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 *barbare@stanford.edu*
The Gladstone Institutes, Stanford University

Abstract

When applying contextual bandit algorithms in high-stakes settings (e.g., medical treatment), practitioners rely on off-policy evaluation (OPE) methods that use historical data to evaluate the behavior of novel policies prior to deployment. Unfortunately, OPE techniques are inherently limited by the breadth of the available data, which may not reflect distribution shifts resulting from the application of a new policy. Recent work attempts to address this challenge by leveraging domain experts to increase dataset coverage by annotating counterfactual samples. However, such annotations are not guaranteed to be free of errors, and incorporating imperfect annotations can lead to worse policy value estimates than not using the annotations at all. To make use of imperfect annotations, we propose a family of OPE estimators based on the doubly robust (DR) principle, which combines importance sampling (IS) with a reward model (direct method, DM) for better statistical guarantees. We introduce three opportunities within the DR estimation framework to incorporate counterfactual annotations. Under mild assumptions, we prove that using annotations within just the DM component yields the most desirable results, providing an unbiased estimator even under noisy annotations. We validate our approaches in several settings, including a real-world medical domain, observing that the theoretical advantages of using annotations within just the DM component hold in practice under realistic conditions. By addressing the challenges posed by imperfect annotations, this work broadens the applicability of OPE methods and facilitates safer and more effective deployment of decision-making systems.

1 Introduction

Contextual bandit methods have been successfully applied to learn optimal decision-making policies across several domains, including healthcare (Yao et al., 2021), recommendation systems (Li et al., 2010), and education (Lan & Baraniuk, 2016). In high-stakes decision-making scenarios, such as designing patient treatment policies in clinical settings, it is critical for practitioners to assess the performance of a new policy prior to deployment. To do so, standard practice consists of applying off-policy evaluation (OPE) methods (Sutton & Barto (2018), Chapter 5), which estimate the value of a new (target) policy using a behavior dataset

collected from a different policy. By facilitating policy evaluations without risky real-world experiments, OPE methods represent a crucial tool for safe policy deployment.

However, OPE is inherently limited by the quality and coverage of the behavior dataset. For instance, the current treatment policy in a hospital may have never recommended a recently developed drug, so no OPE method can reliably evaluate a policy that recommends this drug as a treatment. To address this issue, Tang & Wiens (2023) proposed an importance sampling (IS)-based OPE estimator called C-IS (referred to in this work as IS^+), in which experts provide annotations (i.e., predicted rewards) for counterfactual actions of samples observed in the behavior dataset. However, their approach relied on the **strong assumption that annotations are free of errors**. Realistically, even expert-generated annotations are prone to imperfections. Determining the optimal way to incorporate potentially imperfect counterfactual annotations into an OPE estimator remains an open challenge.

To address this challenge, we propose a family of OPE estimators based on the doubly robust (DR) principle (Cassel et al., 1976). Compared to IS estimators, DR estimators offer provable reductions in variance while remaining unbiased. It is not immediately obvious how to use the additional data coverage provided by potentially imperfect counterfactual annotations while retaining the desirable properties of DR estimators. In this work, we introduce three ways of modifying DR estimators to include counterfactual annotations, each of which impacts the estimator performance in a different way. Through a rigorous analysis of the bias-variance trade-off of each approach, **in the face of imperfect annotations**, we identify one estimator that successfully leverages information from counterfactual annotations to improve coverage without compounding error in those annotations. In contrast, the other two estimators compound error proportionally to the annotation error, resulting in worse policy estimates than ignoring the annotations altogether.

In summary, our contributions are the following:

- **We propose a family of OPE estimators** inspired by the DR principle that incorporate counterfactual annotations while accounting for potential errors in the annotations. We perform a thorough theoretical analysis of our proposed estimators, finding that how annotations are incorporated into the estimator has a substantial impact on the estimator's performance (Section 3).
- **We evaluate our estimators on three synthetic contextual bandit environments and a real medical dataset.** We use the synthetic settings to empirically verify our theoretical insights, and use the medical domain to demonstrate the potential utility of our proposed approaches in high-stakes problems (Section 4).
- **We provide practical considerations for choosing the best OPE estimator** in the presence of imperfect counterfactual annotations, which, to our knowledge, is currently missing from the OPE literature. This systematic guide further facilitates the deployment of contextual bandit policies in high-stakes settings (Section 5).

