Concept-Based Off-Policy Evaluation

Anonymous authors
Paper under double-blind review

Keywords: Off Policy Evalutaion, Interpretability, Concept Bottleneck Models, Reliable OPE

Summary

Evaluating off-policy decisions using batch data is challenging because of limited sample sizes which lead to high variance. Identifying and addressing the sources of this variance is crucial to improve off-policy evaluation in practice. Recent research on Concept Bottleneck Models (CBMs) shows that using human-explainable concepts can improve predictions and provide additional context for understanding decisions. In this paper, we propose incorporating an analogous notion of concepts into OPE to provide additional context that may help us identify specific areas where variance is high. We introduce a family of new concept-based OPE estimators and show that these estimators have two key properties when the concepts are known in advance: they remain unbiased whilst reducing variance of overall estimates. Since real-world applications often lack predefined concepts, we further develop an end-to-end algorithm to learn interpretable, concise, and diverse concepts optimized for variance reduction in OPE. Our experiments on synthetic and real-world datasets show that both known and learnt concept-based estimators significantly improve OPE performance. Crucially, our concept-based estimators offer two advantages over existing OPE methods. First, they are easily interpretable. Second, they allow us to isolate specific concepts contributing to variance. Upon performing targeted interventions on these concepts, we can further enhance the quality of OPE estimators.

Contribution(s)

1. We introduce a new family of IS estimators based on interpretable concepts. [Section 4]
   Context: Previous works perform IS in the state representations, we explicitly define what is a concept representation and tie the original definition of IS under concepts.
2. We derive theoretical conditions ensuring lower variance compared to existing IS estimators. [Section 5]
   Context: We compare the variance of the Concept-OPE estimators with traditional IS/PDIS and MIS estimators and devise conditions under which the variance is reduced.
3. We propose an end-to-end algorithm for optimizing parameterized concepts when concepts are unknown, using OPE characteristics like variance. [Section 6]
   Context: Under real-world scenarios, the concepts are typically unknown or hard to define, which adds to the complexity of performing OPE. In this section, we propose a novel algorithm which learns concepts that satisfy the desidarata: Explainability, Conciseness, Diversity while optimizing for variance.
4. We show, through synthetic and real experiments, that our estimators for both known and unknown concepts outperform existing ones. [Sections 5,6]
   Context: None
5. We interpret the learned concepts to explain OPE characteristics and suggest intervention strategies to further improve OPE estimates. [Section 7]
   Context: Interventions have been typically studied in the context of improving the CBM performance in a supervised learning regime, we instead use interpretations to explain where a concept-OPE estimator has high variance and intervene to reduce variance.

Concept-Based Off-Policy Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Evaluating off-policy decisions using batch data is challenging because of limited sample sizes which lead to high variance. Identifying and addressing the sources of this variance is crucial to improve off-policy evaluation in practice. Recent research on Concept Bottleneck Models (CBMs) shows that using human-explainable concepts can improve predictions and provide additional context for understanding decisions. In this paper, we propose incorporating an analogous notion of concepts into OPE to provide additional context that may help us identify specific areas where variance is high. We introduce a family of new concept-based OPE estimators and show that these estimators have two key properties when the concepts are known in advance: they remain unbiased whilst reducing variance of overall estimates. Since real-world applications often lack predefined concepts, we further develop an end-to-end algorithm to learn interpretable, concise, and diverse concepts optimized for variance reduction in OPE. Our experiments on synthetic and real-world datasets show that both known and learnt concept-based estimators significantly improve OPE performance. Crucially, our concept-based estimators offer two advantages over existing OPE methods. First, they are easily interpretable. Second, they allow us to isolate specific concepts contributing to variance. Upon performing targeted interventions on these concepts, we can further enhance the quality of OPE estimators.

1 Introduction

In domains like healthcare, education, and public policy, where interacting with the environment can be risky, prohibitively expensive, or unethical (Sutton & Barto, 2018; Murphy et al., 2001; Mandel et al., 2014), estimating the value of a policy from batch data before deployment is essential for the practical application of RL. OPE aims to estimate the effectiveness of a specific policy, known as the evaluation or target policy, using offline data collected beforehand from a different policy, known as the behavior policy (e.g., Komorowski et al. (2018a); Precup et al. (2000); Thomas & Brunskill (2016); Jiang & Li (2016)).

Importance sampling (IS) methods are a popular class of methods for OPE which adjust for distributional mismatches between behavior and target policies by reweighting historical data, yielding generally unbiased and consistent estimates (Precup et al., 2000). Despite their desirable properties (Thomas & Brunskill, 2016; Jiang & Li, 2016; Farajtabar et al., 2018), IS methods often face high variance, especially with limited overlap between behavioral samples and evaluation targets or in data-scarce conditions. Evaluation policies may outperform behavior policies for specific individuals or subgroups (Keramati et al., 2021b), making it misleading to rely solely on aggregate policy value estimates. In practice however, these groups are often unknown, prompting the need for methods to learn interpretable characterizations of the circumstances where the evaluation policy benefits certain individuals over others.

