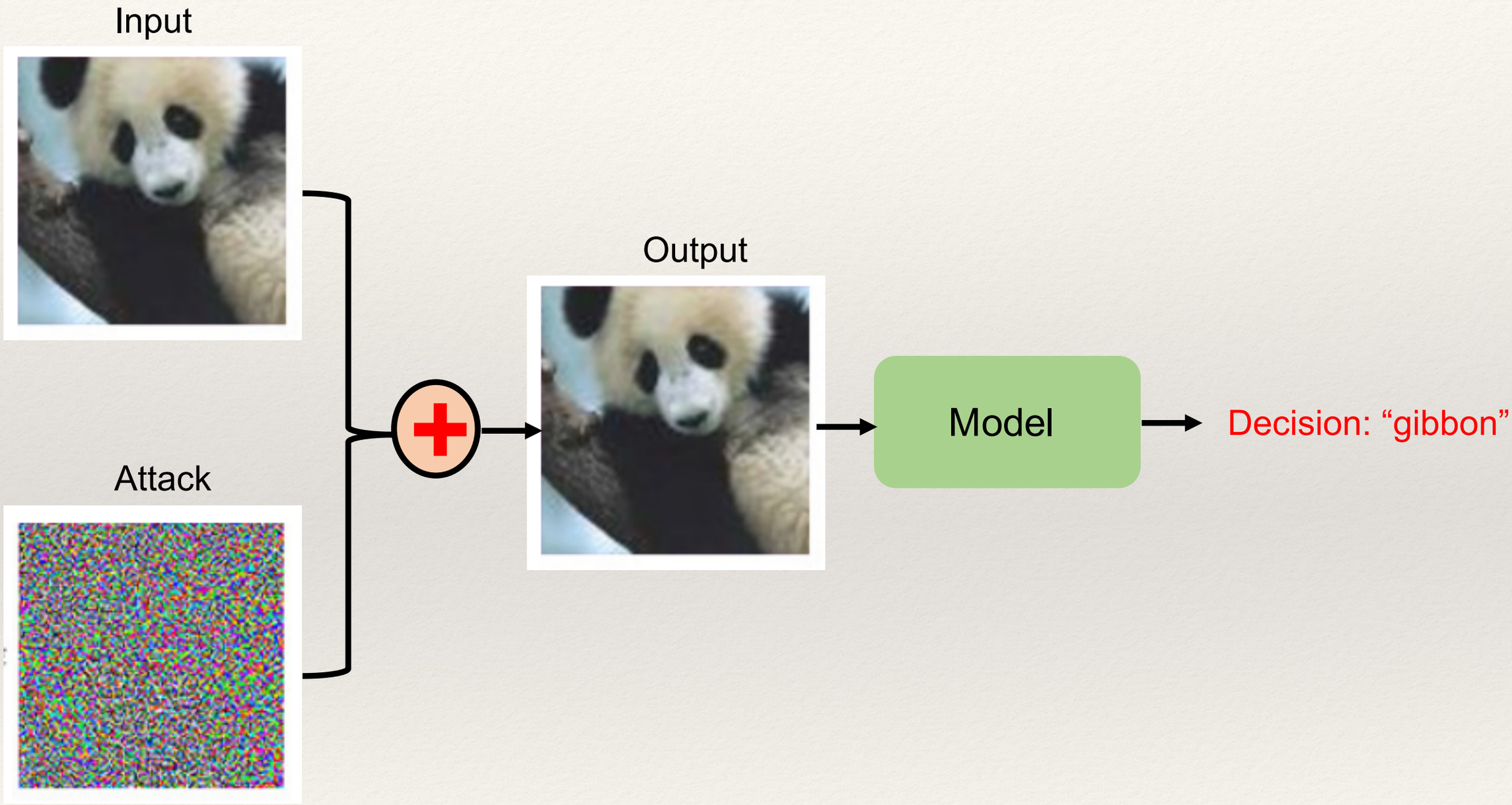
Entropic Regularization for Adversarial Robust Learning

Jie Wang
Georgia Institute of Technology

April 25th, 2023

Joint work with Yifan Lin (Gatech), and
Song Wei (Gatech)

On the Robustness of ML Models



[Goodfellow et al. 2015]

