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

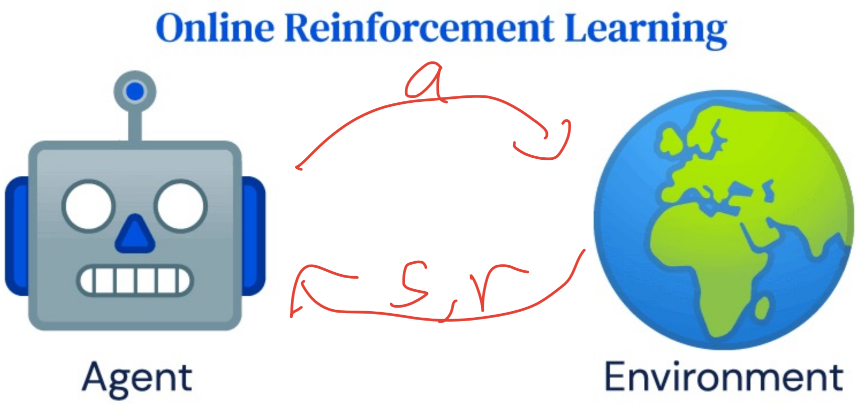
- Short introduction to offline RL

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

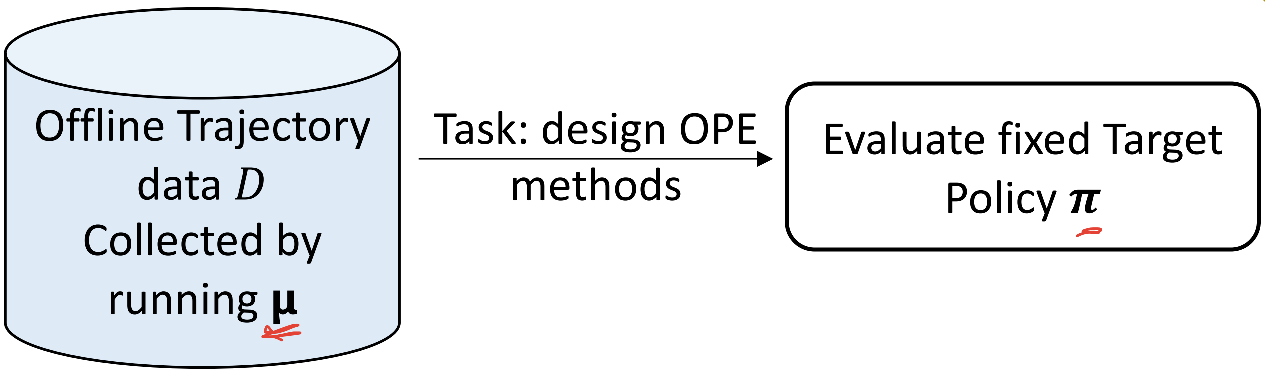


Exploration is often **expensive**, **unsafe**, **unethical** or **illegal** in practice, e.g., in self-driving cars, or in medical applications.

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

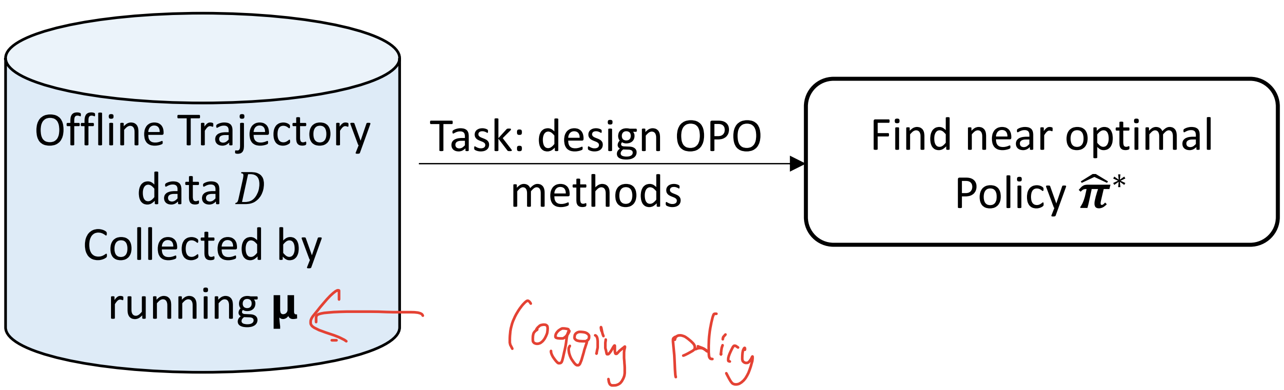
Recap: Offline Reinforcement Learning, aka. Batch RL

- Task 1: Offline Policy Evaluation. (OPE)



Via Uniform 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

$E[R_i | S_i = s, A_i = a] = r(s,a)$ R_i $S_i = s$ $A_i = a$

- 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.

