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/ji-zhu/kdro
Robust optimization

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

Minimize \( \min_{\theta} \mathbb{E}_{\xi \sim \hat{P}} [l(\theta, \xi)] \)

- 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)

Minimize \( \min_{\theta} \sup_{\xi \in \mathcal{U}} [l(\theta, \xi)] \)

- 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)

\[
\begin{align*}
(\text{ERM}) \min_{\theta} & \quad \mathbb{E}_{\xi \sim \hat{P}} l(\theta, \xi) \\
(\text{RO}) \min_{\theta} & \quad \sup_{\xi \in \mathcal{U}} l(\theta, \xi)
\end{align*}
\]

\[\min_{\theta} \sup_{P \in \mathcal{H}} \mathbb{E}_l l(\theta, \xi)\] (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 \( \mathcal{H} \) (ambiguity set), e.g.,
  • \( \mathcal{H} \) can be a metric-ball centered at \( \hat{P} \), e.g., using the popular 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: choosing better ambiguity regions
  • This talk provides a functional analysis and optimization perspective instead of statistics
Learning with kernels

• A kernel is a symmetric function
  \[ k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \text{ e.g., Gaussian kernel} \]
  \[ k(x, x') = \exp \left( -\frac{\|x - x'\|^2}{2\sigma^2} \right). \]

• 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)
  \[ \text{IPM}(\mathcal{F}; P, Q) := \sup_{f \in \mathcal{F}} \int f (P - Q). \]

  Special cases:
  \[ \mathcal{F} = \{ f : \|f\|_{\mathcal{F}} \leq 1 \} \to \text{Maximum Mean Discrepancy (MMD)} \]
  \[ \mathcal{F} = \{ f : \|f\|_{\text{lip}} \leq 1 \} \to \text{Wasserstein (type-1)} \]

\[ \mu := \int \phi \, dP \text{ is the (kernel) mean embedding of } P \text{ in } \mathcal{H}. \]

\[ \mu \text{ can be viewed as a generalized moment vector} \]
\[ \text{e.g., let } \phi(x) = [x, x^2]^T \text{ (related: Lasserre moment-SOS)} \]
Theorem (Kernel DRO duality, Zhu et al. ’20). DRO (P) is equivalent to solving

\[(D) \quad \min_{\theta, f} \delta^*_\mathcal{C}(f) \quad \text{subject to} \quad l(\theta, \cdot) \leq f,\]

\[\delta^*_\mathcal{C}(f) \text{ is the support function, e.g., } \mathbb{E}_\tilde{P} f + \epsilon \|f\|_\mathcal{C}.\]

(Note: no need to estimate \(\|l(\theta, \cdot)\|_\mathcal{C}\))

Test loss

| Perturbation \(\Delta\) (test distribution) |
|----------------|----------------|----------------|----------------|
| 0              | 1              | 2              | 3              | 4              |
| 0.0            | 1.0            | 2.0            | 3.0            | 4.0            |

Robustifying with kernels

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

Geometric intuition

Smoothness of \(f \mapsto \text{Distributional robustness} \quad (\mapsto \text{Size of } \mathcal{H})\)

Intuition: flatten the curve, smooth is robust

\[\mathcal{H} : \text{special case SDP/SOS; generalization to IPM, e.g., W-1}\]
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, \(l \notin \mathcal{H}\). Kernel DRO handles this by finding a majorant \(f \in \mathcal{H}\), with no need to estimate \(\|l(\theta, \cdot)\|_{\mathcal{H}}\).

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

[NSZ '21]

Adversarially Robust Kernel Smoothing

[ZKNS '21]
Conclusions

• A generalized dual program for solving DRO with general ambiguity sets and IPM-balls, with weak assumptions on the loss function (no need to estimate \( \| l(\theta, \cdot) \|_\mathcal{H} \))

• Kernel DRO: Maximizing w.r.t. a distribution \( \rightarrow \) finding a smooth surrogate function. For example,

\[
\begin{align*}
(D) \quad \min_{\theta, f \in \mathcal{H}} \mathbb{E}_{\hat{p}} \left[ f + \epsilon \| f \|_{\mathcal{H}} \right] \quad \text{s.t.} \quad l(\theta, \cdot) \leq f
\end{align*}
\]

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

Future directions

• Generalization and statistical bounds of Kernel DRO
  • Lam-Zeng 2021, Zhu in prep

• Kernel SIP, chance constraints…
  • Marteau-Ferey-Bach-Rudi 2020, Zhu et al. 2021, in prep (related: Lasserre moment-SOS)

• Applications to high-dim. data, deep models, adversarial learning, fairness, control…
  • Kernel DRO offers unique benefits but is not nearly as popular as the Wasserstein distance.

Code: jj-zhu.github.io/research
Related references


Thank you!

Co-authors

Jia-Jie Zhu
jj-zhu.github.io
Weierstrass Institute, Berlin &
Max Planck Institute, Tübingen
Germany

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

Code: jj-zhu.github.io/research