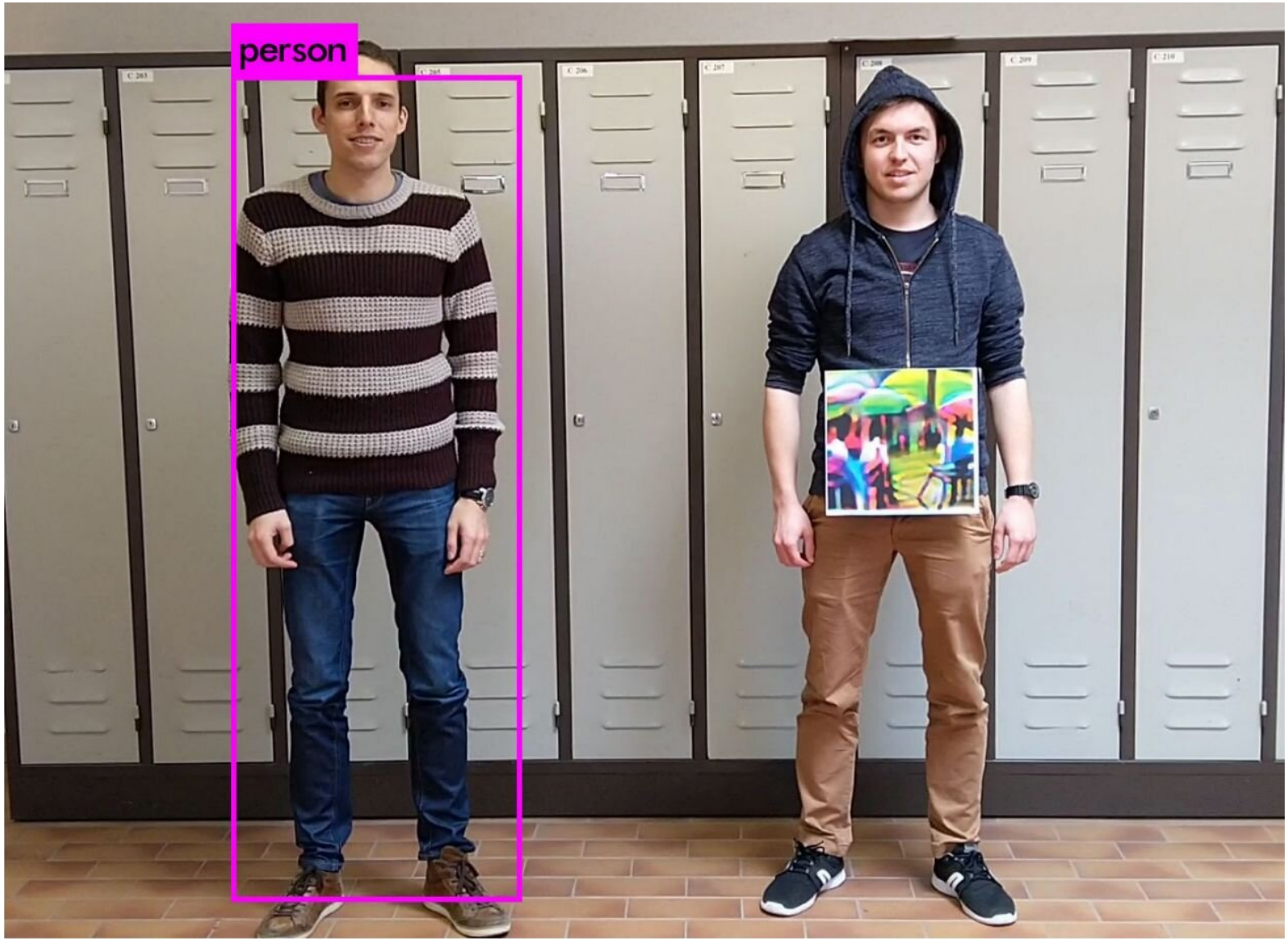


Figure: Thys et al. Fooling automated surveillance cameras: adversarial patches to attack person detection, arXiv:1904.08653

Unprecedented interest

Towards deep learning models resistant to adversarial attacks

A Madry, A Makelov, L Schmidt, D Tsipras… - arXiv preprint arXiv …, 2017 - arxiv.org

Recent work has demonstrated that deep neural networks are vulnerable to adversarial examples---inputs that are almost indistinguishable from natural data and yet classified incorrectly by the network. In fact, some of the latest findings suggest that the existence of …

☆ Cited by 2234 Related articles All 7 versions »

Key questions

- Is there a fundamental (unavoidable) tradeoff between robustness and accuracy?
- What are the optimally robust statistical learning methods?
- How does this depend on the data distribution?

Overview

Model

- Classification problem: features $x \in \mathbb{R}^p$, label $y \in \{\pm 1\}$.
- Classification error:

$$\Pr_{x,y}\{\hat{y}(x) \neq y\} = \mathbb{E}_x \Pr_{y|x}\{\hat{y}(x) \neq y\}$$

- At test time adversary perturbs input by δ, where $\|\delta\| \leq \varepsilon$, ($\|\cdot\|$ is norm).
- Goal: construct classifier $\hat{y} : \mathbb{R}^p \to \{\pm 1\}$ minimizing robust risk: probability of mistake after adversarial perturbation

$$R(\hat{y}, \varepsilon, \|\cdot\|) = \Pr_{x,y}\{\exists_{\delta:\|\delta\|\leq\varepsilon}\ \hat{y}(x+\delta) \neq y\}$$

- Standard risk: $\varepsilon = 0$: $R_s(\hat{y}) = R(\hat{y}, 0, \|\cdot\|) = \Pr_{x,y}\{\hat{y}(x) \neq y\}$
- Bayes optimal classifier: optimal choice for each x

$$\hat{y}(x) \in \operatorname{argmax}_{c \in C} \Pr_{y|x}(y = c)$$

- *Cannot* make separate optimal choices for each x when $\varepsilon > 0$

Model

- Classification problem: features $x \in \mathbb{R}^p$, label $y \in \{\pm 1\}$.
- Classification error:

$$\Pr_{x,y}\{\hat{y}(x) \neq y\} = \mathbb{E}_x \Pr_{y|x}\{\hat{y}(x) \neq y\}$$

- At test time adversary perturbs input by δ, where $\|\delta\| \leq \varepsilon$, ($\|\cdot\|$ is norm).
- Goal: construct classifier $\hat{y} : \mathbb{R}^p \to \{\pm 1\}$ minimizing robust risk: probability of mistake after adversarial perturbation

$$R(\hat{y}, \varepsilon, \|\cdot\|) = \Pr_{x,y}\{\exists_{\delta:\|\delta\|\leq\varepsilon} \ \hat{y}(x+\delta) \neq y\}$$

- Standard risk: $\varepsilon = 0$: $R_s(\hat{y}) = R(\hat{y}, 0, \|\cdot\|) = \Pr_{x,y}\{\hat{y}(x) \neq y\}$
- Bayes optimal classifier: optimal choice for each x

$$\hat{y}(x) \in \operatorname{argmax}_{c \in C} \Pr_{y|x}(y = c)$$

- *Cannot* make separate optimal choices for each x when $\varepsilon > 0$

Model

- Classification problem: features $x \in \mathbb{R}^p$, label $y \in \{\pm 1\}$.
- Classification error:

$$\Pr_{x,y}\{\hat{y}(x) \neq y\} = \mathbb{E}_x \Pr_{y|x}\{\hat{y}(x) \neq y\}$$

- At test time adversary perturbs input by δ, where $\|\delta\| \leq \varepsilon$, ($\|\cdot\|$ is norm).
- Goal: construct classifier $\hat{y} : \mathbb{R}^p \to \{\pm 1\}$ minimizing robust risk: probability of mistake after adversarial perturbation

$$R(\hat{y}, \varepsilon, \|\cdot\|) = \Pr_{x,y}\{\exists_{\delta:\|\delta\|\leq\varepsilon}\ \hat{y}(x+\delta) \neq y\}$$

- Standard risk: $\varepsilon = 0$: $R_s(\hat{y}) = R(\hat{y}, 0, \|\cdot\|) = \Pr_{x,y}\{\hat{y}(x) \neq y\}$
- Bayes optimal classifier: optimal choice for each x

$$\hat{y}(x) \in \operatorname{argmax}_{c \in C} \Pr_{y|x}(y = c)$$

- *Cannot* make separate optimal choices for each x when $\varepsilon > 0$

Model

- Classification problem: features $x \in \mathbb{R}^p$, label $y \in \{\pm 1\}$.
- Classification error:

$$\Pr_{x,y}\{\hat{y}(x) \neq y\} = \mathbb{E}_x \Pr_{y|x}\{\hat{y}(x) \neq y\}$$

- At test time adversary perturbs input by δ, where $\|\delta\| \leq \varepsilon$, ($\|\cdot\|$ is norm).
- Goal: construct classifier $\hat{y} : \mathbb{R}^p \to \{\pm 1\}$ minimizing robust risk: probability of mistake after adversarial perturbation

$$R(\hat{y}, \varepsilon, \|\cdot\|) = \Pr_{x,y}\{\exists_{\delta:\|\delta\|\leq\varepsilon} \ \hat{y}(x+\delta) \neq y\}$$