In this paper, we propose performing OPE using interpretable concepts (Koh et al., 2020; Madeira et al., 2023) instead of relying solely on state and action information. We demonstrate that this approach offers significant practical benefits for evaluation. These concepts can capture critical aspects in historical data, such as key transitions in a patient's treatment or features affecting short-term outcomes that serve as proxies for long-term results. By learning interpretable concepts from data, we introduce a new family of concept-based IS estimators that provide more accurate value estimates and stronger statistical guarantees. Additionally, these estimators allow us to identify which concepts contribute most to variance in evaluation. When the evaluation is unreliable, we can modify, intervene on, or remove these high-variance concepts to assess how the resulting evaluation improves (Marcinkevičs et al., 2024; Madeira et al., 2023).

A physician treating two patients infected with the same virus, with similar disease dynamics focuses on overall trends rather than precise viral load values when administering treatments. That is, if a drug lowers one patient's viral load below a threshold, it may also help the other, whereas it may be ineffective for a patient with a different disease trajectory. This distinction between a *concept* — a generalizable trend, such as viral load reduction — and a *state* — specific measurements at individual time points e.g. viral load — is key. Learning concepts that capture these trends, rather than isolated values, can better guide treatment decisions and evaluation. This idea is illustrated in Figure 1.

Our work makes the following contributions: i) We introduce a new family of IS estimators based on interpretable concepts; ii) We derive theoretical conditions ensuring lower variance compared to existing IS estimators; iii) We propose an end-to-end algorithm for optimizing parameterized concepts when concepts are unknown, using OPE characteristics like variance; iv) We show, through synthetic and real experiments, that our estimators for both known and unknown concepts outperform existing ones; v) We interpret the learned concepts to explain OPE characteristics and suggest intervention strategies to further improve OPE estimates.

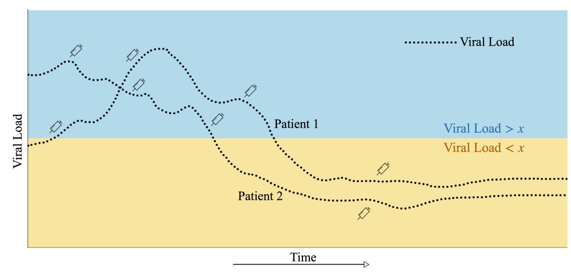


Figure 1: Simple example of a state vs concept. In this scenario, the state is the viral load in a patient's blood, whereas the concept is defined as the viral load being above or below a certain threshold x. The concept divides patients into two groups, in which different treatments are administered, indicated by the frequency of syringes. We do evaluation based on these two concepts, rather than the unique values of the viral loads.

2 Related Work

Off-Policy Evaluation. There is a long history of methods for performing OPE, broadly categorized into model-based or model-free (Sutton & Barto, 2018). Model-based methods, such as the Direct Method (DM), learn a model of the environment to simulate trajectories and estimate the policy value (Paduraru, 2013; Chow et al., 2015; Hanna et al., 2017; Fonteneau et al., 2013; Liu et al., 2018b). These methods often rely on strong assumptions about the parametric model for statistical guarantees. Model-free methods, like IS, correct sampling bias in off-policy data through reweighting to obtain unbiased estimates (e.g., Precup et al. (2000); Horvitz & Thompson (1952); Thomas & Brunskill (2016)). Doubly robust (DR) estimators (e.g., Jiang & Li (2016); Farajtabar et al. (2018)) combine model-based DM and model-free IS for OPE but may fail to reduce variance when both DM and IS have high variance. Various methods have been developed to refine estimation accuracy in IS, such as truncating importance weights and estimating weights from steady-state visitation distributions (Liu et al., 2018a; Xie et al., 2019; Doroudi et al., 2017; Bossens & Thomas, 2024).

Off-Policy Evaluation based on Subgroups. Keramati et al. (2021b) extend OPE to estimate treatment effects for subgroups and provide actionable insights on which subgroups may benefit from specific treatments, assuming subgroups are known or identified using regression trees. Unlike regression trees, which are limited in scalability, our approach learns interpretable concepts to characterize individuals, on the basis of which we introduce a new family of IS estimators. Similarly, Shen et al. (2021) propose reducing variance by omitting likelihood ratios for certain states. Our work complements this by summarizing relevant trajectory information using concepts, rather than explicitly omitting states irrelevant to the return. The advantage of using concepts as opposed to states is that we can easily interpret and intervene on these concepts unlike the states.

Marginalized Importance Sampling (MIS) estimators (Uehara et al., 2020; Liu et al., 2018a; Nachum et al., 2019; Zhang et al., 2020b;a) mitigate the high variance of traditional IS by reweighting data tuples using density ratios computed from the state visitation at each time step. These estimators enhance robustness by focusing on states with high visitation density ratios, thereby marginalizing out less visited states. However, MIS has its challenges: computing density ratios can introduce high variance, particularly in complex state spaces, and it obscures which aspects of the state space contribute directly to variance. Some studies, such as Katdare et al. (2023) and Fujimoto et al. (2023), improve MIS by decomposing density ratio estimation into components like large density ratio mismatch and transition probability mismatch. Our work differs from MIS by characterizing trends in a trajectory using interpretable concepts rather than solely relying on density ratios. This approach enables targeted interventions on specific concepts of interest, leading to more accurate return estimates and reduced variance in OPE. Unlike MIS, our method provides interpretability, which becomes increasingly important as problem complexity grows. Proposals for hybrid estimators, such as those in Pavse & Hanna (2022a), suggest using low-dimensional abstractions of state spaces with MIS to manage high-dimensional spaces more effectively. Our work differs in the sense that we use concepts instead of state abstractions which can be easily plugged into the existing IS OPE definitions as elaborated in Sections 4.2 and Supplementary Material C.

