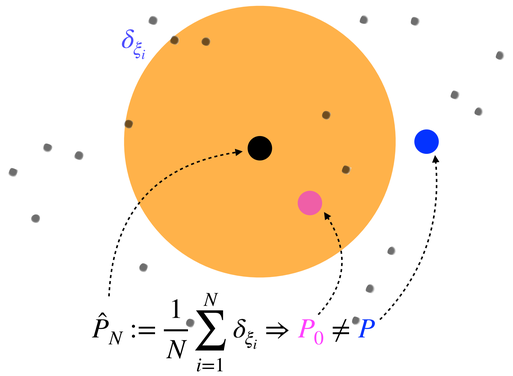
Classical Empirical Risk Minimization

$$\min_\theta \frac{1}{N}\sum_{i=1}^{N} l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- Do well on average; can bound e.g., $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$
- Not robust to data distribution shifts, when $P \neq P_0$



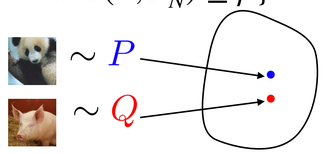
Distributionally Robust Optimization (DRO)

$$\text{(DRO)} = \min_\theta \sup_{P \in \mathscr{M}} \mathbb{E}_P l(\theta, \xi)$$

- Do well under a local worst-case distribution P
- Distribution shift described by an ambiguity set $\mathscr{M}$

Example: MMD-ball $\mathscr{M} = \{P : \text{MMD}(P, \hat{P}_N) \leq \rho\}$

$$\begin{aligned}\text{MMD}_{\mathscr{H}}(Q, P) &:= \sup_{\|f\|_{\mathscr{H}} \leq 1} \int f d(Q - P) \\ &= \mathbb{E}_{x,x' \sim Q} k(x, x') + \mathbb{E}_{y,y' \sim P} k(y, y') \\ &\quad - 2\mathbb{E}_{x \sim Q, y \sim P} k(x, y)\,.\end{aligned}$$



- Bound performance under P ($\neq P_0$) beyond statistical fluctuation (classical learning theory)

Kernel Distributionally Robust Optimization

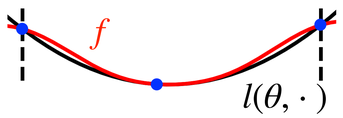
(Kernel DRO) [Zhu et al. AISTATS 2021]

How to solve MMD-constrained (DRO)?

Theorem (simplified). *Primal DRO problem is equivalent to the following dual kernel machine learning problem, i.e., (DRO)=(K-DRO).*

$$\text{(K-DRO)} \min_{\theta, f \in \mathscr{H}} \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon \|f\|_{\mathscr{H}} \quad \text{subject to } l(\theta, \cdot) \leq f$$

Geometric intuition: using kernel approximations as robust surrogate losses



Cf. Kantorovich duality in optimal transport (OT)

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

InstaDeep™

WIAS

Adversarially Robust Kernel Smoothing

Jia-Jie Zhu[1,3], Christina Kouridi[2,3], Yassine Nemmour[3], Bernhard Schölkopf[3]

[1]Weierstrass Institute for Applied Analysis and Stochastics
Berlin, Germany

[2]InstaDeep Ltd.
London, United Kingdom

[3]Max Planck Institute for Intelligent Systems
Tübingen, Germany

Email: zhu@wias-berlin.de
Website: jj-zhu.github.io

AISTATS

Adversarially Robust Kernel Smoothing

(ARKS) [Zhu et al. AISTATS 2022]

A constructive feasible solution to Kernel DRO:

$$f(x) = \sup_z \{l(\theta, z) k(z, x)\}$$

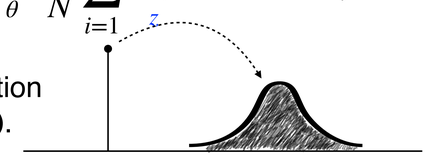
kernel choice: $k(x, x') := e^{-c(x,x')/\sigma}$.
c: transport cost in OT, $\sigma > 0$: bandwidth

✔infinite constraint satisfied: $l(\theta, x) \leq f(x), \forall x$

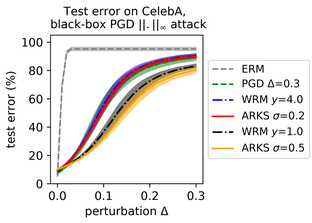
✔applies to loss with practical models, e.g., DNN

$$\text{(ARKS)} \quad \min_\theta \frac{1}{N}\sum_{i=1}^{N} \sup_z \{l(\theta, z) k(z, \xi_i)\}$$

Intuition: modeling adversarial perturbation using density $k(z, \xi_i)$.



Examples. Certified adversarially robust deep learning



Distributional robustness certificate.

$$\sup_{\mathscr{W}_c(P, P_0) \leq \rho} \mathbb{E}_P \ln l(\hat{\theta}, \xi) \leq \ln \underbrace{\left\{ \frac{1}{N}\sum_{i=1}^{N} \sup_z \{l(\hat{\theta}, z) k(z, \xi_i)\} \right\}}_{\text{ARKS objective}} + \frac{\rho}{\sigma} + \epsilon_N$$

$\mathscr{W}_c(\cdot, \cdot)$: OT metric associated with transport cost c

Future directions

- Design specific kernels for robustness beyond norm-balls
- Physics, information geometry, and general dynamic OT
- Causal inference via distributional robustness

References

- Zhu, J.-J., Jitkrittum, W., Diehl, M. & Schölkopf, B. Kernel Distributionally Robust Optimization. AISTATS 2021
- Zhu, J.-J., Kouridi, C., Nemmour, Y. & Schölkopf, B. Adversarially Robust Kernel Smoothing. AISTATS 2022

Code

- KDRO: https://github.com/jj-zhu/kdro
- ARKS: https://github.com/christinakouridi/arks