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 initial context distribution. Given a behavior dataset $D = \{(s_i, a_i, r_i)\}_{i=1}^N$ generated from a behavior policy π_b, we aim to evaluate a different target policy π_e by estimating its value $v(\pi_e) = \mathbb{E}_{s\sim d_0, a\sim\pi_e(\cdot|s), r\sim R(s,a)}[r]$.

2.1 Off-Policy Evaluation

We give an overview of three common types of OPE approaches in the context of contextual bandit. Importance sampling (IS), $\hat{V}^{\text{IS}} = \frac{1}{N}\sum_{i=1}^N \rho_{s_i}(a_i) r_i$, 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 (Horvitz & Thompson, 1952; Precup et al., 2000). Similar to prior work, we assume that the IPS ratio ρ is known (Farajtabar et al., 2018; Thomas & Brunskill, 2016). IS results in an unbiased estimate of the value of the target policy, $v(\pi_e)$, when π_e is well supported by the behavior dataset (Precup et al., 2000), i.e., has sufficient "coverage" (Assumption 1).

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

The variance of the IS estimator is (Tang & Wiens, 2023)

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

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.

Another approach to OPE is the direct method (DM) (Li et al., 2010; Beygelzimer & 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 estimate a reward model, $\hat{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, to predict the mean reward, and then uses $\hat{R}$ to directly compute the target policy value as $\hat{V}^{\text{DM}} = \sum_s d_0(s) \sum_a \pi_e(a|s)\hat{R}(s,a)$. $\hat{R}$ can vary in complexity, ranging from regression models to neural networks. If the reward model is fully realizable and there is full coverage in the behavior dataset, then DM has zero bias and favorable variance in its estimate of the target policy value. Typically, DM estimators have a lower variance than IS (Dudik et al., 2011) when the size of the behavior dataset is sufficiently large to learn an accurate reward model.

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 & Li, 2016). These methods are termed "doubly robust" because they maintain strong theoretical guarantees even when either the IPS ratio ρ, or the estimated reward function $\hat{R}$, is inaccurate. As such, the DR estimator is robust to two sources of error (the IPS ratio and the reward model). A related line of literature is in one-step estimators, which strongly resemble DR estimators (Kennedy, 2023). The asymptotic properties of these estimators is studied in semiparametric efficient statistics and efficient influence functions. The standard DR estimator is

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

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 Equation (1) as the *DM part* and the *IS part*, respectively. Under standard coverage assumptions (Assumption 1), the DR estimator produces an unbiased estimate of $v(\pi_e)$ under standard sample splitting assumptions (e.g., the reward model and OPE estimate are learned using independent splits of the data). DR methods also see a reduction in variance in comparison to IS-based methods; the variance can be written as

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

The reduction in variance relative to the IS estimator rests in the second term, in which ρ is scaled by $\bar{R}(s,a) - \hat{R}(s,a)$ instead of $\bar{R}(s,a)$, which is close to 0 if the estimated reward model $\hat{R}$ is accurate.

