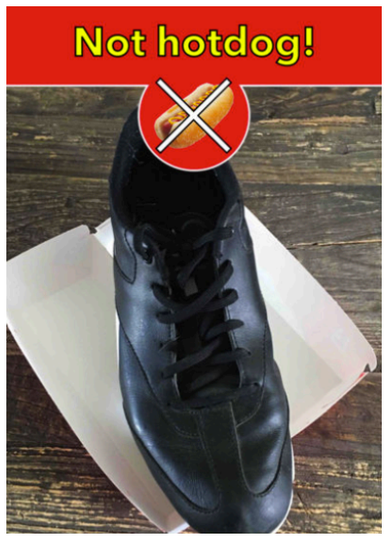
Lec 21: Adversarial Robustness

Yaoliang Yu

July 21, 2020

Supervised Learning



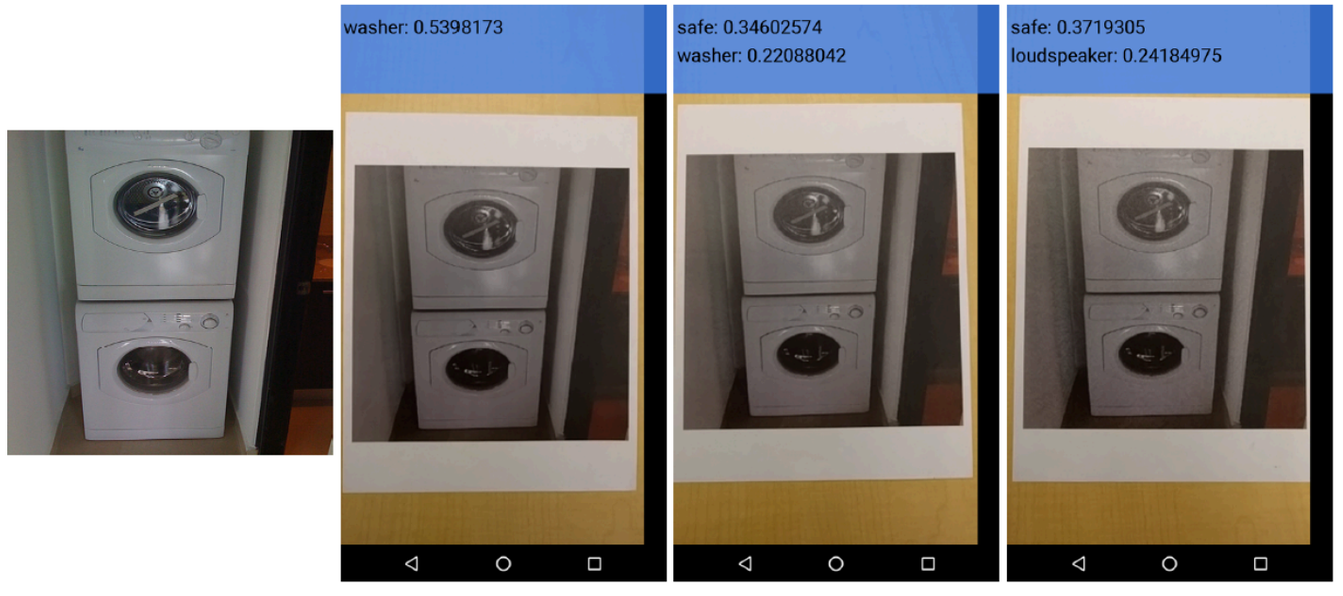
Formally

- Given a training set of pairs of examples $(\mathbf{x}_i, y_i) \in \mathsf{X} \times \mathsf{Y}$
- Return a function (classifier) $f : \mathsf{X} \rightarrow \mathsf{Y}$
- On an unseen test example x, output f(x)
- The goal is to do well on unseen test data
  - usually do not care about performance on training set

Performance Metric

- Accuracy (top-1, top-10 error, precision, recall, etc.)
- Training time
- Memory
- Test time
- Robustness
- Privacy
- Fairness
- Interpretability

And then the surprise



Szegedy et al. Intriguing properties of neural networks, ICLR 2014
Athalye et al. Synthesizing Robust Adversarial Examples, ICML 2018
Kurakin, Goodfellow, Bengio. Adversarial Examples in the Physical World, ICLR workshop 2017

Why should we care?

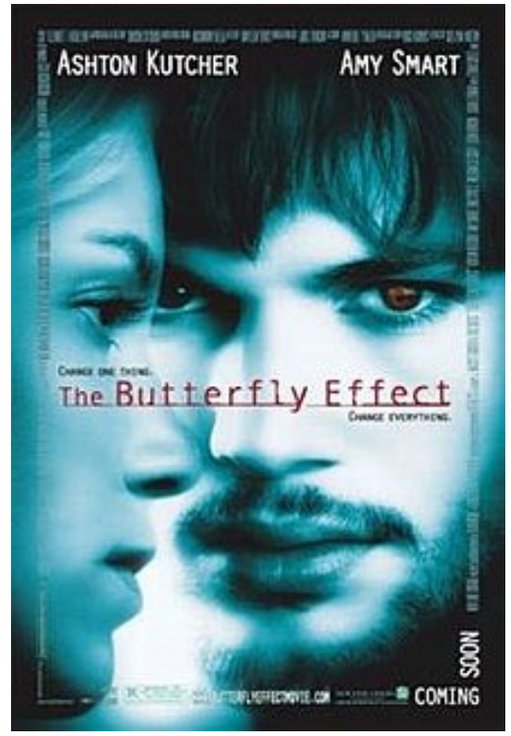
Goodfellow, P. McDaniel and N. Papernot. Making machine learning robust against adversarial inputs, CACM (2018)
Gilmer et al. Motivating the Rules of the Game for Adversarial Example Research, arXiv:1807.06732 (2018)

Formally

- Exist small $\Delta\mathbf{x}$ such that $f(\mathbf{x} + \Delta x) \neq y(\mathbf{x})$
- Practically, exist small $\Delta\mathbf{x}$ such that $f(\mathbf{x} + \Delta x) \neq f(\mathbf{x})$
  - similar if f is very accurate
  - such examples **x** are called “adversarial”
- Intuitive explanation:

  f is not sufficiently smooth (continuous)

