Off-Policy Evaluation from Logged Human Feedback

Aniruddha Bhargava [1] **Lalit Jain** [2] **Branislav Kveton** [3] **Ge Liu** [4] **Subhojyoti Mukherjee** [5]

Abstract

Learning from human feedback has been central to recent advances in artificial intelligence and machine learning. Since the collection of human feedback is costly, a natural question to ask is if the new feedback always needs to collected. Or could we evaluate a new model with the human feedback on responses of another model? This motivates us to study off-policy evaluation from logged human feedback. We formalize the problem, propose both model-based and model-free estimators for policy values, and show how to optimize them. We analyze unbiasedness of our estimators and evaluate them empirically. Our estimators can predict the absolute values of evaluated policies, rank them, and be optimized.

1. Introduction

Large language models (LLMs) (Bommasani et al., 2021) have recently emerged as general purpose inference machines that achieve human-level performance on a wide range of tasks (Brown et al., 2020; Mirchandani et al., 2023). The key step in training them is *reinforcement learning with human feedback (RLHF)*, where these models are aligned to generate human-preferred text (Ouyang et al., 2022). The key idea in RLHF is to use human feedback to learn a latent reward model. After this, a policy is optimized to maximize the reward under the reward model. In DPO (Rafailov et al., 2023), the reward model is reparameterized using the policy, which is then optimized. One common property of RLHF and DPO is their reliance on human feedback. After each alignment, a new dataset with human feedback is generated. A natural question to ask is: do new datasets always need to be collected or could we evaluate a new LLM with human feedback on responses of another LLM?

Motivated by this question, we study off-policy evaluation (Li et al., 2010; Bottou et al., 2013; Li et al., 2015; Hofmann et al., 2016) with human feedback. We formulate the problem as follows. Given a logged dataset of n lists of responses, which are generated by an LLM and re-ranked by humans according to their preferences, we want to estimate how another LLM would align with human feedback without another human study. We say that the LLM aligns with human feedback when its first response is the same as the most preferred human response. We envision two main use cases for our approach. The first use case is counterfactual evaluation (Bottou et al., 2013; Li et al., 2015), akin to those in learning to rank (Joachims et al., 2017; Swaminathan et al., 2017; Li et al., 2018). The second use case is LLM alignment (Ouyang et al., 2022; Rafailov et al., 2023) with an interpretable objective that represents counterfactual human feedback. To allow these, we make algorithmic contributions in off-policy evaluation and optimization, and summarize them below:

1. We formalize the problem of off-policy evaluation from logged human feedback as offline evaluation with ranked lists (Joachims et al., 2017; Swaminathan et al., 2017; Li et al., 2018). Specifically, given a logged dataset of n lists of length K, which are generated by the logging policy and re-ranked by humans according to their preferences, we want to estimate how another policy would align with human feedback without another human study. The novelty in our work is in a new feedback model induced by the *Plackett-Luce (PL) model* (Plackett, 1975; Luce, 2005; Zhu et al., 2023b). The *Bradley-Terry-Luce (BTL) model* (Bradley and Terry, 1952; Sekhari et al., 2024) is a special case of $K = 2$. From the feedback point of view, our problem is both a bandit (Lattimore and Szepesvari, 2019), because only K responses are ranked by a human; and full-information (Auer et al., 1995), because the full ranking of the K responses is observed.

2. We propose model-based (Robins et al., 1994; Dudik et al., 2014) and model-free (Horvitz and Thompson, 1952; Strehl et al., 2010) estimators. The model-based

[1] Amazon [2] University of Washington [3] AWS AI Labs, Santa Clara, USA [4] University of Illinois Urbana-Champaign [5] Department of Electrical and Computer Engineering, University of Wisconsin-Madison.. Correspondence to: Branislav Kveton <bkveton@amazon.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024, Workshop on Models of Human Feedback for AI Alignment. Copyright 2024 by the author(s).

[*] This work was done outside of Amazon. The author names are listed alphabetically.

estimator relies on a learned reward model, similarly to RLHF. However, unlike in RLHF, the mean reward is the probability that the policy aligns with human feedback, instead of being latent. This makes our estimator interpretable. The model-free approach is based on *inverse propensity scores (IPS)* (Horvitz and Thompson, 1952; Ionides, 2008). We improve upon a naive application of IPS over ranked lists by IPS over sets of responses (Section 3.3). This is only possible because of the PL feedback model, which depends on the set of ranked responses but not their order. We analyze basic properties of our estimators, such as unbiasedness. We also show how to optimize them in a practical way (Swaminathan and Joachims, 2015).

3. We comprehensively evaluate our estimators in multiple experiments. First, we show that they can estimate the absolute value of evaluated policies. Second, we show that they can rank policies better than the latent reward functions that RLHF and DPO optimize. Third, we confirm our findings on a real-world dataset where the reward model is misspecified. Finally, we show that optimization of our estimators leads to comparable policies to those learned by RLHF and DPO.

This paper is organized as follows. In Section 2, we introduce the problem of off-policy evaluation from logged human feedback. In Section 3, we present our model-based and model-free estimators. In Section 4, we show how to optimize our estimators. In Section 5, we empirically evaluate our estimators. We review prior works in Section 6 and conclude in Section 7.

2. Setting

We study the following setting. A policy, for instance given by an LLM, interacts with a human for n rounds. In round $t \in [n]$, the policy generates a ranked list of K responses A_t to the query in round t. The human critiques A_t and provides their preferred order of the responses $A_{t,*}$, represented as a permutation of A_t. The policy *aligns with human feedback* in round t when A_t and $A_{t,*}$ are *similar*. We want to reuse previously logged data to estimate alignment of another policy without collecting additional human feedback.

We formalize our problem as follows. The query in round t is $x_t \in \mathcal{X}$, where $\mathcal{X}$ is the query set. There are L potential responses to any query and we denote the set of integers that indexes them by $\mathcal{A} = [L]$. We do not assume that L is small and comment on its impact on our estimators throughout the paper. A logging policy π_0 generates a ranked list of K responses $A_t = (a_{t,i})_{i=1}^K$ from $\mathcal{A}$, where $a_{t,i} \in \mathcal{A}$ is the i-th response. The human responds with a permutation of A_t, $A_{t,*} = (a_{t,*,i})_{i=1}^K$, that represents their preferred order of the responses.

