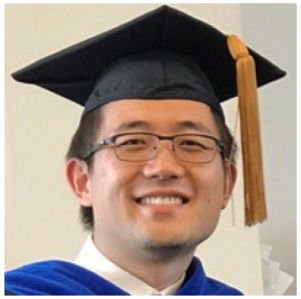
Off-policy Learning in Theory and in the Wild

Yu-Xiang Wang

Based on joint works with

Alekh Agarwal, Miro Dudik, Yifei Ma, Murali Narayanaswamy

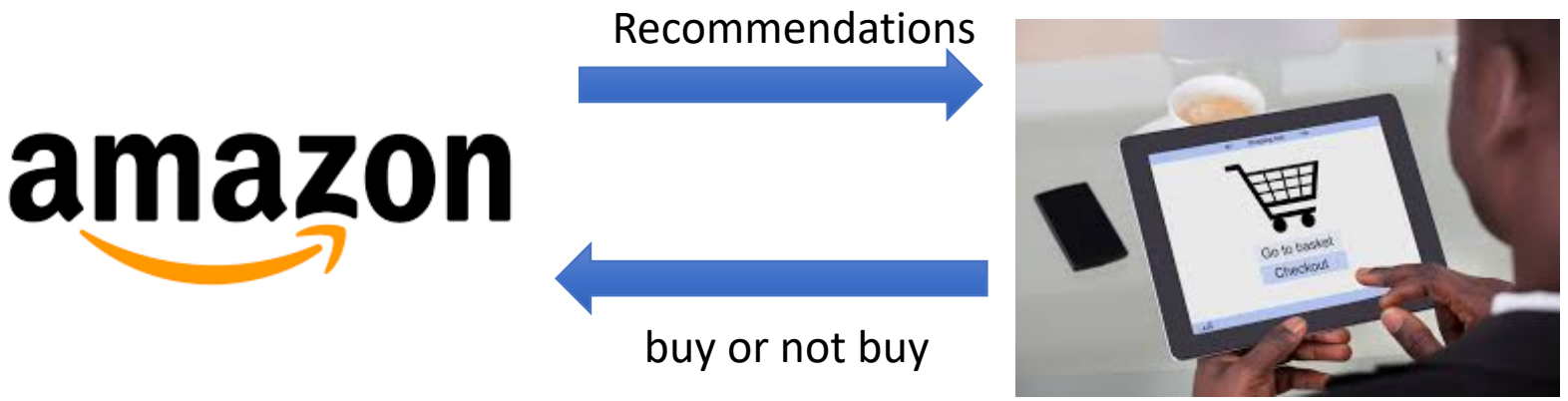





Outline

- Off-policy evaluation and ATE estimation
  - A finite sample optimality theory
  - The SWITCH estimator
- Off-policy learning in the real world
  - Challenges: missing logging probabilities, confounders, model misspecification, complex action spaces
  - Solutions

Off-Policy learning: an example



How to evaluate a new algorithm without actually running it live?

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.

Off-policy Evaluation and Learning

Off-policy evaluation

Estimate the value of a fixed target policy π

$$v_\pi := \mathbb{E}_\pi[\text{Reward}]$$

Off-policy learning

Find $\pi \in \Pi$

that maximizes v_π

- Using data $(x_i, a_i, r_i)_{i=1}^n$
- often the policy μ or logged propensities $(\mu_i)_{i=1}^n$

ATE estimation is a special case of off-policy evaluation

- a: Action ⇔ T: Treatment {0,1}
- r: Reward ⇔ Y: Response variable
- x: Contexts ⇔ X: covariates

- Take a = {0,1}, π = [0.5,0.5]
- r (x,a) = [2Y(X,T=1), -2Y(X,T=-1)]

- Then, the value of π = ATE

Direct Method / Regression estimator

- 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

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

Pros:

- Low-variance.
- Can evaluate on unseen contexts

Cons:

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

Inverse propensity scoring / Importance sampling

(Horvitz & Thompson, 1952)

Importance weights

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

$\therefore \rho_i$

Pros:

- No assumption on rewards
- Unbiased
- Computationally efficient

Cons:

- High variance when the weight is large

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 out there. Are they optimal? How good is good enough?

Our results in (W., Agarwal, Dudik, ICML-17):

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?

- A minimax formulation

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

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: 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

This implies that IPS is optimal!

- The high variance is required.
  - In contextual bandits with large context spaces and non-degenerate context distribution.

