2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Doubly Robust Estimator for Off-Policy Evaluation with Large Action Spaces

Tatsuhiro Shimizu
Department of Political Science and Economics
Waseda University
Tokyo, Japan
t.shimizu432@akane.waseda.jp

Laura Forastiere
Department of Biostatistics
Yale School of Public Health
New Haven, Connecticut, USA
laura.forastiere@yale.edu

***Abstract*—We study Off-Policy Evaluation (OPE) in contextual bandit settings with large action spaces. The benchmark estimators suffer from severe bias and variance tradeoffs. Parametric approaches suffer from bias due to difficulty specifying the correct model, whereas ones with importance weight suffer from variance. To overcome these limitations, Marginalized Inverse Propensity Scoring (MIPS) was proposed to mitigate the estimator's variance via embeddings of an action. To make the estimator more accurate, we propose the doubly robust estimator of MIPS called the *Marginalized Doubly Robust (MDR)* estimator. Theoretical analysis shows that the proposed estimator is unbiased under weaker assumptions than MIPS while maintaining variance reduction against IPS, which was the main advantage of MIPS. The empirical experiment verifies the supremacy of MDR against existing estimators.**

***Index Terms*—Off-Policy Evaluation, Counterfactual Machine Learning, Doubly Robust Estimator**

I. Introduction

Many intelligent systems like recommendation systems [1]–[5] and personalized medicine [6] gradually utilize individual-level data to optimize their decision-making for individuals to enhance the users' experience. One of the ways to attain such an aim is to conduct an A/B test online, but the drawback of such an online algorithm [7], [8] is that it is costly and may harm the user experience. Therefore, much research [9]–[16] has been done to use the past data to evaluate the intervention accurately, which is called *Off-Policy Evaluation (OPE)*. Most of the problems in OPE are formulated in the contextual bandits' settings [11] where we observe context (e.g., the demography of the users), an action collected by the already implemented policy in the system called *behavior policy*, and reward (e.g., the click, purchase of the product, recovery of the patient). In such settings, the goal of OPE is to evaluate the counterfactual performance of the evaluation policy that the system had not implemented. Unfortunately, with large action spaces, which is often the case in the real world, existing estimators suffer from bias due to misspecification of the model or variance due to the wide range of importance weight. To circumvent these limitations, Saito and Joachims [16] proposed Marginalized Inverse Propensity Scoring (MIPS), utilizing importance weight on the space of action embedding to have lower variance. However, the assumption required for the unbiasedness of MIPS does not necessarily hold, as it requires that the embedding of the action fully mediates the effect of the action on the reward, which is not realistic in practice. To alleviate this limitation of MIPS, we develop a doubly robust estimator called MDR, which is unbiased under either the assumption required for MIPS or that on the model while preserving the preferable variance reduction of MIPS. Synthetic data analysis validates that MDR is the most accurate estimator of existing estimators.

II. Background

The data we consider is the contextual vector $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, action $a \in \mathcal{A}$, and reward $r \in [0, r_{\max}]$. We assume that we sample the context vector from an unknown distribution $x \sim p(x)$, action $a \sim \pi(a|x)$ from the stochastic intervention called *policy* $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ given contextual vector x, and reward r from an unknown distribution $r \sim p(r|x, a)$ given contextual vector x and action a. We observe independent and identically distributed n samples collected by the behavior policy π_b [17], [18]. Thus, the actually observed data called *logged bandit data* D is given by $D = \{(x_i, a_i, r_i)\}_{i=1}^n$ where $x_i \sim p(x_i), a_i \sim \pi_b(a_i|x_i), r_i \sim p(r_i|x_i, a_i)$ for all $i \in [n]$. If we know the distribution of contextual vector $p(x)$ and reward $p(r|x, a)$, then we can obtain the performance of the policy called *value function* $V(\pi)$ of π, which means how good the policy π is by

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

where $q(x, a) = \mathbb{E}_{p(r|x,a)}[r|x, a]$ is the expected reward given contextual vector x and action a. However, as we do not know the distribution of contextual vector $p(x)$ and reward $p(r|x, a)$ in practice, we need to estimate the value function to evaluate the performance of a 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 $\widehat{V}(\pi_e; D) \approx V(\pi_e)$ where we use the Mean Squared Error (MSE)

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

as the quantity to measure how good the estimator is.

