Optimal and Adaptive Off-Policy Evaluation in Contextual Bandits

Yu-Xiang Wang

Joint work with Alekh Agarwal, Miro Dudik

Off-Policy Evaluation: Answering the "what-if" question

- Targeted advertisement
  - A "policy" decides which ad to show based on "context"
  - Then the user may click or not click
  - The click-through rate measures how good the policy is

- **What if** I ran a different policy instead?
  - a.k.a., Counterfactual reasoning

Many applications

- For safe policy deployment
- For policy optimization

Contextual bandits

- 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:
  - $v^{\boxed{\mu}} = \mathbb{E}_{x\sim\lambda}\mathbb{E}_{\boxed{a\sim\mu(\cdot|x)}}\mathbb{E}_D[r|x,a]$
- We collect data $(x_i, a_i, r_i)_{i=1}^n$ by the above processes.

- **What if** we use a different policy π (the "Target" policy)?
  - How do we estimate its value?

Importance sampling/Inverse propensity scoring

(Horvitz & Thompson, 1952)

Importance weights

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

$=: \rho_i$

Pros:

- No assumption on rewards
- Unbiased
- Computationally efficient

Cons:

- High variance when the weight is large

Model-based approach

- Fit a regression model of the reward

$$\hat{r}(x,a) \approx \mathbb{E}(r|x,a)$$ using the data

- Then for any target policy

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

Pros:

- Low-variance.
- Can evaluate on unseen contexts

Cons:

- Often high bias
- The model can be wrong/ hard to learn

Variants and combinations

- Modifying importance weights:
  - Trimmed IPS (Bottou et. al. 2013)
  - Truncated/Reweighted IPS (Bembom and van der Laan,2008)

- Doubly Robust estimators:
  - A systematic way of incorporating DM into IPS
  - Originated in statistics (see e.g., Robins and Rotnitzky, 1995; Bang and Robins, 2005)
  - Used for off-policy evaluation (Dudík et al., 2014)

Many estimators are proposed. Are they optimal? How good is good enough?

In this work, we formally address these problems.

1. Minimax lower bound: IPS is optimal in the general case.

2. A new estimator --- SWITCH --- that can be even better than IPS in some cases.

What do we mean by "optimal"?

- Minimax theory
  - Find an estimator that works well for ALL problem within a class of problems.
  - An estimator $\hat{v} : (\mathcal{X} \times \mathcal{A} \times \mathbb{R})^n \rightarrow \mathbb{R}$
  - Minimax risk / rate:

$$\inf_{\hat{v}} \sup_{\text{a class of problems}} \boxed{\mathbb{E}}(\hat{v}(\text{Data}) - v^{\pi})^2$$

Taken over data ~ μ

  - Fix context distribution and policies (λ, μ, π)
  - A class of problems = a class of reward distributions.

What do we mean by "optimal"?

- The class of problems: (generalizing Li et. al. 2015)

$$\mathcal{R}(\sigma, R_{\max}) := \Big\{ D(r|x,a) : 0 \le \mathbb{E}_D[r|x,a] \le R_{\max}(x,a) \text{ and } \mathrm{Var}_D[r|x,a] \le \sigma^2(x,a) \text{ for all } x, a \Big\}.$$

- The minimax risk

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

Lower bounding the minimax risk

- Our main theorem: under mild conditions

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

$$= \Omega\left[\frac{1}{n}\left(\underbrace{\mathbb{E}_\mu[\rho^2\sigma^2]}_{\text{Randomness in reward}} + \underbrace{\mathbb{E}_\mu[\rho^2 R_{\max}^2]}_{\text{Randomness due to context distribution}}(1-\tilde{O}(n\boxed{\lambda_0}))\right)\right]$$

Max prob. of a single x

- Subsumes lower bound for multi-arm bandit.

