BIS 631: Advanced Techniques in Causal Inference Due: May 9th, 2023

Final Project -Stochastic Intervention with Large Action Space-

Lecturer: Laura Forastiere *Name: Tatsuhiro Shimizu*

1 Introduction

In our lecture, we focused on the irregular setting of the Causal Inference. Primarily, we covered the cases where the assignment mechanism does not satisfy the strong ignorability assumption. In such settings, we mostly focused on deterministic intervention. Nevertheless, theoretically, it is worthwhile to study the stochastic intervention since it is more general than a deterministic one. Furthermore, practically, from the perspective of Reinforcement Learning or sequential decision-making, it would be risky if we only choose the deterministic intervention which maximizes some outcomes of our interest since the variance of the estimate could be large. Therefore, in this final project, I extend the intervention class from deterministic to stochastic. Furthermore, I consider the irregular setting where the number of possible interventions is large and most of the existing estimators like Direct Method (DM) Beygelzimer and Langford [2009], Inverse Propensity Score (IPS) Horvitz and Thompson [1952], and Doubly Robust (DR) Dudik et al. [2011] do not work well due to the large variance. In such a setting, Marginalized Inverse Propensity Score (MIPS) Saito and Joachims [2022] works well. Combining the good properties of MIPS and DR, I will propose Marginalized Doubly Robust (MDR) with good theoretical and empirical guarantees. The Python code for the simulation can be found in the appendix A.

2 Setting of the Problem

Since the stochastic intervention case is studied in the contextual bandit setting in Off Policy Evaluation (OPE) and Off Policy Learning (OPL), we introduce new notation sometimes but I try to use the notation used in the class as much as possible.

2.1 Data

The data we consider is the contextual vector $x \in \mathcal{X} = \mathbb{R}^{d_x}$, which is like covariates, action $a \in \mathcal{A}$, and outcome $y \in [0, y_{\max}]$. We assume that the context vector is sampled from the unknown distribution $x \sim p(x)$, action a is chosen given contextual vector x by the stochastic intervention called policy $a \sim \pi(a|x)$ where $\pi : \mathcal{X} \to \Delta(\mathcal{A})$, and outcome y is sampled from the unknown distribution $y \sim p(y|x, a)$ given contextual vector x and action a. We observe independent and identically distributed n samples collected by the already implemented stochastic intervention called behavior policy π_b Strehl et al. [2011], Langford et al. [2008]. Thus, the actually observed data D is given by $D = \{(x_i, a_i, y_i)\}_{i=1}^n$ where $x_i \sim p(x_i), a_i \sim \pi_b(a_i|x_i), y_i \sim p(y_i|x_i, a_i) \quad \forall i \in [n]$.

2.2 Problem

If we know the distribution of contextual vector $p(x)$ and outcome $p(y|x,a)$, then we can define value function $V(\pi)$ of π, which means how good the policy π is by

$$V(\pi) := \mathbb{E}_{p(x)\pi(a|x)p(y|x,a)}[y] = \mathbb{E}_{p(x)\pi(a|x)}[q(x,a)] \tag{1}$$

where $q(x,a) = \mathbb{E}_{p(y|x,a)}[y|x,a]$ is the expected outcome given contextual vector x and action a. Then, we can choose the policy π which maximizes the value function $V(\pi)$. This is called Off Policy Learning (OPL). However, as we assumed that we do not know the distribution of contextual vector $p(x)$ and outcome $p(y|x,a)$, which is a realistic assumption, we need to estimate the value function so that we can choose the best policy π. We define the evaluation policy π_e whose value function should be estimated to distinguish it from the behavior policy π_b. Then, the problem of our interest is how to construct the estimator $\widehat{V}(\pi_e; D) \approx V(\pi)$ where we use the Mean Squared Error (MSE)

$$\mathrm{MSE}(\widehat{V}(\pi_e)) = \mathbb{E}_D\left[(V(\pi_e) - \widehat{V}(\pi_e; D))^2\right] \tag{2}$$

as the quantity to measure how good the estimator is.

3 Literature Review

There are three classical estimators to estimate the value function $V(\pi_e)$: Direct Method (DM), Inverse Propensity Score (IPS), and Doubly Robust (DR). I introduce them with drawback when the cardinality of the action space is large $|\mathcal{A}| \gg 1$. In such a setting, the Marginalized Inverse Propensity Score (MIPS) plays a significant role.

3.1 Direct Method (DM)

Definition 1 (Direct Method (DM))**.** Direct Method Beygelzimer and Langford [2009] $\widehat{V}_{\mathrm{DM}}(\pi_e; D, \widehat{q})$ is given by

$$\widehat{V}_{\mathrm{DM}}(\pi_e; D, \widehat{q}) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i, a)] \tag{3}$$

where $\widehat{q}(x,a)$ is the estimated expected outcome given contextual vector x and a as follows.

$$\widehat{q} = \operatorname*{argmin}_{q' \in \mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n} \left(y_i - q'(x_i, a_i)\right)^2 \tag{4}$$

Having the definition of DM, we have the following proposition about the unbiasedness of DM Dudik et al. [2011] under the unbiasedness of $\widehat{q}(x,a)$.

Assumption 2 (Unbiasedness of Estimated Outcome given x, a)**.**

$$\widehat{q}(x,a) = q(x,a) \quad \forall x \in \mathcal{X}, a \in \mathcal{A} \tag{5}$$

Proposition 3 (Unbiasedness of DM)**.** DM 1 is unbiased under the unbiasedness of the estimated expected outcome 2.