  - Or in fancier words, f is not robust



WORLD MAP

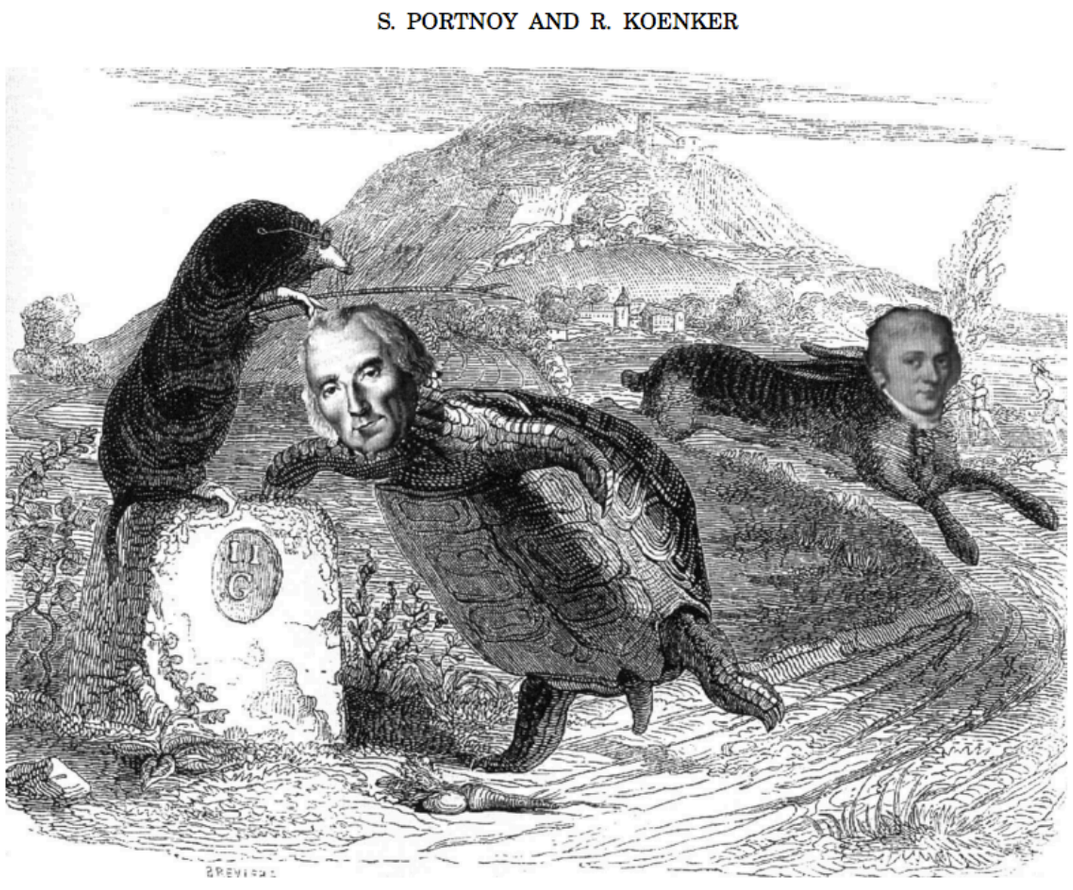
“Adversarial” examples exist for any (non-constant) classifier

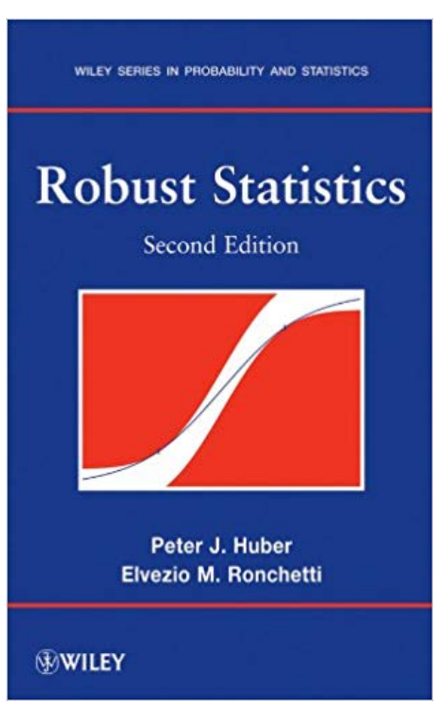
Existence is not surprising; universality is

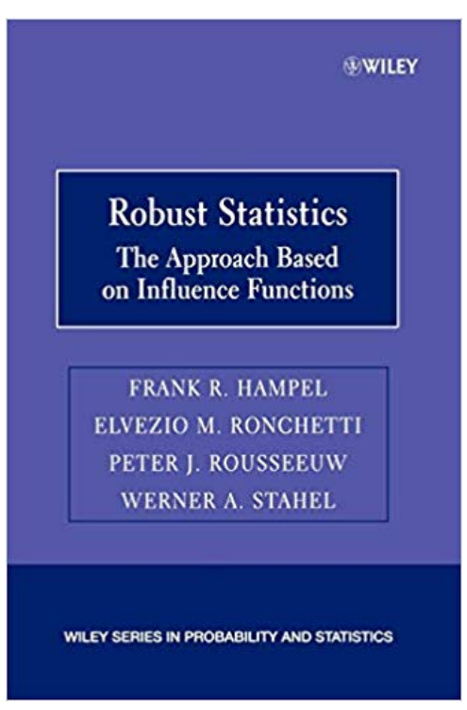
Robustness is not new

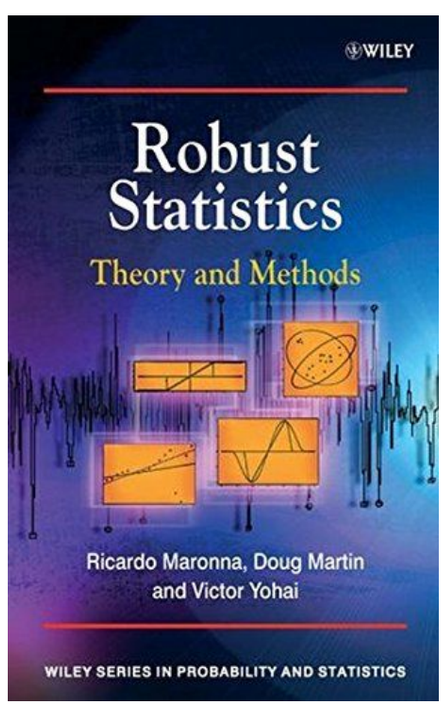
least-squares vs least absolute deviation

$$\|X\mathbf{w} - \mathbf{y}\|_2 \quad \textit{VS.} \quad \|X\mathbf{w} - \mathbf{y}\|_1$$

