Distributionally Robust Optimization using Integral Probability Metrics and Reproducing Kernel Hilbert Spaces

Jia-Jie Zhu

jj-zhu.github.io

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

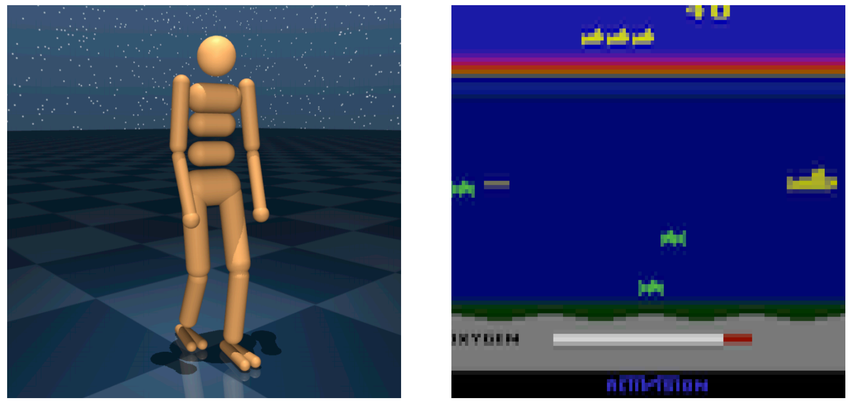
Based on joint work with
Wittawat Jitkrittum (Google Research), **Moritz Diehl** (Uni. Freiburg), **Bernhard Schölkopf** (MPI Tübingen)

SIAM Conference on Optimization (OP21)
July 21, 2021

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

Robust optimization

Empirical risk minimization (ERM)
(sample average approximation (SAA))



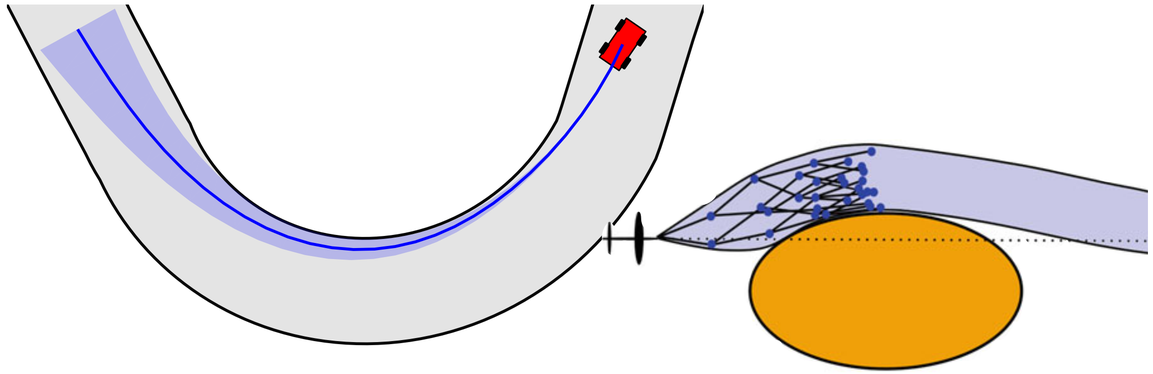
loss/cost, uncertain

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

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

- Do well on average
- Strength: high-performance (optimal)
- Weakness: fragile — adversarial attacks, sim2real transfer, safety/off-policy in RL

Robust optimization (RO)
(robust control, games)



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

- Do well in the worst case
- Strength: robustness
- Weakness: conservative — worst case doesn't often happen

Image credit: Mnih'13, MuJuCo, Houska and Villanueva '19, Hewing et al.'18

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

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

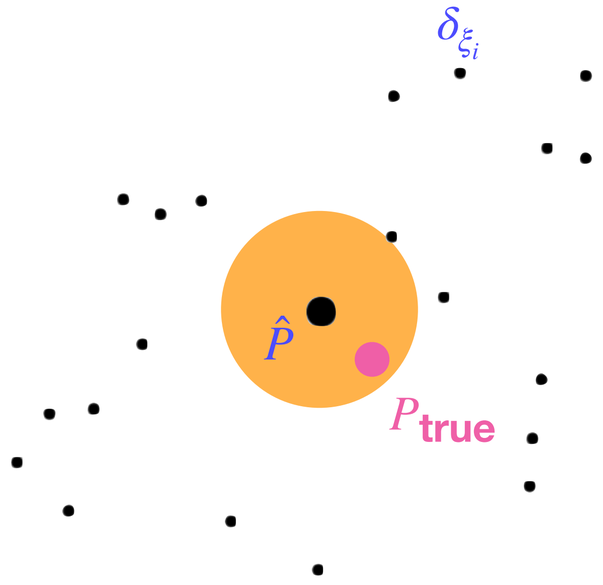
(RO) $\min_{\theta} \sup_{\xi\in\mathcal{U}} l(\theta,\xi)$