Concept Bottleneck Models. Concept Bottleneck Models (CBMs) (Koh et al., 2020) are a class of prediction models that first predict a set of human interpretable concepts, and subsequently use these concepts to predict a downstream label. Variations of these models include learning soft probabilistic concepts (Mahinpei et al., 2021), learning hierarchical concepts (Panousis et al., 2023) and learning concepts in a semi-supervised manner (Sawada & Nakamura, 2022). The key advantage of these models is they allow us to explicitly intervene on concepts and interpret what might happen to a downstream label if certain concepts were changed (Marcinkevičs et al., 2024). Unlike previous works, we leverage this idea to introduce a new class of estimators for OPE where we group trajectories based on interpretable concepts which are relevant for the downstream evaluation task.

3 Preliminaries

Concept Bottleneck Models Conventional CBMs learn a mapping from some input features $x \in \mathbb{R}^d$ to targets y via some interpretable concepts $c \in \mathbb{R}^k$ based on training data of the form $\{x_n, c_n, y_n\}_{n=1}^N$. This mapping is a composition of a mapping from inputs to concepts, $f : \mathbb{R}^d \to \mathbb{R}^k$, and a mapping from concepts to targets, $g : \mathbb{R}^k \to \mathbb{R}$. These may be trained via independent, sequential or joint training (Marcinkevičs et al., 2024). Variations which consider learning concepts in a greedy fashion or in a semi-supervised way include Wu et al. (2022); Havasi et al. (2022).

Markov Decision Processes (MDP). An MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, T)$. $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ are the transition and reward functions, $\gamma \in [0, 1]$ is the discount factor, $T \in \mathbb{Z}^+$ is the fixed time horizon. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a mapping from each state to a probability distribution over actions in $\mathcal{A}$. A T-step trajectory following policy π is denoted by $\tau = [(s_t, a_t, r_t, s_{t+1})]_{t=1}^T$ where $s_1 \sim d_1, a_t \sim \pi(s_t), r_t \sim r(s_t, a_t), s_{t+1} \sim p(s_t, a_t)$. The value function of policy π, denoted by $V_\pi : \mathcal{S} \to \mathbb{R}$, maps each state to the expected discounted sum of rewards starting from that state following policy π. That is, $V_\pi(s) = \mathbb{E}_\pi[\sum_{t=1}^T \gamma^{t-1} r_t | s_1 = s]$.

Off-Policy Evaluation. In OPE, we have a dataset of T-step trajectories $\mathcal{D} = \{\tau^{(n)}\}_{n=1}^N$ independently generated by a *behaviour policy* π_b. Our goal is to estimate the value function of another *evaluation policy*, π_e. We aim to use $\mathcal{D}$ to produce an estimator, $\hat{V}_{\pi_e}$, that has low mean squared error, $MSE(V_{\pi_e}, \hat{V}_{\pi_e}) = \mathbb{E}_{\mathcal{D} \sim P^\tau_{\pi_b}}[(V_{\pi_e} - \hat{V}_{\pi_e})^2]$. Here, $P^\tau_{\pi_b}$ denotes the distribution of trajectories τ, under π_b, from which $\mathcal{D}$ is sampled.

4 Concept-Based Off-Policy Evaluation

The goal of our work is to incorporate the notion of concepts into off-policy evaluation for improved interpretability and variance reduction. In this section, we formally define the notion of a concept, outline its desiderata and formally introduce a class of OPE estimators. In subsequent sections, we discuss how these estimators can be used when a) concepts are known from domain expertise (see Section 5), and b) concepts are unknown and must be learnt using a parametric representation (see Section 6).

4.1 Defining a Concept for OPE

Given a dataset $\mathcal{D} = \{\tau^{(n)}\}_{n=1}^N$ of N T-step trajectories, let $\phi : \mathcal{S} \times \mathcal{A} \times R \times \mathcal{S} \to \mathcal{C} \in \mathbb{R}^d$ denote a function that maps trajectory histories h_t to interpretable concepts in d-dimensional concept space $\mathcal{C}$. This mapping results in the concept vector $c_t = [c_t^1, c_t^2, ..., c_t^d]$ at time t, defined by $\phi(h_t)$. These concepts can capture various vital information in the history h_t, such as transition dynamics, short-term rewards, influential states, inter-dependencies in actions across time steps, etc. Without loss of generality, we consider concepts c_t as functions of current state s_t, however this could be extended to include historical information. This considers the scenario where concepts capture important information based on the criticality of the state. The concept function ϕ satisfies the following desiderata: explainability, conciseness, better trajectory coverage and diversity. A detailed description of these desiderata is provided in Supplementary Material A.

4.2 Concept-Based Estimators for OPE.

We propose a new class of concept-based OPE estimators, adapting existing non-concept-based methods to integrate concepts into OPE. Here, we present the results specifically for per-decision IS and standard IS estimators, as these serve as the foundation for several other estimators. We also demonstrate in Supplementary Material C how these methods can be extended to other estimators.