- Standard risk: $\varepsilon = 0$: $R_s(\hat{y}) = R(\hat{y}, 0, \|\cdot\|) = \Pr_{x,y}\{\hat{y}(x) \neq y\}$
- Bayes optimal classifier: optimal choice for each x

$$\hat{y}(x) \in \operatorname{argmax}_{c \in C} \Pr_{y|x}(y = c)$$

- *Cannot* make separate optimal choices for each x when $\varepsilon > 0$

Model

- Classification problem: features $x \in \mathbb{R}^p$, label $y \in \{\pm 1\}$.
- Classification error:

$$\Pr_{x,y}\{\hat{y}(x) \neq y\} = \mathbb{E}_x \Pr_{y|x}\{\hat{y}(x) \neq y\}$$

- At test time adversary perturbs input by δ, where $\|\delta\| \leq \varepsilon$, ($\|\cdot\|$ is norm).
- Goal: construct classifier $\hat{y} : \mathbb{R}^p \to \{\pm 1\}$ minimizing robust risk: probability of mistake after adversarial perturbation

$$R(\hat{y}, \varepsilon, \|\cdot\|) = \Pr_{x,y}\{\exists_{\delta:\|\delta\|\leq\varepsilon} \ \hat{y}(x+\delta) \neq y\}$$

- Standard risk: $\varepsilon = 0$: $R_s(\hat{y}) = R(\hat{y}, 0, \|\cdot\|) = \Pr_{x,y}\{\hat{y}(x) \neq y\}$
- Bayes optimal classifier: optimal choice for each x

$$\hat{y}(x) \in \operatorname{argmax}_{c \in C} \Pr_{y|x}(y = c)$$

- *Cannot* make separate optimal choices for each x when $\varepsilon > 0$

Model

- Classification problem: features $x \in \mathbb{R}^p$, label $y \in \{\pm 1\}$.
- Classification error:

$$\Pr_{x,y}\{\hat{y}(x) \neq y\} = \mathbb{E}_x \Pr_{y|x}\{\hat{y}(x) \neq y\}$$

- At test time adversary perturbs input by δ, where $\|\delta\| \leq \varepsilon$, ($\|\cdot\|$ is norm).
- Goal: construct classifier $\hat{y} : \mathbb{R}^p \to \{\pm 1\}$ minimizing robust risk: probability of mistake after adversarial perturbation

$$R(\hat{y}, \varepsilon, \|\cdot\|) = \Pr_{x,y}\{\exists_{\delta:\|\delta\|\leq\varepsilon}\ \hat{y}(x+\delta) \neq y\}$$

- Standard risk: $\varepsilon = 0$: $R_s(\hat{y}) = R(\hat{y}, 0, \|\cdot\|) = \Pr_{x,y}\{\hat{y}(x) \neq y\}$
- Bayes optimal classifier: optimal choice for each x

$$\hat{y}(x) \in \operatorname{argmax}_{c \in C} \Pr_{y|x}(y = c)$$

- *Cannot* make separate optimal choices for each x when $\varepsilon > 0$

Model

- Classification problem: features $x \in \mathbb{R}^p$, label $y \in \{\pm 1\}$.
- Classification error:

$$\Pr_{x,y}\{\hat{y}(x) \neq y\} = \mathbb{E}_x \Pr_{y|x}\{\hat{y}(x) \neq y\}$$

- At test time adversary perturbs input by δ, where $\|\delta\| \leq \varepsilon$, ($\|\cdot\|$ is norm).
- Goal: construct classifier $\hat{y} : \mathbb{R}^p \to \{\pm 1\}$ minimizing robust risk: probability of mistake after adversarial perturbation

$$R(\hat{y}, \varepsilon, \|\cdot\|) = \Pr_{x,y}\{\exists_{\delta:\|\delta\|\leq\varepsilon}\ \hat{y}(x+\delta) \neq y\}$$

- Standard risk: $\varepsilon = 0$: $R_s(\hat{y}) = R(\hat{y}, 0, \|\cdot\|) = \Pr_{x,y}\{\hat{y}(x) \neq y\}$
- Bayes optimal classifier: optimal choice for each x

$$\hat{y}(x) \in \operatorname{argmax}_{c \in C} \Pr_{y|x}(y = c)$$

- *Cannot* make separate optimal choices for each x when $\varepsilon > 0$

Geometry of robust classifiers

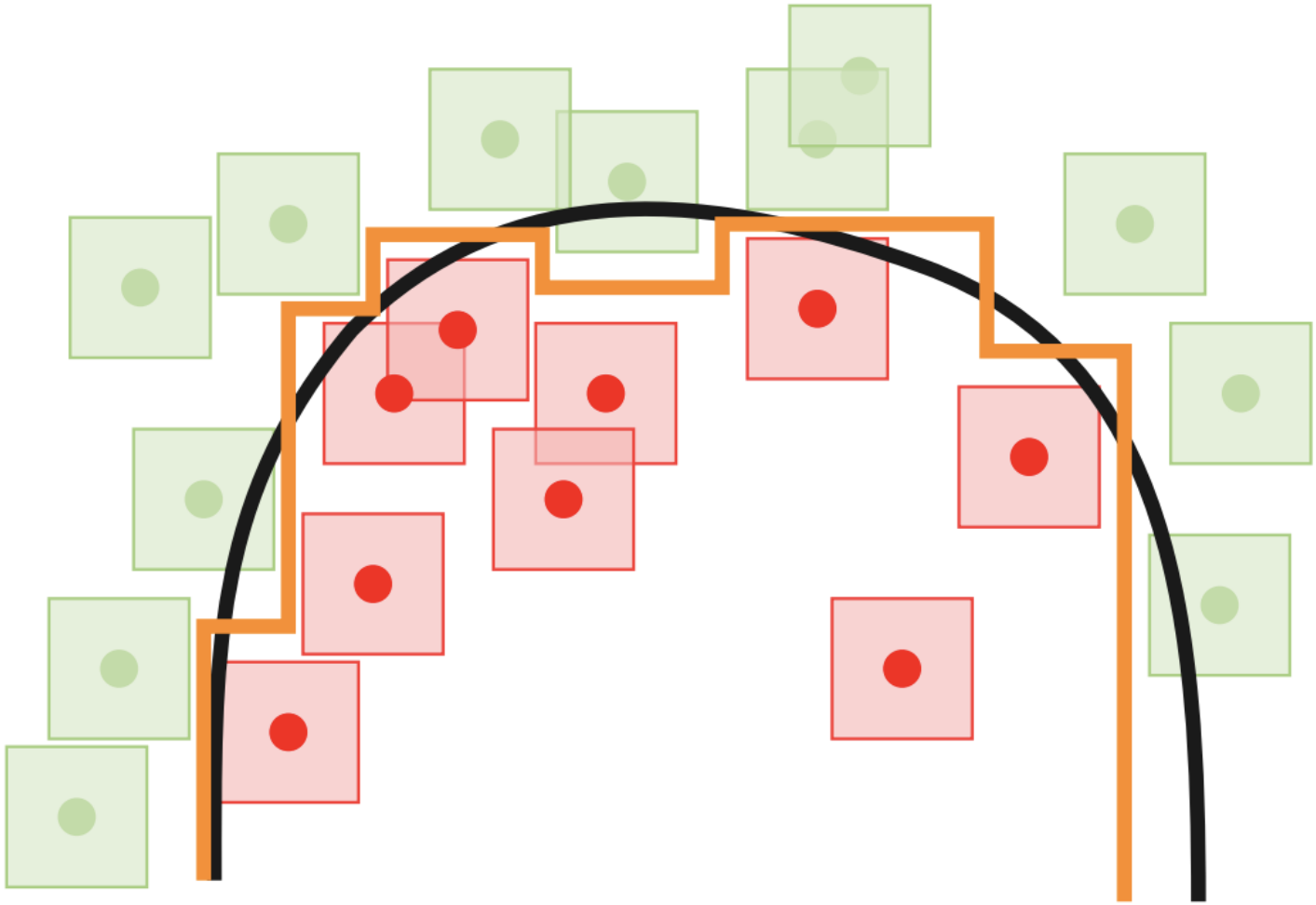


Figure: Figure from Yang el al., 2020

Two-class Gaussian Model

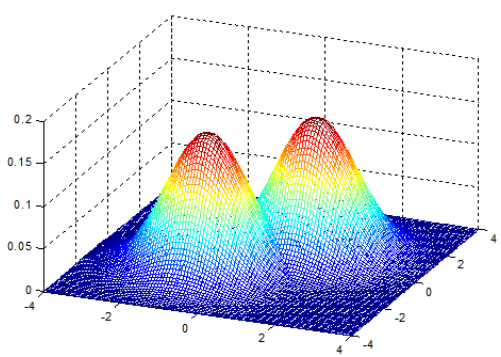
- $y \in \{\pm 1\}$, $P(y = 1) = \pi$
- Isotropic Gaussian class-conditional distribution (wlog $\sigma^2 = 1$):