III. Existing Estimators

There are three classical estimators to estimate the value function $V(\pi_e)$: Direct Method (DM) [19], Inverse Propensity Score (IPS) [20], and Doubly Robust (DR) [9]. These estimators have drawbacks when the cardinality of the action space is large. Marginalized Inverse Propensity Scoring (MIPS) plays a significant role in large action spaces to mitigate such shortcomings.

A. Direct Method

DM [19] uses the estimated expected reward function $\widehat{q}$ to estimate the value function as follows.

$$\begin{aligned}\widehat{V}_{\text{DM}}(\pi_e; D, \widehat{q}) &:= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i, a)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\pi_e(a|x_i)\widehat{q}(x_i, a).\end{aligned}$$

where $\widehat{q}(x, a)$ is the estimated expected reward given contextual vector x and a. For instance, we can consider the following function: $\widehat{q} = \operatorname{argmin}_{q'\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - q'(x_i, a_i)\right)^2$ where $\mathcal{Q}$ is some space of the model where we want to optimize q'. DM is unbiased under the perfect estimation of the expected reward [9] as follows.

Assumption 1 (Perfect Estimation of Expected Reward Function)**.** We say that the regression model $\widehat{q}$ has perfection estimation if $\widehat{q}(x, a) = q(x, a)$ for all context $x \in \mathcal{X}$ and action $a \in \mathcal{A}$.

In practice, it is significantly difficult to accurately estimate the regression model $\widehat{q}$, so DM incurs a significant bias in the event of a large action space.

B. Inverse Propensity Scoring

Unlike the parametric approach like DM, IPS [20] re-weights the reward by the ratio of the propensity scores of behavior and evaluation policies as follows.

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

where $w(x, a)$ is a *vanilla importance weight* defined as $w(x, a) := \pi_e(a|x)/\pi_b(a|x)$. IPS is unbiased [16] under the common support defined as follows, which often holds in practice as long as the behavior policy assigns a non-zero probability to the action whose probability in evaluation policy is non-zero, defined as follows.

Assumption 2 (Common Support)**.** We say behavior policy π_b satisfies the common support for the evaluation policy π_e if $\pi_e(a|x) > 0 \implies \pi_b(a|x) > 0$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$.

Furthermore, the variance of IPS [16] is as follows.

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

where $\sigma(x, a) := \mathbb{V}_{p(r|x,a)}[r|x, a]$ is the variance of the reward. Though the unbiasedness of IPS is preferable, the importance weight $w(x, a)$ has a wide range, resulting in a significant variance by the first and third terms when the cardinality of the action space is large.

C. Doubly Robust

DR [9] combines the preferable properties of DM and IPS, defined as follows.

$$\begin{aligned}&\widehat{V}_{\text{DR}}(\pi_e; D, \widehat{q}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i, a)] + w(x_i, a_i)\left(r_i - \widehat{q}(x_i, a_i)\right)\right\}\end{aligned}$$

DR guarantees unbiasedness under either Assumption 1 or 2. Furthermore, we have the variance of DR:

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

where $\Delta(x, a) := q(x, a) - \widehat{q}(x, a)$ is the prediction error of the expected reward. When the cardinality of the action space is large, the variance of DR gets large due to the first and third terms.

D. Marginalized Inverse Propensity Scoring

To tackle the problem of the significant bias of DM and variance of IPS and DR under large action spaces, MIPS [16] was proposed. 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 action a is the movie we recommend, the embedding e can be the genre, directors, and actors which categorize the movie. If the embedding characterizes the film well, 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. To use action embedding, we define the new data-generating process and value function as follows.

We assume that we sample the action embedding e from an unknown distribution $e \sim p(e|x, a)$ given x and a, and the reward from an unknown distribution $r \sim p(r|x, a, e)$ given x, a, and e. Thus, the logged data D is

$$\begin{aligned}D &= \{(x_i, a_i, e_i, r_i)\}_{i=1}^{n} \\ &\sim \prod_{i=1}^{n} p(x_i)\pi_b(a_i|x_i)p(e_i|x_i, a_i)p(r_i|x_i, a_i, e_i)\end{aligned}$$

. Furthermore, we define the value function $V(\pi)$ as follows.