Having the proposition and definition of DM, we can say that DM works well when the estimated expected outcome predicts the actually expected outcome function well.

3.2 Inverse Propensity Score (IPS)

We have another estimator called IPS, which is similar to the IPS we used in our class.

Definition 4 (Inverse Propensity Score (IPS))**.** IPS Horvitz and Thompson [1952] $\widehat{V}_{\mathrm{IPS}}(\pi_e; D)$ is given by

$$\widehat{V}_{\mathrm{IPS}}(\pi_e; D) := \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i) y_i \tag{6}$$

where importance weight $w(x_i, a_i)$ is defined by

$$w(x_i, a_i) := \frac{\pi_e(a_i|x_i)}{\pi_b(a_i|x_i)} \tag{7}$$

IPS is unbiased under the common support Saito and Joachims [2022] defined as follows.

Assumption 5 (Common Support)**.**

$$\pi_e(a|x) > 0 \implies \pi_b(a|x) > 0 \quad \forall x \in \mathcal{X}, a \in \mathcal{A} \tag{8}$$

Proposition 6 (Unbiasedness of IPS)**.** IPS is unbiased under the common support 5.

Furthermore, we have the proposition on the variance of IPS Saito and Joachims [2022] as follows.

Proposition 7 (Variance of IPS)**.** The variance of IPS is given by

$$\begin{aligned} &n\mathbb{V}_D\left[\widehat{V}_{\mathrm{IPS}}(\pi_e; D)\right] \\ &= \mathbb{E}_{p(x)\pi_b(a|x)}[w(x,a)^2\sigma(x,a)^2] + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_b(a|x)}[w(x,a)q(x,a)]\right] + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_b(a|x)}[w(x,a)q(x,a)]\right] \end{aligned}$$

where $\sigma(x,a) := \mathbb{V}_{p(r|x,a)}[y|x,a]$

Having the proposition of the variance of IPS, we can say that when the cardinality of the action space is large $|\mathcal{A}| \gg 1$, the importance weight $w(x,a)$ has a wide range, resulting in the large variance by the first and third terms.

3.3 Doubly Robust (DR)

Dudik et al. [2011] combine the good properties of both DM and IPS and they proposed DR which was proposed by Cassel et al. [1976]. DR is defined as follows.

Definition 8 (Doubly Robust (DR))**.**

$$\widehat{V}_{\mathrm{DR}}(\pi_e; D, \widehat{q}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i, a)] + w(x_i, a_i)\left(y_i - \widehat{q}(x_i, a_i)\right)\right\} \tag{9}$$

DR guarantees the unbiasedness under either the unbiasedness of the estimated expected outcome 2 or the common support 5.

Proposition 9 (Unbiasedness of DR)**.** DR is unbiased under either the unbiasedness of the estimated expected outcome 2 or the common support 5.

Furthermore, we have the proposition on the variance of DR Huang et al. [2021] as follows.

Proposition 10 (Variance of DR)**.** The variance of IPS is given by

$$\begin{aligned}
&n\mathbb{V}_D\left[\widehat{V}_{\mathrm{DR}}(\pi_e; D, \widehat{q})\right] \\
&= \mathbb{E}_{p(x)\pi_b(a|x)}[w(x,a)^2\sigma(x,a)^2] + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_b(a|x)}[w(x,a)q(x,a)]\right] + \mathbb{E}_{p(x)}\left[\mathbb{V}_{\pi_b(a|x)}[w(x,a)\Delta(x,a)]\right]
\end{aligned}$$

where $\Delta(x,a) := q(x,a) - \widehat{q}(x,a)$ is the prediction error of the expected outcome.

The variance of DR is the same as the one of IPS except for the third term, which is often lower than IPS. When the cardinality of the action space is large $|\mathcal{A}| \gg 1$, the variance of DR gets large as well due to the first and third terms though the deterioration of the estimation is mitigated by the reduction of the third term with compared to IPS.

3.4 Marginalized Inverse Propensity Score (MIPS)

To tackle the problem of the large bias and variance of existing estimators when there are lots of possible interventions to be considered, Saito and Joachims [2022] proposed Marginalized Inverse Propensity Score (MIPS). Instead of using the importance weight $w(x,a)$ used in IPS, MIPS uses the marginal importance weight $w(x,e)$ where $e \in \mathcal{E} \subset \mathbb{R}^{d_e}$ is the embedding of the action. For instance, if we want to construct the best movie recommendation system where policy π is the stochastic recommendation, action a is the movie we recommend to the customers where there are lots of movies we can recommend $|\mathcal{A}| \gg 1$, we find the embedding e including the genre of the movies like comedy, love story, action movie, directors of the movies, actors of the movies, which categorize the movie. If the embedding characterizes the movie very well, then, we can reduce the cardinality of embedding space $|\mathcal{E}|$. Therefore, using the embedding for the marginal importance weight improves the variance of the MIPS. Since we use the action embedding, for the MIPS, we need to define the new data-generating process, and value function as follows.