- Previously, IPS is known to be asymptotically inefficient
  - for multi-arm bandit (Li et. al., 2015)
  - for ATE. (Hahn, Hirano, Imbens)

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

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

- Our lower bound is bigger!

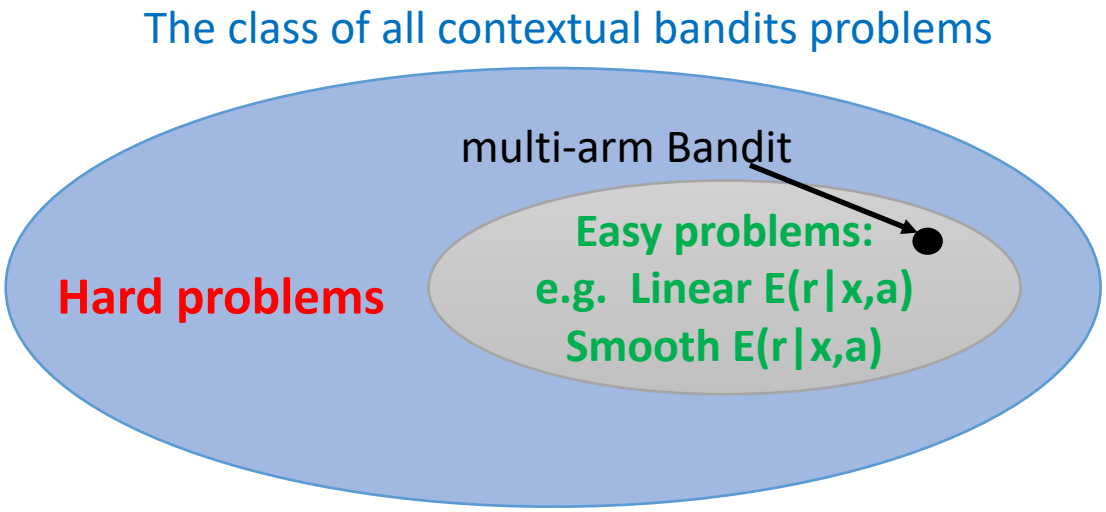
$$\mathbb{E}_{\mu}\left[\rho^2\sigma^2\right]+\mathbb{E}_{\mu}\left[\rho^2 R_{\max}^2\right]$$

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!

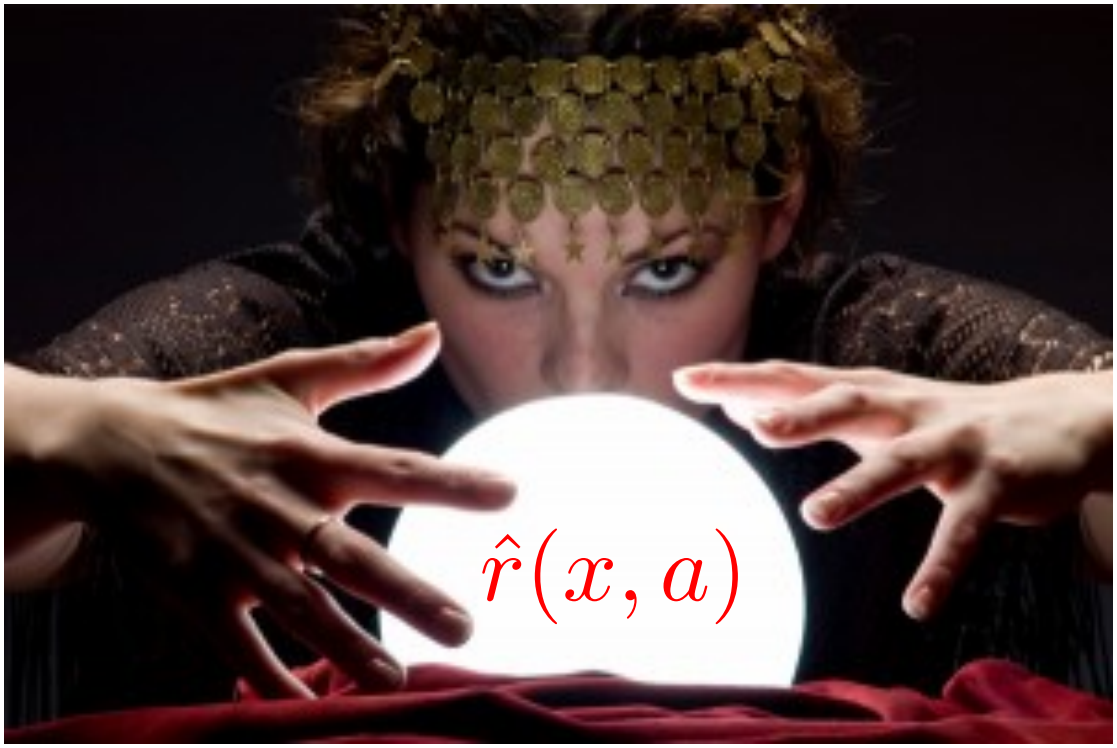
Assumption:	**Realizable assumption:** **E[r|x,a] is differentiable in x for each a.**	**No assumption on E[r|x,a] excepts boundedness.**
Consequences	Hirano et. al. is optimal. Imbens et. al. is optimal. IPS is suboptimal!	IPS is optimal (up to a universal constant)
Caveat	Poor finite sample performance.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

