Proceedings of Machine Learning Research 333:644–687, 2026 Conference on Health, Inference, and Learning (CHIL) 2026

CANDOR: Counterfactual ANnotated DOubly Robust Off-Policy Evaluation

Aishwarya Mandyam am2@stanford.edu
Stanford University

Shengpu Tang shengpu.tang@emory.edu
Emory University

Jiayu Yao jy3491@columbia.edu
Columbia University

Jenna Wiens wiensj@umich.edu
University of Michigan

Barbara E. Engelhardt barbarae@stanford.edu
Stanford University, The Gladstone Institutes

Abstract

Off-policy evaluation (OPE) is critical for applying contextual bandit algorithms to high-stakes decision-making settings such as healthcare, where new treatment policies must be evaluated prior to deployment. Unfortunately, OPE techniques are inherently limited by the breadth of the available data, which may not be sufficient to evaluate the performance of a new policy. Recent work attempts to improve dataset coverage by adding expert-annotated counterfactual samples. However, such annotations are often imperfect and can lead to worse estimator performance than using no annotations at all. To better leverage imperfect annotations, we propose a family of OPE estimators grounded in the doubly robust (DR) framework, which combines importance sampling (IS) with a reward model (direct method, DM) for better statistical guarantees. We study three ways of incorporating counterfactual annotations. Under mild assumptions, we prove that using annotations within just the DM component yields the most desirable theoretical results. Experiments on multiple healthcare tasks, including real-world electronic health records (EHR) data, show that this strategy is most robust under misspecified reward models and inaccurate annotations. By addressing the challenges posed by imperfect annotations, this work broadens the applicability of OPE methods and facilitates safer deployment of decision-making policies in healthcare.

Data and Code Availability We consider three synthetic or semi-synthetic domains (Multi-Armed Bandit (Tang and Wiens, 2023), HeartSteps (Mandyam et al., 2024), Sepsis (Oberst and Sontag, 2019)) and data from the MIMIC-IV dataset (Johnson et al., 2020). All data sources are publicly available. We include the code for our experiments on Github.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Contextual bandit methods have been widely applied to learn optimal sequential decision-making policies across several domains. In healthcare, they have been used for high-stakes tasks that directly affect patient outcomes, such as personalized treatment design and adaptive clinical decision support (Yao et al., 2021). Given the nature of these applications, it is critical for practitioners to assess the performance of a new policy prior to deployment.

A common approach to assess policy performance is off-policy evaluation (OPE) (Sutton and Barto (2018), Chapter 5), which estimates the value of a new policy (the target policy) using a behavior dataset retrospectively collected from a different policy. For example, a hospital may wish to evaluate a new potassium repletion policy that adjusts intravenous potassium dosing, using only retrospective data from patients treated under the hospital's existing treatment protocol. OPE methods estimate how the new policy would perform (e.g., its treatment effect on patient outcomes) without ever deploying it on real patients. This makes these methods crucial for safe policy deployment.

However, OPE is inherently limited by the quality and coverage of the behavior dataset. If the behavior dataset is collected from a hospital that has never prescribed a recently developed drug, then no OPE method can reliably evaluate a policy that recommends the new drug since there are no observed outcomes for that treatment.

To address this issue, Tang and Wiens (2023) proposed an importance sampling (IS)-based OPE estimator called C-IS (referred to in this work as IS^+). IS^+ improves dataset coverage by augmenting the behavior dataset with expert-sourced annotations (i.e., predicted rewards) of counterfactual actions. However, IS^+ relies on the strong assumption that counterfactual annotations are **free of errors**. In practice, even expert-generated annotations are prone to errors, and their accuracy is typically unknown a priori. For example, clinicians may systematically over- or underestimate the effects of an uncommon treatment, or disagree due to limited clinical experience. Determining the optimal way to incorporate annotations of unknown quality into an OPE estimator remains an open challenge.

In light of this shortcoming, we propose a family of OPE estimators based on the standard doubly robust (DR) estimator (Dudík et al., 2014). Compared to IS estimators, DR estimators offer provable reductions in variance while remaining unbiased. However, incorporating potentially imperfect counterfactual annotations into DR estimators without sacrificing these desirable theoretical properties is nontrivial.

In this work, we study three ways of modifying DR estimators to include counterfactual annotations, each of which impacts the estimator performance in a different way. We perform a theoretical analysis of the bias-variance trade-off and an empirical robustness analysis of the estimator performance. Under **imperfect annotations**, we identify one estimator that effectively leverages counterfactual annotations to improve coverage without compounding error from the annotations. In contrast, the other two estimators accumulate error in proportion to the magnitude of error in the annotations, resulting in policy value estimates that are worse than simply ignoring the annotations altogether. In summary, our contributions are the following:

- **We propose a family of DR-inspired OPE estimators** that can leverage counterfactual annotations. We theoretically analyze our proposed estimators, showing that the manner in which annotations are incorporated into the estimator substantially affects estimator performance (Section 3). Among these estimators, one is robust to both reward model misspecification and inaccurate annotations.
- **We evaluate our estimators using one synthetic and two semi-synthetic healthcare environments, as well as a real-world EHR dataset.** The synthetic and semi-synthetic settings are used to empirically validate our theoretical insights, while the EHR dataset demonstrates the practical utility of our method in a high-stakes clinical decision-making setting (Section 4).
- **We conduct robustness analyses** of the proposed OPE estimators under realistic clinical conditions, where both annotation quality and reward model accuracy are unknown. We find that one of the proposed estimators is more robust than the others, making it the most suitable choice for reliable policy evaluation in healthcare settings (Section 4).

2. Background

We consider a contextual bandit setting defined by $(\mathcal{S}, \mathcal{A}, R, d_0)$, where $\mathcal{S}$ is the discrete context space, $\mathcal{A}$ is the discrete action space, $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is the reward function, and d_0 is the context distribution. We observe a behavior dataset $D = \{(s_i, a_i, r_i)\}_{i=1}^N$ collected under a behavior policy π_b, with samples drawn i.i.d. from an underlying data-generating distribution $\mathcal{D}$. We aim to estimate the value of a target policy π_e, defined as $v(\pi_e) = \mathbb{E}_{s \sim d_0, a \sim \pi_e(\cdot|s), r \sim R(s,a)}[r]$.

2.1. Off-Policy Evaluation

We briefly review three classes of OPE methods for contextual bandits. Importance sampling (IS) (Horvitz and Thompson, 1952; Precup et al., 2000) assigns an inverse propensity score (IPS), $\rho_s(a) = \frac{\pi_e(a|s)}{\pi_b(a|s)}$, to each sample (s_i, a_i, r_i) in the behavior dataset and estimates the target policy value as,

$$\hat{V}^{\text{IS}} = \frac{1}{N} \sum_{i=1}^{N} \rho_{s_i}(a_i) r_i.$$

Following prior work, we assume that the IPS ρ is known (Farajtabar et al., 2018; Thomas and Brunskill, 2016).

Assumption 1 (Common support)**.** $\pi_e(a|s) > 0 \to \pi_b(a|s) > 0$.

Under a standard support assumption (Assumption 1), IS yields an unbiased estimate of the target policy value, $v(\pi_e)$ (Precup et al., 2000). Under the same assumption, the variance of the estimator is

$$\begin{aligned} N \cdot \mathbb{V}[\hat{V}^{\text{IS}}] = {} & \mathbb{V}_{s\sim d_0}[v^{\pi_e}(s)] \\ & + \mathbb{E}_{s\sim d_0}[\mathbb{V}_{a\sim\pi_b(\cdot|s)}[\rho_s(a)\bar{R}(s,a)]] \\ & + \mathbb{E}_{s\sim d_0}[\mathbb{E}_{a\sim\pi_b(\cdot|s)}[\rho_s(a)^2\sigma_R(s,a)^2]], \end{aligned} \tag{1}$$

where $\bar{R}(s,a) = \mathbb{E}[R(s,a)]$ and $\sigma_R(s,a)^2 = \mathbb{V}[R(s,a)]$ are the mean and variance of the reward distribution, respectively (Tang and Wiens, 2023).

Another approach to OPE is the direct method (DM) (Li et al., 2010; Beygelzimer and Langford, 2009; van Seijen et al., 2009; Harutyunyan et al., 2016; Le et al., 2019; Voloshin et al., 2021). DM first uses the behavior dataset to fit a reward model, $\hat{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, to predict the conditional mean reward. It then uses $\hat{R}$ to directly compute the target policy value as

$$\hat{V}^{\text{DM}} = \sum_s d_0(s) \sum_a \pi_e(a|s)\hat{R}(s,a).$$

The reward model $\hat{R}$ may range in complexity from simple regression models to neural networks. When the reward model is realizable and the behavior dataset provides full coverage, the DM estimator has zero bias and favorable variance. In practice, DM estimators often have lower variance than IS (Dudik et al., 2011) when the reward model is accurate.

The last category of OPE approaches consists of doubly robust (DR) methods (Dudik et al., 2011; Dudík et al., 2014; Farajtabar et al., 2018; Jiang and Li, 2016). These methods are termed "doubly robust" because they maintain strong theoretical guarantees when either the IPS ratio ρ, or the estimated reward function $\hat{R}$, is inaccurate, thereby providing robustness to both sources of error. In this work, we aim to design an OPE estimator that is robust to misspecified reward models and inaccurate annotations.

The standard DR estimator is defined as

$$\hat{V}^{\text{DR}} = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\hat{R}(s_i, \pi_e)}_{\text{DM part}} + \underbrace{\rho_{s_i}(a_i)(r_i - \hat{R}(s_i, a_i))}_{\text{IS part}}, \tag{2}$$

where $\hat{R}(s, \pi_e) = \sum_{a\in\mathcal{A}} \pi_e(a|s)\hat{R}(s,a)$ is the estimated value of state s under the target policy π_e using the reward model $\hat{R}$. We refer to the first and second term in Eqn (2) as the *DM part* and the *IS part*, respectively.

Under the common support assumption (Assumption 1), the DR estimator produces an unbiased estimate of $v(\pi_e)$ (if the reward model and OPE estimate are learned using independent splits of the data). DR methods often exhibit lower variance than IS-based methods; the variance can be written as

$$\begin{aligned} N\cdot\mathbb{V}[\hat{V}^{\text{DR}}] = {} & \mathbb{V}_{s\sim d_0}[v^{\pi_e}(s)] \\ & + \mathbb{E}_{s\sim d_0}\left[\mathbb{V}_{a\sim\pi_b(\cdot|s)}\left[\rho_s(a)(\bar{R}(s,a) - \hat{R}(s,a))\right]\right] \\ & + \mathbb{E}_{s\sim d_0}\left[\mathbb{E}_{a\sim\pi_b(\cdot|s)}\left[\rho_s(a)^2\sigma_R(s,a)^2\right]\right]. \end{aligned}$$

Comparing to Eqn (1), the variance reduction relative to the IS estimator rests in the second term: the IPS ρ is scaled by the residual $\bar{R}(s,a) - \hat{R}(s,a)$, while the standard IS estimator scales ρ using $\bar{R}(s,a)$ alone. When the reward model $\hat{R}$ is well-specified, this residual approaches 0, shrinking the impact of the IPS term and reducing overall variance.

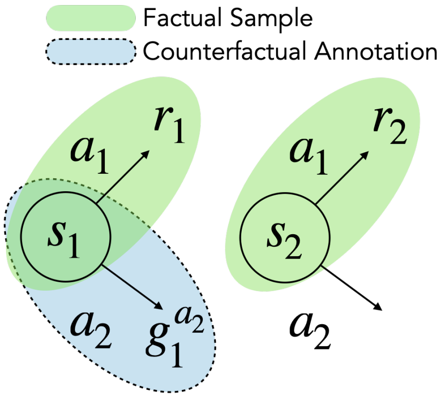


Figure 1: **A counterfactual-annotated dataset with two contexts and two actions.** Two factual samples, (s_1, a_1, r_1) (left) and (s_2, a_1, r_2) (right), are observed. The left sample has one counterfactual annotation, $(s_1, a_2, g_1^{a_2})$, while the right one has none.

2.2. Counterfactual Annotations

In our work, we consider incorporating counterfactual annotations to increase data coverage. Specifically, we assume that each factual sample (s_i, a_i) in the behavior dataset D is associated with a set of counterfactual annotations $\mathbf{g}_i = \{g_i^{\tilde{a}} \mid \tilde{a} \in \mathcal{A} \setminus \{a_i\}\}$. Note that $\mathbf{g}_i$ may be empty. We assume that the annotation of the counterfactual action $\tilde{a}$ is drawn from a distribution $G : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, $g_i^{\tilde{a}} \sim G(s_i, \tilde{a})$.

Because counterfactual annotations are often expensive to obtain, we do not assume full annotation coverage. We refer to the dataset that augments factual samples and counterfactual annotations as the *counterfactual-annotated* dataset, denoted by $D^+ \sim \mathcal{D}^+$, where $\mathcal{D}^+$ is the distribution over such datasets. Figure 1 illustrates a simple example of a counterfactual-annotated dataset with partial coverage, where two factual samples are observed and only one has an associated counterfactual annotation.

In Section 3.2, we study three scenarios for the annotation distribution G (perfect, biased, or noisy annotations). Additionally, for ease of notation, we use c_i^a to refer to either the reward or the counterfactual annotation of the factual sample (s_i, a_i), i.e., $c_i^a = r_i$ when $a = a_i$ and $c_i^a = g_i^a$ when $a \neq a_i$.

2.3. The IS$^+$ Estimator

Naively incorporating counterfactual annotations into IS estimators introduces bias, as the context distribution in D^+ may differ from that of the behavior dataset D. To address this, Tang and Wiens (2023) introduced IS$^+$, which reweights factual and counterfactual samples to align these two distributions.

Specifically, let $\{w_i^a\}$ be a set of user-defined weights for the i-th factual sample (s_i, a_i) and its associated counterfactual annotations $\{(s_i, g_i^a)\}$, with the constraint $\sum_{a \in \mathcal{A}} w_i^a = 1$ for each sample i (i.e., the weights for all actions associated with a factual sample s_i must add up to 1). We set $w_i^{\tilde{a}} = 0$ if the annotation of the counterfactual action $\tilde{a}$ for the i-th sample's state is unavailable. Define the augmented IPS ratio as $\rho_s^+(a) = \frac{\pi_e(a|s)}{\pi_b^+(a|s)}$ and the augmented behavior policy as,

$$\pi_b^+(a|s) = \bar{W}(a|s,a)\pi_b(a|s) + \sum_{\check{a} \in \mathcal{A} \setminus \{a\}} \bar{W}(a|s,\check{a})\pi_b(\check{a}|s),$$

where $\bar{W}(\tilde{a}|s,a) = \mathbb{E}[w^{\tilde{a}}]$ is the average weight of action $\tilde{a}$ for the factual context-action pair (s, a). The IS$^+$ estimator is defined as,

$$\hat{V}^{\text{IS}^+} = \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in \mathcal{A}} w_i^a \rho_{s_i}^+(a) c_i^a.$$

3. Methods

When the behavior dataset has limited coverage, IS estimators are known to have high variance (Jiang and Li, 2016). In contrast, DM estimators have high bias when the reward model is misspecified. We therefore explore how to introduce counterfactual annotations into a DR estimator, which retains beneficial theoretical properties even when the reward model is misspecified, a common challenge in healthcare settings. While DR estimators are well-understood in a setting with only factual samples, we seek to incorporate counterfactual annotations such that we can improve behavior dataset coverage.

A naive approach is to apply a standard DR estimator directly to the counterfactual-annotated dataset D^+, treating the counterfactual annotations as additional samples. However, as we discuss in Section C, this approach can produce arbitrarily biased estimates of $v(\pi_e)$ because it alters the context distribution of the behavior dataset, an issue that arises regardless of annotation quality. As such, we explore three new estimators that incorporate counterfactual annotations into the DR framework.

3.1. Proposed DR Estimators with Counterfactual Annotations

The standard DR estimator (Eqn (2)) can be broken down into two components: the direct method (DM) part and the importance sampling (IS) part. We observe that counterfactual annotations can be incorporated independently into either component, or into both. Based on this insight, we propose three new DR-inspired estimators that can leverage counterfactual annotations (Eqn (3) to (5)).

$$\hat{V}^{\text{DM}^+\text{-IS}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{R}^+(s_i, \pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i, a_i)) \Big) \tag{3}$$

$$\hat{V}^{\text{DM-IS}^+} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{R}(s_i, \pi_e) + \sum_{a \in A} w_i^a \rho_{s_i}^+(a)(c_i^a - \hat{R}(s_i, a)) \Big) \tag{4}$$

$$\hat{V}^{\text{DM}^+\text{-IS}^+} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{R}^+(s_i, \pi_e) + \sum_{a \in A} w_i^a \rho_{s_i}^+(a)(c_i^a - \hat{R}^+(s_i, a)) \Big). \tag{5}$$

Here, $\hat{R}^+(s, a)$ is the reward function estimate learned using the counterfactual-annotated dataset D^+ (further discussion in Section E), and

$\hat{R}^+(s,\pi_e) = \sum_{a\in\mathcal{A}} \pi_e(a|s)\hat{R}^+(s,a)$. The first estimator, **DM$^+$-IS** (Eqn (3)), uses the counterfactual-annotated dataset to estimate the reward model and combines it with standard IS. The second estimator, **DM-IS$^+$** (Eqn (4)), instead uses counterfactual annotations to augment the IS part (as in IS$^+$) and combines it with a standard DM estimator. The third estimator, **DM$^+$-IS$^+$** (Eqn (5)) uses counterfactual annotations in both the DM and IS parts.

3.2. Theoretical Analyses under Imperfect Annotations

In our problem setting, there are three possible sources of error: incorrect estimates of the behavior policy π_b, a misspecified reward model, and imperfect (biased or noisy) annotations. Much like prior work (Farajtabar et al., 2018; Thomas and Brunskill, 2016), we assume that the IPS ratio ρ is known, and instead focus on identifying an OPE estimator that is robust to the last two sources of error. Accordingly, we analyze our proposed estimators under reward model misspecification and imperfect annotations, characterizing how these errors affect their performance. This analysis also provides insight into the robustness of the estimators, which we study empirically in Section 4.

The novelty of our theoretical results is two-fold. First, prior work on DR estimators typically treats the reward model error as fixed. For example, Dudik et al. (2011) assumed that $\hat{R}(s,a) = \bar{R}(s,a) + \epsilon(s,a)$ without modeling how $\epsilon(s,a)$ depends on the data used to estimate $\hat{R}$. In contrast, we treat $\hat{R}$ as a random function whose distribution is induced by the dataset used to fit the reward model. Specifically, the reward model, $\hat{R}$, is estimated from a separate dataset, which we refer to as $D_{\hat{R}} \sim \mathcal{D}$ or $D_{\hat{R}^+} \sim \mathcal{D}^+$, depending on if counterfactual annotations are included. Importantly, we do not assume that the reward model class is well specified.

Second, we derive the bias and variance of our proposed DR estimators under imperfect annotations, a setting that more closely reflects practical applications. We formalize the quality of counterfactual annotations through the following three assumptions.

Assumption 2 (Perfect annotations). $\mathbb{E}_{g^a\sim G(s,a)}[g^a] = \bar{R}(s,a)$, $\mathbb{V}_{g^a\sim G(s,a)}[g^a] = \sigma_R^2(s,a)$.

Assumption 3 (Biased annotations). $\mathbb{E}_{g^a\sim G(s,a)}[g^a] = \bar{R}(s,a) + \epsilon_G(s,a),\ \epsilon_G(s,a) \neq 0$.

Assumption 4 (Noisy annotations). $\mathbb{V}_{g^a\sim G(s,a)}[g^a] = \sigma_R(s,a)^2 + \Delta_G(s,a),\ \Delta_G(s,a) > 0$.

In Assumptions 3 and 4, annotation bias and variance, quantified by the terms $\epsilon_G(s,a)$ and $\Delta_G(s,a)$ respectively, are used to study the effect of biased and noisy (i.e., higher variance) annotations. Assumption 2 represents the idealized setting of perfect annotations and is mutually exclusive with Assumptions 3 and 4. Finally, following Tang and Wiens (2023), we adopt a relaxed data coverage assumption (Assumption 1) to analyze the properties of DM-IS$^+$ and DM$^+$-IS$^+$.

Assumption 5 (Common support with counterfactual annotations). $\pi_e(a|s) > 0 \to \pi_b^+(a|s) > 0$.

With these assumptions in place, we now summarize the main theoretical implications for the bias and variance of the proposed estimators under both idealized and realistic annotation regimes. First, we consider the setting with perfect annotations (Assumption 2). Under appropriate coverage assumptions (Assumption 1 or 5), all three proposed estimators are unbiased (Propositions S12, S14 and S17 in Section G). Additionally, when all counterfactual actions are annotated and uniform weights are used (i.e., $w^a = 1/|\mathcal{A}|$), both DM-IS$^+$ and DM$^+$-IS$^+$ are equivalent to the IS$^+$ estimator (Corollaries S20 and S21 in Section H). These results suggest that in the absence of annotation error, counterfactual annotations can be viewed as an additional high-quality dataset and yield unbiased OPE estimates.

We next analyze the more realistic setting in which the annotations are imperfect (i.e., Assumption 2 is violated). We begin by characterizing the bias of our proposed estimators. Note that the bias derivations only rely on Assumption 3, which captures annotation bias and do not require Assumption 4, which governs annotation noise.

