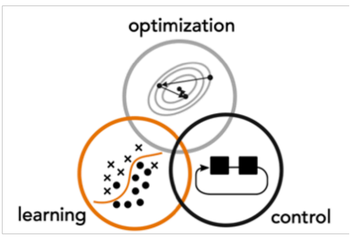
Regularization for Adversarial Robust Learning

Jie Wang †, Rui Gao ‡, Yao Xie †

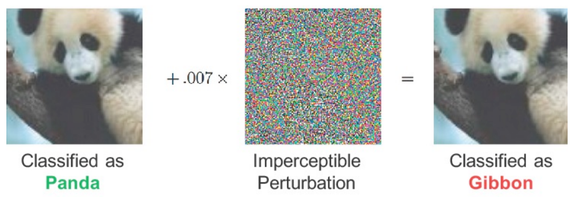
† H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

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



Motivation and Background

In Picture



In Text

A **inspire** movie. It is entire of feelings and **wonderful** behaving. I could have sat through it a seconds period. → **Positive**

A **touch** movie. It is entire of feelings and **good** behaving. I could have sat through it a seconds period. → **Negative**

Adversarial Training (Goodfellow et al, 2014)

$$\min_{\theta\in\Theta}\left\{\mathbb{E}_{z\sim\mathbb{P}_n}\left[\sup_{z':\ \|z-z'\|\le\rho}\ell(z';\theta)\right]\right\}$$

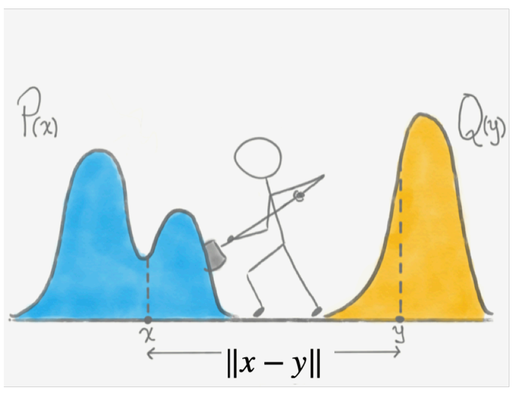
- $z\sim\mathbb{P}_n$: Data (e.g., feature-label pair)
- $\|z-z'\|\le\rho$: Perturbation Constriants
- $\ell(z';\theta)$: Loss Function

- **Cons**: Inner supremum is generally **nonconcave** in z
- **Heuristics**: Fast but **no theoretical guarantees of robustness**

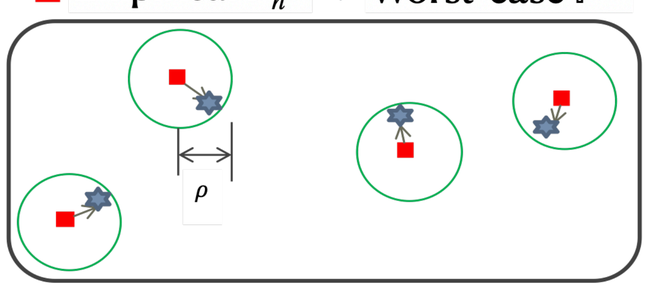
$$\Delta(\theta;z):=\arg\max_{\|\Delta\|\le\rho}\left\{\ell(z;\theta)+\nabla_z\ell(z;\theta)^\top\Delta\right\},\qquad\text{then perturb } z\leftarrow z+\Delta(\theta;z).$$

Wasserstein Distributionally Robust Optimization (DRO)

$$\begin{aligned}(\mathrm{AT})&=\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P}:\ \mathcal{W}_\infty(\mathbb{P},\mathbb{P}_n)\le\rho}\mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]\right\}\\&=\min_{\theta\in\Theta}\left\{\sup_{\mathbb{P},\gamma}\ \mathbb{E}_{z\sim\mathbb{P}}[\ell(z;\theta)]:\ \begin{array}{l}\gamma\in\Gamma(\mathbb{P}_n,\mathbb{P})\\ \gamma\text{-esssup}\ \|\zeta_1-\zeta_2\|\le\rho\end{array}\right\}\end{aligned}$$



$$\mathcal{W}_\infty(\mathbb{P},\mathbb{Q})=\inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})}\left\{\gamma\text{-esssup}\ \|x-y\|\right\}$$