Definition 4.1 (Concept-Based Importance Sampling (CIS))**.**

$$\hat{V}_{\pi_e}^{CIS} = \frac{1}{N}\sum_{n=1}^{N} \rho_{0:T}^{(n)} \sum_{t=0}^{T} \gamma^t r_t^{(n)}; \quad \rho_{0:T}^{(n)} = \prod_{t'=0}^{T} \frac{\pi_e^c(a_{t'}^{(n)}|c_{t'}^{(n)})}{\pi_b^c(a_{t'}^{(n)}|c_{t'}^{(n)})}$$

Definition 4.2 (Concept-based Per-Decision Importance Sampling, (CPDIS))**.**

$$\hat{V}_{\pi_e}^{CPDIS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T} \gamma^t \rho_{0:t}^{(n)} r_t^{(n)}; \quad \rho_{0:t}^{(n)} = \prod_{t'=0}^{t} \frac{\pi_e^c(a_{t'}^{(n)}|c_{t'}^{(n)})}{\pi_b^c(a_{t'}^{(n)}|c_{t'}^{(n)})}$$

Concept-based variants of IS replace the traditional IS ratio with one that leverages the concept c_t at time t instead of the state s_t. This enables customized evaluations for various concept types, such as: 1) subgroups with similar short-term outcomes, 2) cases with comparable state-visitation densities, and 3) subjects with high-variance transitions. Details on selecting concept types are provided in Supplementary Material B.

5 Concept-based OPE under Known Concepts

We first consider the scenario where the concepts are known apriori using domain knowledge and human expertise. These concepts must satisfy the desiderata defined in Supplementary Material A.

5.1 Theoretical Analysis of Known Concepts

In this subsection, we discuss the theoretical guarantees of OPE under known concepts. We make the completeness assumption where every action of a particular state has a non-zero probability of appearing in the batch data. When this assumption is satisfied, we obtain unbiasedness and lower variance when compared with traditional estimators. Proofs follow in Supplementary Material D.

Assumption 5.1 (Completeness)**.** *$\forall s \in S, a \in A$, if $\pi_b(a|s), \pi_b^c(a|c) > 0$ then $\pi_e(a|s), \pi_e^c(a|c) > 0$.*

This assumption states that if an action appears in the batch data with some probability, it also has a chance of being evaluated with some probability.

Assumption 5.2. *$\forall s \in S, a \in A, |\pi_e^c(a|c) - \pi_e(a|s)| < \beta$ and $|\pi_b^c(a|c) - \pi_b(a|s)| < \beta$. This assumption states that for all states s, the policies conditioned on concepts are allowed to differ from the state policies by atmost β, which is defined by the practitioner.*

This assumption constrains concept-based policies to be close to state-based policies, with a maximum allowable difference of β, defined by the practitioner. This is to ensure that the evaluation policy π_e^c under concepts is reflective of the original policy π_e. If the practitioner is confident in the state representation, they may set a lower β to find concepts that align closely with state policies. Conversely, a higher β allows for more deviation between concept and state policies.

Theorem 5.3 (Bias)**.** *Under known-concepts, when assumption 5.1 holds, both $\hat{V}_{\pi_e}^{CIS}$ and $\hat{V}_{\pi_e}^{CPDIS}$ are unbiased estimators of the true value function V_{π_e}.*

Theorem 5.4 (Variance comparison with traditional OPE estimators)**.** *When $Cov(\rho_{0:t}^c r_t, \rho_{0:k}^c r_k) \leq Cov(\rho_{0:t} r_t, \rho_{0:k} r_k)$, the variance of known concept-based IS estimators is lower than traditional estimators, i.e. $\mathbb{V}_{\pi_b}[\hat{V}^{CIS}] \leq \mathbb{V}_{\pi_b}[\hat{V}^{IS}]$, $\mathbb{V}_{\pi_b}[\hat{V}^{CPDIS}] \leq \mathbb{V}_{\pi_b}[\hat{V}^{PDIS}]$.*

As noted in Jiang & Li (2016), the covariance assumption across timesteps is crucial yet challenging for OPE variance comparisons. Concepts being interpretable allows a user to design policies which align with this assumption, thereby reducing variance. We also compare concept-based estimators to the MIS estimator, the gold standard for minimizing variance via steady-state distribution ratios.

Theorem 5.5 (Variance comparison with MIS estimator)**.** *When $Cov(\rho_{0:t}^c r_t, \rho_{0:k}^c r_k) \leq Cov(\frac{d^{\pi_e}(s_t,a_t)}{d^{\pi_b}(s_t,a_t)} r_t, \frac{d^{\pi_e}(s_k,a_k)}{d^{\pi_b}(s_k,a_k)} r_k)$, the variance of known concept-based IS estimators is lower than the Variance of MIS estimator, i.e. $\mathbb{V}_{\pi_b}[\hat{V}^{CIS}] \leq \mathbb{V}_{\pi_b}[\hat{V}^{MIS}]$, $\mathbb{V}_{\pi_b}[\hat{V}^{CPDIS}] \leq \mathbb{V}_{\pi_b}[\hat{V}^{MIS}]$.*

Finally, we evaluate the CR-bounds on the MSE and quantify the tightness achieved using concepts.

Theorem 5.6 (Confidence bounds for Concept-based estimators)**.** *The Cramer-Rao bound on the Mean-Square Error of CIS and CPDIS estimator under known-concepts is tightened by a factor of K^{2T}, where K is the ratio of the cardinality of the concept-space and state-space.*

With limited samples, certain relevant states are underrepresented in the behavior policy, leading to a low $\pi_b(.|s)$ and corresponding high IS ratio. However, an alternative state s' in the data may closely resemble s (e.g., similar blood pressure values). Thus, even if s is missing, it can be characterized by s' through the concept function $\phi(s')$, as $\phi(s) \approx \phi(s')$. Consequently, while $\pi_b(.|s)$ is low, $\pi_b^c(.|s)$ is higher, as s is effectively represented via s'. This reduces the IS ratio in concept space and tightens the overall bounds.

