Cornell Bowers C·IS
College of Computing
and Information Science

CornellEngineering
Operations Research
and Information Engineering

Doubly Robust Distributional Robust Off-Policy Evaluation and Learning

Kaiwen Wang
Cornell University

Joint work with Nathan Kallus, Xiaojie Mao and Zhengyuan Zhou

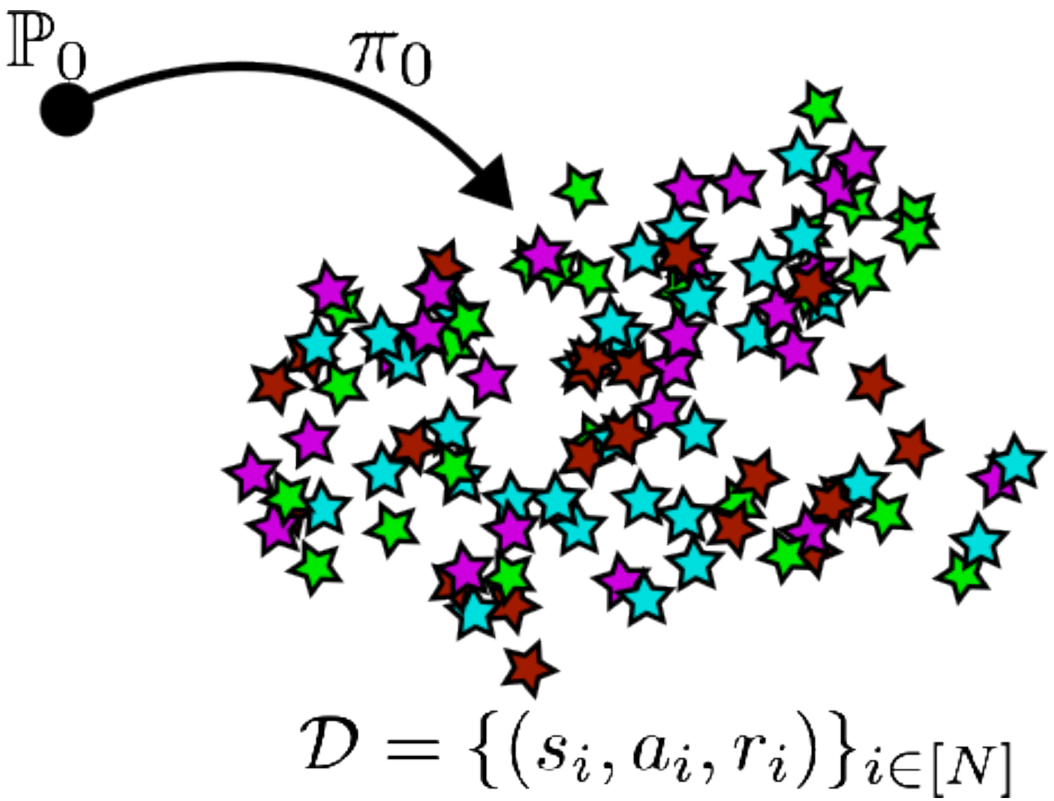
arena



NYU STERN

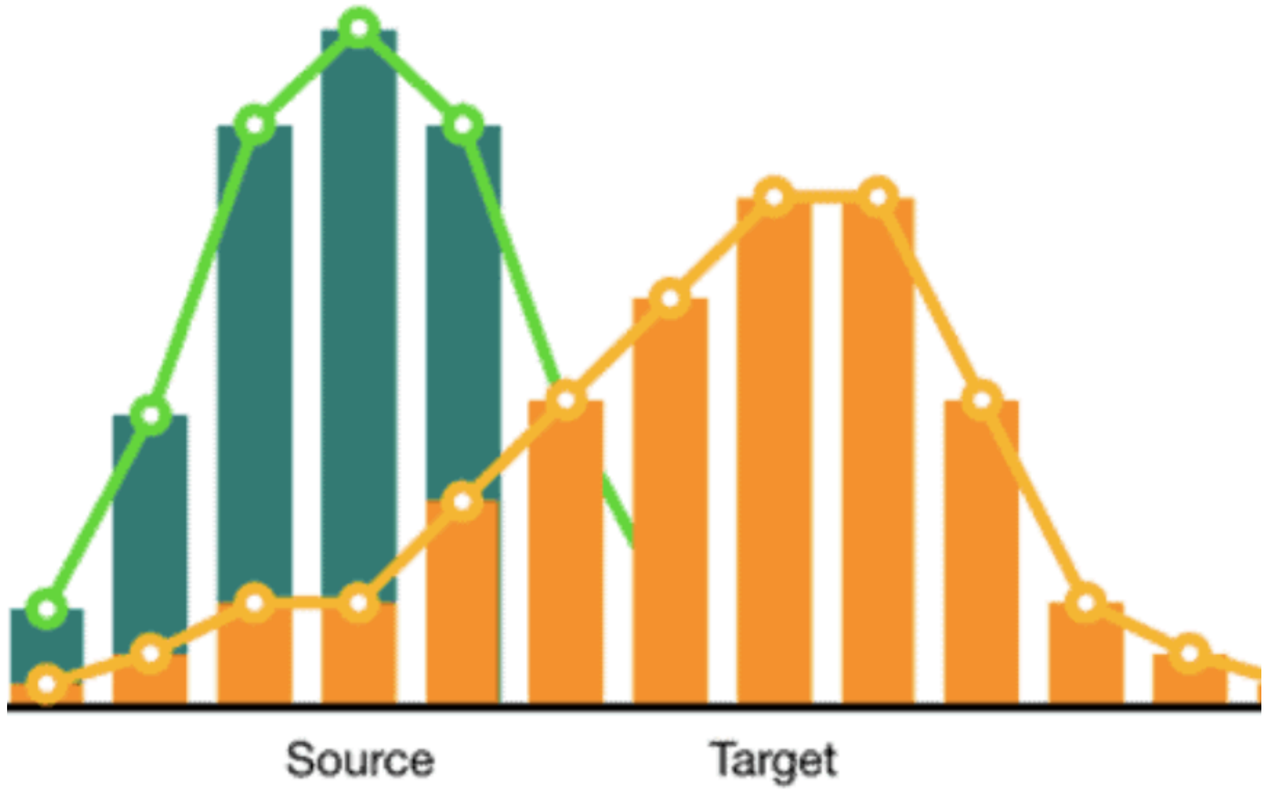
Off-Policy Evaluation (OPE)