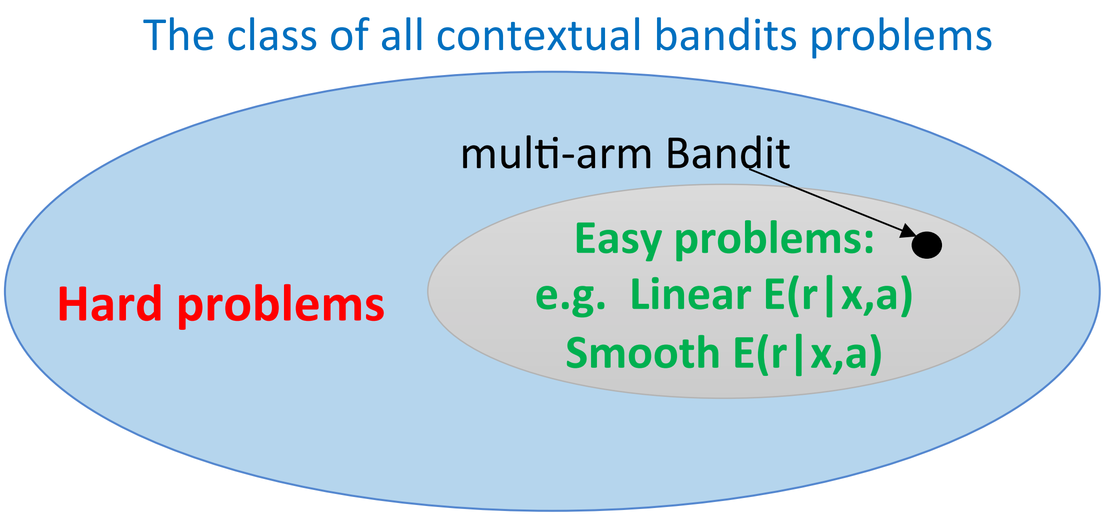
Li, Lihong, Rémi Munos, and Csaba Szepesvári. "Toward Minimax Off-policy Value Estimation." *AISTATS*. 2015.

This implies that IPS is optimal!

- The high variance is required.
  - In contextual bandits with large context spaces and non-degenerate context distribution.
  - Model-free approach is fundamentally limited.

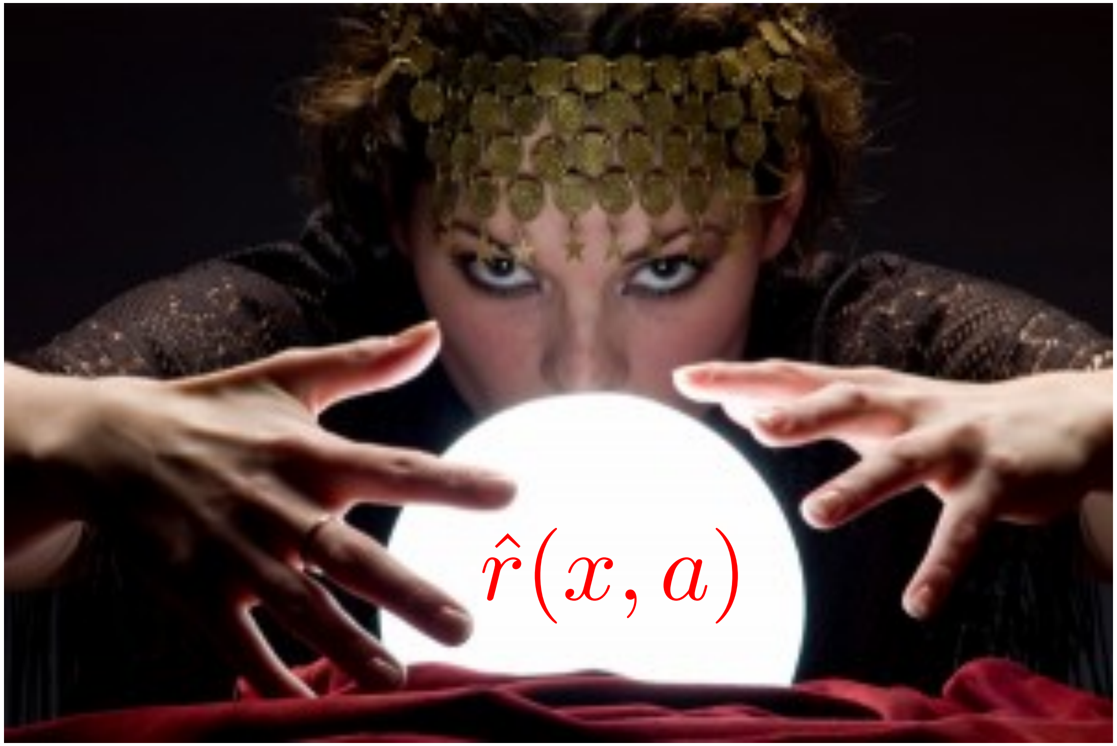
- Different from multi-arm bandit
  - Li et. al. (2015) showed that in k-arm bandit, IPS is strictly suboptimal.