5.2 Experimental Setup and Metrics

Environments: We consider a synthetic domain: WindyGridworld and the real world MIMIC-III dataset for acutely hypotensive ICU patients as our experiment domains for the rest of the paper.

WindyGridworld: The goal is to reach the top-right corner of the grid. The states are defined by the 2D co-ordinates, and actions are directions up, down, left, right. We (as human experts) define a concept $c_t = \phi(\text{distance to target}, \text{wind})$ as a function of the distance to the target and the wind acting on the agent at a given state. This concept can take 25 unique values, ranging from 0 to 24. For example: $c_t = 0$ when distance to target $\in [15, 19] \times [15, 19]$ and wind $= [0, 0]$. The first and second co-ordinates represent the horizontal and vertical features respectively. Detailed description of known concepts in Supplementary Material G.

MIMIC: The goal is to treat and manage hypotensive patients. The state space consists of the physiological quantities of the patient while actions correspond to quantities of IV-fluids and vasopressors. Concepts $c_t \in \mathbb{Z}^{15}$ are a function of 15 different vital signs (interpretable features) of a patient at a given timestep. The vital signs considered are: Creatinine, FiO_2, Lactate, Partial Pressure of Oxygen (PaO_2), Partial Pressure of CO_2, Urine Output, GCS score, and electrolytes such as Calcium, Chloride, Glucose, HCO_3, Magnesium, Potassium, Sodium, and SpO_2. Each vital sign is binned into 10 discrete levels, ranging from 0 (very low) to 9 (very high).

For example, a patient with the concept representation $[0, 2, 1, 1, 2, 0, 9, 5, 2, 0, 6, 2, 1, 5, 9]$ shows the following conditions: acute kidney injury-AKI (very low creatinine), severe hypoxemia (very low PaO_2), metabolic alkalosis (very high SpO_2), and critical electrolyte imbalances (low potassium and magnesium), along with severe hypoglycemia. The normal GCS score indicates preserved neurological function, but over-oxygenation and potential respiratory failure are likely. The combination of anuria, AKI, and hypoglycemia points strongly toward hypotension or shock as underlying causes.

Policy descriptions: In the case of WindyGridworld, we run a PPO Schulman et al. (2017) algorithm for 10k epochs and consider the evaluation policy π_e as the policy at epoch 10k, while the behavior policy π_b is taken as the policy at epoch 5k. For the MIMIC case, we generate the behavior policy π_b by running an Approximate Nearest Neighbors algorithm with 200 neighbors, using Manhattan distance as the distance metric. The evaluation policy π_e involves a more aggressive use of vasopressors (10% more) compared to the behavior policy. See Supplementary Material F for further details.

Metrics: In the case of the synthetic domain, we measure bias, variance, MSE, and the effective sample size (ESS) to assess the quality of our concept-based OPE estimates. The ESS is defined as $N \times \frac{\mathbb{V}_{\pi_e}[\hat{V}^{on-policy}_{\pi_e}]}{\mathbb{V}_{\pi_b}[\hat{V}_{\pi_e}]}$, where N is the number of trajectories in the off-policy data, and $\hat{V}^{on-policy}_{\pi_e}$ and $\hat{V}_{\pi_e}$ are the on-policy and OPE estimates of the value function, respectively. For MIMIC, where the true on-policy estimate is unknown due to the unknown transition dynamics and environment model, we only consider variance as the metric. Additionally, we compare the Inverse Propensity scores (IPS) under concepts and states to better underscore the reasons for variance reduction.

5.3 Results and Discussion

Good concept-based estimators demonstrate reduced variance, improved ESS, and lower MSE compared to traditional estimators, although they come with slightly higher bias. Figure 2 compares known-concept and traditional OPE estimators. We observe a consistent reduction in variance and an increase in ESS across all sample sizes for the concept-based estimators. Although our theoretical analysis suggests that known-concept estimators are unbiased in the asymptotic case, practical results indicate some bias due to finite sample size. While unbiased estimates are generally preferred, they can lead to higher errors when the behavior policy does not cover all states. This issue is especially pronounced in limited data settings, which are common in medical applications. Despite this bias-variance trade-off, the MSE for concept-based OPE estimators shows a 1-2 order of magnitude improvement over traditional estimators due to significant variance reduction. In the real-world MIMIC example, concept-based estimators exhibit a variance reduction of one order of magnitude compared to traditional OPE estimators. This shows that characterizing diverse states, such as varying grid world positions or patient vital signs, in terms of shared concepts based on common attributes, improves OPE characterization.

The frequency of higher IPS scores is reduced in good concepts compared to states. Figure 2c compares IPS scores in good concept and state estimators. We observe, the frequency of lower IPS scores is higher under concepts as opposed to states. This indicates the source of variance reduction in Concept-based OPE lies in the lowering of the IPS scores, which is also backed theoretically in Theorem 5.4 when the rewards r_t equal 1.