(S_i, A_i, R_i)

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

$$\hat{V}^{\pi}_{IS} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)} R_i$$

- Mean:

$$\mathbb{E}[\hat{V}^{\pi}_{IS}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\frac{\pi(A_i|S_i)}{\mu(A_i|S_i)} R_i\right] = \mathbb{E}_{S_1\sim\lambda}\left[\mathbb{E}_{A_1\sim\mu(\cdot|S_1)}\left[\underbrace{\mathbb{E}[R_1|S_1,A_1]}_{r(S_1,A_1)} \frac{\pi(A_1|S_1)}{\mu(A_1|S_1)} \Big| S_1\right]\right]$$

$$= \mathbb{E}_{S_1\sim\lambda}\left[\sum_{a\in A} \mu(a|S_1)\cdot \frac{\pi(a|S_1)}{\mu(a|S_1)} r(S_1,a)\right]$$

$$= V^{\pi}$$

- Variance:

$$\mathrm{Var}\left[\frac{1}{n}\sum_i \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)} R_i\right] = \frac{1}{n}\mathrm{Var}\left[\frac{\pi(A_1|S_1)}{\mu(A_1|S_1)} R_1\right]$$

$$= \frac{1}{n}\left(\mathbb{E}_S\left[\mathrm{Var}_{\mu}\left[\frac{\pi(A_1|S_1)}{\mu(A_1|S_1)} R_1 \Big| S_1\right]\right] + \mathrm{Var}_S\left[\mathbb{E}_{\mu}\left[\frac{\pi(A_1|S_1)}{\mu(A_1|S_1)} R_1 \Big| S_1\right]\right]\right)$$

$$= \frac{1}{n}\left(\mathbb{E}_{S_1}\left[\mathbb{E}_{A_1\sim\mu(\cdot|S_1)}\left[\rho_1^2 \mathrm{Var}[R_1|S_1,A_1]\right] \Big| S_1\right] + \mathrm{Var}_{A_1\sim\mu(\cdot|S_1)}\left(\mathbb{E}[\rho_1 R_1|S_1,A_1]\right)\Big|S_1 + \mathrm{Var}_{S_1}\left[\mathbb{E}_{\pi}[r(S_1,A_1)|S_1]\right]\right)$$

$$= \frac{1}{n}\left(\underbrace{\mathbb{E}_{\mu}[\rho^2 \sigma^2(S,A)]}_{\text{Reward variance}} + \underbrace{\mathbb{E}\left[\mathrm{Var}_{\mu}[\rho_1 r(S,A_1)]\big|S\right]}_{\text{Logging policy's variance}} + \underbrace{\mathrm{Var}\left[\mathbb{E}_{\pi}(r(S,A_1)|S_1)\right]}_{\text{Variance of the context}}\right)$$

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

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

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

Importance Sampling and Direct Method are surprisingly similar in some cases

- Consider just MAB:

$$\hat{r}_{IS}(a)_i = \begin{cases} 0 & \text{if } A_i \neq a \\ \frac{R_i}{\mu(A_i)} & \text{if } A_i = a \end{cases}$$

$$\hat{r}_{IS}(a) = \frac{1}{n}\sum_i R_i \frac{\mathbb{1}(A_i=a)}{\mu(A_i)} = \frac{1}{n}\sum_i \hat{r}_{IS}(a)_i$$

  - Importance sampling as a regression estimator

$$\hat{V}^{\pi}_{IS} = \frac{1}{n}\sum_i \frac{\pi(A_i)}{\mu(A_i)} R_i = \sum_{a\in A} \pi(a) \frac{1}{n}\sum_i \frac{R_i \mathbb{1}(A_i=a)}{\mu(A_i)} = \sum_{a\in A} \pi(a)\, \hat{r}_{IS}(a)$$

