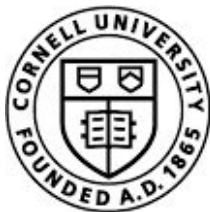


Off-Policy Evaluation for Large Action Spaces via Embeddings (ICML2022)

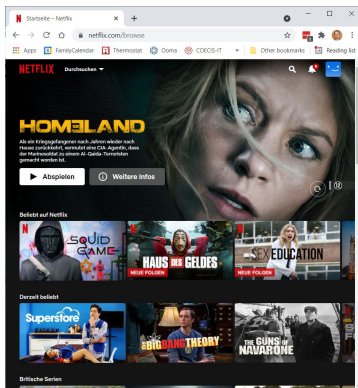
Yuta Saito and Thorsten Joachims



Cornell Bowers C-IS
Computer Science

Off-Policy Evaluation (OPE) of Contextual Bandits

OPE aims at estimating the **value** of *evaluation (new) bandit policy* π_e



身長と体重で選ぶマルチサイズアイテム

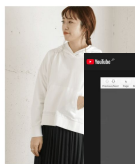
人気ブランドのアイテムをあなたに理想のサイズで



ITEMS URBANRESEARCH

¥4,290

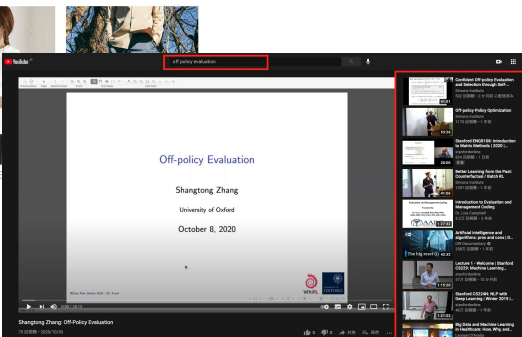
MS マルチサイズ



ITEMS URBANRESEARCH

¥4,950

MS マルチサイズ



What would the system have performed if a new policy was deployed instead?

logged bandit data

$$\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^n \sim \pi_0$$


collected by
a “logging” policy

Benchmark Estimator: Inverse Propensity Score (IPS)

IPS provides an unbiased estimation of the policy value

$$\hat{V}_{\text{IPS}}(\pi_e; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{\pi_e(a_i|x_i)}{\pi_0(a_i|x_i)}}_{w(x_i, a_i)} \cdot r_i$$