Proposition 1 (Unbiasedness of DM$^+$-IS under imperfect annotations). *Under common support (Assumption 1) and biased annotations (Assumption 3):* $\mathbb{E}[\hat{V}^{\text{DM}^+\text{-IS}}] = v(\pi_e)$.

Theorem 2 (Bias of DM-IS$^+$ and DM$^+$-IS$^+$ under imperfect annotations). *Under common support (Assumption 5) and biased annotations (Assumption 3),* DM-IS$^+$ *and* DM$^+$-IS$^+$ *have the same expectation:*

$$\mathbb{E}[\hat{V}^{\text{DM-IS}^+}] = \mathbb{E}[\hat{V}^{\text{DM}^+\text{-IS}^+}] \tag{6}$$

$$= v(\pi_e) + \mathbb{E}_{\substack{s\sim d_0\\ a\sim\pi_e(s)}}\left[\left(1 - \frac{\bar{W}(a|s,a)\pi_b(a|s)}{\pi_b^+(a|s)}\right)\epsilon_G(s,a)\right].$$

Remark. Proposition 1 establishes that when ρ is known, DM$^+$-IS is an unbiased estimator of the target policy value $v(\pi_e)$, even with biased annotations. In contrast, Theorem 2 shows that both DM-IS$^+$ and DM$^+$-IS$^+$ will produce biased estimates of $v(\pi_e)$. Note that the last term in Eqn (6) is identical to the expectation derivation for IS$^+$ (Tang and Wiens, 2023).

Having characterized the bias, we now study the variance of our proposed estimators under imperfect annotations. We begin with DM$^+$-IS.

Theorem 3 (Variance of DM$^+$-IS under imperfect annotations)**.** *Under Assumptions 1, 3 and 4,*

$$
\begin{aligned}
N\cdot\mathbb{V}[\hat{V}^{\mathrm{DM^+\text{-}IS}}] &= \mathbb{V}_{s\sim d_0}[v^{\pi_e}(s)] \\
&+ \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b(s)}[\rho_s(a)^2\sigma_R^2(s,a)] \\
&+ \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}\Big[\big(\rho_s(a)^2 - \tfrac{1}{\pi_b(a|s)}\big)\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]\Big] \\
&+ \mathbb{E}_{s\sim d_0}\Big[\mathbb{E}_{a\sim\pi_b}[\rho_s(a)^2\varepsilon_{\hat{R}^+}(s,a)^2] - \varepsilon^{\pi_e}_{\hat{R}^+}(s)^2\Big],
\end{aligned}
$$

where $\varepsilon_{\hat{R}^+}(s,a) = \mathbb{E}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)] - \bar{R}(s,a)$, *and* $\varepsilon^{\pi_e}_{\hat{R}^+}(s) = \mathbb{E}_{a\sim\pi_e}[\varepsilon_{\hat{R}^+}(s,a)]$.

Remark. Theorem 3 characterizes the variance of DM$^+$-IS under imperfect counterfactual annotations, showing that it depends on both the noise and the bias of the counterfactual annotations as reflected in the last two terms. The third term, which depends on $\hat{R}^+$, can dominate the overall variance when annotations are highly noisy. The last term emerges from the systematic estimation error of the reward model due to biased annotations.

We also derive the variance of DM-IS$^+$ and DM$^+$-IS$^+$ under imperfect annotations (Proposition S15 and Proposition S18). Despite their closed-form analytical expressions, two key questions remain difficult to answer analytically: (1) whether the proposed estimators reduce variance relative to the standard DR estimator, and (2) how the three proposed estimators compare to one another in terms of variance. Answering these questions depends critically on how the annotation noise ($\Delta_G(s,a)$) and bias ($\epsilon_G(s,a)$) vary across the state-action space; these quantities are highly application-dependent and resistant to general theoretical analysis. We therefore turn to empirical studies in Section 4.

We summarize our theorems in Appendix Table 3, with full proofs provided in Section G. In short, under perfect annotations, all of our proposed DR estimators are unbiased. Under imperfect annotations, DM$^+$-IS$^+$ and DM-IS$^+$ share the same bias, while DM$^+$-IS remains unbiased. We expect imperfect annotations to increase the variance of all three proposed estimators due to the increased bias and variance of the estimated reward-model. Now, we turn to empirical analysis to understand how these estimators perform in practice.

4. Experiments

Our experiments seek to answer the following questions: **1)** How do imperfect annotations alone empirically affect the performance of the proposed methods? **2)** How do the proposed methods perform under compounding errors from imperfect annotations and misspecified reward models? **(3)** Which estimator is most robust when the quality of the annotations and reward model is unknown?

4.1. Experimental Domains

To answer these questions, we investigate three healthcare-inspired contextual bandit domains with progressively increasing state and action space sizes. The first two domains are semi-synthetic, thus enabling direct control over the data-generating process. This allows us to validate hypotheses derived from our theoretical analyses under controlled conditions. The last domain uses real-world clinical data, where counterfactual annotations are generated using a large language model (LLM), reflecting a more realistic and noisy deployment setting. We additionally verify our proofs in a fully synthetic environment under idealized conditions, with results deferred to Section A.1 and Section A.2.

Heartsteps (Mandyam et al., 2024): This semi-synthetic mobile health simulator models the user's physical activities given mobile interventions based on the Heartsteps study (Klasnja et al., 2019). The context is a three-dimensional vector that includes a treatment effect term and the step count of the previous day. At each decision time, the agent selects between two actions (either *send an intervention* or *do nothing*), and the reward is drawn from a normal distribution with the mean being the square root of the user's observed step count.

Sepsis (Oberst and Sontag, 2019): We adapt the semi-synthetic sepsis simulator used in prior work (originally built for the sequential Markov decision process (MDP) setting) to a contextual bandit setting by interacting with the environment for only one step. The patient context is an 8-dimensional vector that contains information about vitals and ongoing treatments. There are 8 treatment options, and the reward is an indicator function of whether the patient is under treatment and has stable vitals.

MIMIC-IV (Johnson et al., 2020; Goldberger et al., 2000): MIMIC-IV contains electronic health records for over 65,000 admitted patients. We study a subset of patients who received intravenous (IV) potassium repletion. Potassium repletion is a common task in critical care settings; imbalanced potassium levels can have severe side effects including cardiac arrest (Prasad et al., 2022).

We created two splits of the dataset based on whether a patient has renal disease (we refer to these splits as "renal" and "non-renal"). The behavior policy π_b and the target policy π_e are defined by the clinician treatment policies for non-renal and renal patients, respectively. In Appendix Figure S6, we see that the treatment policies for these subgroups are different. Specifically, patients with renal disease are administered lower dosages to account for their impaired kidney function (Shrimanker and Bhattarai, 2025). Our goal is to estimate the value of the target policy using data generated under the behavior policy.

In this setting, the patient context is a 20-dimensional vector that contains information about vitals, administered medications, and static covariates. The actions are five discretized potassium dosage levels. The reward is an indicator function of whether the patient's lab potassium value is within the reference range 2 hours after administering a given dosage. We used linear regression to fit our estimated reward model. Distinct from the prior settings, neither π_b nor π_e is known and both are instead estimated using behavior cloning. While alternatives such as inverse reinforcement learning exist, they are generally more computationally demanding and require additional assumptions, making behavior cloning the more practical choice.

4.2. Generating Counterfactual Annotations and Misspecified Reward Models

For the semi-synthetic domains, to produce perfect counterfactual annotations of state s and counterfactual action $\tilde{a}$, we sample from the true reward model, i.e., $G(s, \tilde{a}) = \mathcal{N}(\bar{R}(s, \tilde{a}), \sigma_R(s, \tilde{a}))$. To produce biased and noisy counterfactual annotation, we sample from $\mathcal{N}(\bar{R}(s, \tilde{a}) + \epsilon_G(s, a), \sigma_R(s, \tilde{a}) + \Delta_G(s, a))$, where $\epsilon_G(s, a)$ and $\Delta_G(s, a)$ refer to the additional bias and variance that compromise the quality of the annotations.

For the MIMIC-IV data, we randomly selected a subset of state-action pairs in the behavior dataset and generated counterfactual annotations for those samples using OpenAI's "o1" model (OpenAI et al., 2024) following prior work (Mandyam et al., 2024). Specifically, "o1" is prompted to predict a patient's blood potassium level after administering a counterfactual dosage of IV potassium. This procedure mimics a setting where counterfactual annotations may be imperfect. Further details regarding the dataset and annotation construction are in Section B.

In addition to imperfect annotations, we study the compounding error of misspecified reward models in the semi-synthetic settings. In our experiments, we create misspecified reward models by either partially observing the state or altering the state representation (Table 1 and Section B).

4.3. Baselines and Metrics

We compare our proposed estimators against two classes of baselines. The first consists of common OPE estimators that do not use counterfactual annotations (IS, DM, and DR). The second includes estimators that leverage counterfactual annotations (IS$^+$ and DM$^+$). DM$^+$ is a direct method estimator that estimates the reward model using the counterfactual-annotated dataset, defined as

$$\hat{V}^{\mathrm{DM}^+} = \sum_s d_0(s) \sum_a \pi_e(a|s) \hat{R}^+(s, a)$$

.

For the semi-synthetic domains, we consider various combinations of stochastic behavior and target policies (details in Section B). Specifically, the behavior policies vary in their coverage of the action space. We present results averaged across these combinations. We estimate the ground truth target policy value using Monte Carlo simulation in the synthetic and semi-synthetic domains, and by averaging observed rewards in the MIMIC-IV dataset. For all settings, we report the root mean squared error (RMSE) of estimated policy values.

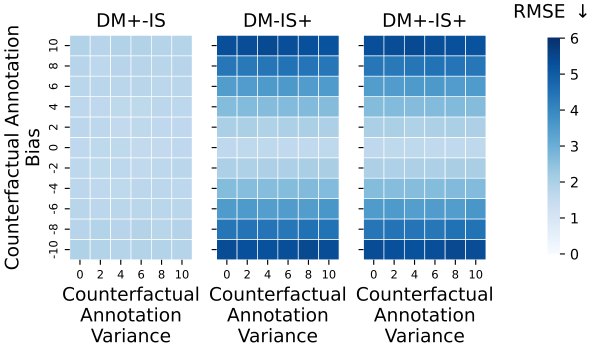


Figure 2: **In Heartsteps, the bias of the counterfactual annotations has a larger impact on RMSE than the variance**. The x, y-axis represents the variance (Δ_G) and the bias (ϵ_G) of the annotations respectively. We report mean RMSE, and error metrics Section A.1. The RMSE remains nearly constant as annotation variance increases, and instead increases proportional to the magnitude of the absolute annotation bias. This trend is particularly noticeable in DM-IS$^+$ and DM$^+$-IS$^+$. The RMSE of DM$^+$-IS is far more consistent regardless of the annotation bias and variance.

4.4. Results

4.4.1. IMPERFECT ANNOTATIONS WITH A WELL-SPECIFIED REWARD MODEL

First, we isolate the impact of imperfect annotations by studying a setting with well-specified reward models, thereby eliminating confounding effects due to reward model misspecification. In the Heartsteps domain, we observe that **annotation bias has a greater effect on the RMSE of the proposed estimators than annotation noise (Figure 2)**. In particular, the RMSE remains nearly unchanged as the annotation variance increases. From observing the results in Figure 2 and Figure S1, we find that this trend holds across all OPE methods that leverage counterfactual annotations and across all synthetic and semi-synthetic domains.

We also note that the effect of imperfect annotations is especially pronounced in methods that incorporate annotations into the "IS" component (e.g., IS$^+$, DM-IS$^+$, DM$^+$-IS$^+$). This observation is consistent with our theoretical results (Proposition 1, Theorem 2), which demonstrate that these estimators are biased when the counterfactual annotations are biased.

Furthermore, Figure 2 and Figure S1 also demonstrate that DM$^+$-IS exhibits greater robustness to annotation quality than baseline approaches. The RMSE of the estimator varies little across different magnitudes of annotation bias and variance. Although our theoretical analysis in Theorem 3 indicates that noisier annotations can, in principle, increase the variance of DM$^+$-IS, we empirically observe that this increase is not substantial.

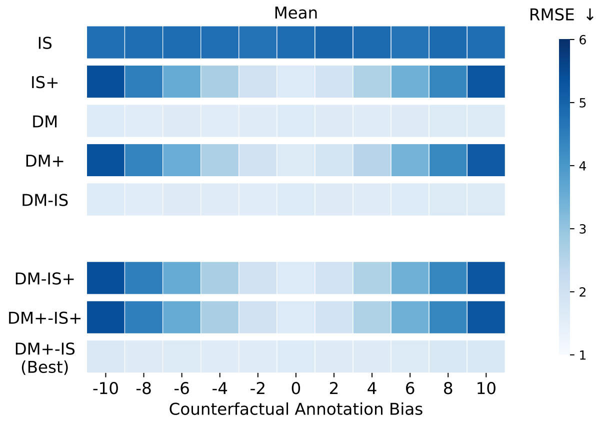
4.4.2. IMPERFECT ANNOTATIONS WITH A MISSPECIFIED REWARD MODEL

In Section 4.4.1, we isolated the effect of imperfect annotations by assuming a well-specified reward model. However, many realistic healthcare settings often involve both reward model misspecification and imperfect annotations. In this section, we demonstrate that **DM$^+$-IS is most robust to these combined sources of error.**

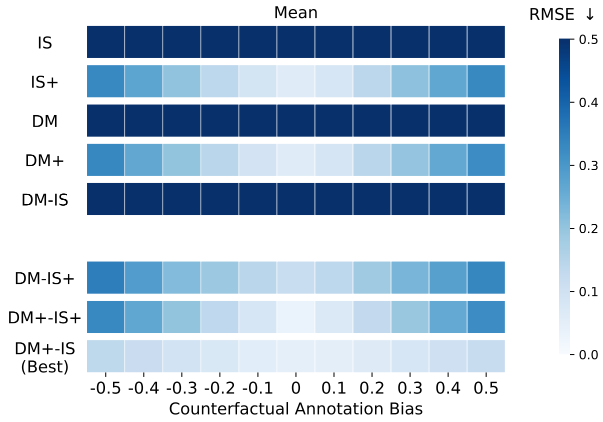
In Section 4.4.1, we observed that the annotation noise has a limited effect on the RMSE of the proposed estimators. We therefore shift our focus to annotation bias. We first examine the Heartsteps and sepsis treatment domains, where the magnitude of bias and noise can be controlled. Figure 3 shows that DM$^+$-IS is most resilient to the compounded errors arising from biased annotations and misspecified reward models. Across both semi-synthetic domains, we see that DM$^+$-IS consistently achieves the lowest RMSE or performs comparably to the best performing baseline. We attribute this robustness to the fact that DM$^+$-IS is the only proposed estimator that is unbiased in the presence of imperfect annotations, and retains beneficial theoretical properties even with a misspecified reward model.

To further demonstrate the robustness of DM$^+$-IS, we compare it to the best-performing OPE estimator under varying degrees of annotation imperfection in the sepsis environment (Figure 4). Here, the best-performing OPE estimator is given access to a well-specified reward model and oracle knowledge of annotation quality; in particular, it uses counterfactual annotations only when they are perfect.

Our results indicate that **DM$^+$-IS remains robust, achieving RMSE within a small margin of the best achievable OPE performance despite these informational disadvantages**. In particular, Δ (the performance gap between DM$^+$-IS and



(a) Heartsteps setting.



(b) Sepsis treatment setting.

Figure 3: **DM$^+$-IS is robust to the joint effects of annotation bias and reward model misspecification.** We report the mean RMSE across the Heartsteps (Figure 3(a)) and Sepsis environments (Figure 3(b)) as a function of annotation bias (x-axis, ϵ_G). Across all settings, DM$^+$-IS consistently matches or exceeds the performance of the best baselines (further results in Section A). Notably, among methods that use counterfactual annotations, DM$^+$-IS exhibits the lowest RMSE across all degrees of annotation bias. When compared to OPE baselines without annotations, DM$^+$-IS achieves a lower or comparable RMSE. Error metrics are provided in Section A.2.

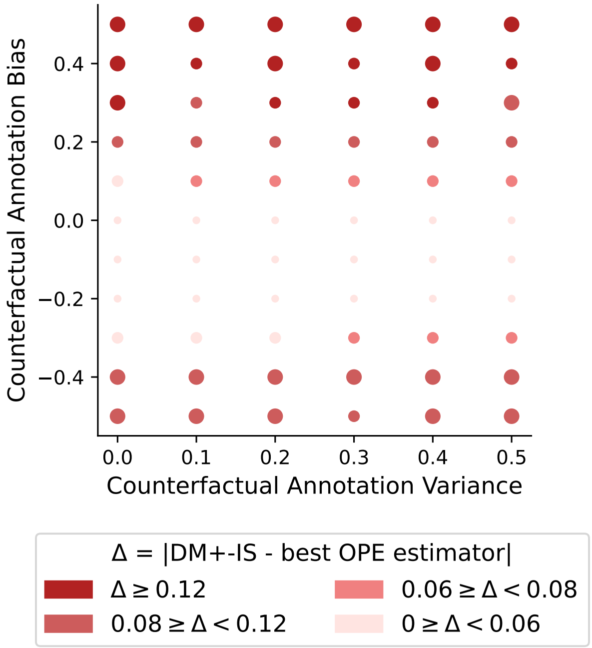
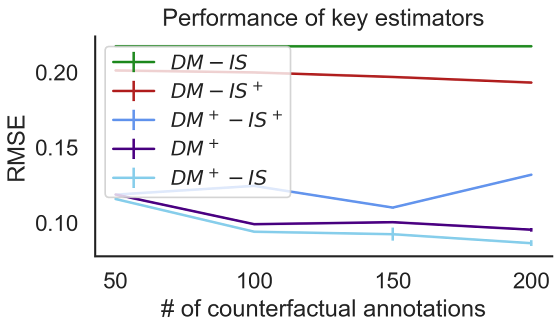


Figure 4: **DM$^+$-IS is a robust estimator across annotations of various qualities.** Here, we study the performance of DM$^+$-IS with a misspecified reward model in comparison to the best-performing OPE method in the sepsis domain. The x, y-axis represents the variance (Δ_G) and the bias (ϵ_G) of the annotations, respectively. We report $\Delta = v^{\text{DM}^+\text{-IS}} - v^{\text{best estimate}}|$ where the best estimate is identified by observing prediction error from all baseline OPE estimators. The color of the dots represents the magnitude of mean Δ, and the size of the dots is proportional to the variance of Δ across 100 iterations. We find that regardless of how imperfect the annotations are, DM$^+$-IS produces estimates with low mean Δ relative to the range of reward in the sepsis environment (5.2). Δ increases in magnitude proportional to the magnitude of the annotation bias, though even in extreme cases, Δ is relatively small.

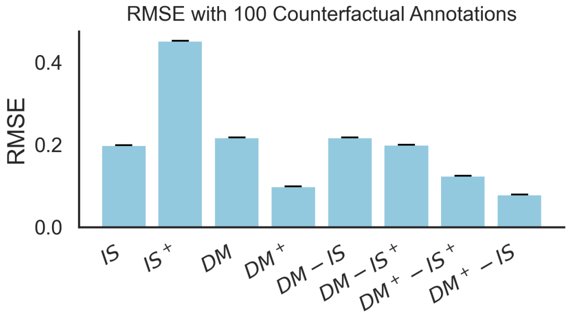
the best-performing OPE method) is small relative to the reward range in the sepsis environment, indicating that DM$^+$-IS produces estimates of $v(\pi_e)$ close to those of the best-performing OPE method.

4.4.3. Robustness analysis of DM$^+$-IS under realistic conditions

In Section 4.4.2, we examined the robustness of the proposed estimators under controlled conditions.



(a) **As the number of counterfactual annotations increases, the performance of DM$^+$-IS initially improves and then plateaus.** Error bars represent 95% confidence intervals. In contrast, other estimators either have higher RMSE or have worse performance as the number of annotations increase.



(b) **DM$^+$-IS outperforms all estimators with 100 counterfactual annotations.** Error bars represent 95% confidence intervals, and DM$^+$-IS outperforms baselines with no overlapping intervals.

Figure 5: **DM$^+$-IS outperforms baselines on MIMIC-IV dataset.** We first study how the estimator performance varies as more annotations are added (Figure 5(a)) and study the RMSE of all estimators given a small number of counterfactual annotations (Figure 5(b)).

However, in many real-world settings, the reward model and annotation quality are typically unknown. To evaluate performance under realistic conditions, we apply our proposed OPE estimators to potassium administration using the MIMIC-IV database.

We first examine how the performance of key OPE estimators varies as the number of counterfactual annotations increases (Figure 5(a)). While the RMSE of DM-IS$^+$ and DM$^+$-IS$^+$ remains high as annotation count increases, **the RMSE of DM$^+$-IS decreases initially and then plateaus with no evidence of degradation.** This pattern suggests that DM$^+$-IS leverages the counterfactual annotations more effectively than baseline estimators. These trends are consistent with those observed in the synthetic or semi-synthetic experiments (Figures 2 and 3).

We next examine all OPE estimators under a limited annotation budget (100 counterfactual annotations). We find that **DM$^+$-IS achieves the lowest RMSE across all baselines (Figure 5(b))**. Notably, IS$^+$ exhibits the highest error, which suggests that the counterfactual annotations are imperfect. The relatively small performance gap between DM$^+$ and DM$^+$-IS implies that the estimated reward model is reasonably accurate in this setting. In summary, these results demonstrate that DM$^+$-IS is well suited to realistic conditions in which the reward model and annotation quality are unknown.