2.2 Counterfactual Annotations

In our work, we consider incorporating counterfactual annotations to increase data coverage. Suppose that we are given a behavior dataset of size N, $D = \{(s_i, a_i, r_i)\}_{i=1}^N$. Each factual sample (s_i, a_i) in the behavior dataset 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 some distribution $G : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, $g_i^{\tilde{a}} \sim G(s_i, \tilde{a})$. We assume that there are a total of M counterfactual annotations. In practice, we expect to collect a small subset of all possible counterfactual annotations because they may be expensive to obtain. We refer to the

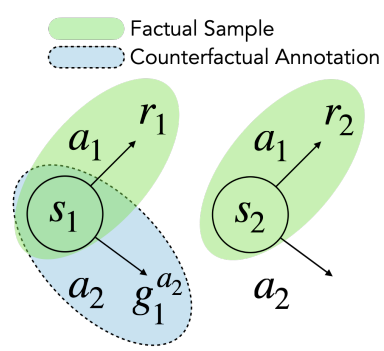


Figure 1: **Counterfactual-annotated dataset with two contexts and two actions.** There are two factual samples, (s_1, a_1, r_1) and (s_2, a_1, r_2). For the first (left) factual sample, we have a corresponding

dataset that combines factual samples and counterfactual annotations as the counterfactual-annotated dataset and denote it by D^+. A simple example of a counterfactual-annotated dataset with two contexts and two actions is visualized in Figure 1. In this example, we observe two factual samples and only one of them has a counterfactual annotation.

In Section 3.2, we discuss three scenarios for the function G (perfect, biased, or noisy annotations). For simplicity, 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

To incorporate counterfactual annotation, Tang & Wiens (2023) introduced IS$^+$, defined as

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

where $\{w_i^a\}$ is a set of user-defined weights for the i-th factual sample (s_i, a_i) and its associated counterfactual annotations. This method requires that $\sum_{a \in \mathcal{A}} w_i^a = 1$ to ensure that IS$^+$ is a convex combination of factual and counterfactual samples. We set $w_i^{\tilde{a}} = 0$ if the annotation of the counterfactual action $\tilde{a}$ is not available. $\rho_s^+(a) = \frac{\pi_e(a|s)}{\pi_b^+(a|s)}$ is the augmented IPS ratio, where the augmented behavior policy $\pi_b^+(a|s)$ is defined as,

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

and $\bar{W}(\tilde{a}|s, a) = \mathbb{E}[w^{\tilde{a}}]$ is the average weight of action $\tilde{a}$ for the factual context-action pair (s, a). The weights and augmented IPS ratios are critical in the construction of IS$^+$, as they ensure the context distribution in the counterfactual-annotated dataset D^+ remains identical to the context distribution in the original factual dataset D. This is because the weights associated with all actions taken in a particular context s_i must sum to 1.

3 Methods

When the behavior dataset has limited coverage (which tends to be true in practice), IS estimators are known to have high variance (Jiang & Li, 2016). In contrast, DM estimators have high bias when the reward model is misspecified. Thus, we explore how to introduce counterfactual annotations into a DR estimator, which retains beneficial theoretical properties even when the reward model is misspecified–a situation that frequently occurs in practice. While DR is well-understood in a setting with only factual samples, we aim to incorporate counterfactual annotations such that we can overcome the limitations of the coverage of the behavior dataset. The most naive approach is to directly use the counterfactual-annotated dataset D^+ in a standard DR estimator, viewing the counterfactual annotations as additional samples. However, as we discuss in Appendix C, this approach can produce arbitrarily biased estimates of $v(\pi_e)$ depending on the number of annotations used, because it alters the context distribution of the behavior dataset, regardless of annotation quality. As a result, we focus on developing new estimators that build on the DR principle. Below, we present three new estimators along with a rigorous theoretical analysis of their bias and variance properties in the presence of imperfect annotations.

3.1 Proposed DR Estimators with Counterfactual Annotations

The standard DR estimator, as shown in Equation (1), 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 independently leveraged in either of these components. Based on this insight, we propose **three**

new DR-inspired estimators leveraging counterfactual annotations. First, **DM$^+$-IS** (Equation (2)) uses the counterfactual-annotated dataset to estimate the reward model and combines it with standard IS. Next, **DM-IS$^+$** (Equation (3)) uses counterfactual annotations to augment the IS part (as in IS$^+$) and combines it with a standard DM estimator. Finally, **DM$^+$-IS$^+$** (Equation (4)) uses counterfactual annotations in both the DM and IS parts.

