CS292F StatRL Lecture 12 OPE in Bandits and Reinforcement Learning

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Recap: Lecture 11

- Exploration in Linear MDPs
  - Finished the regret analysis
  - Uniform convergence with a covering number argument

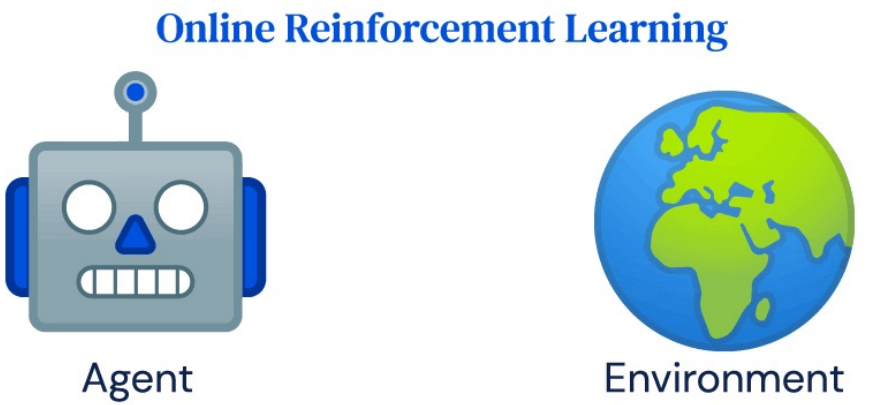
- Short introduction to offline RL

What I did not cover about exploration?

- Theoretical driven techniques
  - We covered: Optimism / UCB / Exploration bonuses
  - In the homework, you will see something else
  - We did not see: Thompson sampling

- Exploration techniques used in Deep RL
  - Curiosity (and other ways to add bonuses)
  - Adding noise (to model parameters / values / actions)
  - Random Network Distillation (user model-uncertainty)

Recap: Online RL vs Offline RL

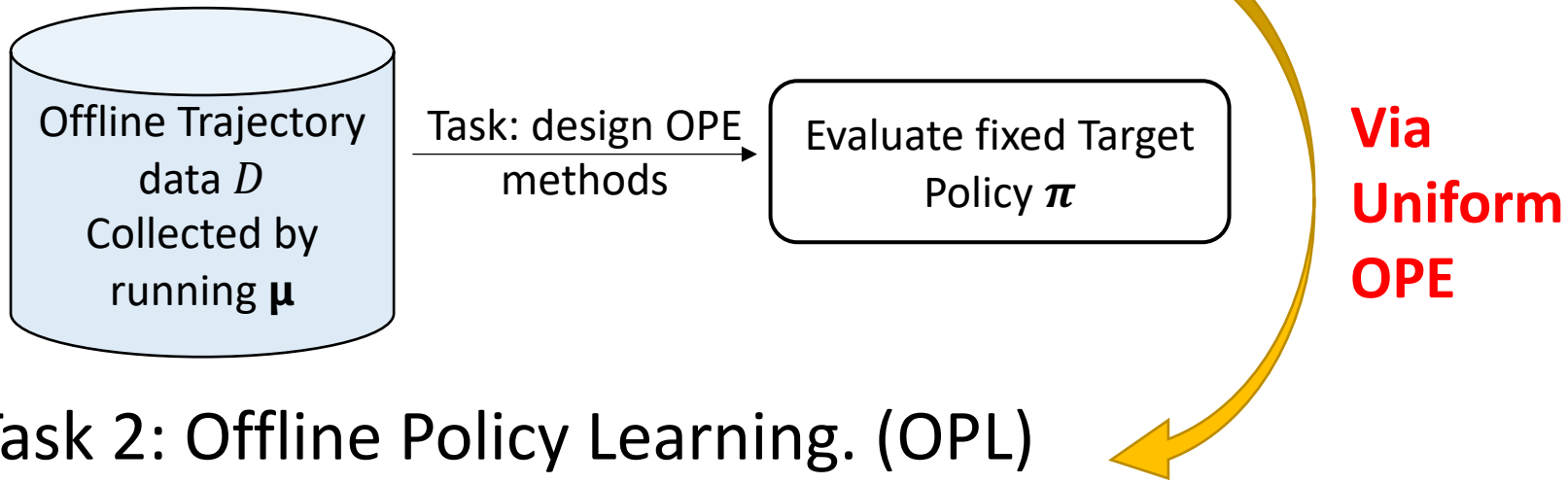


Exploration is often **expensive**, **unsafe**, **unethical** or **illegal** in practice, e.g., in self-driving cars, or in medical applications.