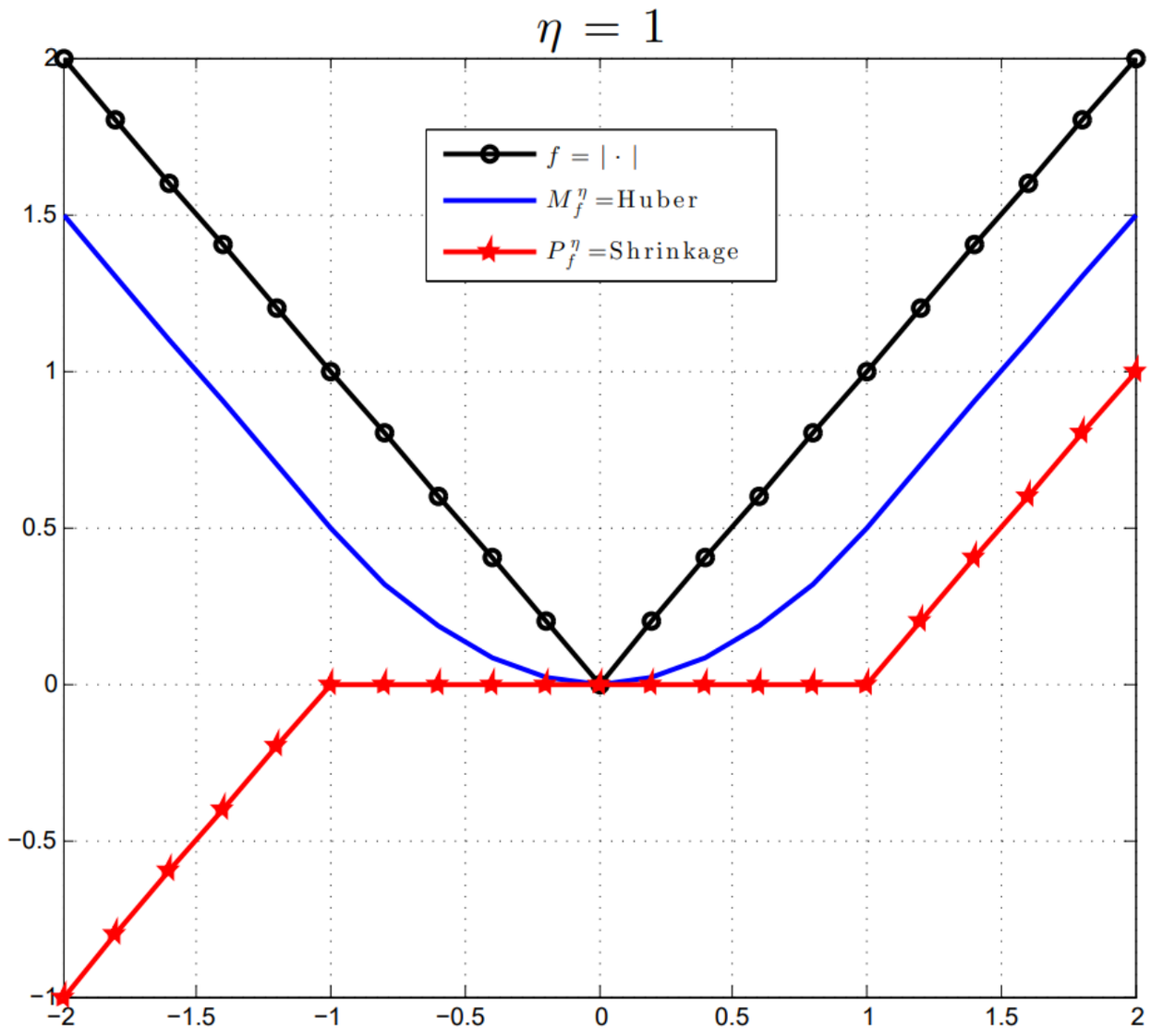


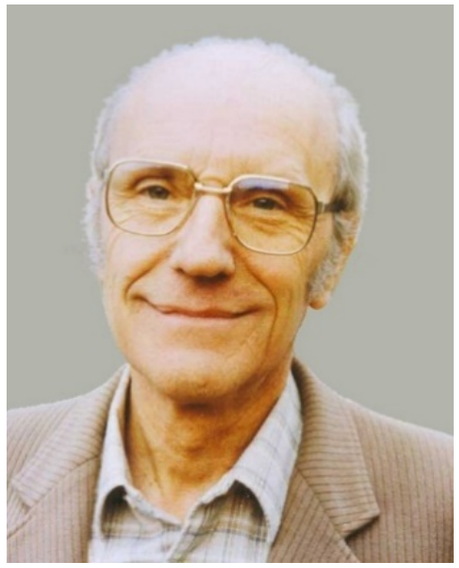






Huber's loss is Moreau's envelope





$$\mathsf{M}_{|\cdot|}^{\eta}(t) = \min_s \tfrac{1}{2}(s-t)^2 + \eta|s|$$

$$= \begin{cases} \frac{1}{2}t^2, & |t| \le \eta \\ \eta|t| - \frac{1}{2}\eta^2, & |t| \ge \eta \end{cases}$$

Peter J. Huber. Robust estimation of a location parameter, Annals of Statistics (1964)
Jean J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. C.R.A.S. (1962)

Huber's Contamination

observation distr.

contamination proportion

true distr.

contamination distr. (arbitrary)

$$F = (1-\epsilon)G + \epsilon H$$

mixture model

$$D_{\text{tv}}(F\|G) \le \epsilon$$

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \sim F$$

- with probability (w.p.) ϵ, $\mathbf{x}_i$ is from contamination H
- and w.p. 1-ϵ, $\mathbf{x}_i$ is from true distribution
- so roughly ϵ proportion of training set is (arbitrarily) contaminated
- difficulty lies in don't know which data example is authentic or not

Score classifier

- A score function (classifier) $f : \mathrm{X} \rightarrow \mathbb{R}$
- Input universe X is usually subset of $\mathbb{R}^d$
- Intuitively, small perturbation on x should result in small perturbation of f(x)
- Lipschitz continuity:

Lipschitz constant

$$|f(\mathbf{x}) - f(\mathbf{z})| \leq L \cdot \|\mathbf{x} - \mathbf{z}\|$$

$$\|\nabla f\| \leq L$$

Attack Algorithms

adversarial example

true label

$$\max_{\|\mathbf{x}-\mathbf{x}_0\|\leq\epsilon} f(\mathbf{x}, y; \mathbf{w})$$

network weights, fixed

attack model

input image

perturbation budget

(indistinguishability)

Fast Gradient Sign Method (FGSM)

$$\mathbf{x} \leftarrow \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} f(\mathbf{x}, y; \mathbf{w}))$$

$$\mathbf{x} \leftarrow \text{Proj}(\mathbf{x})$$

Projected Gradient Method (PGM)

$$\mathbf{x} \leftarrow \mathbf{x} + \eta \cdot \nabla_{\mathbf{x}} f(\mathbf{x}, y; \mathbf{w})$$

$$\mathbf{x} \leftarrow \text{Proj}(\mathbf{x})$$

Goodfellow, Shlens, Szegedy. Explaining and Harnessing Adversarial Examples, ICLR (2015)

Targeted Attack

adversarial example

target label

$$\min_{\|\mathbf{x}-\mathbf{x}_0\|\leq\epsilon} f(\mathbf{x}, y_{\text{target}}; \mathbf{w})$$

network weights, fixed

attack model

input image

perturbation budget

(indistinguishability)