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

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

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

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

3.2 Theoretical Analyses under Imperfect Annotations

Now, we examine the performance of our proposed estimators in the presence of imperfect annotations, offering insights into how these limitations affect the estimators. This analysis also provides theoretical support for our guidance on selecting a robust OPE estimator, which we discuss further in Section 5. 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. Like prior work (Farajtabar et al., 2018; Thomas & 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. These theoretical results inform our hypotheses and help ensure that our empirical findings align with the expected behavior (Section 4.2).

The novelty of our theoretical results are two-fold. First, prior work on DR estimators provided expectation and variance derivations assuming a prespecified error term in $\hat{R}$ (e.g., (Dudik et al., 2011) assumed that $\hat{R}(s,a) = \bar{R}(s,a) + \epsilon(s,a)$). In contrast, our analysis accounts for the stochasticity in $\hat{R}$ arising from the dataset used to fit the reward model, since our proposed approaches explicitly modify what data is used to fit the reward model. We assume that the reward model is estimated from a separate dataset, which we refer to as $D_{\hat{R}}$ or $D_{\hat{R}^+}$, depending on if counterfactual annotations are incorporated. We assume that $D_{\hat{R}}$ and $D_{\hat{R}^+}$ are drawn from the same data distributions as D and D^+, respectively. Second, we derive the bias and variance of our proposed DR estimators under imperfect annotations, which is arguably more realistic in practice. We do not make assumptions about the reward model class being well speficified. We use the following three assumptions to quantify the quality of counterfactual annotations.

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

Assumption 3 and Assumption 4 are used to study the effect of biased and noisy (i.e., higher variance) annotations. The additional bias and noise are captured in the terms ϵ_G and Δ_G, respectively. Assumption 2 is contrary to Assumptions 3 and 4, and as such never invoked together. Similar to Tang & Wiens (2023), we use the following assumption on dataset support.

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

First, we show that, with perfect annotations (Assumption 2) and appropriate coverage assumptions (Assumption 1 or 5), all three proposed estimators are unbiased (Propositions 12, 14 and 16 in Appendix G). Additionally, when all counterfactual actions are annotated and $w^a = 1/|\mathcal{A}|$, DM-IS$^+$ and DM$^+$-IS$^+$ are both equivalent to IS$^+$(Corollaries 18 and 19 in Appendix H).

Now, we derive the bias of the proposed estimators when Assumption 2 (perfect annotations) is violated. We only rely on Assumption 3 (annotation bias) but not Assumption 4 (annotation variance).

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

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

$$\mathbb{E}[\hat{V}^{\text{DM-IS}^+}] = \mathbb{E}[\hat{V}^{\text{DM}^+\text{-IS}^+}] = v(\pi_e) + \mathbb{E}_{\substack{s\sim d_0 \\ a\sim\pi_e(s)}}\left[\left(1 - \frac{\bar{W}(a|s,a)\pi_b(a|s)}{\pi_b^+(a|s)}\right)\epsilon_G(s,a)\right]. \tag{5}$$

Proposition 1 establishes that, with biased annotations, DM$^+$-IS is an unbiased estimator of the target policy value $v(\pi_e)$. 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 Equation (5) is identical to the expectation derivation for IS$^+$ (Tang & Wiens, 2023).

For variance analysis, we focus on **DM$^+$-IS as it is the only estimator that remains unbiased with biased annotations**. The variance decompositions for DM-IS$^+$ and DM$^+$-IS$^+$ under imperfect annotations are nontrivial extensions, and we instead focus on their empirical evaluations (Figure 2).

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

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

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

Theorem 3 characterizes the variance of DM$^+$-IS under biased and noisy counterfactual annotations. The first two terms of the variance remain identical to those derived under perfect annotations (see Appendix Proposition 13). However, the third term (highlighted in purple), which depends on $\hat{R}^+$, can be dominant when noisy annotations introduce additional variance in the estimate of the reward model. The last term highlighted in green emerges from the possible estimation error of the reward model due to imperfect annotations.