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 it?

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 the oracle estimator.

Error bounds for SWITCH

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

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

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

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

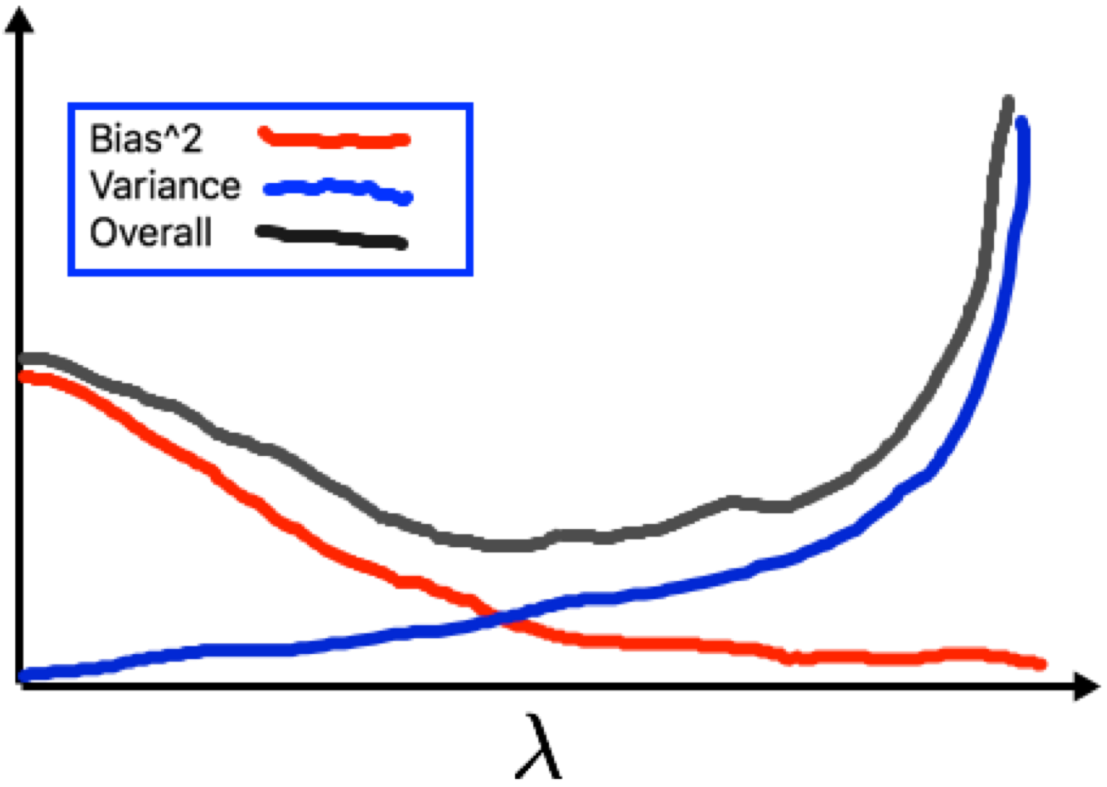
1) Variance from IPS (reduced by truncation)