3.4.1 New Data generating process

We assume that the context vector x is sampled from the unknown distribution $x \sim p(x)$ as we assumed before, action a is chosen by the stochastic intervention $a \sim \pi(a|x)$, action embedding e is sampled from the unknown distribution $e \sim p(e|x,a)$ given x and a, and the outcome is sampled from the unknown distribution $y \sim p(y|x,a,e)$ given x, a, and e. Thus, the actually observed data D is given by $D = \{(x_i, a_i, e_i, y_i)\}_{i=1}^n$ where $x_i \sim p(x_i), a_i \sim \pi_b(a_i|x_i), e_i \sim p(e_i|x_i,a_i), y_i \sim p(y_i|x_i,a_i,e_i) \quad \forall i \in [n]$ i.i.d. Saito and Joachims [2022]. Furthermore, we define the value function $V(\pi)$ as follows.

$$V(\pi) := \mathbb{E}_{p(x)\pi(a|x)p(e|x,a)p(y|x,a,e)}[y] = \mathbb{E}_{p(x)\pi(a|x)p(e|x,a)}[q(x,a,e)] = \mathbb{E}_{p(x)\pi(a|x)}[q(x,a)] \tag{10}$$

where $q(x,a,e) := \mathbb{E}_{p(y|x,a,e)}[y|x,a,e]$ is the expected outcome function given x, a, and e and $q(x,a) := \mathbb{E}_{p(e|x,a)}[q(x,a,e)]$ is the expected outcome function given x and a.

3.4.2 definition and properties of MIPS

Having the new data generating process and the definition of the value function, Saito and Joachims [2022] propose MIPS as follows.

Definition 11 (Marginalized Inverse Propensity Score (MIPS))**.**

$$\widehat{V}_{\mathrm{MIPS}}(\pi_e; D) := \frac{1}{n}\sum_{i=1}^{n} w(x_i, e_i) y_i \tag{11}$$

where $w(x_i, e_i) := \frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}$ is the marginal importance weight and $p(e|x,\pi) := \sum_{a\in\mathcal{A}} \pi(a|x)p(e|x,a)$ is the marginal distribution of e given the context vector x and policy π.

MIPS is unbiased under two assumptions as follows.

Assumption 12 (No Direct Effect of a on y)**.** Given context x and action embedding e, action a and outcome y are independent ($a \perp\!\!\!\perp y|x,e$)

No direct effect assumption means that the effect of action a on the outcome y is fully mediated by the action embedding e.

Assumption 13 (Common Embedding Support)**.**

$$P(e|x,\pi_e) \implies P(e|x,\pi_b) \quad \forall x \in \mathcal{X}, e \in \mathcal{E} \tag{12}$$

Note that common embedding support 13 is a weaker assumption than the common support 5 which is necessary for the unbiasedness of IPS.

Proposition 14 (MIPS is unbiased)**.** MIPS is unbiased under 1. no direct effect 12 and 2. common embedding support 13.

4 New estimator Marginalized Doubly Robust (MDR)

Even though MIPS has good properties in that it can estimate the value function without having a large variance, which was the problem of the existing estimators: DM, IPS, and DR, it would be possible to construct a better estimator by combining MIPS and DR. Using the good properties of MIPS and DR, I propose Marginalized Doubly Robust (MDR)

4.1 Definition of MDR

Definition 15 (Marginalized Doubly Robust (MDR))**.**

$$\widehat{V}_{\mathrm{MDR}}(\pi_e; D, \widehat{q}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i, a)] + w(x_i, e_i)\left(y_i - \widehat{q}(x_i, a_i, e_i)\right)\right\} \tag{13}$$

4.2 Theoretical Analysis

Theoretically, MDR has a doubly robust unbiasedness under either common embedding support or the following assumption about the precision of the prediction of the outcome function.

Assumption 16 (Unbiasedness of Estimated Outcome given x, a, and e)**.**

$$\widehat{q}(x, a, e) = q(x, a, e) \quad \forall x \in \mathcal{X}, a \in \mathcal{A}, e \in \mathcal{E} \tag{14}$$

4.2.1 Unbiasedness

Proposition 17 (MDR is unbiased)**.** MDR is unbiased under 1. no direct effect 12 and 2. either common embedding support 13 or the unbiasedness of the estimated expected outcome given x, a, and e 16.

Proof. See Appendix B. □

4.2.2 Variance Reduction

For the variance, as we know the variance of DR 10, we only focus on the reduction of the variance when we compare DR and MDR as follows.

Proposition 18 (Variance Reduction of MDR)**.** Under 1. no direct effect 12, 2. common embedding support 13, 3. and common support 5. The difference between the variances of DR and MDR is

$$n\left(\mathbb{V}_D\left[\widehat{V}_{\mathrm{DR}}(\pi_e; D, \widehat{q})\right] - \mathbb{V}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e; D, \widehat{q})\right]\right) \tag{15}$$

$$= \mathbb{E}_{\bar{d}_{\pi_b}}\left[w(x, a)^2\Delta(x, a)^2 - w(x, e)^2\Delta(x, a, e)^2\right] \tag{16}$$

where

$$\bar{d}_{\pi_b}(x,a,e) := \sum_y d_{\pi_b}(x,a,e,y) \tag{17}$$

$$d_{\pi_b}(x,a,e,y) := p(x)\pi_b(a|x)p(e|x,a)p(y|x,a,e) \tag{18}$$

$$\Delta(x,a) := q(x,a) - \widehat{q}(x,a) \tag{19}$$

$$\Delta(x,a,e) := q(x,a,e) - \widehat{q}(x,a,e) \tag{20}$$

Proof. See Appendix C. □

By the formula of the variance reduction, if we assume that action embedding represents the action well, vanilla importance weight $w(x,a)$ is often larger than marginal importance weight $w(x,e)$. In addition, $\Delta(x,a)$ is larger than $\Delta(x,a,e)$ for most of the cases. Thus, we have the variance reduction $n\left(\mathbb{V}_D\left[\widehat{V}_{\mathrm{DR}}(\pi_e;D,\widehat{q})\right] - \mathbb{V}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e;D,\widehat{q})\right]\right) > 0$