Evaluate the performance of π using data collected from behavior policy π_0.



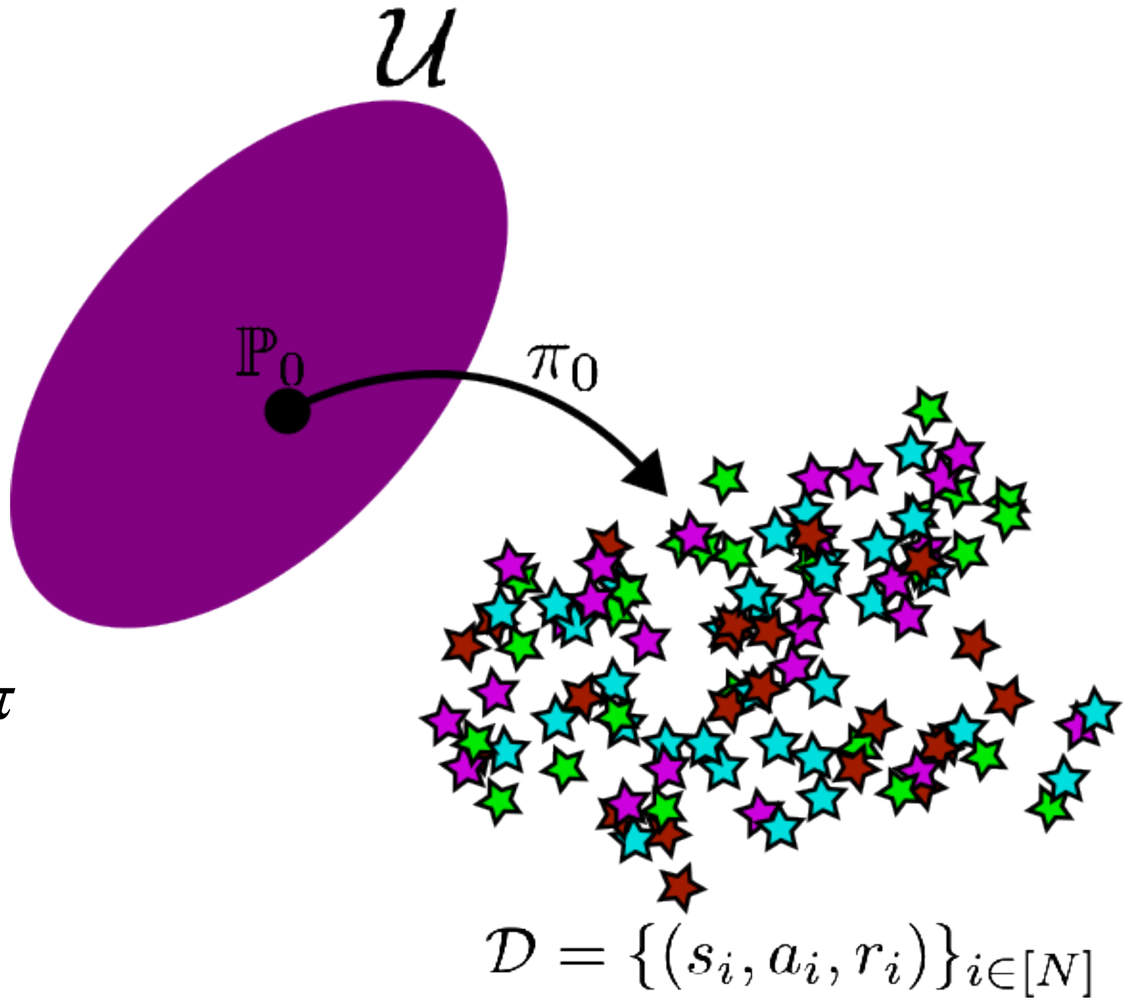
Distribution Shift

- Covariate shift, in $\mathbb{P}_0(S)$.
- Concept shift, in $\mathbb{P}_0(R(a) \mid S)$.



Distributionally Robust OPE (DROPE)

Evaluate the **worst-case** performance of π **in an uncertainty set** $\mathcal{U}$ using the same data as in OPE.

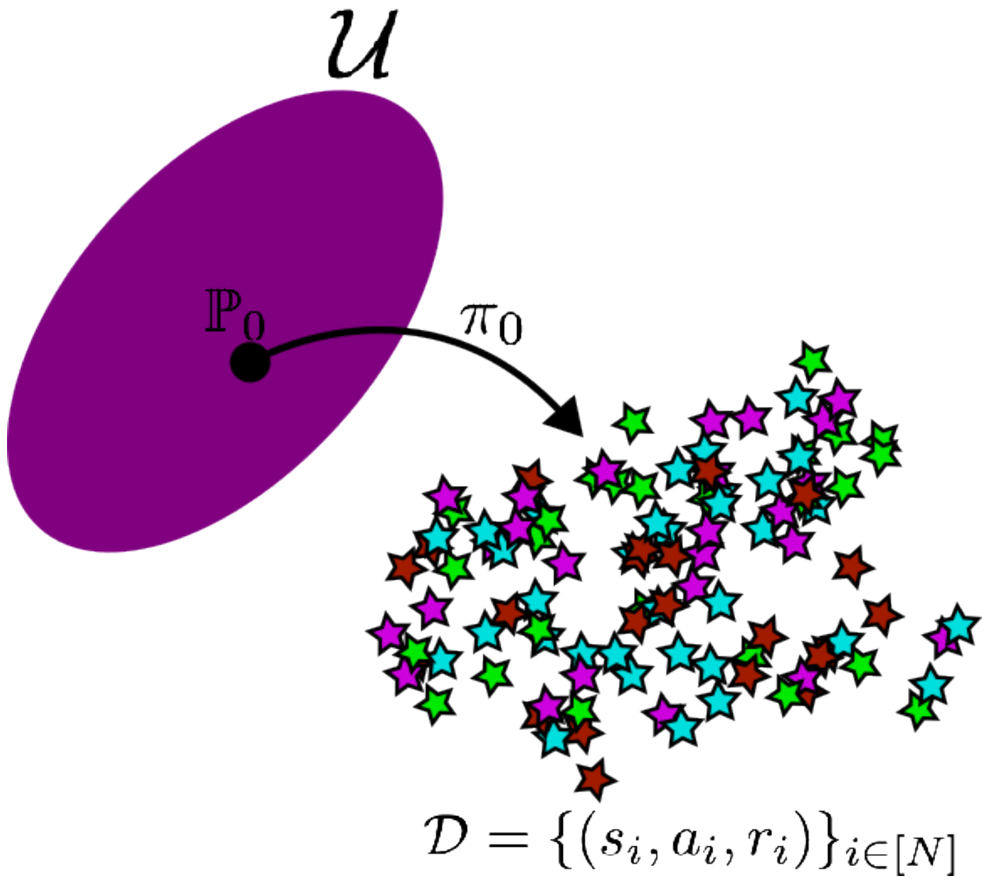


Distributionally Robust Value

For a given radius δ,

$$\mathcal{U}(\delta) = \{\mathbb{P}_1 \ll \mathbb{P}_0 : D_{KL}(\mathbb{P}_1 \| \mathbb{P}_0) \leq \delta\},$$

$$\mathcal{V}_\delta(\pi) = \inf_{\mathbb{P}_1 \in \mathcal{U}(\delta)} \mathbb{E}_{\mathbb{P}_1}[R(\pi(S))].$$