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

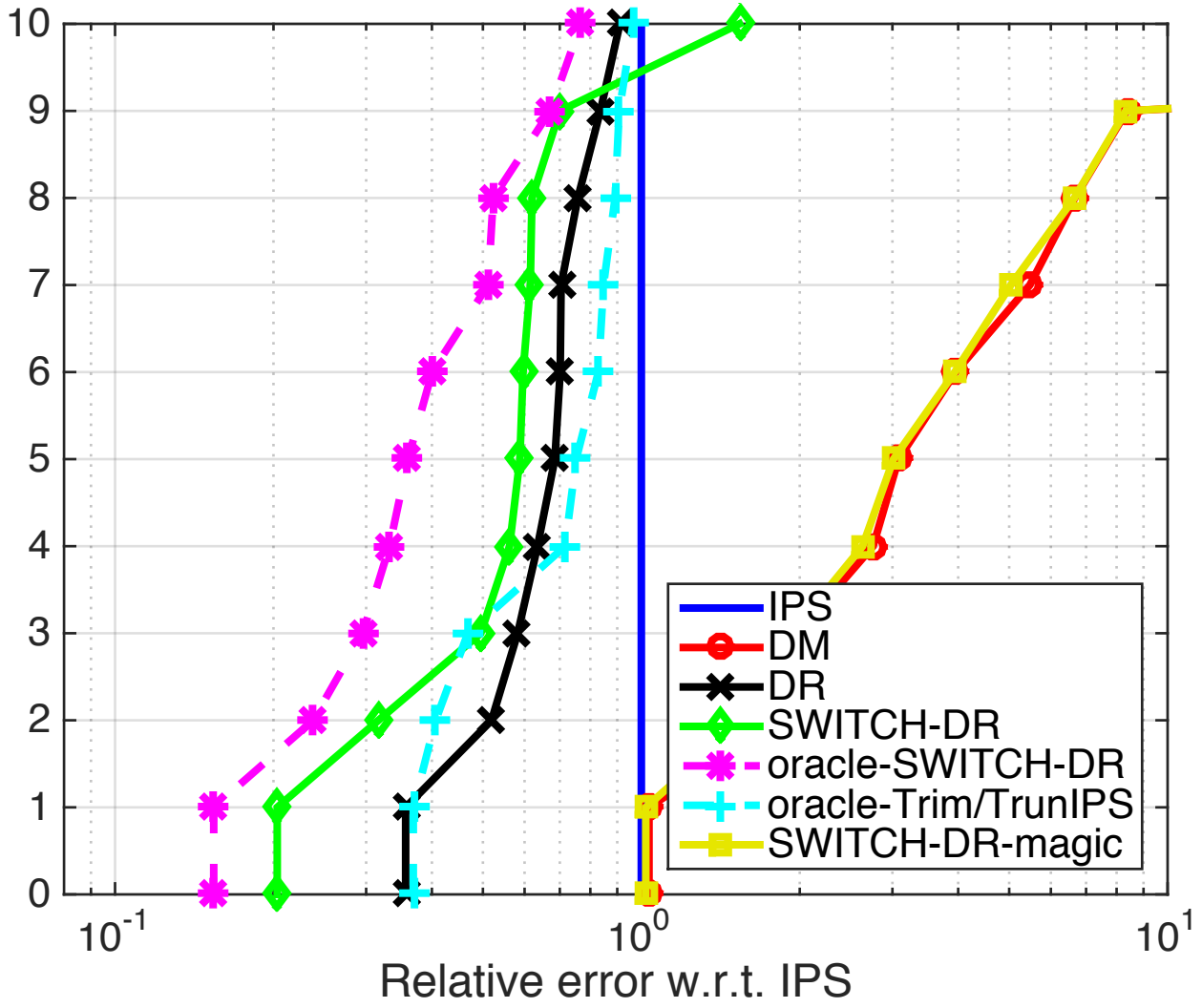
1) Bias from the oracle.

How to choose the threshold?