The pursuit of adaptive estimators



- Minimaxity: perform optimally on hard problems.
- Adaptivity: perform better on easier problems.

Suppose we are given an oracle



- Could be very good, or completely off.
- How to **make the best use** of the predictions?

Why not just use doubly robust?

- Originated in statistics (see e.g.: Robins and Rotnitzky, 1995; Bang and Robins, 2005)
- Proposed for off-policy evaluation previously:

  Dudık, Langford and Li. "Doubly Robust Policy Evaluation and Learning.“ *ICML*-11.
  Jiang and Li. "Doubly Robust Off-policy Value Evaluation for Reinforcement Learning." *ICML*-2016.

- We show that: DR can be as bad as IPS
- Does not adapt even with perfect oracle:

$$\hat{r}(x,a) = \mathbb{E}(r|x,a)$$

$$\mathrm{MSE}(\hat{v}_{\mathrm{DR}}) \leq \frac{1}{n}\left(\boxed{\mathbb{E}_\mu(\rho^2\sigma^2)} + \mathbb{E}_\pi(R^2_{\max})\right)$$

DR can suffer from high variance just like IPS!

SWITCH estimator

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

- 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 the oracle estimator.

The approach is related to MAGIC estimator (Thomas & Brunskill, 2016), but with important difference.

Error bounds for SWITCH

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

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

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

$$+ \ \frac{2}{n}\mathbb{E}_\pi\underbrace{[\epsilon|\rho > \tau]^2 \mathbb{P}_\pi(\rho > \tau)^2}_{(3)}$$

1) Variance from IPS (reduced truncation)

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

3) Bias from the oracle.

Error bounds for SWITCH

- **For appropriately tuned ``threshold'' parameter, SWITCH is**
  - **Independent to ρ when oracle is perfect.**
  - **Minimax when oracle is horrible.**
  - **Robust to large importance weight.**