Distributionally Robust Value

For a given radius δ,

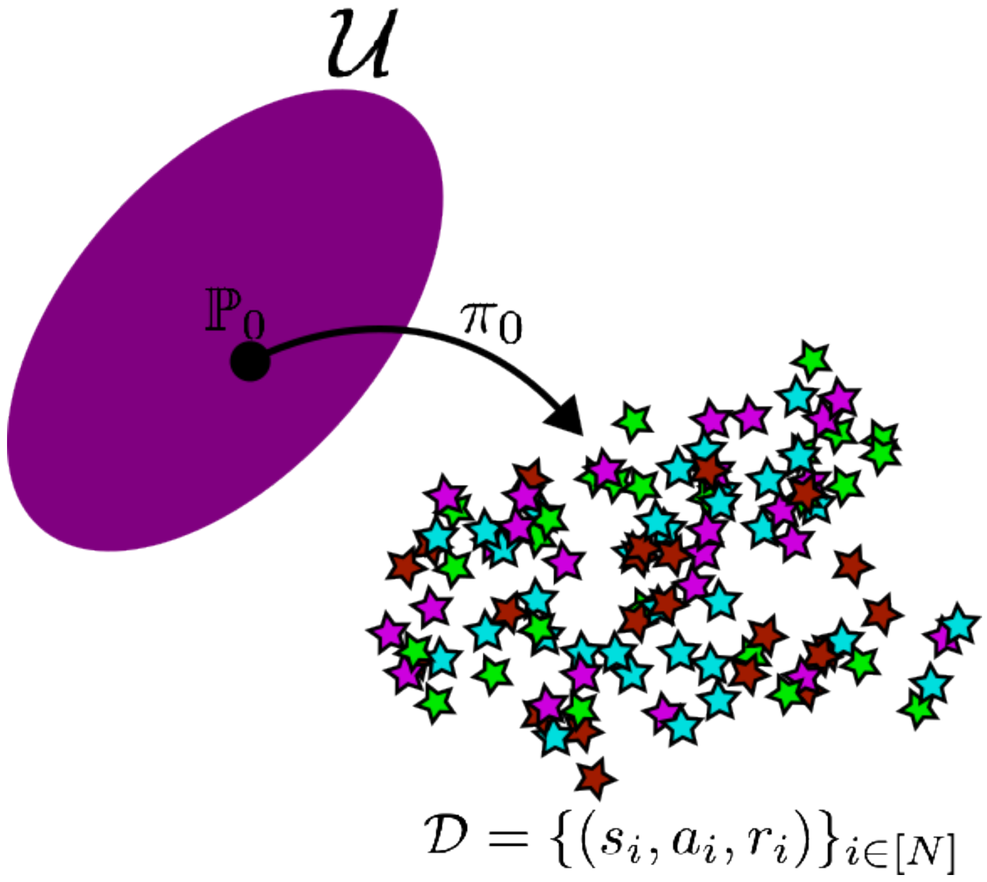
$$\mathcal{U}(\delta) = \{\mathbb{P}_1 \ll \mathbb{P}_0 : D_{KL}(\mathbb{P}_1 \| \mathbb{P}_0) \leq \delta\},$$

$$\mathcal{V}_\delta(\pi) = \inf_{\mathbb{P}_1 \in \mathcal{U}(\delta)} \mathbb{E}_{\mathbb{P}_1}[R(\pi(S))].$$

By strong duality,

$$\mathcal{V}_\delta(\pi) = \max_{\alpha > 0} -\alpha \log W(\pi, \alpha) - \alpha\delta,$$

$$W(\pi, \alpha) = \mathbb{E}_{\mathbb{P}_0}[\exp(-R(\pi(S))/\alpha)].$$



Distributionally Robust Value

For a given radius δ,

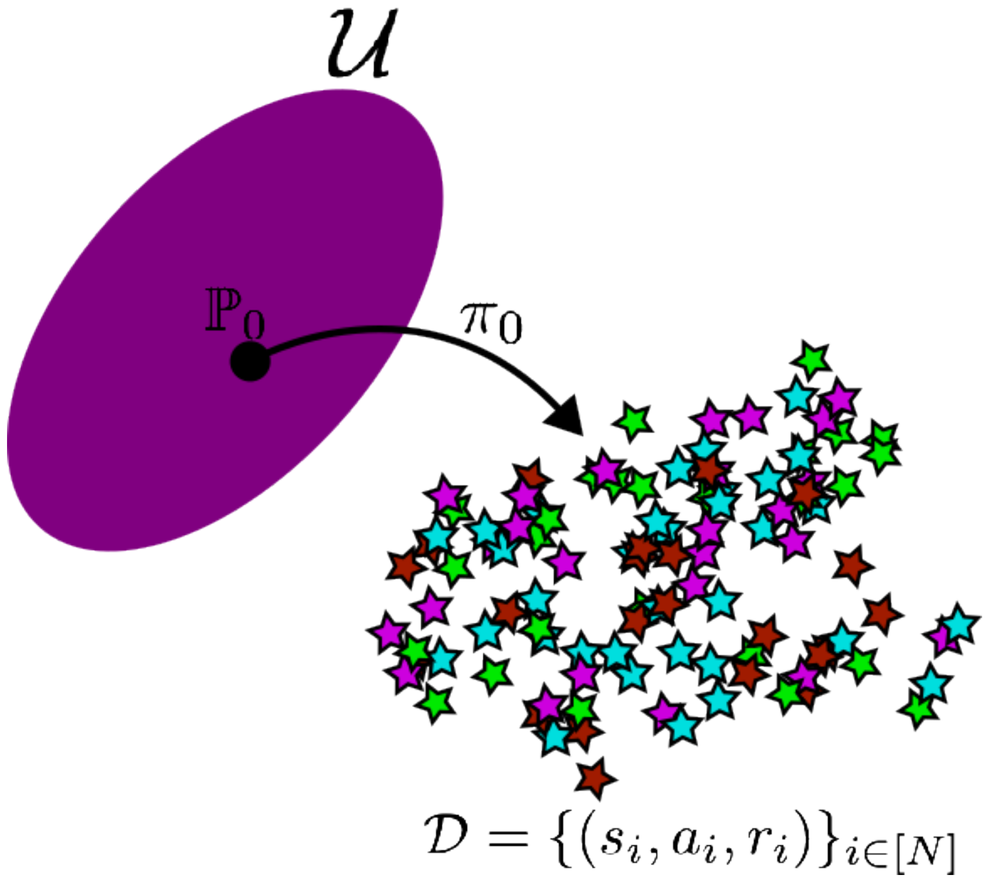
$$\mathcal{U}(\delta) = \{\mathbb{P}_1 \ll \mathbb{P}_0 : D_{KL}(\mathbb{P}_1 \| \mathbb{P}_0) \leq \delta\},$$