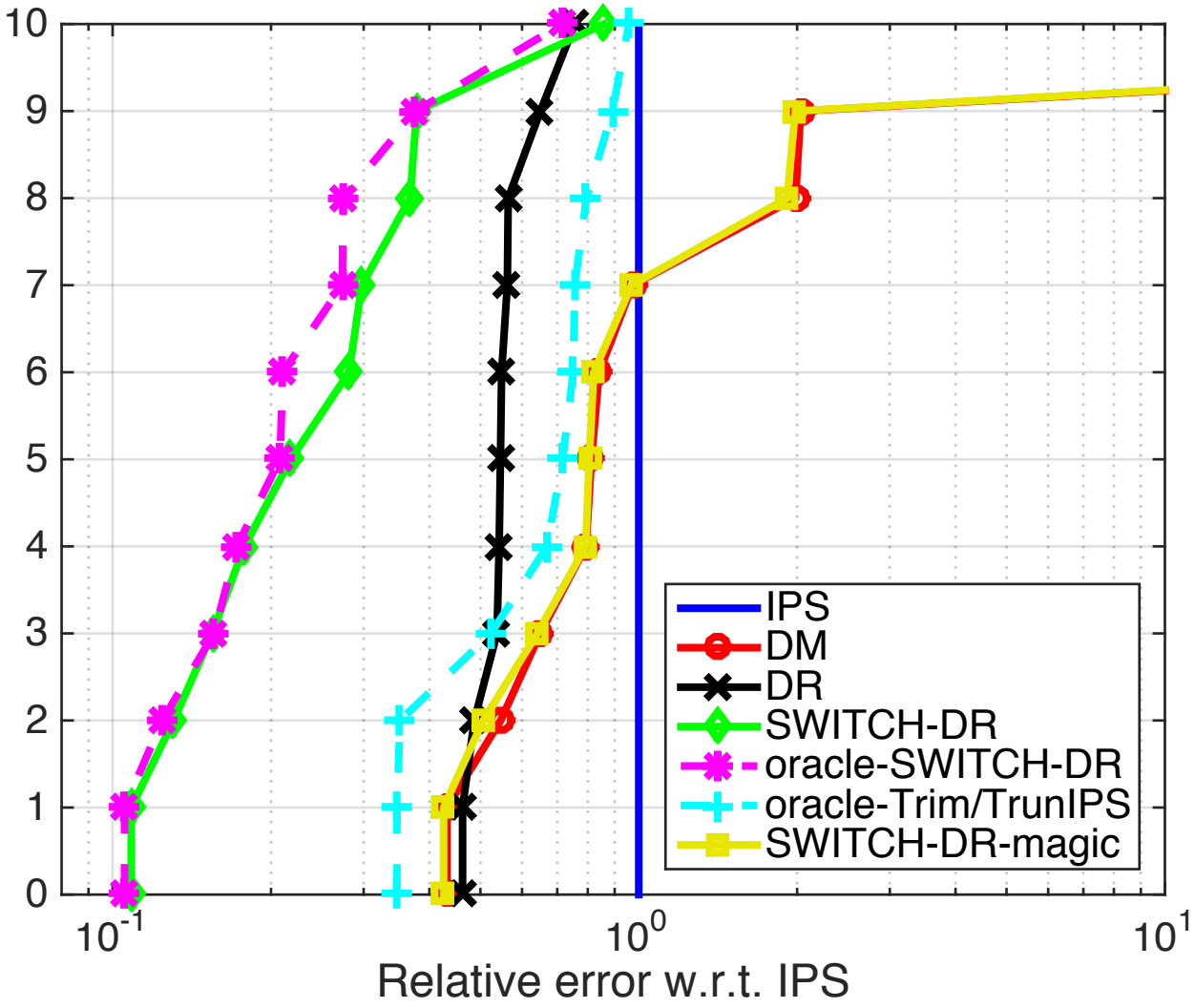
- Be conservative:
  - Minimize the variance + square bias upper bound.



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



With additional label noise



Quick summary of Part I

- Off-policy evaluation ⇔ a generalized ATE estimation
- Simple IPS cannot be improved except when having access to a realizable model.
- Best of both world: doubly robust and SWITCH

Part 2: Off-Policy learning in the Wild!