Fast Gradient Sign Method (FGSM)	Projected Gradient Method (PGM)
$\mathbf{x} \leftarrow \mathbf{x} - \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} f(\mathbf{x}, y_{\text{target}}; \mathbf{w}))$	$\mathbf{x} \leftarrow \mathbf{x} - \eta \cdot \nabla_{\mathbf{x}} f(\mathbf{x}, y_{\text{target}}; \mathbf{w})$
$\mathbf{x} \leftarrow \text{Proj}(\mathbf{x})$	$\mathbf{x} \leftarrow \text{Proj}(\mathbf{x})$

Lipschitz regularization

$$\mathscr{F}_\epsilon = \{F : \mathsf{W}(F\|G) \le \epsilon\}$$

$$\max_{f:\mathsf{Lip}(f)\le 1} \mathbb{E}_{X\sim F} f(X) - \mathbb{E}_{Z\sim G} f(Z)$$

$$\min_{\mathbf{w}} \max_{F\in\mathscr{F}_\epsilon} \mathbb{E}\ell(\mathbf{w}^\top X), \quad X \sim F$$

$$\Vert$$

$$\min_{\gamma\ge 0} \gamma\epsilon^p - \mathbb{E}\mathsf{M}^\gamma_{-\rho}(X;\mathbf{w}), \quad \rho(\mathbf{x}) = \ell(\mathbf{w}^\top\mathbf{x}), \quad X\sim G$$

$$\Vert\!\Lambda$$

$$\mathbb{E}\ell(\mathbf{w}^\top\mathbf{x}) + \epsilon\cdot\mathsf{Lip}(\ell_{\mathbf{w}}), \quad X\sim G$$

P. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric, Mathematical Programming (2018)
Sinha, H. Namkoong, J. Duchi. Certifying some distributional robustness with principled adversarial training, ICLR (2018)
Z. Cranko, S. Kornblith, Z. Shi and R. Nock. Lipschitz Networks and Distributional Robustness, arXiv:1809.01129 (2018)

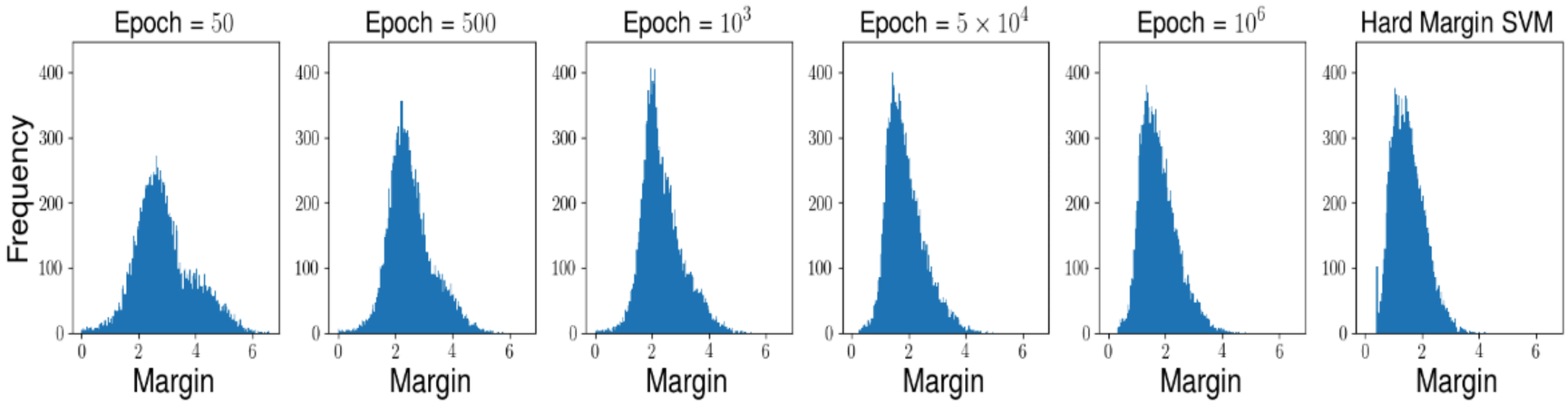
Margin story

$$\mathrm{m}(\mathbf{x}, y; \{\mathsf{F}_k\}) := \mathrm{sign}(\hat{y}(\mathbf{x}), y) \cdot \mathrm{d}(\mathbf{x}, \partial F_{\hat{y}})$$

Theorem 1 (Soudry et al. 2018 [3]) For almost all linearly separable binary datasets and any smooth decreasing loss with an exponential tail, **gradient descent** with small constant step size and any starting point $\mathbf{w}_0$ converges to the (unique) solution $\widehat{\mathbf{w}}$ of hard-margin SVM, *i.e.*

$$\lim_{t\to\infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} = \frac{\widehat{\mathbf{w}}}{\|\widehat{\mathbf{w}}\|}.$$

■

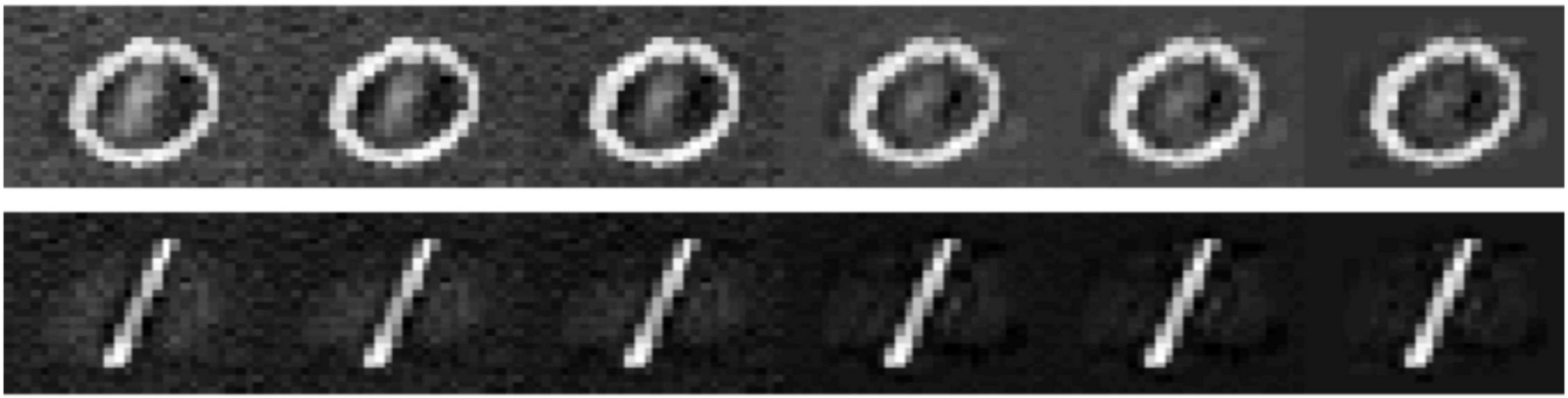


D. Soudry, E. Hoffer and N. Srebro. The Implicit Bias of Gradient Descent on Separable Data, ICLR (2018)
[WY]. Understanding Adversarial Robustness: The Trade-off between Minimum and Average Margin, (2019)