We summarize our theorems in Appendix Table 3 with full proofs provided in Appendix 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 reward function estimate.

4 Experiments

We now empirically evaluate the performance of our proposed estimators, focusing on settings with imperfect annotations and a misspecified reward model, which reflect real-world scenarios. Prior work demonstrated that counterfactual annotations can improve the variance of an IS-based OPE estimator, suggesting that incorporating these annotations into both the IS and DM part of the DR estimator should lead to an even larger reduction in the estimator error (Tang & Wiens, 2023). However, our findings reveal a nuance: the improvement in estimator error depends critically on the quality of the annotations as well as how they are incorporated into the estimator.

Our experiments seek to answer the following questions: **1)** How do imperfect annotations empirically affect the proposed OPE methods? **2)** How do our proposed methods perform with compounding errors from imperfect annotations and a misspecified reward model?

4.1 Experimental Setup

To answer these questions, we investigated three synthetic settings with progressively increasing state and action space sizes, and one real-world medical domain. Key characteristics of these domains are summarized in Appendix Table 1 with further details in Appendix B.

4.1.1 Synthetic Domains

Two Context Bandit (Tang & Wiens, 2023): This setting is visualized in Figure 1 and has two contexts, and two actions. Without loss of generality, the reward of taking either action from the first context is sampled from a normal distribution and set to 0 for the second context.

Heartsteps (Mandyam et al., 2024): This realistic 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 3-dimensional vector that includes a treatment effect term and the step count of the previous day. There are two actions (either *send an intervention* or *do nothing*) at each decision time, 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 & Sontag, 2019): In this setting, we adapt the sepsis simulator in (Oberst & Sontag, 2019), which is originally built for a 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.

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 ϵ_G and Δ_G refer to the additional bias and variance that compromise the quality of the annotations.

In addition to imperfect annotations, we study the compounding error of misspecified reward models. A misspecified reward model cannot perfectly capture the environment's true reward function, regardless of the training data size. Such misspecification is common in practice. For instance, in a clinical setting, the true reward model guiding a clinician's treatment decisions is often unknown and approximated by a simpler model (e.g., maintaining the patient's vitals within a safe range). In our experiments, we create misspecified reward models across all three synthetic settings by either partially observing the state or altering the state representation (Table 1 and Appendix B).

For these three synthetic domains, we consider various combinations of stochastic behavior and target policies (details in Appendix B). Specifically, the behavior policies vary in their coverage of the action space. We present the averaged results across these combinations. We calculate the value of the target policy using Monte Carlo estimates, and report the root mean squared error (RMSE) of estimated policy values.

4.1.2 Real-World Clinical Data

MIMIC-IV, Potassium Administration (Johnson et al., 2020; Goldberger et al., 2000): MIMIC-IV contains electronic health records for over 65,000 admitted patients. In this domain, we study a subset of patients from MIMIC-IV that received potassium repletion through an intravenous line. 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 is the clinician's treatment policy for the "non-renal" patients and the target policy is the clinician's policy for "renal" patients. In Appendix Figure 14, we see that patients with renal disease are given lower dosages to account for their impaired kidney function (Shrimanker & Bhattarai, 2025). Our goal is to estimate the value of the target policy using data from the behavior policy.

Domain Setup: The patient context is a 20-dimensional vector containing information about vitals, administered medications, and static covariates; the actions are five possible dosages of potassium; and 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. Distinct from the synthetic settings, π_b and π_e are not given and are instead estimated using behavior cloning. We use linear regression to fit our estimated reward model. We measure estimators' performance using RMSE.

Counterfactual Annotations: We randomly selected a subset of state-action pairs in the behavior data split and generated counterfactual annotations for those samples. The annotations are produced using OpenAI

"o1" (OpenAI et al., 2024), which is prompted to predict a patient's blood potassium level after administering a dosage that is different from what that patient actually received in the behavior dataset. This procedure mimics a setting where counterfactual annotations may be imperfect. Further details regarding the dataset and annotation construction are in Appendix B.