Proposed Formulation

$$\min_{\theta\in\Theta}\sup_{\mathbb{P},\gamma}\left\{\mathbb{E}_{z\sim\mathbb{P}}\ [\ell(z;\theta)]-\eta\mathbb{D}_f(\gamma,\gamma_0):\ \begin{array}{l}\gamma\in\Gamma(\mathbb{P}_n,\mathbb{P})\\ \gamma\text{-esssup}\ \|x-y\|\le\rho\end{array}\right\}\triangleq V_{\mathrm{Primal}}$$

- f-**divergence (e.g., KL or χ^2):** $\mathbb{D}_f(\gamma,\gamma_0)=\int f\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\gamma_0}\right)\mathrm{d}\gamma_0$
- **Reference transport γ_0 is uniform**: For each $z\in\mathrm{supp}(\mathbb{P}_n)$,

  $\mathbb{Q}_z(\cdot)\triangleq\gamma_0(\cdot\mid z)$ is an uniform measure on $\mathbb{B}_\rho(z)$.

- **Under mild conditions, $V_{\mathrm{Primal}}=V_{\mathrm{dual}}$:**

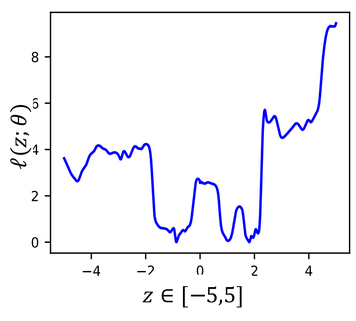
$$V_{\mathrm{Dual}}=\mathbb{E}_{\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}}\left\{\mu+\mathbb{E}_{z'\sim\mathbb{Q}_z}[(\eta f)^*(\ell(z';\theta)-\mu)]\right\}\right]$$

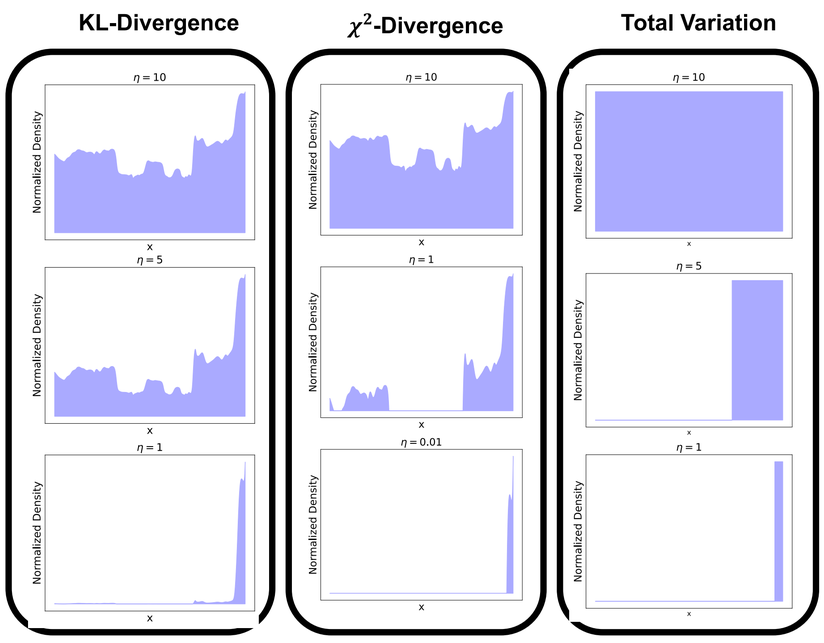
Penalized f-divergence DRO with reference measure $\mathbb{Q}_z$

$$\frac{\mathrm{d}\mathbb{P}^*(\omega)}{\mathrm{d}\omega}=\mathbb{E}_{z\sim\mathbb{P}_n}\left[\alpha_z\cdot\mathbf{1}\{\|\omega-z\|\le\rho\}\cdot(\eta f)^{*\prime}(\ell(\omega;\theta)-\mu_z^*)\right]$$