The setting of $K < L$ is motivated by the limited capacity of humans to provide preferential feedback on too many choices L. When $K = 2$, we have a relative feedback over two responses, as in the Bradley-Terry-Luce model (Bradley and Terry, 1952). When $K > 2$, we have a ranking feedback over K responses, as in the Plackett-Luce model (Plackett, 1975; Luce, 2005). We define the policies and human alignment next.

Policy. We denote by $\pi(a \mid x)$ the probability that policy π generates a response $a \in \mathcal{A}$ to query $x \in \mathcal{X}$. A ranked list of K responses $A = (a_i)_{i=1}^K$ is sampled from $\pi(\cdot \mid x)$ with probability

$$\pi(A \mid x) = \prod_{i=1}^{K} \frac{\pi(a_i \mid x)}{\sum_{j=i}^{L} \pi(a_j \mid x)}\,, \tag{1}$$

where $a_{k+1}, \dots, a_L$ are arbitrarily ordered responses from $\mathcal{A} \setminus A$. In plain English, the first response is sampled with probability $\pi(a_1 \mid x)$, the second with probability $\pi(a_2 \mid x) / \sum_{j=2}^{L} \pi(a_j \mid x)$, and the i-th with probability $\pi(a_i \mid x) / \sum_{j=i}^{L} \pi(a_j \mid x)$. This is similar to the PL model (Plackett, 1975; Luce, 2005).

The probabilities of individual responses can be obtained from an LLM as follows. Take L most frequent responses to each query x and let $\tilde{\pi}(a \mid x)$ be the probability (potentially unnormalized) of the tokens corresponding to response a. Then $\pi(a \mid x) = \tilde{\pi}(a \mid x) / \sum_{a' \in \mathcal{A}} \tilde{\pi}(a' \mid x)$.

Reward. The alignment of the policy with human feedback can be defined in many ways. In this work, we say that the policy *aligns with human feedback* in round t when the first responses in A_t and $A_{t,*}$ are identical. We represent the alignment using a numerical reward. Specifically, the *reward* in round t is $\mathbb{1}\{a_{t,1} = a_{t,*,1}\}$ and the *mean reward* is

$$r(x_t, A_t) = \mathbb{P}\left(a_{t,1} = a_{t,*,1} \,|\, x_t, A_t\right)\,, \tag{2}$$

where $A_{t,*}$ is the human-preferred order of responses A_t. We wanted to comment on two aspects of the mean reward. First, note that $A_{t,*}$ is a random variable that depends on A_t and x_t. Second, the mean reward is essentially the probability that the most preferred human response $a_{t,*,1}$ is the first logged response $a_{t,1}$.

The mean reward in (2) can be justified from many points of view. First, it is the probability that the first option out of K is chosen in a discrete choice model (Train, 2009; Ben-Akiva and Lerman, 2018), a broad class of classic models of human preferences. Second, it is analogous to Precision@K in ranking (Manning et al., 2008) for $K = 1$, where the human feedback is the ground truth and the policy is a ranker. Finally, as we shall see in Section 3, this quantity can be estimated in a model-free fashion from our feedback, both on- and off-policy. For concreteness, one can think of the mean reward as

$$r(x, A; w) = \frac{\exp[\phi(x, a_1)^\top w]}{\sum_{i=1}^{K} \exp[\phi(x, a_i)^\top w]} \,, \tag{3}$$

where $A = (a_i)_{i=1}^{K}$ is a ranked list of K responses, $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is a feature map, and $w \in \mathbb{R}^d$ is a reward model parameter. We denote the true unknown parameter by w_*. A natural way of thinking of (3) is as the probability that the first response A_1 is preferred over the rest. The algebraic form is borrowed from the PL model. This has two implications. First, w can be estimated using existing techniques for the PL model (Zhu et al., 2023b). Second, the probability depends on all responses in A but not their order. We exploit this property in the design of our more advanced estimators in Section 3.3.

With the definitions of policies and human-feedback alignment in hand, we can define the value of a policy. The value of policy π is the probability that its first response aligns with the human-preferred response. Formally, over n rounds with queries $(x_t)_{t=1}^{n}$, this is

$$V(\pi) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{A_t \sim \pi(\cdot \mid x_t)} \left[r(x_t, A_t) \right] . \tag{4}$$

Our goal is to estimate this quantity from human feedback on responses generated by another policy π_0. Note that we could have assumed that $x_t \sim p$ for some distribution p. We do not because the stochasticity of queries is not necessary to derive any result in this work.

3. Off-Policy Evaluation

The key idea in our work is to pose the problem of off-policy evaluation from logged human feedback as a counterfactual evaluation problem (Li et al., 2010; Bottou et al., 2013; Li et al., 2015; Hofmann et al., 2016) over ranked lists (Joachims et al., 2017; Swaminathan et al., 2017; Li et al., 2018). Before we get to the details, note that this problem is particularly easy when the human feedback is collected using policy π. Then an unbiased estimate of $V(\pi)$ in (4) is the frequency that the first responses in A_t and $A_{t,*}$ are identical,

$$\hat{V}(\pi) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{a_{t,1} = a_{t,*,1}\} . \tag{5}$$

We prove this in Lemma 1 (Appendix A). Now suppose that $A_t \sim \pi_0(\cdot \mid x_t)$, where π_0 is the logging policy. Then $V(\pi)$ can be estimated using either model-based or model-free techniques.

3.1. Direct Method

A popular approach to off-policy evaluation is the *direct method (DM)* (Dudik et al., 2014). The key idea in the DM is to learn a reward model and then use it to estimate the expected value of a policy. A natural choice in our setting is (3). We estimate w by solving the maximum likelihood estimation (MLE) problem

$$\hat{w} = \arg\max_{w} \sum_{t=1}^{n} \sum_{i=1}^{K} \log \left(\frac{\exp[\phi(x_t, a_{t,*,i})^\top w]}{\sum_{j=i}^{K} \exp[\phi(x_t, a_{t,*,j})^\top w]} \right) . \tag{6}$$

Note that this estimator is defined over K responses while the reward model in (3) is the probability of the first response over the rest. This is not harmful since all sampling stages in the PL model share the same parameter. With the reward model in hand, we estimate the value of policy π as

$$\hat{V}_{\text{DM}}(\pi) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{A \sim \pi(\cdot \mid x_t)} \left[r(x_t, A; \hat{w}) \right] , \tag{7}$$

where $r(x, A; \hat{w})$ is the estimated preference for response a_1 from A. Also note that (6) is analogous to learning the reward model in RLHF (Section 6). The difference in (7) is in how the reward model is used. In (17), the latent preference for a single human response a is estimated. In (7), we estimate the probability that the first response in A is preferred by the human.

