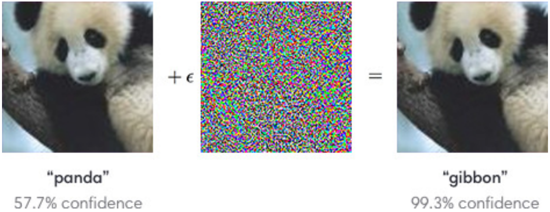
Certifying Some Distributional Robustness with Principled Adversarial Training

Aman Sinha*, Hongseok Namkoong*, & John Duchi

Stanford University

ICLR 2018

Motivation



[Goodfellow et al. 2015]

[Athalye et al. 2017]

We want to increase the robustness of machine-learned systems

Current Approaches

- Adversarial training heuristics: Fast but no theoretical guarantees of robustness
  - Goodfellow et al. 2015, Kurakin et al. 2016, Papernot et al. 2016, He et al. 2017, Carlini & Wagner 2017, Tramer et al. 2017, Madry et al. 2018, etc.

- Formal verification: Rigorous guarantees but very slow
  - Huang et al. 2017, Katz et al. 2017, Kolter & Wong 2017, Tjeng & Tedrake 2017, Raghunathan et al. 2018

Our goal: balance efficiency with robustness guarantees

Our Work: Principled adversarial training

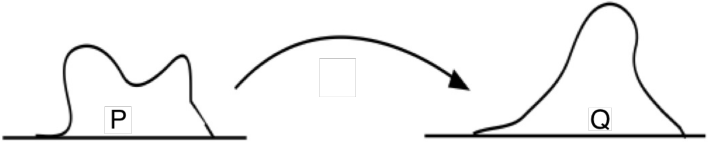
- Setup: model/network weights $\theta \in \Theta$, feature vector X, label Y, and loss function $\ell(\theta; X, Y)$

  Overall idea: replace $\ell(\theta; X, Y)$ with robust surrogate $\phi_\gamma(\theta; X, Y)$

- For moderate levels of desired robustness and smooth losses ℓ:
  - Provably fast convergence, 5-10x as fast as ERM
  - Statistical guarantees for performance on (perturbations to) the test set

Distributionally robust optimization (DRO)

- Goal: robustness to ρ-perturbations in a Wasserstein ball
  - $c(x, x_0)$: "cost" to perturb x_0 to x (e.g. $\|x - x_0\|^2$)
  - Wasserstein distance
    $D_c(Q, P) := \min_{M: M_X = Q, M_{X'} = P} \mathbb{E}_M[\|X - X'\|^2]$
- Generally intractable for arbitrary ρ

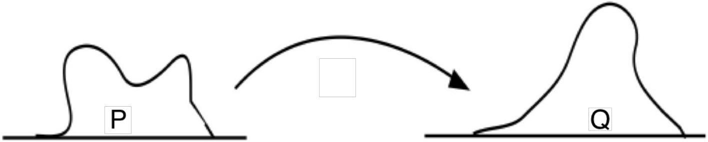


[Esfahani & Kuhn 2015; Shafieezadeh-Abadeh et al. 2015; Blanchet et al. 2016, Lee & Raginsky 2017]

Distributionally robust optimization (DRO)

$$\underset{\theta \in \Theta}{\text{minimize}} \ \mathbb{E}_{P_0}[\ell(\theta; X, Y)]$$

- Goal: robustness to ρ-perturbations in a Wasserstein ball
  - $c(x, x_0)$: "cost" to perturb x_0 to x (e.g. $\|x - x_0\|^2$)
  - Wasserstein distance $D_c(Q, P) := \min_{M: M_X = Q, M_{X'} = P} \mathbb{E}_M[\|X - X'\|^2]$
- Generally intractable for arbitrary ρ

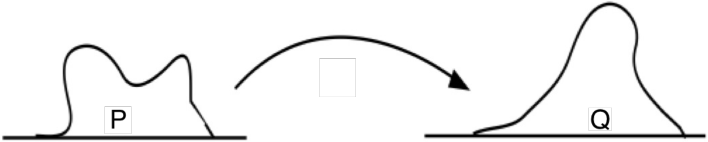


[Esfahani & Kuhn 2015; Shafieezadeh-Abadeh et al. 2015; Blanchet et al. 2016, Lee & Raginsky 2017]

Distributionally robust optimization (DRO)

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \max_{Q} \{\mathbb{E}_Q[\ell(\theta; X, Y)] : D_c(Q, P_0) \leq \rho\}$$

- Goal: robustness to ρ-perturbations in a Wasserstein ball
  - $c(x, x_0)$: "cost" to perturb x_0 to x (e.g. $\|x - x_0\|^2$)
  - Wasserstein distance
    $D_c(Q, P) := \min_{M : M_X = Q, M_{X'} = P} \mathbb{E}_M[\|X - X'\|^2]$
- Generally intractable for arbitrary ρ