  - Regression estimator as an importance sampling

$$\hat{V}^{\pi}_{DM} = \sum_a \hat{r}(a)\cdot\pi(a) = \sum_a \frac{1}{N_a}\sum_i R_i \mathbb{1}(A_i=a)\cdot\pi(a) = \sum_a \frac{1}{n}\frac{1}{\hat{\mu}(a)}\sum_{i=1}^{n} R_i \mathbb{1}(A_i=a)\cdot\pi(a)$$

$$\hat{r}(a) = \begin{cases} \frac{1}{N_a}\sum_i R_i \mathbb{1}(A_i=a) & \text{if } N_a > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$N_a = n\cdot\frac{\sum \mathbb{1}(A_i=a)}{n} = n\cdot\hat{\mu}(a)$$

- Which one is better?

$$\mathbb{E}[R \mid A=a] = r(a) \qquad \hat{r}(a) \xrightarrow{N_a\to\infty} r(a)$$

Comparing the MSE of DM and IS

In Simple MAB: IS

$$MSE(\hat V_{IS}) = Var[\hat V_{IS}] = \frac{1}{n}\left[\underbrace{E_\mu\left(\left(\frac{\pi(A)}{\mu(A)}\right)^2 \sigma^2(A)\right)}_{V_1} + \underbrace{Var_\mu\left[\frac{\pi(A)}{\mu(A)} \cdot r(A)\right]}_{V_2}\right]$$

- Mean Square Error and Bias-Variance Decomposition

$$MSE(\hat V) = E[(\hat V - V^\pi)^2] = E[(\hat V - E[\hat V] + E[\hat V] - V^\pi)^2]$$

$$= \underbrace{E[(\hat V - E\hat V)^2]}_{Var(\hat V)} + \underbrace{E(E\hat V - V^\pi)^2}_{(bias(\hat V))^2} + 2E[(\hat V - E\hat V)](E\hat V - V^\pi) \quad (=0)$$

- Analyzing DM with plug-in estimator

$$\hat V_{DM} = \sum_a \pi(a)\hat r(a). \quad E[\hat V_{DM}] = \sum_a \pi(a) E[\hat r(a)] = \sum_a \pi(a) E\left[\frac{1}{N_a}\sum R_i \mathbb{1}(A_i=a)\mathbb{1}(N_a>0)\right]$$

$$= \sum_a \pi(a) P(N_a>0)\, \underbrace{E\left[\frac{1}{N_a}\sum R_i \mathbb{1}(A_i=a) \mid N_a>0\right]}_{r(a)}$$

$$bias(\hat V_{DM}) = E\hat V_{DM} - V^\pi = \sum_a \pi(a) r(a)[P(N_a>0) - 1] \xrightarrow{n\to\infty} 0 \qquad (1-\mu(a))^n$$

$$Var\left[\sum_a \pi(a)\hat r(a)\right] = E\left[Var\left[\sum_a \pi(a)\hat r(a) \mid N_a, a\in A\right]\right] + Var\left[E\left[\sum_a \pi(a)\hat r(a) \mid N_a, a\in A\right]\right]$$

$$= E\left[\sum_a \pi(a)^2 Var(\hat r(a) \mid N_a)\right] + Var\left[\sum_a \pi(a) r(a)\mathbb{1}(N_a>0)\right]$$

$$= E\left[\sum_a \frac{\pi(a)^2}{N_a}\sigma^2(a)\mathbb{1}(N_a>0)\right]$$

$$\le E\left[\sum_a \frac{\pi(a)^2}{N_a+1}\sigma^2(a)\right] = \sum_a \frac{\pi(a)^2}{(n+1)\mu(a)}\sigma^2(a)\left(1-(1-\mu(a))^{n+1}\right) + R_{max}^2 \cdot P(\exists a, N_a=0) \xrightarrow{n\to\infty} 0$$

$$\approx \frac{1}{n+1}\sum_a \mu(a)\frac{\pi(a)^2}{\mu(a)^2}\sigma^2(a) = \frac{1}{n+1} E\rho^2\sigma^2(a) < V_1$$