5. Conclusion

In this work, we address the open problem of incorporating imperfect counterfactual annotations into an OPE estimator. This is a research question of growing importance: as LLMs become increasingly capable of generating realistic synthetic samples, a broader challenge emerges of how to integrate such samples into statistical estimators without inheriting data biases. We systematically study a range of design choices for integrating annotations into a DR-style estimators. We find that imperfect counterfactual annotations are most beneficial when incorporated into just the DM part of such an estimator. Through theoretical and empirical analyses, we find that the performance of an OPE estimator relies most on two critical factors: (1) whether the reward model is well-specified, and (2) the annotation quality. We conclude that under the most realistic conditions, where the reward model and annotation quality are unknown, our DM$^+$-IS estimator is most robust in comparison to baseline approaches.

Overall, our approach relaxes restrictive assumptions about annotation quality and facilitates more reliable use of bandit algorithms in high-stakes applications through improved off-policy evaluation.

Limitations and Future Work. This work focuses on the contextual bandit setting, with future directions including extensions to the MDP setting. Additionally, this work considers a subset of possible reward function parameterizations. A promising avenue for future work includes optimizing the use of a limited budget of counterfactual annotations, and identifying which counterfactual actions to label, in the spirit of Zrnic and Candès (2026). Furthermore, future work can consider identifying practical ways to pre-process a set of counterfactual annotations to mitigate the effect of bias and noise.

Acknowledgments

The authors would like to thank Emma Brunskill, Matthew Jörke, Yash Chandak, Allen Nie, Alex Nam, and Jared Moore for feedback on versions of this manuscript. AM was funded in part by a Stanford Data Science Fellowship. ST was funded in part by the Emory URC Research Award. JY was funded by Columbia University Data Science Institute. BEE was funded in part by grants from the Parker Institute for Cancer Immunology (PICI), the Chan-Zuckerberg Institute (CZI), the Biswas Family Foundation, NIH NHGRI R01 HG012967, and NIH NHGRI R01 HG013736. BEE is a CIFAR Fellow in the Multiscale Human Program. BEE is on the Scientific Advisory Board for ArrePath Inc, GSK AI for Cancer, and Freenome; she consults for Neumora.

References

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 129–138, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584959. doi: 10.1145/1557019.1557040. URL https://doi.org/10.1145/1557019.1557040.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning, 2011.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4), November 2014. ISSN 0883-4237. doi: 10.1214/14-sts500. URL http://dx.doi.org/10.1214/14-STS500.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 2000.

Anna Harutyunyan, Marc G. Bellemare, Tom Stepleton, and Remi Munos. Q(λ) with off-policy corrections, 2016. URL https://arxiv.org/abs/1602.04951.

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. ISSN 01621459. URL http://www.jstor.org/stable/2280784.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning, 2016.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and R Mark IV. Mimic-iv (version 0.4). *PhysioNet*, 2020.

Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of heartsteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019.

Hoang M. Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints, 2019. URL https://arxiv.org/abs/1903.08738.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, WWW '10. ACM, April 2010. doi: 10.1145/1772690.1772758. URL http://dx.doi.org/10.1145/1772690.1772758.

Aishwarya Mandyam, Matthew Jörke, William Denton, Barbara E. Engelhardt, and Emma Brunskill. Adaptive interventions with user-defined goals for health behavior change, 2024.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models, 2019. URL https://arxiv.org/abs/1905.05824.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Niranjani Prasad, Aishwarya Mandyam, Corey Chivers, Michael Draugelis, Clarence Hanson, Barbara Engelhardt, and Krzysztof Laudanski. Guiding efficient, effective, and patient-oriented electrolyte replacement in critical care: An artificial intelligence reinforcement learning approach. *Journal of Personalized Medicine*, 12:661, 04 2022. doi:

10.3390/jpm12050661.

Doina Precup, Richard Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, 06 2000.

Ishaan Shrimanker and Suman Bhattarai. Potassium. https://www.ncbi.nlm.nih.gov/books/NBK539791/, 2025. URL https://www.ncbi.nlm.nih.gov/books/NBK539791/. Accessed: 2025-05-15.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, second edition edition, 2018. ISBN 9780262039246.

Shengpu Tang and Jenna Wiens. Counterfactual-augmented importance sampling for semi-offline policy evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dsH244r9fA.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning, 2016.

Harm van Seijen, H. V. Hasselt, Shimon Whiteson, and Marco A Wiering. A theoretical and empirical analysis of expected sarsa. *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 177–184, 2009. URL https://api.semanticscholar.org/CorpusID:6230754.

Cameron Voloshin, Hoang M. Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning, 2021.

Jiayu Yao, Emma Brunskill, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. Power constrained bandits. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, pages 209–259, 2021.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference, 2026. URL https://arxiv.org/abs/2403.03208.

Appendix A. Additional Empirical Results

In the main text, we primarily focus on the case when the annotations are imperfect and the reward model is misspecified. We characterize the performance of the baseline OPE methods and our proposed approaches in four settings: either a well-specified or misspecified reward model, and either perfect or imperfect annotations. Here, we report additional results for each setting including error metrics on plots reported in the main text.

A.1. Well-Specified Reward

In Figure 2, we note that the bias of the counterfactual annotation plays a larger role in affecting the RMSE of the proposed methods than the variance of the counterfactual annotation. Here, we report the mean and standard deviation of the RMSE across all datasets (Figure S1). Our results indicate that DM^{+}-IS has the least fluctuation in standard deviation across the range of counterfactual annotation bias and variance in comparison to all methods that use counterfactual annotations.

A.2. Misspecified Reward

In Figure 3 we report mean RMSE across a variety of counterfactual annotation bias values in the Heartsteps and sepsis settings. In Figure S2, we report the mean and standard deviation of RMSE across all three datasets.

In Figure 4, we find that DM^{+}-IS is robust to annotation quality and reward model misspecification. We report the same plot for DM-IS^{+} and DM^{+}-IS^{+} in Figure S5. Our conclusions remain identical to those discussed in Section 4.

Finally, as discussed in Section 3, our goal is to define an estimator that is “doubly robust” to two sources of error, namely the error of the annotation and the error of the reward model. As such, we do not account for imperfect IPS ratios and assume that the estimates of IPS ratios are fairly accurate. In the case that the IPS ratio is inaccurate, all proposed estimators will be biased. The estimation error of the IPS ratio will thus propagate through the bias and variance reductions introduced in our work. To further illustrate this, we report the performance of DM^{+}-IS with varying degrees of incorrectly estimated behavior policies ϵ (Figure S3). We leave further evaluation of this setting to future work.

A.3. Additional MIMIC-IV Results

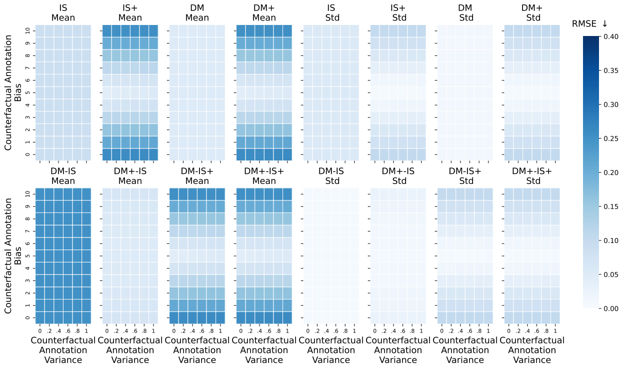
In Section 4.4, we compared the performance on MIMIC-IV data of all estimators with 100 counterfactual annotations. Now, we demonstrate that DM^{+}-IS has the most favorable performance across a wider range of m, the number of counterfactual annotations available (Figure S4). In particular, we note that the performance of DM^{+}-IS stays consistently better in comparison to the other estimators across all values of m and that the performance either improves or remains consistent, but does not degrade if you increase the number of counterfactual annotations.

Appendix B. Simulator environments

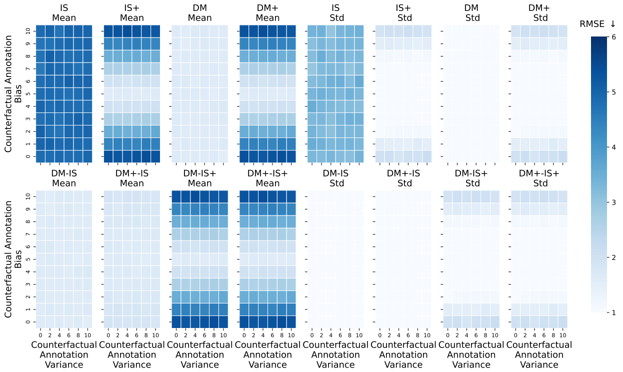
In all of our experiments, we use a separate random seed to generate datasets. We report results across 50 runs, each of which is a sampled dataset. Our results across all experiments take approximately 100 hours of compute, which was run on a local university cluster. Our code is attached as a portion of the supplement material and will be made public upon acceptance via a Github link. Our code is covered under the MIT license. Here we discuss the implementation details for the three simulator settings we use in our empirical results. Key details of the settings can be observed in Table 1.

B.1. Two-context bandit

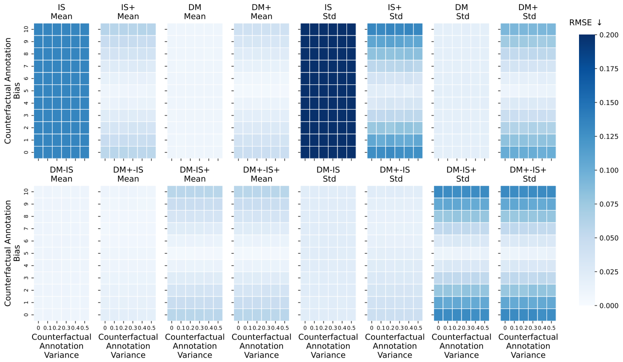
The two-context bandit setting contains two contexts, each with two actions. The bandit receives reward for taking either action from the first context, and no reward for taking any action from the second context.



(a) 2-context bandit.

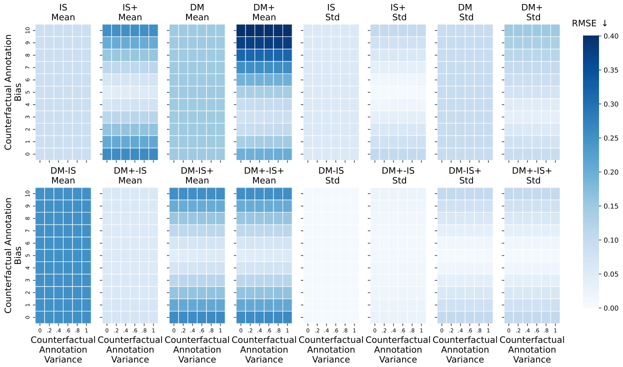


(b) Heartsteps.

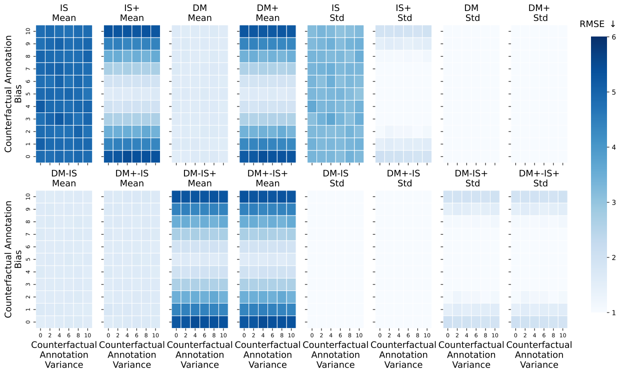


(c) Sepsis.

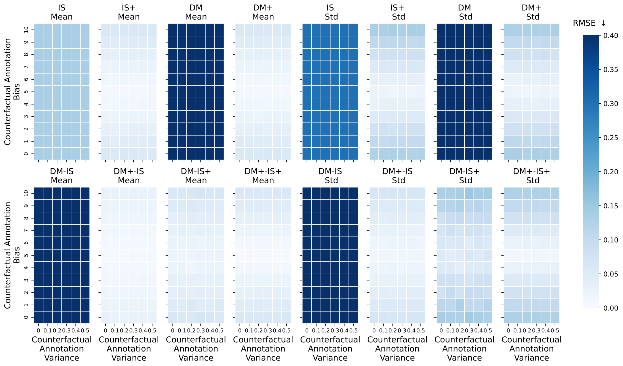
Figure S1: **Heatmap of mean and standard deviation of RMSE with a well-specified reward model (lower mean/standard deviation is represented lighter)**: The mean and standard deviation are reported with respect to the different combinations of behavior and target policies we consider. The x, y-axis represents the variance (Δ_G) and the bias (ϵ_G) of the annotations, respectively. In a well-specified reward setting, DM and DR perform comparably and best. Of note, IS sometimes performs well in this setting largely due to high coverage in the behavior policy for most pairings. In terms of standard deviation of RMSE, all methods that use counterfactual annotations with the exception of DR experience high standard deviation for the most imperfect annotations.



(a) 2-context bandit.



(b) Heartsteps.



(c) Sepsis.

Figure S2: **Heatmap of mean and standard deviation of RMSE with a misspecified reward model (lower mean/standard deviation is represented lighter)**: The x, y-axis represents the variance (Δ_G) and the bias (ϵ_G) of the annotations, respectively. In a misspecified reward setting, DM and DM$^+$ tend to suffer. In comparison, DM$^+$-IS outperforms all method or performs comparably to the best performing methods for both RMSE mean and standard deviation.

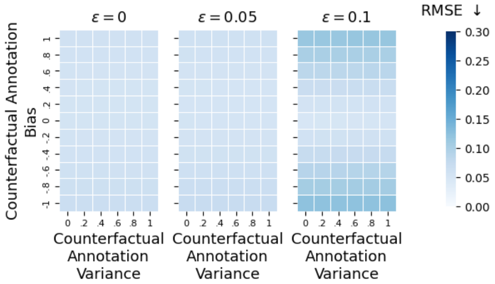


Figure S3: **Analyzing the consequences of incorrect IPS ratios in DM$^+$-IS in the Two-context Bandit:** The x, y axes correspond to the annotation variance and bias respectively. The heatmap color corresponds to the mean RMSE of the DM$^+$-IS estimator. We study three settings of ϵ, where $\hat{\pi_b} = \pi_b + \epsilon$ is the estimated behavior policy. As ϵ increases, DM$^+$-IS becomes a more biased estimator.

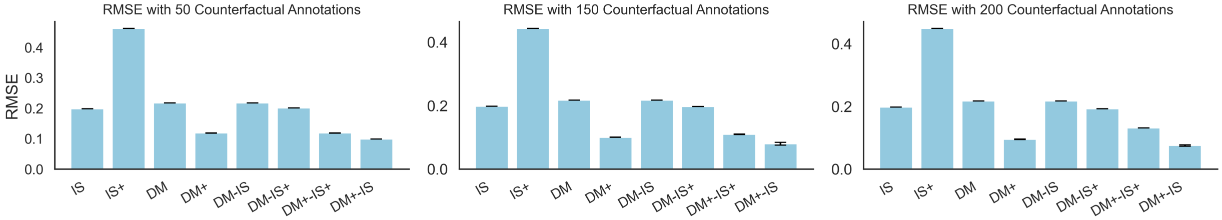
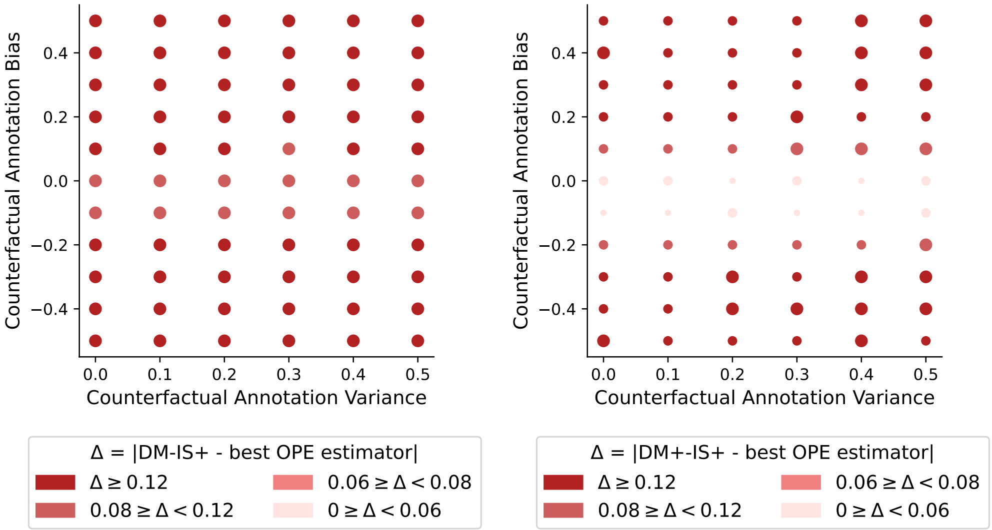


Figure S4: **Comparing the performance of all estimators in MIMIC-IV as we increase the number of counterfactual annotations.** The x-axis enumerates the possible estimators, and the y-axis measures the RMSE of the learned policy estimate. Error bars represent standard deviation. Each plot corresponds to the performance at different number of counterfactual annotations. We note that DM$^+$-IS has the most favorable performance across all settings.

In particular, the reward is sampled from the normal distribution $\mathcal{N}(1, 0.5)$ for the first action in the first context and the distribution $\mathcal{N}(2, 0.5)$ for the second action in the first context. In this domain, we sample a counterfactual annotation for every sample. Factual rewards are sampled from the reward distribution. We use a uniform initial context distribution, and report results over 100 runs for all estimators. We use a dataset of size $N = 100$ to fit the OPE estimate and $M = 100$ counterfactual annotations. The same dataset size and distribution is used to learn a reward function estimate if necessary depending on the OPE method. We equally weight all samples (0.5). The reward function is represented as a sample mean. A misspecified reward function uses a partially observed context. In particular, for 50% of the samples, the context is randomly selected. There are no hyperparameters to be tuned. We report results averaged across 9 possible combinations of π_b, π_e. We use all combinatorial combinations of the policies $[0.1, 0.9]$, $[0.5, 0.5]$, $[0.9, 0.1]$.

B.2. Heartsteps

The Heartsteps simulator is a step count simulator based off earlier work (Mandyam et al., 2024). All samples are sampled randomly and independently from an initial context distribution d_0 where each context is the square root of the prior day's step count. There are two possible actions: send a notification, or do not send a notification. The context and action are projected into $\mathbb{R}^3$ using a function $\phi(s, a)$, which outputs a vector that contains a scalar to represent the eventual decrease in step count over time, the previous day's step count, and a treatment effect term that is nonzero when the action is nonzero. Then, the step count is calculated as $\phi(s, a) \cdot \theta^T$ where $\theta = [-0.04, 0.9999, 0.3]$. The well-specified reward model is a linear function



(a) Exploring the consequences of choosing DM-IS$^+$ when annotation and reward model quality are unknown.

(b) Exploring the consequences of choosing DM$^+$-IS$^+$ when annotation and reward model quality are unknown.

Figure S5: Here we compare the consequences of choosing the other two proposed estimators (DM-IS$^+$ and DM$^+$-IS$^+$). Note that across the range of possible annotation quality, DM-IS$^+$ has a larger error (darker color) in comparison to DM$^+$-IS as displayed in Figure 4. While DM$^+$-IS$^+$ has smaller Δ when the counterfactual annotations have low variance, Δ more rapidly increases as the annotations become more imperfect.

of $\phi(s, a)$. The misspecified reward model is a linear function of only the first two indices of the vector $\phi(s, a)$. We report results averaged across 9 possible combinations of π_b, π_e. We use all combinatorial combinations of the policies $[0.1, 0.9], [0.5, 0.5], [0.9, 0.1]$.

B.3. Sepsis Treatment

The sepsis treatment simulator is based off prior work (Oberst and Sontag, 2019). While the original simulator assumes a Markov Decision Process (MDP) setting, we adapt the simulator to accommodate a contextual bandit setting. To do this, we sample one-step transitions rather than full trajectories. There are 1442 possible contexts and 8 possible actions in the environment. The initial context distribution d_0 samples uniformly across the possible contexts. Each context is represented as a vector of length 8, where the features describe patient heart rate, systolic blood pressure, blood glucose level, percentage oxygen, and the presence of three treatments including antibiotics, vasopressors, and ventilation. The 8 possible actions represent every combinatorial combination of three binary treatments: antibiotics, vasopressors, and ventilation. The well-specified reward function is a linear function of the number of abnormal vitals, and whether the patient is on treatment or not, with $\theta = [-1, -1]$. The misspecified reward model one-hot-encodes the context and action and projects into a vector of length 168. The reward function is then a linear function of the vector of length 168. In this setting, we use one target policy, $\pi_e = [0.3, 0.2, 0, 0, 0.2, 0.1, 0.1, 0.1]$. We report results averaged across the following behavior policies:

$$\begin{aligned}
\pi_{b1} &= [0.1, 0.1, 0.4, 0.3, 0.1, 0.0, 0.0, 0.0] \\
\pi_{b2} &= [0.1, 0.1, 0.4, 0.2, 0.1, 0.1, 0.0, 0.0] \\
\pi_{b3} &= [0.1, 0.1, 0.4, 0.1, 0.1, 0.1, 0.0, 0.1] \\
\pi_{b4} &= [0.1, 0.1, 0.3, 0.1, 0.1, 0.1, 0.1, 0.1] \\
\pi_{b5} &= [0.2, 0.1, 0.2, 0.1, 0.1, 0.1, 0.1, 0.1] \\
\pi_{b6} &= [0.3, 0.1, 0.2, 0.0, 0.1, 0.1, 0.1, 0.1]
\end{aligned}$$

B.4. MIMIC-IV

Here, we use data from MIMIC-IV (Johnson et al., 2020), an electronic health records dataset sourced from the Beth Israel Deaconess Medical Center in Boston, MA. We consider a subset of the patients who receive potassium repletion. That is, we include all patients who have received at least one instance of potassium administration through an IV. Furthermore, we treat this setting as a one-step contextual bandit setting. A patient context is represented as a 20-dimensional vector containing information about static covariates (e.g., age, gender), aggregated lab values observed in the previous four hour window, aggregated medicines administered in the previous four hour window, and any indication of procedures that were undertaken (e.g., the patient was placed on a ventilator). There are five possible actions, each corresponding to a dosage of potassium (units are mEq). After administering a dosage of potassium, we observe a reward that is a function of the patient's updated context.

