Kernel Distributionally Robust Optimization

Generalized Duality Theorem and Stochastic Approximation

Jia-Jie Zhu*, Wittawat Jitkrittum*,
Moritz Diehl**, Bernhard Schölkopf*

*Empirical Inference Department
Max Planck Institute for Intelligent Systems
Tübingen, Germany

**Department of Microsystems Engineering
Department of Mathematics, University of Freiburg
Freiburg, Germany

The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)
April 13 - 15, 2021

Code: https://github.com/jj-zhu/kdro

Combine the strengths of ERM and RO: distributionally robust optimization (DRO)

loss/cost ... uncertain

(ERM) $$\min_{\theta} \mathbb{E}_{\xi \sim \hat{P}} l(\theta, \xi)$$

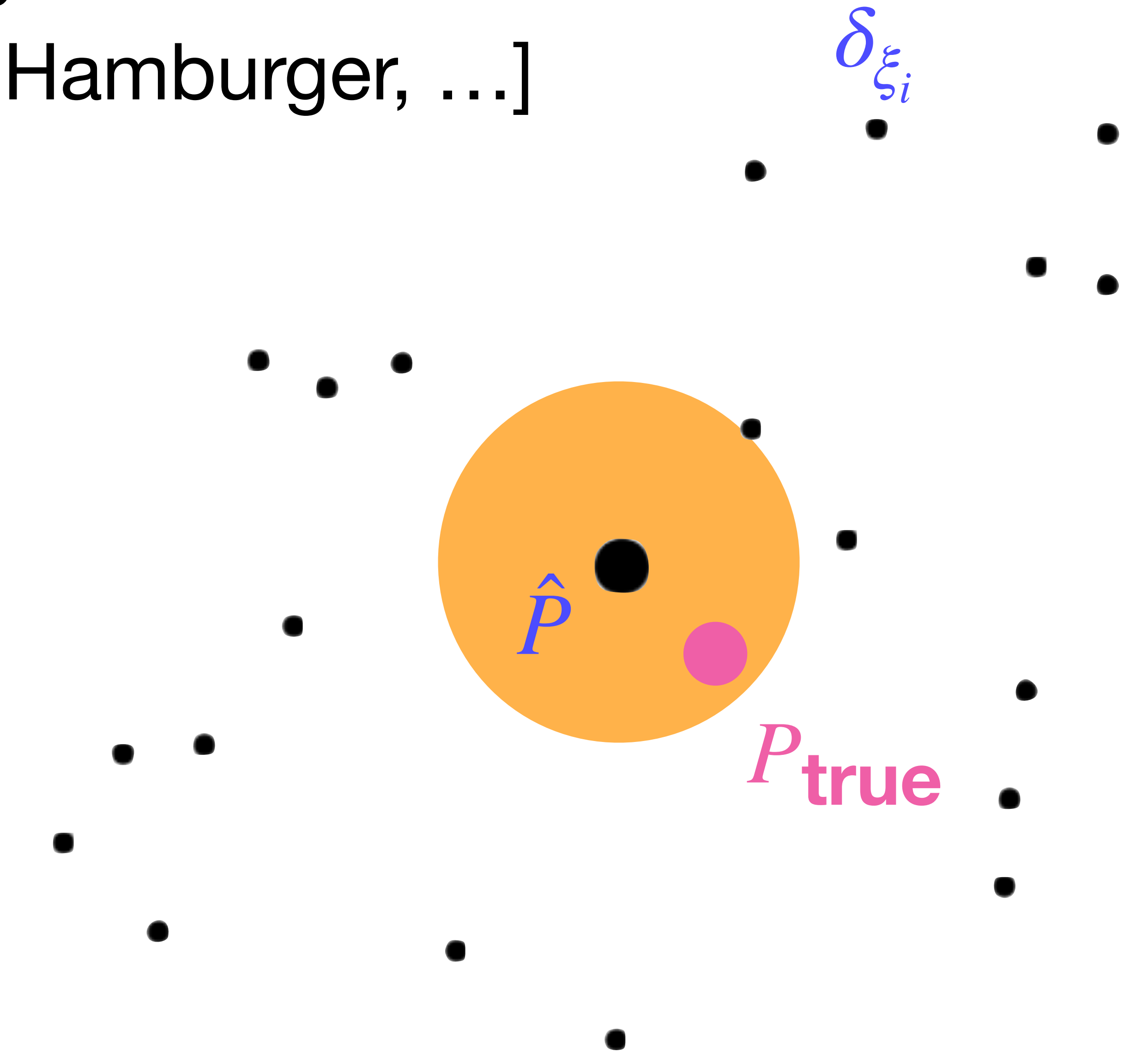
Empirical data dist. $\hat{P} = \sum_{i=1}^{N} \frac{1}{N} \delta_{\xi_i}$

