RL Paper Presentation

Off-Policy Evaluation for Large Action Spaces via Embeddings

Sahas Marwah 2020237
Yash Thakran
Sparsh Mehrotra 2020248

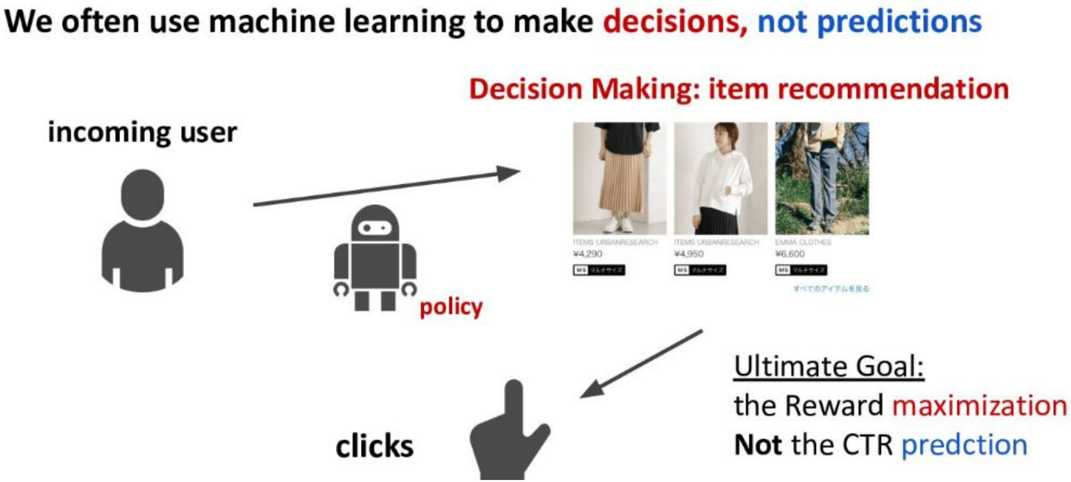
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Introduction

We are talking about ‘Off-Policy Evaluation for Large Action Spaces via Embeddings’.

In this paper the author starts by showing the existing models that exist fail in the case when action space is large. These cases include recommendations systems for netflix, spotify, etc.

In this cases there are millions of movies, and songs to choose from.



We often use machine learning to make decisions, not predictions

Decision Making: item recommendation

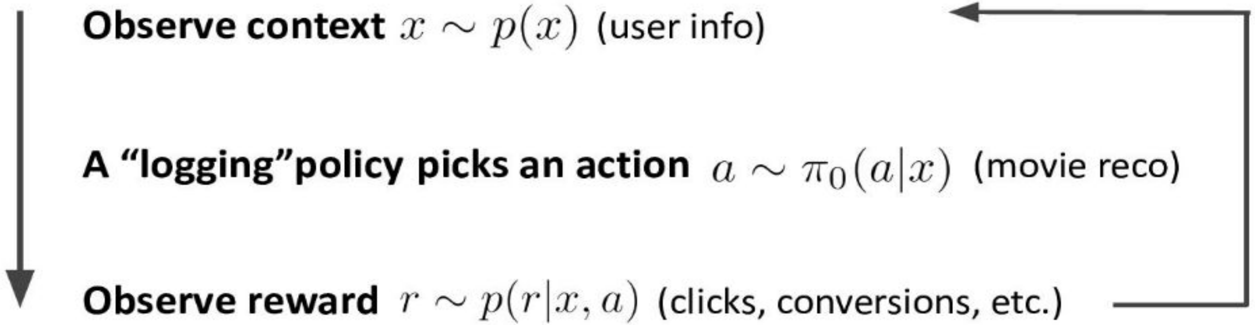
Ultimate Goal:
the Reward maximization
Not the CTR predction

OPE setting - Off policy evaluation

The paper starts by using estimators to make the predictions, in off-policy setting.

Motivation for OPE:

How can we evaluate the performance of a new decision making policy using only data collected by logging, past policy?