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

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

Having the new data-generating process and the definition of the value function, MIPS [16] is defined as follows.

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

where $w(x,e) := p(e|x,\pi_e)/p(e|x,\pi_b)$ 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 3 (No Direct Effect of Action on Reward)**.** Given context x and action embedding e, action a and reward r are independent $(a \perp\!\!\!\perp r|x,e)$

No direct effect assumption means that the action embedding e fully mediates the effect of action a on the reward r.

Assumption 4 (Common Embedding Support)**.** We say that the data-generating process satisfies the common embedding support if $p(e|x,\pi_e) \implies p(e|x,\pi_b)$ for all $x \in \mathcal{X}$ and $e \in \mathcal{E}$.

Common embedding support is a weaker assumption than the common support necessary for the IPS's unbiasedness. Even with the considerable variance reduction of MIPS against IPS, the unbiasedness of MIPS is not necessarily guaranteed in practice as Assumption 3 usually does not hold due to the difficulty finding the perfect embedding of the action.

IV. Proposed Estimator

In this section, we combine MIPS and DR to construct a more accurate estimator to overcome the shortcomings of DM, IPS, DR, and MIPS.

A. Marginalized Doubly Robust

Using the preferable properties of MIPS and DR, we propose the novel estimator called ***M**arginalized **D**oubly **R**obust (**MDR**)* as follows.

$$\begin{aligned}&\widehat{V}_{\text{MDR}}(\pi_e; D, \widehat{q})\\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i,a)] + w(x_i,e_i)\left(r_i - \widehat{q}(x_i,a_i,e_i)\right)\right\}\end{aligned}$$

The first term of MDR is the baseline estimator, which uses the regression model $\widehat{q}$ to incorporate the parametric approach. The second term incorporates the importance weight $w(x_i, e_i)$ on the action embedding space to weight the residuals of the estimated expected reward function, having the doubly robust structure. Intuitively, MDR is supposed to estimate the value function of the evaluation policy more accurately than IPS since MDR has a lighter importance weight than IPS to mitigate the high variance. At the same time, it has doubly robust properties, which require weaker assumptions to be unbiased than MIPS. We theoretically analyze the statistical properties of MDR in the following subsection.

B. Theoretical Analysis

MDR has a doubly robust property under either the no direct effect and the common embedding support or the following assumption about the precision of the prediction of the reward function.

Assumption 5 (Perfect Estimation of Expected Reward given x, a, and e)**.** We say regression model $\widehat{q}(x,a,e)$ perfectly estimates the expected reward function q if $\widehat{q}(x,a,e) = q(x,a,e)$ for all $x \in \mathcal{X}, a \in \mathcal{A}$, and $e \in \mathcal{E}$

Proposition 6 (Unbiasedness of MDR)**.** MDR is unbiased under either Assumptions 3 and 4, or Assumption 5. See Appendix A for the proof.

Proposition 6 shows that MDR is more likely to be unbiased than MIPS as MDR requires weaker assumptions than MIPS due to its doubly robust structure. As we know the variance of DR, we only focus on reducing the variance of MDR against DR as follows.

Proposition 7 (Variance Reduction of MDR against DR)**.** Under Assumptions 2, 3, and 4, the difference between the variances of DR and MDR is

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

where $\bar{d}_{\pi_b}(x,a,e) := \sum_r d_{\pi_b}(x,a,e,r)$ and $d_{\pi_b}(x,a,e,r) := p(x)\pi_b(a|x)p(e|x,a)p(r|x,a,e)$ are visitation measures and $\Delta(x,a) := q(x,a) - \widehat{q}(x,a)$ and $\Delta(x,a,e) := q(x,a,e) - \widehat{q}(x,a,e)$ estimation errors. See Appendix B for the proof.

If we assume that action embedding represents the action well, which can be guaranteed using the estimator selection method called SLOPE [14], [21] or PAS-IF [22] to execute the data-driven embedding selection, 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 cases. Thus, we have the variance reduction.

$$\mathbb{V}_D\left[\widehat{V}_{\text{DR}}(\pi_e; D, \widehat{q})\right] > \mathbb{V}_D\left[\widehat{V}_{\text{MDR}}(\pi_e; D, \widehat{q})\right]$$

Note that as the variance of DR is smaller than IPS, that of MDR is also smaller than IPS.