Off-policy evaluation

Estimate the value of a fixed target policy π

$$v_\pi := \mathbb{E}_\pi[\text{Reward}]$$

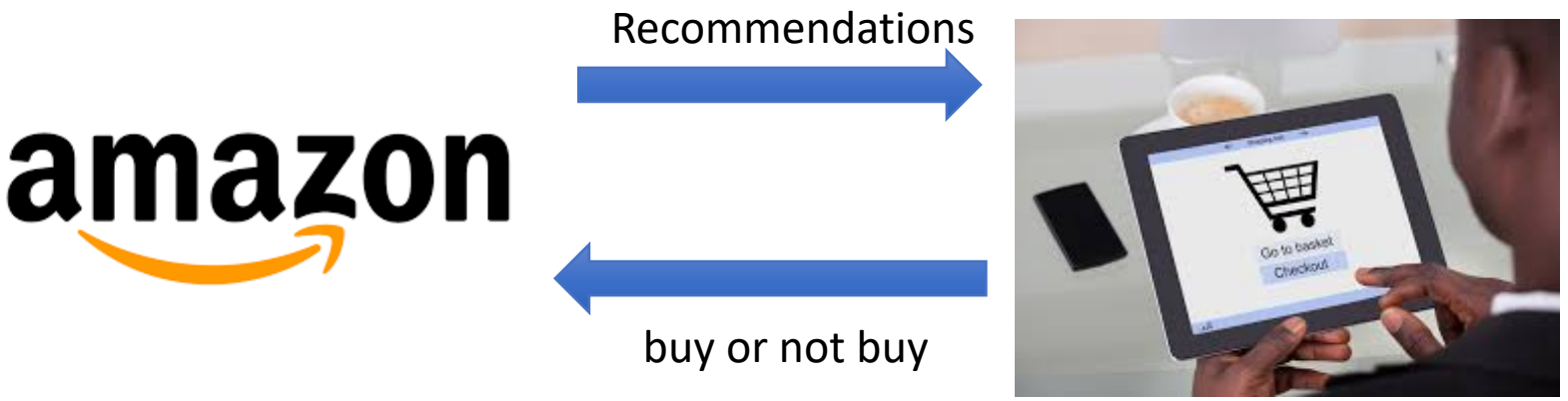
Off-policy learning

Find $\pi \in \Pi$

that maximizes v_π

- Data set $(x_i, a_i, r_i)_{i=1}^n$
- Logging policy $(\mu_i)_{i=1}^n$

Recommendation systems

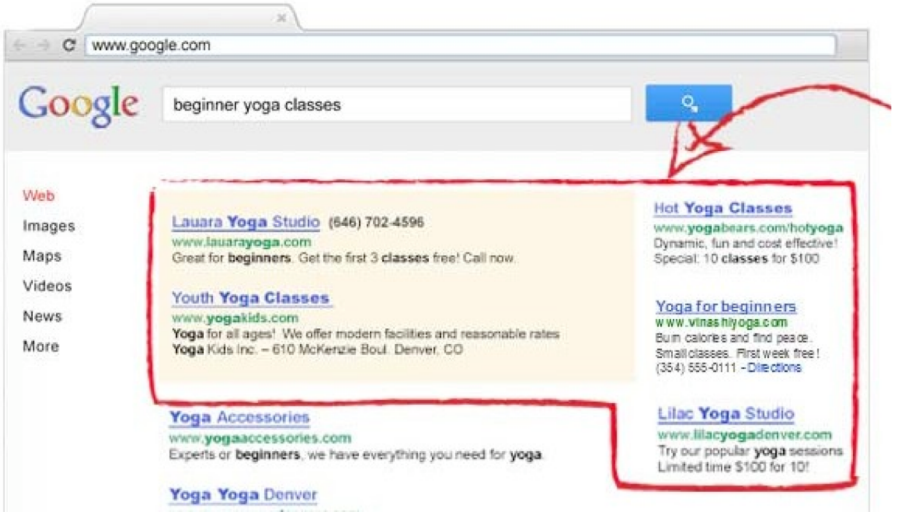


What's commonly being done in industry is collaborative filtering.

This is a direct method!

Serving ads: Google/Criteo/Facebook

- x = context/user features
- a = ads features
- r = click or not.



- Typical approach:
  - Some feature embedding of (x,a) into a really high-dimensional $\phi(x,a)$
  - L1-regularized logistic regression to predict r

This is again a direct method!

Challenges of conducting offline learning in the wild

1. Reward models are always non-realizable
   - Direct methods are expected to have nontrivial bias.

2. Large action space, large importance weight
   - Think how many webpages are out there!

3. Missing logging probabilities
   - Even if randomized, we might not not know the logging policy

4. Confounders: unrecorded common cause of action and reward!
   - Ad-hoc promotion, humans operator overruling the system.

Ma, W. and Narayanaswamy (2018) "Imitation-Regularized Offline Learning." under review.

Two additional assumptions

- The expected reward obeys that

$$0 \leq \mathbb{E}[r|x,a] \leq R, \quad \forall x, a$$

- Not a strong assumption.
  - Satisfied in most applications, e.g., Click-Through-Rate optimization.
  - Bottou et. al. used this to construct lower bounds.
  - The reason why weight clipping / SWITCH works.

