

RL Project Monsoon 2022

Off-Policy Evaluation for Large Action Spaces via Embeddings

Sahas Marwah 2020237

Sparsh Mehrotra 2020248

Yash Thakran 2020269

Repository Link:

<https://github.com/YashThakran/Off-Policy-Evaluation-for-large-action-spaces-via-embeddings>

Paper Link: <https://arxiv.org/pdf/2202.06317.pdf>



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**



Introduction

- Off-policy evaluation (OPE) in contextual bandits has seen rapid adoption in real-world systems, since it enables offline evaluation of new policies using only historic log data.
- But when the number of actions is large, existing OPE estimators degrade severely and can suffer from extreme bias and variance.
- To overcome this issue, a new OPE estimator is proposed that leverages additional information about the actions in the form of **action embeddings**. The estimator used the **marginal distribution** of action embeddings, to define a new type of **importance weights**.
- We compare the Bias, Variance, and Mean Squared Error of the proposed estimator and analyze the conditions under which the action embedding provides statistical benefits.

Research Paper

Yuta Saito and Thorsten Joachims, “Off-Policy Evaluation for Large Action Spaces via Embeddings”. International Conference on Machine Learning, 2022.

Paper Link: <https://arxiv.org/pdf/2202.06317.pdf>

Introduction and Key Terms

- **Action Embeddings:** In RL, action embeddings can be used to represent the actions available to an agent in a continuous, rather than discrete space. This can make it easier to learn good policies using techniques such as gradient descent, as the action space is no longer discrete and the optimization problem becomes continuous, hence smoother.
- **Importance Weights:** are used to correct for bias that can arise when using sample-based estimates to estimate the value of an action or state.
- **Marginal Distribution:** it is defined as

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

Off Policy Evaluation

- Off-Policy:

$$V(\pi_e) := \mathbb{E}_{p(x)\pi_e(a|x)p(r|x,a)}[r]$$

value of eval policy
(our "estimand")

where,

$$(x, a, r) \sim p(x)\pi_0(a|x)p(r|x, a)$$

unknown

known

unknown

- Action embeddings in Logged Bandit dataset in set "D", where
- MSE:

$$\begin{aligned}\text{MSE}(\hat{V}; \pi_e) &:= \mathbb{E}_{\mathcal{D}} \left[\left(V(\pi_e) - \hat{V}(\pi_e; \mathcal{D}) \right)^2 \right] \\ &= \text{Bias}(\hat{V}(\pi_e; \mathcal{D}))^2 + \mathbb{V}_{\mathcal{D}}[\hat{V}(\pi_e; \mathcal{D})]\end{aligned}$$

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

- Bias: refers to the deviation from the true value that can occur when estimating the value of an action or state. In the paper, if the action has no effect on the reward, MIPS is regarded to be Unbiased (Assumption 3.2).

Inverse Propensity Score (IPS) Estimator

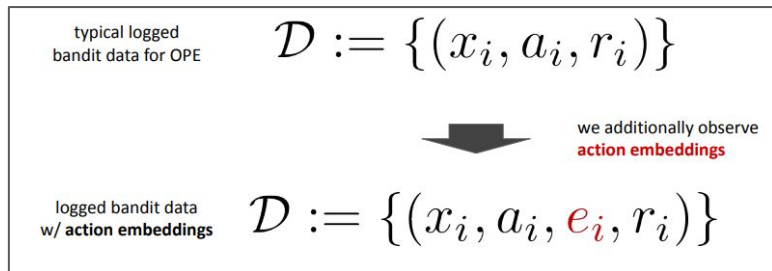
$$\hat{V}_{\text{IPS}}(\pi_e; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{\pi_e(a_i|x_i)}{\pi_0(a_i|x_i)}}_{w(x_i, a_i)} \cdot r_i$$

