The Many Faces of Adversarial Risk

Muni Sreenivas Pydi
ECE, University of Wisconsin-Madison

Varun Jog
DPMMS, University of Cambridge

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

UNIVERSITY OF
CAMBRIDGE

Summary

- We explore the “many faces” of adversarial risk and optimal adversarial risk, which measure the robustness of algorithms to adversarial perturbations.
- Our contributions:
  - A rigorous foundation for adversarial risk, fixing the issues of measurability
  - Equivalences between various definitions of adversarial risk
    - Equivalence between adversarial robustness and robust hypothesis testing with ∞- Wasserstein uncertainty sets
  - Various characterizations of optimal adversarial risk based on:
    - Optimal transport
    - Distributionally robust optimization
    - Game theory
  - Existence of a Nash equilibrium in game between adversary and algorithm.

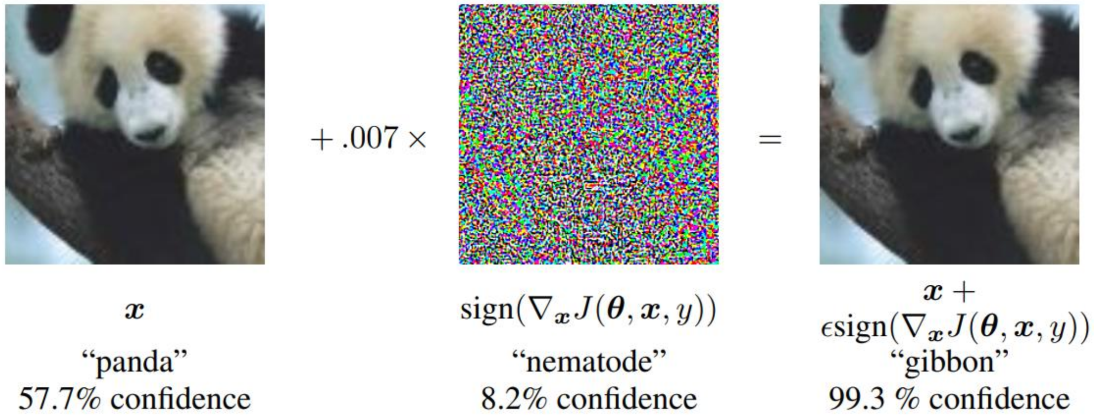
Adversarial Attacks

Perturbed data point

Maximize loss at x’

$$x \mapsto x' \in \underset{d(x,x') \le \epsilon}{\operatorname{argmax}}\, \ell((x',y),w).$$

Budget constraint: Perturbation is “small”



Source: Goodfellow et al. ICLR 2015

01 Apr 2019 | 16:56 GMT

Three Small Stickers in Intersection Can Cause Tesla Autopilot to Swerve Into Wrong Lane

Security researchers from Tencent have demonstrated a way to use physical attacks to spoof Tesla's autopilot

By **Evan Ackerman**

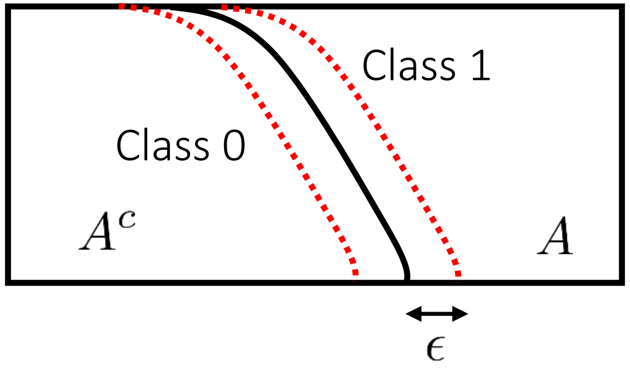
Source: IEEE Spectrum

Adversarial attacks are a security risk for safety-critical applications!