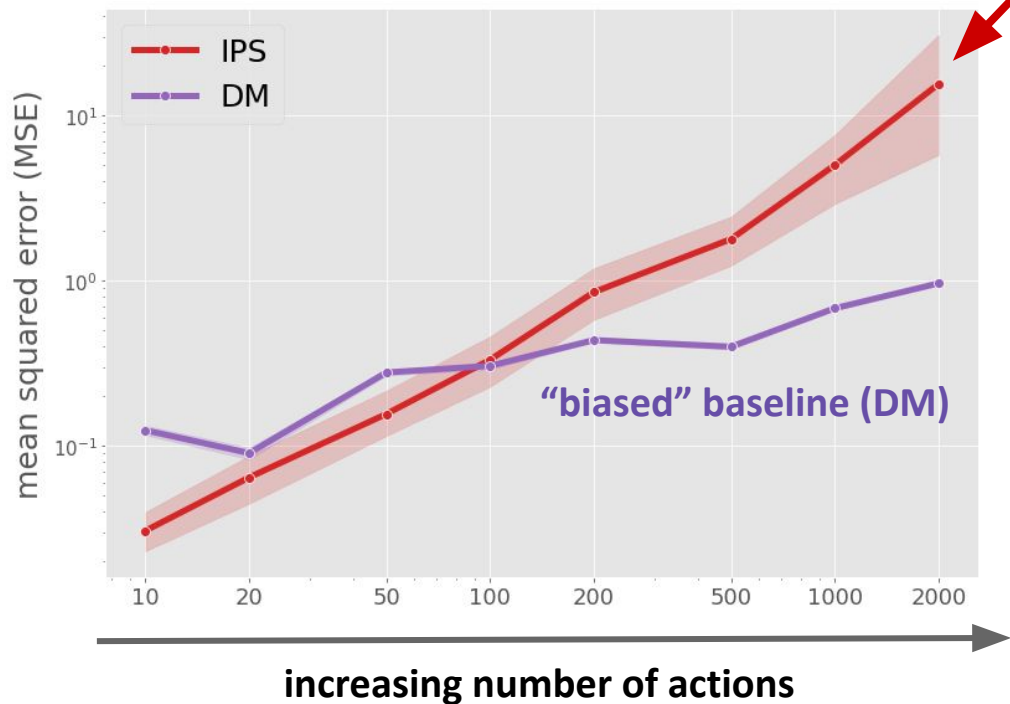
(vanilla)
importance weight



Many recent estimators are based on IPS, however,
its variance becomes very large in large action spaces

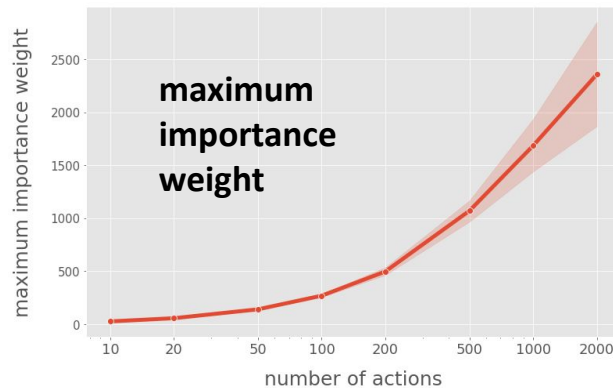
High Variance Issue of IPS for Large Action Spaces

number of data=3000



IPS is getting significantly worse with growing number of actions

This is simply due to the use of the vanilla importance weight $w(x, a)$



Large-Scale Applications of Off-Policy Evaluation

- video recommendation (Youtube)
- playlist recommendation (Spotify)
- artwork personalization (Netflix)
- search optimization (Amazon)

Large Action Spaces

thousands/millions
(or even more) of actions



**How can we achieve a large variance reduction
allowing only minimal bias even in large action spaces?**

Idea: Auxiliary Information about the Actions

The key idea: **why not leveraging auxiliary data about the actions?**

typical logged
bandit data for OPE

$$\mathcal{D} := \{(x_i, a_i, r_i)\}$$



we additionally observe
action embeddings

logged bandit data
w/ **action embeddings**

$$\mathcal{D} := \{(x_i, a_i, \textcolor{red}{e}_i, r_i)\}$$

category, price, actor, review comments, image, etc..

Marginalized Inverse Propensity Score (MIPS)

Leveraging the action embeddings, we propose the following new estimator

$$\hat{V}_{\text{MIPS}}(\pi_e; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{p(e_i | x_i, \pi_e)}{p(e_i | x_i, \pi_0)}}_{w(x_i, e_i)} \cdot r_i$$

vanilla importance weight of IPS

$$w(x, a) := \frac{\pi_e(a|x)}{\pi_0(a|x)}$$

marginal importance weight computed
with the **marginal embedding distribution**

$$p(e|x, \pi) = \sum_{a \in \mathcal{A}} p(e|x, a) \pi(a|x)$$

Highlights of Theoretical Analysis

- **Unbiased** under **No Direct Effect + Common Embed Support**

$$\mathbb{E}_{\mathcal{D}} [\hat{V}_{\text{MIPS}}(\pi_e; \mathcal{D})] = V(\pi_e)$$