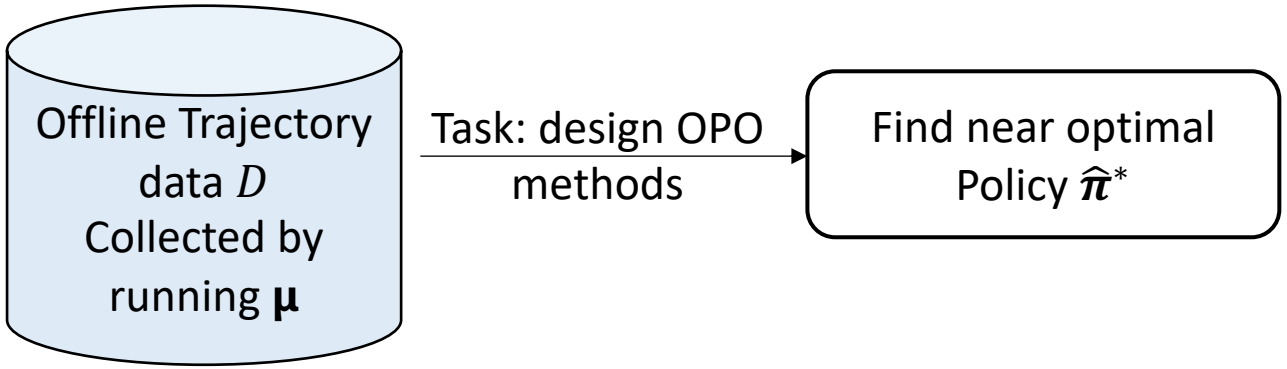
Can we learn a policy from already **logged interaction data**?

Recap: Offline Reinforcement Learning, aka. Batch RL

- Task 1: Offline Policy Evaluation. (OPE)



- Task 2: Offline Policy Learning. (OPL)



Recap: Contextual bandits model

- Contexts:
  - $x_1, \ldots, x_n \sim \lambda$ drawn iid, possibly infinite domain
- Actions:
  - $a_i \sim \mu(a|x_i)$ Taken by a randomized "Logging" policy
- Reward:
  - $r_i \sim D(r|x_i, a_i)$ Revealed only for the action taken
- Value:
  - $v^{\mu} = \mathbb{E}_{x\sim\lambda}\mathbb{E}_{a\sim\mu(\cdot|x)}\mathbb{E}_D[r|x,a]$
- We collect data $(x_i, a_i, r_i)_{i=1}^n$ by the above processes.

Recap: Two standard approaches

- Direct method / regression estimator

$$\hat{v}^{\pi}_{\mathrm{DM}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\hat{r}(x_i, a)\pi(a|x_i)$$

- Importance sampling / Inverse Propensity Score /

$$\hat{v}^{\pi}_{\mathrm{IPS}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\mu(a_i|x_i)}r_i$$

This lecture

- Continue with the estimators for OPEs in Bandits
  - Ideas to improve upon IS / DM
  - Some statistical analysis / comparisons

- OPE estimators for RL

Analyzing the performance of importance sampling estimator

- Mean:

- Variance:

Consider an even simpler setting: Multi-armed bandits with K-arms

$$\hat{v}^{\pi}_{\mathrm{DM}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\hat{r}(x_i,a)\pi(a|x_i)$$

$$\hat{v}^{\pi}_{\mathrm{IPS}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\mu(a_i|x_i)}r_i$$

Importance Sampling and Direct Method are surprisingly similar in some cases

- Consider just MAB:
  - Importance sampling as a regression estimator
  - Regression estimator as an importance sampling
- Which one is better?