Observe context $x \sim p(x)$ (user info)

A "logging"policy picks an action $a \sim \pi_0(a|x)$ (movie reco)

Observe reward $r \sim p(r|x,a)$ (clicks, conversions, etc.)

the *logging policy* interacts with the environment

and produces the *log data*: $\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^n$

Off-Policy Evaluation: Logged Bandit Data

We are given ***logged bandit data*** collected by ***logging policy*** π_0

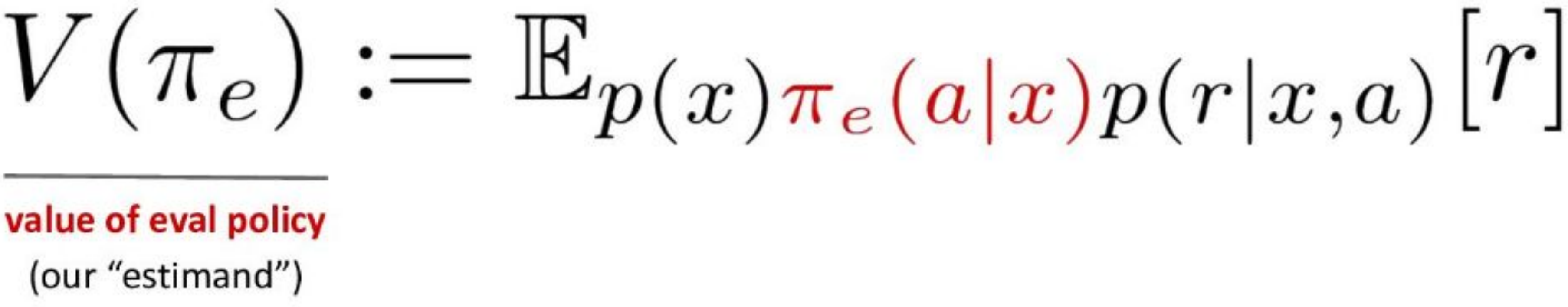
$$\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^n$$

where

$$(x, a, r) \sim \underbrace{p(x)}_{\text{unknown}}\underbrace{\pi_0(a|x)}_{\text{known}}\underbrace{p(r|x,a)}_{\text{unknown}}$$

Off policy

Now let the evaluation function be:



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

value of eval policy

(our “estimand”)

And the evaluation policy is (pi)e

Mean Square error.

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

$$\mathrm{Bias}(\hat{V}(\pi)) := \mathbb{E}_{\mathcal{D}}[\hat{V}(\pi;\mathcal{D})] - V(\pi),$$

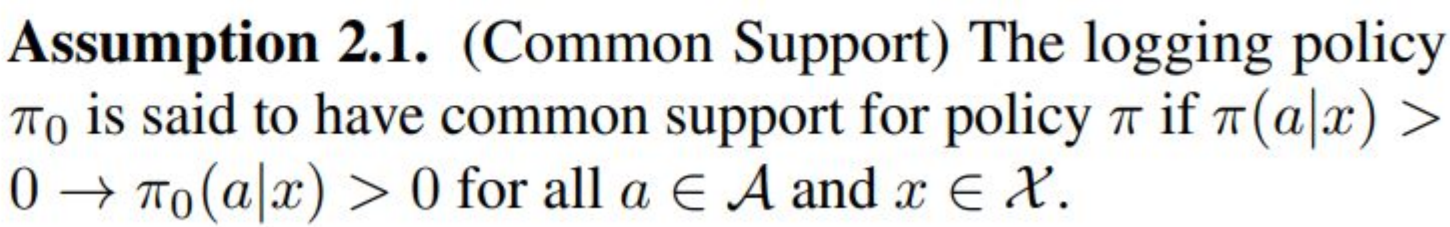
$$\mathbb{V}_{\mathcal{D}}\left[\hat{V}(\pi;\mathcal{D})\right] := \mathbb{E}_{\mathcal{D}}\left[\left(\hat{V}(\pi;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\hat{V}(\pi;\mathcal{D})]\right)^2\right].$$

Estimators IPS

We started by using IPS (inverse propensity score), and the formula for that is:

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

(vanilla)
importance weight



Assumption 2.1. (Common Support) The logging policy π_0 is said to have common support for policy π if $\pi(a|x) > 0 \rightarrow \pi_0(a|x) > 0$ for all $a \in \mathcal{A}$ and $x \in \mathcal{X}$.

This is same to regular offline evaluation, where we say that b>0 for all values where a>0.