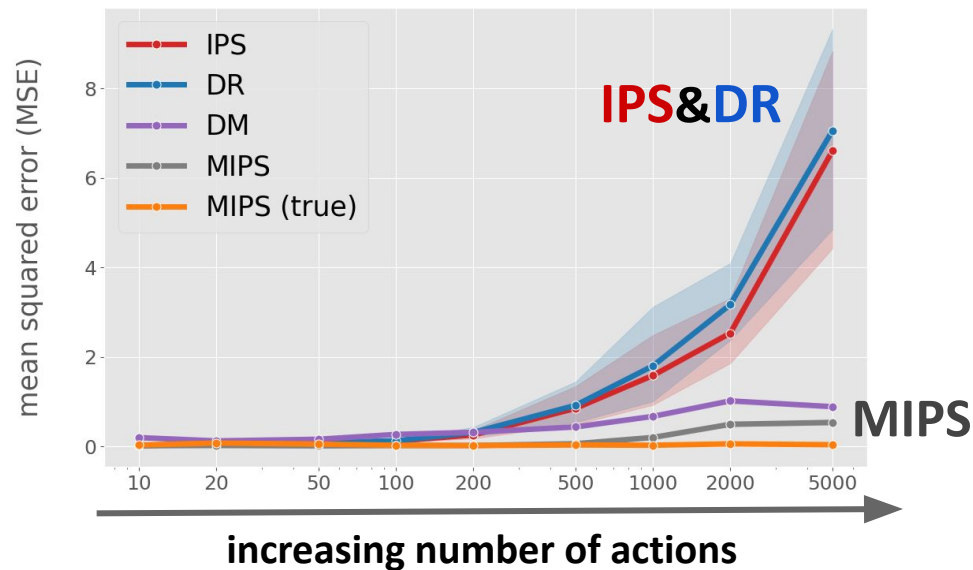
- **Larger Variance Reduction for Larger Actions Spaces** (compared to IPS)

$$\begin{aligned} & \mathbb{V}_{\mathcal{D}} [\hat{V}_{\text{IPS}}(\pi_e; \mathcal{D})] - \mathbb{V}_{\mathcal{D}} [\hat{V}_{\text{MIPS}}(\pi_e; \mathcal{D})] \\ & \propto \mathbb{E}_{x,e} [\mathbb{E}_{p(r|x,e)} [r^2] \cdot \mathbb{V}_{\pi_0(a|x,e)} [w(x,a)]] \geq 0 \end{aligned}$$

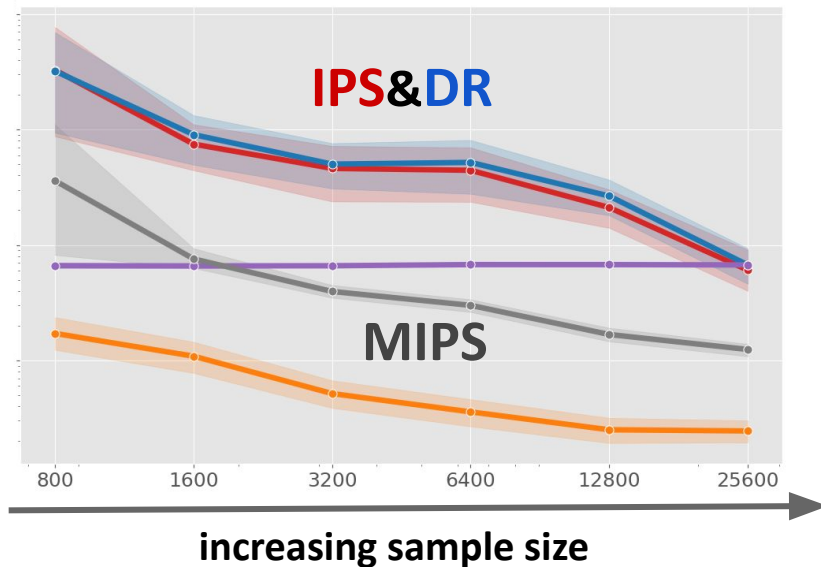
- **Characterizing the MSE, which is controlled by the quality of the action embeddings, and might be minimal when no direct effect is NOT true**

Strong Empirical Performance

more robust to growing action sets



converges much faster



MIPS enables effective OPE even in large action spaces

Summary



Long Talk
(on Youtube)

- **Our “Mrginalized IPS” enables effective OPE even in large action spaces**
- **Many other interesting theoretical/empirical results are in the paper!**
- **Come to our poster!**
Poster Session 3, **Today from 6 p.m. to 8 p.m.**