[Esfahani & Kuhn 2015; Shafieezadeh-Abadeh et al. 2015; Blanchet et al. 2016, Lee & Raginsky 2017]

Distributionally robust optimization (DRO)

- Lagrangian relaxation and its dual formulation (more robustness $\leftrightarrow$ larger $\rho \leftrightarrow$ smaller γ)

$$\underset{\theta \in \Theta}{\text{minimize}} \ \max_{Q} \left\{ \mathbb{E}_Q[\ell(\theta; X, Y)] - \underbrace{\gamma D_c(Q, P_0)}_{\text{penalty}} \right\} =$$

$$\underset{\theta \in \Theta}{\text{minimize}} \ \mathbb{E}_{P_0}[\phi_\gamma(\theta; X, Y)]$$

$$\text{where} \quad \phi_\gamma(\theta; x, y) := \max_{x' \in \mathcal{X}} \left\{ \ell(\theta; x', y) - \underbrace{\gamma \|x' - x\|^2}_{\text{penalty}} \right\}$$

- Compare to ERM: $\text{minimize}_{\theta \in \Theta} \ E_{P_0}[\ell(\theta; X, Y)]$

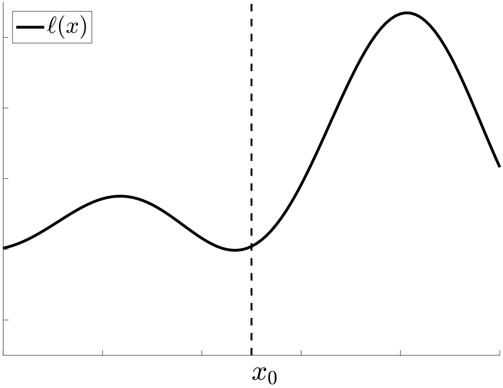
[Blanchet et al. 2016]

Solving the optimization problem

$$\phi_\gamma(\theta; x_0, y_0) := \max_{x \in \mathcal{X}} \left\{ \ell(\theta; x, y_0) - \gamma \|x - x_0\|^2 \right\}$$

Key insight: $(x, y) \mapsto \ell(\theta; x, y) - \gamma \|x - x_0\|^2$ is strongly concave for **smooth** ℓ and large enough γ

- Curvature in $\|\cdot\|^2$ dwarfs out non-concavity of $\ell(\theta; \cdot)$

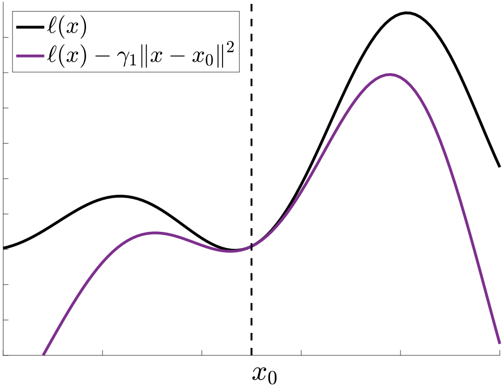


Solving the optimization problem

$$\phi_\gamma(\theta; x_0, y_0) := \max_{x \in \mathcal{X}} \left\{ \ell(\theta; x, y_0) - \gamma \|x - x_0\|^2 \right\}$$

Key insight: $(x, y) \mapsto \ell(\theta; x, y) - \gamma \|x - x_0\|^2$ is strongly concave for **smooth** ℓ and large enough γ

- Curvature in $\|\cdot\|^2$ dwarfs out non-concavity of $\ell(\theta; \cdot)$

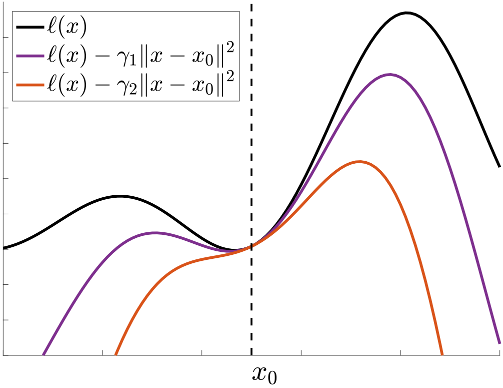


Solving the optimization problem

$$\phi_\gamma(\theta; x_0, y_0) := \max_{x \in \mathcal{X}} \left\{ \ell(\theta; x, y_0) - \gamma \|x - x_0\|^2 \right\}$$

Key insight: $(x, y) \mapsto \ell(\theta; x, y) - \gamma \|x - x_0\|^2$ is strongly concave for **smooth** ℓ and large enough γ