In our other empirical results, we report RMSE as the key metric; this hinges on the fact that we know the reward function for this setting. To emulate this in MIMIC-IV, a setting in which there is no reported reward signal, we construct a specific reward function. In particular, this reward function is a binary indicator of whether a patient's potassium lab value observed after potassium administration is within the potassium reference range (3.5-5 mmol/L). That is, the reward function $R(s_{t+1})$ is a function of the patient's next observed context. It is reasonable to assume that the reward function is shared across all patients because the potassium reference range is shared across all patients.

To emulate a setting in which we have a behavior policy and a target policy, we further split the cohort of patients that receive potassium repletion into two sub-cohorts. In particular, one sub-cohort does not have renal disease, and one sub-cohort does have renal disease. The repletion policies are different between these cohorts because the policy must consider the inability for the renal disease patients' kidneys to properly filter potassium. We treat the non-renal population as the behavior cohort and the renal population as the target cohort. The value of the target policy can be calculated by treating the target samples as Monte-Carlo samples. We estimate the policies from the data.

We expect the repletion policies between the two groups to differ since the policy for renal patients must account for the inability of their kidneys to efficiently absorb potassium. This setup allows us to calculate the ground-truth value of the target policy using the returns of the target trajectories. As shown in Figure S6, patients with renal disease are administered a variety of dosages, including ones that are under-observed in the behavior dataset (e.g., 10 mEq); as a result, this is a setting in which OPE can be useful. Finally, we construct annotations using LLMs for the behavior cohort. The annotations are sourced from a powerful LLM, OpenAI 'o1'. LLMs have been successful in reasoning about medical domains, presenting a cheaper alternative to soliciting an expert (i.e., doctor). To obtain annotations, we ask the LLM to complete two tasks. The first is to summarize the patient context, highlighting covariates relevant for predicting blood potassium level. Then, we ask the LLM to predict the patient's blood potassium if they were administered a given dosage. Note that we have ground truth lab values for the dosages administered in MIMIC-IV; in our experiments we found 'o1' predictions to largely align with ground truth lab values. Annotations are obtained by soliciting lab value predictions for unobserved dosages and feeding these values into the known reward function.

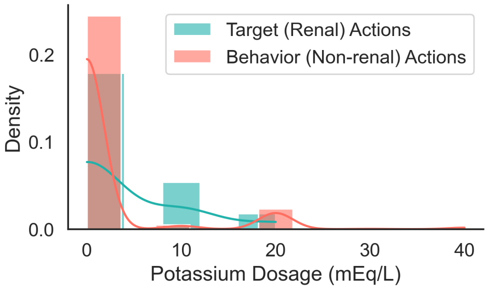


Figure S6: **Repletion policies differ between the behavior and target cohorts**, making this a setting in which OPE can be helpful. In particular, we note that the target dataset samples actions (e.g., 10 mEq of potassium) that are under-observed in the behavior dataset.

Appendix C. Naive Doubly Robust Estimator

As discussed in Section 1, a naive doubly robust estimator uses the dataset D^+ within the context of a standard doubly robust estimator. The definition of the naive doubly robust estimator is

$$\hat{V}^{\text{Naive-DR}} = \sum_{i=1}^{N+M} \Big(\hat{R}^+(s_i, \pi_e) + \rho_{s_i}(a_i)(c_i^{a_i} - \hat{R}^+(s_i, a_i) \Big),$$

where $c_i^{a_i}$ is the factual reward or counterfactual annotation depending on if (s_i, a_i) is a factual or counterfactual sample. We claim in the main text that this approach will always lead to an arbitrarily biased estimate of $v(\pi_e)$ because it does not preserve the context distribution in D. To demonstrate this, we report results (Table 2) for the naive doubly robust estimator in the 2-context bandit environment under a well-specified reward model and a perfect annotation setting (Assumption 2). Note that the standard doubly robust estimator (DR) is unbiased here, and in contrast, the naive doubly robust estimator is biased. This is because the naive doubly robust estimator alters the context distribution d_0, which is assumed to be constant between the behavior dataset and the dataset used to calculate the target policy value $v(\pi_e)$.

Dataset	**N**	**M**	**% annotatec**	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{S}\rvert$	# of π_b, π_e	**Well-specified reward**	**Misspecified reward**
2-context Bandit	100	100	100	2	2	9	Sample mean	A random state is observed 50% of the time, creating partial observability like ModelFail in Thomas and Brunskill (2016).
Heartsteps	200	200	100	2	80	9	Linear regression (3 features)	Linear regression (2 features)
Sepsis	700	700	12.5	8	1442	6	Linear regression (top 2 features)	Linear regression (168 least relevant features)
MIMIC-IV	652	Varies	Varies	5	> 13000	1	N/A	Indicator function of whether the lab is within the reference range

Table 1: Key characteristics of contextual bandits settings used in empirical results.

Table 2: Naive doubly robust estimator performance. We report mean ± standard error for all values.

Method	**RMSE**	**Bias**	**Std**
Naive DR	0.317 ± 0.0007	0.3038 ± 0.0922	0.092 ± 0.0001
DR	0.108 ± 0.0002	-0.016± 0.1066	0.1067 ± 0.0001

Appendix D. Table of theoretical results

We provide a summary for our theoretical results to guide the reader to the appropriate proofs (Table 3). Earlier work (Tang and Wiens, 2023) introduced an IS-based estimator that incorporates counterfactual annotations, and found that re-weighting the samples was required to maintain the context-action distribution defined by the original factual dataset. In our work, we note that this re-weighting is not necessary when building a reward function, which we discuss in Section E. On the journey to constructing a doubly robust estimator, we derive the bias and variance of the standard direct method OPE estimator in Section F. Our work identifies three opportunities to incorporate counterfactual annotations into a doubly robust estimator. In this work, we investigate the theoretical properties of these estimators under three annotation settings: perfect annotations, biased annotations, and higher variance annotations. We prove expectation and variance terms for each of our three estimators with and without assumptions on the annotation quality in Section G. Note that we avoid deriving the variance for DM-IS$^+$ and DM$^+$-IS$^+$ because these terms are very complicated. Instead, we analyze these variance terms empirically in our simulated settings. Finally, we establish an equivalence between two of our doubly robust approaches and IS$^+$ (Tang and Wiens, 2023) under an equal weighting scheme in Section H.

Appendix E. Weighted vs. unweighted reward function

In the main text, we claim that we do not need to re-weight samples when constructing a reward function to produce an unbiased estimate, like earlier work had to do with IS$^+$ (Tang and Wiens, 2023). To demonstrate this, we first derive the expectation and variance of the weighted reward function, and identify that the unweighted (or equally weighted) reward function is a special case. We then compare the variance terms for the unweighted and weighted reward functions, and prove that the variance of the unweighted reward

Table 3: Summary of theoretical results and associated proofs. Note that only Assumption 3 is required for the corresponding bias proofs.

Method	Assumption Annotation	Assumption Coverage	Bias	Std
DM$^+$-IS	2	1	Proposition S12	Proposition S13
DM-IS$^+$	2	5	Proposition S14	Proposition S16
DM$^+$-IS$^+$	2	5	Proposition S17	Proposition S19
DM$^+$-IS	3,4	1	Proposition 1	Theorem 3
DM-IS$^+$	3,4	5	Theorem 2	Proposition S15
DM$^+$-IS$^+$	3,4	5	Theorem 2	Proposition S18

function is lower. This result suggests that not weighting the samples when constructing the reward function is a superior strategy in the case that the annotations are perfect.

E.1. Weighted Reward Function

The weighted reward function is

$$\hat{R}^+(s,a) = \frac{\sum_{i=1}^{N} \mathbb{1}(s_i = s, a_i = a) * w_i^a * r_i + \sum_{j=1}^{M} \mathbb{1}(s_j = s, a_j = a) * w_j^a * g_j}{\sum_{i=1}^{N} \mathbb{1}(s_i = s, a_i = a) * w_i^a + \sum_{j=1}^{M} \mathbb{1}(s_j = s, a_j = a) * w_j^a}$$

where $w_i^a \sim W(s_i, a_i)$ is a known weight associated with the sample s_i, a_i that arises from some function W.

E.1.1. BIAS

Proposition S4 (Unbiasedness of weighted reward function)**.** *Under Assumption 2, the weighted reward function is unbiased.* $\mathbb{E}[\hat{R}^+(s,a)] = \bar{R}(s,a)$.

Proof: Let $N(s,a) = \sum_{i=1}^{M+N} \mathbf{1}(s_i = s, a_i = a)$ be the number of times the sample s,a appears in the counterfactual-annotated dataset D^+. We can re-write the weighted reward function as a function of $N(s,a)$. Let $c_i = r_i$ if the sample is in the factual dataset, and $c_i = g_i$ if the sample is a counterfactual annotation.

$$\hat{R}^+(s,a) = \frac{1}{\sum_{i=1}^{N(s,a)} w_i^a} \sum_{i=1}^{N(s,a)} w_i^a c_i$$

To calculate expectation, we first use the law of total expectation and the expectation of the scalar reward or annotation.

$$\mathbb{E}_{D^+ \sim \mathcal{D}^+}[\hat{R}^+(s,a)] = \mathbb{E}_{\substack{N_{s,a} \sim D^+, c_i \sim R(s_i,a_i), \\ w_i^a \sim W(s_i,a)}} \left[\frac{1}{\sum_{i=1}^{N(s,a)} w_i^a} \sum_{i=1}^{N(s,a)} w_i^a c_i \right] \tag{7}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{E}_{c_i \sim R(s_i,a_i), w_i^a \sim W(s_i,a)} \left[\frac{1}{\sum_{i=1}^{N(s,a)} w_i^a} \sum_{i=1}^{N(s,a)} w_i^a c_i \right] \right] \tag{8}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{E}_{w_i^a \sim W(s_i,a)} \left[\frac{1}{\sum_{i=1}^{N(s,a)} w_i^a} \sum_{i=1}^{N(s,a)} w_i^a \mathbb{E}_{c_i \sim R(s_i,a_i)}[c_i] \right] \right] \tag{9}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{E}_{w_i^a \sim W(s_i,a)} \left[\frac{1}{\sum_{i=1}^{N(s,a)} w_i^a} \sum_{i=1}^{N(s,a)} w_i^a \bar{R}(s,a) \right] \right] \tag{10}$$

Let $W_i^a = \frac{w_i^a}{\sum_{i=1}^{N(s,a)} w_i^a}$. Because each w_i^a is normalized by the sum of all the weights across all the samples, the sum of all weights is 1. Now, we can write the sum as a weighted mean:

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{E}_{w_i^a \sim W(s_i,a)} \left[\sum_{i=1}^{N(s,a)} W_i^a \bar{R}(s,a) \right] \right] \tag{11}$$

$$\tag{12}$$

Because $\bar{R}(s_i, a_i)$ is a constant, we can pull it out of the expectation.

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{E}_{w_i^a \sim W(s_i,a)} \left[\bar{R}(s,a) \sum_{i=1}^{N(s,a)} W_i^a \right] \right] \tag{13}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{E}_{w_i^a \sim W(s_i,a)} \left[\bar{R}(s,a) \right] \right] \tag{14}$$

$$= \bar{R}(s,a) \tag{15}$$

E.1.2. VARIANCE

Proposition S5. *Under Assumption 2, the weighted reward function has variance,* $\mathbb{V}[\hat{R}^+(s,a)] = \mathbb{E}_{N_{s,a} \sim D^+} \left[\sum_{i=1}^{N(s,a)} (W_i^a)^2 \sigma_R^2(s,a) \right]$.

Proof: We use the law of total variance and the definition of W_i^a.

$$\mathbb{V}_{D^+ \sim \mathcal{D}^+}[\hat{R}^+(s,a)] = \mathbb{V}_{\substack{N_{s,a} \sim D^+, \\ c_i \sim R(s_i,a_i)}} \left[\sum_{i=1}^{N(s,a)} \frac{w_i^a c_i}{\sum_{i=1}^{N(s,a)} w_i^a} \right] \tag{16}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{V}_{c_i \sim R(s_i,a_i)} \left[\sum_{i=1}^{N(s,a)} \frac{w_i^a c_i}{\sum_{i=1}^{N(s,a)} w_i^a} \right] \right] \tag{17}$$

$$+ \mathbb{V}_{N_{s,a} \sim D^+} \left[\mathbb{E}_{c_i \sim R(s_i,a_i)} \left[\sum_{i=1}^{N(s,a)} \frac{w_i^a c_i}{\sum_{i=1}^{N(s,a)} w_i^a} \right] \right] \tag{18}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{V}_{c_i \sim R(s_i,a_i)} \left[\sum_{i=1}^{N(s,a)} \frac{w_i^a c_i}{\sum_{i=1}^{N(s,a)} w_i^a} \right] \right] \tag{19}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\mathbb{V}_{c_i \sim R(s_i,a_i)} \left[\sum_{i=1}^{N(s,a)} W_i^a c_i \right] \right] \tag{20}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\sum_{i=1}^{N(s,a)} (W_i^a)^2 \mathbb{V}_{c_i \sim R(s,a)} \left[c_i \right] \right] \tag{21}$$

$$= \mathbb{E}_{N_{s,a} \sim D^+} \left[\sum_{i=1}^{N(s,a)} (W_i^a)^2 \sigma_R^2(s,a) \right] \tag{22}$$

$$\tag{23}$$

E.2. Unweighted Reward Function

Here the unweighted reward function is identical to the weighted one, except without weights. This can also be seen as a special case of the weighted reward function with equal weights. The unweighted reward function is :

$$\hat{R}^{+}(s,a) = \frac{\sum_{i=1}^{N} \mathbb{1}(s_i = s, a_i = a) * r_i + \sum_{j=1}^{M} \mathbb{1}(s_j = s, a_j = a) * g_j}{\sum_{i=1}^{N} \mathbb{1}(s_i = s, a_i = a) + \sum_{j=1}^{M} \mathbb{1}(s_j = s, a_j = a)} \tag{24}$$

E.2.1. BIAS

Proposition S6. *Under Assumption 3, the unweighted reward function has expectation* $\mathbb{E}[\hat{R}^{+}(s,a)] = \mathbb{E}_{N_{s,a},M_{s,a}}\left[\bar{R}(s_i,a_i) + \frac{M_{s,a}\epsilon_G}{N_{s,a}+M_{s,a}}\right]$ *where* $N_{s,a}, M_{s,a}$ *are the number of factual and counterfactual samples with context-action equivalent to* s, a *in the counterfactual augmented dataset.*

Proof: Define $N_{s,a} = \sum_{i=1}^{N} \mathbb{1}(s_i = s, a_i = a)$ and $M_{s,a} = \sum_{j=1}^{M} \mathbb{1}(s_j = s, a_j = a)$ where N is the number of total factual samples and M is the total number of counterfactual annotations.

Now, we can re-write the reward function as $\hat{R}^{+}(s,a) = \frac{\sum_{i=1}^{N_{s,a}} r_i + \sum_{j=1}^{M_{s,a}} g_j}{N_{s,a}+M_{s,a}}$.

We calculate the expectation of this reward function under the dataset distribution, which is the joint distribution across $N_{s,a} \sim D^{+}, M_{s,a} \sim D^{+}, r \sim R(s,a), g \sim G(s,a)$.

$$\mathbb{E}_{D^{+}\sim\mathcal{D}^{+}}[\hat{R}^{+}(s,a)] = \mathbb{E}_{N_{s,a},M_{s,a}\sim D^{+}, r\sim R(s,a), g\sim G(s,a)}\left[\frac{\sum_{i=1}^{N_{s,a}} r_i + \sum_{j=1}^{M_{s,a}} g_j}{N_{s,a}+M_{s,a}}\right] \tag{25}$$

We can use the law of total expectation to separate out the joint distribution:

$$= \mathbb{E}_{\substack{N_{s,a}\\ \sim D^{+}}}\left[\mathbb{E}_{M_{s,a}\sim D^{+}, r\sim R(s,a), g\sim G(s,a)}\left[\frac{\sum_{i=1}^{N_{s,a}} r_i + \sum_{j=1}^{M_{s,a}} g_j}{N_{s,a}+M_{s,a}}\right]\right] \tag{26}$$

$$= \mathbb{E}_{\substack{N_{s,a}\\ \sim D^{+}}}\left[\mathbb{E}_{M_{s,a}\sim D^{+}}\left[\mathbb{E}_{r\sim R(s,a), g\sim G(s,a)}\left[\frac{\sum_{i=1}^{N_{s,a}} r_i + \sum_{j=1}^{M_{s,a}} g_j}{N_{s,a}+M_{s,a}}\right]\right]\right] \tag{27}$$

$$= \mathbb{E}_{\substack{N_{s,a}\\ \sim D^{+}}}\left[\mathbb{E}_{M_{s,a}\sim D^{+}}\left[\frac{1}{N_{s,a}+M_{s,a}}\sum_{i=1}^{N_{s,a}} \bar{R}(s_i,a_i) + \sum_{j=1}^{M_{s,a}} \bar{R}(s_j,a_j) + \epsilon_G\right]\right] \tag{28}$$

$$= \mathbb{E}_{N_{s,a}\sim D^{+}}\left[\mathbb{E}_{M_{s,a}\sim D^{+}}\left[\frac{1}{N_{s,a}+M_{s,a}}\sum_{i=1}^{N_{s,a}} \bar{R}(s_i,a_i) + \sum_{j=1}^{M_{s,a}} \bar{R}(s_j,a_j) + \epsilon_G\right]\right] \tag{29}$$

$$= \mathbb{E}_{N_{s,a}\sim D^{+}}\left[\mathbb{E}_{M_{s,a}\sim D^{+}}\left[\frac{1}{N_{s,a}+M_{s,a}} N_{s,a} \times \bar{R}(s_i,a_i) + M_{s,a} \times (\bar{R}(s_j,a_j) + \epsilon_G)\right]\right] \tag{30}$$

$$= \mathbb{E}_{N_{s,a}\sim D^{+}}\left[\mathbb{E}_{M_{s,a}\sim D^{+}}\left[\bar{R}(s_i,a_i) + \frac{M_{s,a}\epsilon_G}{N_{s,a}+M_{s,a}}\right]\right] \tag{31}$$

$$= \mathbb{E}_{N_{s,a},M_{s,a}\sim D^{+}}\left[\bar{R}(s_i,a_i) + \frac{M_{s,a}\epsilon_G}{N_{s,a}+M_{s,a}}\right] \tag{32}$$

$$\tag{33}$$

Proposition S7. *Under Assumption 2, the unweighted reward function is unbiased,* $\mathbb{E}[\hat{R}^{+}(s,a)] = \bar{R}(s,a)$.

Proof: If we set $\epsilon_G = 0$ in Proposition S6, the expected value of the unweighted reward function is $\bar{R}(s,a)$, which means that it is an unbiased estimator.

E.2.2. VARIANCE

Proposition S8. *Under Assumption 2, the unweighted reward function has variance,* $\mathbb{V}[\hat{R}^{+}(s,a)] = \sigma_R^2(s,a)\mathbb{E}_{N_{s,a}}\left[\frac{1}{N_{s,a}}\right]$.