$$\mathcal{V}_\delta(\pi) = \inf_{\mathbb{P}_1 \in \mathcal{U}(\delta)} \mathbb{E}_{\mathbb{P}_1}[R(\pi(S))].$$

By strong duality,

$$\mathcal{V}_\delta(\pi) = \max_{\alpha > 0} -\alpha \log W(\pi, \alpha) - \alpha\delta,$$

$$W(\pi, \alpha) = \mathbb{E}_{\mathbb{P}_0}[\exp(-R(\pi(S))/\alpha)].$$



[1] Si, N., Zhang, F., Zhou, Z., and Blanchet, J. Distributionally robust policy evaluation and learning in offline contextual bandits. In *ICML*, pp. 8884–8894. PMLR, 2020

Prior Work

Assuming that propensities $\pi_0(a_i \mid s_i)$ are known,

$$\widehat{\mathscr{V}}_{\delta}^{SNIPS}(\pi) = \max_{\alpha>0} -\alpha \log \widehat{W}^{SNIPS}(\pi, \alpha) - \alpha\delta,$$

$$\widehat{W}^{SNIPS}(\pi, \alpha) = \sum_{i=1}^{N} \frac{w_i}{\sum_j w_j} \exp(-r_i/\alpha),\ w_i := \frac{\pi(a_i \mid s_i)}{\pi_0(a_i \mid s_i)}.$$

Prior Work

Assuming that propensities $\pi_0(a_i \mid s_i)$ are known,

$$\widehat{\mathscr{V}}_{\delta}^{SNIPS}(\pi) = \max_{\alpha>0} -\alpha \log \widehat{W}^{SNIPS}(\pi, \alpha) - \alpha\delta,$$

$$\widehat{W}^{SNIPS}(\pi, \alpha) = \sum_{i=1}^{N} \frac{w_i}{\sum_j w_j} \exp(-r_i/\alpha),\ w_i := \frac{\pi(a_i \mid s_i)}{\pi_0(a_i \mid s_i)}.$$

! Sub-optimal variance.

!! If π_0 is unknown, the propensity estimation error is first-order!

Double Robustness resolves both shortcomings for offline RL, but...

How can we apply Doubly Robust methods to the more complex, distributionally robust offline RL?

Moment Equation Formulation

Reframe as finding the root of the objective's gradient.

$$-\log W_0(\pi, \alpha) - \frac{W_1(\pi, \alpha)}{\alpha W_0(\pi, \alpha)} - \delta = 0,$$

$$\text{where } W_j(\pi, \alpha) := \mathbb{E}_{\mathbb{P}_0}\left[R(\pi(S))^j \exp(-R(\pi(S))/\alpha)\right].$$

Then apply Localized Double Machine Learning.

*Cross-fitted appropriately, as detailed in the paper.

Statistical Optimality

Theorem (Informal)

Denote $\theta^\star = (\alpha^\star, W_0^\star, W_1^\star, \mathscr{V}_\delta)$. Then,

$$\sqrt{N}(\hat{\theta}^{LDROPE} - \theta^\star) \rightsquigarrow \mathcal{N}(0, \Sigma),$$

where Σ is the optimal covariance, i.e. achieves semi-parametric efficiency, provided that...