4.3 Simulation Study

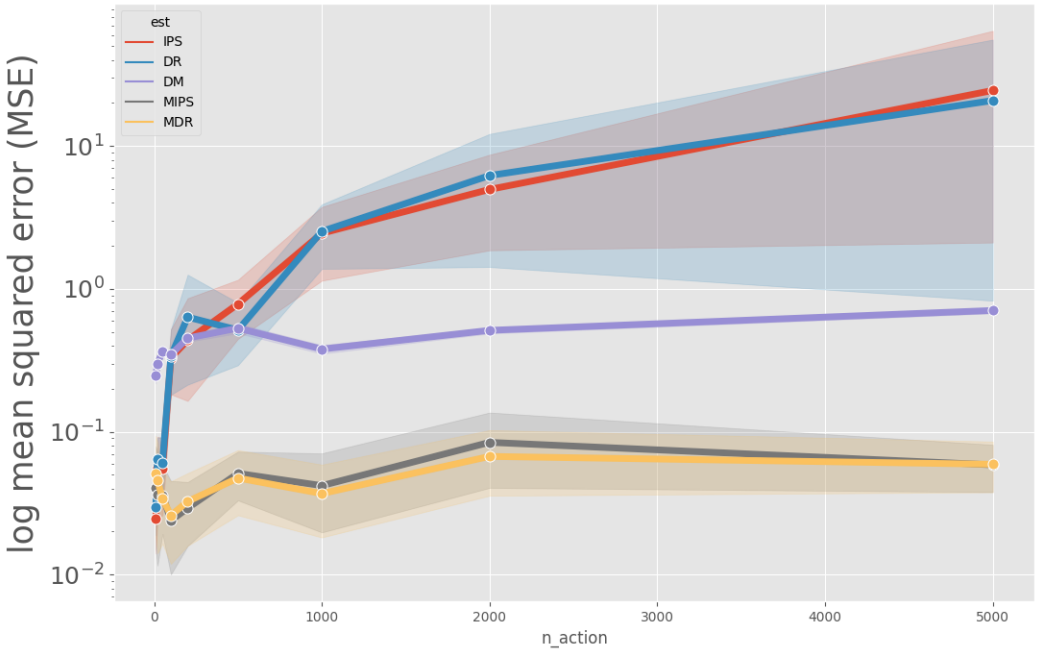

Figure 1: log MSE

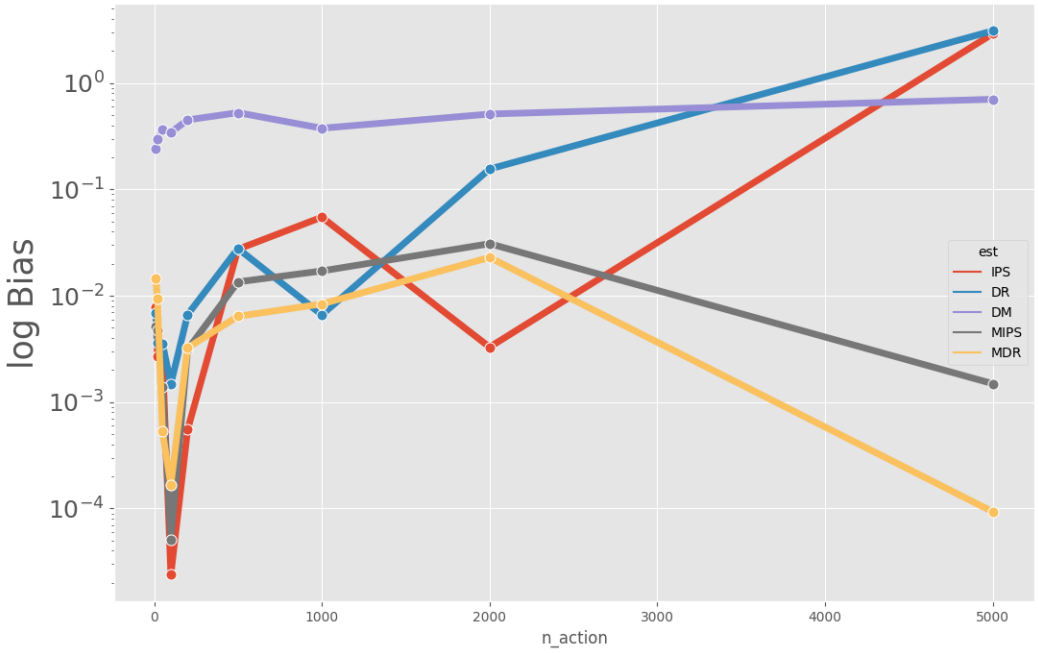

Figure 2: log Bias

To ensure the theoretical guarantee and applicability of MDR, I conducted the simulation study by implementing MDR I proposed by using the Open Bandit Pipeline (Saito et al. [2020]. The comprehensive results can be found in the appendix D The environment of the simulation is mostly the same as Saito and Joachims [2022]. Figure 1 4.3 shows MSE of DM, IPS, DR, MIPS, and MDR. As the number of action space $|\mathcal{A}|$ gets large, DM, IPS, and DR have high MSE. For DM, this is primarily because of the large bias whereas for IPS and DR, this is mainly due to the large variance caused by the wide importance weight $w(x, a)$. MIPS and MDR overcome this problem by using the marginalized importance weight $w(x, e)$. Furthermore, MDR has a lower bias than MIPS in this setting because as I showed in the theoretical analysis, MDR has doubly robust unbiasedness 17 as shown in Figure 2 4.3. Moreover, as I showed the variance reduction of MDR compared to DR, Figure 3 shows the reduction of the variance of MDR. Thus, I empirically show that MDR works better than DM Beygelzimer and Langford [2009], IPS Horvitz and Thompson [1952], DR Dudik et al. [2011], and MIPS Saito and Joachims [2022].