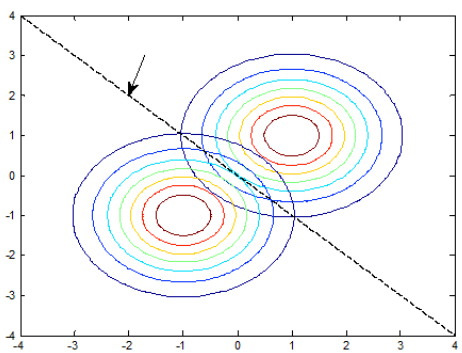
$$x|y \sim \mathcal{N}(y\mu, \sigma^2 I_p),$$

- Bayes optimal (non-robust) classifier is linear (special case of Fisher's LDA):

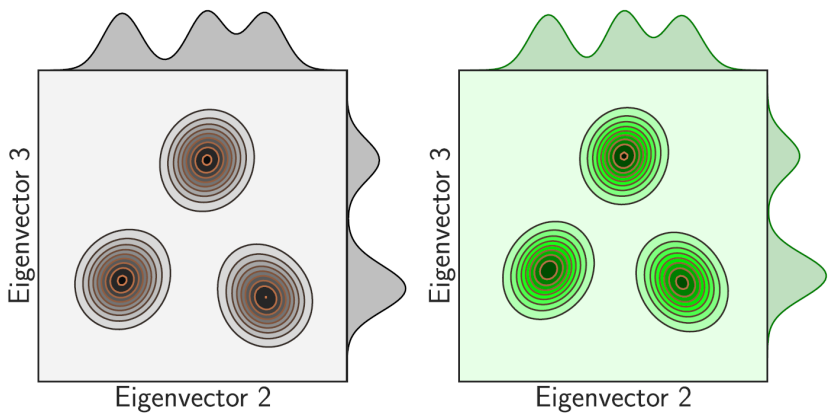
$$\hat{y}_b(x) = \text{argmax}_{c \in C} \Pr_{y|x}(y = c) = \text{sign}(x^\top \mu - q/2),$$

$q := \ln\{(1 - \pi)/\pi\}$ is the log-odds-ratio





(a) Puente et al., 2010



(b) Seddik et al. 2020: GAN $\approx$ GMM

Two-class Gaussian Model

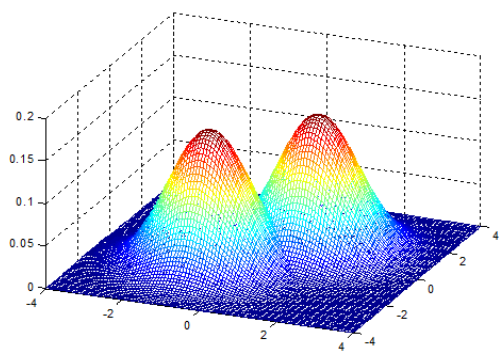
- $y \in \{\pm 1\}$, $P(y = 1) = \pi$
- Isotropic Gaussian class-conditional distribution (wlog $\sigma^2 = 1$):

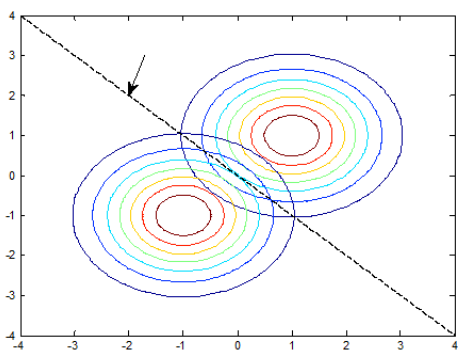
$$x|y \sim \mathcal{N}(y\mu, \sigma^2 I_p),$$

- Bayes optimal (non-robust) classifier is linear (special case of Fisher's LDA):

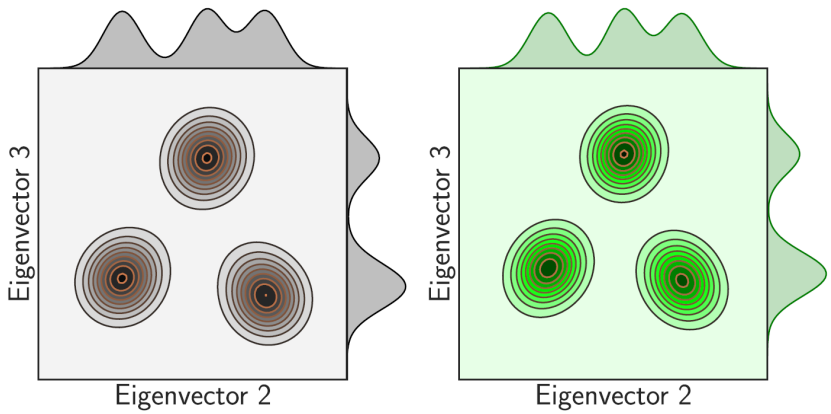
$$\hat{y}_b(x) = \text{argmax}_{c \in C} \Pr_{y|x}(y = c) = \text{sign}(x^\top \mu - q/2),$$

$q := \ln\{(1 - \pi)/\pi\}$ is the log-odds-ratio





(a) Puente et al., 2010



(b) Seddik et al. 2020: GAN $\approx$ GMM

Two-class Gaussian Model

- Bayes risk

$$R_{\mathrm{Bay}}(\mu, \pi) = \pi \cdot \Phi\left(\frac{q}{2\|\mu\|_2} - \|\mu\|_2\right) + (1-\pi) \cdot \bar{\Phi}\left(\frac{q}{2\|\mu\|_2} + \|\mu\|_2\right),$$

  Φ normal c.d.f, $\bar{\Phi} := 1 - \Phi$, normal tail d.f.

- Two-class Gaussian model fundamental and has rich history (Pearson, Fisher, Wald, ...). Serves as basis for many modern high-dim developments

Two-class Gaussian Model

- Bayes risk

$$R_{\mathrm{Bay}}(\mu, \pi) = \pi \cdot \Phi\left(\frac{q}{2\|\mu\|_2} - \|\mu\|_2\right) + (1-\pi) \cdot \bar{\Phi}\left(\frac{q}{2\|\mu\|_2} + \|\mu\|_2\right),$$

  Φ normal c.d.f, $\bar{\Phi} := 1 - \Phi$, normal tail d.f.

- Two-class Gaussian model fundamental and has rich history (Pearson, Fisher, Wald, ...). Serves as basis for many modern high-dim developments

Questions

- Can we find the optimal robust classifiers?
- Challenge: decisions made at each point influence neighborhood, does not decouple
- Tradeoffs? Dependence on data distribution?

Optimal ℓ_2-robust two-class classifiers

Take $\|\cdot\| = \|\cdot\|_2$ to be the ℓ_2 norm.

Theorem

In the two-class Gaussian model, the optimal ℓ_2 robust classifier is the linear classifier:

$$\hat{y}^*(x) := \text{sign}\left\{x^\top \mu \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+ - q/2\right\},$$

where $q = \ln\{(1-\pi)/\pi\}$ and $(x)_+ = \max(x, 0)$. Moreover, the corresponding optimal robust risk is

$$R^*_{\text{rob}}(\mu, \pi; \varepsilon) := R_{\text{Bay}}\left\{\mu\left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+, \pi\right\}.$$

Optimal ℓ_2-robust two-class classifiers

Take $\|\cdot\| = \|\cdot\|_2$ to be the ℓ_2 norm.

Theorem

In the two-class Gaussian model, the optimal ℓ_2 robust classifier is the linear classifier:

$$\hat{y}^*(x) := \text{sign}\left\{x^\top \mu \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+ - q/2\right\},$$

where $q = \ln\{(1-\pi)/\pi\}$ and $(x)_+ = \max(x, 0)$. Moreover, the corresponding optimal robust risk is

$$R^*_{\text{rob}}(\mu, \pi; \varepsilon) := R_{\text{Bay}}\left\{\mu\left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+, \pi\right\}.$$

Optimal ℓ_2-robust two-class classifiers

Take $\|\cdot\| = \|\cdot\|_2$ to be the ℓ_2 norm.

Theorem

In the two-class Gaussian model, the optimal ℓ_2 robust classifier is the linear classifier:

$$\hat{y}^*(x) := \text{sign}\left\{x^\top \mu \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+ - q/2\right\},$$

where $q = \ln\{(1-\pi)/\pi\}$ and $(x)_+ = \max(x, 0)$. Moreover, the corresponding optimal robust risk is

$$R^*_{\text{rob}}(\mu, \pi; \varepsilon) := R_{\text{Bay}}\left\{\mu\left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+, \pi\right\}.$$

Optimal ℓ_2-robust two-class classifiers: Discussion

- ℓ_2 robust problem equivalent to standard prob. with reduced effective mean $\mu \mapsto \mu \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+$
  - $\|\mu\|_2 \geq \varepsilon \implies$ nontrivial classification impossible
  - consistent with "adv. rob. generalization requires more data" (Schmidt et al., 2018)
- also equivalent to standard prob. with amplified class imbalance $q \mapsto q / \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+$

$$\hat{y}^*(x) = \operatorname{sign}\left\{ x^\top \mu - \frac{q}{2\left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+} \right\},$$

  1. robustness requires predicting majority $\hat{y}$ more often
- subtle tradeoff depends on class balance:
  - classes balanced ($q = \ln\{(1-\pi)/\pi\} = 0$): same classifier optimal for all $\varepsilon \geq 0$; no tradeoff
  - classes imbalanced: tradeoff