(RO) $$\min_{\theta} \sup_{\xi \in \mathcal{U}} l(\theta, \xi)$$

Uncertainty set

$$\min_{\theta} \boxed{\sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi) \text{ (DRO)} \quad \text{[Delage and Ye 2010, Scarf 1958]}}$$

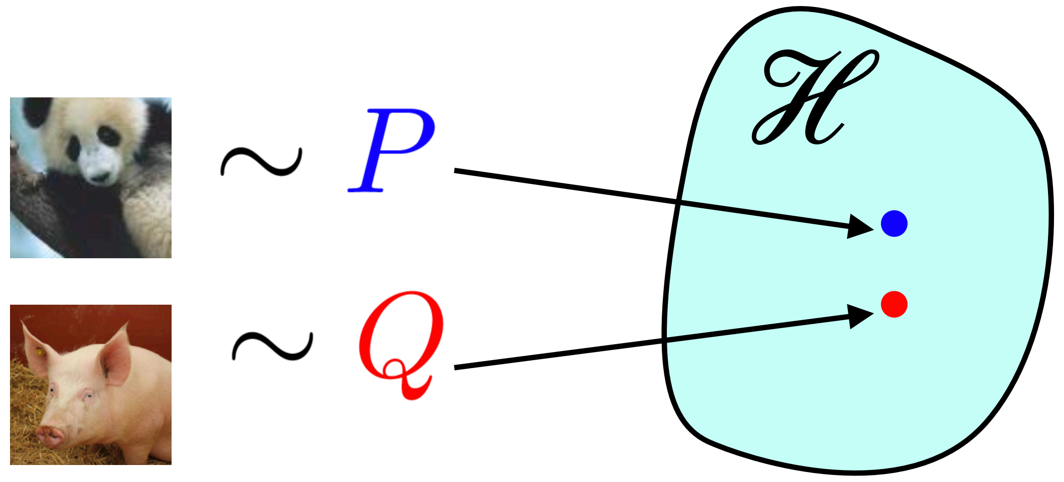
Find the worst-case distribution!
Problem of Moments
[Stieltjes, Hausdorff, Hamburger, …]



- Robustifies against a set of probability measures $\mathcal{K}$ (***ambiguity set***), e.g.,
  - $\mathcal{K}$ can be a metric-ball centered at $\hat{P}$, e.g., using Wasserstein metric [Esfahani&Kuhn'18, Zhao&Guan'18, Gao&Kleywegt'16, …], sets in RKHSs [**this paper**].
    - Relevance to machine learning: one can quantify the empirical mean convergence rate $\gamma(\hat{P}, P_{\text{true}}) \leq \epsilon$, e.g., [Tolstikhin et al.'17].
    - **Active research area. Also related to data-driven RO.**

Smooth is robust: Kernel DRO

(DRO) $\min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$



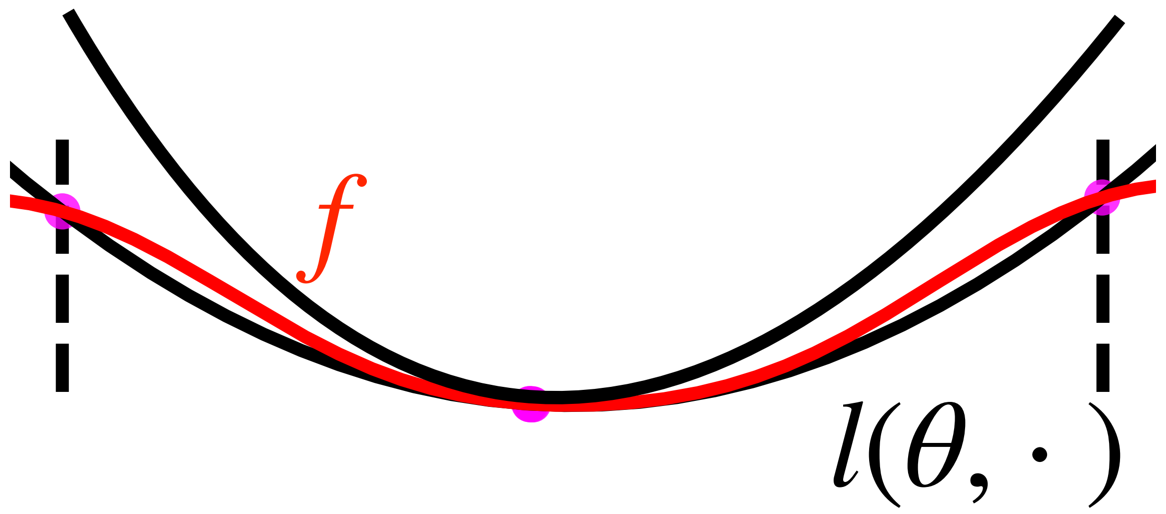
(P) $\min_{\theta} \sup_{P,\mu} \left\{ \mathbb{E}_P l(\theta, \xi) : \int \phi \, dP = \mu, \mu \in \mathscr{C} \right\}$

