# Revisiting Adversarial Risk

Arun Sai Suggala, Adarsh Prasad, Vaishnavh Nagarajan, Pradeep Ravikumar

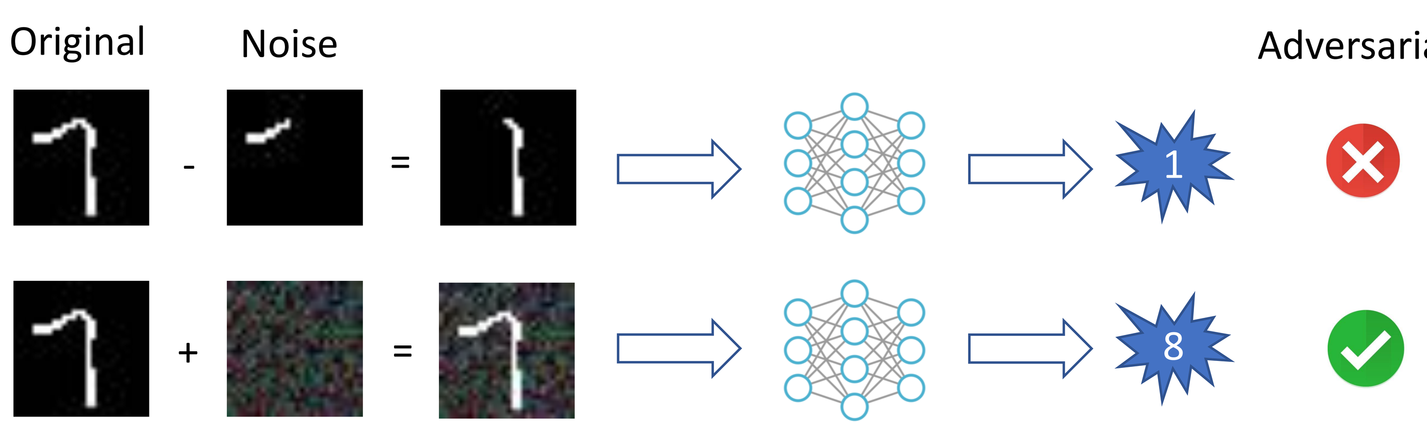
Carnegie Mellon University

## Abstract

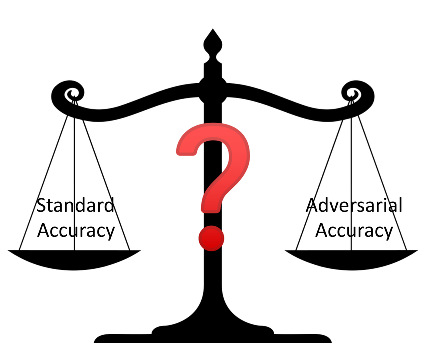
- **Motivation:** Existing definition of adversarial risk is not accurate.
  - Assumes the true label doesn't change after perturbation.
  - Resulted in counter-intuitive claims about adversarial risk.
- **Contributions:** Study a new definition of adversarial risk which is more accurate
  - Incorporates **perceptual similarity**
- **No trade-off** between standard risk and the more accurate notion of adversarial risk.
- Understand conditions under which existing definition of adversarial risk is accurate
  - Existing adversarial risk is **equivalent** to the new definition when the data has **margin**.
- When the data doesn't have a margin, adversarial training using *existing definition* can result in **loss of standard accuracy**.

## Motivation

- Need for incorporation of **perceptual similarity** in the definition of adversarial perturbation




- Counterintuitive conclusions using existing definition of adversarial risk




## Setup

- Binary classification: features $\mathbf{x}$, label $y \in \{-1, 1\}$, classifier $f$.
- Standard risk

$$R(f) = \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\ell(f(\mathbf{x}), y)\right]$$

- Existing adversarial risk

$$G_{\text{adv}}(f) = \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|\leq\epsilon} \ell(f(\mathbf{x}+\boldsymbol{\delta}), y)\right]$$