DM

Weighted importance sampling

- Self-normalization

$$\hat{V}_{WIS} = \frac{1}{\sum_{i=1}^{n} \rho_i} \sum_{i=1}^{n} \rho_i R_i \qquad \rho_i = \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}$$

$$E\left[\sum_{i=1}^{n} \rho_i\right] = E_{\mu} \sum_{i=1}^{n} \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)} = \sum_{i=1}^{n} \sum_{a} \mu(A_i|S_i) \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}$$

$$= n$$

$$\hat{V}_{WIS} \xrightarrow{n \to \infty} V^{\pi}$$

Experiment 1: Facebook data

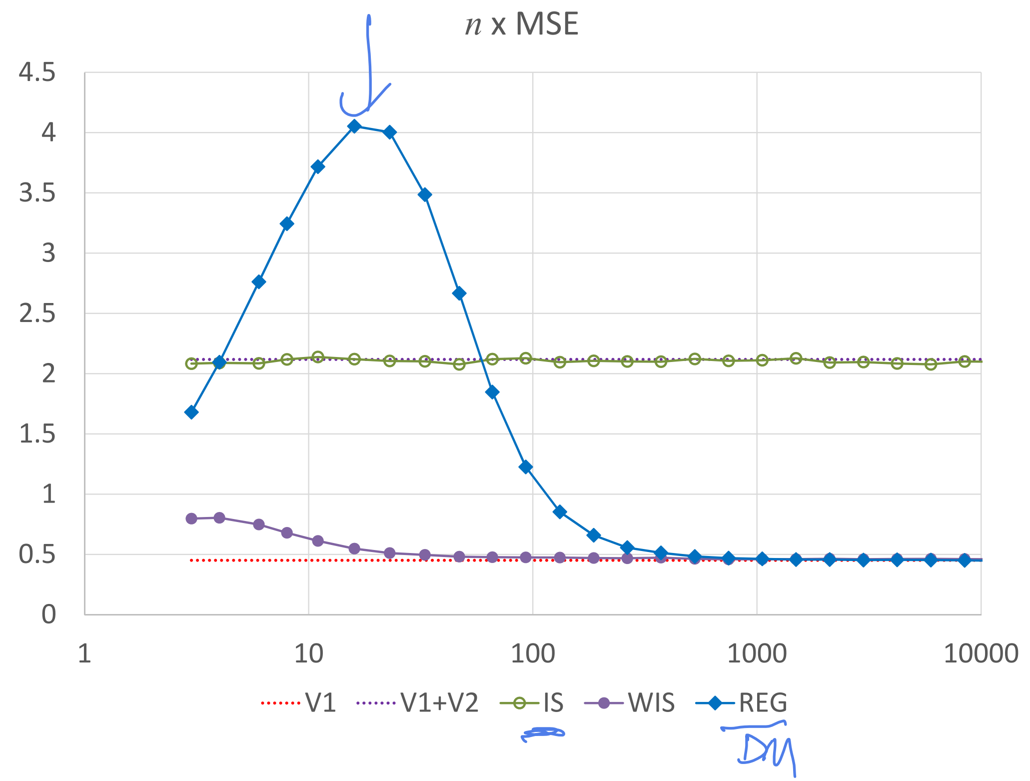


Figure 3: nMSE for query "facebook" ($K = 2178$). The asymptotic rates V_1 and $V_1 + V_2$ are provided for reference.