Imperfect concepts baseline: While known concepts display superior OPE performance, in real world scenarios, concepts are often poorly described. We thus, perform an experiment where the concepts known but poor and study the resulting OPE performance. For Windygridworld, we define concepts as functions solely of the horizontal distance to the target. This approach neglects critical information such as vertical distance to the target, wind effects, and region penalties. As a result, these concepts violate one of the primary desiderata: diversity. By capturing only one important

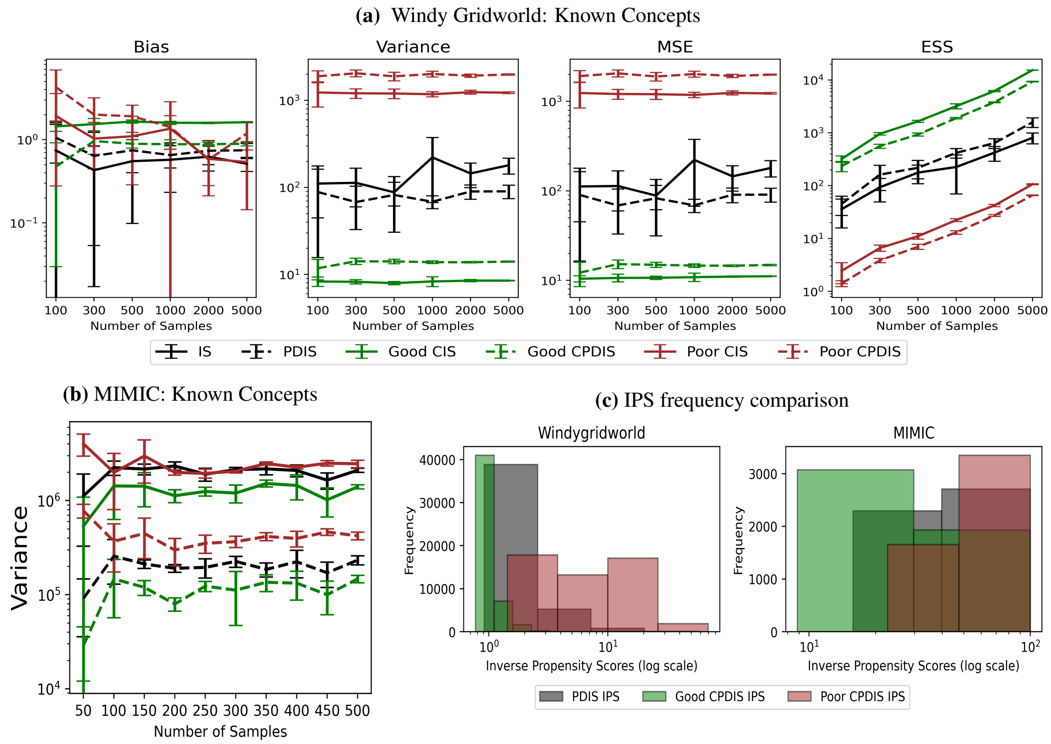


Figure 2: *Known Concepts.* (a) *Windy Gridworld:* Concept-based estimators with good concepts have lower variance and MSE, and higher ESS compared to traditional OPE estimators, with a higher bias. For poor concepts, we observe the OPE performance to be poor across all metrics compared to traditional estimators. (b) *MIMIC:* Good Concept-based estimators have lower variance compared to traditional OPE estimators, while poor concepts have higher variance. (c) Under good concepts, the frequency of high IPS scores is lower compared to traditional estimators, whereas for poor concepts, the frequency is higher across both domains. This indicates the source of variance reduction in good concept estimators lies in the lowered IPS scores.

concept dimension while disregarding others, these poor concepts fail to represent the full complexity of the environment.

Poor concepts exhibit inferior OPE characteristics across all metrics. From figures 2a and 2b, we observe that poor concepts exhibit higher bias, variance, and MSE, along with lower ESS, compared to traditional OPE estimators. Additionally, figure 2c shows an increased frequency of high IPS scores for poor concepts. This demonstrates that not all concept-based estimators improve performance; their quality is crucial and depends on the desiderata they satisfy. It also underscores the need for an algorithm that learns concepts with favorable OPE characteristics, particularly in complex domains or scenarios with imperfect experts. We explore this in the next section. Nonetheless, poor concepts still allow for interventions, as their impact on OPE metrics can be systematically analyzed and addressed.

6 Concept-based OPE under Unknown Concepts

While domain knowledge and predefined concepts can enhance OPE, in real-world situations concepts are typically unknown. In this section, we address cases where concepts are unknown and must be estimated. We learn a parametric representation of concepts via CBMs, which initially may not meet the required desiderata. This section introduces a methodology to optimize parameterized concepts to meet explicitly these desiderata, alongside improving OPE metrics like variance.