V. Simulation Study

We conducted the synthetic data experiment by implementing MDR we proposed using the Open Bandit Pipeline [23]. The Python code for the simulation can be found in the https://github.com/tatsu432/DR-estimator-OPE-large-action. The simulation environment is mostly the same as MIPS [16].

A. Synthetic Data

In the synthetic data, we define the data-generating process as follows. The contextual vector x is drawn i.i.d from the 10-

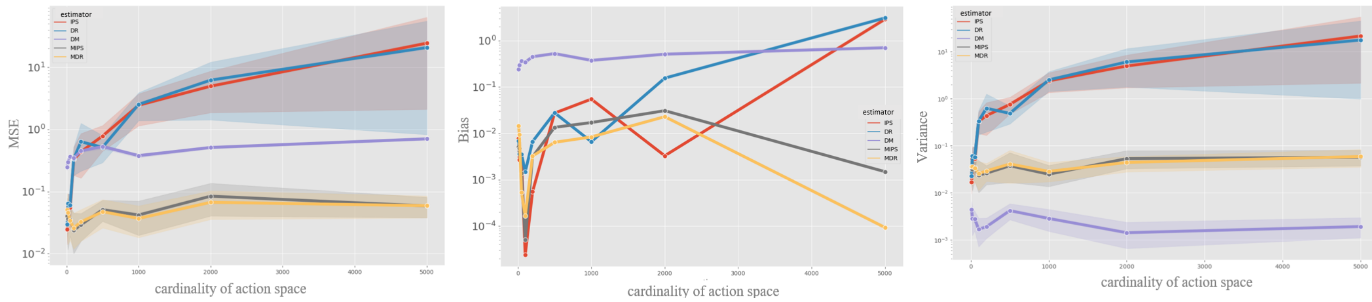

Fig. 1. MSE (left), Bias (center), and variance (right) of DM, IPS, DR, MIPS, and MDR(ours) when we change the cardinality of the action space

dimensional standard normal distribution. Given action a, the embedding of the action is drawn i.i.d from the distribution

$$p(e|x,a) = \prod_{k\in[d_e]} \frac{\exp(\alpha_{a,e_k})}{\sum_{e'\in\mathcal{E}_k}\exp(\alpha_{a,e'_k})}$$

where α_{a,e_k} is a set of parameters drawn from the standard normal distribution $\mathcal{N}(0,1)$ and the cardinality of the action embedding space is 10; $\mathcal{E}_k = [10]$. The behavior policy π_b is then defined as

$$\pi_b(a|x) = \frac{\exp(\beta\cdot q(x,a))}{\sum_{a'\in\mathcal{A}}\exp(\beta\cdot q(x,a'))}$$

where β is the parameter that handles the optimality of the behavior policy and $q(x,a) := \mathbb{E}_{p(e|x,a)}[q(x,e)]$. The expected reward function $q(x,e)$ given context x and action embedding e is defined as

$$q(x,e) = \sum_{k\in[d_e]} \eta_k \cdot (x^\top M x_{e_k} + \theta_x^\top x + \theta_e^\top x_{e_k})$$

where M, θ_x, and θ_e are parameters whose elements we sample from the uniform distribution whose range is $[-1,1]$ and η_k represents the importance of the k-th dimension of the action embedding sampled from Dirichlet distribution such that $\sum_{k\in[d_e]}\eta_k = 1$. Then the evaluation policy π_e whose value function we want to estimate is defined as

$$\pi_e(a|x) := (1-\epsilon)\cdot \mathbb{1}\{a = \operatorname*{argmax}_{a'\in\mathcal{A}} q(x,a')\} + \epsilon/|\mathcal{A}|$$

where $\epsilon\in[0,1]$ represents the quality of the evaluation policy π_e and we set it to $\epsilon = 0.05$.