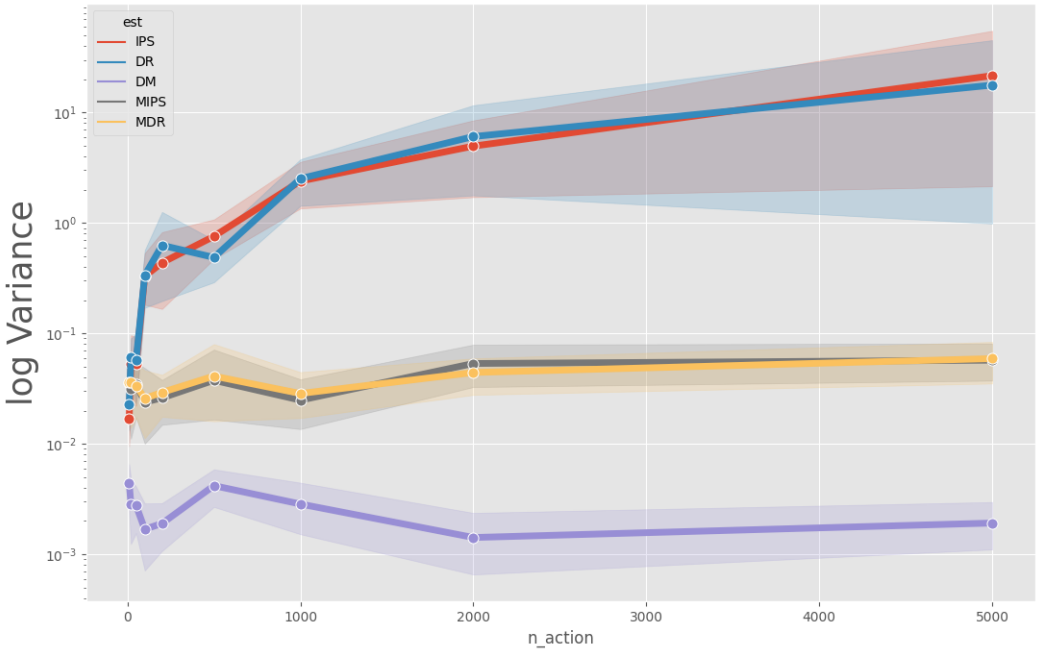

Figure 3: log Variance

5 Limitation and Future work

For the limitation of the paper, though I conducted the simulation study, I did not empirically analyze every situation but the same environment as Saito and Joachims [2022]. Therefore, for future work, it would be better if we conduct a comprehensive experiment such as the cases where the assumption for the unbiasedness of MDR 16 13 12 is fully satisfied or not satisfied so that we can analyze how robust MDR is when the assumption is violated.

Furthermore, it is important to find the better action embedding e to have the common embedding support 13, which is one of the assumptions for the unbiasedness of MDR, so in the future, it would be interesting to find the algorithm to construct the better embedding.

Moreover, when I was trying to create MDR to combine the properties of MIPS and DR, I come up with three estimators and two of them have doubly robust unbiasedness under different assumptions. Thus, it would be interesting to compare the candidates of MDR by simulation study or combine them to construct the triply robust estimator.

References

A. Beygelzimer and J. Langford. The offset tree for learning with partial labels. *KDD*, pp. 129–138, 2009. doi: https://doi.org/10.48550/arXiv.0812.4044.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: http://www.jstor.org/stable/2280784.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *ICML 2011*, arXiv:1103.4601 [cs.LG], 2011. doi: https://doi.org/10.48550/arXiv.1103.4601.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. *ICML 2022*, arXiv:2202.06317 [cs.LG], 2022. doi: https://doi.org/10.48550/arXiv.2202.06317.

A. Strehl, J. Langford, L. Li, and S. Kakade. Learning from logged implicit exploration data. *NuerIPS*, pp. 2217–2225, 2011. doi: https://hunch.net/~jl/projects/interactive/scavenging/scavenging.pdf.

J. Langford, A. L. Strehl, and J. Wortman. Exploration scavenging. *ICML*, pp. 528–535, 2008. doi: https://hunch.net/~jl/projects/interactive/scavenging/scavenging.pdf.

C. M. Cassel, C. E. S¨arndal, and J. H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620, 1976.

A. Huang, L. Leqi, Z. C. Lipton, and K. Azizzadenesheli. Off-policy risk assessment in contextual bandits. *NeurIPS 2021*, arXiv:2104.08977 [cs.LG], 2021. doi: https://doi.org/10.48550/arXiv.2104.08977.

Y. Saito, S. Aihara, M. Matsutani, and Y. Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv*, arXiv:2008.07146 [cs.LG], 2020. doi: https://doi.org/10.48550/arXiv.2008.07146.