The strength of the DM estimator is that the policy π only needs to be sampled from. The probability $\pi(A \mid x_t)$, which requires normalization and thus a summation over L terms in (1), is not needed. On the other hand, the estimator may perform poorly when the reward model is biased.

3.2. Propensity-Based Methods

Another popular approach to off-policy evaluation are *inverse propensity scores (IPS)* (Horvitz and Thompson, 1952; Ionides, 2008). The key idea in IPS is to reweigh the data collected by the logging policy π_0 as if they were logged by the evaluated policy π. Specifically, let $A_t \sim \pi(\cdot \mid x_t)$. Then based on the definition of the expected value of a policy in (4), the IPS estimator is

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n} \sum_{t=1}^{n} \frac{\pi(A_t \mid x_t)}{\pi_0(A_t \mid x_t)} \mathbb{1}\{a_{t,1} = a_{t,*,1}\} . \tag{8}$$

We prove that this estimator is unbiased in Lemma 2 (Appendix A).

The strength of the IPS estimator is that it does not make any assumptions about the reward model. It only reweighs the logged rewards $\mathbb{1}\{a_{t,1} = a_{t,*,1}\}$ by $\pi(A_t \mid x_t)/\pi_0(A_t \mid x_t)$. The estimator has two shortcomings. First, its variance can be high when the policies π and π_0 are not close, because $\pi(A_t \mid x_t)/\pi_0(A_t \mid x_t)$ is high. Second, the computation of $\pi(A \mid x_t)$ requires normalization over L terms in (1), which may be challenging when L is large.

The *doubly-robust method (DR)* (Robins et al., 1994; Dudik et al., 2014; Jiang and Li, 2016) combines the advantages of the DM and IPS. Specifically, it uses the model as a variance-reduction techniques in the IPS. For the DM and IPS estimators in (7) and (8), respectively, it is

$$\begin{aligned}\hat{V}_{\text{DR}}(\pi) = \frac{1}{n}\sum_{t=1}^{n} \frac{\pi(A_t \mid x_t)}{\pi_0(A_t \mid x_t)} (\mathbb{1}\{a_{t,1} = a_{t,*,1}\} \\ - r(x_t, A_t; \hat{w})) + \hat{V}_{\text{DM}}(\pi)\,.\end{aligned} \tag{9}$$

The DR estimator is unbiased when the DM is unbiased or the propensities in the IPS estimator are correctly specified. We prove this in Lemma 3 (Appendix A)

The DR estimator tends to have a lower variance than the IPS estimator when the original rewards $\mathbb{1}\{a_{t,1} = a_{t,*,1}\}$ have high variances but the centered rewards $\mathbb{1}\{a_{t,1} = a_{t,*,1}\} - r(x_t, A_t; \hat{w})$ do not. The estimator also inherits the computational limitation of IPS, that the normalization over L terms in (1) is needed to compute the propensity scores.

3.3. Advanced Propensity-Based Methods

The additional structure of our problem allows us to improve the IPS estimator in (8). To define the improved estimator, we need additional notation. Let S_t be the set of responses in A_t forgetting the ranking and $\nu(S_t \mid x_t)$ be the probability that the set S_t is generated in round t by the policy π. The quantities $\pi(A \mid x)$ and $\nu(S \mid x)$ are related as

$$\nu(S \mid x) = \textstyle\sum_{A \in \Pi(S)} \pi(A \mid x)\,, \tag{10}$$

for any x and S, where $\Pi(S)$ is the set of all permutations of S. Moreover, let

$$\pi(a \mid S, x) = \frac{\pi(a \mid x)}{\sum_{a' \in S} \pi(a' \mid x)}$$

be the probability that policy π generates a response a from set S to query x.

The key insight is that the human feedback tells us which response in A_t is preferred. As a result, we can do counterfactual reasoning conditioned on x_t and S_t. Specifically, let $\mathcal{S}$ be the set of all subsets of $\mathcal{A}$ of size K. Then the value of policy π in (4) can be rewritten as

$$\begin{aligned}
V(\pi) &= \frac{1}{n}\sum_{t=1}^{n} \sum_{S_t \in \mathcal{S}} \sum_{A_t \in \Pi(S_t)} \pi(A_t \mid x_t)\, r(x_t, A_t) \\
&= \frac{1}{n}\sum_{t=1}^{n} \sum_{S_t \in \mathcal{S}} \nu(S_t \mid x_t) \sum_{A_t \in \Pi(S_t)} \frac{\pi(A_t \mid x_t)}{\nu(S_t \mid x_t)}\, r(x_t, A_t) \\
&= \frac{1}{n}\sum_{t=1}^{n} \sum_{S_t \in \mathcal{S}} \nu(S_t \mid x_t)\, \mathbb{E}\left[\mathbb{1}\{a_{t,1} = a_{t,*,1}\} \mid x_t, S_t\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \sum_{S_t \in \mathcal{S}} \nu(S_t \mid x_t) \cdot \\
&\qquad \mathbb{E}\left[\sum_{a \in \mathcal{A}} \pi(a \mid S_t, x_t)\, \mathbb{1}\{a_{t,*,1} = a\} \,\middle|\, x_t, S_t\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \sum_{S_t \in \mathcal{S}} \nu(S_t \mid x_t)\, \mathbb{E}\left[\pi(a_{t,*,1} \mid S_t, x_t) \mid x_t, S_t\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{S_t \sim \nu(\cdot \mid x_t)}\left[\pi(a_{t,*,1} \mid S_t, x_t)\right]\,.
\end{aligned}$$

The fourth equality follows from the assumption that $A_{t,*}$ depends only on the set of responses S_t. The expectation in the last step is needed because $A_{t,*}$ remains to be a random variable.