4.1.3 Baselines

We compare our proposed estimators to standard OPE estimators that do not use counterfactual annotations (IS, DM, and DR) and IS$^+$, which uses counterfactual annotations. We also compare to a direct method estimator that estimates the reward model using the counterfactual-annotated dataset, defined as $\hat{V}^{\text{DM}^+} = \sum_s d_0(s) \sum_a \pi_e(a|s) \hat{R}^+(s,a)$.

4.2 Results

4.2.1 Imperfect annotations and well-specified reward models

Biased annotations affect the RMSE of OPE estimators more than higher variance annotations. First, we examine the impact of biased and noisy annotations under a well-specified reward model. Focusing on the two context bandit setting, we demonstrate that biased annotations have a greater effect on the RMSE of the proposed estimators than noisy annotations (Figure 2). Error metrics for this setting are provided in Appendix A.1. While Figure 2 reports only the proposed OPE methods, this trend holds across all OPE methods that use counterfactual annotations, across all synthetic environments. The effect is most pronounced in methods that incorporate annotations in the IS component (IS$^+$, DM-IS$^+$, DM$^+$-IS$^+$).

This observation aligns with our theoretical results (Proposition 1, Theorem 2), which suggest that estimators using IS$^+$ are biased with biased annotations. The RMSE of DM$^+$-IS remains far more stable across biased and noisy annotations, and is in general more invariant to annotation quality than baselines. From Theorem 3, we know that noisier annotations increase the RMSE of DM$^+$-IS; empirically, we find that this increase is not substantial.

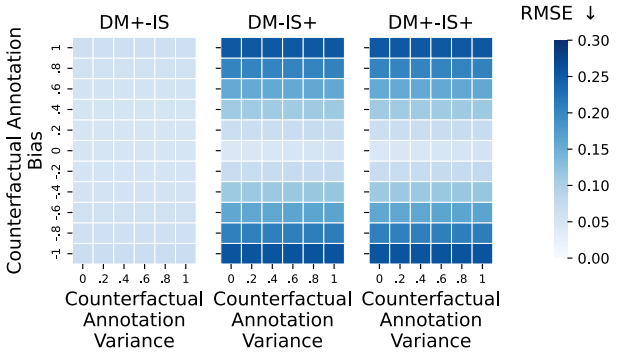


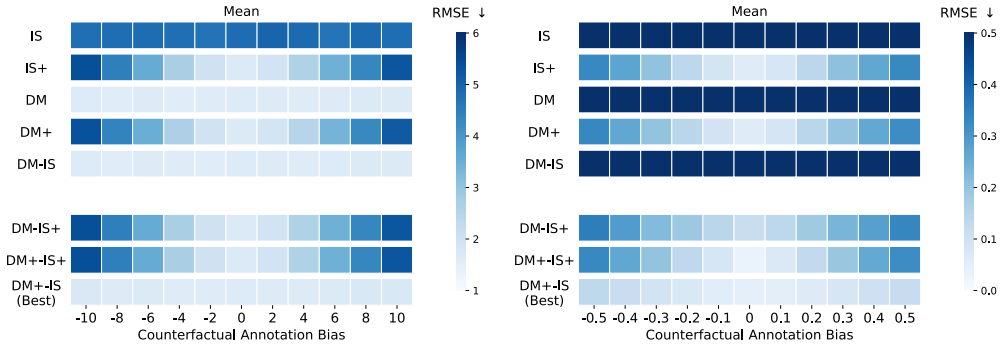
Figure 2: **Heatmaps of mean RMSE with a well-specified reward model on Two-Context Bandit (lower RMSE is represented lighter):** The bias of the counterfactual annotations has a larger impact on RMSE than the variance. The x, y-axis represents the variance (Δ_G) and the bias (ϵ_G) of the annotations, respectively. The RMSE hardly varies across the x-axis, but increases proportionally to the magnitude of the 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.