Adversarial Risk Minimization

- Formulation:

$$\min_{\theta\in\Theta} \left\{ F(\theta;\rho,\widehat{\mathbb{P}}) = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\sup_{z\in\mathbb{B}_\rho(x)} f_\theta(z) \right] \right\}$$

- **Intractability**:

When $f_\theta(z)$ is convex in θ: Convex-Nonconcave Game

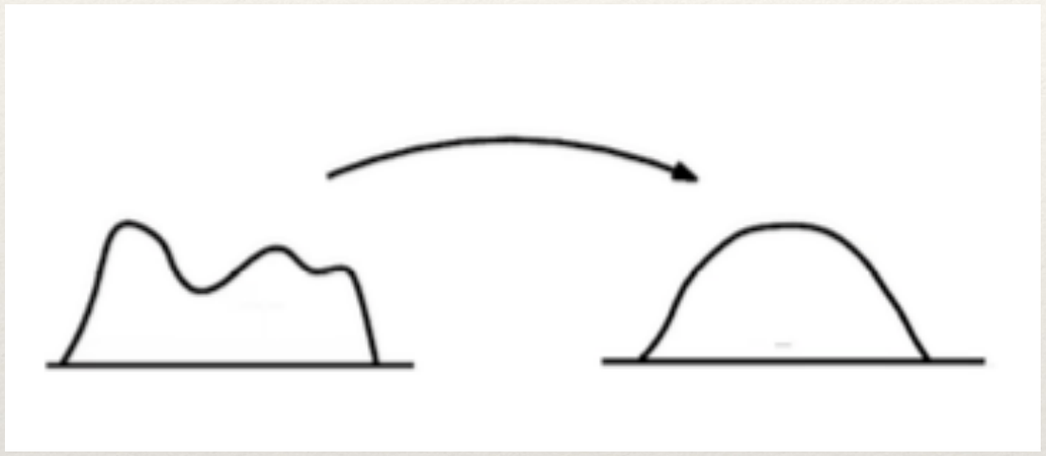
When $f_\theta(z)$ is nonconvex in θ: Nonconvex-Nonconcave Game

- Connections with Optimal Transport:

$$\min_{\theta\in\Theta} \left\{ \sup_{\mathbb{P}} \ \mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)] : \ \mathcal{W}_\infty(\mathbb{P},\widehat{\mathbb{P}}) \le \rho \right\}$$

p-Type Wasserstein Distance

$$\mathcal{W}_p(\mathbb{P},\mathbb{Q}) := \left(\min_{\gamma\in\mathcal{P}(\Omega^2)} \left\{ \mathbb{E}_{(\omega,\omega')\sim\gamma}[\|\omega-\omega'\|^p] : \right.\right.$$



$$\left.\left. \gamma \text{ has marginal distributions } \mathbb{P} \text{ and } \mathbb{Q} \right\} \right)^{1/p}.$$

$$\mathcal{W}_\infty(\mathbb{P},\mathbb{Q}) = \inf_{\gamma} \left\{ \text{ess.sup } \mathsf{d}(\zeta_1,\zeta_2) : \begin{array}{l} \gamma \text{ is a joint distribution of } \zeta_1 \text{ and } \zeta_2 \\ \text{with marginals } \mathbb{P} \text{ and } \mathbb{Q}, \text{ respectively} \end{array} \right\}$$

- ✓ Geometric properties;
- ✓ Flexibility: non-overlapping support, discrete and continuous.

Our Proposed Formulation

- Original Formulation:

$$\min_{\theta\in\Theta} \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{\mathbb{P}}[f_\theta(z)] : \begin{array}{l} \mathrm{Proj}_{1\#\gamma} = \widehat{\mathbb{P}}, \mathrm{Proj}_{2\#\gamma} = \mathbb{P} \\ \mathrm{ess.sup}_\gamma \mathsf{d}(\zeta_1, \zeta_2) \leq \rho \end{array} \right\}$$

- Formulation with Entropic Regularization:

$$\min_{\theta\in\Theta} \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{\mathbb{P}}[f_\theta(z)] - \eta H(\gamma \mid \pi) : \begin{array}{l} \mathrm{Proj}_{1\#\gamma} = \widehat{\mathbb{P}}, \mathrm{Proj}_{2\#\gamma} = \mathbb{P} \\ \mathrm{ess.sup}_\gamma \mathsf{d}(\zeta_1, \zeta_2) \leq \rho \end{array} \right\}$$