B. Results

Fig. 1 shows the MSE, bias, and variance of DM, IPS, DR, MIPS, and MDR. As the number of action spaces $|\mathcal{A}|$ gets large, DM, IPS, and DR have high MSE. For DM, this is primarily because of the significant bias, whereas for IPS and DR, we can attribute this to the substantial variance caused by the wide importance weight. MIPS and MDR overcome this problem by using the marginalized importance weight. Furthermore, MDR has a lower bias than MIPS in this settingthis favorable property of MDR against MIPS results from the doubly robust property of MDR. Moreover, the variance of MDR is reduced against IPS and DR, as we derived in the theoretical analysis. Thus, the empirical experiment demonstrates that MDR enables us to evaluate the value function more accurately than DM, IPS [20], DR [9], and MIPS [16].

VI. Conclusion and Future Work

We studied OPE with large action spaces and proposed the MDR, which has the doubly robust property to be unbiased under weaker assumptions than MIPS and has a significantly lower variance than IPS. The simulation study demonstrates that MDR outperforms other existing estimators, including MIPS.

Our study gives rise to several interesting future directions. It is crucial to find the better action embedding e to have the standard embedding support, 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, other estimators might be doubly robust under different assumptions. Thus, it would be intriguing to compare the candidates of MDR by simulation study or combine them to construct the triply robust estimator.

Acknowledgements

We thank Yuta Saito at Cornell University Department of Computer Science for his support and guidance.

References

[1] A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, and B. Carterette, "Offline evaluation to make decisions about playlist recommendation algorithms," in *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, 2019, pp. 420–428.

[2] Y. Narita, S. Yasui, and K. Yata, "Debiased off-policy evaluation for recommendation systems," in *ACM Conference on Recommender Systems*, 2020.

[3] Y. Saito, S. Yaginuma, Y. Nishino, H. Sakata, and K. Nakata, "Unbiased recommender learning from missing-not-at-random implicit feedback," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 501–509.

[4] Y. Saito and T. Joachims, "Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances," in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, p. 828830.

[5] T. Joachims, B. London, Y. Su, A. Swaminathan, and L. Wang, "Recommendations as treatments," in *AAAI AI Magazine*, 2021, pp. vol 42, number 3, pages 19–30.

[6] O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. Celi, "Guidelines for reinforcement learning in helathcare," *Nature Medicine*, 2019.

[7] A. Agarwal, N. Jiang, and M. Kakade, S, "Reinforcement learning: Theory and algorithms," *MIT*, 2020.

[8] T. Lattimore and G. Hinton, "Bandit algorithms," *Cambridge*, 2020.

[9] M. Dudík, D. Erhan, J. Langford, and L. Li, "Doubly robust policy evaluation and optimization," *Statistical Science*, vol. 29, no. 4, pp. 485–511, 2014.

[10] A. Swaminathan and T. Joachims, "Counterfactual risk minimization: Learning from logged bandit feedback," in *International Conference on Machine Learning*. PMLR, 2015, pp. 814–823.

[11] Y.-X. Wang, A. Agarwal, and M. Dudık, "Optimal and adaptive off-policy evaluation in contextual bandits," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3589–3597.

[12] M. Farajtabar, Y. Chow, and M. Ghavamzadeh, "More robust doubly robust off-policy evaluation," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. PMLR, 2018, pp. 1447–1456.

[13] Y. Su, L. Wang, M. Santacatterina, and T. Joachims, "Cab: Continuous adaptive blending for policy evaluation and learning," in *International Conference on Machine Learning*, vol. 84, 2019, pp. 6005–6014.

[14] Y. Su, P. Srinath, and A. Krishnamurthy, "Adaptive estimator selection for off-policy evaluation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9196–9205.

[15] A. M. Metelli, A. Russo, and M. Restelli, "Subgaussian and differentiable importance sampling for off-policy evaluation and learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[16] Y. Saito and T. Joachims, "Off-policy evaluation for large action spaces via embeddings," in *International Conference on Machine Learning*. PMLR, 2022, pp. 19 089–19 122.

[17] A. Strehl, J. Langford, L. Li, and S. M. Kakade, "Learning from logged implicit exploration data," in *Advances in Neural Information Processing Systems*, vol. 23, 2010, pp. 2217–2225.

[18] J. Langford, A. Strehl, and J. Wortman, "Exploration scavenging," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 528–535.

[19] A. Beygelzimer and J. Langford, "The offset tree for learning with partial labels," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 129–138.

[20] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 663–685, 1952.

