Entropic Regularization for Adversarial Robust Learning

Jie Wang[†], Yifan Lin[†], Song Wei[†], Rui Gao[‡], Yao Xie[†]

† H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

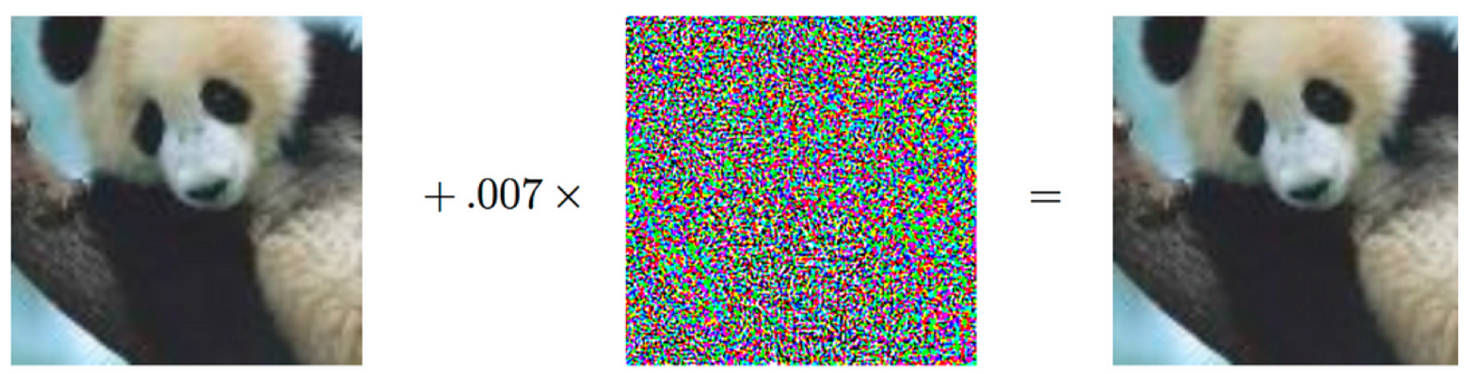
‡ Department of Information, Risk, and Operations Management, McCombs School of Business, University of Texas at Austin

Summary of Contributions

- Novel adversarial robust training framwork integrating **distributionally robust optimization** and **entropic regularization**.
- **Near-optimal** stochastic gradient methods with biased gradient oracles.
- Connections with **regularized** empirical risk minimization training.

Motivation and Background

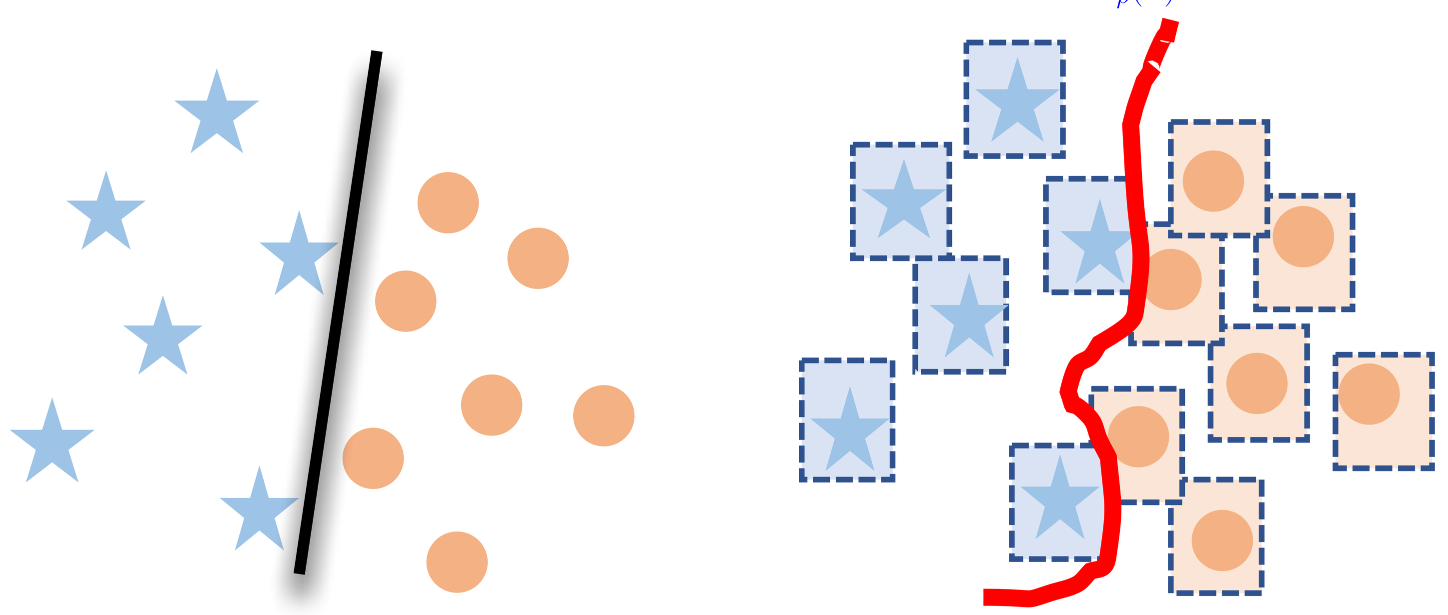
- Neural networks are vulnerable to adversarial attacks (Goodfellow et al, 2015).