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

[Delage and Ye 2010, Scarf 1958]

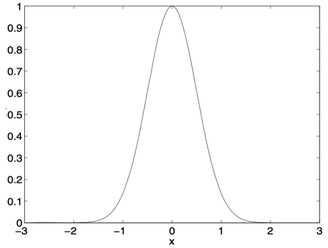
Find the worst-case distribution!
Problem of Moments [Stieltjes, Hausdorff, Hamburger, …]

- 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, sets in RKHSs [**this talk**].
    - One way of constructing ambiguity region: one can quantify the empirical mean convergence rate $\gamma(\hat{P}, P_{\text{true}}) \leq \epsilon$.
      - **Active research area: choose better ambiguity regions**
    - **This talk focuses on optimization and functional analysis**



Preliminary: learning with kernels (functions)

- A kernel is a symmetric function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, e.g., Gaussian kernel $k(x, x') = \exp\left(-\|x - x'\|_2^2 \,/\, 2\sigma^2\right)$.



- A p.d. k corresponds to a Hilbert space $\mathcal{H}$ (RKHS), which satisfies the **reproducing property** $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$, $\phi(x) := k(x, \cdot\,)$ is the **canonical feature** of $\mathcal{H}$.
- If $\mathcal{H}$ is a large (dense in C), $\gamma_{\mathcal{H}}$ is a metric on $\mathcal{P}$.
- We can generalize to the more general **integral probability metric** (IPM)

$$\mathrm{IPM}(\mathcal{F}; P, Q) := \sup_{f \in \mathcal{F}} \int f \, d(P - Q).$$

$\sim P$

$\sim Q$

$\gamma_{\mathcal{H}}(P, Q)$

$\mathcal{P}$ $\quad \mu_P := \int \phi \, dP \quad$ $\mathcal{H}$

duality

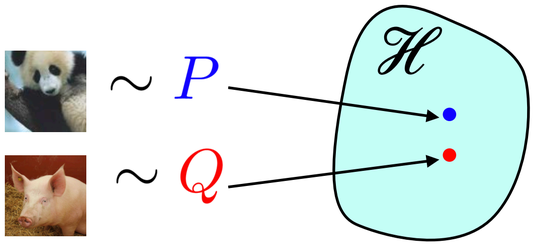
$\mu := \int \phi \, dP$ is the *(kernel)* ***mean embedding*** of P in $\mathcal{H}$

μ can be viewed as a generalized moment vector
e.g., let $\phi(x) = [x, x^2]^\top$ (related: Lasserre moment-SOS)

Illustration design inspired by Gretton, Sutherland, Jitkrittum NeurIPS 2019 tutorial

Smooth is robust: Kernel DRO

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



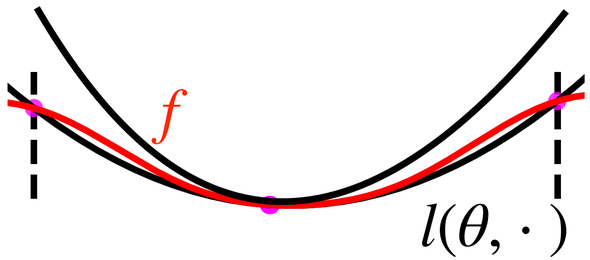
(P) $\min_{\theta} \sup_{P,\mu} \left\{ \mathbb{E}_P l(\theta, \xi) : \int \phi \, dP = \mu, \mu \in \mathscr{C} \right\}$

Theorem (**Kernel DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving

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

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

Geometric intuition



Smoothness of f ↭ Distributional robustness (↭ Size of $\mathcal{H}$)