Proof: Let $N_{s,a}$ denote how many samples in the total dataset have the same state-action as s, a. We now use the law of total variance, the earlier bias result, and the expectation of c_i.

$$\mathbb{V}_{D^{+}\sim\mathcal{D}^{+}}[\hat{R}^{+}(s,a)] = \mathbb{V}_{\substack{N_{s,a} \\ c_i\sim R(s,a)}}\left[\frac{1}{N_{s,a}}\sum_{i=1}^{N_{s,a}} c_i\right] \tag{34}$$

$$= \mathbb{E}_{N_{s,a}}\left[\mathbb{V}_{c\sim R}\left[\frac{1}{N_{s,a}}\sum_{i=1}^{N_{s,a}} c_i\right]\right] \tag{35}$$

$$+ \mathbb{V}_{N_{s,a}}\left[\mathbb{E}_{c_i\sim R(s,a)}\left[\frac{1}{N_{s,a}}\sum_{i=1}^{N_{s,a}} c_i\right]\right] \tag{36}$$

$$= \mathbb{E}_{N_{s,a}}\left[\mathbb{V}_{c_i\sim R(s,a)}\left[\frac{1}{N_{s,a}}\sum_{i=1}^{N_{s,a}} c_i\right]\right] \tag{37}$$

$$= \mathbb{E}_{N_{s,a}}\left[\frac{1}{{N_{s,a}}^2}\sum_{i=1}^{N_{s,a}} \mathbb{V}_{c_i\sim R(s,a)}\left[c_i\right]\right] \tag{38}$$

$$= \mathbb{E}_{N_{s,a}}\left[\frac{1}{{N_{s,a}}^2}\sum_{i=1}^{N_{s,a}} \sigma_R^2(s,a)\right] \tag{39}$$

$$= \mathbb{E}_{N_{s,a}}\left[\frac{1}{{N_{s,a}}^2} \times N_{s,a} \times \sigma_R^2(s,a)\right] \tag{40}$$

$$= \mathbb{E}_{N_{s,a}}\left[\frac{\sigma_R^2(s,a)}{N_{s,a}}\right] \tag{41}$$

$$= \sigma_R^2(s,a)\mathbb{E}_{N_{s,a}}\left[\frac{1}{N_{s,a}}\right] \tag{42}$$

$$\tag{43}$$

E.3. Comparing variance terms

Proposition S9. *The variance of the unweighted reward function is less than or equal to the variance of the weighted reward function.*

Proof: We can prove that the variance of the weighted reward function is higher than that of the unweighted reward function using Jensen's inequality. Since both the variance terms for the weighted and unweighted reward function have a $\sigma_R^2(s,a)$, we consider only the coefficient. We start with the coefficient for the weighted reward function on the LHS of the first line. We know by Jensen's inequality that:

$$\frac{1}{N_{s,a}}\sum_{i=1}^{N_{s,a}}(W_i^a)^2 \geq \frac{1}{N_{s,a}^2}\left(\sum_{i=1}^{N_{s,a}} W_i^a\right)^2 \tag{44}$$

$$\frac{1}{N_{s,a}}\sum_{i=1}^{N_{s,a}}(W_i^a)^2 \geq \frac{1}{N_{s,a}^2} \times 1 \tag{45}$$

$$\sum_{i=1}^{N_{s,a}} (W_i^a)^2 \geq \frac{1}{N_{s,a}} \tag{46}$$

Note that the coefficient of the variance term for the unweighted reward function is $\frac{1}{N_{s,a}}$. This result suggests that the variance of the randomly weighted reward function is always at least as high as the variance of the uniformly weighted reward function. As a result, we recommend using a uniformly weighted (or unweighted) reward function estimator when we have perfect annotations.

In the case that we know the annotations are imperfect, we recommend using a weighted reward function that down-weights the imperfect annotation (and consequently up-weights the corresponding factual sample). We hypothesize that this procedure will result in an improved estimator in both the misspecified and well-specified reward function scenarios because it limits the effect of the imperfect annotation.

We include a baseline which augments the standard DM estimator using counterfactual annotations. This estimator is $\hat{V}^{DM^+} = \sum_s d_0(s) \sum_a \pi(a|s)\hat{R}^+(s,a)$.

Appendix F. Bias and variance of the standard direct method (DM) OPE estimator

In the main text, we reference the variance term for the standard direct method (DM) estimator. Here, we derive the expectation and variance of the DM estimator, which is defined as:

$$\hat{V}^{\mathrm{DM}} = \frac{1}{N} \sum_{s_i \in D_{\hat{R}}} \sum_{a \in A} \pi_e(a|s_i)\hat{R}(s_i, a) \tag{47}$$

For the purposes of the expectation and variance derivations below, we assume that the reward model is fully realizable according to Assumption 6.

Assumption 6 (Realizability)**.** $R^* \in \mathcal{F}$.

We derive the expectation and variance with respect to two datasets. D_0 is used to estimate the OPE value, and $D_{\hat{R}}$ is used to estimate the reward function.

Proposition S10 (Unbiasedness of $\hat{V}^{\mathrm{DM}}$)**.** *Under Assumption 6,* $\mathbb{E}_{D_0\sim\mathcal{D}, D_{\hat{R}}\sim\mathcal{D}}[\hat{V}^{DM}] = v(\pi_e)$ *if* $N_{s,a} > 0$ *for all* (s, a) *where* $d(s) > 0$ *and* $\pi_e(a|s) > 0$ *in the dataset* $D_{\hat{R}}$.

Proof:

$$\begin{aligned}
\mathbb{E}_{D_0\sim\mathcal{D}, D_{\hat{R}}\sim\mathcal{D}}[\hat{V}^{\mathrm{DM}}] &= \mathbb{E}_{D_0, D_{\hat{R}}\sim\mathcal{D}}\left[\frac{1}{N}\sum_{s_i\in D_0}\sum_{a\in A}\pi_e(a|s_i)\hat{R}(s_i,a)\right] && (48)\\
&= \mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{a\in A}\pi_e(a|s_i)\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}[\hat{R}(s_i,a)]\right] && (49)\\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\left[\sum_{a\in A}\pi_e(a|s_i)\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}[\hat{R}(s_i,a)]\right] && (50)\\
& && (51)
\end{aligned}$$

Now, because we consider every sample as independent, we consider just the term inside the summation without $\frac{1}{N}$.

$$= \sum_{i=1}^{N}\sum_{s\in\mathcal{S}} d(s)\left(\sum_{a\in A}\pi_e(a|s)\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}[\hat{R}(s,a)]\right) \tag{52}$$

$$= \sum_{i=1}^{N} \sum_{s \in \mathcal{S}} d(s) \left(\sum_{a \in A} \pi_e(a|s) \bar{R}(s,a) \right) \tag{53}$$

$$= \sum_{i=1}^{N} \sum_{s \in \mathcal{S}} d(s) v^{\pi_e}(s_i) \tag{54}$$

$$= \sum_{i=1}^{N} v(\pi_e) \tag{55}$$

$$= v(\pi_e) \tag{56}$$

where we apply linearity of expectation, definition of expectation over $D_{\hat{R}}$, substitute the expectation of $\hat{R}(s,a)$ where $d(s) > 0$ and $\pi_e(a|s) > 0$, definition of value function and policy value.

Proposition S11 (Variance of $\hat{V}^{\mathrm{DM}}$)**.** *Under Assumption 6,* $\mathbb{V}_{D\sim\mathcal{D},D_{\hat{R}}\sim\mathcal{D}}[\hat{V}^{DM}] = \frac{1}{N}\mathbb{V}_{s\sim d_0}\left[V^{\pi_e}(s)\right] + \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_e}\left[\left(\frac{1}{N} + \left(1-\frac{1}{N}\right)d(s)\right)\pi_e(a|s)\,\sigma_R^2(s,a)\,\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\left[\frac{1}{N_{s,a}(D_{\hat{R}})}\right]\right]$.

Proof: By law of total variance, we can decompose the variance into two terms:

$$\mathbb{V}_{D_{\hat{R}}\sim\mathcal{D},D_{\hat{R}}\sim\mathcal{D}}[\hat{V}^{\mathrm{DM}}] = \underbrace{\mathbb{V}_{D_{\hat{R}}\sim\mathcal{D}}\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\left[\hat{V}^{\mathrm{DM}}\right]}_{(1)} + \underbrace{\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\mathbb{V}_{D_{\hat{R}}\sim\mathcal{D}}\left[\hat{V}^{\mathrm{DM}}\right]}_{(1')} \tag{57}$$

We can substitute the intermediate result from the proof of Proposition S10 into (1):

$$(1) = \mathbb{V}_{D_{\hat{R}}\sim\mathcal{D}}\left[\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\left[\hat{V}^{\mathrm{DM}}\right]\right] = \mathbb{V}_{D_{\hat{R}}\sim\mathcal{D}}\left[\hat{R}(d,\pi_e)\right] \tag{58}$$

For $(1')$, we first consider the inner variance with respect to $D_{\hat{R}}$ assuming $\hat{R}$ is given:

$$\mathbb{V}_{D_{\hat{R}}\sim\mathcal{D}}\left[\hat{V}^{\mathrm{DM}}\right] = \mathbb{V}_{D_{\hat{R}}\sim\mathcal{D}}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{a\in A}\pi_e(a|s_i)\hat{R}(s_i,a)\right] \tag{59}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}_{s_i\sim d_0}\left[\hat{R}(s_i,\pi_e)\right] \qquad \text{var of sum of iid} \tag{60}$$

$$= \frac{1}{N}\mathbb{V}_{s\sim d_0}\left[\hat{R}(s,\pi_e)\right] \qquad \text{iid sample average} \tag{61}$$

Substituting this into $(1')$:

$$(1') = \mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\mathbb{V}_{D_{\hat{R}}\sim\mathcal{D}}\left[\hat{V}^{\mathrm{DM}}\right] \tag{62}$$

$$= \mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\left[\frac{1}{N}\mathbb{V}_{s\sim d_0}\left[\hat{R}(s,\pi_e)\right]\right] \tag{63}$$

$$= \frac{1}{N}\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\left[\mathbb{E}_{s\sim d_0}\left[\hat{R}(s,\pi_e)^2\right] - \mathbb{E}_{s\sim d_0}\left[\hat{R}(s,\pi_e)\right]^2\right] \qquad \text{definition of var} \tag{64}$$

$$= \frac{1}{N}\left(\underbrace{\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\mathbb{E}_{s\sim d_0}\left[\hat{R}(s,\pi_e)^2\right]}_{(2)} - \underbrace{\mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\left[\mathbb{E}_{s\sim d_0}\left[\hat{R}(s,\pi_e)\right]^2\right]}_{(2')}\right) \tag{65}$$

$$(2) = \mathbb{E}_{D_{\hat{R}}\sim\mathcal{D}}\mathbb{E}_{s\sim d_0}\left[\hat{R}(s,\pi_e)^2\right] \tag{66}$$

$$= \mathbb{E}_{s \sim d_0} \mathbb{E}_{D_{\hat{R}} \sim \mathcal{D}} \left[\hat{R}(s, \pi_e)^2 \right] \tag{67}$$

$$= \mathbb{E}_{s \sim d_0} \left[v^{\pi_e}(s)^2 + \mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}}[\hat{R}(s, \pi_e)] \right] \qquad \text{substitute corollary} \tag{68}$$

$$= \mathbb{E}_{s \sim d_0} \left[v^{\pi_e}(s)^2 \right] + \mathbb{E}_{s \sim d_0} \left[\mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}}[\hat{R}(s, \pi_e)] \right] \qquad \text{linearity of expectation} \tag{69}$$

$$= \mathbb{E}_{s \sim d_0} \left[v^{\pi_e}(s) \right]^2 + \mathbb{V}_{s \sim d_0} \left[v^{\pi_e}(s) \right] + \mathbb{E}_{s \sim d_0} \left[\mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}}[\hat{R}(s, \pi_e)] \right] \qquad \text{definition of variance} \tag{70}$$

$$= v(\pi_e)^2 + \mathbb{V}_{s \sim d_0} \left[v^{\pi_e}(s) \right] + \mathbb{E}_{s \sim d_0} \left[\mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}}[\hat{R}(s, \pi_e)] \right] \qquad \text{definition of value function} \tag{71}$$

$$\tag{72}$$

$$(2') = \mathbb{E}_{D_{\hat{R}} \sim \mathcal{D}} \left[\mathbb{E}_{s \sim d_0} \left[\hat{R}(s, \pi_e) \right]^2 \right] \tag{73}$$

$$= \mathbb{E}_{D_{\hat{R}} \sim \mathcal{D}} \left[\hat{R}(d, \pi_e)^2 \right] \tag{74}$$

$$= v(\pi_e)^2 + \mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}}[\hat{R}(d, \pi_e)] \tag{75}$$

Thus,

$$(2) - (2') = \mathbb{V}_{s \sim d_0} \left[v^{\pi_e}(s) \right] + \mathbb{E}_{s \sim d_0} \left[\mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}}[\hat{R}(s, \pi_e)] \right] - \mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}} \left[\hat{R}(d, \pi_e) \right] \tag{76}$$

Putting everything together, we have:

$$\mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}, D_{\hat{R}} \sim \mathcal{D}}[\hat{V}^{\mathrm{DM}}] = (1) + (1') \tag{77}$$

$$= \mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}} \left[\hat{R}(d, \pi_e) \right] \tag{78}$$

$$+ \frac{1}{N} \left(\mathbb{V}_{s \sim d_0} \left[v^{\pi_e}(s) \right] + \mathbb{E}_{s \sim d_0} \left[\mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}}[\hat{R}(s, \pi_e)] \right] - \mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}} \left[\hat{R}(d, \pi_e) \right] \right) \tag{79}$$

$$= \frac{1}{N} \mathbb{V}_{s \sim d_0} \left[v^{\pi_e}(s) \right] + \frac{1}{N} \mathbb{E}_{s \sim d_0} \left[\mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}}[\hat{R}(s, \pi_e)] \right] + \big(1 - \frac{1}{N}\big) \mathbb{V}_{D_{\hat{R}} \sim \mathcal{D}} \left[\hat{R}(d, \pi_e) \right] \tag{80}$$

$$= \frac{1}{N} \mathbb{V}_{s \sim d_0} \left[v^{\pi_e}(s) \right] + \frac{1}{N} \mathbb{E}_{s \sim d_0} \left[\mathbb{E}_{a \sim \pi_e(\cdot|s)} \left[\pi_e(a|s)\, \sigma_R^2(s, a)\, \mathbb{E}_{D_{\hat{R}} \sim \mathcal{D}} \Big[\frac{1}{N_{s,a}(D_{\hat{R}})} \Big] \right] \right] \tag{81}$$

$$+ \big(1 - \frac{1}{N}\big) \mathbb{E}_{s \sim d_0} \mathbb{E}_{a \sim \pi_e(\cdot|s)} \left[d(s)\, \pi_e(a|s)\, \sigma_R^2(s, a)\, \mathbb{E}_{D_{\hat{R}} \sim \mathcal{D}} \Big[\frac{1}{N_{s,a}(D_{\hat{R}})} \Big] \right] \tag{82}$$

$$= \frac{1}{N} \mathbb{V}_{s \sim d_0} \left[v^{\pi_e}(s) \right] \tag{83}$$

$$+ \mathbb{E}_{s \sim d_0} \mathbb{E}_{a \sim \pi_e(\cdot|s)} \left[\Big(\frac{1}{N} + \big(1 - \frac{1}{N}\big) d(s) \Big)\, \pi_e(a|s)\, \sigma_R^2(s, a)\, \mathbb{E}_{D_{\hat{R}} \sim \mathcal{D}} \Big[\frac{1}{N_{s,a}(D_{\hat{R}})} \Big] \right] \tag{84}$$

Appendix G. Expectation and variance of doubly robust estimators under different annotation conditions

Now we derive the expectation and variance of the introduced doubly robust estimators under different annotation conditions. In the main text, we note that the annotations can be perfect (i.e. $\mathbb{E}[g_i] = \mathbb{E}[r_i], \mathbb{V}[g_i] = \mathbb{V}[r_i]$), biased (i.e. $\mathbb{E}[g_i] = \mathbb{E}[r_i] + \epsilon_G(s_i, a_i)$), and have higher variance (i.e. $\mathbb{V}[g_i] = \mathbb{V}[r_i] + \Delta_G(s_i, a_i)$). In our proofs, we first derive the expectation and variance under the imperfect annotations condition. Then, we show that the perfect annotation condition is a special case of these derivations.

G.1. Expectation and variance of DM$^+$-IS

The DR estimator is defined as

$$\hat{V}^{DM^+-IS} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{R}^+(s_i,\pi_e) + \frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)}(r_i - \hat{R}^+(s_i,a_i))\right)$$

G.1.1. Expectation

We now prove Proposition 1, which is restated below.

Proposition (Expectation of DM$^+$-IS under imperfect annotations)**.** *Under Assumptions 1 and 3,* $\mathbb{E}[\hat{V}^{\mathrm{DM^+\text{-}IS}}] = v(\pi_e)$.

Proof: The expectation is taken over the dataset D_0, which is used to fit the OPE estimate, and $D_{\hat{R}}$, which is used to learn the reward function estimate. We first use the linearity of expectation.

$$\mathbb{E}_{D_0,D_{\hat{R}}\sim\mathcal{D}}[\hat{V}^{DM^+-IS}] \tag{85}$$

$$= \mathbb{E}_{\substack{D_{\hat{R}},s\sim d_0\\ a\sim\pi_b(\cdot|s),r\sim R(s,a)}}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{R}^+(s_i,\pi_e) + \frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)}(r_i - \hat{R}^+(s_i,a_i))\right)\right] \tag{86}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\substack{D_{\hat{R}},s\sim d_0\\ a\sim\pi_b(\cdot|s),r\sim R(s,a)}}\left[\left(\hat{R}^+(s_i,\pi_e) + \frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)}(r_i - \hat{R}^+(s_i,a_i))\right)\right] \tag{87}$$

We now split the expectation into two terms:

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{D_{\hat{R}},s\sim d_0}[\hat{R}^+(s_i,\pi_e)] + \mathbb{E}_{\substack{D_{\hat{R}},s\sim d_0\\ a\sim\pi_b(\cdot|s),r\sim R(s,a)}}\left[\frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)}(r_i - \hat{R}^+(s_i,a_i))\right] \tag{88}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{D_{\hat{R}},s\sim d_0}[\hat{R}^+(s_i,\pi_e)] + \mathbb{E}_{\substack{D_{\hat{R}},s\sim d_0\\ r\sim R(s,a)}}\left[\pi_b(a_i|s_i)\frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)}(r_i - \hat{R}^+(s_i,a_i))\right] \tag{89}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_0}[\hat{R}^+(s_i,\pi_e)] + \mathbb{E}_{s\sim d_0 a\sim\pi_e(\cdot|s),r\sim R(s,a)}\left[(r_i - \hat{R}^+(s_i,a_i))\right] \tag{90}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{D_{\hat{R}},s\sim d_0}[\hat{R}^+(s_i,\pi_e)] + \mathbb{E}_{D_{\hat{R}},s\sim d_0 a\sim\pi_e(\cdot|s)}\left[(\bar{R}(s_i,a_i) - \hat{R}^+(s_i,a_i))\right] \tag{91}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_0 a\sim\pi_e}[\bar{R}(s_i,a_i)] \tag{92}$$

$$= v(\pi_e) \tag{93}$$

The expectation of DM$^+$-IS when the annotations are biased is the value of the target policy. The variance of the annotation has no effect on the expectation of this estimator.

Proposition S12 (Unbiasedness of DM$^+$-IS)**.** *If both Assumptions 2 and 5 hold, the* DM$^+$-IS *estimator is unbiased,* $\mathbb{E}[\hat{V}^{\mathrm{DM^+\text{-}IS}}] = v(\pi_e)$.

Proof: Under imperfect annotations, the estimator is an unbiased estimator of the value of the target policy. If we have additional assumptions about perfect annotations, this estimator is also unbiased.

G.1.2. VARIANCE

We now prove Theorem 3, which is restated below.

Theorem (Variance of DM$^+$-IS under annotations with higher variance)**.** *Under Assumption 4 and Assumption 1,*

$$N \cdot \mathbb{V}[\hat{V}^{\text{DM}^+\text{-IS}}] = \mathbb{V}_{s\sim d_0}[v^{\pi_e}(s)] + \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b(s)}[\rho_s(a)^2\sigma_R^2(s,a)]$$
$$+ \mathbb{E}_{s\sim d_0}\Big[\mathbb{E}_{a\sim\pi_b}[\rho_s(a)^2\varepsilon_{\hat{R}^+}(s,a)^2] - \varepsilon^{\pi_e}_{\hat{R}^+}(s)^2\Big] + \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}\Big[\big(\rho_s(a)^2 - \tfrac{1}{\pi_b(a|s)}\big)\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]]\Big]$$

where $\varepsilon^+_{\hat{R}}(s,a) = \mathbb{E}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)] - \bar{R}(s,a)$ *and* $\varepsilon^{\pi_e}_{\hat{R}^+}(s) = \mathbb{E}_{a\sim\pi_e}[\varepsilon_{\hat{R}^+}(s,a)]$.