Binary linear classifiers

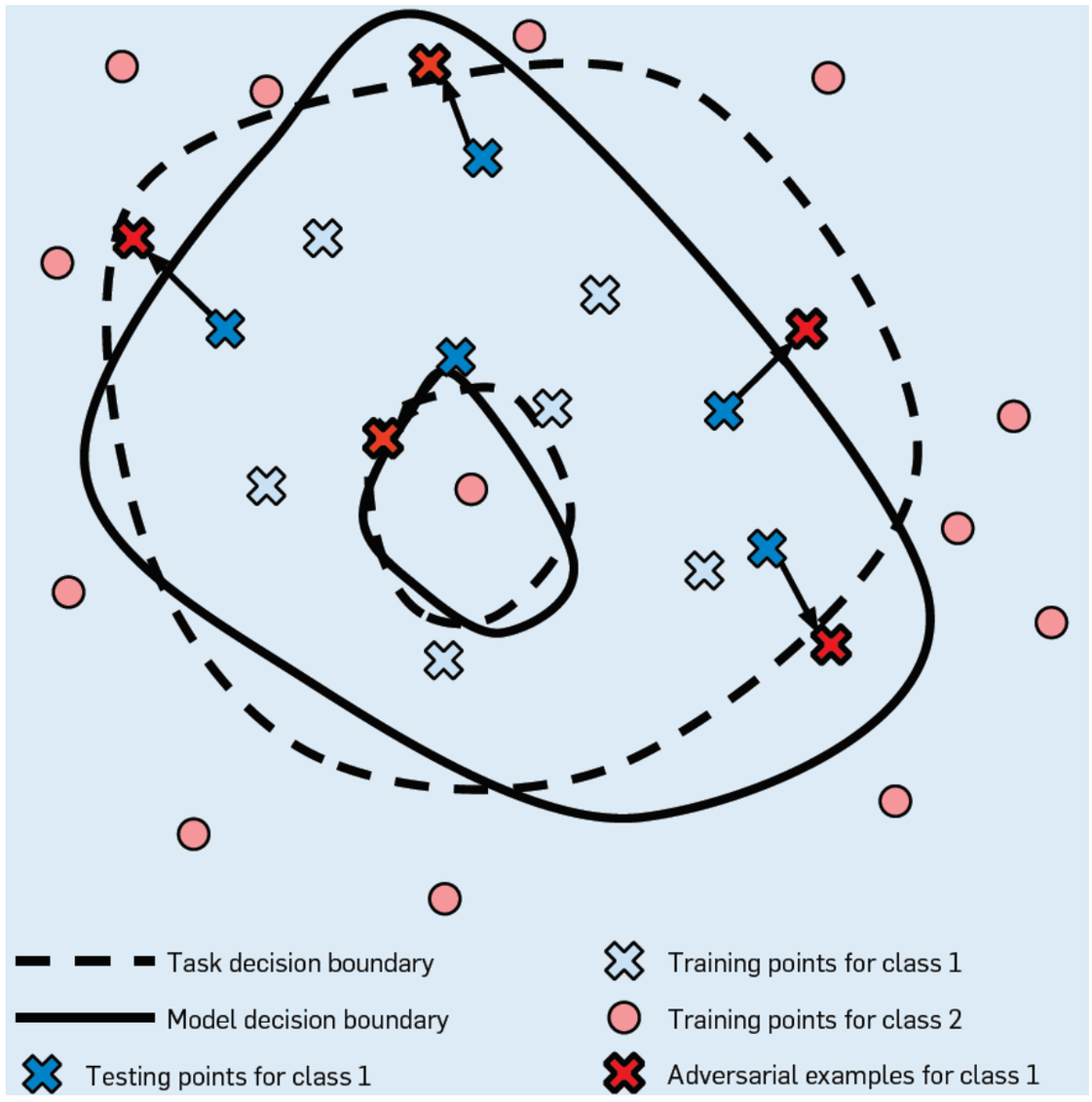
For a binary linear classifier which predicts positive if $\mathbf{w}^\top \mathbf{x} > 0$, can construct

$$\mathbf{x}^{\text{adv}} = \mathbf{x} - \frac{\mathbf{w}^\top \mathbf{x}}{\|\mathbf{w}\|^2}\mathbf{w}$$



[WY]. Understanding Adversarial Robustness: The Trade-off between Minimum and Average Margin, (2019)

The deep challenge



Deep Learning

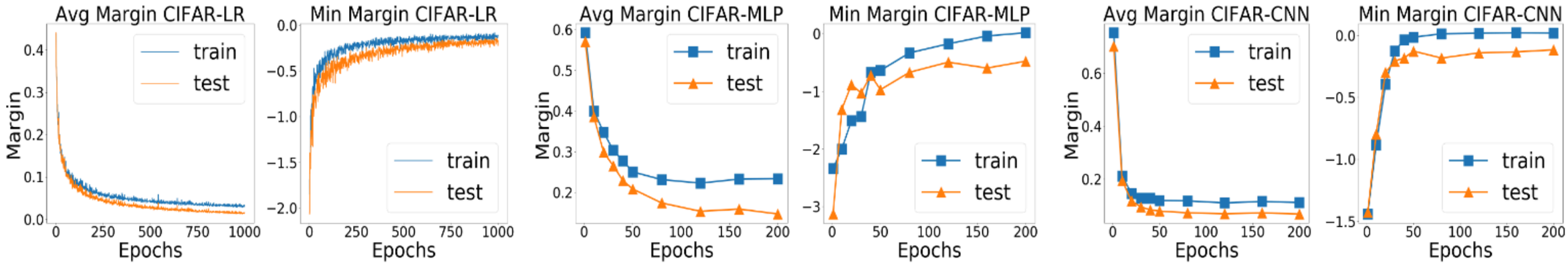
$$\mathbf{x} \xrightarrow{\varphi} \varphi(\mathbf{x}) \xrightarrow{\mathbf{w}} \mathbf{w}^\top \varphi(\mathbf{x}) =: \hat{y}$$