Now note that the process of sampling ranked lists can also be viewed as sampling sets. Therefore, when $S_t \sim \nu(\cdot \mid x_t)$, $V(\pi)$ can be estimated as

$$\hat{V}_{\text{SET}}(\pi) = \frac{1}{n}\sum_{t=1}^{n} \pi(a_{t,*,1} \mid S_t, x_t)\,. \tag{11}$$

We prove that this estimator is unbiased in Lemma 4 (Appendix A). The main improvement over (5) is due to reducing variance, by eliminating the randomness in $A_t \mid S_t$. Similarly to the IPS estimator in (8), when $S_t \sim \nu_0(\cdot \mid x_t)$, the IPS estimator becomes

$$\hat{V}_{\text{SET-IPS}}(\pi) = \frac{1}{n}\sum_{t=1}^{n} \frac{\nu(S_t \mid x_t)}{\nu_0(S_t \mid x_t)} \pi(a_{t,*,1} \mid S_t, x_t)\,. \tag{12}$$

We call it `SetIPS` because the propensity scores are over sets of responses instead of ranked lists. We prove that this estimator is unbiased in Lemma 5 (Appendix A).

The main improvement in (12) over (8) is due to reducing variance. In particular, the propensities over a larger set, of all lists of length K, are replaced with a smaller set, of all sets of size K. Due to the relation in (10), this leads to higher propensities and therefore a better control of their ratios. The new estimator has an interesting behavior as $K \to L$, and in particular $K = L$. In the latter case, the propensities vanish because $\nu(S_t \mid x_t) = \nu_0(S_t \mid x_t) =$

1. The classic IPS estimator in (8) does not poses this property and would be cursed by low propensities. `SetIPS` also inherits the limitations of IPS, that the variance can be high and that the normalization over L terms in (1) can be challenging.

Similarly to (9), we can define a *set doubly-robust method (*`SetDR`*)*. Specifically, the probability of aligning with human feedback under policy π given set S and reward model $r(x, A; w)$ is

$$r(x, S; \pi, w) = \sum_{A \in \Pi(S)} \frac{\pi(A \mid x)}{\nu(S \mid x)} r(x, A; w)\,.$$

When this model is used as a variance-reduction techniques in `SetIPS`, we get

$$\hat{V}_{\text{SET-DR}}(\pi) = \frac{1}{n} \sum_{t=1}^{n} \frac{\nu(S_t \mid x_t)}{\nu_0(S_t \mid x_t)} (\pi(a_{t,*,1} \mid S_t, x_t) - r(x_t, S_t; \pi, \hat{w})) + \hat{V}_{\text{DM}}(\pi) \tag{13}$$

The DR estimator is unbiased when the DM is unbiased or the propensities in the IPS estimator are correctly specified. We prove this in Lemma 6 (Appendix A)

4. Off-Policy Optimization

The strength of our approach is that any estimator from Section 3 can be used for optimization. In particular, let $\pi(a \mid x; \theta)$ be a parameterization of policy π by $\theta \in \mathbb{R}^d$. Then the optimization of its estimated value can be viewed as maximizing

$$\hat{V}(\pi) - \gamma \frac{1}{n} \sum_{t=1}^{n} d(\pi(\cdot \mid x_t; \theta), \pi_0(\cdot \mid x_t))\,, \tag{14}$$

where $d(p, q)$ is the KL divergence between distributions p and q with support $\mathcal{A}$ and $\gamma > 0$ is a tunable parameter. The KL term can be viewed as a constraint that forces π to be close to π_0, and plays the same role as in RLHF and DPO (Section 6). The gradient of (14) with respect to θ is

$$\nabla \hat{V}(\pi) - \gamma \frac{1}{n} \sum_{t=1}^{n} \nabla d(\pi(\cdot \mid x_t; \theta), \pi_0(\cdot \mid x_t))\,. \tag{15}$$

We derive $\nabla \hat{V}(\pi)$ for all of our estimators in Appendix B. To compute the sum, we sample $\tilde{A}_t \sim \pi(\cdot \mid x_t; \theta)$ and replace $\nabla d(\pi(\cdot \mid x_t; \theta), \pi_0(\cdot \mid x_t))$ with its unbiased estimate

$$(\nabla \log \pi(\tilde{A}_t \mid x_t; \theta))(1 + \log \pi(\tilde{A}_t \mid x_t; \theta) - \log \pi_0(\tilde{A}_t \mid x_t))\,.$$

We properly derive this gradient in Appendix B. As in RLHF, the normalization over L terms in (1) is needed to compute the above log-probabilities.

5. Experiments

We conduct four experiments. In Section 5.1, we evaluate our off-policy estimators on predicting the absolute value of a policy. In Section 5.2, we evaluate them on ranking policies. In Section 5.3, we apply our estimators to large language models. Finally, in Section 5.4, we optimize our off-policy estimators. In all plots but in Figure 4, we report standard errors of the estimates.

5.1. Absolute Error

We first evaluate how good our off-policy estimators are in predicting the absolute value of a policy. This experiment is conducted on synthetic problems, which are generated as follows. The number of potential responses is $L = 7$. We choose this value because the maximum number of unique ranked lists $7! = 5\,040$ when $K = L$. This is more than enough to illustrate the impracticality of naive IPS estimators and gains due to more advanced estimators (Section 3.3). For each query $x \in \mathcal{X}$, we generate a random vector $u_x \in [-1, 1]^4$. For each response $a \in \mathcal{A}$, we generate a random vector $v_a \in [-1, 1]^4$. The feature vector of query x and response a is $\phi(x, a) = \text{vec}(u_x v_a^\top)$ and has length $d = 16$. The reward model parameter is sampled as $w_* \sim \mathcal{N}(\mathbf{0}_d, 10^2 I_d)$.