- α_z: Normalizing Constant
- $\mathbf{1}\{\|\omega-z\|\le\rho\}$: Support Constraint
- $(\eta f)^{*\prime}(\ell(\omega;\theta)-\mu_z^*)$: Density contributed by z

- Visualization of worst-case distribution
- $\ell(z;\theta)$ is a feed-forward neural network





Optimization Algorithm

- (Regularized-AT) as a bilevel optimization with many lower level constraints:

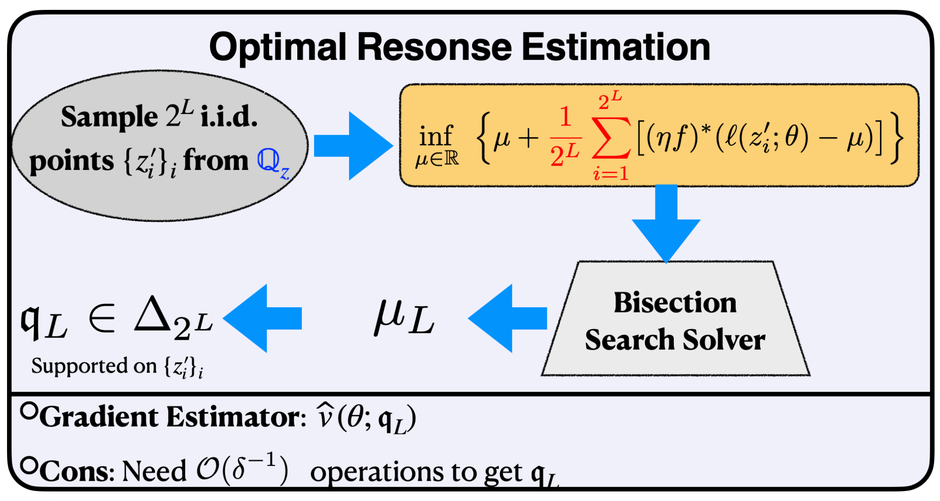
$$\begin{aligned}&\min_{\theta\in\Theta}\left\{F(\theta)=\mathbb{E}_{z\sim\mathbb{P}_n}\left[R(\theta;z)\right]\right\}\\&\text{where}\quad R(\theta;z)=\inf_{\mu\in\mathbb{R}}\left\{\mu+\mathbb{E}_{z'\sim\mathbb{Q}_z}\left[(\eta f)^*(\ell(z';\theta)-\mu)\right]\right\},\quad\forall z\\&\qquad\qquad=\sup_{\mathrm{q}\in\mathcal{P}}\left[\mathbb{E}_{z'\sim\mathrm{q}}[\ell(z';\theta)]-\eta\mathbb{E}_{z'\sim\mathbb{Q}_z}\left[\phi\left(\frac{\mathrm{d}\mathrm{q}(z')}{\mathrm{d}\mathbb{Q}_z(z')}\right)\right]\right],\quad\forall z\end{aligned}$$

- For iterations $t=0,1,\ldots,T-1$:

  (i) Estimate biased (sub-)gradient; (ii) Perform projected gradient descent.