[21] G. Tucker and J. Lee, "Improved estimator selection for off-policy evaluation," *Workshop on Reinforcement Learning Theory at the 38th International Conference on Machine Learning*, 2021.

[22] T. Udagawa, H. Kiyohara, Y. Narita, Y. Saito, and K. Tateno, "Policy-adaptive estimator selection for off-policy evaluation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2023.

[23] Y. Saito, S. Aihara, M. Matsutani, and Y. Narita, "Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation," *arXiv preprint arXiv:2008.07146*, 2020.

Appendix A
Proof of the unbiasedness of MDR

Proof. First, when Assumption 4 is satisfied, we have

$$
\begin{aligned}
&\mathbb{E}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e; D, \widehat{q})\right] \\
&= V(\pi) + \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)p(r|x,a,e)}\Big[\big\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] \\
&\quad - w(x,e)\widehat{q}(x,a,e)\big\}\Big] \qquad (1) \\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] \\
&\quad - \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}} \pi_b(a|x) \sum_{e\in\mathcal{E}} p(e|x,a)\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\widehat{q}(x,e)\right] \qquad (2) \\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] \\
&\quad - \mathbb{E}_{p(x)}\left[\sum_{e\in\mathcal{E}} \frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\widehat{q}(x,e)p(e|x,\pi_b)\right] \\
&= V(\pi) + \mathbb{E}_{p(x)\pi_e(a'|x)}\left[\widehat{q}(x,a')\right] \\
&\quad - \mathbb{E}_{p(x)}\left[\sum_{a\in\mathcal{A}} \pi_e(a|x)\widehat{q}(x,a)\right] \\
&= V(\pi)
\end{aligned}
$$

For (1), we use the i.i.d assumption of the data. For (2), we use the no direct effect $q(x,a,e) = q(x,e)$.

Secondly, when the Assumption 5 is satisfied, we have

$$
\begin{aligned}
&\mathbb{E}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e; D, \widehat{q})\right] \\
&= \mathbb{E}_D\Big[\frac{1}{n}\sum_{i=1}^{n}\big\{\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i,a)] \\
&\quad + w(x_i,e_i)\left(r_i - \widehat{q}(x_i,a_i,e_i)\right)\big\}\Big] \\
&= \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi_e(a|x_i)}[\widehat{q}(x_i,a)]\right] \\
&\quad + \mathbb{E}_D\left[w(x_i,e_i)\left(r_i - q(x_i,a_i,e_i)\right)\right] \qquad (3) \\
&= \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\pi_e(a|x_i)}[q(x_i,a)]\right] \\
&\quad + \mathbb{E}_D\left[w(x_i,e_i)\left(r_i - q(x_i,a_i,e_i)\right)\right] \qquad (4) \\
&= V(\pi) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{p(x_i)\pi_b(a_i|x_i)p(e_i|x_i,a_i)p(r_i|x_i,a_i,e_i)}\Big[\\
&\quad w(x_i,e_i)\big(r_i - q(x_i,a_i,e_i)\big)\Big] \\
&= \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)p(r|x,a,e)}\left[w(x,e)\left(r - q(x,a,e)\right)\right] \\
&\quad + V(\pi) \qquad (5) \\
&= V(\pi)
\end{aligned}
$$

For (3) and (4), we use the assumption of the unbiasedness of the estimated expected outcome function. For (5), we use the i.i.d. assumption. □

Appendix B
Proof of the Variance Reduction of MDR

Proof. Under Assumptions 2, 3, and 4 (i.e., the common support, no direct effect of action on reward, and the common embedding support), we can derive the variance reduction of MDR against DR as follows.