(vanilla)
importance weight

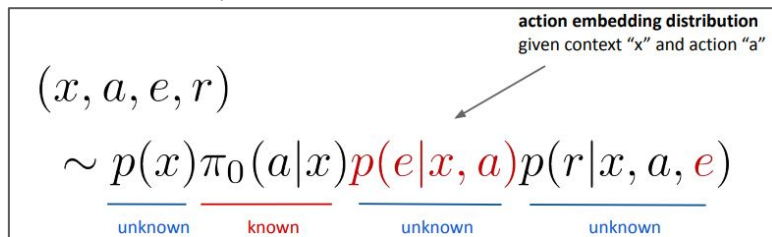
where, Vanilla Importance weights of IPS is

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

New Embeddings and Assumptions



And now,



Assumption 3.1. (Common Embedding Support) The logging policy π_0 is said to have common embedding support for policy π if $p(e|x, \pi) > 0 \rightarrow p(e|x, \pi_0) > 0$ for all $e \in \mathcal{E}$ and $x \in \mathcal{X}$, where $p(e|x, \pi) := \sum_{a \in \mathcal{A}} p(e|x, a)\pi(a|x)$ is the *marginal* distribution over the action embedding space given context x and policy π .

Assumption 3.2. (No Direct Effect) Action a has no direct effect on the reward r , i.e., $a \perp r \mid x, e$.

So the policy becomes:

$$\begin{aligned}
 V(\pi_e) &:= \mathbb{E}_{p(x)\pi_e(a|x)p(r|x, a)}[r] \\
 &= \mathbb{E}_{p(x)p(e|x, \pi_e)p(r|x, e)}[r]
 \end{aligned}$$

where $p(e|x, \pi) = \sum_{a \in \mathcal{A}} p(e|x, a)\pi(a|x)$ is the *marginal distribution of action embeddings* induced by a particular policy π

Marginalized IPS Estimator

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

Marginalized IPS (MIPS)

where, Vanilla Importance weights of IPS is

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

and, new Importance weights for MIPS is

$$w(x, e) = \mathbb{E}_{\pi_0(a|x, e)} [w(x, a)] .$$

Variance, Bias and MSE Calculations

Variance:

$$\begin{aligned} & n \left(\mathbb{V}_{\mathcal{D}}[\hat{V}_{\text{IPS}}(\pi; \mathcal{D})] - \mathbb{V}_{\mathcal{D}}[\hat{V}_{\text{MIPS}}(\pi; \mathcal{D})] \right) \\ &= \mathbb{E}_{p(x)p(e|x, \pi_0)} \left[\mathbb{E}_{p(r|x, e)} [r^2] \mathbb{V}_{\pi_0(a|x, e)} [w(x, a)] \right], \end{aligned}$$

Bias:

Theorem 3.5. (*Bias of MIPS*) If Assumption 3.1 is true, but Assumption 3.2 is violated, MIPS has the following bias.

$$\begin{aligned} & \text{Bias}(\hat{V}_{\text{MIPS}}(\pi)) \\ &= \mathbb{E}_{p(x)p(e|x, \pi_0)} \left[\sum_{a < b} \pi_0(a|x, e) \pi_0(b|x, e) \right. \\ & \quad \times (q(x, a, e) - q(x, b, e)) \\ & \quad \times (w(x, b) - w(x, a)) \left. \right], \end{aligned}$$


MSE:

$$\begin{aligned} & n \left(\text{MSE}(\hat{V}_{\text{IPS}}(\pi)) - \text{MSE}(\hat{V}_{\text{MIPS}}(\pi)) \right) \\ &= \mathbb{E}_{x, a, e \sim \pi_0} \left[(w(x, a)^2 - w(x, e)^2) \cdot \mathbb{E}_{p(r|x, a, e)} [r^2] \right] \\ & \quad + 2V(\pi) \text{Bias}(\hat{V}_{\text{MIPS}}(\pi)) + (1 - n) \text{Bias}(\hat{V}_{\text{MIPS}}(\pi))^2. \end{aligned}$$