We consider a parametric class of policies

$$\pi(a \mid x; \theta) = \frac{\exp[\phi(x, a)^\top \theta]}{\sum_{i=1}^{L} \exp[\phi(x, i)^\top \theta]}\,. \tag{16}$$

The logging policy is specified by $\theta_0 = w_* + \varepsilon_0$ for $\varepsilon_0 \sim \mathcal{N}(\mathbf{0}_d, 5^2 I_d)$. Because its per-dimension variance 5^2 is much lower than that of w_*, the policy is likely to have a high reward. We evaluate $N = 5$ policies per run. Each evaluated policy is defined as $\theta_i = \theta_0 + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 I)$ is its perturbation. The parameter $\sigma_e > 0$ defines how close the evaluated policy θ_i is to the logging policy θ_0, and thus the hardness of the problem. We set $\sigma_e = 5$. The evaluation metric is the *absolute error* $\frac{1}{N} \sum_{i=1}^{N} |V(\theta_i) - \hat{V}(\theta_i)|$. All experiments are averaged over 50 runs, where we randomize w_*, θ_0, and all θ_i. The default values of the parameters are $K = 2$ and $n = 3\,000$.

Our results are reported in Figure 1. In Figures 1a-b, the model is correctly specified. In Figure 1a, we vary K. The best estimator is `SetDR` and the second best is DR. This does not come as a surprise since the superiority of DR estimators has long been recognized (Dudik et al., 2014). The IPS estimator performs the worst at $K = L$. `SetIPS` leverages the full observability at $K = L$ (Section 3.3) and thus performs well. In Figure 1b, all estimators improve with more logged interactions n.

In the next experiments, the reward model can be misspecified. The misspecification is obtained by adding indepen-

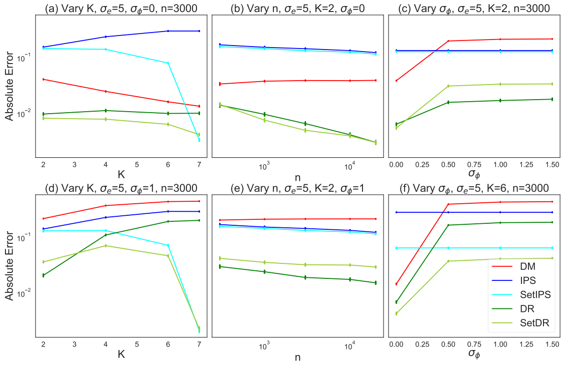

Figure 1: Absolute error of estimated policy values as we vary K, n, and σ_ϕ.

dent $\mathcal{N}(0, \sigma_\phi^2)$ noise to $\phi(x, a)$ in the reward model in (3). In Figure 1c, we vary σ_ϕ and observe its adverse effect on model-based estimators. Specifically, the DM becomes worse than IPS estimators, while the DR estimators remain more robust. In Figures 1d-e, we set $\sigma_\phi = 1$. In Figure 1d, we vary K. We observe that the DM performs the worst and SetIPS is the second best estimator for $K > 4$. In Figure 1e, all estimators improve with more logged interactions n. Finally, in Figure 1f, we vary σ_ϕ and observe its adverse effect on model-based estimators, at $K = 6$.

5.2. Relative Error

We left out RLHF and DPO from the absolute error comparison because they perform poorly. This is because they optimize a different notion of reward, which can be roughly viewed as the exponent in (3). However, since they are popular in optimization, they should be good in ranking policies. To test this, we conduct a relative comparison. Specifically, we repeat all experiments from Section 5.1 but change the metric to the number of incorrectly ranked evaluated policies,

$$\frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathbb{1}\left\{\mathrm{sgn}(\hat{V}(\theta_i) - \hat{V}(\theta_j)) \neq \mathrm{sgn}(V(\theta_i) - V(\theta_j))\right\}$$

We call this metric a *relative error*. Our results are reported in Figure 2. We observe two general trends. First, when the model is correctly specified, DM is among the best methods. Second, SetDR is always among the best methods.

5.3. Large Language Models

This experiment showcases our methods on a real-world Nectar dataset (Zhu et al., 2023a). We take 500 prompts from the dataset, each with 7 responses generated by popular LLMs, and treat it as a logged dataset with $n = 500$ and $L = 7$. The feature vectors in (3) are 768-dimensional Instructor embeddings (Su et al., 2022) of the prompt and its response. We compute the propensities of individual

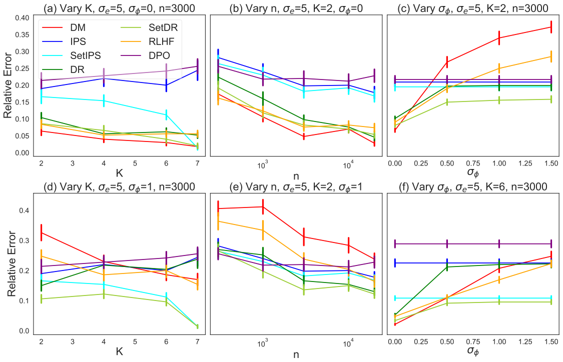

Figure 2: Ranking error of estimated policy values as we vary K, n, and σ_ϕ.

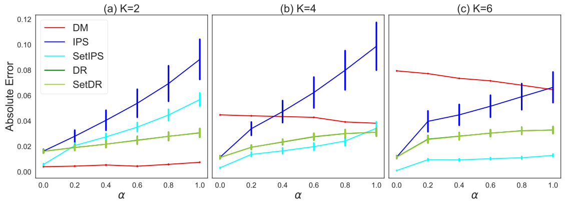

Figure 3: Evaluation of large language model policies on Nectar dataset by the absolute error. The parameter $\alpha \in [0, 1]$ interpolates between Phi3 and Llama3 policies.

responses using two large language models: $\pi_L(a \mid x)$ for LLama3 (Meta, 2024) and $\pi_P(a \mid x)$ for Phi3 (Abdin et al., 2024). We treat $\pi_P(a \mid x)$ as the logging policy and evaluate convex combinations of policies

$$\pi_\alpha(a \mid x) = (1 - \alpha)\pi_P(a \mid x) + \alpha\pi_L(a \mid x)$$

for $\alpha \in [0, 1]$. When $\alpha = 0$, the evaluated policy is the logging policy. When $\alpha = 1$, the evaluated policy is Llama3. The evaluation becomes progressively harder as $\alpha \to 1$. We use NVIDIA GeForce RTX 3090 GPU with 24GB RAM to load the large language models for $\pi_L(a \mid x)$ and $\pi_P(a \mid x)$. Phi3 requires less than 3GB RAM and LLama3 requires less than 7 GB RAM.