Classified as **Panda** | Imperceptible Perturbation | Classified as **Gibbon**

- Adversarial training (Aleksander et al, 2018):

$$\min_{\theta\in\Theta}\left\{\mathbb{E}_{x\sim\widehat{\mathbb{P}}}[R_\rho(\theta;x)]\right\},\quad \text{where } R_\rho(\theta;x)\triangleq\sup_{z\in\mathbb{B}_\rho(x)} f_\theta(z).\qquad \text{(AT)}$$



Concern: Inner supremum of (AT) is generally **nonconcave** in z!
Literature: Approximately solves $R_\rho(\theta;x)$ by **linear approximation** of $f_\theta(z)$ around x.

- Distributionally robust optimization (DRO) point of view:

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

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

where $\mathcal{W}_\infty(\cdot,\cdot)$ is the ∞-Wasserstein metric:

$$\mathcal{W}_\infty(\mathbb{P},\mathbb{Q})=\inf_{\gamma:\ \text{Proj}_{1\#}\gamma=\mathbb{P},\text{Proj}_{2\#}\gamma=\mathbb{Q}}\left\{\text{ess.sup}_\gamma\ \|\zeta_1-\zeta_2\|\right\}.$$

Proposed Formulation

- Entropic-regularized formulation:

$$\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P},\gamma}\left\{\mathbb{E}_{z\sim\mathbb{P}}[f_\theta(z)]-\eta\mathcal{H}(\gamma):\ \begin{array}{l}\text{Proj}_{1\#\gamma}=\widehat{\mathbb{P}},\ \text{Proj}_{2\#\gamma}=\mathbb{P}\\ \text{ess.sup}_\gamma\|\zeta_1-\zeta_2\|\le\rho\end{array}\right\}\right\}.\qquad\text{(Entropy-AT)}$$

The entropy term $\mathcal{H}(\gamma)\triangleq\int\log\left(\frac{\mathrm{d}\gamma(x,z)}{\mathrm{d}\widehat{\mathbb{P}}(x)\,\mathrm{d}z}\right)\mathrm{d}\gamma(x,z)$.