Algorithm	Naive Estimator		Random Sampling Estimator	
Loss $\ell(z,\cdot)$	Convex	Nonconvex Smooth	Convex	Nonconvex Smooth
Choice of f-divergence	Arbitrary	**KL-divergence**	Arbitrary	**KL-divergence**
Complexity	$\tilde{O}(\delta^{-3})$	$\tilde{O}(\delta^{-6})$	$\tilde{O}(\delta^{-2})$	$\tilde{O}(\delta^{-4})$

- $\nabla F(\theta)=\mathbb{E}_{z\sim\mathbb{P}_n}\mathbb{E}_{z'\sim\mathrm{q}^*(z;\theta)}[\nabla_\theta\ell(z';\theta)]$, where optimal response $\mathrm{q}^*(z;\theta)$ solves $R(\theta;z)$.



Random Sampling Gradient Estimator

$$\widehat{v}(\theta;\mathrm{q}_{L+1})=\widehat{v}(\theta;\mathrm{q}_1)+\sum_{\ell=1}^{L}\left[\widehat{v}(\theta;\mathrm{q}_{\ell+1})-\widehat{v}(\theta;\mathrm{q}_\ell)\right]$$

$$=\widehat{v}(\theta;\mathrm{q}_1)+\sum_{\ell=1}^{L}p_\ell\frac{\widehat{v}(\theta;\mathrm{q}_{\ell+1})-\widehat{v}(\theta;\mathrm{q}_\ell)}{p_\ell}=\mathbb{E}_{\ell\sim\mathbb{P}_\ell}\left[\widehat{v}(\theta;\mathrm{q}_1)+\frac{\widehat{v}(\theta;\mathrm{q}_{\ell+1})-\widehat{v}(\theta;\mathrm{q}_\ell)}{p_\ell}\right].$$

- Sample ℓ according to pmf $p_\ell\propto 2^{-\ell}$, $\sum_\ell p_\ell=1$, then construct

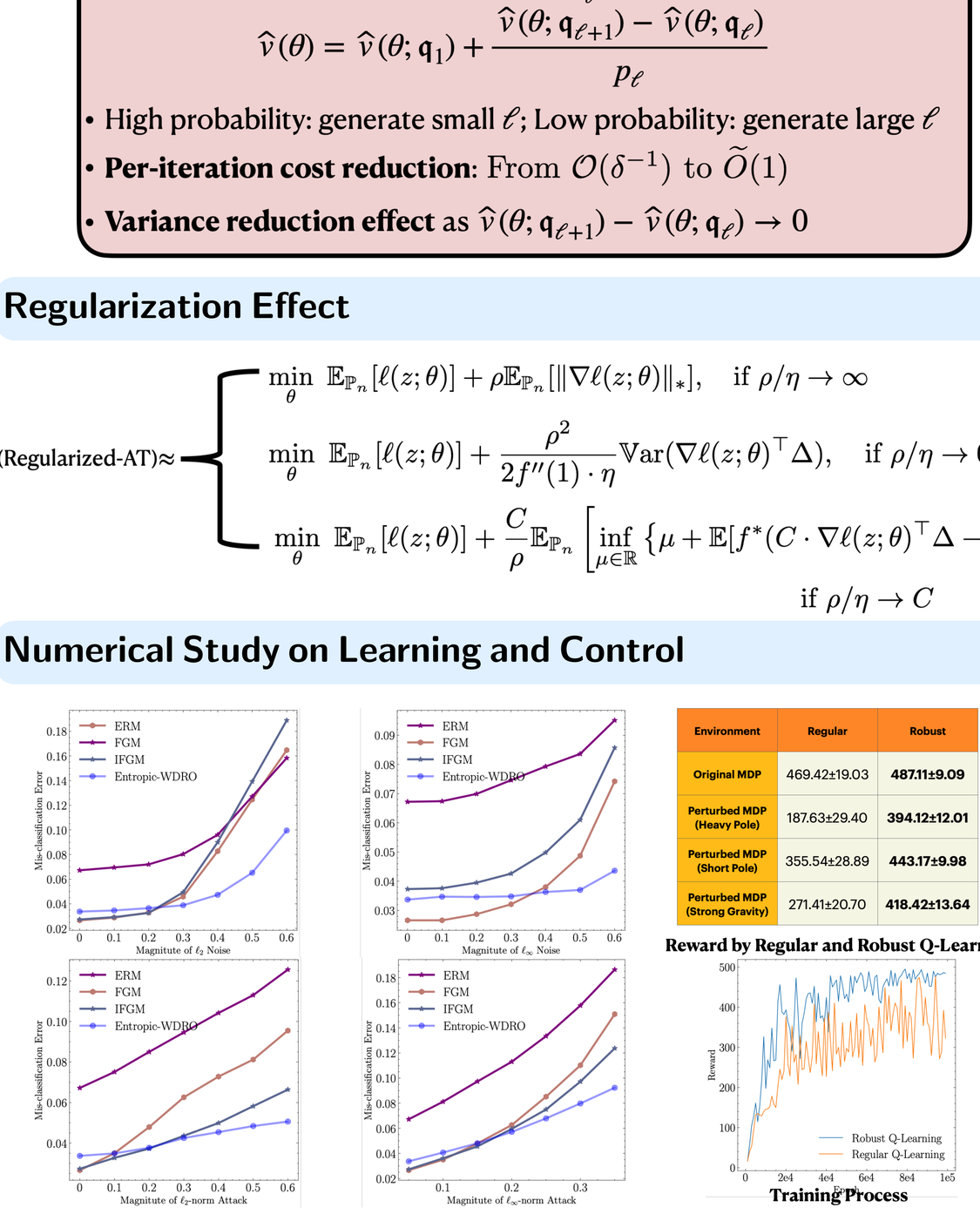
$$\widehat{v}(\theta)=\widehat{v}(\theta;\mathrm{q}_1)+\frac{\widehat{v}(\theta;\mathrm{q}_{\ell+1})-\widehat{v}(\theta;\mathrm{q}_\ell)}{p_\ell}$$