Click-prediction by multiclass classification

$$\underset{\pi}{\operatorname{argmin}} \operatorname{CE}(\pi; r) = -\frac{1}{n} \sum_{i=1}^{n} r_i \log \pi(a_i \mid x_i).$$

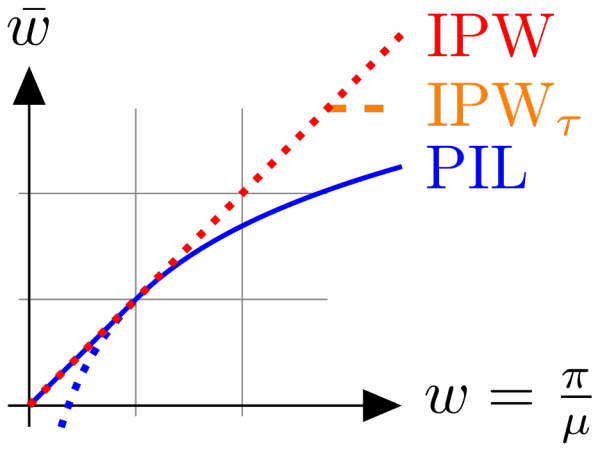
- Only when the user clicked, do we have a label.
- Train a multiclass classifier using the labeled examples.

Causal interpretation of cross entropy-based direct click prediction

$$\frac{1}{n}\sum_{i=1}^{n} r_i \log \pi(a_i \mid x_i) - \log(\mu(a_i|x_i))$$

$$= \frac{1}{n}\sum_{i=1}^{n} r_i \log\left(\frac{\pi(a_i|x_i)}{\mu(a_i|x_i)}\right)$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} r_i \left(\frac{\pi(a_i|x_i)}{\mu(a_i|x_i)} - 1\right)$$



Policy Improvement Lower bound (PIL)

Causal interpretation of cross entropy-based direct click prediction

$$\frac{1}{n}\sum_{i=1}^{n} r_i \log \pi(a_i \mid x_i) - \log(\mu(a_i|x_i))$$

- Implicitly maximize a lower bound of the counterfactual objective **without** having the logging probabilities!
- **Adaptive** to any unknown logging probabilities.
- We **can optimize but cannot evaluate** the lower bound!

Our solution: Imitate the policy!

$$\mathrm{KL}(\mu \| \pi) = \mathbb{E}_\mu \log \frac{\mu}{\pi} = -\mathbb{E}_\mu \log w.$$

Fully observed:

$$\mathrm{IML}_{\mathrm{full}}(\pi) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}(x_i)} \mu(a \mid x_i) \log w(a \mid x_i)$$

Partially observed:

$$\mathrm{IML}_{\mathrm{part}}(\pi) = -\frac{1}{n} \sum_{i=1}^{n} \log w_i;$$

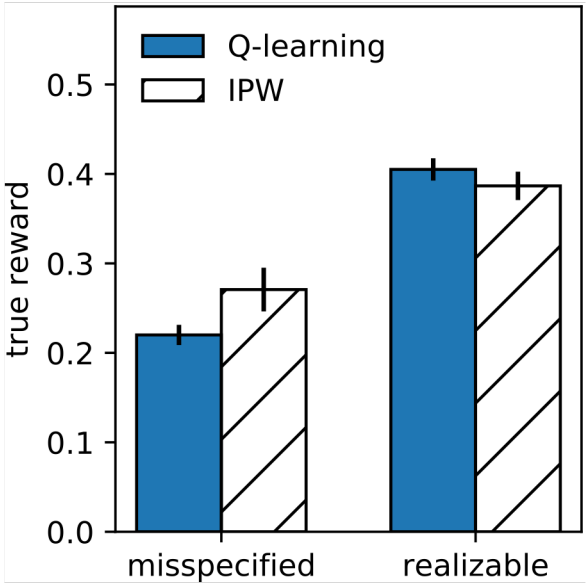
Completely missing:

$$\mathrm{IML}_{\mathrm{miss}}(\pi) = -\frac{1}{n} \sum_{i=1}^{n} \log \pi(a_i \mid x_i) - \mathrm{CE}(\mu; 1)$$