Theorem (**Generalized variational duality**). DRO (P) is equivalent to solving

(D) $\min_{\theta, f \in \mathscr{H}} \delta^*_{\mathscr{C}}(f) \quad \text{subject to } l(\theta, \cdot) \le f,$

$\delta^*_{\mathscr{C}}(f)$ is the support function, e.g., $\mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}}$.

Geometric intuition



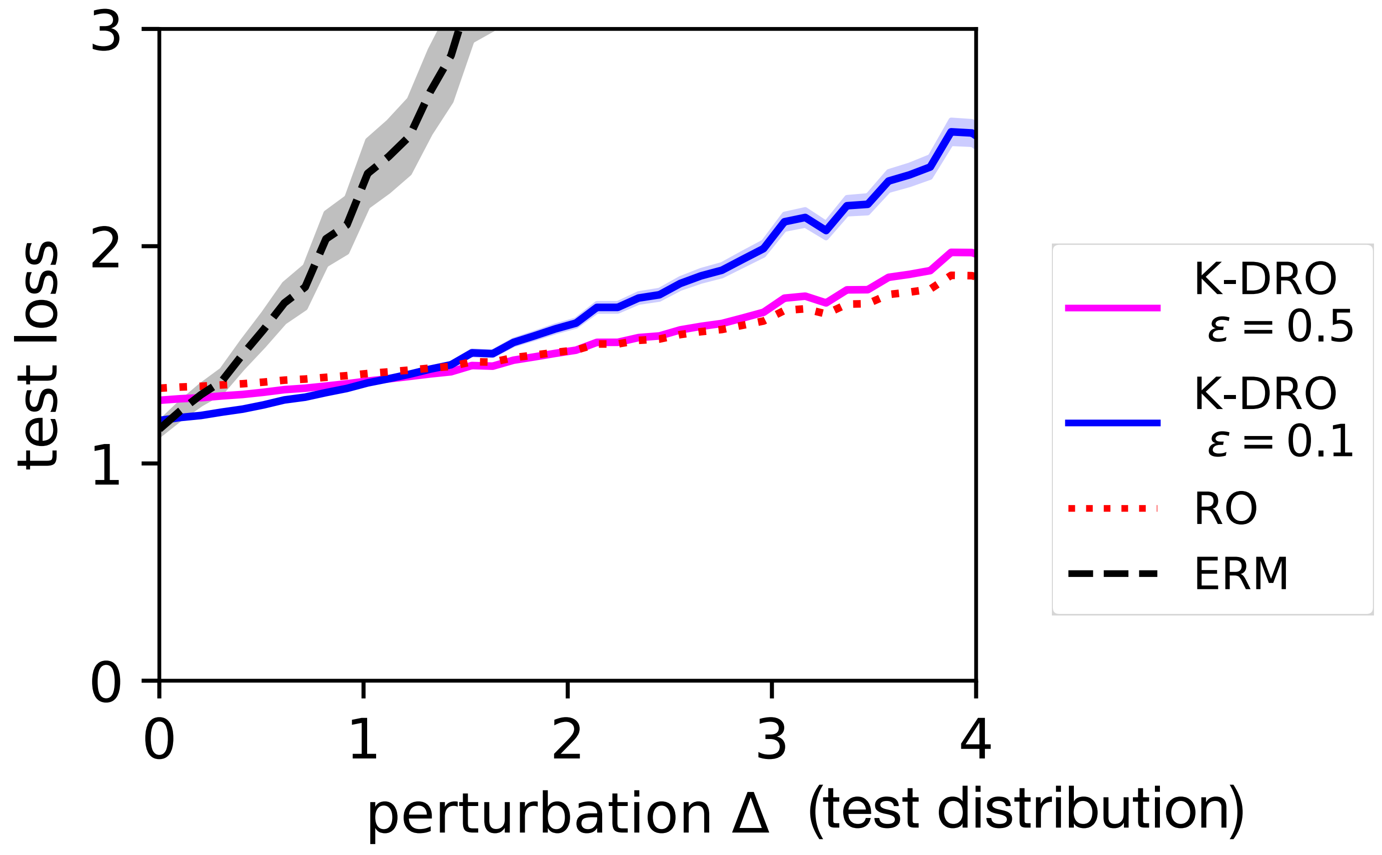
Smoothness of f ↔ Distributional robustness (↔ Size of $\mathscr{H}$)