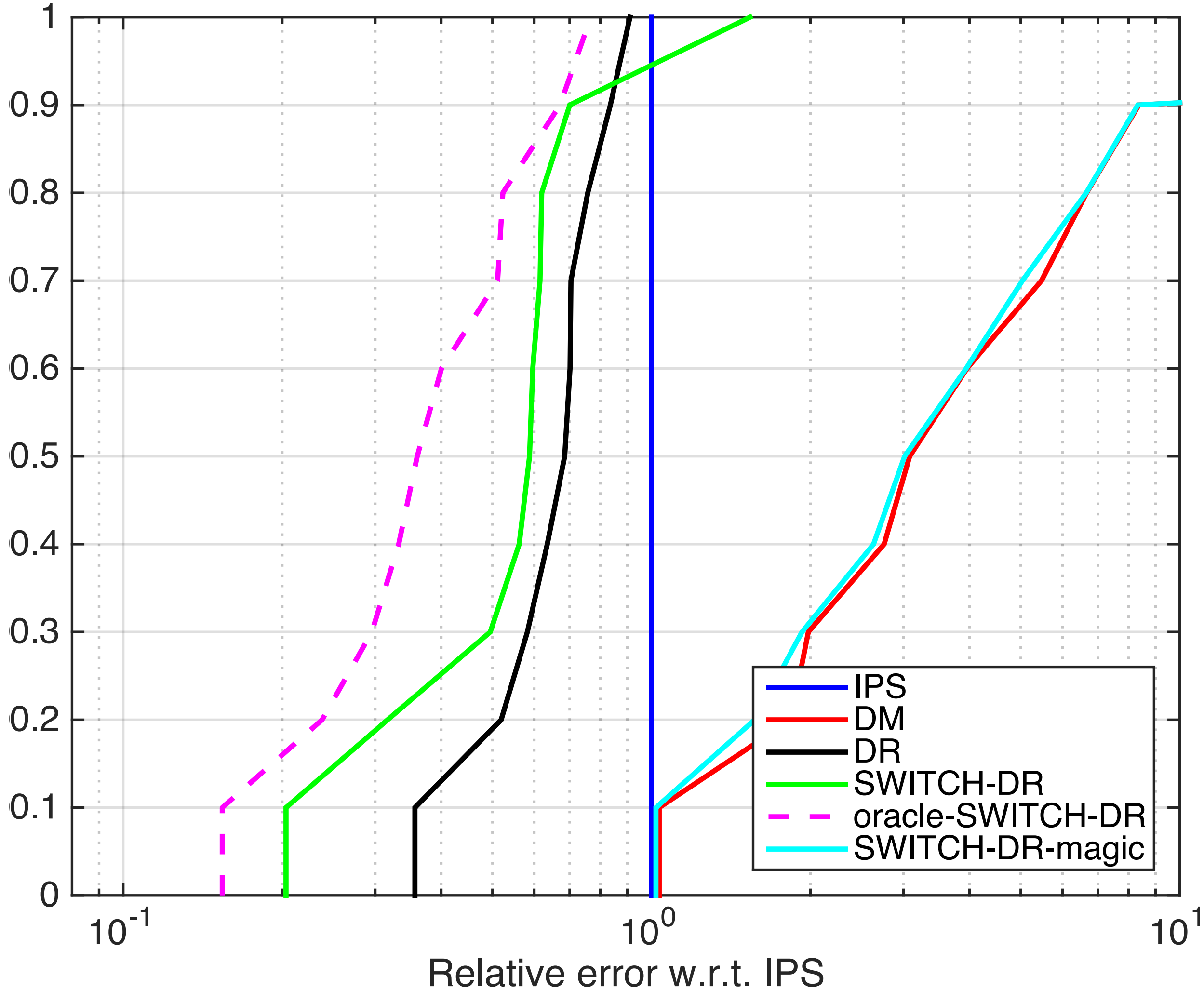
- Data dependent tuning of parameter? Check out our paper!
- Different from MAGIC (Thomas and Brunskill, 2016)