Data-Driven Embedding Selection

- We want to identify a set of action embeddings/features that minimizes the MSE of the resulting MIPS
- The problem is that estimating the bias is equally difficult as OPE itself
- So, we adjust “SLOPE” to our setup. SLOPE is originally developed to tune hyperparameters of OPE and does not need to estimate the bias of the estimator

$$\min_{\mathcal{E} \subseteq \mathcal{V}} \text{Bias} \left(\hat{V}_{\text{MIPS}} (\pi_e; \mathcal{E}) \right)^2 + \mathbb{V}_{\mathcal{D}} \left[\hat{V}_{\text{MIPS}} (\pi_e; \mathcal{D}, \mathcal{E}) \right]$$


$$\text{Bias} \left(\hat{V}_{\text{MIPS}} (\pi; \mathcal{E}) \right) = \mathbb{E}_{\mathcal{D}} \left[\hat{V}_{\text{MIPS}} (\pi; \mathcal{D}, \mathcal{E}) \right] - \underline{V(\pi)}$$

depends on the true policy value

Estimating the Marginal Importance Weights

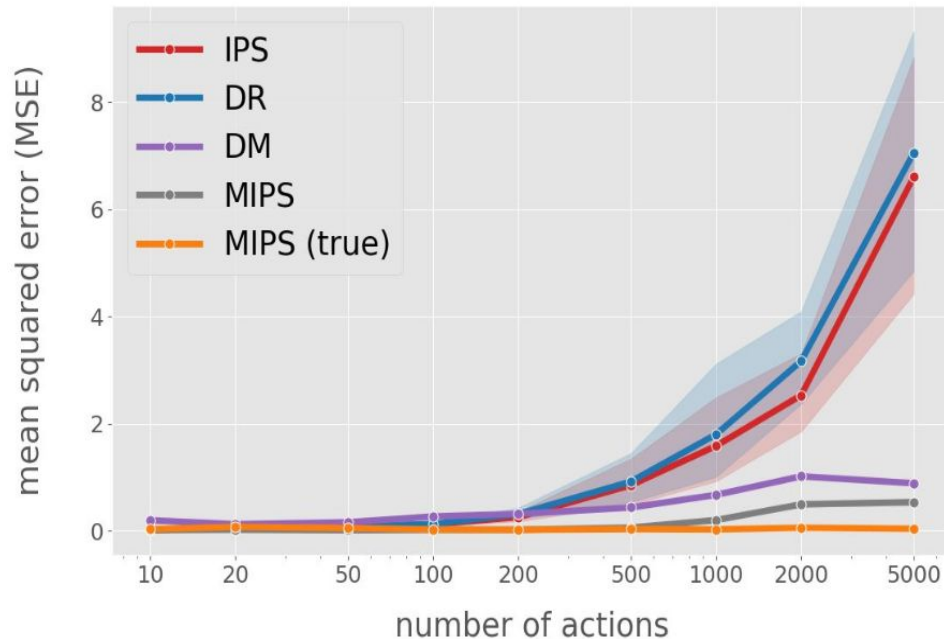
- Even if we know the logging policy, we may have to estimate because we do not know the true distribution
- A simple procedure is to utilize the following transformation

$$\underline{w(x, e) = \mathbb{E}_{\pi_0(a|x, e)}[w(x, a)]}$$

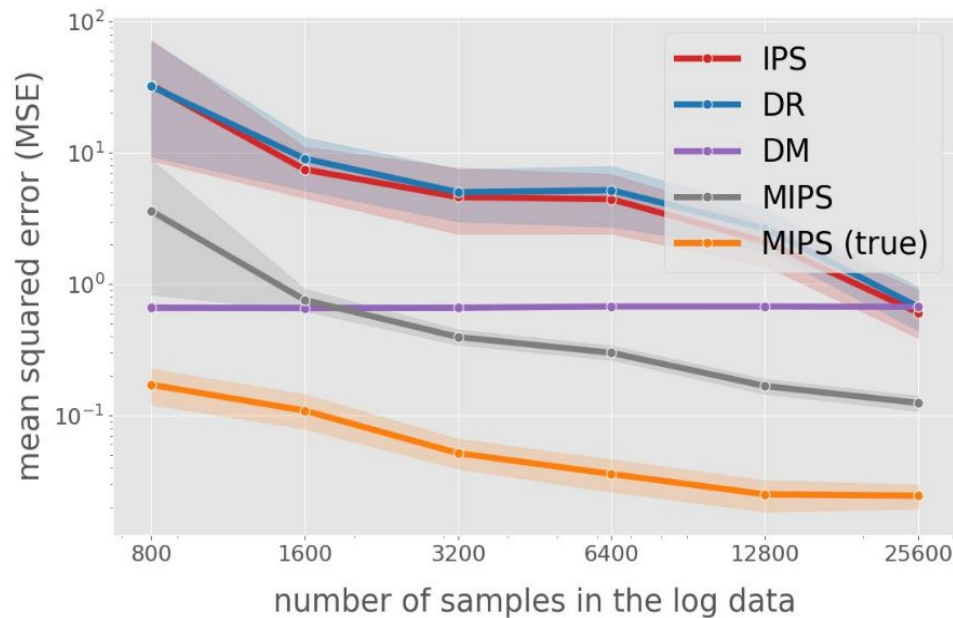
So, our task is to estimate $\pi_0(a|x, e) \approx \hat{\pi}_0(a|x, e)$