Comparing the MSE of DM and IS

- Mean Square Error and Bias-Variance Decomposition
- Analyzing DM with plug-in estimator

Weighted importance sampling

- Self-normalization

Experiment 1: Facebook data

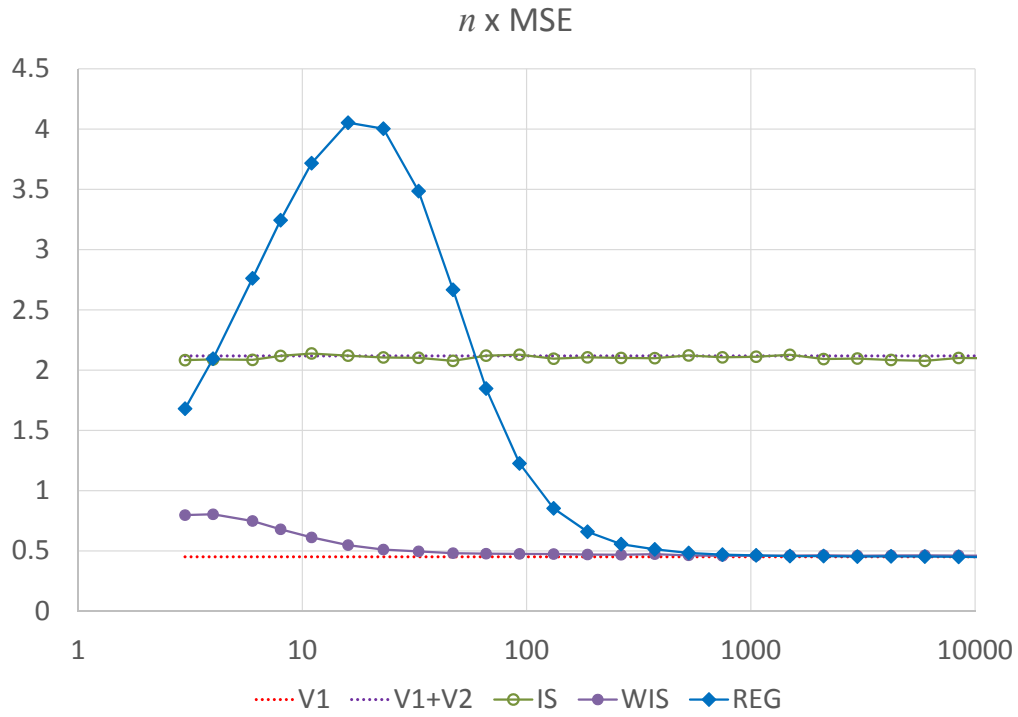


Figure 3: nMSE for query "facebook" ($K = 2178$). The asymptotic rates V_1 and $V_1 + V_2$ are provided for reference.