Statistical Optimality

Theorem (Informal)

Denote $\theta^\star = (\alpha^\star, W_0^\star, W_1^\star, \mathscr{V}_\delta)$. Then,

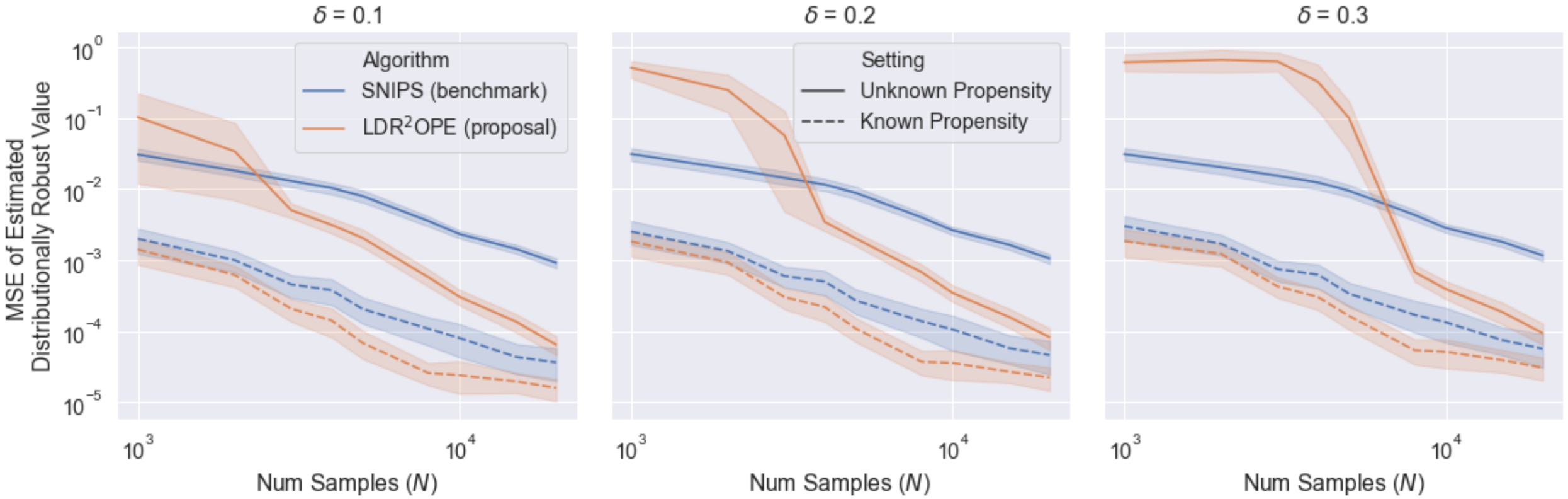
$$\sqrt{N}(\hat{\theta}^{LDROPE} - \theta^\star) \rightsquigarrow \mathcal{N}(0, \Sigma),$$

where Σ is the optimal covariance, i.e. achieves semi-parametric efficiency, provided that...

$$\text{Rate}(\pi_0) \cdot \Big(\text{Rate}(f_0, f_1) + \text{Rate}(\pi_0)\Big) = o_p(N^{-1/2}).$$

Note: $\hat{\pi}_0$ and $\hat{f}_0, \hat{f}_1$ can have slow non-parametric $o_p(N^{-1/4})$ rates!