and use it to compute $\hat{w}(x, e) = \mathbb{E}_{\hat{\pi}_0(a|x, e)}[w(x, a)]$

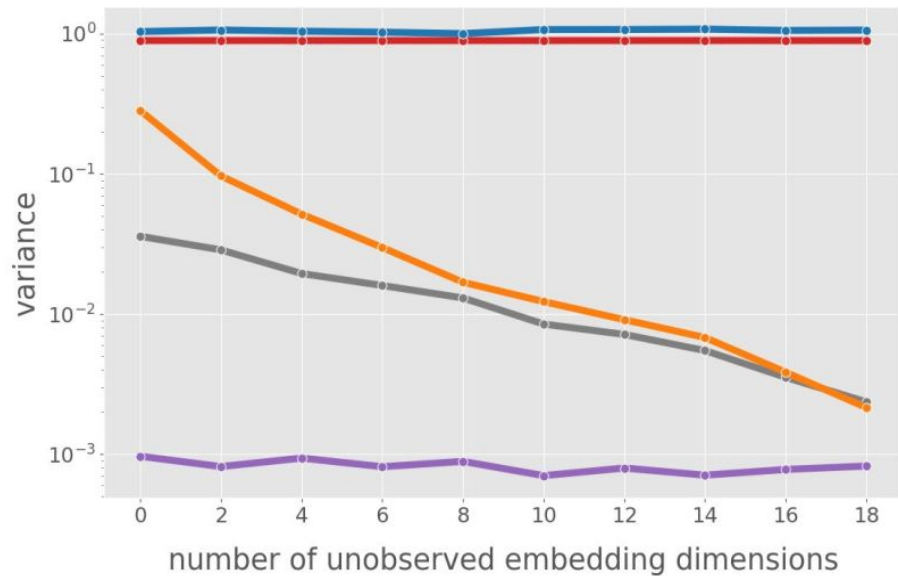
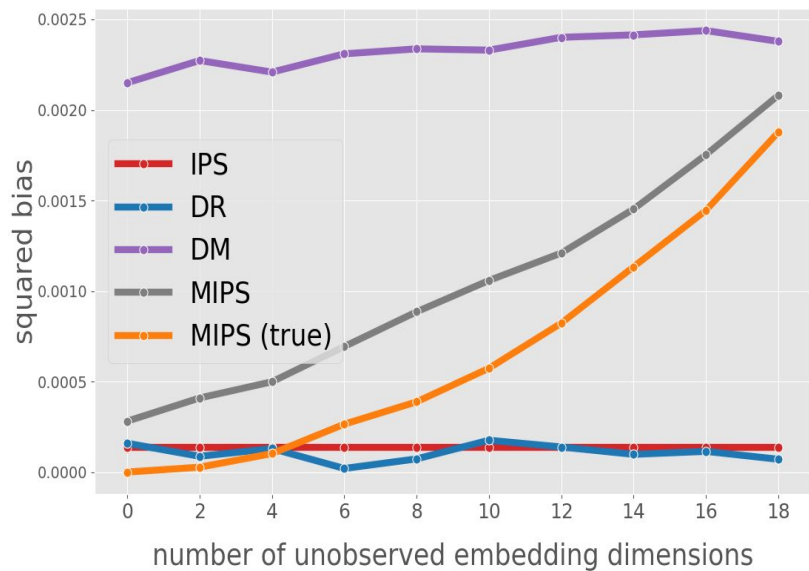
Empirical Evaluation on Synthetic Data