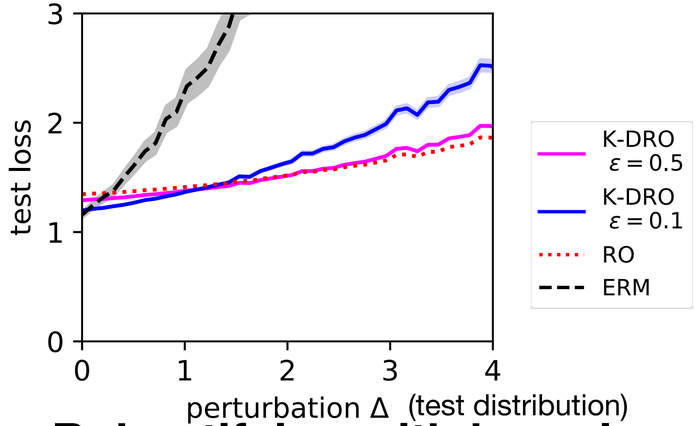
Intuition: flatten the curve, smooth is robust

Example. Uncertain least squares

[El Ghaoui Lebret '97]

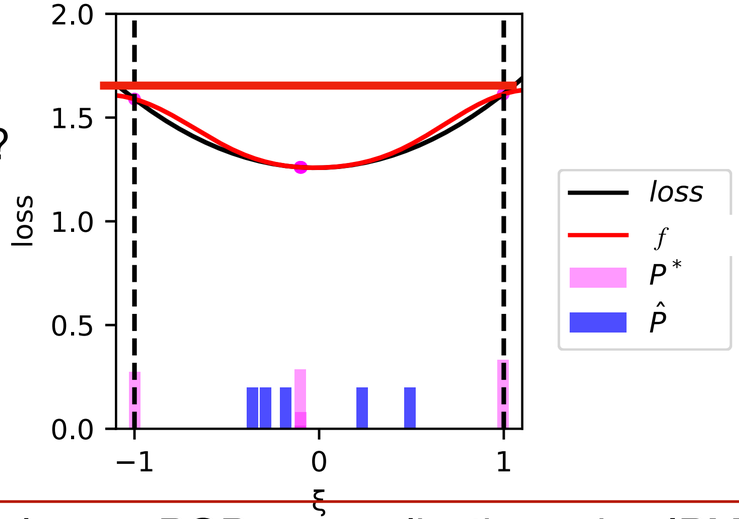
$\text{minimize } l(\theta, \xi) := \|A(\xi) \cdot \theta - b\|_2^2$

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



Robustifying with kernels

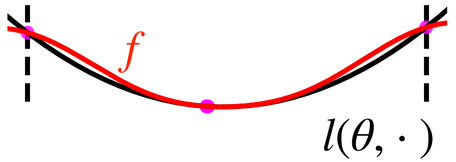
What if $f \equiv c \in \mathbb{R}$?



Space $\mathcal{H}$: special case: POP; generalization using IPM, e.g., W-1

Distributionally robust nonlinear optimization for machine learning and control

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

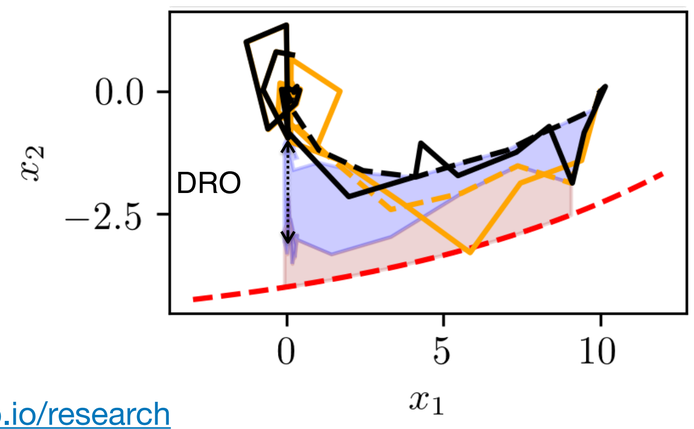


l: general nonlinear function, i.e., loss with DNN, $l \notin \mathscr{H}$. Kernel DRO handles this by finding a majorant $f \in \mathscr{H}$

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

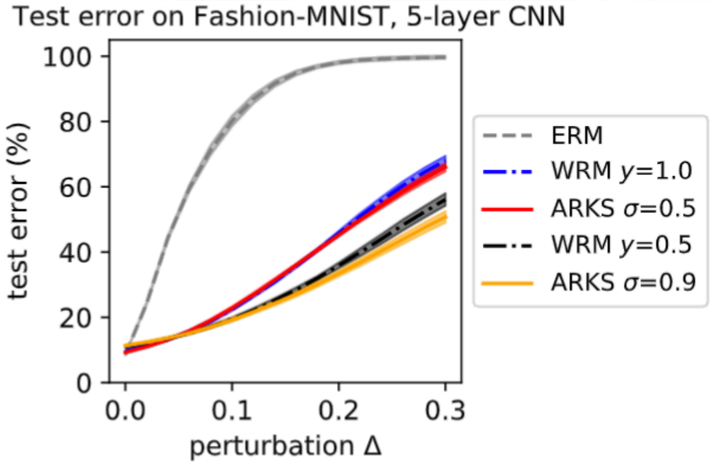
DRO for stochastic model predictive control (MPC) with nonlinear constraints