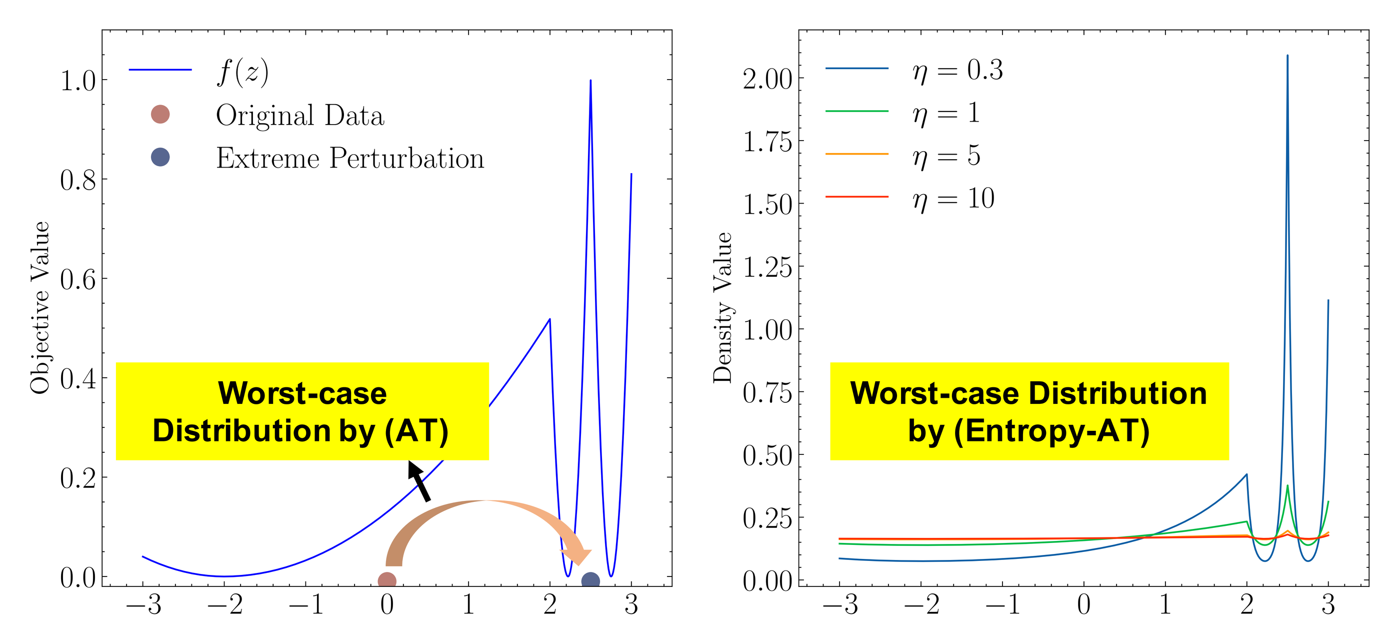
Under mild assumptions it holds that $V_{\mathrm{P}}=V_{\mathrm{D}}$:

$$V_{\mathrm{P}}=\sup_{\mathbb{P},\gamma}\left\{\mathbb{E}_{z\sim\mathbb{P}}[f(z)]-\eta\mathcal{H}(\gamma):\ \begin{array}{l}\text{Proj}_{1\#\gamma}=\widehat{\mathbb{P}},\ \text{Proj}_{2\#\gamma}=\mathbb{P}\\ \text{ess.sup}_\gamma\|\zeta_1-\zeta_2\|\le\rho\end{array}\right\},$$

$$V_{\mathrm{D}}=\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\eta\log\mathbb{E}_{z\sim\mathbb{Q}_x}\left[\exp\left(\frac{f(z)}{\eta}\right)\right]\right],$$

where $\mathbb{Q}_x$ **is an uniform distribution on** $\mathbb{B}_\rho(x)$.

- **Geometry of Worst-Case Distribution:**
  - For each $x\in\text{supp}(\widehat{\mathbb{P}})$, optimal transport maps it to a (conditional) distribution γ_x: $\frac{\mathrm{d}\gamma_x(z)}{\mathrm{d}z}=\alpha_x\cdot e^{f(z)/\eta},\quad z\in\mathbb{B}_\rho(x)$.
  - Worst-case distribution $\widetilde{\mathbb{P}}=\int\gamma_x\,\mathrm{d}\widehat{\mathbb{P}}(x)$.
- When $f(z)$ is a quadratic loss with 1-dimensional input neural network, $\widehat{\mathbb{P}}=\delta_{x=0}$, and $\rho=3$:



Optimization Algorithm

- Reformulate (Entropy-AT) as a single minimization:

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

- Biased Stochastic Mirror Descent (BSMD): for $t=1,\ldots,T$,

$$\begin{cases}v(\theta_t)\leftarrow\text{(biased) gradient/subgradient estimate of } F(\theta_t)\\ \theta_{t+1}\leftarrow\textbf{Prox}_{\theta_t}(\tau v(\theta_t))\end{cases}$$

Scenarios	Computation Cost	Memory Cost
Nonsmooth Convex Optimization	$\tilde{O}(\epsilon^{-2})$	$\tilde{O}(1)$
Constrained Smooth Nonconvex Optimization	$\tilde{O}(\epsilon^{-4})$	$\tilde{O}(\epsilon^{-2})$
Unconstrained Nonconvex Optimization	$\tilde{O}(\epsilon^{-4})$	$\tilde{O}(1)$

Gradient Estimator using Multi-level Monte-Carlo (MLMC):

- Consider $O(2^{-\ell})$-approximation function of $F(\theta)$:

$$F^\ell(\theta)=\mathbb{E}_{x^\ell\sim\widehat{\mathbb{P}}}\mathbb{E}_{\{z_j^\ell\}\sim\mathbb{Q}_{x^\ell}}\left[\eta\log\left(\frac{1}{2^\ell}\sum_j\exp\left(\frac{f_\theta(z_j^\ell)}{\eta}\right)\right)\right].$$

Define samples $\zeta^\ell=(x^\ell,\{z_j^\ell\}_{j\in[2^\ell]})$, and

$$U_{n_1:n_2}(\theta,\zeta^\ell)=\eta\log\left(\frac{1}{n_2-n_1+1}\sum_{j\in[n_1:n_2]}\exp\left(\frac{f_\theta(z_j^\ell)}{\eta}\right)\right),$$

$$G^\ell(\theta,\zeta^\ell)=\nabla_\theta\left[U_{1:2^\ell}(\theta,\zeta^\ell)-\frac{1}{2}U_{1:2^{\ell-1}}(\theta,\zeta^\ell)-\frac{1}{2}U_{2^{\ell-1}+1:2^\ell}(\theta,\zeta^\ell)\right].$$

(a) Sample random level $\iota\sim\mathbb{Q}_{\mathrm{RT}}$ with

$$\mathbb{Q}_{\mathrm{RT}}(\iota=\ell)=q_\ell\propto 2^{-\ell},\ell=0,\ldots,L.$$

(b) Construct $v^{\mathrm{MLMC}}(\theta)=\frac{1}{q_\iota}\cdot G^\iota(\theta;\zeta^\iota)$.

- MLMC estimator $v^{\mathrm{MLMC}}(\theta)$ is an unbiased estimator of $\nabla F^L(\theta)$:

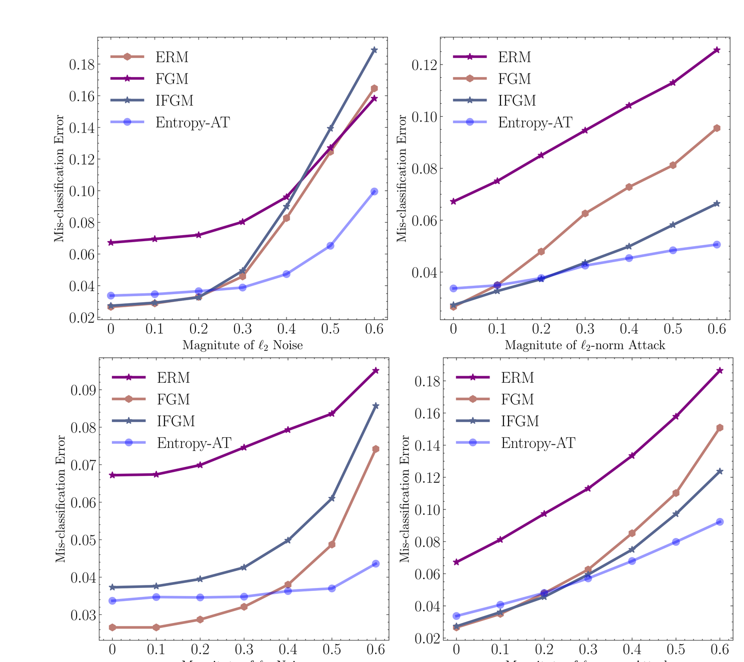
$$\mathbb{E}[v^{\text{RT-MLMC}}(\theta)]=\mathbb{E}_{\iota_1}\left[\frac{1}{q_{\iota_1}}\mathbb{E}_{\zeta^{\iota_1}}[G^{\iota_1}(\theta,\zeta^{\iota_1})]\right]=\sum_{\ell=0}^L q_\ell\cdot\left[\frac{1}{q_\ell}\mathbb{E}_{\zeta^\ell}[G^\ell(\theta,\zeta^\ell)]\right]=\nabla F^L(\theta).$$