$$
\begin{aligned}
& n\left(\mathbb{V}_D\left[\widehat{V}_{\mathrm{DR}}(\pi_e; D, \widehat{q})\right] - \mathbb{V}_D\left[\widehat{V}_{\mathrm{MDR}}(\pi_e; D, \widehat{q})\right]\right) \\
&= \mathbb{V}_{p(x)\pi_b(a|x)p(e|x,a)p(r|x,a,e)}\big[\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] \\
&\quad + w(x,a)\left(r - \widehat{q}(x,a)\right)\big] \\
&\quad - \mathbb{V}_{p(x)\pi_b(a|x)p(e|x,a)p(r|x,a,e)}\big[\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] \\
&\quad + w(x,e)\left(r - \widehat{q}(x,a,e)\right)\big] \qquad (6) \\
&= \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)p(r|x,a,e)}\Big[\big\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] \\
&\quad + w(x,a)\left(r - \widehat{q}(x,a)\right)\big\}^2\Big] \\
&\quad - \mathbb{E}_{p(x)\pi_b(a|x)p(e|x,a)p(r|x,a,e)}\Big[\big\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] \\
&\quad + w(x,e)\left(r - \widehat{q}(x,a,e)\right)\big\}^2\Big] \qquad (7) \\
&= \mathbb{E}_{d_{\pi_b}}\Big[\Big\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] + w(x,a)\big(r - \\
&\quad \widehat{q}(x,a)\big)\Big\}^2\Big] - \mathbb{E}_{d_{\pi_b}}\Big[\big\{\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')] + w(x,e) \\
&\quad \times \left(r - \widehat{q}(x,a,e)\right)\big\}^2\Big] \\
&= \mathbb{E}_{d_{\pi_b}}\Big[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\big\{w(x,a)(r - \widehat{q}(x,a)) \\
&\quad - w(x,e)(r - \widehat{q}(x,a,e))\big\}\Big] - \mathbb{E}_{d_{\pi_b}}\Big[w(x,a)^2(r \\
&\quad - \widehat{q}(x,a))^2 - w(x,e)^2(r - \widehat{q}(x,a,e))^2\Big] \\
&= \mathbb{E}_{\bar{d}_{\pi_b}}\Big[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\big\{w(x,a)\Delta(x,a) \\
&\quad - w(x,e)\Delta(x,a,e)\big\}\Big] \\
&\quad - \mathbb{E}_{\bar{d}_{\pi_b}}\left[w(x,a)^2\Delta(x,a)^2 - w(x,e)^2\Delta(x,a,e)^2\right]
\end{aligned}
$$

Equation (6) is from the independent and identically distributed data-generating process of context, action, action embedding, reward, and the definition of DR and MDR. For (7), 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

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

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

$$
\begin{aligned}
& \mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,e)\Delta(x,a,e)\right] \\
&= \mathbb{E}_{p(x)}\Bigg[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{a\in\mathcal{A}}\pi_b(a|x) \\
&\quad \times \sum_{e\in\mathcal{E}}p(e|x,a)\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\Delta(x,e)\Bigg] \\
&= \mathbb{E}_{p(x)}\Bigg[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{e\in\mathcal{E}}\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)}\Delta(x,e) \\
&\quad \times \sum_{a\in\mathcal{A}}\pi_b(a|x)p(e|x,a)\Bigg] \\
&= \mathbb{E}_{p(x)}\Bigg[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{e\in\mathcal{E}}\frac{p(e|x,\pi_e)}{p(e|x,\pi_b)} \\
&\quad \times \Delta(x,e)p(e|x,\pi_b)\Bigg] \\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{e\in\mathcal{E}}p(e|x,\pi_e)\Delta(x,e)\right] \\
&= \mathbb{E}_{p(x)}\Bigg[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{e\in\mathcal{E}}\sum_{a\in\mathcal{A}}\pi_e(a|x) \\
&\quad \times p(e|x,a)\Delta(x,a,e)\Bigg] \qquad (8) \\
&= \mathbb{E}_{p(x)}\Bigg[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{a\in\mathcal{A}}\pi_e(a|x) \\
&\quad \times \sum_{e\in\mathcal{E}}p(e|x,a)\Delta(x,a,e)\Bigg] \\
&= \mathbb{E}_{p(x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\sum_{a\in\mathcal{A}}\pi_e(a|x)\Delta(x,a)\right] \\
&= \mathbb{E}_{p(x)\pi_e(a|x)}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]\Delta(x,a)\right]
\end{aligned}
$$

For (8), we used the no direct effect. Therefore,

$$
\begin{aligned}
& \mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,a)\Delta(x,a)\right] \\
&= \mathbb{E}_{\bar{d}_{\pi_b}}\left[2\mathbb{E}_{\pi_e(a'|x)}[\widehat{q}(x,a')]w(x,e)\Delta(x,a,e)\right]
\end{aligned}
$$

holds, so we have

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

□