Drawbacks of IPS and other filters

$$
\begin{aligned}
n\mathbb{V}_{\mathcal{D}}\left[\hat{V}_{\mathrm{IPS}}(\pi;\mathcal{D})\right] &= \mathbb{E}_{p(x)\pi_0(a|x)}[w(x,a)^2\sigma^2(x,a)] \\
&+ \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_0(a|x)}[w(x,a)q(x,a)]\right] \\
&+ \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_0(a|x)}[w(x,a)q(x,a)]\right],
\end{aligned} \tag{2}
$$

Here it is made by three terms

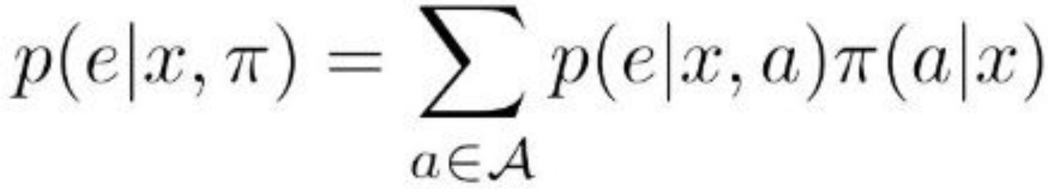
- The first term reflects the randomness in the rewards.
- The second term represents the variance due to the randomness over the contexts.
- The final term is the penalty arising from the use of IPS weighting, and it is proportional to the weights and the true expected reward

MIPS Intro

Action Embeddings: is the additional information that we extract from the action space. This is prior information stored in logged bandit dataset “D”. The intuition is that this can help the estimator transfer information between similar actions.

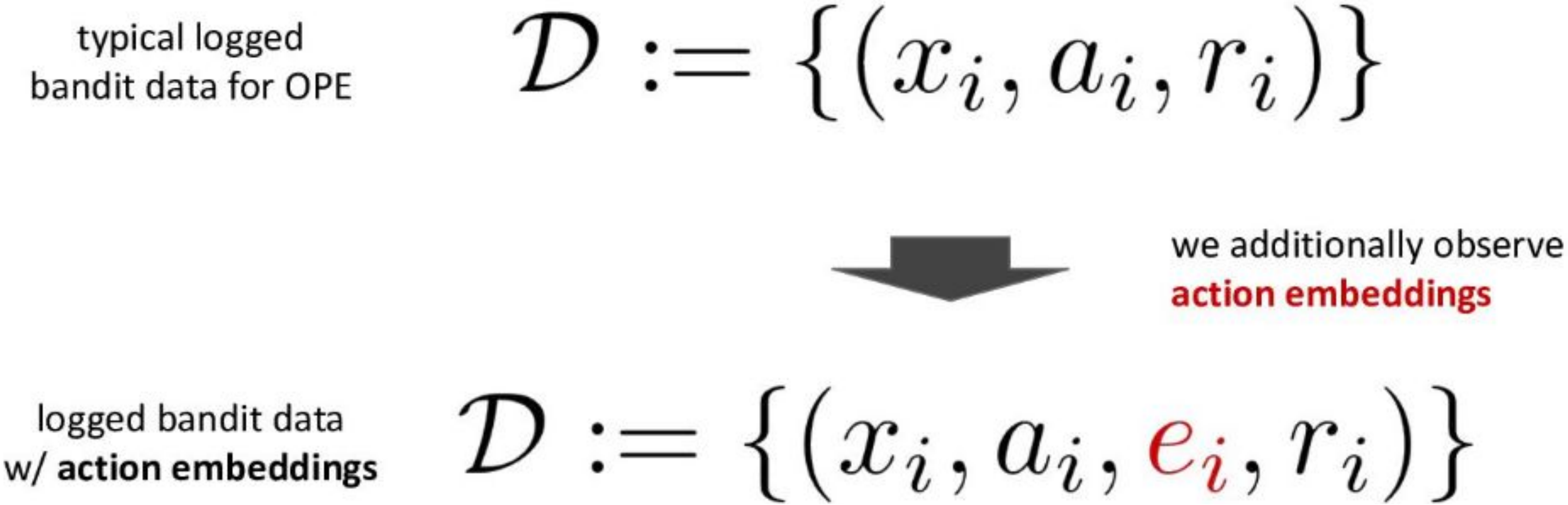
MIPS Estimator: uses the marginal distribution of action embeddings, rather than actual actions, to define a new type of importance weights.