(Li, Munos, Szepesvari, 2015) Toward Minimax Offline Policy Evaluation

Experiment 2: Gmail data

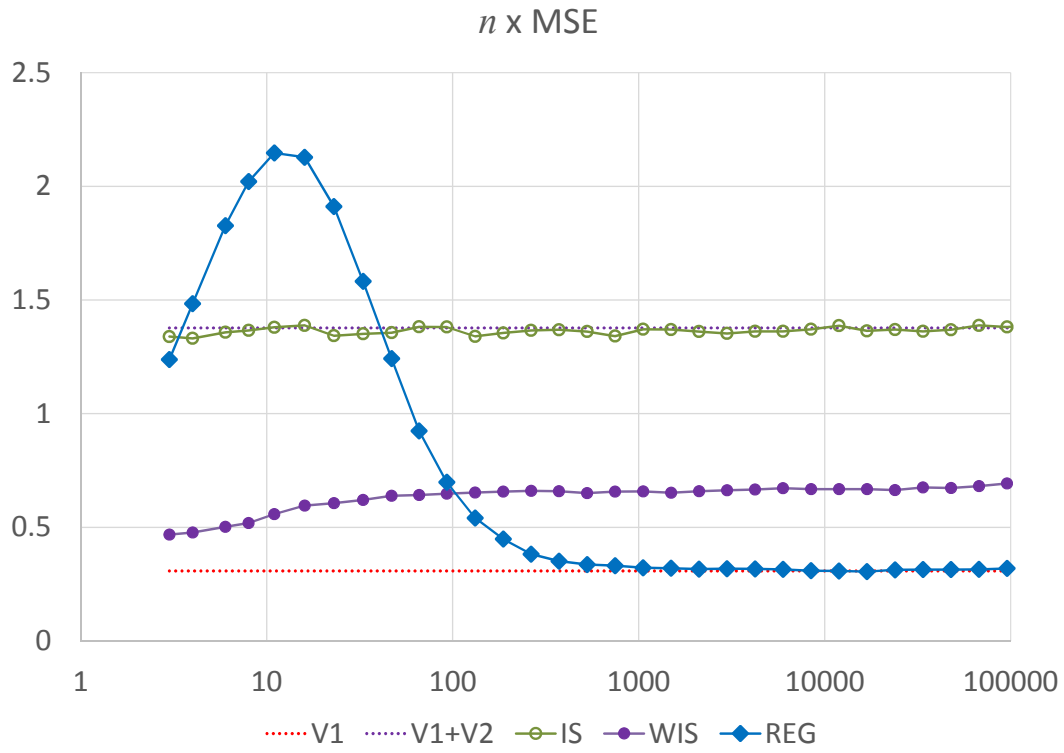


Figure 4: nMSE for query "gmail" ($K = 648$). The asymptotic rates V_1 and $V_1 + V_2$ are provided for reference.

(Li, Munos, Szepesvari, 2015) Toward Minimax Offline Policy Evaluation

Doubly robust estimator for OPE

- We are using the regression estimator as a baseline.

(Dudík et al., 2014)

Theory of doubly robust estimator

- Double robustness in model-misspecification

- Variance reduction (sometimes)

(Robins and Rotnitzky, 1995; Bang and Robins,2005)

Lower bounding the minimax risk

- Our main theorem: assume λ is a probability density, then under mild moment conditions

$$\inf_{\hat{v}} \sup_{D(r|a,x)\in\mathcal{R}(\sigma^2,R_{\max})} \mathbb{E}(\hat{v}-v^\pi)^2$$

$$= \Omega\left[\frac{1}{n}\left(\underbrace{\mathbb{E}_\mu[\rho^2\sigma^2]}_{\text{Randomness in reward}} + \underbrace{\mathbb{E}_\mu[\rho^2 R_{\max}^2]}_{\text{Randomness due to context distribution}}\right)\right]$$

W., Agarwal, Dudik (2017) Optimal and adaptive off-policy evaluation in contextual bandits. ICML'17

Classical optimality theory (Hahn, 1998)

- n* Var[any LAN estimator] is greater than:

$$\mathbb{E}_{x\sim\mathcal{D}}\left\{\mathbb{E}_{\mu}[\rho^2\mathrm{Var}(r|x,a)|x]\right\}+\mathrm{Var}_{x\sim\mathcal{D}}\left\{\mathbb{E}_{\mu}[\rho r|x]\right\}.$$

Take ⬇ supremum

$$\mathbb{E}_{\mu}[\rho^2\sigma^2]+\mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{\mu}[\rho R_{\max}|x]^2\right].$$

- The minimax lower bound is bigger!

$$\mathbb{E}_{\mu}\left[\rho^2\sigma^2\right]+\mathbb{E}_{\mu}\left[\rho^2 R_{\max}^2\right]$$

How could that be? There are estimators that achieve asymptotic efficiency.

- e.g., Robins, Hahn, Hirano, Imbens, and many others in the semiparametric efficiency industry!

Assumption:	**Realizable assumption:** E[r|x,a] is differentiable in x for each a.	No assumption on E[r|x,a] except boundedness.
Consequences	Hirano et. al. is optimal. Imbens et. al. is optimal. IPS is suboptimal!	IPS is optimal (up to a universal constant)
Caveat	Poor finite sample performance. Exponential dependence in d.	Does NOT adapt to easier problems.

SWITCH estimator

- Recall that IPS is bad because: $\hat{v}^{\pi}_{\mathrm{IPS}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\mu(a_i|x_i)} r_i$

- SWITCH estimator:

For each $i = 1, \ldots, n$, for each action $a \in: \mathcal{A}$

if $\pi(a|x_i)/\mu(a|x_i) \leq \tau$:

Use IPS (or DR).

else:

Use regression estimator.