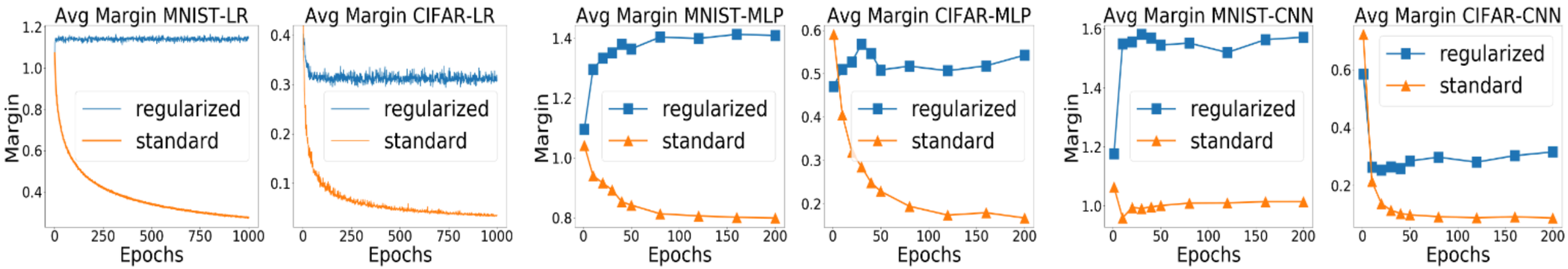
$$\min_{\mathbf{w},\varphi} \quad \ell(y, \hat{y})$$

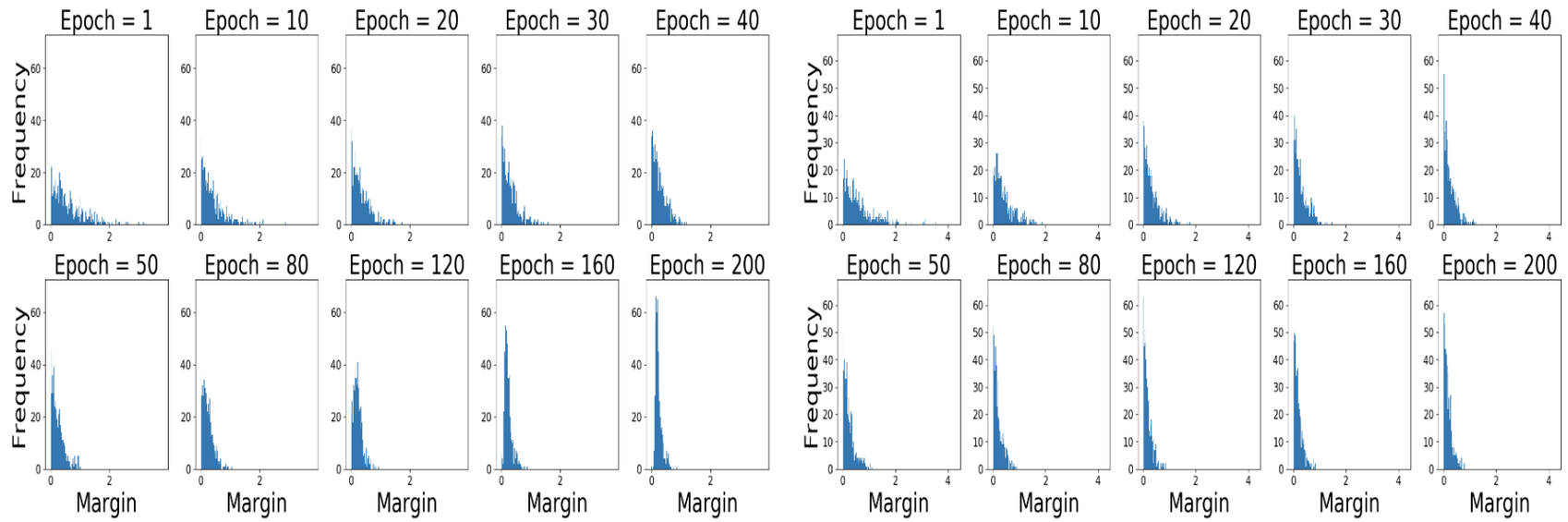
- nonlinear transformation φ
- linear classifier w
- trained jointly by SGD

I. Goodfellow, P. McDaniel and N. Papernot. Making machine learning robust against adversarial inputs, CACM (2018)

Observation & Solution?



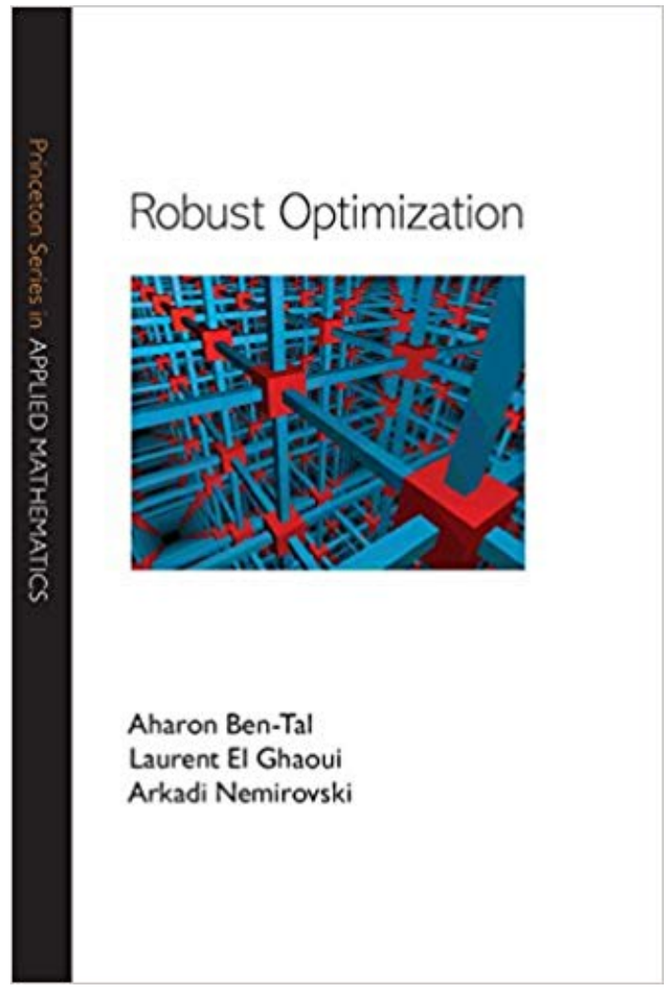




[WY]. Understanding Adversarial Robustness: The Trade-off between Minimum and Average Margin, (2019)