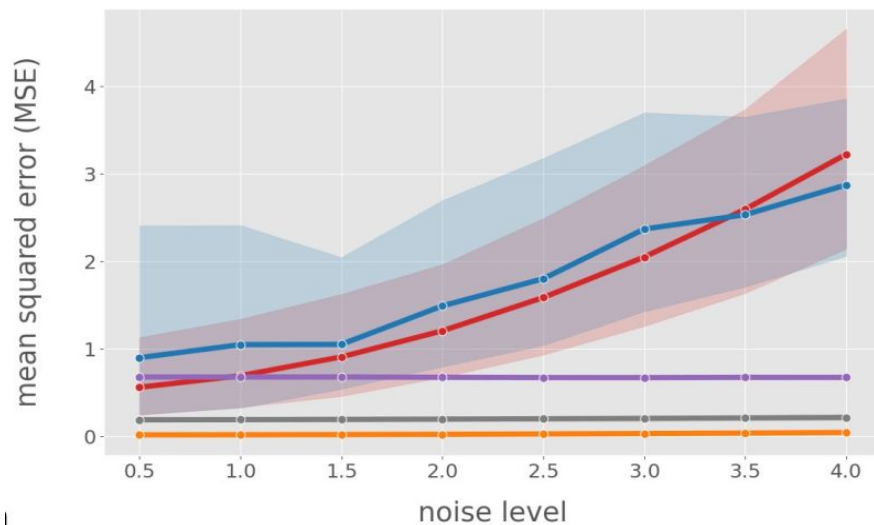
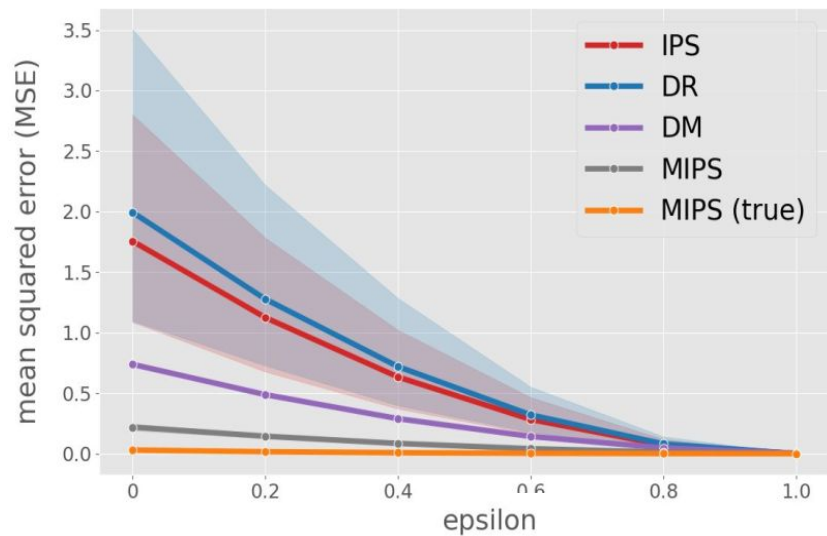
- MIPS is More Robust to Increasing Number of Actions
- Achieving Large Variance Reduction



- MIPS Makes Full Use of Increasing Data
- MIPS works like IPS/DR, and is much better in small sample size
- MIPS is also increasingly better than DM for larger sample sizes



missing dimensions introduce some bias but, it also reduces the variance.