Evaluation Comparison

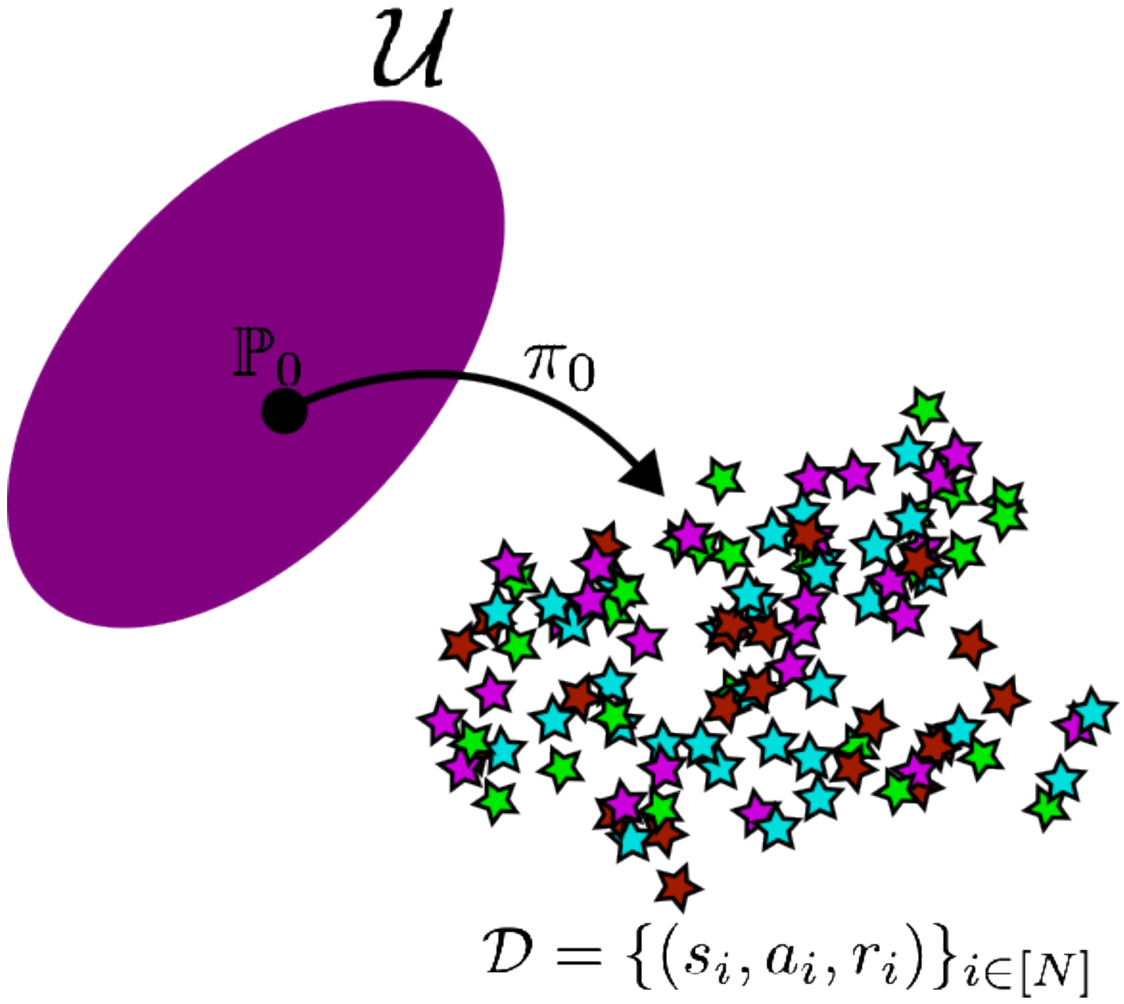


- Up to 10x reduction in MSE for large enough N!
- May have worse MSE when N is small.

Distributionally Robust Learning (DROPL)

Learn a robust policy $\widehat{\pi}$ that maximizes the worst-case policy value, i.e. try to solve

$$\arg\max_{\pi \in \Pi} \mathscr{V}_{\delta}(\pi).$$



Doubly Robust Method for Learning

Train a continuum of regression functions $\{\hat{f}_0(\,\cdot\,;\alpha), \alpha > 0\}$.