4.2.2 Imperfect annotations and misspecified reward models

In many realistic settings, such as those involving clinical data, we are likely to have both a misspecified reward model and imperfect annotations. **Our results demonstrate that, across all synthetic datasets, DM$^+$-IS is most robust to these two sources of error.** Intuition suggests that, under a misspecified reward model, DM and DM$^+$ will suffer, since a misspecified reward violates the accurate model assumption (Dudik et al., 2011). However, a misspecified reward model should not substantially affect any of the approaches that uses the DR principle, because these estimators can rely on the their IS component to still produce favorable results.

In Section 4.2.1, we noted that the annotation variance does not highly impact the RMSE of the proposed estimators. Thus, we focus on the effect of biased annotations in this set of results. We report mean RMSE across all three synthetic environments with varying degrees of annotation bias. Error metrics are available in Appendix A.2. Our results show that DM$^+$-IS is most resilient to the compounding errors from both

imperfect annotations and a misspecified reward model (Figure 3), showing the lowest RMSE across all magnitudes of annotation bias. Across all synthetic domains, we see that DM$^+$-IS consistently has either the lowest RMSE, or performs comparably to the best performing method. We hypothesize that this is because DM$^+$-IS is the only proposed estimator that is unbiased in the presence of imperfect annotations, and does not suffer from a misspecified reward model due to its DR properties.



(b) Heartsteps. (c) Sepsis treatment.

Figure 3: **Heatmaps of mean RMSE with a misspecified reward model and imperfect annotations (lower RMSE is represented lighter):** The x-axis represents annotation bias, ϵ_G. Across all datasets (two context results reported in Appendix A), DM$^+$-IS performs either better than all baselines, or comparably to the best-performing baseline. Among all methods that use counterfactual annotations, DM$^+$-IS is most robust to biased annotations and a misspecified reward model. In comparison to baselines that do not use counterfactual annotations, DM$^+$-IS frequently produces a lower RMSE.

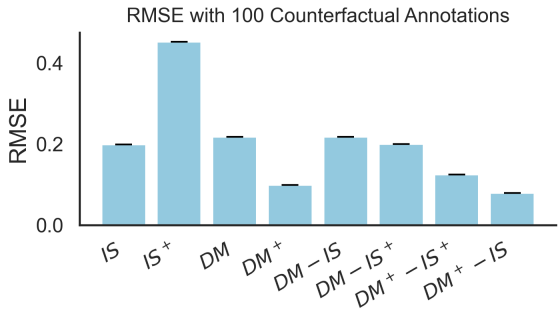
4.3 DM$^+$-IS outperforms baselines on MIMIC-IV data

Finally, we evaluate our methods using offline data from MIMIC-IV. With 100 counterfactual annotations, we find that DM$^+$-IS outperforms all baselines (Figure 4a). Notably, IS$^+$ exhibits the highest error, suggesting that the counterfactual annotations may be imperfect. The relatively small difference in performance between DM$^+$ and DM$^+$-IS implies that the estimated reward model is reasonably accurate. We also examine how the performance of key OPE estimators varies as the number of counterfactual annotations increases (Figure 4b). While the RMSE of DM-IS$^+$ and DM$^+$-IS$^+$ remains high with additional annotations—indicating that the incorporation of imperfect annotations introduces bias—the RMSE of DM$^+$-IS initially decreases and then plateaus. These trends are consistent with our observations in the synthetic experiments (Figures 2 and 3).