Optimal ℓ_2-robust two-class classifiers: Discussion

- ℓ_2 robust problem equivalent to standard prob. with reduced effective mean $\mu \mapsto \mu \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+$
  - $\|\mu\|_2 \geq \varepsilon \implies$ nontrivial classification impossible
  - consistent with "adv. rob. generalization requires more data" (Schmidt et al., 2018)
- also equivalent to standard prob. with amplified class imbalance $q \mapsto q / \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+$

$$\hat{y}^*(x) = \text{sign}\left\{x^\top \mu - \frac{q}{2\left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+}\right\},$$

1. robustness requires predicting majority $\hat{y}$ more often

- subtle tradeoff depends on class balance:
  - classes balanced ($q = \ln\{(1-\pi)/\pi\} = 0$): same classifier optimal for all $\varepsilon \geq 0$; no tradeoff
  - classes imbalanced: tradeoff

Optimal ℓ_2-robust two-class classifiers: Discussion

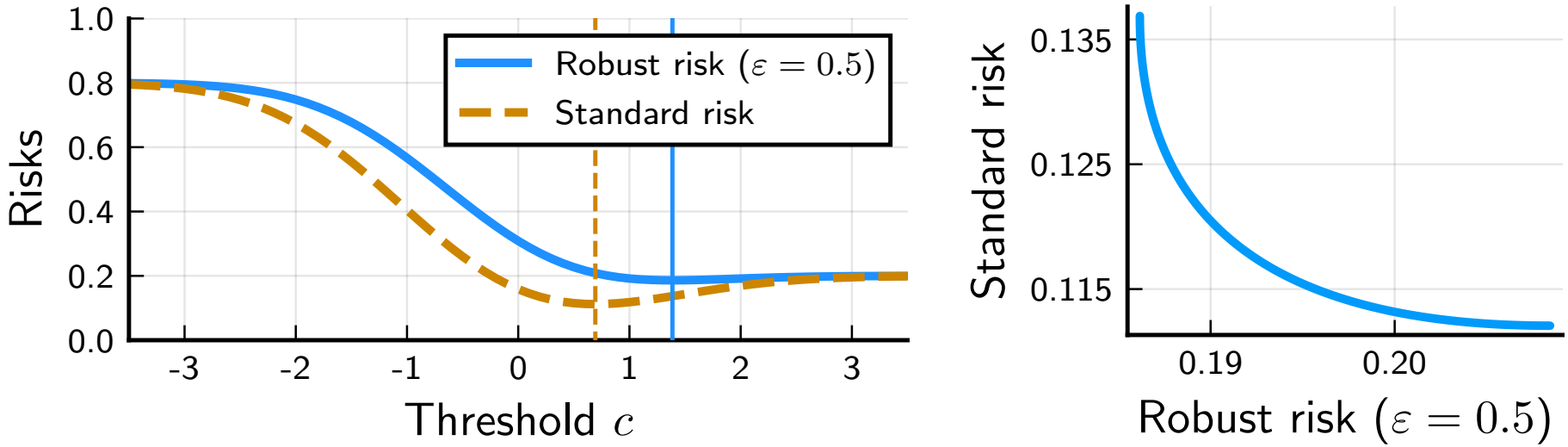
- ℓ_2 robust problem equivalent to standard prob. with reduced effective mean $\mu \mapsto \mu \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+$
  - $\|\mu\|_2 \geq \varepsilon \implies$ nontrivial classification impossible
  - consistent with "adv. rob. generalization requires more data" (Schmidt et al., 2018)
- also equivalent to standard prob. with amplified class imbalance $q \mapsto q / \left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+$

$$\hat{y}^*(x) = \text{sign}\left\{ x^\top \mu - \frac{q}{2\left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+} \right\},$$

1. robustness requires predicting majority $\hat{y}$ more often

- subtle tradeoff depends on class balance:
  - classes balanced ($q = \ln\{(1-\pi)/\pi\} = 0$): same classifier optimal for all $\varepsilon \geq 0$; no tradeoff
  - classes imbalanced: tradeoff

Tradeoffs between robustness and accuracy



(a) Risks as functions of threshold c; vertical lines at optimal thresholds.

(b) Standard v.s. robust risk for thresholds between minima.

Figure: Tradeoffs between standard and robust risks. $\|\mu\|_2 = 1$, $\pi = 0.2$, $\varepsilon = 0.5$. Risks of the linear classifier $\hat{y}(x) = \text{sign}(x^\top \mu - c)$ as a function of the threshold c.

Optimal ℓ_2-robust two-class classifiers: Discussion

- Robust problem equivalent to adding noise

$$\mathcal{N}\left(0, \left[\frac{1}{\left(1-\frac{\varepsilon}{\|\mu\|_2}\right)_+} - 1\right] \cdot I_p\right)$$

  to the data.

- Such algorithms were proposed in (Xie et al., 2018; Athalye et al., 2017).

Optimal ℓ_2-robust two-class classifiers: Discussion

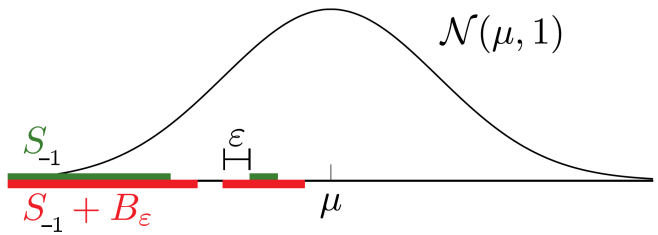
- Robust problem equivalent to adding noise

$$\mathcal{N}\left(0, \left[\frac{1}{\left(1-\frac{\varepsilon}{\|\mu\|_2}\right)_+} - 1\right] \cdot I_p\right)$$

  to the data.
- Such algorithms were proposed in (Xie et al., 2018; Athalye et al., 2017).

Optimal ℓ_2-robust two-class classifiers: Proof (wlog $\mu > 0$)

$$R(\hat{y}) = \pi \cdot P_{x|y=+1}(S_{-1} + B_\varepsilon) + (1-\pi) \cdot P_{x|y=-1}(S_{+1} + B_\varepsilon)$$
$$S_{\pm 1} = \{x : \hat{y}(x) = \pm 1\}$$



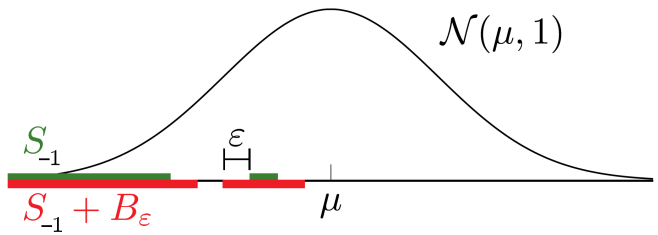
- Step 1: linear classifiers are admissible
  - can project to 1-D
  - Gaussian concentration of measure (Borell, 1975; Sudakov and Tsirel'son, 1978): There is a half-line $\tilde{S}_{-1} = (-\infty, c]$ st (& analogue for $+1$)

$$P_+(\tilde{S}_{-1} + B_\varepsilon) \leq P_+(S_{-1} + B_\varepsilon), \qquad \text{and} \qquad P_+(\tilde{S}_{-1}) = P_+(S_{-1}).$$

  - Neyman-Pearson: $\tilde{S}_{\pm 1}$ overlap, can shrink them to form linear classif.
- Step 2: find optimal linear classifiers

Optimal ℓ_2-robust two-class classifiers: Proof (wlog $\mu > 0$)

$$R(\hat{y}) = \pi \cdot P_{x|y=+1}(S_{-1} + B_\varepsilon) + (1-\pi) \cdot P_{x|y=-1}(S_{+1} + B_\varepsilon)$$
$$S_{\pm 1} = \{x : \hat{y}(x) = \pm 1\}$$



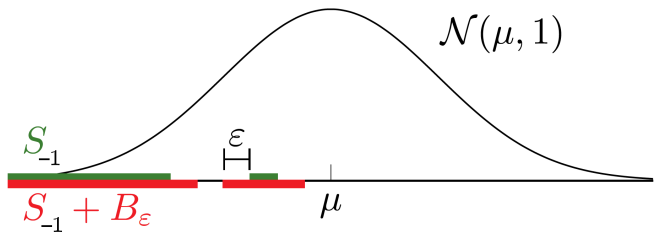
- Step 1: linear classifiers are admissible
  - can project to 1-D
  - Gaussian concentration of measure (Borell, 1975; Sudakov and Tsirel'son, 1978): There is a half-line $\tilde{S}_{-1} = (-\infty, c]$ st (& analogue for $+1$)

$$P_+(\tilde{S}_{-1} + B_\varepsilon) \leq P_+(S_{-1} + B_\varepsilon), \qquad \text{and} \qquad P_+(\tilde{S}_{-1}) = P_+(S_{-1}).$$

  - Neyman-Pearson: $\tilde{S}_{\pm 1}$ overlap, can shrink them to form linear classif.
- Step 2: find optimal linear classifiers

Optimal ℓ_2-robust two-class classifiers: Proof (wlog $\mu > 0$)

$$R(\hat{y}) = \pi \cdot P_{x|y=+1}(S_{-1} + B_\varepsilon) + (1-\pi) \cdot P_{x|y=-1}(S_{+1} + B_\varepsilon)$$
$$S_{\pm 1} = \{x : \hat{y}(x) = \pm 1\}$$