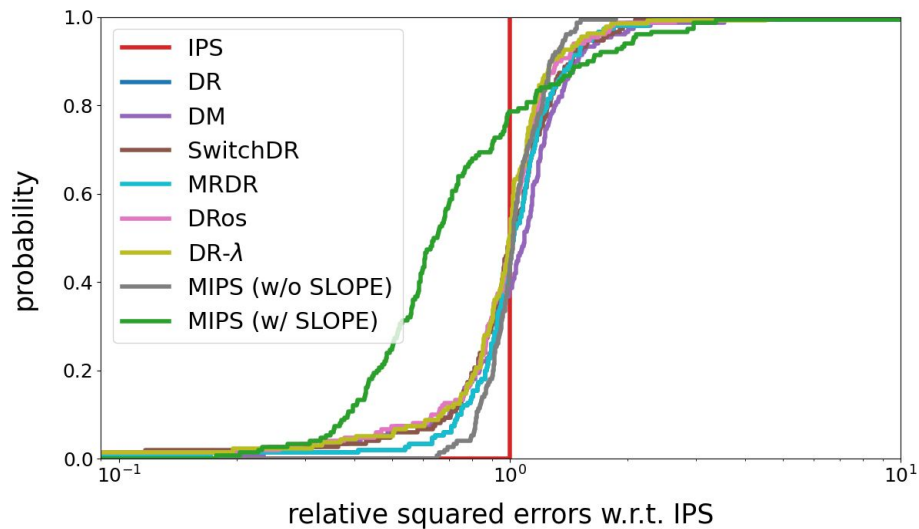
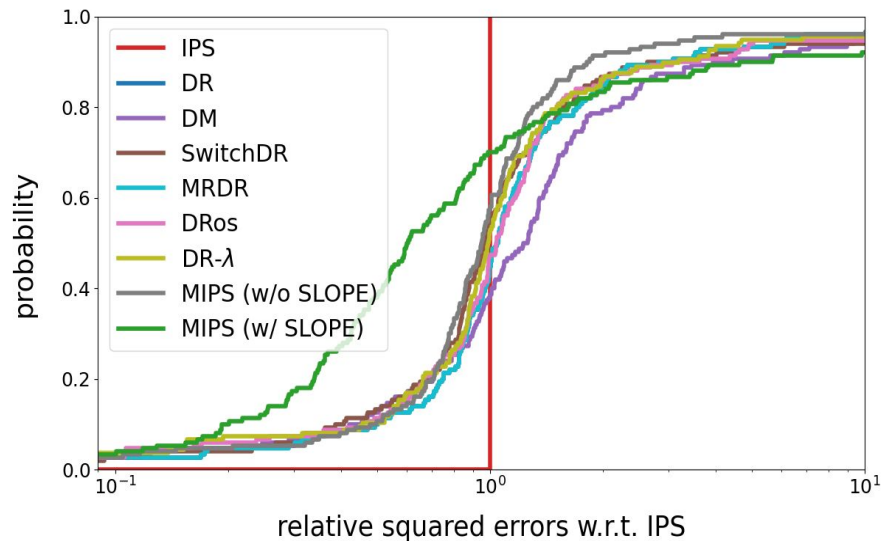


MIPS is robust to deterministic evaluation policies and noisy reward.

Results from Graphs

- With growing number of actions, MIPS provides a large variance reduction (works even better than DM), while the MSEs of IPS&DR inflate.
- With growing sample size, MIPS works like an unbiased/consistent estimator (similarly to IPS/DR), while DM remains highly biased.
- Even if we violate the assumption and introduce some bias, intentionally dropping some embedding dimensions might provide a greater MSE gain.

Empirical Evaluation on Real-World Data



Observations

- **Bias:** MIPS is unbiased if action has no effect on reward (Assumption 3.2). If this assumption is violated, MIPS produces less bias as compared to IPS.
- **Variance:** MIPS has a lower variance than IPS, especially when there is a large number of actions.
- **MSE:** MIPS has a significant gain in MSE when run in large action spaces (gain is good). Thus outperforming IPS and related estimators.

Experiments on synthetic and real-world bandit data verify the theoretical findings, indicating that MIPS can provide an effective bias-variance trade-off in the presence of many actions.

Conclusion and Future Work

- We explored the problem of OPE for large action spaces. In this, existing estimators based on IPS suffer from high variance, which limits its usability.
- As a solution, we used the MIPS estimator, which builds on the marginal importance weights computed with action embeddings.
- As a result, the Variance decreased and showed significant gain in MSE, thus outperforming IPS and other related estimators.

- Developing a method for accurately estimating the marginal importance weight would also be crucial to fill the gap between MIPS and MIPS (true) observed in our experiment.

Thank You