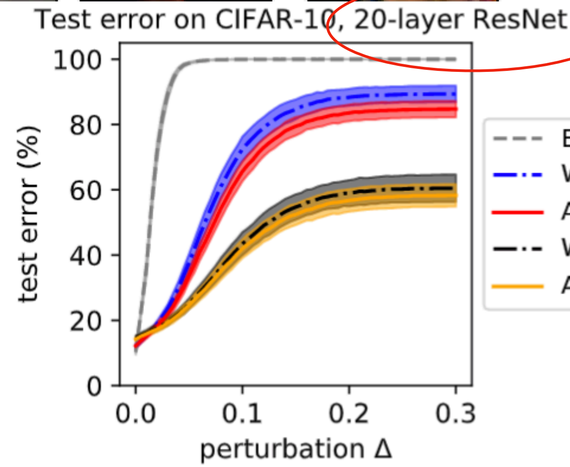
[NSZ '21]



Adversarially Robust Kernel Smoothing

[ZKNS '21]



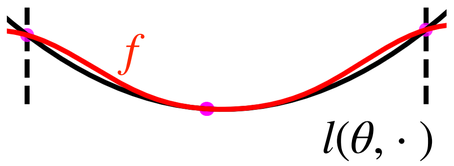


Code: jj-zhu.github.io/research

Conclusions

- A generalized duality theorem for solving DRO with general ambiguity sets and IPM-balls, with weak assumptions on the loss
  - Kernel DRO: Maximizing w.r.t. a distribution → finding a smooth surrogate function. For example,

(D) $\min_{\theta, f \in \mathscr{H}} \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}} \quad \text{s.t.}\ l(\theta, \cdot) \le f$



- Takeaway
  - Large (universal) RKHSs as dual spaces for DRO
  - **Flatten the curve, smooth is robust**

Future directions

- Generalizaiton and risk bound of Kernel DRO
  - Lam-Zeng 2021, Zhu in prep
- Kernel semi-infinite program
  - Zhu et al. 2021, in prep, Marteau-Ferey-Bach-Rudi 2020, (related: Lasserre moment-SOS)
- Applications to high-dim. data, deep models, adversarial learning, fairness, control…

Code: jj-zhu.github.io/research

Related references

1. Zhu, J.-J., Kouridi, C., Nemmour, Y. & Schölkopf, B. Adversarially Robust Kernel Smoothing. arXiv:2102.08474 [cs, math, stat] (2021). https://arxiv.org/abs/2102.08474

2. Zhu, J.-J., Jitkrittum, W., Diehl, M. & Schölkopf, B. Kernel Distributionally Robust Optimization. Proceedings of the 24thInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego,California, USA. PMLR: Volume 130. (2020). https://arxiv.org/abs/2006.06981

3. Nemmour, Y., Schölkopf, B. & Zhu, J.-J. Approximate Distributionally Robust Nonlinear Optimization with Application to Model Predictive Control: A Functional Approach. in Learning for Dynamics and Control 1255–1269 (PMLR, 2021). http://proceedings.mlr.press/v144/nemmour21a.html

4. Marteau-Ferey, U., Bach, F. & Rudi, A. Non-parametric Models for Non-negative Functions. arXiv:2007.03926 [cs, math, stat] (2020).

5. Lasserre, J.-B. The Moment-SOS hierarchy and the Christoffel-Darboux kernel. arXiv:2011.08566 [math, stat] (2020).

6. Lam, H. & Zeng, Y. Complexity-Free Generalization via Distributionally Robust Optimization. arXiv:2106.11180 [cs, math, stat] (2021).

Thank you!

Code: jj-zhu.github.io/research

Jia-Jie Zhu

jj-zhu.github.io

Weierstrass Institute, Berlin &
Max Planck Institute, Tübingen
Germany

Co-authors

Wittawat Jitkrittum (Google Research)
Moritz Diehl (Uni. Freiburg)
Bernhard Schölkopf (MPI Tübingen)

SIAM OP21

Ph.D. positions available in Berlin, Germany
Robust machine learning and data-driven optimization & control