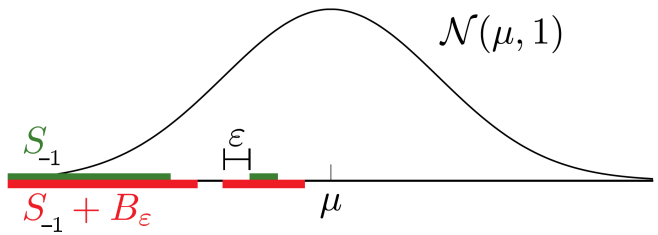
- Step 1: linear classifiers are admissible
  - can project to 1-D
  - Gaussian concentration of measure (Borell, 1975; Sudakov and Tsirel'son, 1978): There is a half-line $\tilde{S}_{-1} = (-\infty, c]$ st (& analogue for $+1$)

    $$P_+(\tilde{S}_{-1} + B_\varepsilon) \leq P_+(S_{-1} + B_\varepsilon), \quad \text{and} \quad P_+(\tilde{S}_{-1}) = P_+(S_{-1}).$$

  - Neyman-Pearson: $\tilde{S}_{\pm 1}$ overlap, can shrink them to form linear classif.
- Step 2: find optimal linear classifiers

Optimal ℓ_2-robust two-class classifiers: Proof (wlog $\mu > 0$)

$$R(\hat{y}) = \pi \cdot P_{x|y=+1}(S_{-1} + B_\varepsilon) + (1-\pi) \cdot P_{x|y=-1}(S_{+1} + B_\varepsilon)$$
$$S_{\pm 1} = \{x : \hat{y}(x) = \pm 1\}$$



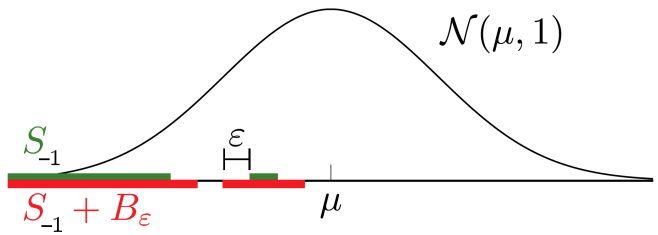
- Step 1: linear classifiers are admissible
  - can project to 1-D
  - Gaussian concentration of measure (Borell, 1975; Sudakov and Tsirel'son, 1978): There is a half-line $\tilde{S}_{-1} = (-\infty, c]$ st (& analogue for $+1$)

$$P_+(\tilde{S}_{-1} + B_\varepsilon) \leq P_+(S_{-1} + B_\varepsilon), \qquad \text{and} \qquad P_+(\tilde{S}_{-1}) = P_+(S_{-1}).$$

  - Neyman-Pearson: $\tilde{S}_{\pm 1}$ overlap, can shrink them to form linear classif.
- Step 2: find optimal linear classifiers

Optimal ℓ_2-robust two-class classifiers: Proof (wlog $\mu > 0$)

$$R(\hat{y}) = \pi \cdot P_{x|y=+1}(S_{-1} + B_\varepsilon) + (1-\pi) \cdot P_{x|y=-1}(S_{+1} + B_\varepsilon)$$
$$S_{\pm 1} = \{x : \hat{y}(x) = \pm 1\}$$



- Step 1: linear classifiers are admissible
  - can project to 1-D
  - Gaussian concentration of measure (Borell, 1975; Sudakov and Tsirel'son, 1978): There is a half-line $\tilde{S}_{-1} = (-\infty, c]$ st (& analogue for $+1$)

$$P_+(\tilde{S}_{-1} + B_\varepsilon) \leq P_+(S_{-1} + B_\varepsilon), \qquad \text{and} \qquad P_+(\tilde{S}_{-1}) = P_+(S_{-1}).$$

  - Neyman-Pearson: $\tilde{S}_{\pm 1}$ overlap, can shrink them to form linear classif.
- Step 2: find optimal linear classifiers

Optimal ℓ_2-robust two-class classifiers: Extensions

- Weighted combinations: $R(\hat{y}, Q) = \mathbb{E}_{\varepsilon \sim Q} R(\hat{y}, \varepsilon)$. e.g., $R(\hat{y}, 0) + \lambda \cdot R(\hat{y}, \varepsilon)$ for $\lambda > 0$.
- Low-dimensional manifold: Data lies on an affine subspace
- General covariance: top eigenvector aligns with diff. in means

Optimal ℓ_2-robust two-class classifiers: Extensions

- Weighted combinations: $R(\hat{y}, Q) = \mathbb{E}_{\varepsilon \sim Q} R(\hat{y}, \varepsilon)$. e.g., $R(\hat{y}, 0) + \lambda \cdot R(\hat{y}, \varepsilon)$ for $\lambda > 0$.
- Low-dimensional manifold: Data lies on an affine subspace
- General covariance: top eigenvector aligns with diff. in means

Optimal ℓ_2-robust two-class classifiers: Extensions

- Weighted combinations: $R(\hat{y}, Q) = \mathbb{E}_{\varepsilon \sim Q} R(\hat{y}, \varepsilon)$. e.g., $R(\hat{y}, 0) + \lambda \cdot R(\hat{y}, \varepsilon)$ for $\lambda > 0$.
- Low-dimensional manifold: Data lies on an affine subspace
- General covariance: top eigenvector aligns with diff. in means

Approximately optimal ℓ_2-robust two-class classifiers

- Important to know what the approximately optimal classifiers are. But: seems very hard
- For now, results in 1-D, and special classifiers
- γ: Gaussian measure
- $\gamma^*(S)$ is Gaussian deficit of S; approximability by half-lines

$$\gamma^*(S) := \inf_{\text{half-lines } H} \gamma(S \Delta H)$$

- $\gamma^*(S) \geq 0$, equality when S is a half-line almost surely

Approximately optimal ℓ_2-robust two-class classifiers

- Important to know what the approximately optimal classifiers are. But: seems very hard
- For now, results in 1-D, and special classifiers
- γ: Gaussian measure
- $\gamma^*(S)$ is Gaussian deficit of S; approximability by half-lines

$$\gamma^*(S) := \inf_{\text{half-lines } H} \gamma(S \Delta H)$$

- $\gamma^*(S) \geq 0$, equality when S is a half-line almost surely

Approximately optimal robust classifiers

1. Two-class Gaussian problem in 1-D
2. Classifiers whose classification regions $S_\pm$ are unions of intervals with endpoints in $[-M, M]$
3. ε is less than the half-width of all intervals

Theorem (All robust classifiers are close to linear)

Define $\tau = \tau(\varepsilon, M, \mu) = \varepsilon \exp\{-[(M+\mu)\varepsilon + \varepsilon^2/2]\}$. *Then, for some universal constant* $c > 0$,

$$R_{\text{rob}}(\hat{y}, \varepsilon) \geq R_{\text{Bay}} + \tau \cdot c \cdot [\pi \cdot \gamma^*(S_- - \mu)^2 + (1-\pi) \cdot \gamma^*(S_+ + \mu)^2].$$

If robust risk $R_{\text{rob}}(\hat{y}, \varepsilon)$ small, then the decision regions $S_\pm$ are near half-lines.

Approximately optimal robust classifiers

1. Two-class Gaussian problem in 1-D
2. Classifiers whose classification regions $S_\pm$ are unions of intervals with endpoints in $[-M, M]$
3. ε is less than the half-width of all intervals

Theorem (All robust classifiers are close to linear)

Define $\tau = \tau(\varepsilon, M, \mu) = \varepsilon \exp\{-[(M+\mu)\varepsilon + \varepsilon^2/2]\}$. *Then, for some universal constant* $c > 0$,

$$R_{\mathrm{rob}}(\hat{y}, \varepsilon) \geq R_{\mathrm{Bay}} + \tau \cdot c \cdot [\pi \cdot \gamma^*(S_- - \mu)^2 + (1-\pi) \cdot \gamma^*(S_+ + \mu)^2].$$

If robust risk $R_{\mathrm{rob}}(\hat{y}, \varepsilon)$ small, then the decision regions $S_\pm$ are near half-lines.

Approximately optimal robust classifiers: Proof

- Hinges on recent results from robust isoperimetry (Cianchi et al., 2011; Mossel and Neeman, 2015).
- "If a set is approximately isoperimetric (its boundary measure is close to minimal for its volume), then it is close to a hyperplane"
- Have not been used in statistics/machine learning/data science before?