Error bounds for SWITCH

$$\mathrm{MSE}(\hat{v}_{\mathrm{SWITCH}}) \leq$$

$$\frac{2}{n}\mathbb{E}_{\mu}[(\sigma^2 + R_{\max}^2)\rho^2 \mathbf{1}(\rho \leq \tau)] \quad (1)$$

$$+ \frac{2}{n}\mathbb{E}_{\pi}[R_{\max}^2 \mathbf{1}(\rho > \tau)] \quad (2)$$

$$+ \mathbb{E}_{\pi}\left[\epsilon \mathbf{1}(\rho > \tau)\right]^2 \quad (3)$$

1) Variance from IPS (reduced by truncation)

2) Variance due to sampling x. Required even with perfect oracle

1) Bias from the oracle.

Automatic parameter tuning

- Conservative approximate MSE minimizing.

$$\widehat{\tau} = \operatorname*{argmin}_{\tau} \widehat{\mathrm{Var}}_\tau + \widehat{\mathrm{Bias}}_\tau^2.$$

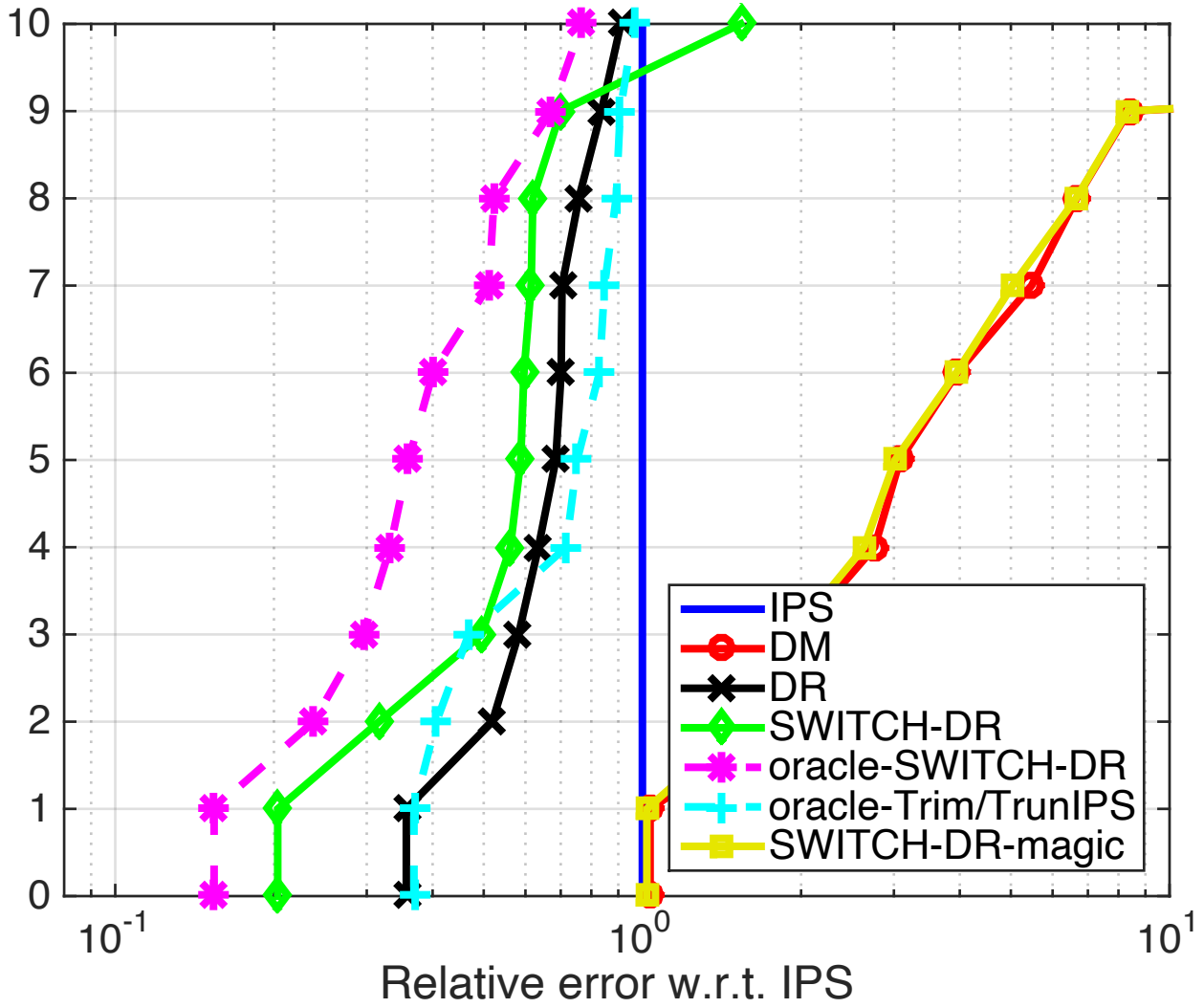
- Details:

$$Y_i(\tau) := r_i \rho_i \mathbf{1}(\rho_i \le \tau) + \sum_{a \in \mathcal{A}} \hat{r}(x_i, a) \pi(a|x_i) \mathbf{1}(\rho(x_i, a) > \tau) \quad \text{and} \quad \bar{Y}(\tau) = \frac{1}{n} \sum_{i=1}^{n} Y_i(\tau),$$

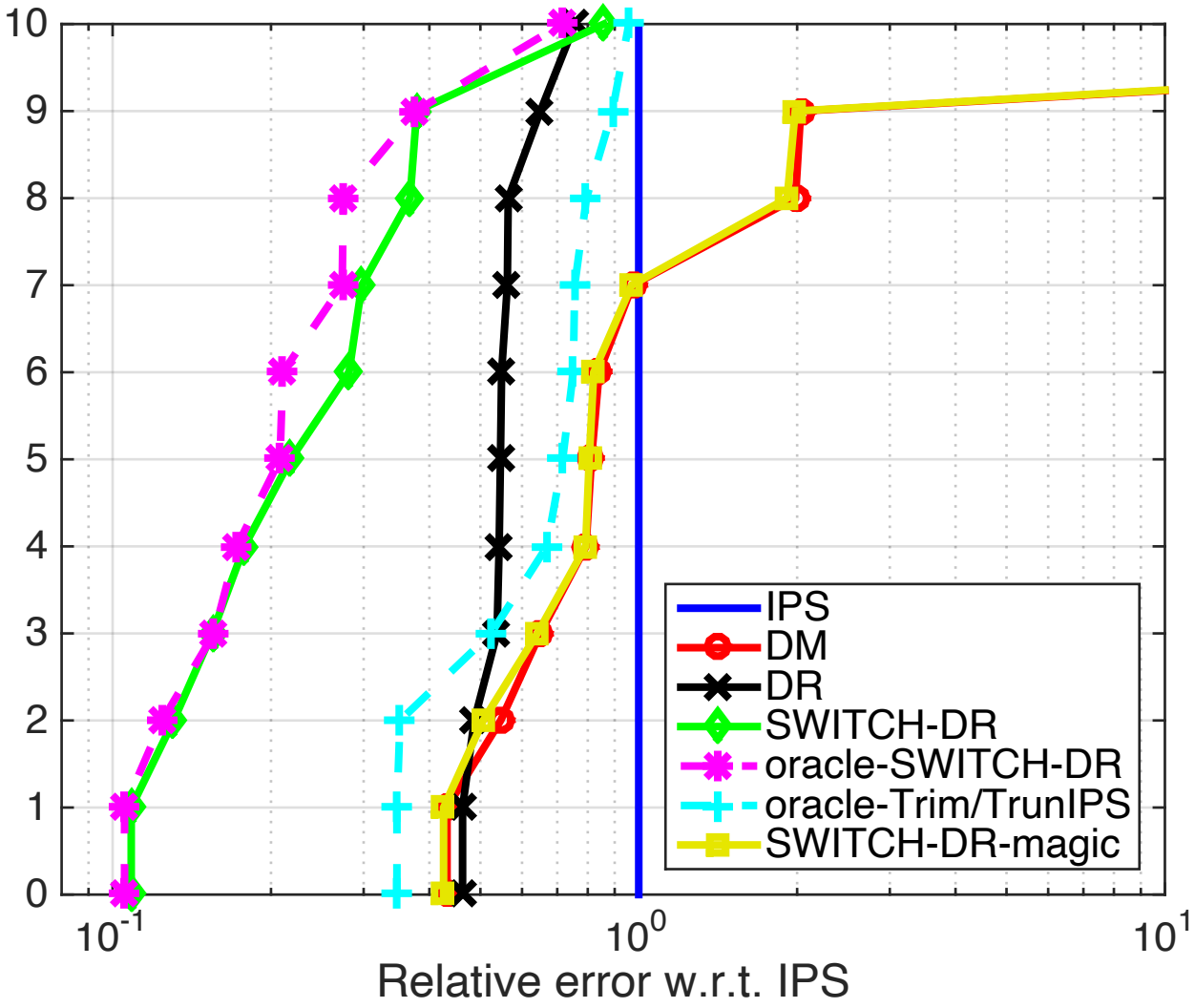
$$\mathrm{Var}(\hat{v}_{\mathrm{SWITCH}-\tau}) = \frac{1}{n} \mathrm{Var}(\hat{v}_{\mathrm{SWITCH}-\tau}(x_1)) \approx \frac{1}{n^2} \sum_{i=1}^{n} (Y_i(\tau) - \bar{Y}(\tau))^2 =: \widehat{\mathrm{Var}}_\tau,$$