Entropy: $$H(\gamma \mid \pi) = \mathbb{E}_{(x,z)\sim\gamma} \left[\log \left(\frac{\mathrm{d}\gamma(x,z)}{\mathrm{d}\pi(x,z)} \right) \right]$$

$$= \mathbb{E}_{x\sim\widehat{\mathbb{P}}} \mathbb{E}_{z\sim\gamma_x} \left[\log \left(\frac{\mathrm{d}\widehat{\mathbb{P}}(x)\mathrm{d}\gamma_x(z)}{\mathrm{d}\widehat{\mathbb{P}}(x)\mathrm{d}\nu_x(z)} \right) \right]$$

Our Contribution (I)

- Strong Dual Formulation:

$$V_{\text{Primal}} = \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{z\sim\mathbb{P}}\left[f_\theta(z)\right] - \eta H(\gamma \mid \pi) : \begin{array}{l} \text{Proj}_{1\#\gamma} = \widehat{\mathbb{P}}, \text{Proj}_{2\#\gamma} = \mathbb{P} \\ \text{ess.sup}_\gamma \mathsf{d}(\zeta_1, \zeta_2) \le \rho \end{array} \right\}.$$

$$V_{\text{Dual}} = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\eta \log \mathbb{E}_{z\sim\mathbb{Q}_x}\left[\exp\left(\frac{f_\theta(z)}{\eta}\right)\right]\right] + \text{Constant}.$$

- Kernel Probability Distribution:

$$\frac{\mathrm{d}\mathbb{Q}_x}{\mathrm{d}\nu_x}(z) = \text{vol}(\nu_x)^{-1}, \quad z \in \mathbb{B}_\rho(x).$$

Our Contribution (II)

$$V_{\text{Primal}} = \sup_{\mathbb{P},\gamma} \left\{ \mathbb{E}_{z\sim\mathbb{P}}\left[f_\theta(z)\right] - \eta H(\gamma \mid \pi) : \begin{array}{l} \text{Proj}_{1\#\gamma} = \widehat{\mathbb{P}}, \text{Proj}_{2\#\gamma} = \mathbb{P} \\ \text{ess.sup}_\gamma \mathsf{d}(\zeta_1, \zeta_2) \le \rho \end{array} \right\}.$$

- Recover Worst-case Distribution:

The worst-case distribution maps each $x \in \text{supp}(\widehat{\mathbb{P}})$ to a conditional distribution γ_x whose density value (with respect to measure ν_x) is proportional to

$$\alpha_x \cdot \exp\left(\frac{f_\theta(z)}{\eta}\right),$$

$$\mathrm{d}\mathbb{P}_*(z) = \mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\alpha_x \cdot \exp\left(\frac{f_\theta(z)}{\eta}\right) \mathrm{d}\nu_x(z)\right]$$

Our Contribution (III)

- Tractable Algorithm for Dual Formulation:

$$\min_{\theta\in\Theta} \left\{ F_\eta(\theta) \triangleq \mathbb{E}_{x\sim\widehat{\mathbb{P}}} \left[\eta \log \mathbb{E}_{z\sim\mathbb{Q}_x} \left[\exp\left(\frac{f_\theta(z)}{\eta} \right) \right] \right] \right\}$$

Algorithm 1 BSMD for solving (8)

Require: maximum iterations T, constant step size γ, initial guess θ_0.

1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: Formulate (biased) gradient estimate of $F_\eta(\theta_t)$, denoted as $v(\theta_t)$.
3: Update $\theta_{t+1} = \mathrm{Prox}_{\theta_t}\left(\gamma v(\theta_t)\right)$.
4: **end for**

Output $\tilde{\theta}$ **randomly selected from** $\{\theta_0, \theta_1, \ldots, \theta_T\}$.