## New Adversarial Risk

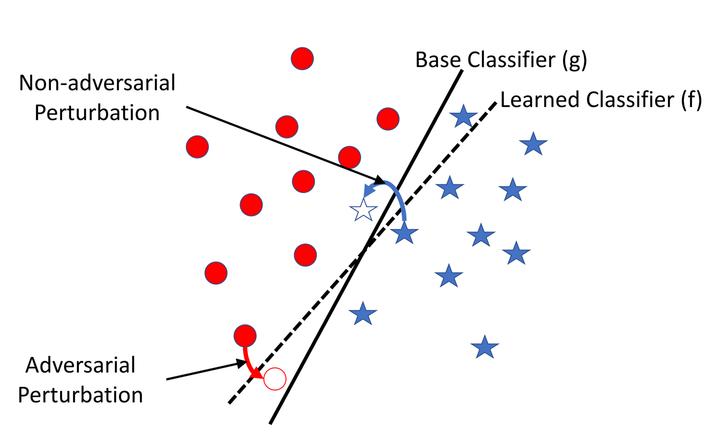
- Measure robustness of any classifier with respect to a **base classifier**.
  - Base classifier is a human classifier in many tasks.
  - Captures the human notion of perceptual similarity in image classification tasks.

**Definition 1 (Adversarial Perturbation)** *Let $g$ be the base classifier. Then the perturbation $\boldsymbol{\delta}_{\mathbf{x}}$ at $\mathbf{x}$ is adversarial for a classifier $f$, w.r.t base classifier $g$, if $\|\boldsymbol{\delta}_{\mathbf{x}}\| \leq \epsilon$ and*

$$f(\mathbf{x}) = g(\mathbf{x}), \quad g(\mathbf{x}) = g(\mathbf{x}+\boldsymbol{\delta}_{\mathbf{x}}),$$

*and*

$$f(\mathbf{x}+\boldsymbol{\delta}_{\mathbf{x}}) \neq g(\mathbf{x}).$$




**Definition 2 (Adversarial Risk)** *The adversarial risk of a classifier $f$ w.r.t base classifier $g$ is the fraction of points which can be adversarially perturbed*

$$R_{adv}(f) = \mathbb{E}\left[\max_{\substack{\|\boldsymbol{\delta}\|\leq\epsilon \\ g(\mathbf{x})=g(\mathbf{x}+\boldsymbol{\delta})}} \ell\left(f(\mathbf{x}+\boldsymbol{\delta}), g(\mathbf{x})\right) - \ell\left(f(\mathbf{x}), g(\mathbf{x})\right)\right].$$

## Adversarial Training

- A robust classifier can be obtained by minimizing the following joint objective

$$\operatorname*{argmin}_{f\in\mathcal{F}} R(f) + \lambda R_{\text{adv}}(f).$$

- The following Theorem shows there is *no trade-off* between standard and adversarial risks.

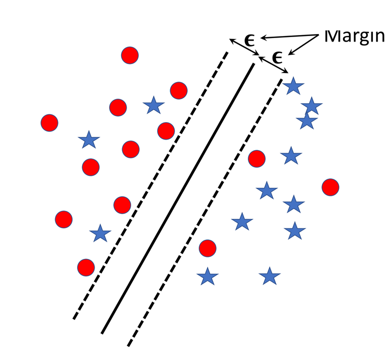
**Theorem 1 (Main Result)** *Suppose the hypothesis class $\mathcal{F}$ is the set of all measurable functions. Let the base classifier $g$ be a Bayes optimal classifier. Then any minimizer of*

$$\min_{f\in\mathcal{F}} R(f) + \lambda R_{adv}(f),$$

*is also a minimizer of standard risk.*

## Relation to Existing Adversarial Risk

- **When is the existing definition accurate?**
  - If the data has *margin*, existing definition is equivalent to the new definition
  - Or else, they are not equivalent.
- *Trade-off* between adversarial and standard risks, if data has no margin.