Then, target the following policy:

$$\widehat{\pi}^{DR} \in \arg\max_{\pi\in\Pi} \max_{\alpha>0} -\alpha \log \widehat{W}^{DR}(\pi,\alpha) - \alpha\delta,$$

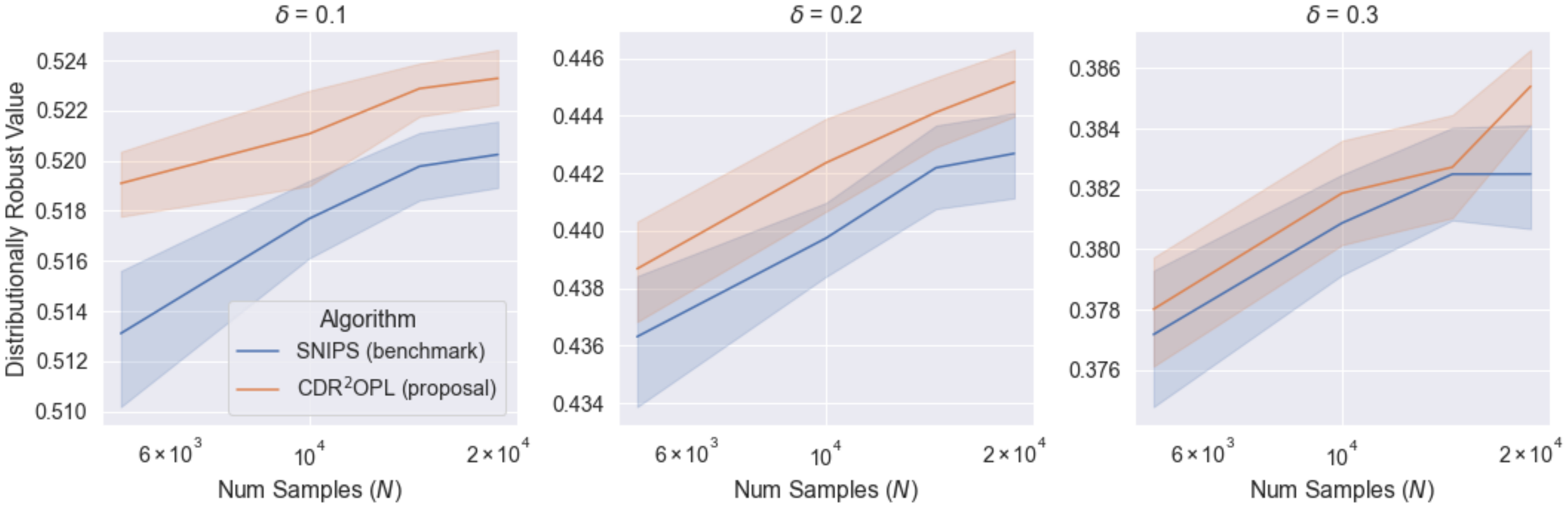
$$\widehat{W}^{DR}(\pi,\alpha) = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi(a_i \mid s_i)}{\widehat{\pi}_0(a_i \mid s_i)} \left(\exp(-r_i/\alpha) - \hat{f}_0(s_i, a_i; \alpha)\right) + \sum_{a\in\mathcal{A}} \pi(a \mid s_i)\hat{f}_0(s_i, a; \alpha)$$