- High probability: generate small ℓ; Low probability: generate large ℓ
- **Per-iteration cost reduction**: From $\mathcal{O}(\delta^{-1})$ to $\widetilde{O}(1)$
- **Variance reduction effect** as $\widehat{v}(\theta;\mathrm{q}_{\ell+1})-\widehat{v}(\theta;\mathrm{q}_\ell)\to 0$

Regularization Effect

$$\text{(Regularized-AT)}\approx\begin{cases}\min_\theta\ \mathbb{E}_{\mathbb{P}_n}[\ell(z;\theta)]+\rho\mathbb{E}_{\mathbb{P}_n}[\|\nabla\ell(z;\theta)\|_*], & \text{if } \rho/\eta\to\infty\\ \min_\theta\ \mathbb{E}_{\mathbb{P}_n}[\ell(z;\theta)]+\frac{\rho^2}{2f''(1)\cdot\eta}\mathbb{V}\mathrm{ar}(\nabla\ell(z;\theta)^\top\Delta), & \text{if } \rho/\eta\to 0\\ \min_\theta\ \mathbb{E}_{\mathbb{P}_n}[\ell(z;\theta)]+\frac{C}{\rho}\mathbb{E}_{\mathbb{P}_n}\left[\inf_{\mu\in\mathbb{R}}\{\mu+\mathbb{E}[f^*(C\cdot\nabla\ell(z;\theta)^\top\Delta-\mu)]\}\right], & \text{if } \rho/\eta\to C\end{cases}$$

Numerical Study on Learning and Control



Environment	Regular	Robust
Original MDP	469.42±19.03	**487.11±9.09**
Perturbed MDP (Heavy Pole)	187.63±29.40	**394.12±12.01**
Perturbed MDP (Short Pole)	355.54±28.89	**443.17±9.98**
Perturbed MDP (Strong Gravity)	271.41±20.70	**418.42±13.64**

Reward by Regular and Robust Q-Learning