Intuition: flatten the curve, smooth is robust

Example. Uncertain least squares

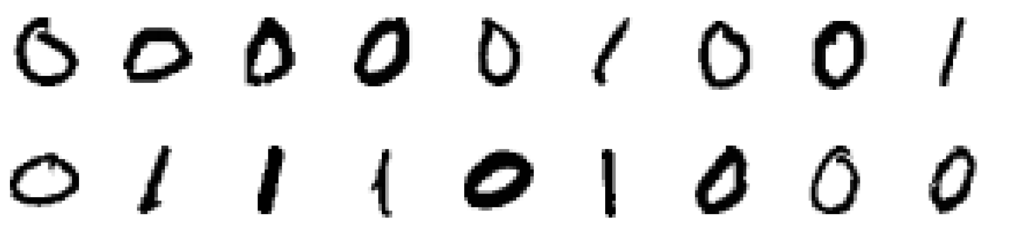
$$\text{minimize } l(\theta, \xi) := \|A(\xi) \cdot \theta - b\|_2^2$$

Given historical samples $\xi_1, \xi_2, \ldots, \xi_N$

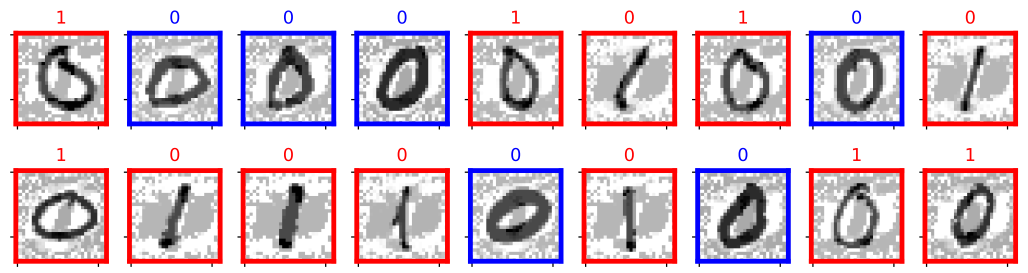


Example. Neural network classification

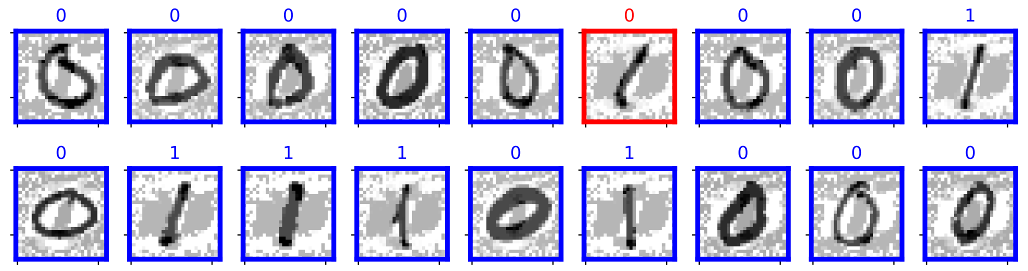
Clean data



Perturbed data



Kernel DRO solution



Conclusions

- Distributional shift is inevitable for machine learning and AI.
    - DRO is a principled tool for decision-making under distribution-shift based on RO.
- We have established a generalized duality theorem for solving DRO with general ambiguity sets and IPM, with weak assumptions on the loss.
    - Maximizing w.r.t. a distribution → finding a smooth function
- Takeaway
    - Use universal RKHSs as dual spaces for DRO
    - **Flatten the curve**
    - **Smooth is robust**

Jia-Jie Zhu
jzhu@tuebingen.mpg.de

Empirical Inference Department
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Co-authors

AISTATS, April 2021

Code: https://github.com/jj-zhu/kdro