Our Contribution (III)

- Tractable Algorithm for Dual Formulation:

$$\min_{\theta\in\Theta} \left\{ F_\eta(\theta) \triangleq \mathbb{E}_{x\sim\widehat{\mathbb{P}}} \left[\eta \log \mathbb{E}_{z\sim\mathbb{Q}_x} \left[\exp\left(\frac{f_\theta(z)}{\eta} \right) \right] \right] \right\}$$

Estimators	Convex Nonsmooth	Convex Smooth	Nonconvex Smooth
Vanilla SGD	$O(\delta^{-3})$	$O(\delta^{-3})$	$O(\delta^{-6})$
V-MLMC	N/A	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$
RT-MLMC	N/A	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$

Vanilla SGD Estimator

$$\min_{\theta\in\Theta} \left\{ F_\eta(\theta) \triangleq \mathbb{E}_{x\sim\widehat{\mathbb{P}}} \left[\eta \log \mathbb{E}_{z\sim\mathbb{Q}_x} \left[\exp\left(\frac{f_\theta(z)}{\eta} \right) \right] \right] \right\}$$

- Construct approximation function:

$$F_\eta^\ell(\theta) = \mathbb{E}_{x^\ell}\ \mathbb{E}_{\{z_j^\ell\}_{j\in[2^\ell]}|x^\ell} \left[\eta \log \left(\frac{1}{2^\ell} \sum_{j\in[2^\ell]} \exp\left(\frac{f_\theta(z_j^\ell)}{\eta} \right) \right) \right].$$

- Oracle: Sample random parameters $\zeta^L = \{x^L\} \cup \{z_j^\ell\}_j$ and compute

$$g(\theta, \zeta^\ell) = \nabla_\theta \left\{ \eta \log \left(\frac{1}{2^L} \sum_{j\in[2^L]} \exp\left(\frac{f_\theta(z_j^\ell)}{\eta} \right) \right) \right\}$$