Approximately optimal robust classifiers: Tools

The Annals of Probability
2015, Vol. 43, No. 3, 971–991
DOI: 10.1214/13-AOP860

ROBUST DIMENSION FREE ISOPERIMETRY IN GAUSSIAN SPACE[1]

By Elchanan Mossel and Joe Neeman

University of California, Berkeley

We prove the first robust dimension free isoperimetric result for the standard Gaussian measure γ_n and the corresponding boundary measure γ_n^+ in $\mathbb{R}^n$. The main result in the theory of Gaussian isoperimetry (proven in the 1970s by Sudakov and Tsirelson, and independently by Borell) states that if $\gamma_n(A) = 1/2$ then the surface area of A is bounded by the surface area of a half-space with the same measure, $\gamma_n^+(A) \leq (2\pi)^{-1/2}$. Our results imply in particular that if $A \subset \mathbb{R}^n$ satisfies $\gamma_n(A) = 1/2$ and $\gamma_n^+(A) \leq (2\pi)^{-1/2} + \delta$ then there exists a half-space $B \subset \mathbb{R}^n$ such that $\gamma_n(A \Delta B) \leq C \log^{-1/2}(1/\delta)$ for an absolute constant C. Since the Gaussian isoperimetric result was established, only recently a robust version of the Gaussian isoperimetric result was obtained by Cianchi et al., who showed that $\gamma_n(A \Delta B) \leq C(n)\sqrt{\delta}$ for some function $C(n)$ with no effective bounds. Compared to the results of Cianchi et al., our results have optimal (i.e., no) dependence on the dimension, but worse dependence on δ.

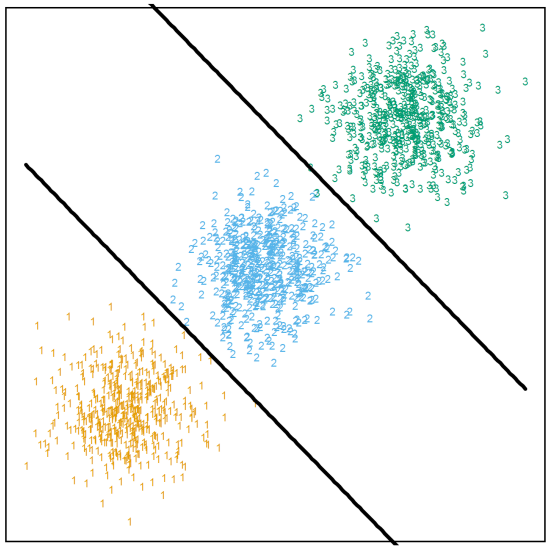
ON THE ISOPERIMETRIC DEFICIT IN GAUSS SPACE

By A. Cianchi, N. Fusco, F. Maggi, and A. Pratelli

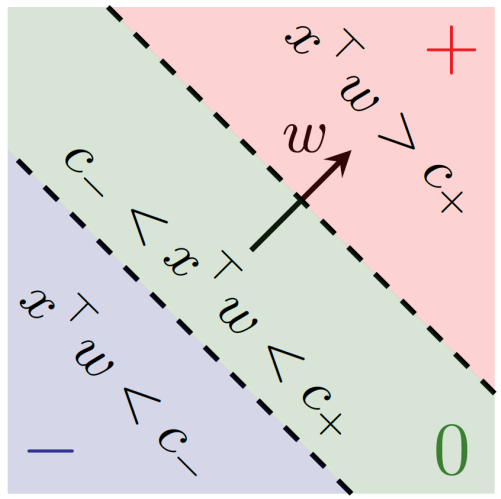
Optimal ℓ_2-robust three-class classifiers

- More general setting:
  1. three classes $C = \{-1, 0, 1\}$: $y = \pm 1$ w.p. $\pi_\pm$ and $y = 0$ w.p. π_0, where $\pi_+, \pi_0, \pi_- \in [0, 1]$ sum to unity and are the class proportions.
  2. $x|y \sim \mathcal{N}(y\mu, I_p)$
- Bayes optimal classif. is a linear-interval classif. ($w = \mu$, $c_+ \geq c_-$):

$$\hat{y}_{int}(x; w, c_+, c_-) := \begin{cases} +1 & \text{if } x^\top w \geq c_+, \\ 0 & \text{if } c_- \leq x^\top w < c_+, \\ -1 & \text{if } x^\top w \leq c_-. \end{cases}$$



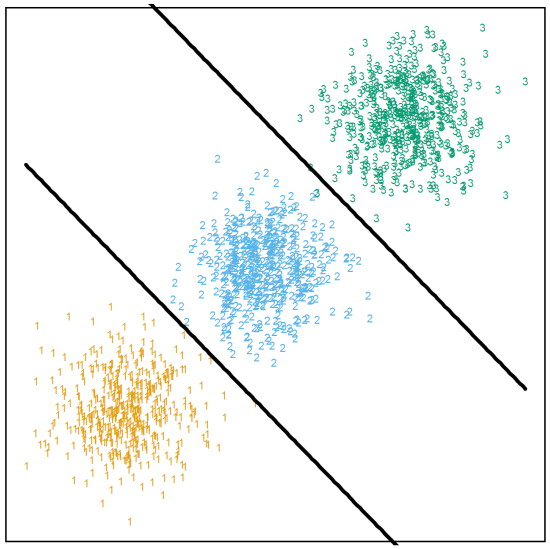
(a) Figure: Hastie et al., 2009



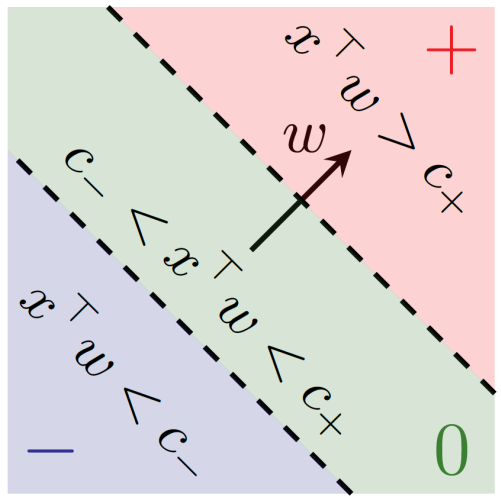
Optimal ℓ_2-robust three-class classifiers

- More general setting:
  1. three classes $C = \{-1, 0, 1\}$: $y = \pm 1$ w.p. $\pi_\pm$ and $y = 0$ w.p. π_0, where $\pi_+, \pi_0, \pi_- \in [0, 1]$ sum to unity and are the class proportions.
  2. $x|y \sim \mathcal{N}(y\mu, I_p)$
- Bayes optimal classif. is a linear-interval classif. ($w = \mu$, $c_+ \geq c_-$):

$$\hat{y}_{int}(x; w, c_+, c_-) := \begin{cases} +1 & \text{if } x^\top w \geq c_+, \\ 0 & \text{if } c_- \leq x^\top w < c_+, \\ -1 & \text{if } x^\top w \leq c_-. \end{cases}$$



(a) Figure: Hastie et al., 2009



Optimal ℓ_2-robust three-class classifiers

Theorem

For $\varepsilon < \|\mu\|_2/2$, an optimal linear interval ℓ_2-robust classifier is

$$\hat{y}_{int} := \hat{y}_{int}(x; \mu, c_+^*, c_-^*),$$

where the thresholds $c_+^* \geq c_-^*$ are as below:

Case 1 (small zero class). *If $\pi_0 \leq \alpha^* \sqrt{\pi_- \pi_+}$, then the optimal classifier ignores the zero class in between. The thresholds are equal:*

$$c_+^* = c_-^* = \frac{\ln(\pi_-/\pi_+)}{2\left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+}$$

Optimal ℓ_2-robust three-class classifiers

Theorem

For $\varepsilon < \|\mu\|_2/2$, an optimal linear interval ℓ_2-robust classifier is

$$\hat{y}_{int} := \hat{y}_{int}(x; \mu, c_+^*, c_-^*),$$

where the thresholds $c_+^* \geq c_-^*$ are as below:

Case 1 (small zero class). *If $\pi_0 \leq \alpha^* \sqrt{\pi_- \pi_+}$, then the optimal classifier ignores the zero class in between. The thresholds are equal:*

$$c_+^* = c_-^* = \frac{\ln(\pi_-/\pi_+)}{2\left(1 - \frac{\varepsilon}{\|\mu\|_2}\right)_+}$$

Optimal ℓ_2-robust three-class classifiers

Theorem (ctd)

Case 2 (large zero class). *Otherwise,* $c_+^* - c_-^* > 2\varepsilon\|\mu\|_2$, *with*

$$c_\pm^* = \pm\frac{\|\mu\|_2^2}{2} \pm \frac{\ln(\pi_0/\pi_\pm)}{\left(1 - \frac{2\varepsilon}{\|\mu\|_2}\right)_+}$$

The cutoff α^* *is the unique solution to* ($\gamma := \sqrt{\pi_+/\pi_-}$):

$$(\gamma + \gamma^{-1}) R^*_{\text{rob}}\{\mu, \frac{\gamma}{\gamma + \gamma^{-1}}; \varepsilon\} \tag{1}$$

$$= (\gamma + \alpha) R^*_{\text{rob}}\{\frac{\mu}{2}, \frac{\gamma}{\gamma + \alpha}; \varepsilon\} + (\gamma^{-1} + \alpha) R^*_{\text{rob}}\{\frac{\mu}{2}, \frac{\gamma^{-1}}{\gamma^{-1} + \alpha}; \varepsilon\} - \alpha,$$

for $\alpha \geq \exp\{-(\|\mu\|_2 - 2\varepsilon)^2/2\}$.