A Python Code for the simulation study

GitHub: BIS631-Advanced-Techniques-in-Causal-Inference-Final-Project-MDR

B Proof of the unbiasedness of MDR

Proof. First, when the common embedding support 13 is satisfied, we have

$$
\begin{aligned}
&\mathbb{E}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e; D, \widehat{q})\right] && (21)\\
&= \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i,a)] + w(x_i,e_i)\left(y_i - \widehat{q}(x_i,a_i,e_i)\right)\right\}\right] && (22)\\
&= \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n} w(x_i,e_i)y_i\right] + \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i,a)] - w(x_i,e_i)\widehat{q}(x_i,a_i,e_i)\right\}\right] && (23)\\
&= \mathbb{E}_D\left[\widehat{V}_{\mathrm{MIPS}}(\pi_e; D)\right] + \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i,a)] - w(x_i,e_i)\widehat{q}(x_i,a_i,e_i)\right\}\right] && (24)\\
&= V(\pi) + \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i,a)] - w(x_i,e_i)\widehat{q}(x_i,a_i,e_i)\right\}\right] && (25)\\
&= V(\pi) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{p(x_i)\pi_b(a_i|x_i)p(e_i|x_i,a_i)p(y_i|x_i,a_i,e_i)}\left[\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i,a)] - w(x_i,e_i)\widehat{q}(x_i,a_i,e_i)\right\}\right] && (26)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)p(y|x,a,e)}\left[\left\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] - w(x,e)\widehat{q}(x,a,e)\right\}\right] && (27)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] - \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)}\left[w(x,e)\widehat{q}(x,a,e)\right] && (28)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] - \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\pi_b(a|x)\sum_{e\in\mathcal{E}}p(e|x,a)\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\widehat{q}(x,e)\right] && (29)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] - \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{E}}\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\widehat{q}(x,e)\sum_{a\in\mathcal{A}}\pi_b(a|x)p(e|x,a)\right] && (30)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] - \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{E}}\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\widehat{q}(x,e)p(e|x,\pi_b)\right] && (31)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] - \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{E}}p(e|x,\pi_e)\widehat{q}(x,e)\right] && (32)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] - \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{E}}\sum_{a\in\mathcal{A}}\pi_e(a|x)p(e|x,a)\widehat{q}(x,e)\right] && (33)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] - \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}}\pi_e(a|x)\widehat{q}(x,a)\right] && (34)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] - \mathbb{E}_{p(x)\pi_e(a|x)}\left[\widehat{q}(x,a)\right] && (35)\\
&= V(\pi) && (36)\\
& && (37)
\end{aligned}
$$

We use the definition of MDR for equation (22). For equation (25), we use the unbiasedness of MIPS. For equation (27), we use the i.i.d assumption of the data. For equation (29) and (34), we use the no direct effect $q(x, a, e) = q(x, e)$. From equation (28) to (34), we change the order of the summation by the linearity.

Secondly, when the unbiasedness of the estimated expected outcome is satisfied, we have

$$
\begin{aligned}
&\mathbb{E}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e; D, \widehat{q})\right] && (38)\\
&= \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i, a)] + w(x_i, e_i)\left(y_i - \widehat{q}(x_i, a_i, e_i)\right)\right\}\right] && (39)\\
&= \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i, a)]\right] + \mathbb{E}_D\left[w(x_i, e_i)\left(y_i - \widehat{q}(x_i, a_i, e_i)\right)\right] && (40)\\
&= \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i, a)]\right] + \mathbb{E}_D\left[w(x_i, e_i)\left(y_i - q(x_i, a_i, e_i)\right)\right] && (41)\\
&= \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi_e(a|x_i)}[q(x_i, a)]\right] + \mathbb{E}_D\left[w(x_i, e_i)\left(y_i - q(x_i, a_i, e_i)\right)\right] && (42)\\
&= \mathbb{E}_D\left[\widehat{V}_{\mathrm{DM}}(\pi_e; D, \widehat{q})\right] + \mathbb{E}_D\left[w(x_i, e_i)\left(y_i - q(x_i, a_i, e_i)\right)\right] && (43)\\
&= V(\pi) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{p(x_i)\pi_b(a_i|x_i)p(e_i|x_i,a_i)p(y_i|x_i,a_i,e_i)}\left[w(x_i, e_i)\left(y_i - q(x_i, a_i, e_i)\right)\right] && (44)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)p(y|x,a,e)}\left[w(x, e)\left(y - q(x, a, e)\right)\right] && (45)\\
&= V(\pi) + \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)}\left[w(x, e)\left(q(x, a, e) - q(x, a, e)\right)\right] && (46)\\
&= V(\pi) && (47)
\end{aligned}
$$

For equations (42) and (43), we use the assumption of the unbiasedness of the estimated expected outcome function. From equation (44) to (45), we use the i.i.d. assumption. □

C Proof of the variance reduction of MDR

Proof.