We showed that $\widehat{\pi}^{DR}$ has regret at most $\mathcal{O}\left(\frac{\text{comp}(\Pi)}{\sqrt{N}}\right)$ assuming

$\text{Rate}(\pi_0) \cdot \text{Rate}(\{f_0(\,\cdot\,;\alpha)\}) = o_p(N^{-1/2})$.

*Cross-fitted appropriately. Practical algorithm detailed in the paper.

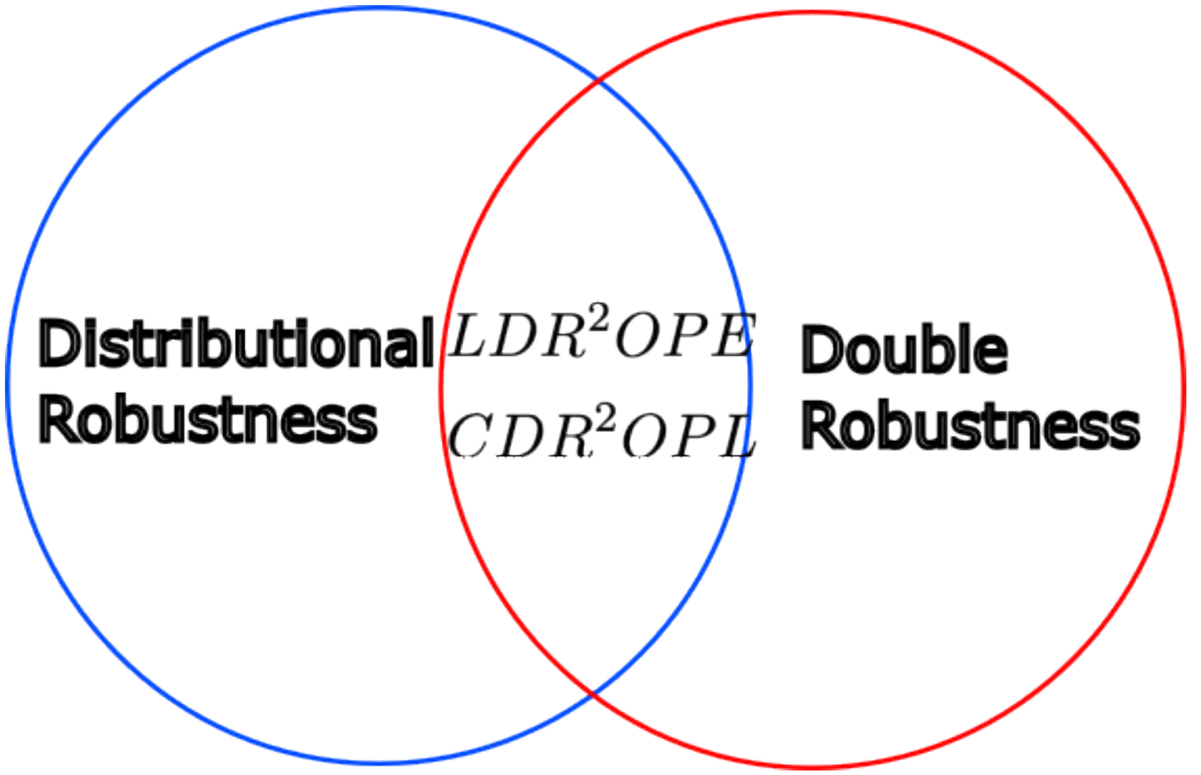
Learning Comparison



- Improves over SNIPS, but only marginally (1%).
- Computational cost is $\mathcal{O}(N^2)$ vs. $\mathcal{O}(N)$.

Key Takeaways

- For DROPE, use LDR²OPE (provided enough data). It is statistically optimal with barely any computational overhead.
- For DROPL, try SNIPS and CDR²OPL. Then, select the better policy with LDR²OPE.



Thank you!

Github Repository: https://github.com/CausalML/doubly-robust-dropel