Distributionally Robust Learning and Optimization

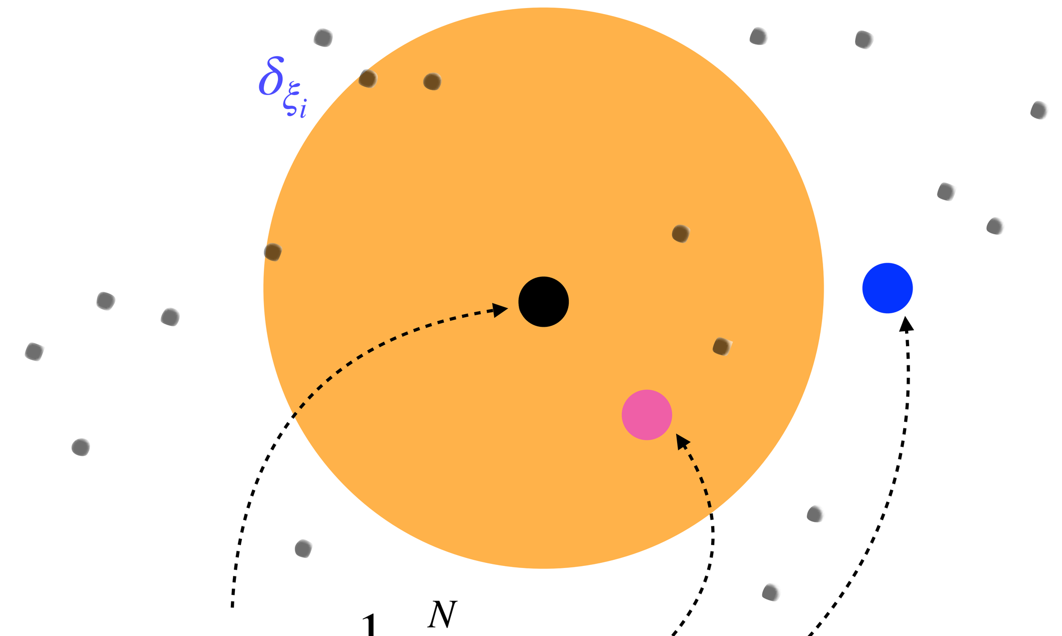
Jia-Jie Zhu

Empirical Inference Department
Max Planck Institute for Intelligent Systems
Tübingen, Germany
&
Weierstrass Institute for Applied Analysis and Stochastics
Berlin, Germany

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 against data distribution shifts, when $P \neq P_0$



$\hat P_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i} \Rightarrow P_0 \neq P$

Distributionally Robust Optimization (DRO)

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

- Do well under a local worst-case distribution P
- Distribution shift described by an ambiguity set $\mathcal{M}$.
  Example: MMD-ball$\{P : \text{MMD}(P, \hat P_N) \leq \rho_N\}$where ρ_N can be chosen using previous works [Tolstikhin et al. 2017, Gretton et al. 2012]
- Bound performance under P ($\neq P_0$) beyond statistical fluctuation (classical learning theory)

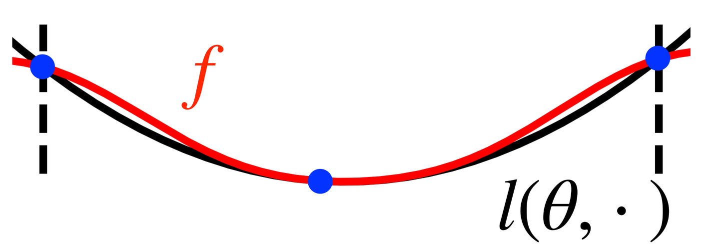
Kernel Distributionally Robust optimization

(Kernel DRO) [Z. et al. 2021]

Theorem (simplified). *Primal DRO problem is equivalent to the following dual problem, i.e., (P)=(D).*

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

Geometric intuition: using kernel approximations as robust surrogate losses



Cf. Kantorovich duality in optimal transport (OT)

Adversarially Robust Kernel Smoothing

(ARKS) [Z. et al. 2022]

A constructive feasible solution to Kernel DRO:

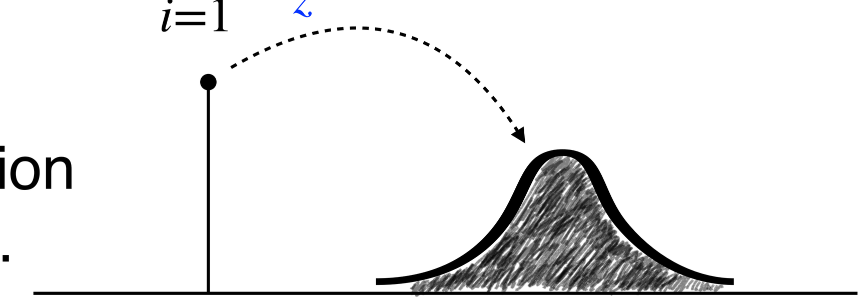
$$f(x) = \sup_z \{l(z)k(z, x)\}$$

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

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

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

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



Example. Certified adversarially robust deep learning

Classify the presence of sunglasses under adversarial attacks (cf. references)

Distributional robustness certificate.

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

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

Future directions

- Causal inference using DRO
- Dynamical systems modeling & control; dynamic OT
- Multi-stage adjustable DRO
- Large-scale DR learning

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
- Tolstikhin, I., Sriperumbudur, B. & Muandet, K. Minimax Estimation of Kernel Mean Embeddings. JMLR 2017
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. JMLR 2012

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