Our results are reported in Figure 3. For $K > 2$, SetIPS performs the best. It is followed the DR and SetDR, which perform almost identically. The IPS estimator performs the worst for all K. The DM works well for $K = 2$. This is because the mean rewards of the best two responses are close to 0.5 in most queries. As a result, they can be estimated well by a constant 0.5, while the propensity scores in IPS methods only introduce unnecessary variance as $\alpha \to 1$.

5.4. Policy Optimization

We experiment with the same problems as in Section 5.1. In the first problem, $K = 2$ and the logging policy is uniform, $\theta_0 = \mathbf{0}_d$. In the second problem, $K = 2$ and the logging policy has a high reward. We generate it as in Section 5.1. In the third problem, $K = 4$ and the logging policy is uniform. We optimize our estimators as described in Section 4. RLHF and DPO are implemented as described in Section 6. We set $\gamma = 0.001$ and optimize the policies by Adam (Kingma and Ba, 2015). The logged dataset size is $n = 1\,000$. We choose these settings because they resulted in stable optimization results.

In Figure 4, we report the values of all optimized policies in a single optimization run. The values are estimated using (11). We observe two main trends. In all plots, DPO is among the best methods. Second, DM sometimes outperforms RLHF, while SetDR is the second best method. Interestingly, while our estimators are not designed for optimization, some of them perform well in this setting.

6. Related Work

The closest related works are off-policy evaluation and optimization for ranking (Joachims et al., 2017; Swaminathan et al., 2017; Li et al., 2018). We differ from them in two key aspects. First, they consider absolute feedback on individual responses, such as clicks. Our feedback model is relative, the preference of one response over another. Second, these works assume that the responses in ranked lists are not affected by the logging policy; only their order is. Our logged lists contain K responses out of $L > K$, and thus are affected by logging. This setting is motivated by the limited capacity of humans to provide relative feedback on too many responses, when $L \gg K$.

While we focus on evaluation, recent works on learning with human feedback primarily focused on optimization. We take RLHF (Ouyang et al., 2022) as an example. In RLHF, the goal is to optimize

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{a \sim \pi(\cdot \mid x_t;\theta)}\left[r_{\text{RLHF}}(x_t, a)\right] - \gamma d(\pi(\cdot \mid x_t;\theta), \pi_0(\cdot \mid x_t))\,, \tag{17}$$

where $r_{\text{RLHF}}(x, a)$ in an estimate of the latent preference for response a to query x, $d(p, q)$ is the KL divergence between distributions p and q with support $\mathcal{A}$, $\gamma > 0$ is a tunable parameter, and θ is an optimized policy parameter. The reward model can be viewed as $r_{\text{RLHF}}(x, a) = \phi(x, a)^\top w$, where ϕ is the same feature map as in (3) and w is learned as in (6). The first term in (17) can be viewed as the value of policy θ in (4). Since $\phi(x, a)^\top w_{\text{RLHF}}$ is latent, the value is not interpretable. This is the main conceptual difference from our work. We also study a plethora of policy value estimators, beyond what would be considered as the direct method in (17).

One shortcoming of RLHF is the estimation of the latent reward model. This motivates DPO (Rafailov et al., 2023), where the latent reward is reparameterized using the optimized policy based on the structure of the maximized objective. For pairwise feedback, DPO maximizes

$$\sum_{t=1}^{n}\log\sigma\left(\gamma\log\left(\frac{\pi(A_{t,*,1} \mid x_t;\theta)}{\pi(A_{t,*,1} \mid x_t;\theta_0)}\right) - \gamma\log\left(\frac{\pi(A_{t,*,2} \mid x_t;\theta)}{\pi_0(A_{t,*,2} \mid x_t)}\right)\right)\,, \tag{18}$$

where $A_{t,*} = (A_{t,*,1}, A_{t,*,2})$ and $\sigma(v) = 1/(1 + \exp[v])$ is the sigmoid.

7. Conclusions

Learning from human feedback has been central to recent advances in artificial intelligence and machine learning. As more datasets with human feedback are collected, a natural question arises: do new datasets always need to be collected or can we reuse the old ones to estimate how a human would respond to a new policy? We propose both model-based and model-free estimators for this setting, analyze their unbiasedness, and show how to optimize them. We evaluate them empirically, on both synthetic and real-world datasets, and on both evaluation and optimization tasks. A natural direction to future work is studying other models of human feedback and reward (Train, 2009; Ben-Akiva and Lerman, 2018).

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331, 1995.

Moshe Ben-Akiva and Steven Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA, 2018.

Rishi Bommasani et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Leon Bottou, Jonas Peters, Joaquin Quinonero-Candela, Denis Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(101):3207–3260, 2013.

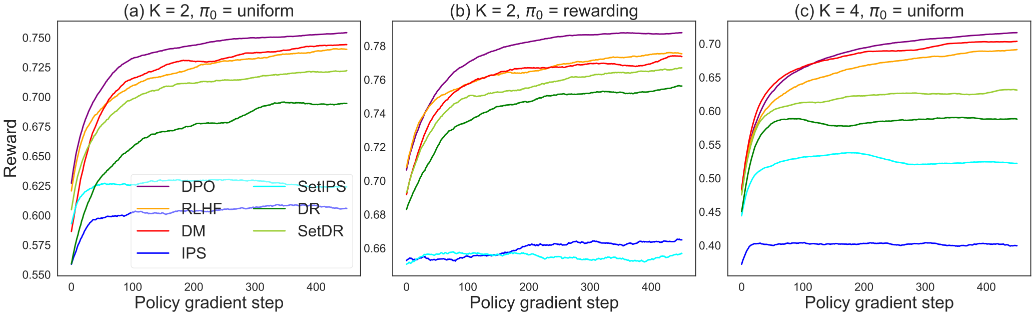

Figure 4: Policy optimization of our estimators, together with RLHF and DPO. The plots are smoothed out by averaging over a window of 50 steps.

Ralph Allan Bradley and Milton Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

Tom Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, 2020.