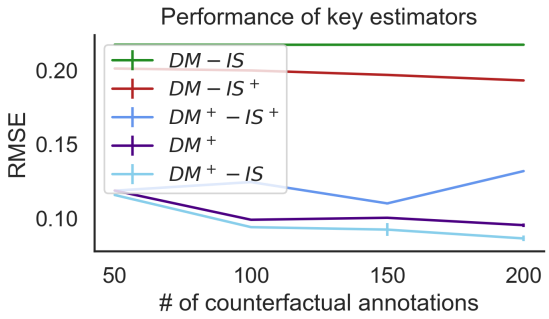
5 Selecting an OPE Estimator

As discussed in Section 4.2, we find the choice of OPE estimator depends most on (1) whether the reward model is misspecified, and (2) the bias of the counterfactual annotations. In the case that the reward model and annotation quality are known, our recommendations are summarized in Figure 5. The empirical results reported in the main text focus on the misspecified reward model and imperfect annotations case; we report results supporting the other three settings in Appendix A.1 and Appendix A.2. However, in the vast majority of real-world settings, the reward model and annotation quality are unknown a priori. In these settings, we recommend using DM$^+$-IS, which is most robust to the compounding errors from imperfect annotations and a misspecified reward model (Figure 3). Particularly, any further use of imperfect annotations (such as in the IS part), can lead to larger compounding errors.

To further emphasize the utility of DM$^+$-IS, we explore the consequences of choosing DM$^+$-IS when both the annotation and reward model quality is unknown in the sepsis treatment environment (Appendix Figure 9). Our results indicate that choosing DM$^+$-IS regardless of annotation or reward model quality will provide OPE estimates that are within a small margin of the best possible OPE method. The best performing OPE method is either DM or DM$^+$ (according to Figure 5), both with a well-specified reward model. Δ, the difference between the DM$^+$-IS estimate and the DM or DM$^+$ estimate (depending on annotation quality) is



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



(b) As the number of counterfactual annotations increases, the performance of DM$^+$-IS initially improves and then stays consistent. Error bars represent 95% confidence intervals.

Figure 4: **DM$^+$-IS performs best on MIMIC-IV**, a setting where annotations are likely imperfect.

small relative to the range of possible reward in the environment. That is, DM$^+$-IS produces estimates of $v(\pi_e)$ that are close to those of the best performing OPE method. This result suggests that, in a setting where it is difficult to assess reward model or annotation quality, DM$^+$-IS is the natural choice.

6 Conclusion

In this work, we address the open problem of incorporating imperfect counterfactual annotations into an OPE estimator and present a practical guide for their integration. We systematically explore various design options for incorporating annotations into a DR-based OPE estimator, and we find that imperfect counterfactual annotations are most beneficial when incorporated into the DM part of a DR estimator. Through comprehensive theoretical analyses and empirical evaluations, we find that selecting the best OPE method hinges 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 (i.e., a misspecified reward model and imperfect annotations), our DM$^+$-IS estimator is most robust.

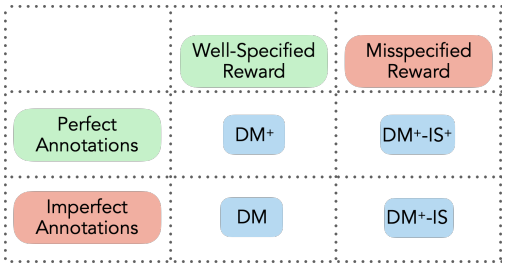


Figure 5: **Lookup table capturing the practical considerations when choosing an OPE estimator:** The most critical factors include (1) whether the reward model is well-specified and (2) the quality of the annotations. If these factors are known a priori, the best OPE estimator can be easily identified.

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 to improve OPE performance. Furthermore, future work can consider identifying practical ways to preprocess a set of counterfactual annotations to mitigate the effect of bias and noise. Overall, our approach relaxes restrictive assumptions about annotation quality, enabling more practical use of bandit algorithms in high-stakes applications.

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, pp. 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.

Claes M. Cassel, Carl-Erik Särndal, and Jan H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620, 1976. URL https://api.semanticscholar.org/CorpusID:120645424.

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.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review, 2023. URL https://arxiv.org/abs/2203.06469.

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.

Andrew S. Lan and Richard Baraniuk. A contextual bandits framework for personalized learning action selection. In *Educational Data Mining*, 2016. URL https://api.semanticscholar.org/CorpusID:15394680.

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*, pp. 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*, pp. 209–259, 2021.