Adversarial Risk

General Loss: $$R_\epsilon(\ell, w) = \mathbb{E}_{(x,y)\sim\rho}\left[\sup_{d(x,x')\leq\epsilon} \ell((x', y), w)\right]$$

Expected value of worst-case loss



Binary Classification 0/1 Loss: $$R_{\oplus\epsilon}(\ell_{0/1}, A) = \frac{T}{T+1} p_0(A^{\oplus\epsilon}) + \frac{1}{T+1} p_1((A^c)^{\oplus\epsilon})$$

Priors in ratio T:1

Expanded error regions

True label distributions

$$A^{\oplus\epsilon} := \cup_{a\in A} B_\epsilon(a)$$

A Variety of Definitions

$R_{\oplus\epsilon}(\ell_{0/1}, A)$ Minkowski set expansion	$R_{\epsilon}(\ell_{0/1}, A)$ Closed set expansion
$\frac{T}{T+1}p_0(A^{\oplus\epsilon}) + \frac{1}{T+1}p_1((A^c)^{\oplus\epsilon})$ $A^{\oplus\epsilon} := \cup_{a\in A} B_\epsilon(a)$ Original definition, measurability issues	$\frac{T}{T+1}p_0(A^{\epsilon}) + \frac{1}{T+1}p_1((A^c)^{\epsilon})$ $A^{\epsilon} := \{x \in \mathcal{X} : d(x, A) \leq \epsilon\}$ Budget constraint violated
$R_{F_\epsilon}(\ell_{0/1}, A)$ Transport maps	$R_{F_\epsilon}(\ell_{0/1}, A)$ Transport couplings
$\sup_{\substack{f_0, f_1 : \mathcal{X}\to\mathcal{X} \\ \forall x\in\mathcal{X}, d(x, f_i(x))\leq\epsilon}} \frac{T}{T+1} f_{0\sharp p_0}(A) + \frac{1}{T+1} f_{1\sharp p_1}((A^c))$ $f_{\sharp\mu}(A) = \mu(f^{-1}(A))$ Deterministic perturbation	$\sup_{\substack{W_\infty(p_1, p_1')\leq\epsilon \\ W_\infty(p_0, p_0')\leq\epsilon}} \frac{T}{T+1}p_0'(A) + \frac{1}{T+1}p_1'((A^c))$ $W_\infty(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \operatorname{ess\,sup}_{(x,x')\sim\pi} d(x,x')$ Budget constraint holds a.s.

The Many Faces of Adversarial Risk

- The diversity of definitions makes it challenging to compare approaches
- Not all definitions are well-defined – issues of measurability persist (for $R_{\oplus\epsilon}(A)$)
- This has led to incorrect proofs and insufficient assumptions

A a mathematically rigorous foundation for adversarial risk is essential for future research.

Our Contributions (part 1 of 4)

Risk	Defining Characteristic	Adversary's action	Perturbation	$d(x,x') \le \epsilon$?
$R_{\oplus\epsilon}$	Minkowski set expansion	$x \in A \to x' \in A^{\oplus\epsilon}$	Random	Yes, $\forall x$
R_{Γ_ϵ}	Transport couplings	$p_0, p_1 \to p_0', p_1'$ $W_\infty(p_i, p_i') \le \epsilon$	Random	Almost surely yes, $\forall x$
R_{F_ϵ}	Transport maps	$x \to x' = f_i(x)$	Deterministic	Yes, $\forall x$
R_ϵ	Closed set expansion	$x \in A \to x' \in A^\epsilon$	Random	No

$(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d))$ — Euclidean space with Lebesgue σ-algebra

Any Polish space — $(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$ or $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$

$(\mathcal{X}, \overline{\mathcal{B}}(\mathcal{X}))$ and $A \in \mathcal{B}(\mathcal{X})$ — Borel sets in universally measurable metric space

$(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d))$ and p_0, p_1 have densities

Conditions for which adversarial risk is well-defined

Conditions for equivalence with adversarial risk

Our Contributions (part 2 of 4)

Optimal Adversarial Risk: $R^*_{\oplus\epsilon} := \inf_{A \in \mathcal{B}(\mathcal{X})} R_{\oplus\epsilon}(\ell_{0/1}, A)$