Miroslav Dudik, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Katja Hofmann, Lihong Li, and Filip Radlinski. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 2016.

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.

Edward Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, 2017.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

Lihong Li, Wei Chu, John Langford, and Robert Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

Lihong Li, Shunbao Chen, Jim Kleban, and Ankur Gupta. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, 2015.

Shuai Li, Yasin Abbasi-Yadkori, Branislav Kveton, S. Muthukrishnan, Vishwa Vinay, and Zheng Wen. Offline evaluation of ranking policies with click models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1685–1694, 2018.

Robert Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Dover Publications, 2005.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Meta. Introducing meta llama 3: The most capable openly available llm to date. 2024. URL https://ai.meta.com/blog/meta-llama-3/.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, 2022.

Robin Lewis Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, 2023.

James Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning with preference-based active queries. *Advances in Neural Information Processing Systems*, 36, 2024.

Alex Strehl, John Langford, Lihong Li, and Sham Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems 23*, 2010.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. 2022. URL https://arxiv.org/abs/2212.09741.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 814–823, 2015.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems 30*, 2017.

Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, New York, NY, 2009.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7b: Improving llm helpfulness & harmlessness with rlaif, November 2023a.

Banghua Zhu, Jiantao Jiao, and Michael Jordan. Principled reinforcement learning with human feedback from pairwise or K-wise comparisons. *CoRR*, abs/2301.11270, 2023b. URL https://arxiv.org/abs/2301.11270.

A. Technical Lemmas

Lemma 1. *Let $\hat{V}(\pi)$ be defined as in (5). Then $\mathbb{E}\left[\hat{V}(\pi)\right] = V(\pi)$.*

Proof. The proof follows from a sequence of identities,

$$\begin{aligned}
\mathbb{E}\left[\hat{V}(\pi)\right] &= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{A_t \sim \pi(\cdot \mid x_t),\, A_{t,*} \sim \cdot \mid x_t, A_t}\left[\mathbb{1}\{a_{t,1} = a_{t,*,1}\}\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{A_t \sim \pi(\cdot \mid x_t)}\left[\mathbb{P}\left(a_{t,1} = a_{t,*,1} \mid x_t, A_t\right)\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{A_t \sim \pi(\cdot \mid x_t)}\left[r(x_t, A_t)\right] = V(\pi)\,.
\end{aligned}$$

This completes the proof. □

Lemma 2. *Let $\hat{V}_{\text{IPS}}(\pi)$ be defined as in (8). Then $\mathbb{E}\left[\hat{V}_{\text{IPS}}(\pi)\right] = V(\pi)$.*

Proof. The proof is similar to Lemma 1, with the addition of propensity scores. In particular,

$$\begin{aligned}
\mathbb{E}\left[\hat{V}_{\text{IPS}}(\pi)\right] &= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{A_t \sim \pi_0(\cdot \mid x_t),\, A_{t,*} \sim \cdot \mid x_t, A_t}\left[\frac{\pi(A_t \mid x_t)}{\pi_0(A_t \mid x_t)}\mathbb{1}\{a_{t,1} = a_{t,*,1}\}\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{A_t \sim \pi_0(\cdot \mid x_t)}\left[\frac{\pi(A_t \mid x_t)}{\pi_0(A_t \mid x_t)}\mathbb{P}\left(a_{t,1} = a_{t,*,1} \mid x_t, A_t\right)\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{A_t \sim \pi(\cdot \mid x_t)}\left[\mathbb{P}\left(a_{t,1} = a_{t,*,1} \mid x_t, A_t\right)\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{A_t \sim \pi(\cdot \mid x_t)}\left[r(x_t, A_t)\right] = V(\pi)\,.
\end{aligned}$$

This completes the proof. □

Lemma 3. *Let $\hat{V}_{\text{DR}}(\pi)$ be defined as in (9). Then $\mathbb{E}\left[\hat{V}_{\text{DR}}(\pi)\right] = V(\pi)$ when the DM is unbiased or the propensities in the IPS estimator are correctly specified.*

Proof. When the reward model is correctly specified,

$$\mathbb{E}\left[\mathbb{1}\{a_{t,1} = a_{t,*,1}\} - r(x_t, A_t; \hat{w}) \mid x_t, A_t\right] = r(x_t, A_t) - r(x_t, A_t; \hat{w}) = 0\,,$$

and the first term in (9) vanishes. This proves the first claim.

When the propensities are correctly specified,

$$\mathbb{E}_{A_t \sim \pi_0(\cdot \mid x_t)}\left[\frac{\pi(A_t \mid x_t)}{\pi_0(A_t \mid x_t)} r(x_t, A_t; \hat{w})\right] = \mathbb{E}_{A_t \sim \pi(\cdot \mid x_t)}\left[r(x_t, A_t; \hat{w})\right]\,.$$

In this case, $r(x_t, A_t; \hat{w})$ and $\hat{V}_{\text{DM}}(\pi)$ cancel out. This proves the second claim. □

Lemma 4. *Let $\hat{V}_{\text{SET}}(\pi)$ be defined as in (11). Then $\mathbb{E}\left[\hat{V}_{\text{SET}}(\pi)\right] = V(\pi)$.*

Proof. The proof follows from introducing a conditional expectation,

$$\begin{aligned}\mathbb{E}\left[\hat{V}_{\text{SET}}(\pi)\right] &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{S_t\sim\nu(\cdot\mid x_t),\, A_{t,*}\sim\cdot\mid x_t,S_t}\left[\pi(a_{t,*,1}\mid S_t,x_t)\right] \\ &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{S_t\sim\nu(\cdot\mid x_t)}\left[\mathbb{E}\left[\pi(a_{t,*,1}\mid S_t,x_t)\,\middle|\,x_t,S_t\right]\right] = V(\pi)\,.\end{aligned}$$

This completes the proof. □

Lemma 5. *Let $\hat{V}_{\text{SET-IPS}}(\pi)$ be defined as in (12). Then $\mathbb{E}\left[\hat{V}_{\text{SET-IPS}}(\pi)\right] = V(\pi)$.*