Say we have a policy pi. The Marginal Distribution by the paper is given as:



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

Embeddings



typical logged
bandit data for OPE

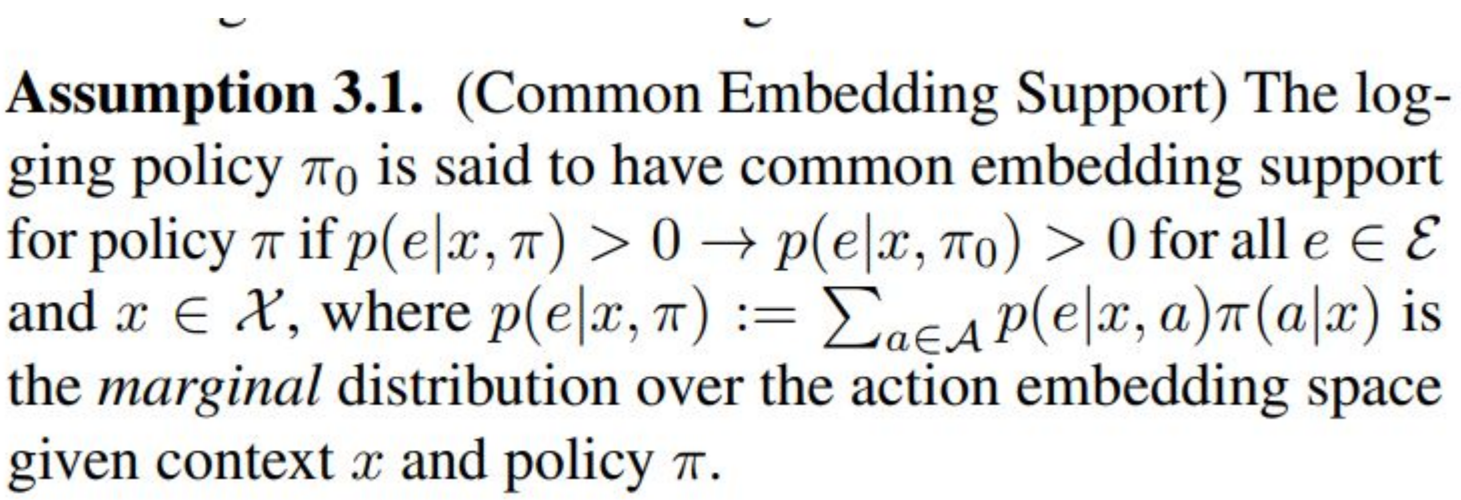
$$\mathcal{D} := \{(x_i, a_i, r_i)\}$$

we additionally observe
action embeddings

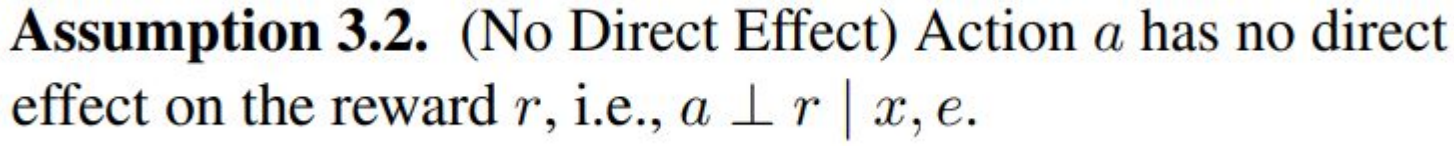
logged bandit data
w/ **action embeddings**

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

New assumptions



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., $a \perp r \mid x, e$.

MIPS (marginalized IPS)