$$
\begin{aligned}
& n\left(\mathbb{V}_D\left[\widehat{V}_{\mathrm{DR}}(\pi_e; D, \widehat{q})\right] - \mathbb{V}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e; D, \widehat{q})\right]\right) && (48)\\
&= \mathbb{V}_{p(x)\pi_b(a|x)p(e|x,a)p(y|x,a,e)}\left[\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] + w(x,a)\left(y - \widehat{q}(x,a)\right)\right] && (49)\\
&\quad - \mathbb{V}_{p(x)\pi_b(a|x)p(e|x,a)p(y|x,a,e)}\left[\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] + w(x,e)\left(y - \widehat{q}(x,a,e)\right)\right] && (50)\\
&= \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)p(y|x,a,e)}\left[\left\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] + w(x,a)\left(y - \widehat{q}(x,a)\right)\right\}^2\right] && (51)\\
&\quad - \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)p(y|x,a,e)}\left[\left\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] + w(x,e)\left(y - \widehat{q}(x,a,e)\right)\right\}^2\right] && (52)\\
&= \mathbb{E}_{d_{\pi_b}}\left[\left\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] + w(x,a)\left(y - \widehat{q}(x,a)\right)\right\}^2\right] && (53)\\
&\quad - \mathbb{E}_{d_{\pi_b}}\left[\left\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] + w(x,e)\left(y - \widehat{q}(x,a,e)\right)\right\}^2\right] && (54)\\
&= \mathbb{E}_{d_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\left\{w(x,a)(y - \widehat{q}(x,a)) - w(x,e)(y - \widehat{q}(x,a,e))\right\}\right] && (55)\\
&\quad - \mathbb{E}_{d_{\pi_b}}\left[w(x,a)^2(y - \widehat{q}(x,a))^2 - w(x,e)^2(y - \widehat{q}(x,a,e))^2\right] && (56)\\
&= \mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\left\{w(x,a)\Delta(x,a) - w(x,e)\Delta(x,a,e)\right\}\right] && (57)\\
&\quad - \mathbb{E}_{\bar{d}_{\pi_b}}\left[w(x,a)^2\Delta(x,a)^2 - w(x,e)^2\Delta(x,a,e)^2\right] && (58)
\end{aligned}
$$

From equations (48) to (49) and (50), we use the i.i.d. assumption of the data-generating process and the definition of DR and MDR. For equations (51) and (52), as DR and MDR are unbiased under the assumptions, we focus on the expectation of the second moment. For $\mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,a)\Delta(x,a)\right]$,, we have

$$
\begin{aligned}
& \mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,a)\Delta(x,a)\right] && (59)\\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{a\in\mathcal{A}}\pi_b(a|x)\frac{\pi_e(a|x)}{\pi_b(a|x)}\Delta(x,a)\right] && (60)\\
&= \mathbb{E}_{p(x)\pi_e(a|x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\Delta(x,a)\right] && (61)\\
& && (62)
\end{aligned}
$$

For $\mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,e)\Delta(x,a,e)\right]$, we have

$$
\begin{aligned}
&\mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,e)\Delta(x,a,e)\right] && (63)\\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{a\in\mathcal{A}}\pi_b(a|x)\sum_{e\in\mathcal{E}}p(e|x,a)\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\Delta(x,e)\right] && (64)\\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{e\in\mathcal{E}}\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\Delta(x,e)\sum_{a\in\mathcal{A}}\pi_b(a|x)p(e|x,a)\right] && (65)\\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{e\in\mathcal{E}}\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\Delta(x,e)p(e|x,\pi_b)\right] && (66)\\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{e\in\mathcal{E}}p(e|x,\pi_e)\Delta(x,e)\right] && (67)\\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{e\in\mathcal{E}}\sum_{a\in\mathcal{A}}\pi_e(a|x)p(e|x,a)\Delta(x,a,e)\right] && (68)\\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{a\in\mathcal{A}}\pi_e(a|x)\sum_{e\in\mathcal{E}}p(e|x,a)\Delta(x,a,e)\right] && (69)\\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{a\in\mathcal{A}}\pi_e(a|x)\Delta(x,a)\right] && (70)\\
&= \mathbb{E}_{p(x)\pi_e(a|x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\Delta(x,a)\right] && (71)\\
& && (72)
\end{aligned}
$$

For equations (64) and (68), we used the no direct effect 12. Therefore, $\mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,a)\Delta(x,a)\right]$ and $\mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,e)\Delta(x,a,e)\right]$ cancel out, so we have

$$
\begin{aligned}
&n\left(\mathbb{V}_D\left[\widehat{V}_{\mathrm{DR}}(\pi_e;D,\widehat{q})\right]-\mathbb{V}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e;D,\widehat{q})\right]\right) && (73)\\
&= \mathbb{E}_{\bar{d}_{\pi_b}}\left[w(x,a)^2\Delta(x,a)^2-w(x,e)^2\Delta(x,a,e)^2\right] && (74)
\end{aligned}
$$

□

D the comprehensive result of the experiment

Here, I also show the simulation study without taking the log of MSE, bias, and variance.