Learning concepts that characterize relevant trajectory information. Algorithm 1 outlines the training methodology. We split the batch of trajectories $\mathcal{D}$ into training trajectories $\mathcal{T}_{\text{train}}$ and evaluation trajectories $\mathcal{T}_{\text{OPE}}$, with the evaluation policy π_e, the behavior policy π_b, and an OPE estimator (e.g. CIS/CPDIS) known beforehand. We aim to learn our concepts using a CBM parameterized by θ. The

Algorithm 1 Unknown Concept-based Off Policy Evaluation

Require: Trajectories $\{\mathcal{T}_{\text{train}}, \mathcal{T}_{\text{OPE}}\}$, Policies $\{\pi_e, \pi_b\}$, OPE Estimator.
Ensure: CBM θ, concept policies $\tilde{\pi}^c$ $\{\theta_b, \theta_e\}$
Loss terms: $\{L_{\text{output}}, L_{\text{interpretability}}, L_{\text{diversity}}, L_{\text{OPE-metric}}, L_{\text{policy}}\} = 0$
1: **while** Not Converged **do**
2: **for** trajectory in $\mathcal{T}_{\text{train}}$ **do**
3: **for** (s, a, r, s', o) in trajectory **do** ▷ Choices for o: s' (Next state) / r (Next reward)
4: $c', o' \leftarrow \text{CBM}(s)$ ▷ CBM predicts concept c' and output label o'
5: $L_{\text{output}} \mathrel{+}= C_{\text{output}}(o, o')$ ▷ Eg: MSE/Cross-entropy between true next state and predicted next state
6: $L_{\text{interpretability}} \mathrel{+}= C_{\text{interpretability}}(c')$ ▷ Eg: L1-loss over weights
7: $L_{\text{diversity}} \mathrel{+}= C_{\text{diversity}}(c')$ ▷ Eg: Cosine distance between sub-concepts
8: $L_{\text{policy}} \mathrel{+}= C_{\text{policy}}(c')$ ▷ Eg: MSE/Cross-entropy between predicted logits and true logits in Assn 5.2
9: **end for**
10: **end for**
11: Returns $\leftarrow$ Estimator$(\mathcal{T}_{train}, \pi_e, \pi_b, \text{CBM})$ ▷ Eg: CIS/CPDIS
12: $\text{Loss}(\theta, \theta_b, \theta_e) = L_{\text{output}} + L_{\text{interpretability}} + L_{\text{diversity}} + C_{\text{OPE-metric}}(\text{Returns})$ ▷ Eg: Variance
13: Gradient Descent on $\{\theta, \theta_b, \theta_e\}$ using $\text{Loss}(\theta, \theta_b, \theta_e)$
14: **end while**
15: **Return** Concept OPE Returns $\leftarrow$ OPE Estimator$(\mathcal{T}_{\text{OPE}}, \pi_e, \pi_b, CBM)$

CBM maps states to outputs through an intermediary concept layer. In this work, the output o is the next state, indicating that the bottleneck concepts capture transition dynamics. Other possible outputs could include short-term rewards, long-term returns, or any user-defined information of interest present in the batch data. In addition to learning concepts, we also learn parameterized concept policies $\tilde{\pi}^c$ which maps concepts to actions parameterized by θ_b, θ_e for π_b, π_e respectively.

Optimizing concepts for variance reduction in OPE. For each transition tuple (s, a, r, s'), the CBM computes a concept vector c' and an output o'. Since the concepts are initially unknown, they do not inherently satisfy the concept desiderata and must be learned through constraints. Lines 5-7 impose soft constraints on the concepts to meet these desiderata using loss functions. The losses are updated based on output, interpretability, and diversity, with MSE used for C_{output}, L1 loss for $C_{\text{interpretability}}$, and cosine distance for $C_{\text{diversity}}$. In Line 8, we constrain the difference between the concept policies and the original policies to satisfy Assumption 5.2. For our experiments, we take maximum allowable difference $\beta = 0$, however a user can choose a different value to allow for more deviation in the concept policies $\tilde{\pi^c}$ and original policies π. In line 11, we evaluate the OPE estimator's returns based on the concepts at the current iteration with metrics like variance. The aggregate loss, Loss(θ), guides gradient descent on CBM parameters θ. Finally, the OPE estimator is applied to $\mathcal{T}_{\text{OPE}}$ using learned concepts, yielding concept-based OPE returns. Integrating multiple competing loss components makes this problem complex, and, to our knowledge, this is the first approach that incorporates the OPE metric directly into the loss function.

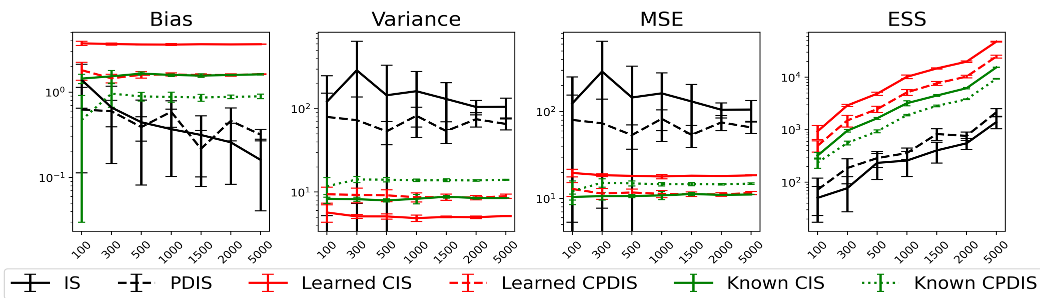
6.1 Theoretical Analysis of Unknown Concepts

The theoretical implications mainly differ in the bias, consequently MSE and their Confidence bounds on moving from known to unknown concepts, as analyzed below. Proofs are listed in Supplementary material E.