Optimal transport characterization of optimal adversarial risk:

$$R^*_{\oplus\epsilon} = \frac{1}{T+1}\left[1 - \inf_{\substack{q \in \mathcal{P}(\mathcal{X}) \\ q \preceq T p_0}} \inf_{\pi \in \Pi(q, p_1)} \mathbb{E}_{(x,x')\sim\pi}[\mathbb{1}\{d(x,x') > 2\epsilon\}]\right]$$

0-1 valued transport cost

Expected transport cost

Optimal transport cost

Optimize over probability measures stochastically dominated by Tp_0 (T>1)

Our Contributions (part 3 of 4)

Optimal Adversarial Risk:

$$R^*_{\oplus\epsilon} := \inf_{A \in \mathcal{B}(\mathcal{X})} R_{\oplus\epsilon}(\ell_{0/1}, A)$$

Distributionally robust optimization based characterization of optimal adversarial risk:

$$R^*_{\oplus\epsilon} = \sup_{\substack{W_\infty(p_1, p_1') \leq \epsilon \\ W_\infty(p_0, p_0') \leq \epsilon}} \underbrace{\frac{1}{T+1}\left[1 - \inf_{\substack{q \in \mathcal{P}(\mathcal{X}) \\ q \preceq T p_0'}} \overbrace{D_{TV}(q, p_1')}^{\text{Total Variation distance}}\right]}_{\text{Bayes risk for binary classification between q and p'}_1}$$

Contamination of true distributions in ∞-Wasserstein metric

Bayes risk for binary classification between q and p'_1

Our Contributions (part 4 of 4)

Optimal Adversarial Risk: $R^*_{\oplus\epsilon} := \inf_{A \in \mathcal{B}(\mathcal{X})} R_{\oplus\epsilon}(\ell_{0/1}, A)$

Game theoretic characterization of optimal adversarial risk:

$$r(A, p_0', p_1') = \frac{T}{T+1} p_0'(A) + \frac{1}{T+1} p_1'((A^c))$$

Payoff function

Player 1: Algorithm $f_A(x) = \mathbb{1}\{x \in A\}$
Action space: decision regions

Player 2: Adversary
Action space: Perturbed distributions in Wasserstein ball

$$R^*_{\oplus\epsilon} = \inf_{A \in \mathcal{B}(\mathcal{X})} \sup_{\substack{W_\infty(p_1, p_1') \le \epsilon \\ W_\infty(p_0, p_0') \le \epsilon}} r(A, p_0', p_1') = \sup_{\substack{W_\infty(p_1, p_1') \le \epsilon \\ W_\infty(p_0, p_0') \le \epsilon}} \inf_{A \in \mathcal{B}(\mathcal{X})} r(A, p_0', p_1')$$

Minimax theorem => Existence of ***Nash Equilibrium***

Summary & Related Works

Our results	**Technical tools**	**Previous works that we generalize/extend/strengthen**
Conditions for which adversarial risk is well-defined Conditions for equivalences between various notions of adversarial risk	Euclidean space: Porous sets Polish space: Analytic sets	• Meunier et al. (ICML, 2021) • Pydi and Jog (IEEE Trans. IT, 2021)
Optimal transport characterization of optimal adversarial risk	Generalized Strassen's theorem Duality in linear programming	• Strassen (Ann. Math. Stat. 1965) • Dohmatob (ICML 2019) • Bhagoji et al. (NeurIPS, 2019) • Pydi and Jog (ICML, 2020)
Distributionally robust optimization based characterization of optimal adversarial risk	Euclidean space: Huber and Strassen's theory of 2-alternating capacities Polish space: measurable selection theorems	• Sinha et al. (ICLR 2018) • Tu et al. (NeurIPS 2019) • Pydi and Jog (IEEE Trans. IT, 2021)
Game theoretic characterization of optimal adversarial risk	All of the above	• Pinot et al. (ICML 2020) • Bose et al. (NeurIPS 2020) • Meunier et al. (ICML, 2021)