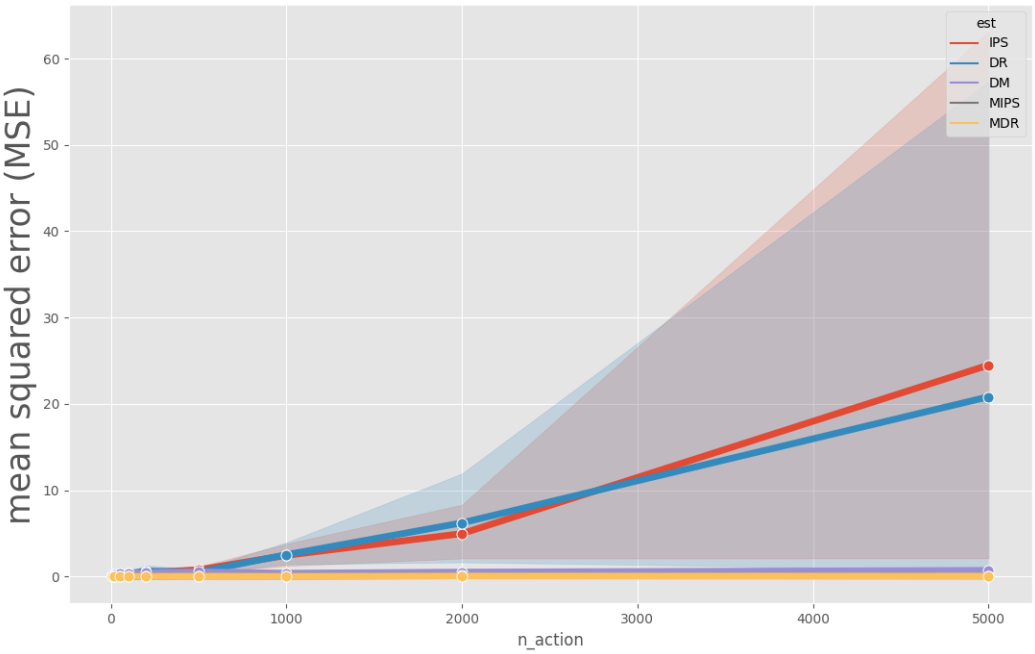

Figure 4: MSE

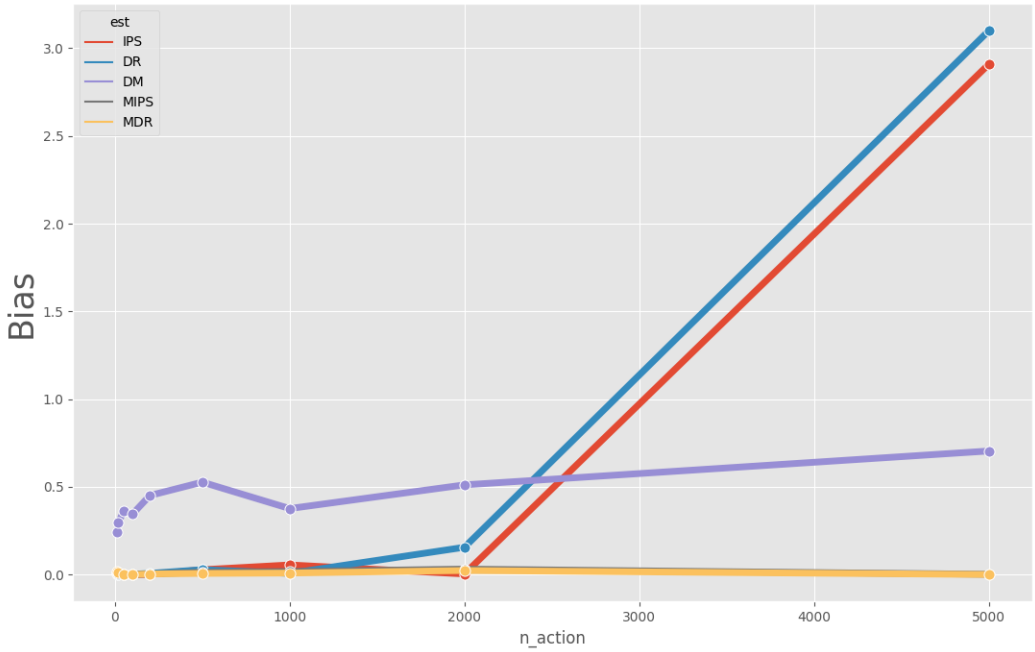

Figure 5: Bias

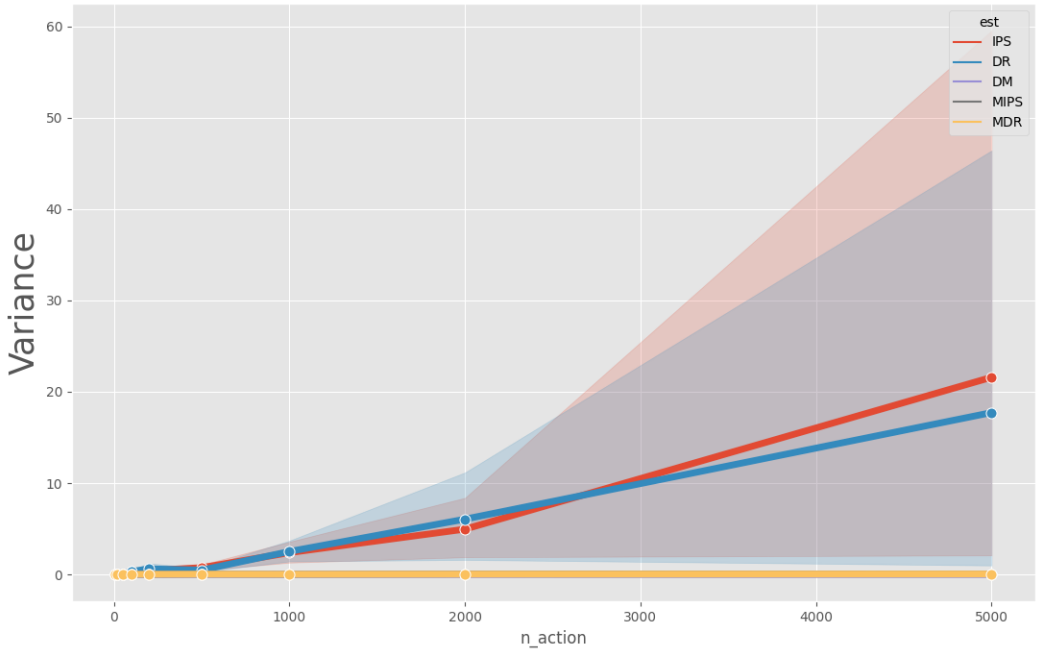

Figure 6: Variance