- Curvature in $\|\cdot\|^2$ dwarfs out non-concavity of $\ell(\theta; \cdot)$

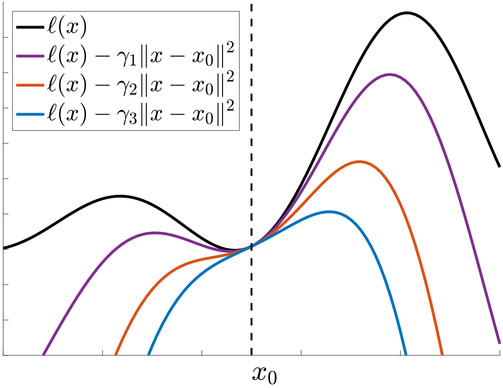


Solving the optimization problem

$$\phi_\gamma(\theta; x_0, y_0) := \max_{x \in \mathcal{X}} \left\{ \ell(\theta; x, y_0) - \gamma \|x - x_0\|^2 \right\}$$

Key insight: $(x, y) \mapsto \ell(\theta; x, y) - \gamma \|x - x_0\|^2$ is strongly concave for **smooth** ℓ and large enough γ

- Curvature in $\|\cdot\|^2$ dwarfs out non-concavity of $\ell(\theta; \cdot)$

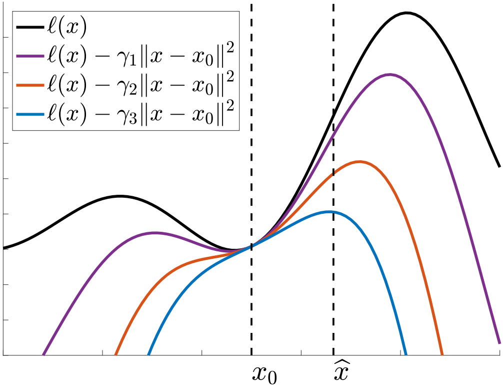


Solving the optimization problem

$$\phi_\gamma(\theta; x_0, y_0) := \max_{x \in \mathcal{X}} \left\{ \ell(\theta; x, y_0) - \gamma \|x - x_0\|^2 \right\}$$

Key insight: $(x, y) \mapsto \ell(\theta; x, y) - \gamma\|x - x_0\|^2$ is strongly concave for **smooth** ℓ and large enough γ

- Curvature in $\|\cdot\|^2$ dwarfs out non-concavity of $\ell(\theta; \cdot)$

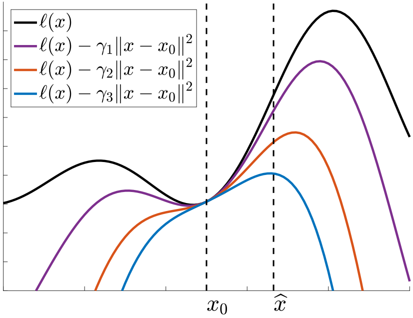


Solving the optimization problem

$$\phi_\gamma(\theta; x_0, y_0) := \max_{x \in \mathcal{X}} \left\{ \ell(\theta; x, y_0) - \gamma \|x - x_0\|^2 \right\}$$

Key insight: $(x, y) \mapsto \ell(\theta; x, y) - \gamma \|x - x_0\|^2$ is strongly concave for **smooth** ℓ and large enough γ

- Curvature in $\|\cdot\|^2$ dwarfs out non-concavity of $\ell(\theta; \cdot)$



Deep nets with smooth activations (ELUs, sigmoid, etc.) are smooth

Optimization guarantees

Algorithm: SGD for $\min_\theta E_{P_0}[\phi_\gamma(\theta; X, Y)]$

- Sample $(x^t, y^t) \sim P_0$
- Compute adversarial example:
  (approximate) maximizer $\widehat{x}^t$ of $\ell(\theta^t; x, y^t) - \gamma\|x - x^t\|^2$
- $\theta^{t+1} \leftarrow \theta^t - \alpha \nabla_\theta \ell(\theta^t; \widehat{x}^t, y^t)$

- So long as $\nabla_x \ell(\theta; \cdot)$ is L_{xx}-Lipschitz and $\gamma > L_{xx}$, we can compute $\widehat{x}^t$ in $10 \sim 20$ gradient ascent steps
- Theorem: converges at standard nonconvex-SGD rate

Certificate of robustness

- Algorithm **generalizes**: we learn to prevent attacks on the **test** set
- θ_{WRM} = output of Algorithm, $\mathfrak{Comp}_n$ = size of Θ, C = problem-dependent constant, $\widehat{P}_n$ = empirical training distribution

Theorem (Robustness Certificate)

With high probability, for any $\rho \geq 0$

$$\max_{Q : D_c(Q, P_0) \leq \rho} \mathbb{E}_Q[\ell(\theta_{\mathrm{WRM}}; X, Y)] \leq \gamma\rho + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta_{\mathrm{WRM}}; X, Y)] + C\frac{\mathfrak{Comp}_n}{\sqrt{n}}$$