Adversarial training

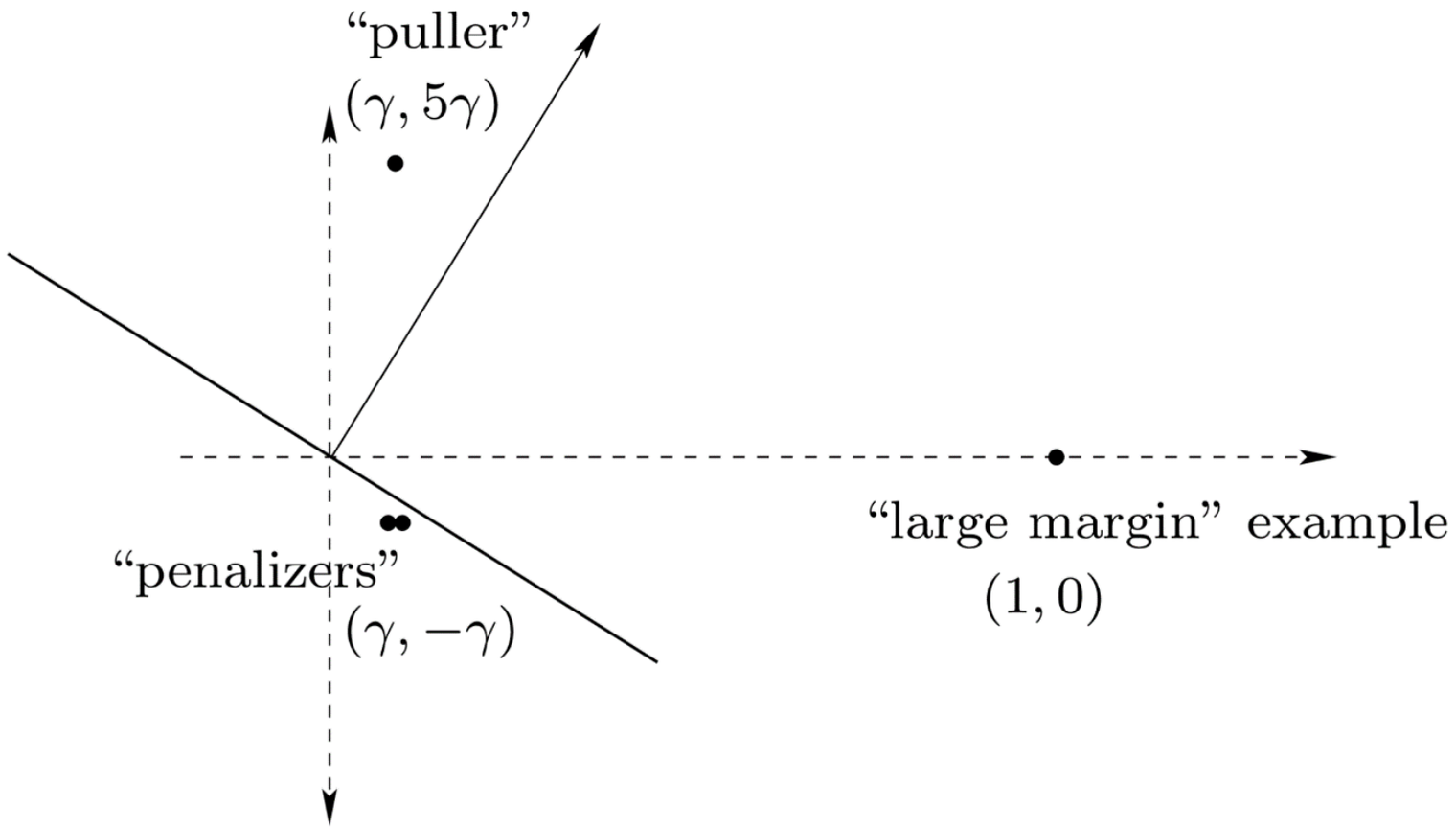
$$\min_{\mathbf{w}} \mathbf{E}\left[\underbrace{\max_{\|\Delta\mathbf{x}\|\leq\epsilon} \ell(\mathbf{x}+\Delta\mathbf{x};\mathbf{w})}_{\bar{\ell}_\epsilon(\mathbf{x};\mathbf{w})}\right]$$



- Amounts to changing the loss…
- One of the best defensive mechanisms
- Min-max formulation
- Inner max solved by an attack algorithm

Shaham, Yamada, Negahban. Understanding adversarial training: Increasing local stability of neural nets through robust optimization, Neurocomputing (2018)
Madry et al. Towards Deep Learning Model Resisstant to Adversarial Attacks, ICLR (2018)

A very negative result



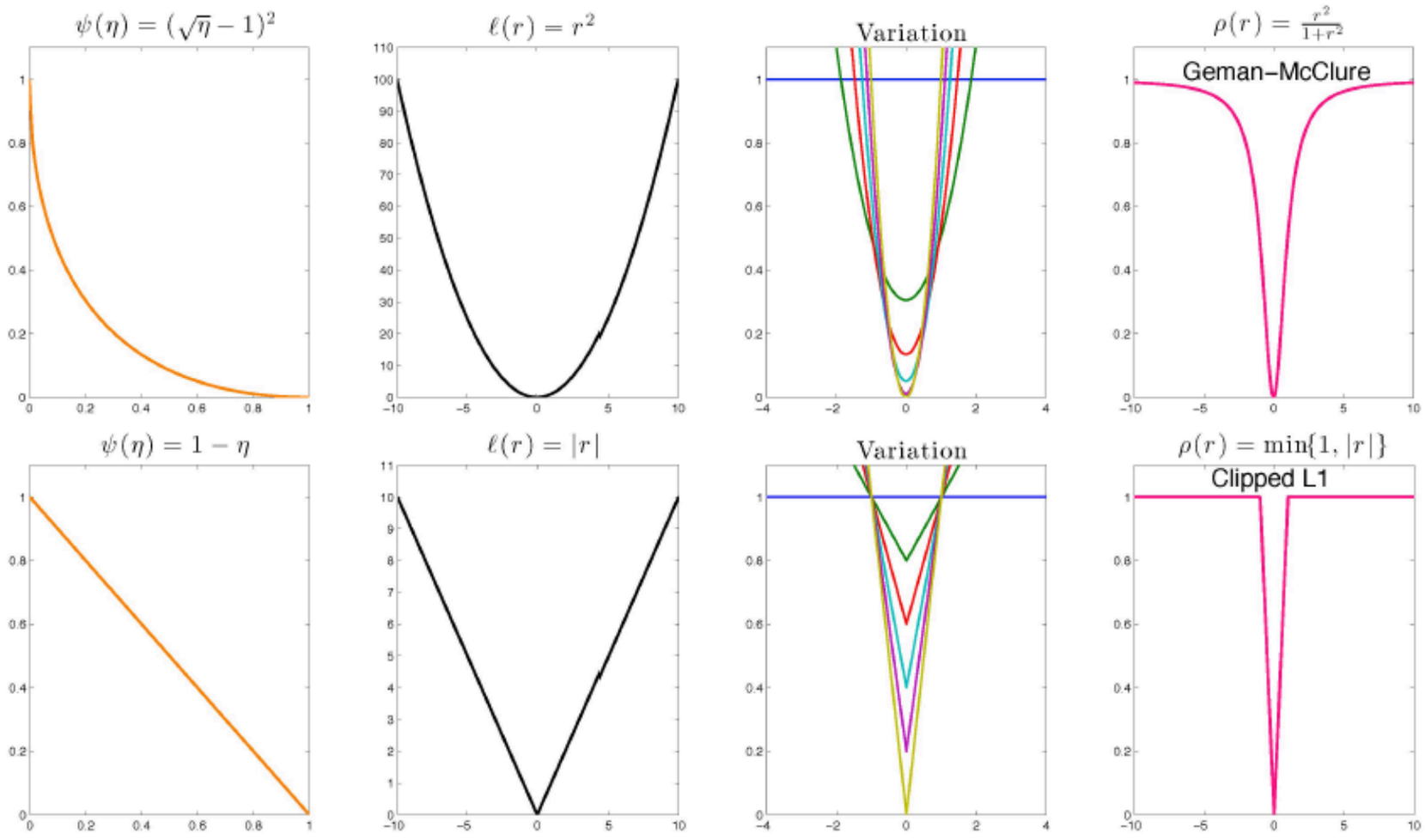
Theorem

All convex potential function based boosters can not tolerate random classification noise at rate $\eta \in (0, 1/2)$.