We use the law of total variance, the earlier expectation term, independence of samples, and the law of total variance.

$$\mathbb{V}_{D_0,D_{\hat{R}^+}}[\hat{V}^{\text{DM}^+\text{-IS}}] = \mathbb{V}_{D_0,D_{\hat{R}^+}}[\frac{1}{N}\sum_{i=1}^{N}\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)] \tag{94}$$

$$= \mathbb{E}_{D_{\hat{R}^+}}[\mathbb{V}_{D_0}[\frac{1}{N}\sum_{i=1}^{N}\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)]] \tag{95}$$

$$+ \mathbb{V}_{D_{\hat{R}^+}}[\mathbb{E}_{D_0}[\frac{1}{N}\sum_{i=1}^{N}\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)]] \tag{96}$$

$$= \mathbb{E}_{D_{\hat{R}^+}}[\frac{1}{N}\sum_{i=1}^{N}\mathbb{V}_{s_i\sim d_0 a_i\sim\pi_b, r_i\sim R(s_i,a_i)}[\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)]] \tag{97}$$

$$= \mathbb{E}_{D_{\hat{R}^+}}[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{s_i\sim d_0}[\mathbb{V}_{a_i\sim\pi_b, r_i\sim R(s_i,a_i)}[\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)]] \tag{98}$$

$$+ \mathbb{V}_{s_i\sim d_0}[\mathbb{E}_{a_i\sim\pi_b, r_i\sim R(s_i,a_i)}[\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)]]] \tag{99}$$

$$= \mathbb{E}_{D_{\hat{R}^+}}[\frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathbb{E}_{s_i\sim d_0}[\mathbb{V}_{a_i\sim\pi_b, r_i\sim R(s_i,a_i)}[\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)]]}_{1} \tag{100}$$

$$+ \mathbb{V}_{s_i\sim d_0}[v(\pi_e)]] \tag{101}$$

$$\text{Term 1} = \mathbb{E}_{D_{\hat{R}^+}}[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a_i\sim\pi_b}[\mathbb{V}_{r_i\sim R(s_i,a_i)}[\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)]] \tag{102}$$

$$+ \mathbb{E}_{s_i\sim d_0}[\mathbb{V}_{a_i\sim\pi_b}[\mathbb{E}_{r_i\sim R(s_i,a_i)}[\hat{R}^+(s_i,\pi_e) + \rho_{s_i}(a_i)(r_i - \hat{R}^+(s_i,a_i)]] \tag{103}$$

$$= \mathbb{E}_{D_{\hat{R}^+}}[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)^2\sigma_R^2]] \tag{104}$$

$$+ \mathbb{E}_{s_i\sim d_0}[\mathbb{V}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)(\bar{R} - \hat{R}^+)]] \tag{105}$$

The total variance term is now:

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{V}_{s_i\sim d_0}[v(\pi_e)]] + \mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)^2\sigma_R^2]] \tag{106}$$

$$+ \mathbb{E}_{D_{\hat{R}^+}}[\mathbb{E}_{s_i\sim d_0}[\mathbb{V}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)(\bar{R} - \hat{R}^+)]]] \tag{107}$$

We simplify the last term, the only term that involves $D_{\hat{R}^+}$.

$$\mathbb{E}_{D_{\hat{R}^+}}[\mathbb{E}_{s_i\sim d_0}[\mathbb{V}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)(\bar{R} - \hat{R}^+)]]] \tag{108}$$

$$= \mathbb{E}_{D_{\hat{R}^+}}[\mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)^2(\bar{R}-\hat{R}^+)^2] - \mathbb{E}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)(\bar{R}-\hat{R}^+)]^2]] \tag{109}$$

$$= \underbrace{\mathbb{E}_{D_{\hat{R}^+}}\mathbb{E}_{s_i\sim d_0}\mathbb{E}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)^2(\bar{R}-\hat{R}^+)^2]}_{1} - \underbrace{\mathbb{E}_{D_{\hat{R}^+}}\mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a_i\sim\pi_b}[\rho_{s_i}(a_i)(\bar{R}-\hat{R}^+)]^2]}_{2} \tag{110}$$

$$\text{Term 1} = \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}[\rho_s(a)^2\mathbb{E}_{D_{\hat{R}^+}}[(\bar{R}-\hat{R}^+)^2]] \tag{111}$$

$$= \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}[\rho_s(a)^2\mathbb{E}_{D_{\hat{R}^+}}[\bar{R}(s,a)^2 - 2\bar{R}(s,a)\hat{R}^+(s,a) + \hat{R}(s,a)^2]] \tag{112}$$

$$= \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}\Big[\rho_s(a)^2\big(\bar{R}(s,a)^2 - 2\bar{R}(s,a)\tilde{R}(s,a) + \tilde{R}(s,a)^2 + \mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]\big)\Big] \tag{113}$$

$$\text{where } \tilde{R}(s,a) = \mathbb{E}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)] \tag{114}$$

$$\text{Term 2} = \mathbb{E}_{D_{\hat{R}^+}}\mathbb{E}_{s\sim d_0}[\mathbb{E}_{a\sim\pi_e}[\bar{R}(s,a) - \hat{R}^+(s,a)]^2] \tag{115}$$

$$= \mathbb{E}_{s\sim d_0}\mathbb{E}_{D_{\hat{R}^+}}\Big[\big(V^{\pi_e}(s) - \hat{R}^+(s,\pi_e)\big)^2\Big] \tag{116}$$

$$= \mathbb{E}_{s\sim d_0}\mathbb{E}_{D_{\hat{R}^+}}\Big[V^{\pi_e}(s)^2 - 2V^{\pi_e}(s)\hat{R}^+(s,\pi_e) + \hat{R}^+(s,\pi_e)^2\Big] \tag{117}$$

$$= \mathbb{E}_{s\sim d_0}\Big[V^{\pi_e}(s)^2 - 2V^{\pi_e}(s)\tilde{R}(s,\pi_e) + \tilde{R}(s,\pi_e)^2 + \mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,\pi_e)]\Big] \tag{118}$$

$$\text{where } \tilde{R}(s,\pi_e) = \mathbb{E}_{D_{\hat{R}^+}}[\hat{R}^+(s,\pi_e)] \tag{119}$$

$$\text{Term 1 - Term 2} = \tag{120}$$

$$\mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}\Big[\rho_s(a)^2\big(\bar{R}(s,a)^2 - 2\bar{R}(s,a)\tilde{R}(s,a) + \tilde{R}(s,a)^2 + \mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]\big)\Big] \tag{121}$$

$$- \mathbb{E}_{s\sim d_0}\Big[V^{\pi_e}(s)^2 - 2V^{\pi_e}(s)\tilde{R}(s,\pi_e) + \tilde{R}(s,\pi_e)^2 + \mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,\pi_e)]\Big] \tag{122}$$

If the DM estimate is unbiased, we have $\tilde{R}(s,a) = \mathbb{E}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)] = \bar{R}(s,a)$. Substituting into the expression above results in cancellations and the expression becomes:

$$\mathbb{E}_{s\sim d_0}\Big[\mathbb{E}_{a\sim\pi_b}[\rho_s(a)^2\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]] - \mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,\pi_e)]\Big] \tag{123}$$

$$= \mathbb{E}_{s\sim d_0}\Big[\sum_a \pi_e(a|s)^2\frac{1}{\pi_b(a|s)}\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)] - \sum_a \pi_e(a|s)^2\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]\Big] \tag{124}$$

$$= \mathbb{E}_{s\sim d_0}\Big[\sum_a \pi_e(a|s)^2\big(\frac{1}{\pi_b(a|s)} - 1\big)\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]\Big] \tag{125}$$

$$= \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}\Big[\big(\rho_s(a)^2 - \tfrac{1}{\pi_b(a|s)}\big)\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]\Big] \tag{126}$$

If the DM estimate is biased, let's say $\tilde{R}(s,a) = \mathbb{E}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)] = \bar{R}(s,a) + \varepsilon_R(s,a)$, those terms no longer cancel and we get additional terms as follows:

$$\mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}\Big[\rho_s(a)^2\big(\bar{R}(s,a)^2 - 2\bar{R}(s,a)\tilde{R}(s,a) + \tilde{R}(s,a)^2\big)\Big] \tag{127}$$

$$-\mathbb{E}_{s\sim d_0}\Big[V^{\pi_e}(s)^2 + 2V^{\pi_e}(s)^2\tilde{R}(s,\pi_e) + \tilde{R}(s,\pi_e)^2\Big] \tag{128}$$

$$= \mathbb{E}_{s\sim d_0}\Big[\mathbb{E}_{a\sim\pi_b}[\rho_s(a)^2\varepsilon(s,a)^2] - \varepsilon^{\pi_e}(s)^2\Big] \tag{129}$$

$$\text{where } \varepsilon^{\pi_e}(s) = \mathbb{E}_{D_{\hat{R}^+}}[\hat{R}(s,\pi_e)] - V^{\pi_e}(s) \tag{130}$$

$$= \mathbb{E}_{s\sim d_0}\Big[\sum_a \pi_e(a|s)^2\frac{1}{\pi_b(a|s)}\varepsilon(s,a)^2 - \Big(\sum_a \pi_e(a|s)\varepsilon(s,a)\Big)^2\Big] \tag{131}$$

If we assume the bias of DM is constant across all (s,a) such that $\varepsilon(s,a) = \varepsilon$, then the above expression becomes

$$= \varepsilon^2\mathbb{E}_{s\sim d_0}\Big[\big(\sum_a \pi_e(a|s)^2\frac{1}{\pi_b(a|s)}\big) - 1\Big] \tag{132}$$

$$\geq 0 \text{ by Sedrakyan's lemma, a special case of Cauchy-Schwarz inequality} \tag{133}$$

The whole variance term is now:

$$N \cdot \mathbb{V}[\hat{V}^{DM^+-IS}] = \mathbb{V}_{s\sim d_0}[v^{\pi_e}(s)] + \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b(s)}[\rho_s(a)^2\sigma_R^2(s,a)] \tag{134}$$

$$+ \mathbb{E}_{s\sim d_0}\Big[\mathbb{E}_{a\sim\pi_b}[\rho_s(a)^2\varepsilon(s,a)^2] - \varepsilon^{\pi_e}(s)^2\Big] \tag{135}$$

$$+ \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}\Big[\big(\rho_s(a)^2 - \tfrac{1}{\pi_b(a|s)}\big)\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]\Big] \tag{136}$$

In the case that the annotations are perfect, the whole variance term is altered to reflect the fact that $\varepsilon = 0$.

Proposition S13 (Variance of DM$^+$-IS)**.** *If Assumptions 2 and 5 hold, and the reward model is well-specified,*

$$N \cdot \mathbb{V}[\hat{V}^{DM^+-IS}] = \mathbb{V}_{s\sim d_0}[v^{\pi_e}(s)] + \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b(s)}[\rho_s(a)^2\sigma_R^2(s,a)] \tag{137}$$

$$+ \mathbb{E}_{s\sim d_0}\mathbb{E}_{a\sim\pi_b}\Big[\big(\rho_s(a)^2 - \tfrac{1}{\pi_b(a|s)}\big)\mathbb{V}_{D_{\hat{R}^+}}[\hat{R}^+(s,a)]\Big] \tag{138}$$

G.2. Expectation and variance of DM-IS$^+$

The DR estimator is defined as

$$\hat{V}^{\text{DM-IS}^+} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{R}(s_i,\pi_e) + \sum_{a\in A} w_i^a \frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(c_i^a - \hat{R}(s_i,a))\right)$$

G.2.1. EXPECTATION

The expectation of the estimator assuming biased annotations is summarized in Theorem 2 which is restated below.

Theorem (Expectation of DM-IS$^+$ and DM$^+$-IS$^+$ under imperfect annotations)**.** *Under Assumptions 3 and 5, the two estimators have the same expectation:*

$$\mathbb{E}[\hat{V}^{\text{DM-IS}^+}] = \mathbb{E}[\hat{V}^{\text{DM}^+\text{-IS}^+}] = v(\pi_e) + \mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a\sim\pi_e(s_i)}[(1 - \frac{\bar{W}(a|s_i,a)\pi_b(a|s_i)}{\pi_b^+(a|s_i)})\epsilon_G(s_i,a)]]$$

Proof: We use the linearity of expectation, and the definition of the expectation.

$$\mathbb{E}_{D_0,D_{\hat{R}}\sim\mathcal{D}}[\hat{V}^{\text{DM-IS}^+}] \tag{139}$$

$$= \mathbb{E}_{\substack{D_{\hat{R}},s_i\sim d_0 a_i\sim\pi_b(s_i),\\ r\sim R(s_i,a_i),w\sim W(s_i,a_i)}}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{R}(s_i,\pi_e) + \sum_{a\in A} w_i^a \frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(c_i^a - \hat{R}(s_i,a))\right)\right] \tag{140}$$

$$= \mathbb{E}_{\substack{D_{\hat{R}},s_i\sim d_0 a_i\sim\pi_b(s_i),\\ r\sim R(s_i,a_i),w\sim W(s_i,a_i),\\ g_i^a\sim G(s_i,a)}}[\frac{1}{N}\sum_{i=1}^{N}(\hat{R}(s_i,\pi_e) + w_i^{a_i}\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(r_i - \hat{R}(s_i,a_i)) \tag{141}$$

$$+ \sum_{a\in A\setminus\{a_i\}} w_i^a \frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(g_i^a - \hat{R}(s_i,a)))] \tag{142}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{D_{\hat{R}},s_i\sim d_0}[\hat{R}(s_i,\pi_e)] \tag{143}$$

$$\underbrace{+ \mathbb{E}_{D_{\hat{R}},s_i\sim d_0 a_i\sim\pi_b(s_i),r_i\sim R(s_i,a_i),w\sim W(s_i,a_i)}[w_i^{a_i}\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(r_i - \hat{R}(s_i,a_i))]}_{2} \tag{144}$$

$$\underbrace{+\,\mathbb{E}_{D_{\hat{R}},s_i\sim d_0 a_i\sim\pi_b(s_i),w\sim W(s_i,a_i),g_i^a\sim G(s_i,a)}[\sum_{a\in A\setminus\{a_i\}} w_i^a\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(g_i^a-\hat{R}(s_i,a))]}_{3} \tag{145}$$

Now we simplify term 2 using the expectation of the weights and the expectation of scalar reward, and the definition of expectation:

$$\text{Term 2} = \mathbb{E}_{D_{\hat{R}},s\sim d_0 a_i\sim\pi_b,r\sim R,w\sim W}[w_i^{a_i}\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(r_i-\hat{R}(s_i,a_i))] \tag{146}$$

$$= \mathbb{E}_{D_{\hat{R}},s\sim d_0 a_i\sim\pi_b,r\sim R}[\mathbb{E}_{w\sim W}[w_i^{a_i}]\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(\mathbb{E}_{r\sim R}[r_i]-\hat{R}(s_i,a_i))] \tag{147}$$

$$= \mathbb{E}_{D_{\hat{R}},s\sim d_0 a_i\sim\pi_b,r\sim R}[\bar{W}(a_i|s_i,a_i)\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i))] \tag{148}$$

(149)

Now we simplify term 3 using a similar procedure:

$$\text{Term 3} = \mathbb{E}_{D_{\hat{R}},s\sim d_0 a\sim\pi_b,w\sim W,g\sim G}[\sum_{a\in A\setminus\{a_i\}} w_i^a\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(g_i^a-\hat{R}(s_i,a))] \tag{150}$$

$$= \mathbb{E}_{D_{\hat{R}},s\sim d_0 a\sim\pi_b}[\sum_{a\in A\setminus\{a_i\}} \mathbb{E}_{w\sim W}[w_i^a]\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(\mathbb{E}_{g\sim G}[g_i^a]-\hat{R}(s_i,a))] \tag{151}$$

$$= \mathbb{E}_{D_{\hat{R}},s\sim d_0 a\sim\pi_b}[\sum_{a\in A\setminus\{a_i\}} \bar{W}(a|s_i,a)\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(\bar{R}(s_i,a)+\epsilon_G-\hat{R}(s_i,a))] \tag{152}$$

(153)

Putting together the terms:

$$\mathbb{E}_{s\sim d_0}[\hat{R}(s_i,\pi_e)] + \mathbb{E}_{s\sim d_0 a_i\sim\pi_b,r\sim R}[\bar{W}(a_i|s_i,a_i)\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i))] \tag{154}$$

$$+\,\mathbb{E}_{s\sim d_0 a\sim\pi_b}[\sum_{a\in A\setminus\{a_i\}} \bar{W}(a|s_i,a)\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(\bar{R}(s_i,a)+\epsilon_G-\hat{R}(s_i,a))] \tag{155}$$

(156)

Note that we can simplify the terms:

$$= \mathbb{E}_{D_{\hat{R}},s\sim d_0}[\hat{R}(s_i,\pi_e)] \tag{157}$$

$$+\,\mathbb{E}_{D_{\hat{R}},s\sim d_0 a_i\sim\pi_b,r\sim R}[\sum_{a\in A}\bar{W}(a_i|s_i,a_i)\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i))] \tag{158}$$

$$+\,\mathbb{E}_{D_{\hat{R}},s\sim d_0 a\sim\pi_b}[\sum_{a\in A\setminus\{a_i\}} \bar{W}(a|s_i,a)\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(\epsilon_G)] \tag{159}$$

The sum of the first two terms becomes the value of the target policy:

$$\mathbb{E}_{D_{\hat{R}},s\sim d_0}[\hat{R}(s_i,\pi_e)] + \mathbb{E}_{D_{\hat{R}},s\sim d_0 a_i\sim\pi_b,r\sim R}[\sum_{a\in A}\bar{W}(a_i|s_i,a_i)\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i))] \tag{160}$$

$$= \mathbb{E}_{D_{\hat{R}},s\sim d_0}[\hat{R}(s_i,\pi_e)] \tag{161}$$

$$+ \mathbb{E}_{D_{\hat{R}}, s\sim d_0 r\sim R}[\sum_{a_i\in A}\pi_b(a_i|s_i)\sum_{a\in A}\bar{W}(a_i|s_i,a_i)\frac{\pi_e(a_i|s_i)}{\pi_{b^+}(a_i|s_i)}(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i))] \tag{162}$$

$$= \mathbb{E}_{D_{\hat{R}}, s\sim d_0}[\hat{R}(s_i,\pi_e)] + \mathbb{E}_{D_{\hat{R}}, s\sim d_0 r\sim R}[\pi_e(a_i|s_i)(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i))] \tag{163}$$

$$= \mathbb{E}_{D_{\hat{R}}, s\sim d_0}[\hat{R}(s_i,\pi_e)] + \mathbb{E}_{s\sim d_0 a_i\sim\pi_e, r\sim R}[(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i))] \tag{164}$$

$$= \mathbb{E}_{s\sim d_0 a_i\sim\pi_e}[(\bar{R}(s_i,a_i)] \tag{165}$$

$$= v(\pi_e) \tag{166}$$

The final term can be simplified as follows:

$$\mathbb{E}_{D_{\hat{R}}, s\sim d_0 a\sim\pi_b}[\sum_{\tilde{a}\in A\setminus\{a_i\}}\bar{W}(\tilde{a}|s_i,a)\frac{\pi_e(\tilde{a}|s_i)}{\pi_{b^+}(\tilde{a}|s_i)}\epsilon_G] \tag{167}$$

$$= \mathbb{E}_{D_{\hat{R}}, s\sim d_0}[\sum_{a\in A}\pi_b(a|s_i)\left(\sum_{\tilde{a}\in A\setminus\{a_i\}}\bar{W}(\tilde{a}|s_i,a)\frac{\pi_e(\tilde{a}|s_i)}{\pi_{b^+}(\tilde{a}|s_i)}\epsilon_G\right)] \tag{168}$$

$$= \mathbb{E}_{D_{\hat{R}}, s\sim d_0}[\sum_{\tilde{a}\in A}((\sum_{a\in A}\pi_b(a|s)\bar{W}(\tilde{a}|s,a))\frac{\pi_e(\tilde{a}|s)}{\pi_b^+(\tilde{a}|s}\epsilon_G))] \tag{169}$$

$$- \mathbb{E}_{D_{\hat{R}}, s\sim d_0}[(\sum_{a\in A}\pi_b(a|s)\bar{W}(a|s,a)\frac{\pi_e(a|s)}{\pi_b^+(a|s)}\epsilon_G)] \tag{170}$$

$$= \mathbb{E}_{D_{\hat{R}}, s\sim d_0}[\sum_{\tilde{a}\in A}\pi_e(\tilde{a}|s)\epsilon_G))] \tag{171}$$

$$- \mathbb{E}_{D_{\hat{R}}, s\sim d_0}[(\sum_{a\in A}\pi_e(a|s)\bar{W}(a|s,a)\frac{\pi_b(a|s)}{\pi_b^+(a|s)}\epsilon_G)] \tag{172}$$

$$= \mathbb{E}_{s\sim d_0}[\mathbb{E}_{a\sim\pi_e}[\left(1-\frac{\bar{W}(a|s,a)\pi_b(a|s)}{\pi_b^+(a|s)}\right)\epsilon_G]] \tag{173}$$

Thus, the final term for the expectation is:

$$v(\pi_e) + \mathbb{E}_{s\sim d_0}\left[\mathbb{E}_{a\sim\pi_e}\left[\left(1-\frac{\bar{W}(a|s,a)\pi_b(a|s)}{\pi_b^+(a|s)}\right)\epsilon_G\right]\right] \tag{174}$$

Proposition S14 (Unbiasedness of DM-IS$^+$)**.** *If both Assumptions 2 and 5 hold, the* DM-IS$^+$ *estimator is unbiased,* $\mathbb{E}[\hat{V}^{\text{DM-IS}^+}] = v(\pi_e)$.

Proof: If $\epsilon_G = 0$, then, the expectation term for the imperfect annotations case reduces to the value of the expert policy $v(\pi_e)$.

G.2.2. VARIANCE

We first study the variance of the DM-IS$^+$ estimator under imperfect annotations.