- $U_{1:2^\ell}(\theta,\zeta^\ell),U_{1:2^{\ell-1}}(\theta,\zeta^\ell)$, and $U_{2^{\ell-1}+1:2^\ell}(\theta,\zeta^\ell)$ are generated using the same ζ^ℓ, implying $G^\ell(\theta,\zeta^\ell)$ and $v^{\mathrm{MLMC}}(\theta)$ has small variance due to **control variate effect**.

Regularization Effect:

$$\text{(Entropy-AT)}\approx\begin{cases}\min_{\theta\in\Theta}\ \mathbb{E}_{\widehat{\mathbb{P}}}[f_\theta(x)]+\rho\mathbb{E}_{x\sim\widehat{\mathbb{P}}}[\|\nabla f_\theta(x)\|_*], & \text{if } \rho/\eta\to\infty,\\ \min_{\theta\in\Theta}\ \mathbb{E}_{\widehat{\mathbb{P}}}[f_\theta(x)]+\frac{\rho^2}{\eta}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\mathsf{Var}_{z\sim\mathbb{Q}_x}[\nabla f_\theta(x)^{\mathrm{T}}z]\right], & \text{if } \rho/\eta\to 0,\\ \min_{\theta\in\Theta}\ \mathbb{E}_{\widehat{\mathbb{P}}}[f_\theta(x)]+\frac{\rho}{C}\mathbb{E}_{x\sim\widehat{\mathbb{P}}}\left[\log\mathbb{E}_{\mathbb{Q}_x}\left[\exp\left(C\nabla f_\theta(x)^{\mathrm{T}}z\right)\right]\right], & \text{if } \rho/\eta\to C.\end{cases}$$

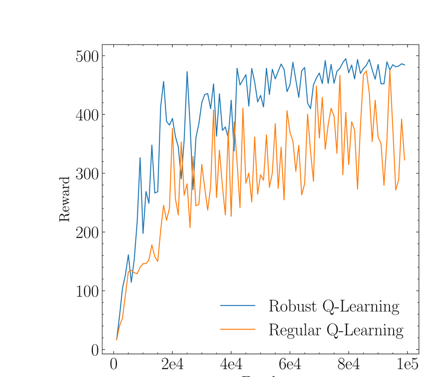
Numerical Study on Supervised and Reinforcement Learning



- Neural network classifier on MNIST dataset;
- Four types of adversarial attack;
- FGM/IFGM are heuristics for solving (AT) based on linear approximation.
- Entropic-AT performs well especially for large adversarial perturbations.

Environment	Regular	Robust
Original MDP	469.42 ± 19.03	$\mathbf{487.11\pm9.09}$
Perturbed MDP (Heavy)	187.63 ± 29.40	$\mathbf{394.12\pm12.01}$
Perturbed MDP (Short)	355.54 ± 28.89	$\mathbf{443.17\pm9.98}$
Perturbed MDP (Strong g)	271.41 ± 20.7	$\mathbf{418.42\pm13.64}$

(a) Performance on Cart-pole MDP Environment



(b) Training process of robust/regular Q-learning