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

Experiment 2: Gmail data

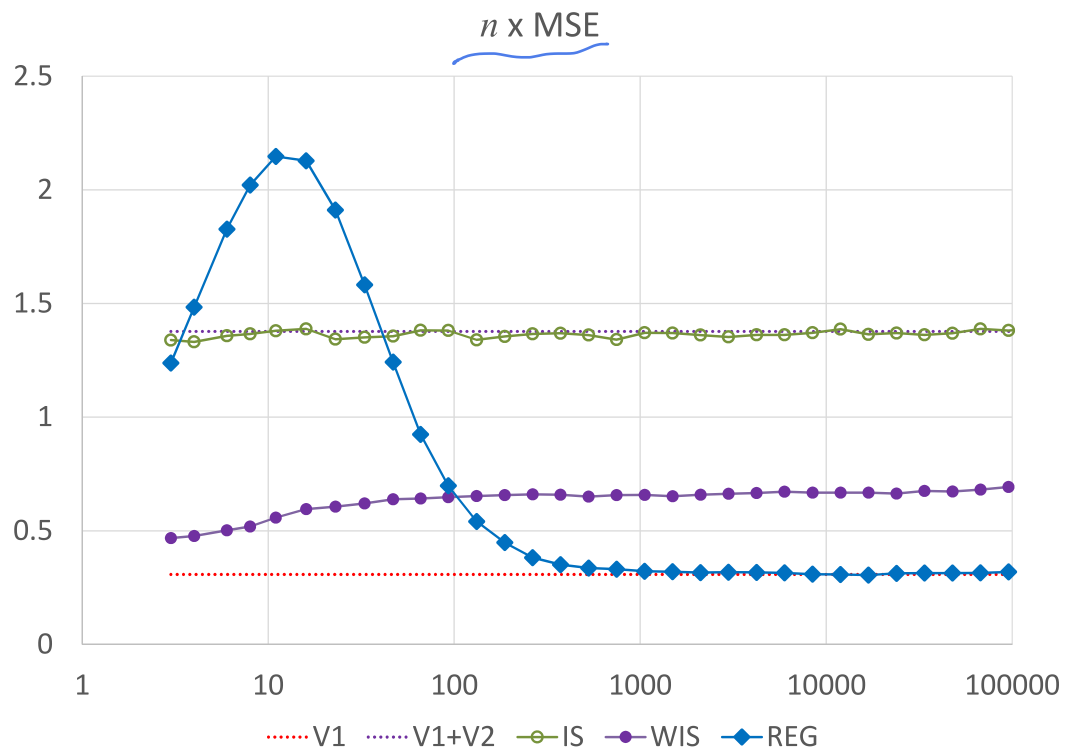


Figure 4: nMSE for query "gmail" ($K = 648$). The asymptotic rates V_1 and $V_1 + V_2$ are provided for reference.

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

Doubly robust estimator for OPE

(x_i, a_i, r_i)

(Contextual bandits) (S_i, A_i, R_i)

- We are using the regression estimator as a baseline.

$$\hat{V}^{\pi}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(E_{\pi}[\hat{r}(S_i, A')|S_i] + \rho_i\left(R_i - \hat{r}(S_i, A_i)\right)\right), \quad \rho_i = \frac{\pi_i}{\mu_i}$$

baseline

$\hat{r}(s,a)$ $V(s,a)$

fixed / given

estimated from a different data

(Dudík et al., 2014)

Theory of doubly robust estimator

unknown $\mu_i = \mu(A_i|S_i) \leftarrow \hat{\mu}(S_i, A_i)$

under $\hat{r}(s,a) \to r(s,a)$, $\hat{\mu}(s,a) \to \mu(s,a)$

- Double robustness in model-misspecification

$$\hat{V}_{DR} = \frac{1}{n}\sum_i \hat{r}^{\pi}(S_i) + \frac{\pi(A_i|S_i)}{\hat{\mu}(S_i,A_i)} \cdot \left(R_i - \hat{r}(S_i,A_i)\right)$$

if either $\hat{\mu}$ or $\hat{r}$ is consistent, then $\hat{V}_{DR} \to V^{\pi}$

if either $E[\hat{\mu}] = \mu$ or $E[\hat{r}] = r$, then $E[\hat{V}_{DR}] = V^{\pi}$

$\hat{r}(S_i,A_i) \to r(S_i,A_i) = E[R_i|S_i,A_i]$

- Variance reduction (sometimes)

$\hat{V}_{DR}$ is asymptotically efficient

if

$$(\hat{r} - r) \cdot \left(\frac{1}{\hat{\mu}} - \frac{1}{\mu}\right) = o\left(\frac{1}{\sqrt{n}}\right)$$

$\hat{r} = o(1)$ if $\hat{\mu} = \mu$

(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

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

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

Randomness in reward

Randomness due to context distribution

IS is optimal in the worst case

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

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

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

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

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

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:

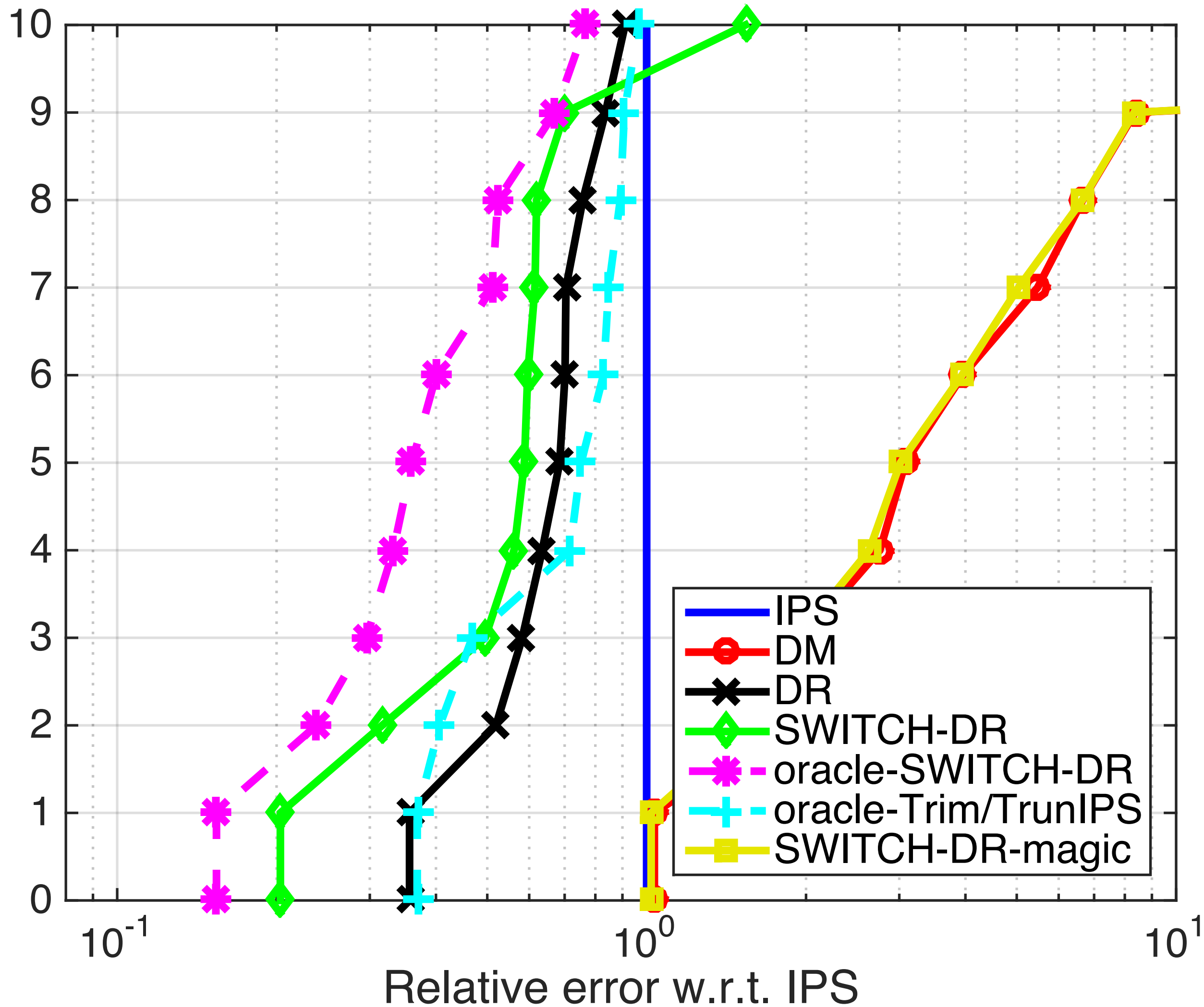
$$Y_i(\tau) := r_i \rho_i \mathbf{1}(\rho_i \le \tau) + \sum_{a \in \mathcal{A}} \hat{r}(x_i, a)\pi(a|x_i)\mathbf{1}(\rho(x_i, a) > \tau) \quad \text{and} \quad \bar{Y}(\tau) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\tau),$$