Proposition S15 (Variance of DM-IS$^+$ under imperfect annotations)**.** *In the case that Assumptions 3 to 5 hold,*

$$\mathbb{V}_{D\sim\mathcal{D}}[\hat{V}^{\text{DM-IS}^+}] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}_{s_i\sim d_0}\left[v(\pi_e) + \mathbb{E}_{\substack{s_i\sim d_0\\ a\sim\pi_e(\cdot|s_i)}}\left[(1-\frac{\bar{W}(a|s_i,a)\pi_b(a|s_i)}{\pi_b^+(a|s_i)})\epsilon_G(s_i,a)\right]\right] \tag{175}$$

$$+ \mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\Big[\frac{\pi_e(s_i|a_i)}{\pi_b^+(s_i|a_i)}\bar{W}(a_i|s_i,a_i)(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i)) \tag{176}$$

$$+ \sum_{a\in A\setminus\{a_i\}} \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)} \bar{W}(a|s_i,a_i)(\bar{R}(s_i,a)+\epsilon_R(s_i,a)-\hat{R}(s_i,a))\Big|a_i\Big]\Big|s_i\Big] \tag{177}$$

$$+ \mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b}\Big[\sum_{a\in A}\sum_{a\in A}\bar{W}(a|s_i,a_i)^2 {\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}}^2 \sigma_R^2(s_i,a) \tag{178}$$

$$+ \sum_{a\in A\setminus\{a_i\}} \bar{W}(a|s_i,a_i)^2 {\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}}^2 \Delta_G(s_i,a)\Big|s_i\Big]\Bigg] \tag{179}$$

$$+ \mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Big[{\frac{\pi_e(a_i|s_i)}{\pi_b^+(a_i|s_i)}}^2 \mathbb{V}_{w_i^a\sim W(s_i,a)}[w_i^{a_i}] \tag{180}$$

$$\times \Bigg(\sigma_R^2(s_i,a)+(\bar{R}(s_i,a)-\hat{R}(s_i,a))^2\Bigg) \tag{181}$$

$$+ \sum_{a\in A\setminus\{a_i\}} {\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}}^2 \mathbb{V}_{w_i^a\sim W(s_i,a)}[w_i^a] \tag{182}$$

$$\times \Bigg(\sigma_R^2(s_i,a)+\Delta_G(s_i,a)+(\bar{R}(s_i,a)+\epsilon_R(s_i,a)-\hat{R}(s_i,a))^2\Bigg) \tag{183}$$

$$+ 2\bar{\psi}_R(s_i,a_i)\sum_{\tilde{a}_j\neq\tilde{a}_i}\bar{\psi}_G(s_i,\tilde{a}_j)\times Cov_{w_i^a\sim W(s_i,a)}\Big(w_i^{\tilde{a}_j}, w_i^{a_i}\Big) \tag{184}$$

$$+ \sum_{\tilde{a}_j\neq a_i}\ \sum_{\tilde{a}_k\neq a_i,\tilde{a}_k\neq\tilde{a}_j}\bar{\psi}_G(s_i,\tilde{a}_j)\bar{\psi}_G(s_i,\tilde{a}_k) \tag{185}$$

$$\times Cov_{w_i^a\sim W(s_i,a)}\Bigg(w_i^{\tilde{a}_j}, w_i^{\tilde{a}_k}\Bigg)\Bigg|a_i\Bigg]\Bigg|s_i\Bigg] \tag{186}$$

Proof: We start by moving the variance inside the summation because the variance of each term that the summation is over is independent. This is a similar strategy proposed in the Let $c_i^a \sim C(s_i,a_i)$ be $r_i^a \sim R(s_i,a)$ if c_i^a is a factual reward and $g_i^a \sim G(s_i,a)$ if c_i^a is a counterfactual annotation.

$$\mathbb{V}_{D\sim\mathcal{D}}[\hat{V}^{\text{DM-IS}^+}] \tag{187}$$

$$= \mathbb{V}_{\substack{s_i\sim d_0 a_i\sim\pi_b(\cdot|s_i),\\ c_i^a\sim C(s_i,a), w_i^a\sim W(s_i,a)}}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a \frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(c_i^a-\hat{R}(s_i,a))\right)\right] \tag{188}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}_{\substack{s_i\sim d_0 a_i\sim\pi_b(\cdot|s_i),\\ c_i^a\sim C(s_i,a), w_i^a\sim W(s_i,a)}}\left[\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a \frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(c_i^a-\hat{R}(s_i,a))\right] \tag{189}$$

Now, we use the law of total variance to separate the joint distribution that the variance is over:

$$= \frac{1}{N^2}\sum_{i=1}^{N}\underbrace{\mathbb{E}_{s_i\sim d_0}\left[\mathbb{V}_{\substack{a_i\sim\pi_b(\cdot|s_i),\\ c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\left[\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Big|s_i\right]\right]}_{2} \tag{190}$$

$$+\mathbb{V}_{s_i\sim d_0}\Bigg[\underbrace{\mathbb{E}_{\substack{a_i\sim\pi_b(\cdot|s_i),\\ c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Bigg[\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Bigg| s_i\Bigg]}_{1}\Bigg] \tag{191}$$

Note that Term 1 is just the value of the target policy conditioned on a specific s_i. We have already proved the bias of the estimator under imperfect annotations. This makes the whole variance now:

$$=\frac{1}{N^2}\sum_{i=1}^{N}\underbrace{\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{\substack{a_i\sim\pi_b(\cdot|s_i),\\ c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Bigg[\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Bigg| s_i\Bigg]\Bigg]}_{2} \tag{192}$$

$$+\mathbb{V}_{s_i\sim d_0}\Bigg[v(\pi_e)+\mathbb{E}_{\substack{s_i\sim d_0\\ a\sim\pi_e(\cdot|s_i)}}\Bigg[\Bigg(1-\frac{\bar{W}(a|s_i,a)\pi_b(a|s_i)}{\pi_b^+(a|s_i)}\Bigg)\epsilon_G(s_i,a)\Bigg]\Bigg] \tag{193}$$

Now we decompose term 2 using the law of total variance:

$$\text{Term } 2=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{\substack{a_i\sim\pi_b(\cdot|s_i),\\ c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Bigg[\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Bigg| s_i\Bigg]\Bigg] \tag{194}$$

$$=\mathbb{E}_{s_i\sim d_0}\Bigg[\underbrace{\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\mathbb{V}_{\substack{c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Bigg[\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Bigg| a_i\Bigg]\Bigg| s_i\Bigg]}_{3} \tag{195}$$

$$+\underbrace{\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\mathbb{E}_{\substack{c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Bigg[\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Bigg| a_i\Bigg]\Bigg| s_i\Bigg]}_{4}\Bigg] \tag{196}$$

Now we decompose term 4 by distributing the inner-most expectation. Note that the inner-most expectation only depends on the distribution over c_i and w_i, which is not in many of the terms:

$$\text{Term } 4=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\mathbb{E}_{\substack{c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Bigg[\hat{R}(s_i,\pi_e)+\sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Bigg| a_i\Bigg]\Bigg| s_i\Bigg]\Bigg] \tag{197}$$

$$=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\hat{R}(s_i,\pi_e)+\mathbb{E}_{\substack{c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Bigg[\sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Bigg| a_i\Bigg]\Bigg| s_i\Bigg]\Bigg] \tag{198}$$

Applying linearity of expectation and the fact that c, w are conditionally independent given s, a:

$$=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\hat{R}(s_i,\pi_e)+\sum_{a\in A}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}\mathbb{E}_{\substack{c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Big[w_i^a(c_i^a-\hat{R}(s_i,a))\Big| a_i\Big]\Bigg| s_i\Bigg]\Bigg] \tag{199}$$

$$=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\sum_{a\in A}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}\mathbb{E}_{\substack{c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\Big[w_i^a(c_i^a-\hat{R}(s_i,a))\Big| a_i\Big]\Bigg| s_i\Bigg]\Bigg] \tag{200}$$

$$\tag{201}$$

Now, separating into factual and counterfactual:

$$=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\Big[\frac{\pi_e(s_i|a_i)}{\pi_b^+(s_i|a_i)}\bar{W}(a_i|s_i,a_i)(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i)) \tag{202}$$

$$+\sum_{a\in A\setminus\{a_i\}}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}\bar{W}(a|s_i,a_i)(\bar{R}(s_i,a)+\epsilon_R(s_i,a)-\hat{R}(s_i,a))\Big|a_i,s_i\Big]\Big] \tag{203}$$

$$\tag{204}$$

We cannot decompose this term any further because each weight is dependent on a_i, and the variance is over the distribution of a_i. Now we decompose term 3 using the law of total variance:

$$\text{Term 3}=\mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\left[\mathbb{V}_{\substack{c_i^a\sim C(s_i,a),\\ w_i^a\sim W(s_i,a)}}\left[\hat{R}(s_i,\pi_e)+\sum_{a\in A}w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\middle|a_i\right]\middle|s_i\right]\right] \tag{205}$$

$$=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\underbrace{\mathbb{E}_{c_i^a\sim C(s_i,a)}\left[\mathbb{V}_{w_i^a\sim W(s_i,a)}\left[\hat{R}(s_i,\pi_e)+\sum_{a\in A}w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\middle|c_i\right]\middle|a_i\right]}_{5} \tag{206}$$

$$\underbrace{+\mathbb{V}_{c_i^a\sim C(s_i,a)}\left[\mathbb{E}_{w_i^a\sim W(s_i,a)}\left[\hat{R}(s_i,\pi_e)+\sum_{a\in A}w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\middle|c_i\right]\middle|a_i\right]}_{6}\Bigg|s_i\Bigg]\Bigg] \tag{207}$$

Now we decompose term 6 by distributing the inner-most expectation. Most terms do not have a weight in them:

$$\text{Term 6}= \tag{208}$$

$$\mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\left[\mathbb{V}_{c_i^a\sim C(s_i,a)}\left[\mathbb{E}_{w_i^a\sim W(s_i,a)}\left[\hat{R}(s_i,\pi_e)+\sum_{a\in A}w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\middle|c_i\right]\middle|a_i\right]\middle|s_i\right]\right] \tag{209}$$

$$=\mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\left[\mathbb{V}_{c_i^a\sim R(s_i,a_i)}\left[\hat{R}(s_i,\pi_e)+\sum_{a\in A}\mathbb{E}_{w_i^a\sim W(s_i,a)}\left[w_i^a\right]\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\middle|a_i\right]\middle|s_i\right]\right] \tag{210}$$

$$=\mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\left[\mathbb{V}_{c_i^a\sim C(s_i,a)}\left[\hat{R}(s_i,\pi_e)+\sum_{a\in A}\bar{W}(a|s_i,a_i)\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\middle|a_i\right]\middle|s_i\right]\right] \tag{211}$$

Notice that the inner-most variance term now is over c_i^a, so all terms that do not consider c_i^a are considered constants.

$$=\mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\left[\sum_{a\in A}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\left(\mathbb{V}_{c_i^a\sim C(s_i,a)}\left[c_i|a_i\right]\right)|s_i\right]\right] \tag{212}$$

Now, because the annotations are assumed to be imperfect, we must split the variance term into a factual sample and several possible counterfactual annotations.

$$=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\bar{W}(a_i|s_i,a_i)^2\frac{\pi_e(a_i|s_i)}{\pi_b^+(a_i|s_i)}^2\sigma_R^2(s_i,a_i) \tag{213}$$

$$+\sum_{a\in A\setminus\{a_i\}}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\left(\sigma_R^2(s_i,a)+\Delta_G(s_i,a)\right)\Bigg|s_i\Bigg]\Bigg] \tag{214}$$

$$=\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\sum_{a\in A}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\sigma_R^2(s_i,a) \tag{215}$$

$$+ \sum_{a\in A\setminus\{a_i\}} \bar{W}(a|s_i,a_i)^2 \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2 \Delta_G(s_i,a)\Bigg| s_i\Bigg]\Bigg] \tag{216}$$

$$\tag{217}$$

Now we decompose term 5. Because the weights are not independent (they need to sum to 1 across all the actions/annotations for a given context), we need to consider the covariance between them. First, we move the variance term outside the summation because $\hat{R}(s_i,\pi_e)$ is considered a constant.

$$\text{Term 5} = \mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)(\cdot|s_i)}\Bigg[\mathbb{E}_{c_i^a\sim C(s_i,a)}\Bigg[\mathbb{V}_{w_i^a\sim W(s_i,a)}\Bigg[\hat{R}(s_i,\pi_e) \tag{218}$$

$$+ \sum_{a\in A} w_i^a \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a - \hat{R}(s_i,a))\Bigg|c_i\Bigg]\Bigg|a_i\Bigg]\Bigg|s_i\Bigg]\Bigg] \tag{219}$$

$$= \mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)(\cdot|s_i)}\Bigg[\mathbb{E}_{c_i^a\sim C(s_i,a)}\Bigg[\mathbb{V}_{w_i^a\sim W(s_i,a)}\Bigg[\sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))\Bigg|c_i\Bigg]\Bigg|a_i\Bigg]\Bigg|s_i\Bigg]\Bigg] \tag{220}$$

When we move the variance inside the summation if we also consider the covariance of each term in the summation. Then, we simplify and consider all terms that don't contain a weight as constant.

$$= \mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\mathbb{E}_{c_i^a\sim C(s_i,a)}\Bigg[\sum_{a\in A}\mathbb{V}_{w_i^a\sim W(s_i,a)}\Bigg[w_i^a\frac{\pi_e(a|s)}{\pi_b^+(a|s)}(c_i^a-\hat{R}(s_i,a))\Bigg|c_i\Bigg] \tag{221}$$

$$+ 2\sum_{\tilde{a}_j,\tilde{a}_k}^{\tilde{a}_j\neq\tilde{a}_k} Cov_{w_i^a\sim W(s_i,a)}\Bigg(w_i^{\tilde{a}_j}\frac{\pi_e(\tilde{a}_j|s_i)}{\pi_b^+(\tilde{a}_j|s_i)}(c_i^{\tilde{a}_j}-\hat{R}(s_i,\tilde{a}_j)), w_i^{\tilde{a}_k}\frac{\pi_e(\tilde{a}_k|s_i)}{\pi_b^+(\tilde{a}_k|s_i)}\Bigg(c_i^{\tilde{a}_k}-\hat{R}(s_i,\tilde{a}_k))\Bigg)\Bigg|a_i\Bigg]\Bigg|s_i\Bigg]\Bigg] \tag{222}$$

$$= \mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\mathbb{E}_{c_i^a\sim C(s_i,a)}\Bigg[\sum_{a\in A}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2(c_i^a-\hat{R}(s_i,a))^2\mathbb{V}_{w_i^a\sim W(s_i,a)}\Big[w_i^a\Big] \tag{223}$$

$$+ 2\sum_{\tilde{a}_j,\tilde{a}_k}^{\tilde{a}_j\neq\tilde{a}_k}\frac{\pi_e(\tilde{a}_j|s_i)}{\pi_b^+(\tilde{a}_j|s_i)}(c_i^{\tilde{a}_j}-\hat{R}(s_i,\tilde{a}_j))\frac{\pi_e(\tilde{a}_k|s_i)}{\pi_b^+(\tilde{a}_k|s_i)}(c_i^{\tilde{a}_k}-\hat{R}(s_i,\tilde{a}_k)) \tag{224}$$

$$\times Cov_{w_i^a\sim W(s_i,a)}\Bigg(w_i^{\tilde{a}_j}, w_i^{\tilde{a}_k}\Bigg)\Bigg|a_i\Bigg]\Bigg|s_i\Bigg]\Bigg] \tag{225}$$

Now, we separate the terms to account for different distributions of the factual rewards and counterfactual annotations. Let $\psi(s_i,a)=\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(c_i^a-\hat{R}(s_i,a))$. Then, we can re-write the expression as:

$$\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\mathbb{E}_{c_i^a\sim C(s_i,a)}\Bigg[\sum_{a\in A}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2(c_i^a-\hat{R}(s_i,a))^2\mathbb{V}_{w_i^a\sim W(s_i,a)}\Big[w_i^a\Big] \tag{226}$$

$$+ 2\sum_{\tilde{a}_j\neq\tilde{a}_k}\psi(s_i,\tilde{a}_j)\psi(s_i,\tilde{a}_k) \tag{227}$$

$$\times Cov_{w_i^a\sim W(s_i,a)}\Bigg(w_i^{\tilde{a}_j}, w_i^{\tilde{a}_k}\Bigg)\Bigg|a_i\Bigg]|s_i\Bigg]\Bigg] \tag{228}$$

Now, splitting into factual and counterfactual:

$$\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\frac{\pi_e(a_i|s_i)}{\pi_b^+(a_i|s_i)}^2\mathbb{E}_{r_i^{a_i}\sim R(s_i,a_i)}\Bigg[(r_i^a-\hat{R}(s_i,a))^2]\mathbb{V}_{w_i^a\sim W(s_i,a)}\Big[w_i^{a_i}\Big] \tag{229}$$

$$+ \sum_{a \in A \setminus \{a_i\}} \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2 \mathbb{E}_{g_i^a \sim G(s_i,a)}\Big[(g_i^a - \hat{R}(s_i,a))^2\Big] \mathbb{V}_{w_i^a \sim W(s_i,a)}\Big[w_i^a\Big] \tag{230}$$

$$+ 2\mathbb{E}_{r_i^{a_i} \sim R(s_i,a_i)}\Big[\psi(s_i,a_i)\Big] \sum_{\tilde{a}_j \neq \tilde{a}_i} \mathbb{E}_{g_i^a \sim G(s_i,a)}\Big[\psi(s_i,\tilde{a}_j)\Big] \times Cov_{w_i^a \sim W(s_i,a)}\Big(w_i^{\tilde{a}_j}, w_i^{a_i}\Big) \tag{231}$$

$$+ \sum_{\tilde{a}_j \neq a_i} \sum_{\tilde{a}_k \neq a_i, \tilde{a}_k \neq \tilde{a}_j} \mathbb{E}_{g_i^a \sim G(s_i,a)}\Big[\psi(s_i,\tilde{a}_j)\psi(s_i,\tilde{a}_k)\Big] \tag{232}$$

$$\times Cov_{w_i^a \sim W(s_i,a)}\Bigg(w_i^{\tilde{a}_j}, w_i^{\tilde{a}_k}\Bigg)\Bigg|a_i\Bigg]\Bigg|s_i\Bigg]\Bigg] \tag{233}$$

Recall that $\mathbb{E}_{r_i^{a_i} \sim R(s_i,a_i)}[\psi(s_i,a)] = \bar{\psi}_R(s_i,a) = \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(\bar{R}(s_i,a) - \hat{R}(s_i,a))$ and $\mathbb{E}_{g_i^a \sim G(s_i,a)}[\psi(s_i,a)] = \bar{\psi}_G(s_i,a) = \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}(\bar{R}(s_i,a) + \epsilon_R(s_i,a) - \hat{R}(s_i,a))$. Also recall that the product of expectation of two independent terms is the expectation of the product. Thus:

$$\mathbb{E}_{s_i \sim d_0}\Bigg[\mathbb{E}_{a_i \sim \pi_b(\cdot|s_i)}\Bigg[\frac{\pi_e(a_i|s_i)}{\pi_b^+(a_i|s_i)}^2 \mathbb{E}_{r_i^{a_i} \sim R(s_i,a_i)}\Big[(r_i^a - \hat{R}(s_i,a))^2\Big] \mathbb{V}_{w_i^a \sim W(s_i,a)}\Big[w_i^{a_i}\Big] \tag{234}$$

$$+ \sum_{a \in A \setminus \{a_i\}} \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2 \mathbb{E}_{g_i^a \sim G(s_i,a)}\Big[(g_i^a - \hat{R}(s_i,a))^2\Big] \mathbb{V}_{w_i^a \sim W(s_i,a)}\Big[w_i^a\Big] \tag{235}$$

$$+ 2\bar{\psi}_R(s_i,a_i) \sum_{\tilde{a}_j \neq \tilde{a}_i} \bar{\psi}_G(s_i,\tilde{a}_j) \times Cov_{w_i^a \sim W(s_i,a)}\Big(w_i^{\tilde{a}_j}, w_i^{a_i}\Big) \tag{236}$$

$$+ \sum_{\tilde{a}_j \neq a_i} \sum_{\tilde{a}_k \neq a_i, \tilde{a}_k \neq \tilde{a}_j} \bar{\psi}_G(s_i,\tilde{a}_j)\bar{\psi}_G(s_i,\tilde{a}_k) \tag{237}$$

$$\times Cov_{w_i^a \sim W(s_i,a)}\Big(w_i^{\tilde{a}_j}, w_i^{\tilde{a}_k}\Big)\Bigg|a_i\Bigg]\Bigg|s_i\Bigg] \tag{238}$$

Now we use the definition of variance to simplify the first two expectation squared terms.