Proof. The proof is similar to Lemma 4, with the addition of propensity scores. In particular,

$$\begin{aligned}\mathbb{E}\left[\hat{V}_{\text{SET-IPS}}(\pi)\right] &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{S_t\sim\nu_0(\cdot\mid x_t),\, A_{t,*}\sim\cdot\mid x_t,S_t}\left[\frac{\nu(S_t\mid x_t)}{\nu_0(S_t\mid x_t)}\pi(a_{t,*,1}\mid S_t,x_t)\right] \\ &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{S_t\sim\nu_0(\cdot\mid x_t)}\left[\frac{\nu(S_t\mid x_t)}{\nu_0(S_t\mid x_t)}\mathbb{E}\left[\pi(a_{t,*,1}\mid S_t,x_t)\,\middle|\,x_t,S_t\right]\right] \\ &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{S_t\sim\nu(\cdot\mid x_t)}\left[\mathbb{E}\left[\pi(a_{t,*,1}\mid S_t,x_t)\,\middle|\,x_t,S_t\right]\right] = V(\pi)\,.\end{aligned}$$

This completes the proof. □

Lemma 6. *Consider $\hat{V}_{\text{SET-DR}}(\pi)$ in (13). Then $\mathbb{E}\left[\hat{V}_{\text{SET-DR}}(\pi)\right] = V(\pi)$ when the DM is unbiased or the propensities in the IPS estimator are correctly specified.*

Proof. We start by noting that

$$\mathbb{E}\left[\pi(a_{t,*,1}\mid S_t,x_t)\,\middle|\,x_t,S_t\right] = \sum_{A\in\Pi(S_t)}\frac{\pi(A\mid x_t)}{\nu(S_t\mid x_t)}r(x_t,A)$$

and

$$r(x_t,S_t;\pi,\hat{w}) = \sum_{A\in\Pi(S_t)}\frac{\pi(A\mid x_t)}{\nu(S_t\mid x_t)}r(x_t,A;\hat{w})\,.$$

Therefore, when the reward model is correctly specified,

$$\mathbb{E}\left[\pi(a_{t,*,1}\mid S_t,x_t) - r(x_t,S_t;\pi,\hat{w})\,\middle|\,x_t,S_t\right] = 0\,,$$

and the first term in (13) vanishes. This proves the first claim.

When the propensities are correctly specified,

$$\begin{aligned}\mathbb{E}_{S_t\sim\nu_0(\cdot\mid x_t)}\left[\frac{\nu(S_t\mid x_t)}{\nu_0(S_t\mid x_t)}r(x_t,S_t;\pi,\hat{w})\right] &= \mathbb{E}_{S_t\sim\nu(\cdot\mid x_t)}\left[r(x_t,S_t;\pi,\hat{w})\right] \\ &= \mathbb{E}_{S_t\sim\nu(\cdot\mid x_t)}\left[\sum_{A\in\Pi(S_t)}\frac{\pi(A\mid x_t)}{\nu(S_t\mid x_t)}r(x_t,A;\hat{w})\right] \\ &= \mathbb{E}_{A\sim\pi(\cdot\mid x_t)}\left[r(x_t,A;\hat{w})\right]\,.\end{aligned}$$

In this case, $r(x_t,S_t;\pi,\hat{w})$ and $\hat{V}_{\text{DM}}(\pi)$ cancel out. This proves the second claim. □

B. Gradients

The gradient of the regularizer is derived as follows. First, we fix interaction $t \in [n]$. Since t is fixed, x_t is fixed, and thus we can write $\pi(\cdot; \theta)$ instead $\pi(\cdot; x_t, \theta)$. Then, using basic algebra, we get

$$\begin{aligned}
&\nabla d(\pi(\cdot; \theta), \pi_0) \\
&= \sum_{a \in \mathcal{A}} \nabla[\pi(a; \theta)(\log \pi(a; \theta) - \log \pi_0(a))] \\
&= \sum_{a \in \mathcal{A}} [\nabla \pi(a; \theta)] \log \pi(a; \theta) + \pi(a; \theta) \nabla \log \pi(a; \theta) - [\nabla \pi(a; \theta)] \log \pi_0(a) \\
&= \sum_{a \in \mathcal{A}} \pi(a; \theta)[\nabla \log \pi(a; \theta)][1 + \log \pi(a; \theta) - \log \pi_0(a)] \,.
\end{aligned}$$

This implies that for $a \sim \pi(\cdot; \theta)$,

$$\nabla d(\pi(\cdot; \theta), \pi_0) = \mathbb{E}\left[(\nabla \log \pi(a; \theta))(1 + \log \pi(a; \theta) - \log \pi_0(a))\right] \,.$$

Therefore, $(\nabla \log \pi(a; \theta))(1 + \log \pi(a; \theta) - \log \pi_0(a))$ is an unbiased single-sample estimate of the gradient.

The DM gradient is computed as follows. We sample $\tilde{A}_t \sim \pi(\cdot \mid x_t; \theta)$ and then use

$$\frac{1}{n} \sum_{t=1}^{n} \nabla \log \pi(\tilde{A}_t \mid x_t; \theta) \, r(x_t, \tilde{A}_t; \hat{w}) \,. \tag{19}$$

This is an unbiased single-sample estimate of $\nabla \hat{V}_{\text{DM}}(\pi)$.

The IPS gradient is computed directly as

$$\nabla \hat{V}_{\text{IPS}}(\pi) = \frac{1}{n} \sum_{t=1}^{n} \frac{\nabla \pi(A_t \mid x_t; \theta)}{\pi_0(A_t \mid x_t)} \mathbb{1}\{a_{t,1} = a_{t,*,1}\} \,. \tag{20}$$

Combining the above, the DR gradient can be computed as follows. We sample $\tilde{A}_t \sim \pi(\cdot \mid x_t; \theta)$ and then use

$$\frac{1}{n} \sum_{t=1}^{n} \frac{\nabla \pi(A_t \mid x_t; \theta)}{\pi_0(A_t \mid x_t)} (\mathbb{1}\{a_{t,1} = a_{t,*,1}\} - r(x_t, A_t; \hat{w})) + \nabla \hat{V}_{\text{DM}}(\pi) \,. \tag{21}$$

This is an unbiased single-sample estimate of $\nabla \hat{V}_{\text{DR}}(\pi)$.