Jeremiah Birrell
Department of Mathematics
Texas State University
Reza Ebrahimi
School of Information Systems and Management
University of South Florida

Adversarially Robust Learning with Optimal-Transport Regularized Divergences

Vulnerability of machine learning methods to adversarial attack: Traditional deep learning methods are known to be **vulnerable to attacks** where a **small perturbation** of the input (e.g., an image) that is **imperceptible** to a human observer can **cause a trained classifier to fail**.

Original (PGD) adversarial training method: PGD[2] trains the loss $\mathcal{L}_\theta$ by **perturbing training samples** z within a metric-space ball of size ϵ to create **adversarial samples** $\tilde{z}$ that are then used in training:

$$\inf_\theta E_{P_n}\left[\sup_{\tilde{z}:d(z,\tilde{z})\leq\epsilon}\mathcal{L}_\theta(\tilde{z})\right],$$

where P_n denotes the empirical distribution of the training samples.

PGD is an example of distributionally robust optimization (DRO): In DRO the empirical distribution is replaced by the **worst-case adversarial distribution** Q in some model neighborhood $\mathcal{U}(P_n)$ around P_n:

$$\inf_\theta \sup_{Q\in\mathcal{U}(P_n)} E_Q[\mathcal{L}_\theta].$$

Robustness is increased by training for the worst case.

When using adversarial training, **the choice of model neighborhood is important for performance!**

Robust Optimization and Simulation of Complex Stochastic Systems
Sep 13 - 15, 2024

ARMOR$_D$ Method[1]**:** Our DRO-based approach **improves model robustness** by both **adversarially transporting** (via an optimal transport cost) and **adversarially re-weighting** (via an information divergence) samples during training.

ARMOR$_D$ can be **combined** with other popular methods, e.g., that modify the training loss, such as PGD[2], TRADES[3], MART[4], and UDR[5] to yield **improved performance** when under adversarial attack:

	CIFAR10 Performance		
Defense	AutoAttack	PGD^{200}	Nat.
PGD	42.5%	46.0%	**86.40%**
$UDR\text{-}PGD$	48.47%	52.95%	81.71%
$ARMOR_\alpha\text{-}UDR\text{-}PGD$	**48.63%**	**53.62%**	80.29%
$TRADES$	49.1%	51.9%	80.8%
$UDR\text{-}TRADES$	49.9%	53.6%	**84.4%**
$ARMOR_\alpha\text{-}TRADES$	**51.4%**	**53.74%**	80.76%
$MART$	48.2%	53.3%	**81.9%**
$UDR\text{-}MART$	49.1%	54.1%	80.1%
$ARMOR_\alpha\text{-}MART$	**50.6%**	**56.22%**	81.03%

Optimal-Transport Regularized Divergences:

$$D^c(\nu\|\mu) := \inf_{\eta\in\mathcal{P}(\mathcal{Z})}\{D(\eta\|\mu)+C(\eta,\nu)\}$$

C is an **optimal transport cost** for a cost function c,

$$C(\mu,\nu) := \inf_{\pi:\pi_1=\mu,\pi_2=\nu}\int c(z,\tilde{z})\pi(dzd\tilde{z})$$

D is an **information divergence**, e.g., an f-divergence,

$$D_f(\mu\|\nu) = E_\nu[f(d\mu/d\nu)].$$

Properties (under appropriate assumptions):

- **Divergence property:** $D^c(\nu\|\mu)\geq 0$ and $D^c(\nu\|\mu)=0$ if and only if $\nu=\mu$. This implies $D^c(\nu\|\mu)$ **quantifies the discrepancy** between ν and μ.
- **Optimizer:** there **exists a unique optimizer**, η_*, with
  $$D^c(\nu\|\mu) = D(\eta_*\|\mu)+C(\eta_*,\nu)$$
- **DRO neighborhoods:** The DRO neighborhoods $\{Q : D^c(Q\|P_n)\leq\epsilon\}$ are **closed convex sets.**
- **Interpolation property:** D^c **interpolates between** D **and** C as follows
  $$\lim_{r\to 0^+} r^{-1}D^{rc}(\nu\|\mu) = C(\mu,\nu), \quad \lim_{r\to\infty} D^{rc}(\nu\|\mu) = D(\nu\|\mu)$$

Computationally-Tractable Dual-Formulation of Adversarially-Robust Training Problem: We use the D^c-DRO neighborhoods to obtain a **novel adversarial training method** and, via convex duality, obtain the **computationally tractable form**

$$\inf_{\theta\in\Theta}\sup_{Q:D^c_f(Q\|P_n)\leq\epsilon} E_Q[\mathcal{L}_\theta]$$
$$=\inf_{\lambda>0,\rho\in\mathbb{R},\theta\in\Theta}\left\{\epsilon\lambda+\rho+\lambda E_{P_n}\left[f^*(\lambda^{-1}(\mathcal{L}^c_{\theta,\lambda}(z_i)-\rho))\right]\right\}$$

where

$$\mathcal{L}^c_{\theta,\lambda}(z) := \sup_{\tilde{z}\in\mathcal{Z}}\{\mathcal{L}_\theta(\tilde{z})-\lambda c(z,\tilde{z})\}.$$

Our work generalizes the optimal-transport DRO results of [6], [7], and is related to the DRO method [8].

References:

[1] J. Birrell, M. Ebrahimi, arXiv:2309.03791, 2023
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, ICLR, 2018
[3] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, ICML, 2019
[4] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, ICLR, 2020
[5] A. T. Bui, T. Le, Q. H. Tran, H. Zhao, and D. Phung, ICLR, 2022
[6] P. Mohajerin Esfahani and D. Kuhn, Mathematical Programming, 2018
[7] J. Blanchet and K. Murthy, Mathematics of Operations Research, 2019
[8] J. Blanchet, D. Kuhn, J. Li, and B. Taskesen, arXiv:2308.05414, 2023