Experiment setup

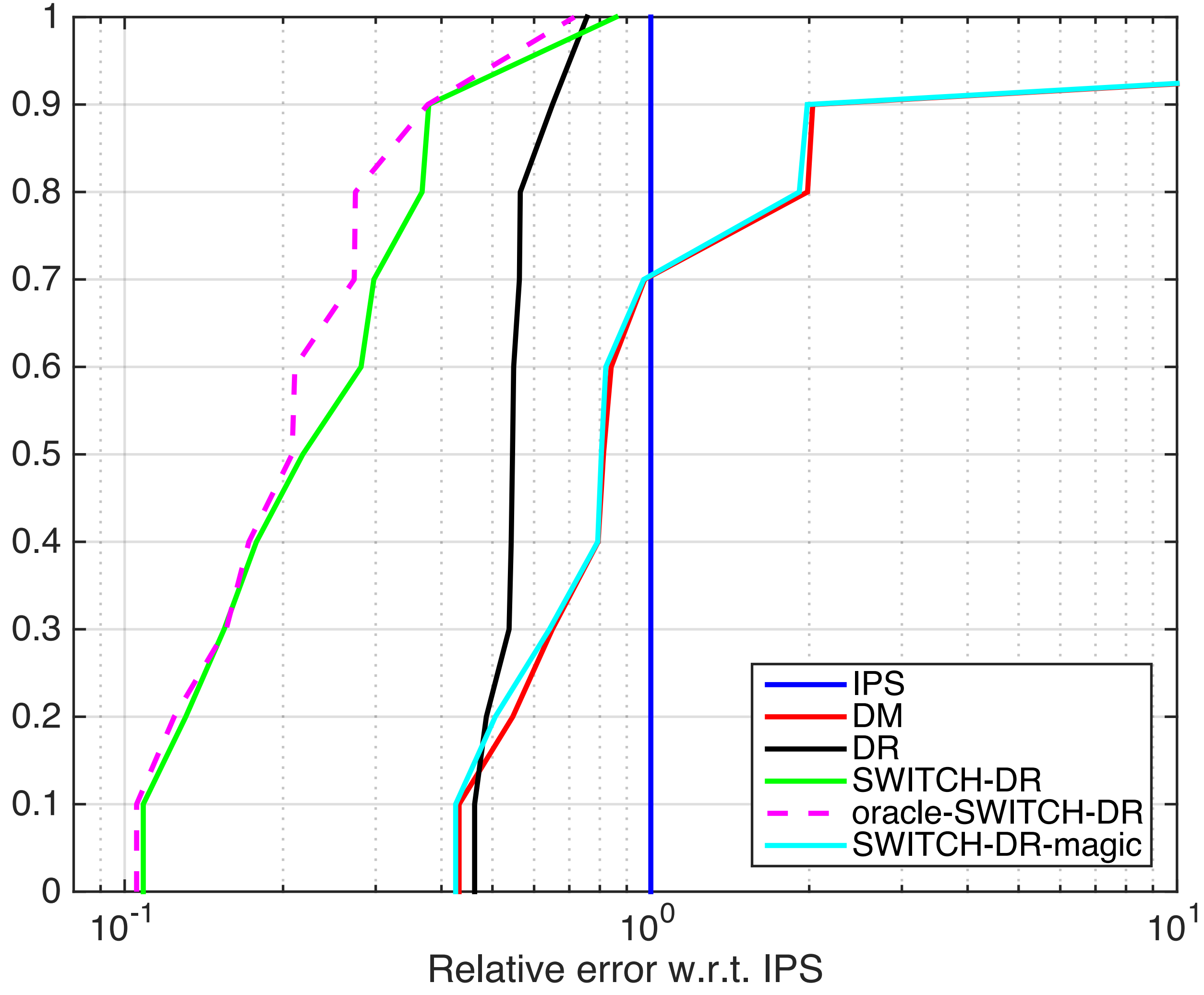
- 10 UCI Classification data sets converted to bandits.
  - Action is to predict labels.
  - Reward is {0,1}, depending on whether the action is correct.

- Follow standard setup in
  - (Beygelzimer & Langford, 2009)
  - (Gretton et. al. 2008)
  - (Dudik et. al. 2011)

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



With additional label noise



Conclusion

- IPS is optimal.
  - Need to go beyond the model-free approach.

- DR is unsatisfactory.

- We propose an new estimator: SWITCH
  - that has good theoretical properties.
  - performs quite well in practice.

Thank you! Any questions?

Connections and future work

- Extension to reinforcement learning
  - Lower bound directly applies in some sense.
  - SWITCH-DR for reinforcement learning?

- Lower bound directly applies to "mean effect" estimation.
  - Basically it corresponds to a different "target policy".

The conditions for the main Theorem

- Moment conditions:

$$\mathbb{E}_\mu[(\rho\sigma)^{2+\epsilon}] \leq \infty$$

$$\mathbb{E}_\mu[(\rho R_{\max})^{2+\epsilon}] \leq \infty$$

$$\mathbb{E}_\mu[\sigma^2/R_{\max}^2] < \infty$$

- If n is sufficiently large

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

$$= \Omega\left[\frac{1}{n}\left(\mathbb{E}_\mu\left[\rho^2\sigma^2\right] + \mathbb{E}_\mu\left[\rho^2 R_{\max}^2\right]\left(1-110\lambda_0\log(4/\lambda_0)\right)\right)\right]$$

Automatic parameter tuning

- Conservative approximate MSE minimizing.

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

- Details:

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

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

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

Experiment setup

- 10 UCI Classification data sets converted to bandits.
  - Action is to predict labels.
  - Reward is {0,1}, depending on whether the action is correct.
  - Target policy is prediction of logistic regression.
  - Logging policy obtained by the label probability of a logistic regression learned from covariate shifted data.

- We sample data of size n = [100, 200,500,1000,...], from discrete distribution of of length N.