$$\begin{aligned} \mathrm{Bias}^2(\hat{v}_{\mathrm{SWITCH}}) &\le \mathbb{E}_\mu[\rho \epsilon^2 | \rho > \tau] \pi(\rho > \tau)^2 \le \mathbb{E}_\mu[\rho R_{\max}^2 | \rho > \tau] \pi(\rho > \tau)^2 \\ &\approx \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\pi \left(R_{\max}^2 | \rho > \tau, x_i \right) \right] \left[\frac{1}{n} \sum_{i=1}^{n} \pi(\rho > \tau | x_i) \right]^2 =: \widehat{\mathrm{Bias}}_\tau^2. \end{aligned}$$

CDF of relative MSE over 10 UCI multiclass classification data sets.



With additional label noise



Checkpoint: OPE for Contextual Bandits

- Estimators: DM, IS, WIS, DR, SWITCH
- Bias-Variance Tradeoff
- Optimality theory
  - Depends on whether you have access to a good regression estimator

OPE in Reinforcement Learning

- Importance sampling on the entire trajectory

- (Per-Step) Importance Sampling

- Exercise:
  - Infinite horizon discounted version?
  - Weighted Importance Sampling Extension?

Doubly Robust OPE in Reinforcement Learning

- An alternative form for the Per-Step IS

$$V^0_{\text{step-IS}} := 0, \text{ and for } t = 1, \ldots, H,$$
$$V^{H+1-t}_{\text{step-IS}} := \rho_t \left(r_t + \gamma V^{H-t}_{\text{step-IS}} \right).$$

- Given a value function approximator

$$V^0_{\text{DR}} := 0, \text{ and for } t = 1, \ldots, H,$$
$$V^{H+1-t}_{\text{DR}} := \widehat{V}(s_t) + \rho_t \left(r_t + \gamma V^{H-t}_{\text{DR}} - \widehat{Q}(s_t, a_t) \right).$$

Jiang, N., & Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In ICML 2016.

Mean and Variance of Doubly Robust OPE in RL

- Doubly Robust OPE in RL is unbiased

- Variance

Theorem 1. *V_{DR} is an unbiased estimator of $v^{\pi_1,H}$, whose variance is given recursively as follows: $\forall t = 1, \ldots, H$,*

$$\mathbb{V}_t\left[V_{DR}^{H+1-t}\right] = \mathbb{V}_t\left[V(s_t)\right] + \mathbb{E}_t\left[\mathbb{V}_t\left[\rho_t \Delta(s_t, a_t) \mid s_t\right]\right] + \mathbb{E}_t\left[\rho_t^2 \, \mathbb{V}_{t+1}\left[r_t\right]\right] + \mathbb{E}_t\left[\gamma^2 \rho_t^2 \, \mathbb{V}_{t+1}\left[V_{DR}^{H-t}\right]\right], \quad (11)$$

where $\Delta(s_t, a_t) := \widehat{Q}(s_t, a_t) - Q(s_t, a_t)$ and $\mathbb{V}_{H+1}\left[V_{DR}^{0} \mid s_H, a_H\right] = 0$.

Main challenge of OPE in RL: The curse of Horizon

$$\widehat{v}_{IS}^{\pi} = \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\left[\prod_{t=1}^{h}\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})}\right] r_h^{(i)}.$$

The curse of horizon. (Liu et al, 2018 NeurIPS)

- The variance is exponential in H!

Next Lecture

- Marginalized Importance Sampling
- Fitted Q-Iterations