$$\mathbb{E}_{s_i \sim d_0}\Bigg[\mathbb{E}_{a_i \sim \pi_b(\cdot|s_i)}\Bigg[\frac{\pi_e(a_i|s_i)}{\pi_b^+(a_i|s_i)}^2 \mathbb{V}_{w_i^a \sim W(s_i,a)}\Big[w_i^{a_i}\Big] \tag{239}$$

$$\times \Bigg(\mathbb{V}_{r_i^{a_i} \sim R(s_i,a_i)}\Big[(r_i^a - \hat{R}(s_i,a))\Big] + \mathbb{E}_{r_i^a \sim R(s_i,a}\Big[(r_i^a - \hat{R}(s_i,a))\Big]^2\Bigg) \tag{240}$$

$$+ \sum_{a \in A \setminus \{a_i\}} \frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2 \mathbb{V}_{w_i^a \sim W(s_i,a)}\Big[w_i^a\Big] \tag{241}$$

$$\times \Bigg(\mathbb{V}_{g_i^a \sim G(s_i,a)}\Big[(g_i^a - \hat{R}(s_i,a))\Big] + \mathbb{E}_{g_i^a \sim G(s_i,a)}\Big[(g_i^a - \hat{R}(s_i,a))\Big]^2\Bigg) \tag{242}$$

$$+ 2\bar{\psi}_R(s_i,a_i) \sum_{\tilde{a}_j \neq \tilde{a}_i} \bar{\psi}_G(s_i,\tilde{a}_j) \times Cov_{w_i^a \sim W(s_i,a)}\Big(w_i^{\tilde{a}_j}, w_i^{a_i}\Big) \tag{243}$$

$$+ \sum_{\tilde{a}_j \neq a_i} \sum_{\tilde{a}_k \neq a_i, \tilde{a}_k \neq \tilde{a}_j} \bar{\psi}_G(s_i,\tilde{a}_j)\bar{\psi}_G(s_i,\tilde{a}_k) \tag{244}$$

$$\times Cov_{w_i^a \sim W(s_i,a)}\Big(w_i^{\tilde{a}_j}, w_i^{\tilde{a}_k}\Big)\Bigg|a_i\Bigg]\Bigg|s_i\Bigg] \tag{245}$$

Simplifying:

$$\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Bigg[\frac{\pi_e(a_i|s_i)}{\pi_b^+(a_i|s_i)}^2\mathbb{V}_{w_i^a\sim W(s_i,a)}\Big[w_i^{a_i}\Big] \tag{246}$$

$$\times\Bigg(\sigma_R^2(s_i,a)+(\bar{R}(s_i,a)-\hat{R}(s_i,a))^2\Bigg) \tag{247}$$

$$+\sum_{a\in A\setminus\{a_i\}}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\mathbb{V}_{w_i^a\sim W(s_i,a)}\Big[w_i^a\Big] \tag{248}$$

$$\times\Bigg(\sigma_R^2(s_i,a)+\Delta_G(s_i,a)+(\bar{R}(s_i,a)+\epsilon_R(s_i,a)-\hat{R}(s_i,a))^2\Bigg) \tag{249}$$

$$+2\bar{\psi}_R(s_i,a_i)\sum_{\tilde{a}_j\neq\tilde{a}_i}\bar{\psi}_G(s_i,\tilde{a}_j)\times Cov_{w_i^a\sim W(s_i,a)}\Big(w_i^{\tilde{a}_j},w_i^{a_i}\Big) \tag{250}$$

$$+\sum_{\tilde{a}_j\neq a_i}\sum_{\tilde{a}_k\neq a_i,\tilde{a}_k\neq\tilde{a}_j}\bar{\psi}_G(s_i,\tilde{a}_j)\bar{\psi}_G(s_i,\tilde{a}_k) \tag{251}$$

$$\times\, Cov_{w_i^a\sim W(s_i,a)}\Big(w_i^{\tilde{a}_j},w_i^{\tilde{a}_k}\Big)\Bigg|a_i\Bigg]\Bigg|s_i\Bigg] \tag{252}$$

Putting together the variance term:

$$\mathbb{V}_{D\sim\mathcal{D}}[\hat{V}^{\text{DM-IS}^+}]= \tag{253}$$

$$\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}_{s_i\sim d_0}\Bigg[v(\pi_e)+\mathbb{E}_{\substack{s_i\sim d_0\\ a\sim\pi_e(\cdot|s_i)}}\Big[(1-\frac{\bar{W}(a|s_i,a)\pi_b(a|s_i)}{\pi_b^+(a|s_i)})\epsilon_G(s_i,a)\Big]\Bigg] \tag{254}$$

$$+\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\Big[\frac{\pi_e(s_i|a_i)}{\pi_b^+(s_i|a_i)}\bar{W}(a_i|s_i,a_i)(\bar{R}(s_i,a_i)-\hat{R}(s_i,a_i)) \tag{255}$$

$$+\sum_{a\in A\setminus\{a_i\}}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}\bar{W}(a|s_i,a_i)(\bar{R}(s_i,a)+\epsilon_R(s_i,a)-\hat{R}(s_i,a))\Big|a_i\Big]\Bigg|s_i\Bigg] \tag{256}$$

$$+\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b}\Big[\sum_{a\in A}\sum_{a\in A}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\sigma_R^2(s_i,a) \tag{257}$$

$$+\sum_{a\in A\setminus\{a_i\}}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\Delta_G(s_i,a)\Big|s_i\Big]\Bigg] \tag{258}$$

$$+\mathbb{E}_{s_i\sim d_0}\Bigg[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\Big[\frac{\pi_e(a_i|s_i)}{\pi_b^+(a_i|s_i)}^2\mathbb{V}_{w_i^a\sim W(s_i,a)}[w_i^{a_i}] \tag{259}$$

$$\times\Bigg(\sigma_R^2(s_i,a)+(\bar{R}(s_i,a)-\hat{R}(s_i,a))^2\Bigg) \tag{260}$$

$$+\sum_{a\in A\setminus\{a_i\}}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\mathbb{V}_{w_i^a\sim W(s_i,a)}[w_i^a] \tag{261}$$

$$\times\Bigg(\sigma_R^2(s_i,a)+\Delta_G(s_i,a)+(\bar{R}(s_i,a)+\epsilon_R(s_i,a)-\hat{R}(s_i,a))^2\Bigg) \tag{262}$$

$$+ 2\bar{\psi}_R(s_i, a_i) \sum_{\tilde{a}_j \neq \tilde{a}_i} \bar{\psi}_G(s_i, \tilde{a}_j) \times Cov_{w_i^a \sim W(s_i,a)}\left(w_i^{\tilde{a}_j}, w_i^{a_i}\right) \tag{263}$$

$$+ \sum_{\tilde{a}_j \neq a_i} \sum_{\tilde{a}_k \neq a_i, \tilde{a}_k \neq \tilde{a}_j} \bar{\psi}_G(s_i, \tilde{a}_j)\bar{\psi}_G(s_i, \tilde{a}_k) \tag{264}$$

$$\times Cov_{w_i^a \sim W(s_i,a)}\left(w_i^{\tilde{a}_j}, w_i^{\tilde{a}_k}\right)\Bigg|a_i\Bigg]\Bigg|s_i\Bigg] \tag{265}$$

The variance of DM-IS$^+$ under perfect annotations is a slight variant to the expression derived under imperfect annotations.

Proposition S16 (Variance of DM-IS$^+$ under perfect annotations)**.** *If Assumptions 2 and 5 hold,*

$$\mathbb{V}_{D\sim\mathcal{D}}[\hat{V}^{\text{DM-IS}^+}] = \frac{1}{N^2}\sum_{i=1}^{N} \mathbb{V}_{s_i\sim d_0}[\hat{V}^{\pi_e}(s_i)] \tag{266}$$

$$+ \mathbb{E}_{s_i\sim d_0}\left[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\left[\sum_{a\in A}\frac{\pi_e(a|s)}{\pi_b^+(a|s)}\bar{W}(a|s_i,a_i)(\bar{R}(s_i,a) - \hat{R}(s_i,a))\Bigg|s_i\right]\right] \tag{267}$$

$$+ \mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b}\left[\sum_{a\in A}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s)}{\pi_b^+(a|s)}^2\sigma_R(s_i,a)^2\Bigg|s_i\right]\right] \tag{268}$$

$$+ \mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}[\sum_{a\in A}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\mathbb{V}_{w_i^a\sim W(s_i,a_i)}[w_i^a] \tag{269}$$

$$\times \left(\sigma_R(s_i,a)^2 + (\bar{R}(s_i,a) - \hat{R}(s_i,a)^2)\right) \tag{270}$$

$$+ 2\sum_{\tilde{a}_j,\tilde{a}_k}^{\tilde{a}_j\neq\tilde{a}_k}\frac{\pi_e(\tilde{a}_j|s_i)}{\pi_b^+(\tilde{a}_j|s_i)}\frac{\pi_e(\tilde{a}_k|s_i)}{\pi_b^+(\tilde{a}_k|s_i)} \tag{271}$$

$$\times Cov_{w_i^a\sim W(s_i,a)}(w_i^{\tilde{a}_j}, w_i^{\tilde{a}_k})(\bar{R}(s_i,\tilde{a}_j) - \hat{R}(s_i,\tilde{a}_j))(\bar{R}(s_i,\tilde{a}_k) - \hat{R}(s_i,\tilde{a}_k))|a_i]|s_i]] \tag{272}$$

Proof. The expression follows from the prior proof for DM-IS$^+$ under imperfect annotations, except with $\Delta_G, \epsilon_G = 0$. □

G.3. Expectation and variance of DM$^+$-IS$^+$

The DR estimator is defined as

$$\hat{V}^{\text{DM}^+\text{-IS}^+} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{R}^+(s_i,\pi_e) + \sum_{a\in A} w_i^a\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}(c_i^a - \hat{R}^+(s_i,a))\right)$$

G.3.1. EXPECTATION

The expectation without making any assumptions about the quality of the counterfactual annotations is summarized in Theorem 2 which is restated below.

Theorem (Expectation of DM-IS$^+$ and DM$^+$-IS$^+$ under imperfect annotations)**.** *Under Assumptions 3 and 5, the two estimators have the same expectation:*

$$\mathbb{E}[\hat{V}^{\text{DM-IS}^+}] = \mathbb{E}[\hat{V}^{\text{DM}^+\text{-IS}^+}] = v(\pi_e) + \mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a\sim\pi_e(s_i)}[(1 - \frac{\bar{W}(a|s_i,a)\pi_b(a|s_i)}{\pi_b^+(a|s_i)})\epsilon_G(s_i,a)]]$$

The proof is nearly identical to the one for the DM-IS$^+$ estimator, except that the reward function $\hat{R}$ is replaced by $\hat{R}^+$.

Proposition S17 (Unbiasedness of DM$^+$-IS$^+$)**.** *If both Assumptions 2 and 5 hold, the* DM$^+$-IS$^+$ *estimator is unbiased,* $\mathbb{E}[\hat{V}^{\text{DM}^+\text{-IS}^+}] = v(\pi_e)$.

Proof: The proof is identical to the unbiasedness proof for the DM-IS$^+$ estimator.

G.3.2. VARIANCE

We first discuss the variance under imperfect annotations.

Proposition S18 (Variance of DM$^+$-IS$^+$ under imperfect annotations)**.**

$$\mathbb{V}_{D\sim\mathcal{D}}\left[\hat{V}^{\text{DM}^+\text{-IS}^+}\right] = \tag{273}$$

$$\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}_{s_i\sim d_0}\left[v(\pi_e) + \mathbb{E}_{\substack{s_i\sim d_0\\ a\sim\pi_e(\cdot|s_i)}}\left[(1-\frac{\bar{W}(a|s_i,a)\pi_b(a|s_i)}{\pi_b^+(a|s_i)})\epsilon_G(s_i,a)\right]\right] \tag{274}$$

$$+\,\mathbb{E}_{s_i\sim d_0}\left[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\left[\frac{\pi_e(s_i|a_i)}{\pi_b^+(s_i|a_i)}\bar{W}(a_i|s_i,a_i)(\bar{R}(s_i,a_i)-\hat{R}^+(s_i,a_i))\right.\right. \tag{275}$$

$$\left.\left.+\sum_{a\in A\setminus\{a_i\}}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}\bar{W}(a|s_i,a_i)(\bar{R}(s_i,a)+\epsilon_R(s_i,a)-\hat{R}^+(s_i,a))\middle|a_i\right]\middle|s_i\right] \tag{276}$$

$$+\,\mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b}\left[\sum_{a\in A}\sum_{a\in A}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\sigma_R^2(s_i,a)\right.\right. \tag{277}$$

$$\left.\left.+\sum_{a\in A\setminus\{a_i\}}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\Delta_G(s_i,a)\middle|s_i\right]\right] \tag{278}$$

$$+\,\mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}\left[\frac{\pi_e(a_i|s_i)}{\pi_b^+(a_i|s_i)}^2\mathbb{V}_{w_i^a\sim W(s_i,a)}[w_i^{a_i}]\right.\right. \tag{279}$$

$$\times\left(\sigma_R^2(s_i,a)+(\bar{R}(s_i,a)-\hat{R}^+(s_i,a))^2\right) \tag{280}$$

$$+\sum_{a\in A\setminus\{a_i\}}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\mathbb{V}_{w_i^a\sim W(s_i,a)}[w_i^a] \tag{281}$$

$$\times\left(\sigma_R^2(s_i,a)+\Delta_G(s_i,a)+(\bar{R}(s_i,a)+\epsilon_R(s_i,a)-\hat{R}^+(s_i,a))^2\right) \tag{282}$$

$$+\,2\bar{\psi}_R(s_i,a_i)\sum_{\tilde{a}_j\neq\tilde{a}_i}\bar{\psi}_G(s_i,\tilde{a}_j)\times Cov_{w_i^a\sim W(s_i,a)}\left(w_i^{\tilde{a}_j},w_i^{a_i}\right) \tag{283}$$

$$+\sum_{\tilde{a}_j\neq a_i}\sum_{\tilde{a}_k\neq a_i,\tilde{a}_k\neq\tilde{a}_j}\bar{\psi}_G(s_i,\tilde{a}_j)\bar{\psi}_G(s_i,\tilde{a}_k) \tag{284}$$

$$\left.\left.\times\, Cov_{w_i^a\sim W(s_i,a)}\left(w_i^{\tilde{a}_j},w_i^{\tilde{a}_k}\right)\middle|a_i\right]\middle|s_i\right] \tag{285}$$

The proof is identical to that of Proposition S15, except the reward model $\hat{R}$ is replaced with $\hat{R}^+$. Now, we derive the variance under the perfect annotation setting.

Proposition S19 (Variance of DM$^+$-IS$^+$ under perfect annotations)**.** *If Assumptions 2 and 5 holds,*

$$\mathbb{V}_{D\sim\mathcal{D}}[\hat{V}^{\text{DM}^+\text{-IS}^+}] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}_{s_i\sim d_0}[\hat{V}^{\pi_e}(s_i)] \tag{286}$$

$$+\,\mathbb{E}_{s_i\sim d_0}\left[\mathbb{V}_{a_i\sim\pi_b(\cdot|s_i)}\left[\sum_{a\in A}\frac{\pi_e(a|s)}{\pi_b^+(a|s)}\bar{W}(a|s_i,a_i)(\bar{R}(s_i,a)-\hat{R}^+(s_i,a))\middle|s_i\right]\right] \tag{287}$$

$$+\,\mathbb{E}_{s_i\sim d_0}\left[\mathbb{E}_{a_i\sim\pi_b}\left[\sum_{a\in A}\bar{W}(a|s_i,a_i)^2\frac{\pi_e(a|s)}{\pi_b^+(a|s)}^2\sigma_R^2(s_i,a)\middle|s_i\right]\right] \tag{288}$$

$$+\,\mathbb{E}_{s_i\sim d_0}[\mathbb{E}_{a_i\sim\pi_b(\cdot|s_i)}[\sum_{a\in A}\frac{\pi_e(a|s_i)}{\pi_b^+(a|s_i)}^2\mathbb{V}_{w_i^a\sim W(s_i,a_i)}[w_i^a] \tag{289}$$

$$\times\left(\sigma_R^2(s_i,a)+(\bar{R}(s_i,a)-\hat{R}^+(s_i,a)^2\right) \tag{290}$$

$$+\,2\sum_{\tilde{a}_j,\tilde{a}_k}^{\tilde{a}_j\neq\tilde{a}_k}\frac{\pi_e(\tilde{a}_j|s_i)}{\pi_b^+(\tilde{a}_j|s_i)}\frac{\pi_e(\tilde{a}_k|s_i)}{\pi_b^+(\tilde{a}_k|s_i)} \tag{291}$$

$$\times\,Cov_{w_i^a\sim W(s_i,a)}(w_i^{\tilde{a}_j},w_i^{\tilde{a}_k})(\bar{R}(s_i,\tilde{a}_j)-\hat{R}^+(s_i,\tilde{a}_j))(\bar{R}(s_i,\tilde{a}_k)-\hat{R}^+(s_i,\tilde{a}_k))|a_i]|s_i]] \tag{292}$$

The proof is identical to the one used to derive variance under perfect conditions for the DM-IS$^+$ estimator, except with $\hat{R}^+$ replacing $\hat{R}$.

Appendix H. Equivalence of IS$^+$, DM-IS$^+$, DM$^+$-IS$^+$ under equal weights

Corollary S20 (Equivalence of IS$^+$, DM$^+$-IS, DM$^+$-IS$^+$ under equal weights)**.** *If we assume that all the weights are equal, and that all OPE methods have access to the same set of counterfactual annotations,* $\hat{V}^{\text{IS}^+}$*,* $\hat{V}^{\text{DM-IS}^+}$*,* $\hat{V}^{\text{DM}^+\text{-IS}^+}$ *are equivalent by definition.*

Proof: First we re-state the definition of the three methods for a single sample s_i, a_i, r_i, and all counterfactual annotations c_i^a, under the assumption that all weights are equal. Under this assumption, each weight is $\frac{1}{|A|}$:

1.

$$\hat{V}^{\text{IS}^+} = \sum_{a\in\mathcal{A}} w_i^a\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}c_i^a \tag{293}$$

$$\hat{V}^{\text{IS}^+*} = \sum_{a\in A}\pi_e(a|s_i)c_i^a \tag{294}$$

$$\hat{V}^{\text{DM-IS}^+} = \hat{R}(s_i,\pi_e)+\sum_{a\in A}\pi_e(a|s_i)(c_i^a-\hat{R}(s_i,a)) \tag{295}$$

$$\hat{V}^{\text{DM}^+\text{-IS}^+*} = \hat{R}^+(s_i,\pi_e)+\sum_{a\in A}\pi_e(a|s_i)(c_i^a-\hat{R}^+(s_i,a)) \tag{296}$$

$\hat{V}^{\text{IS}^+*}$ is $\hat{V}^{\text{IS}^+}$ under equal weights. The following is a derivation of how we can reach $\hat{V}^{\text{IS}^+*}$ under equal weights using the definition of the augmented behavior policy:

$$\hat{V}^{\text{IS}^+} = \sum_{a\in\mathcal{A}} w_i^a\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}c_i^a \tag{297}$$

$$= \sum_{a\in\mathcal{A}}\bar{W}(a|s_i,a_i)\frac{\pi_e(a|s_i)}{\pi_{b^+}(a|s_i)}c_i^a \tag{298}$$

$$= \sum_{a\in\mathcal{A}} \pi_e(a|s_i)c_i^a \tag{299}$$

$$= \hat{V}^{C*-IS} \tag{300}$$

2. Now, we write the definition for $\hat{R}(s_i, \pi_e)$ and $\hat{R}^+(s_i, \pi_e)$.

$$\hat{R}(s_i, \pi_e) = \sum_{a\in A} \pi_e(a|s_i)\hat{R}(s_i, a) \tag{301}$$

$$\hat{R}^+(s_i, \pi_e) = \sum_{a\in A} \pi_e(a|s_i)\hat{R}^+(s_i, a) \tag{302}$$

3. Let us start by re-writing the definition of $\hat{V}^{\text{DM-IS}^+}$ using the definition of $\hat{R}(s_i, \pi_e)$:

$$\hat{V}^{\text{DM-IS}^+} = \sum_{a\in A} \pi_e(a|s_i)\hat{R}(s_i, a) + \sum_{a\in A} \pi_e(a|s_i)(c_i^a - \hat{R}(s_i, a)) \tag{303}$$

$$= \sum_{a\in A} \pi_e(a|s_i)\hat{R}(s_i, a) + \sum_{a\in A} \pi_e(a|s_i)c_i^a - \sum_{a\in A} \pi_e(a|s_i)\hat{R}(s_i, a) \tag{304}$$

$$\tag{305}$$

4. Note that the second term is the definition of *AugIS* for a single sample. Also note that subtracting term 3 from term 1 is equal to 0 by definition.

This shows that IS$^+$ is equivalent to DM-IS$^+$. Note that we can replace the estimate of the reward function with $\hat{R}^+$ and produce the same derivation for DM$^+$-IS$^+$. Thus, we have shown that the three estimators IS$^+$, DM-IS$^+$ DM$^+$-IS$^+$ are equivalent when the weights are equal and all three methods have access to the same counterfactual annotations.

Corollary S21 (Equivalence of Variance of IS$^+$)**.**

Now, we want to verify that our variance decomposition for $\hat{V}^{\text{IS}^+}, \hat{V}^{\text{DM-IS}^+*}, \hat{V}^{\text{DM}^+\text{-IS}^+*}$ are equivalent to that of IS$^+*$. Recall that we are in the setting where the weights are equal. Here, we also assume that $\hat{R}, \hat{R}^+$ are constant and have been learned already.

1. First we note that the variance decomposition for DM-IS$^+$ and DM$^+$-IS$^+$ each have 7 terms. The last 4 terms reason about the covariance of the weights. Because our weights are constant and equal, the last four terms become 0.

2. Next, we note that the first term is identical in all of the variance decompositions. Namely, this is $\mathbb{V}_{s_i\sim d_0}[V^{\pi_e}(s_i)]$.

3. Now, we note that if $\hat{R}, \hat{R}^+$ are constants, the second term in the variance of DM-IS$^+$ and DM$^+$-IS$^+$ becomes 0.

4. All we need to do is establish that that the third term is equivalent to the third term for the variance decomposition. This is true by definition.

Thus, all three variance decompositions are the same under this setting.