Optimal ℓ_2-robust three-class classifiers

Theorem (ctd)

Case 2 (large zero class). *Otherwise,* $c_+^* - c_-^* > 2\varepsilon\|\mu\|_2$, *with*

$$c_\pm^* = \pm\frac{\|\mu\|_2^2}{2} \pm \frac{\ln(\pi_0/\pi_\pm)}{\left(1 - \frac{2\varepsilon}{\|\mu\|_2}\right)_+}$$

The cutoff α^* *is the unique solution to* ($\gamma := \sqrt{\pi_+/\pi_-}$):

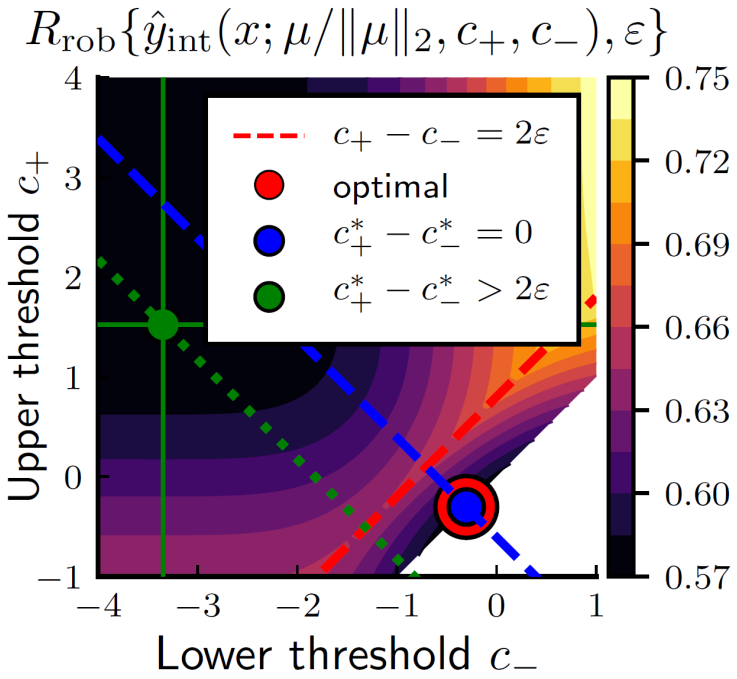
$$(\gamma + \gamma^{-1})R^*_{\text{rob}}\{\mu, \frac{\gamma}{\gamma + \gamma^{-1}}; \varepsilon\} \tag{1}$$

$$= (\gamma + \alpha)R^*_{\text{rob}}\{\frac{\mu}{2}, \frac{\gamma}{\gamma + \alpha}; \varepsilon\} + (\gamma^{-1} + \alpha)R^*_{\text{rob}}\{\frac{\mu}{2}, \frac{\gamma^{-1}}{\gamma^{-1} + \alpha}; \varepsilon\} - \alpha,$$

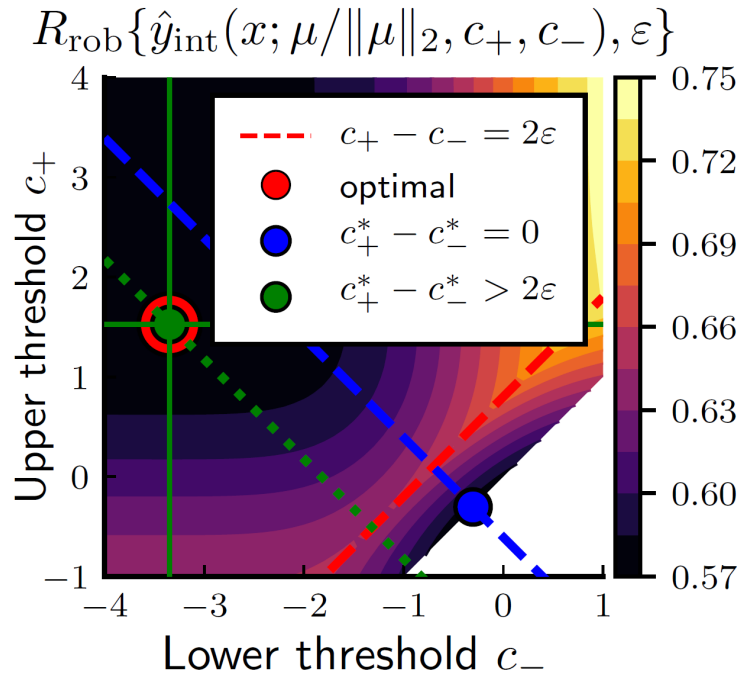
for $\alpha \geq \exp\{-(\|\mu\|_2 - 2\varepsilon)^2/2\}$.

Optimal ℓ_2-robust three-class classifiers: Discussion

- Class (im)balance crucial. Zero class:
  1. Small: thresholds coincide with two-class robust classif (-1 vs 1)
  2. Large: thresholds coincide with two-class robust classif (-1 vs 0), (0 vs 1)
- Discontinuity: optimal thresholds can be discontinuous in problem parameters



(b) $\pi_0 = 42.00\%$

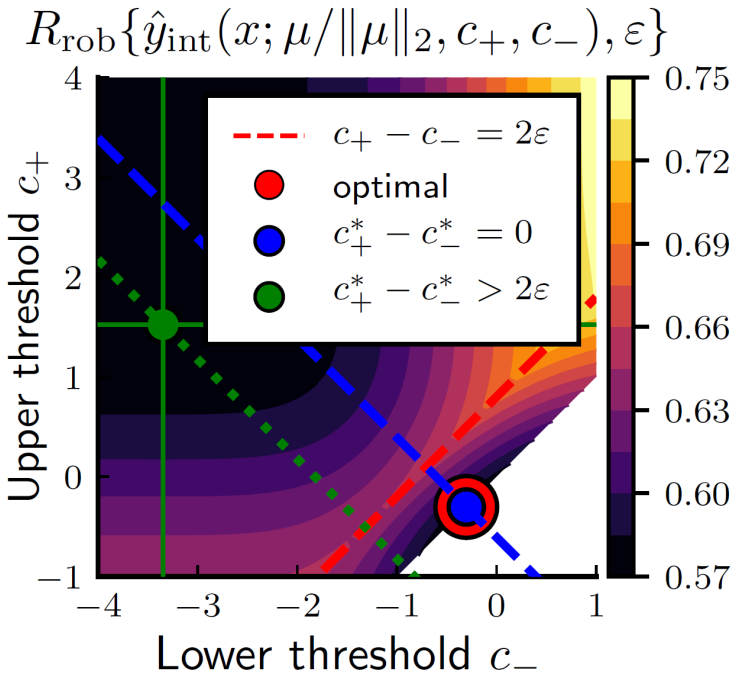


(c) $\pi_0 = 42.01\%$

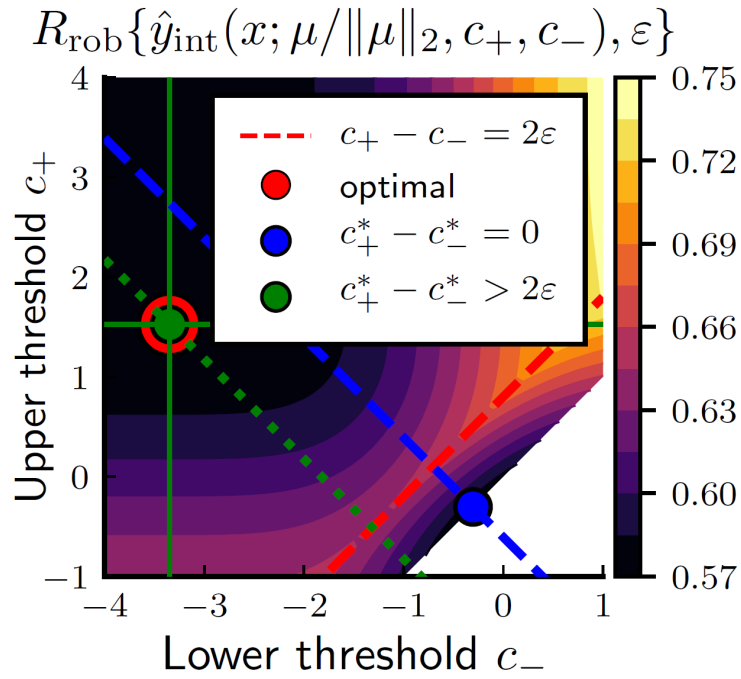
- $0 < c_+ - c_- < 2\varepsilon\|\mu\|_2$ can be improved by moving closer ($S_\pm$ overlap).

Optimal ℓ_2-robust three-class classifiers: Discussion

- Class (im)balance crucial. Zero class:
  1. Small: thresholds coincide with two-class robust classif (-1 vs 1)
  2. Large: thresholds coincide with two-class robust classif (-1 vs 0), (0 vs 1)
- Discontinuity: optimal thresholds can be discontinuous in problem parameters



(b) $\pi_0 = 42.00\%$



(c) $\pi_0 = 42.01\%$

- $0 < c_+ - c_- < 2\varepsilon\|\mu\|_2$ can be improved by moving closer ($S_\pm$ overlap).

Optimal ℓ_2-robust three-class classifiers: Discussion

- Tradeoff in most regimes
  1. Except possibly if extremal classes have equal weight, for small $\pi_0 \leq \exp(-\|\mu\|_2^2/2)\pi_+$
- Linear classifiers are admissible among ε-separated classif.