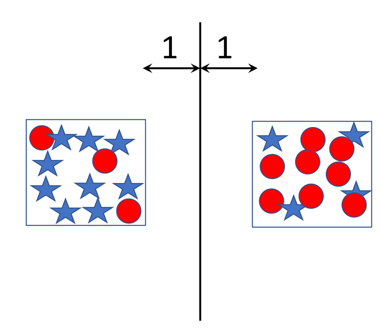


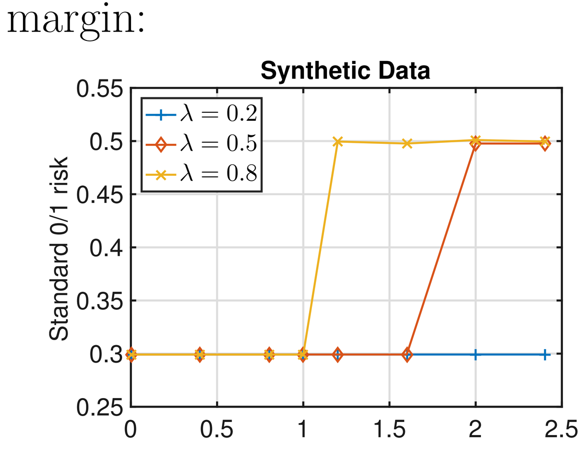

**Theorem 2 (Informal)** *Suppose the hypothesis class $\mathcal{F}$ is the set of all measurable functions. Then any minimizer of*

$$\min_{f\in\mathcal{F}} R(f) + \lambda G_{adv}(f)$$

*for any $\lambda \geq 0$, is also a minimizer of standard risk* ***iff*** *the data has margin.*

- A simple example illustrating importance of margin:



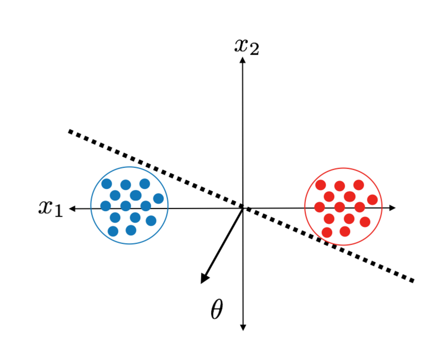

## Importance of Adversarial Training

- Theoream 1 shows that the minimizers of adversarial training objective are also the minimizers of standard risk.
- **Question: Do we really need to perform adversarial training?**

**Yes!!**

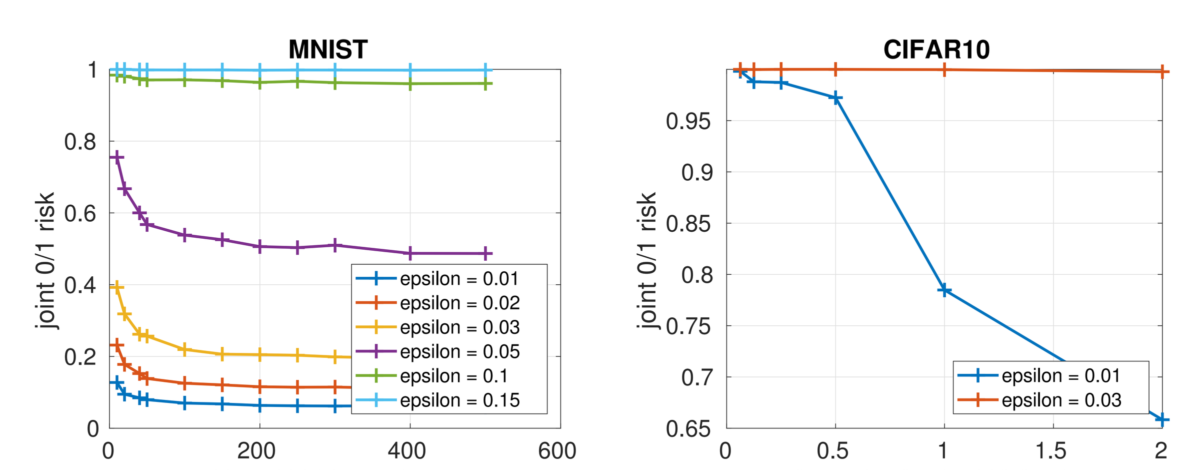
A simple example:

- Data is separable and lies in a low dimensional space.
- There exist classifiers with 0 standard risk but with very high adversarial risk.




## Robustness of Complex Models

- Use insights from Theorem 1 to explain an interesting practical phenomenon.
- Standard training with increasing model complexity can result in more robust models.




- Adversarial training with increasing model complexity can result in more accurate models.