Certificate of robustness

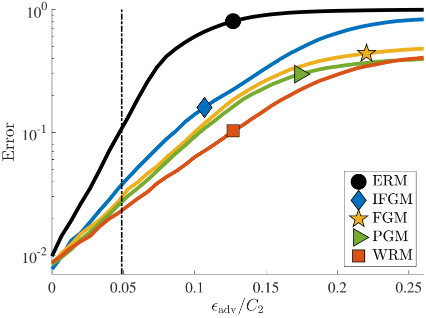
- Bounds can be large in practical applications due to dimension/covering-number dependence

- Alternative bound for any empirical test set $\widehat{P}_{\text{test}}$ and test examples $(x_i^{\text{test}}, y_i^{\text{test}})$

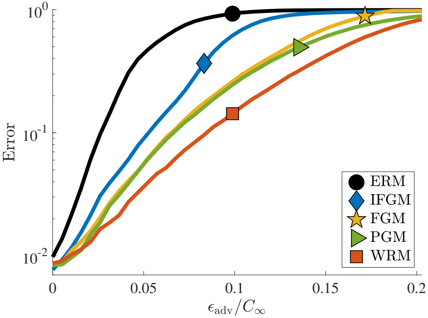
$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n} \max_{x : \|x - x_i^{\text{test}}\|^2 \le \rho} \left\{ \ell(\theta; x, y_i^{\text{test}}) \right\} \le \max_{Q : D_c(Q, \widehat{P}_{\text{test}}) \le \rho} \mathbb{E}_P[\ell(\theta; X, Y)]$$
$$\le \gamma\rho + \mathbb{E}_{\widehat{P}_{\text{test}}}[\phi_\gamma(\theta; X, Y)]$$

MNIST classification

- Compare our method (WRM) with fast-gradient method (FGM), iterated FGM (IFGM), and projected gradient method (PGM)
- All models trained with 2-norm adversary



Test error vs. ϵ_{adv} for PGM $\|\cdot\|_2$ attack

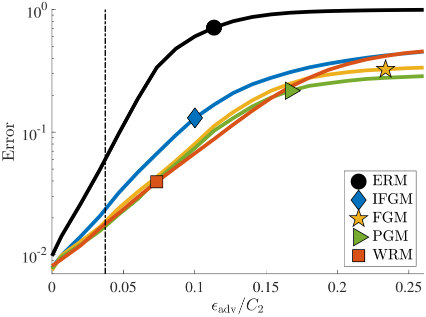


Test error vs. ϵ_{adv} for PGM $\|\cdot\|_\infty$ attack

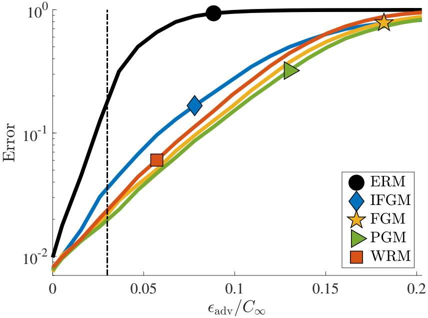
[Goodfellow et al. 2015, Kurakin et al. 2016, Madry et al. 2017]

MNIST classification

- All models except WRM trained with ∞-norm adversary



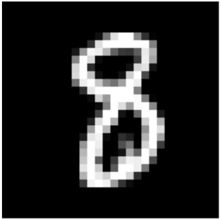
Test error vs. ϵ_{adv} for PGM $\|\cdot\|_2$ attack



Test error vs. ϵ_{adv} for PGM $\|\cdot\|_\infty$ attack

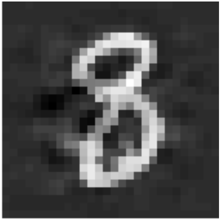
When the model misclassifies

- Minimum perturbation forcing WRM to misclassify is perceptible

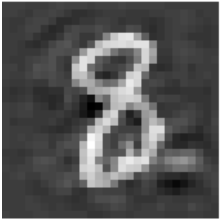


Original

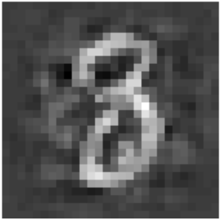
ERM



FGM



IFGM



PGM

WRM

Conclusions & Future Work

- Optimization and robustness guarantees for small adversarial budgets (imperceptible perturbations)
- More empirical comparisons needed on larger models/datasets
- Statistical guarantees can be loose due to covering-number arguments
  - Recent developments: Bartlett et al. 2017, Dziugaite & Roy 2017, Neyshabur et al. 2017

Poster 7, Wednesday 11am-1pm

Code: https://github.com/duchi-lab/certifiable-distributional-robustness