Optimal ℓ_2-robust three-class classifiers: Discussion

- Tradeoff in most regimes
  1. Except possibly if extremal classes have equal weight, for small $\pi_0 \leq \exp(-\|\mu\|_2^2/2)\pi_+$
- Linear classifiers are admissible among ε-separated classif.

Other results

- Thm: Opt. robust linear classifiers for ℓ_∞ perturbations are $\hat{y}^*(x) = \text{sign}\{x^\top \eta_\varepsilon(\mu) - q/2\}$, where η is soft-thresholding

$$\eta_\varepsilon(x) := \begin{cases} x - \varepsilon, & \text{if } x \geq \varepsilon, \\ 0, & \text{if } x \in (-\varepsilon, \varepsilon), \\ x + \varepsilon, & \text{if } x \leq -\varepsilon, \end{cases}$$

- Thm: For any norm-bounded perturbations, classification calibration for linear classif. and decreasing surrogate loss ℓ holds if
  1. ℓ is convex, or
  2. classes are balanced.
- Thm: Favorable finite-sample convergence rates of the robust empirical risk for certain geometric classifiers

Other results

- Thm: Opt. robust linear classifiers for ℓ_∞ perturbations are $\hat{y}^*(x) = \text{sign}\{x^\top \eta_\varepsilon(\mu) - q/2\}$, where η is soft-thresholding

$$\eta_\varepsilon(x) := \begin{cases} x - \varepsilon, & \text{if } x \geq \varepsilon, \\ 0, & \text{if } x \in (-\varepsilon, \varepsilon), \\ x + \varepsilon, & \text{if } x \leq -\varepsilon, \end{cases}$$

- Thm: For any norm-bounded perturbations, classification calibration for linear classif. and decreasing surrogate loss ℓ holds if
  1. ℓ is convex, or
  2. classes are balanced.
- Thm: Favorable finite-sample convergence rates of the robust empirical risk for certain geometric classifiers

Other results

- Thm: Opt. robust linear classifiers for ℓ_∞ perturbations are $\hat{y}^*(x) = \text{sign}\{x^\top \eta_\varepsilon(\mu) - q/2\}$, where η is soft-thresholding

$$\eta_\varepsilon(x) := \begin{cases} x - \varepsilon, & \text{if } x \geq \varepsilon, \\ 0, & \text{if } x \in (-\varepsilon, \varepsilon), \\ x + \varepsilon, & \text{if } x \leq -\varepsilon, \end{cases}$$

- Thm: For any norm-bounded perturbations, classification calibration for linear classif. and decreasing surrogate loss ℓ holds if
  1. ℓ is convex, or
  2. classes are balanced.
- Thm: Favorable finite-sample convergence rates of the robust empirical risk for certain geometric classifiers

Related works: Can only mention a few. See paper

- Adv. ex. are unavoidable
  1. Gilmer et al. (2019); Shafahi et al. (2019): bounds, sphere
  2. Dohmatob (2019): bounds, W_2 Talagrand transportation-cost ineq.
  3. Bhagoji et al. (2019): balanced; two Gaussians: symm pert. linear is optimal
  4. Pydi and Jog (2020): optimal transport, balanced, two spherical Gaussians
  5. Raghunathan et al. (2020): linear regression, geometry
- two-class Gaussian problems
  1. Schmidt et al. (2018): balanced, large SNR; specific classif have tradeoff
  2. Richardson and Weiss (2020) unequal cov; strong asymm, noise $\to 0$
  3. Tsipras et al. (2018) 1-sparse;
  4. Dan et al. (2020): general covariance; balanced classes; minimax excess risk.
- Other criteria:
  1. Wang et al. (2018): max astuteness $= 1-$ rob. risk $\iff$ r-optimality
  2. Fawzi et al. (2018): minimal perturbation length
- Other results: Montasser et al. (2019, 2020); Wang et al. (2019, 2020); Gao et al. (2019); Wu et al. (2020)...

Related works: Can only mention a few. See paper

- Adv. ex. are unavoidable
  1. Gilmer et al. (2019); Shafahi et al. (2019): bounds, sphere
  2. Dohmatob (2019): bounds, W_2 Talagrand transportation-cost ineq.
  3. Bhagoji et al. (2019): balanced; two Gaussians: symm pert. linear is optimal
  4. Pydi and Jog (2020): optimal transport, balanced, two spherical Gaussians
  5. Raghunathan et al. (2020): linear regression, geometry
- two-class Gaussian problems
  1. Schmidt et al. (2018): balanced, large SNR; specific classif have tradeoff
  2. Richardson and Weiss (2020) unequal cov; strong asymm, noise $\to 0$
  3. Tsipras et al. (2018) 1-sparse;
  4. Dan et al. (2020): general covariance; balanced classes; minimax excess risk.
- Other criteria:
  1. Wang et al. (2018): max astuteness $= 1-$ rob. risk $\iff$ r-optimality
  2. Fawzi et al. (2018): minimal perturbation length
- Other results: Montasser et al. (2019, 2020); Wang et al. (2019, 2020); Gao et al. (2019); Wu et al. (2020)...

Related works: Can only mention a few. See paper

- Adv. ex. are unavoidable
  1. Gilmer et al. (2019); Shafahi et al. (2019): bounds, sphere
  2. Dohmatob (2019): bounds, W_2 Talagrand transportation-cost ineq.
  3. Bhagoji et al. (2019): balanced; two Gaussians: symm pert. linear is optimal
  4. Pydi and Jog (2020): optimal transport, balanced, two spherical Gaussians
  5. Raghunathan et al. (2020): linear regression, geometry
- two-class Gaussian problems
  1. Schmidt et al. (2018): balanced, large SNR; specific classif have tradeoff
  2. Richardson and Weiss (2020) unequal cov; strong asymm, noise $\to 0$
  3. Tsipras et al. (2018) 1-sparse;
  4. Dan et al. (2020): general covariance; balanced classes; minimax excess risk.
- Other criteria:
  1. Wang et al. (2018): max astuteness $= 1-$ rob. risk $\iff$ r-optimality
  2. Fawzi et al. (2018): minimal perturbation length
- Other results: Montasser et al. (2019, 2020); Wang et al. (2019, 2020); Gao et al. (2019); Wu et al. (2020)...

Related works: Can only mention a few. See paper

- Adv. ex. are unavoidable
  1. Gilmer et al. (2019); Shafahi et al. (2019): bounds, sphere
  2. Dohmatob (2019): bounds, W_2 Talagrand transportation-cost ineq.
  3. Bhagoji et al. (2019): balanced; two Gaussians: symm pert. linear is optimal
  4. Pydi and Jog (2020): optimal transport, balanced, two spherical Gaussians
  5. Raghunathan et al. (2020): linear regression, geometry
- two-class Gaussian problems
  1. Schmidt et al. (2018): balanced, large SNR; specific classif have tradeoff
  2. Richardson and Weiss (2020) unequal cov; strong asymm, noise $\to 0$
  3. Tsipras et al. (2018) 1-sparse;
  4. Dan et al. (2020): general covariance; balanced classes; minimax excess risk.
- Other criteria:
  1. Wang et al. (2018): max astuteness $= 1-$ rob. risk $\iff$ r-optimality
  2. Fawzi et al. (2018): minimal perturbation length
- Other results: Montasser et al. (2019, 2020); Wang et al. (2019, 2020); Gao et al. (2019); Wu et al. (2020)...

Summary

- Adversarial robustness: active, important area
- Multi-class Gaussian models:
  1. Bayes-optimal classifiers (new proof techniques)
  2. Implications: tradeoffs, imbalance
  3. Classification calibration, Conv. rates
- Many problems remain
- slides at: `github.com/dobriban/talks`
- Thanks!

References I

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pages 7496–7508, 2019.

Christer Borell. The brunn-minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975.

Andrea Cianchi, Nicola Fusco, Francesco Maggi, and Aldo Pratelli. On the isoperimetric deficit in gauss space. *American Journal of Mathematics*, pages 131–186, 2011.

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guarantees for adversarially robust gaussian classification. In *ICML*, 2020.

Elvis Dohmatob. Limitations of adversarial robustness: Strong no free lunch theorem. In *ICML*, 2019.

References II

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems*, pages 13029–13040, 2019.

Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2280–2289, Long Beach, California, USA, 09–15 June 2019. PMLR. URL `http://proceedings.mlr.press/v97/gilmer19a.html`.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

References III

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. *arXiv preprint arXiv:2005.07652*, 2020.

Elchanan Mossel and Joe Neeman. Robust dimension free isoperimetry in gaussian space. *The Annals of Probability*, 43(3):971–991, 2015.

Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. In *ICML*, 2020.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Eitan Richardson and Yair Weiss. A bayes-optimal view on adversarial examples. *arXiv preprint arXiv:2002.08859*, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

References IV

Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1lWUoA9FQ`.

Vladimir N Sudakov and Boris S Tsirel'son. Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 9(1):9–18, 1978.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, volume 1, page 2, 2019.

References V

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rklOg6EFwS`.

Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning*, pages 5133–5142, 2018.

Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Does network width really help adversarial robustness? *arXiv preprint arXiv:2010.01279*, 2020.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Sk9yuql0Z`.