P. Long and R. Servedio. Random classification noise defeats all convex potential boosters, Machine Learning (2010)
van Rooyen, Menon, Williamson. Learning with symmetric label noise: The importance of being unhinged, NeurIPS (2015)

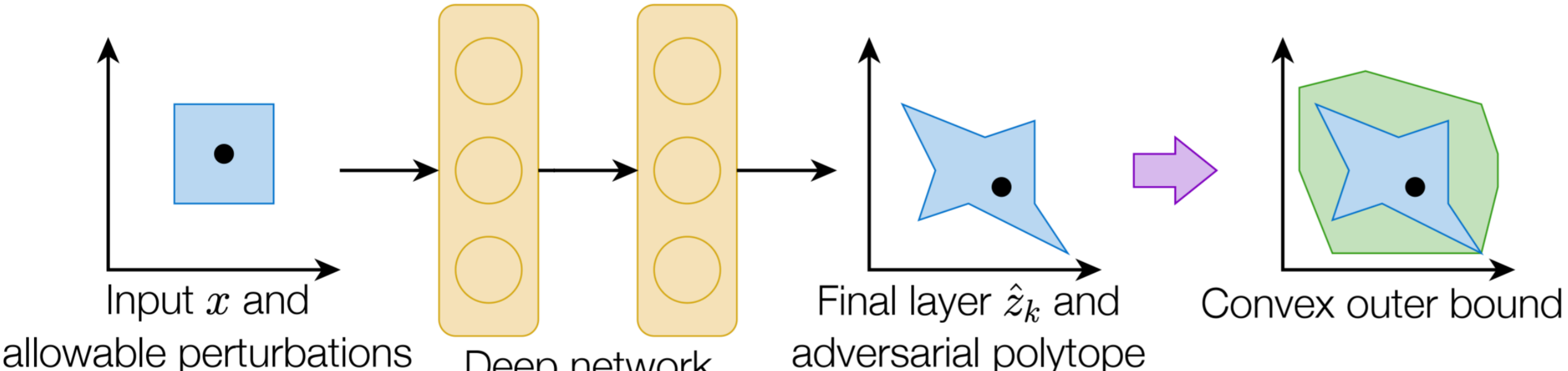
Variational loss

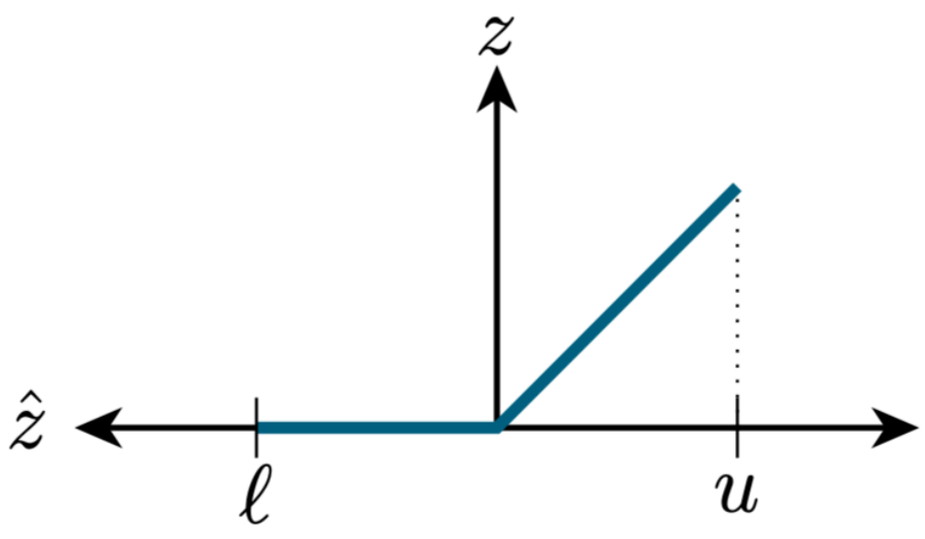
$$r(t) = \min_{0 \le \eta \le 1} \eta \ell(t) + \psi(\eta)$$



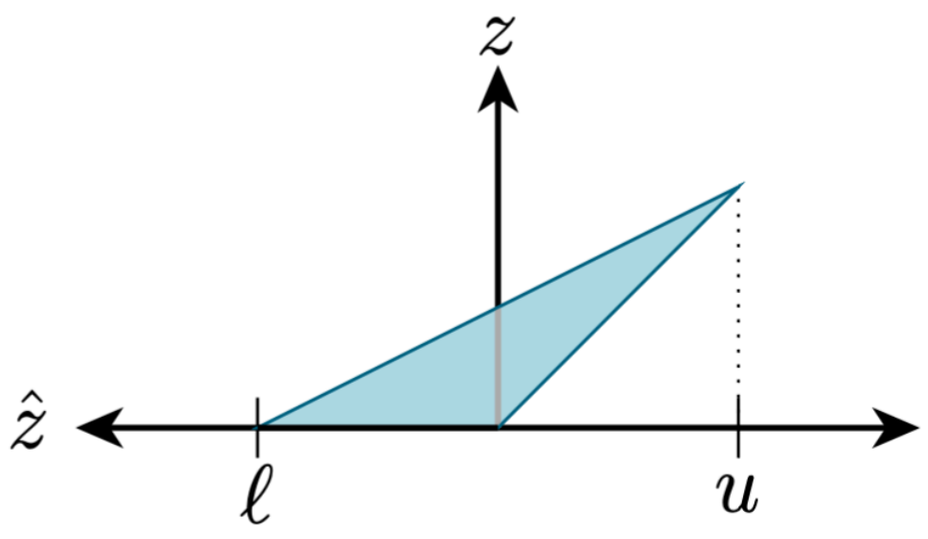
M. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision, IJCV (1996)
[YAS]. A Polynomial-time Form of Robust Regression, NeurIPS (2012)

Certification





Bounded ReLU set



Convex relaxation

Wong and Kolter. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope, ICML (2018)