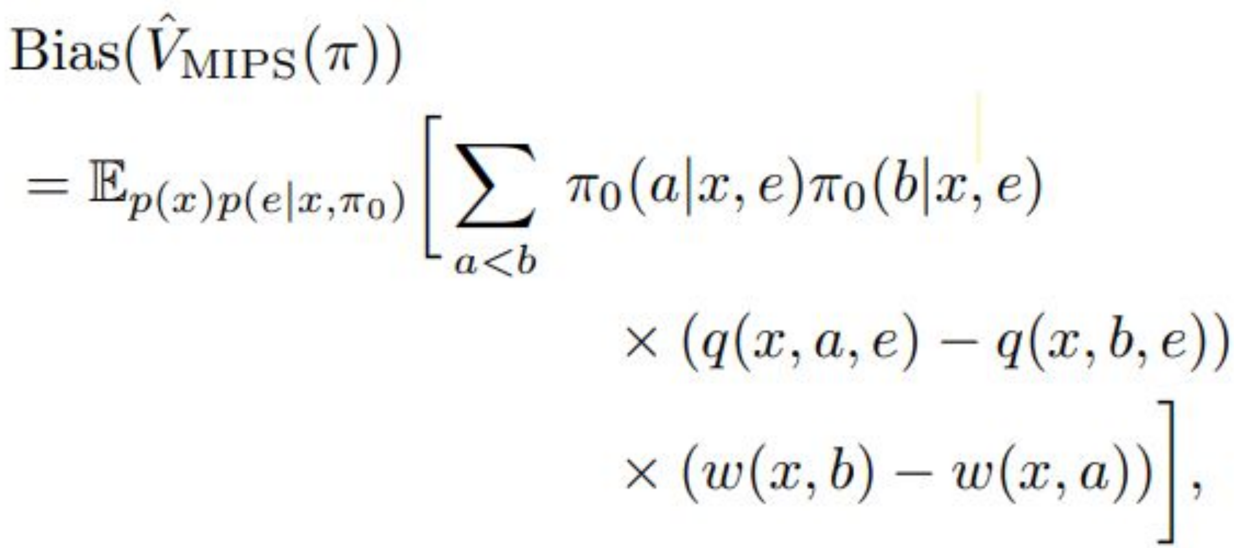
The paper defines the MIPS Policy Value as:

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

Marginalized IPS (MIPS)

Bias

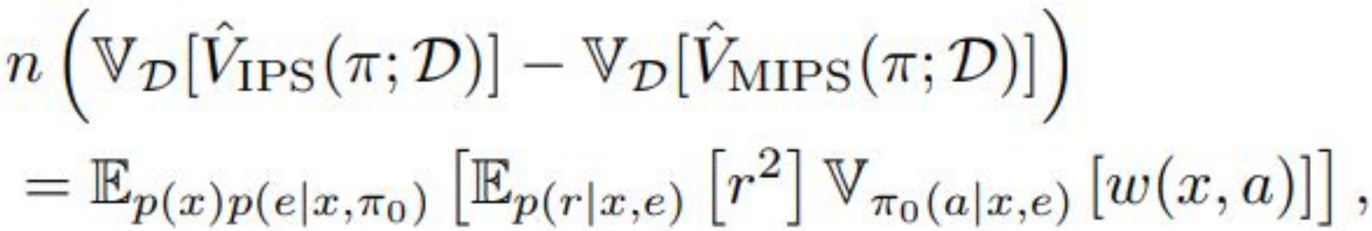
The paper defines Bias as the violations of assumptions taken.



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

Variance

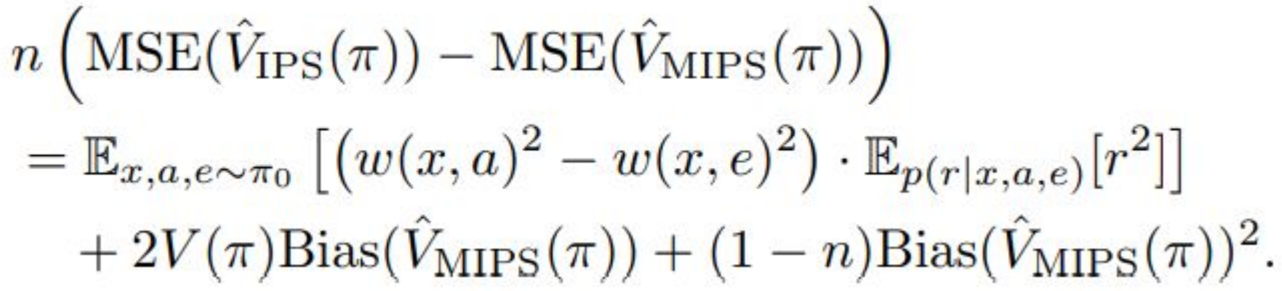
The paper has compared variance of MIPS and IPS and set a value which must be greater than or equal to zero in order to have a variance reduction.



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

MSE

MSE Gain is defined as the following equation, in case of violation of any assumption.



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

Dilemma

There is a Bias and Variance. Bias reduces when Action Embeddings are informative. On the other hand, Variance reduces when AE are not informative.

A balance of Bias can improve MSE and have high variance reduction.

The paper has proposed an Action Embedding Selector method called SLOPE.