- V-SGD: call oracle for several times and return avg

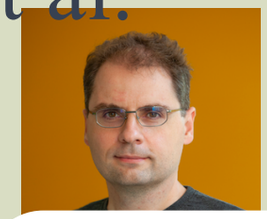
Connections with Existing Literature

Approximation using p-Wasserstein DRO

Sinha, et al. 2020

Origin of Adversarial Training

Aleksander, et al. 2017

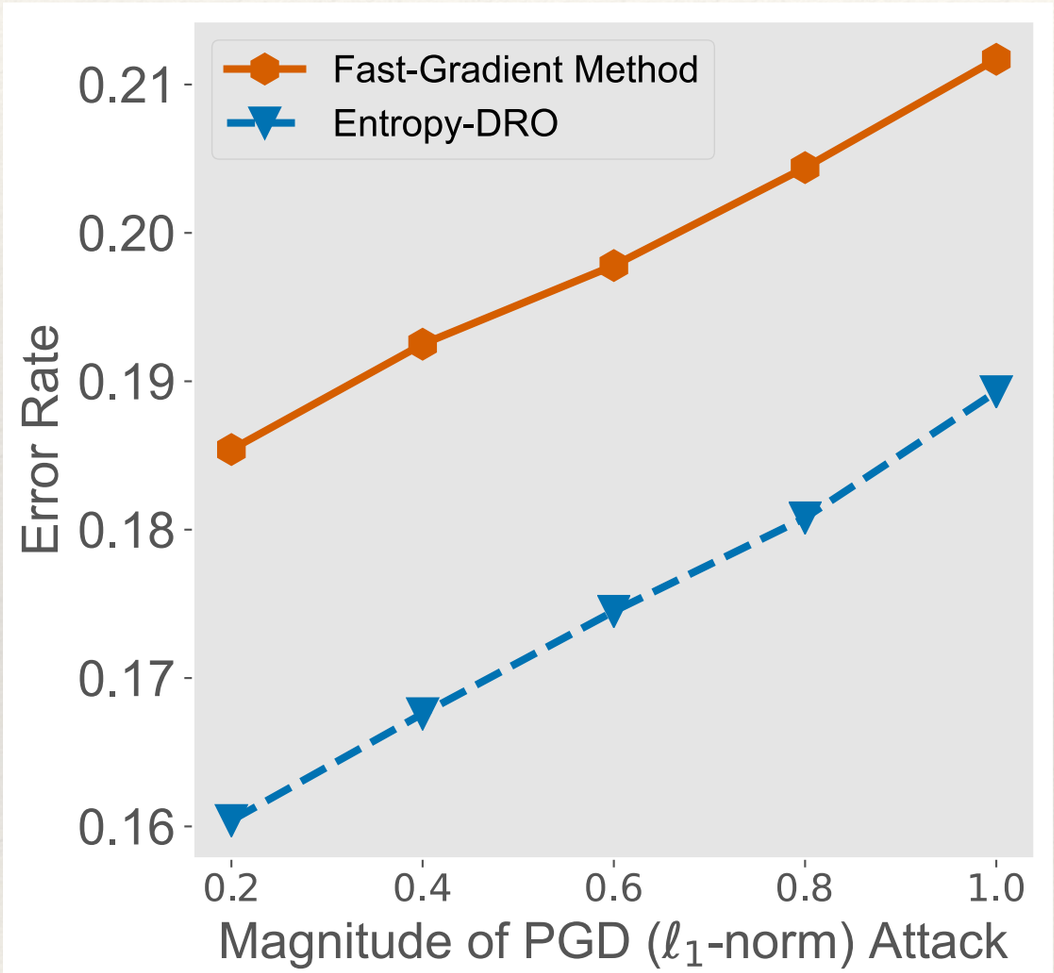


Adding Entropic Regularization Directly

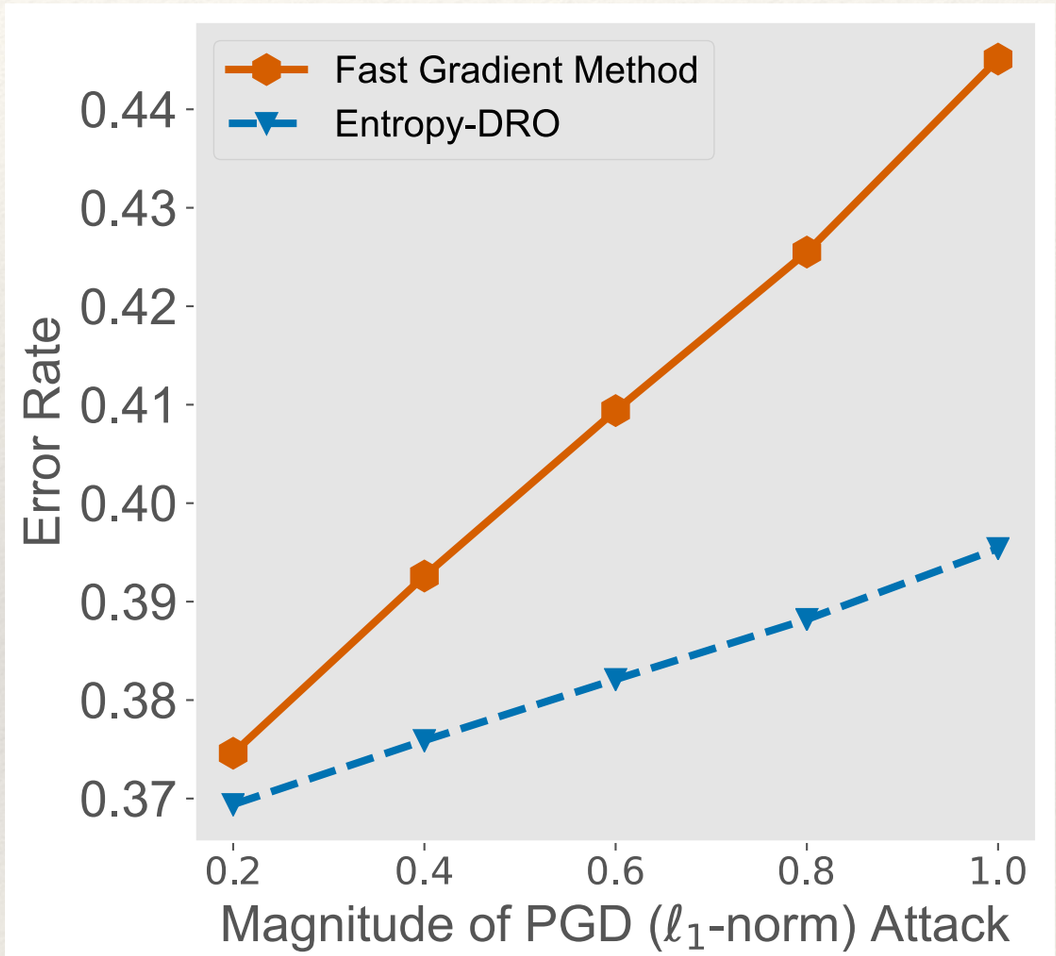
Entropic Regularization for p-Wasserstein DRO

Wang, et al. 2021

Numerical Study



Cifar-10 Dataset



Cifar-100 Dataset