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

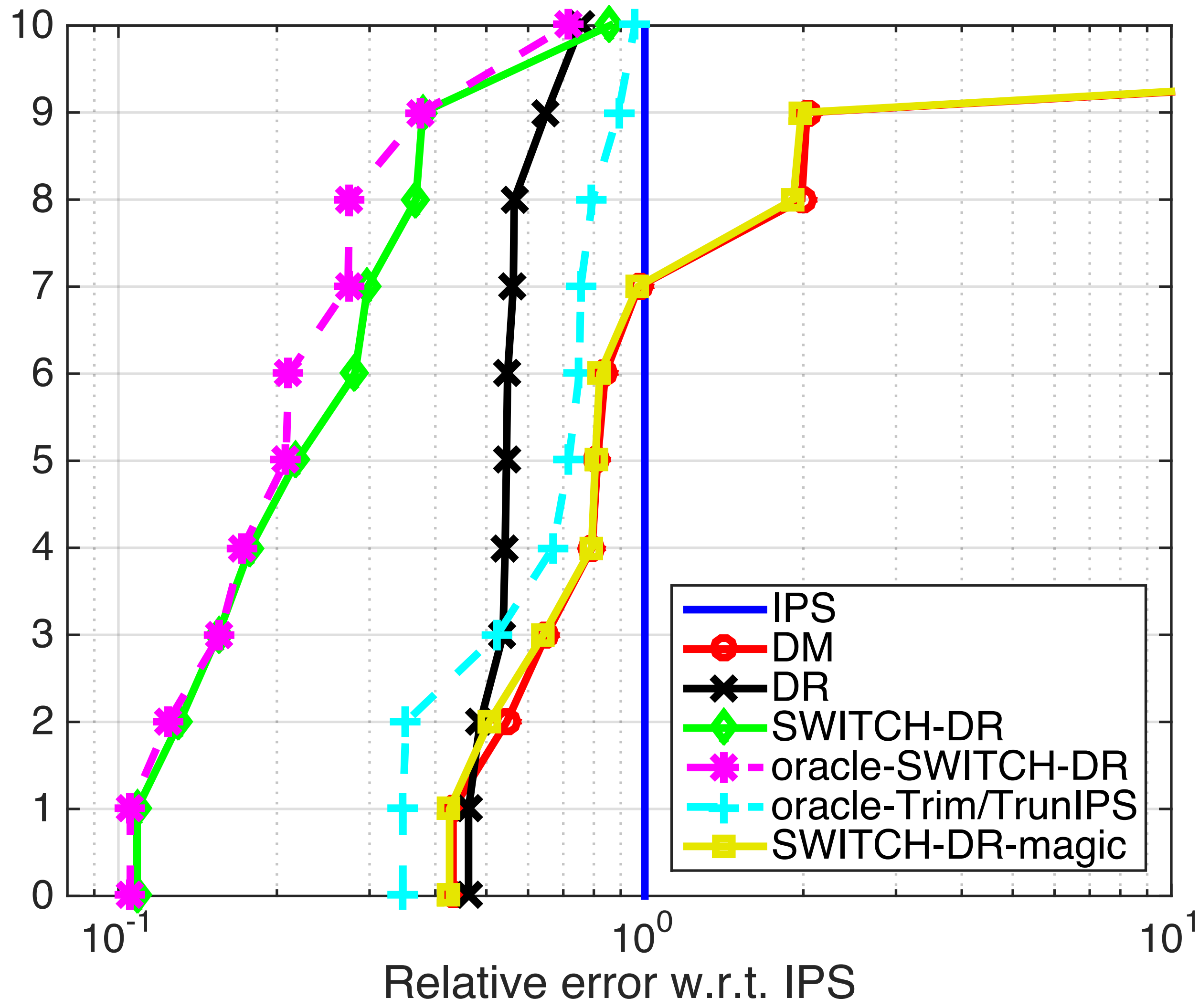
$$\mathrm{Bias}^2(\hat{v}_{\mathrm{SWITCH}}) \le \mathbb{E}_\mu[\rho\epsilon^2|\rho > \tau]\pi(\rho > \tau)^2 \le \mathbb{E}_\mu[\rho R_{\max}^2|\rho > \tau]\pi(\rho > \tau)^2$$

$$\approx \left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\pi\left(R_{\max}^2|\rho > \tau, x_i\right)\right]\left[\frac{1}{n}\sum_{i=1}^{n}\pi(\rho > \tau|x_i)\right]^2 =: \widehat{\mathrm{Bias}}_\tau^2.$$

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