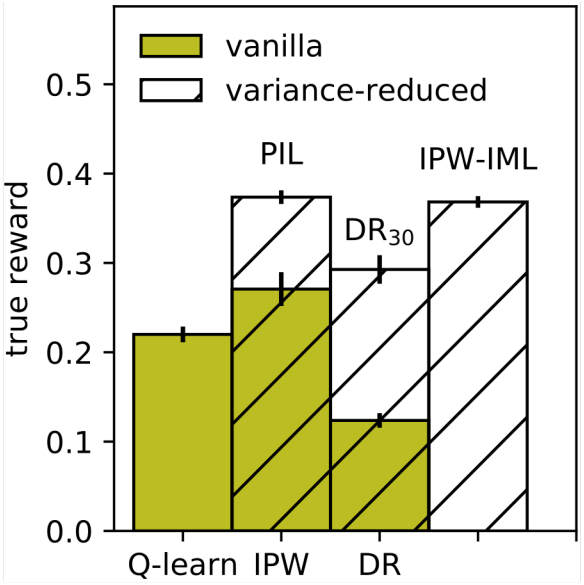
Usage of IML

- To be use as a regularization
  - Closely related to safe-policy improvements.
  - Natural policy gradient
- To diagnose whether there is a confounder
  - If IML solution is nearly 0, then we have good evidence that there is no confounder
- To collect new data using IML policy

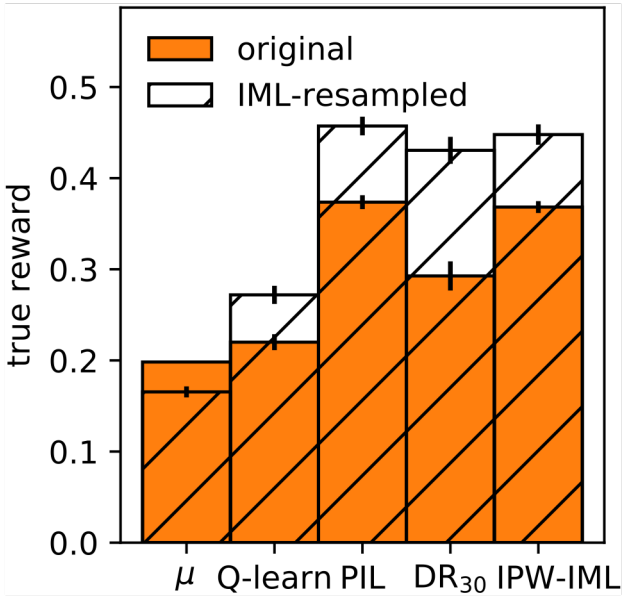
Results on UCI Data Set: Optdigit



(a) Unbiased IPW is better than Q-learning with misspecified models.



(b) Variance reduction techniques further improve offline learning.



(c) Online application of IML improves future offline learning.

The Criteo Counterfactual Dataset

- 1 million records on Ad impressions.
- 250 GB hosted on AWS

$$(x_i, a_i, r_i)_{i=1}^n$$

- They actually ran a randomized policy. And have logged $(\mu_i)_{i=1}^n$

Nothing gets better than their logging policy, but we are close…

Table 5: Criteo counterfactual analysis dataset.

Approach	IPW (bp $\times 10^4$)	Gap (%)
Logging policy	53.4 ± 0.2	0.0 ± 0.0
Uniform random	43.6 ± 3.4	5.6 ± 0.6
Q-learning	50.5 ± 3.4	4.1 ± 0.6
Vanilla IPW+	49.9 ± 1.8	0.1 ± 1.6
IPW_{500}	52.0 ± 3.0	0.9 ± 0.5
DR+	53.7 ± 14.4	-1.5 ± 5.7
IML	52.5 ± 2.4	0.3 ± 0.4
IPW_{500}-10^{-3}IML	53.5 ± 2.8	0.3 ± 0.4
POEM+ [27]	52.7 ± 1.6	0.3 ± 0.6

Summary of Part II

1. Reward models are always non-realizable
   - Use a counterfactual objective!
2. Large action space, large importance weight
   - Optimize a low-variance lower bound instead!
3. Missing logging probabilities
   - You might not need them. If you do, use policy imitation!
4. Confounders
   - Can be detected using policy imitation!

Take home messages

- Off-policy evaluation under the contextual bandits models is closely related to causal inference.
- Minimax optimality depends on assumptions, but ultimately we need estimators that are adaptive in finite sample.
- Offline policy learning is often an orthogonal problem. We can optimize a counterfactual lower bound without knowing the propensities.
- Policy imitation as a regularization and as a diagnosis tool.

Thank you for your attention!

Reference:

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

Ma, W., Narayanaswamy (2018) Imitation-Regularized Offline Learning. Under review.