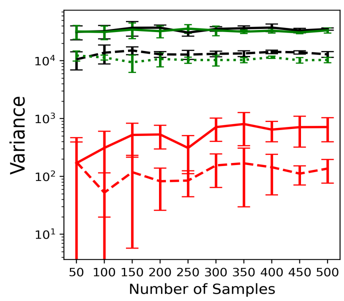
Theorem 6.1 (Bias)**.** *Under Assumptions 5.1, 5.2, the unknown concept-based estimators are biased.* The change of measure theorem from probability distributions π_b to π_b^c is not applicable on moving from known to unknown concepts, leading to bias. In the special case where $\pi_b^c(.|c_t) = \pi_b(.|s_t)$, the estimator is unbiased.

Theorem 6.2 (Variance comparison with traditional OPE estimators)**.** *Under Assumption 5.2, when* $Cov(\rho^c_{0:t} r_t, \rho^c_{0:k} r_k) \leq Cov(\rho_{0:t} r_t, \rho_{0:k} r_k)$*, the variance of concept-based IS estimators is lower than the traditional estimators, i.e.* $\mathbb{V}_{\pi_b}[\hat{V}^{CIS}] \leq \mathbb{V}_{\pi_b}[\hat{V}^{IS}]$, $\mathbb{V}_{\pi_b}[\hat{V}^{CPDIS}] \leq \mathbb{V}_{\pi_b}[\hat{V}^{PDIS}]$.

(a) Windy Gridworld: Learned Concepts



(b) MIMIC: Learned Concepts



(c) IPS frequency comparison

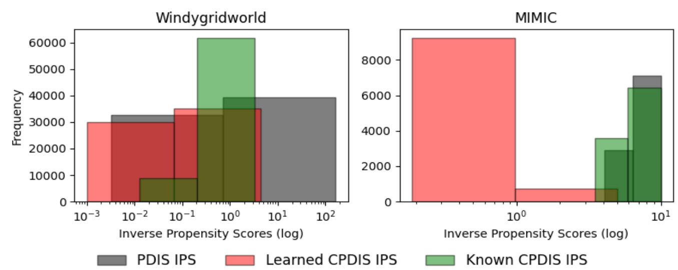


Figure 3: *Learned Concepts*. (a-b) *Windy Gridworld and MIMIC:* In both environments, we see a reduction in variance in learned concepts compared to both traditional and known concept estimators, at a cost of higher bias. (c) The frequency of high IPS scores is lower for learned concepts compared to traditional and known concept estimators. This indicates our proposed algorithm learns alternative concepts which further reduce variance.

Theorem 6.3 (Variance comparison with MIS estimator)**.** *Under Assumption 5.2, when* $Cov(\rho^c_{0:t} r_t, \rho^c_{0:k} r_k) \leq Cov(\frac{d^{\pi_e}(s_t,a_t)}{d^{\pi_b}(s_t,a_t)} r_t, \frac{d^{\pi_e}(s_k,a_k)}{d^{\pi_b}(s_k,a_k)} r_k)$*, like known concepts, the variance is lower than the Variance of MIS estimator, i.e.* $\mathbb{V}_{\pi_b}[\hat{V}^{CIS}] \leq \mathbb{V}_{\pi_b}[\hat{V}^{MIS}]$, $\mathbb{V}_{\pi_b}[\hat{V}^{CPDIS}] \leq \mathbb{V}_{\pi_b}[\hat{V}^{MIS}]$.

Similar to known concepts, when the covariance assumption is satisfied, even unknown concept-based estimators can provide lower variances than traditional and MIS estimators. In known concepts however, this assumption has to be satisfied by the practitioner, whereas in unknown concepts, this assumption can be used as a loss function in our methodology to implicitly reduce variance. (Line 12)

Theorem 6.4 (Confidence bounds for Concept-based estimators)**.** *The Cramer-Rao bound on the Mean-Square Error of CIS and CPDIS estimator loosen by* $\epsilon(|\mathbb{E}_{\pi_e^c}[\hat{V}_{\pi_e}]|^2)$*, under unknown concepts over known-concepts. Here,* $\mathbb{E}_{\pi_e^c}[\hat{V}_{\pi_e}]$ *is the on-policy estimate of concept-based IS (PDIS) estimator.*

The confidence bounds of unknown concepts mirror that of known-concepts, with the addition of the bias term whose maximum value is the true on-policy estimate of the estimator. This is typically unknown in real-world scenarios and requires additional domain knowledge to mitigate.

6.2 Experimental setup

Environments, Policy descriptions, Metrics: Same as those in known concepts section.

Concept representation: In both examples, we use a 4-dimensional concept $c_t \in \mathcal{R}^4$, where each sub-concept is a linear weighted function of human-interpretable features f, i.e., $c^i_t = w \cdot f(s_t)$, with w optimized as previously discussed. Detailed descriptions of the features and optimized concepts after CBM training are provided in Supplementary material H. For MIMIC, features f are normalized vital signs, as threshold information for discretization is unavailable. In brevity of space, we move the training and hyperparameter details to Supplementary material G.

6.3 Results and Discussion

Learned concepts using Algorithm 1 yield improvements across all metrics except bias compared to traditional OPE estimators. Significant improvements in variance, MSE, and ESS are observed for the Windy Gridworld and MIMIC datasets, with gains of 1-2 and 2-3 orders of magnitude, respectively. This improvement is due to our algorithm's ability to identify concepts that satisfy the desiderata, including achieving variance reduction as specified in line 12 of the algorithm. However, like known concepts, optimized concepts show a higher bias than traditional estimators. This is because, unlike variance, bias cannot be optimized in the loss function without the true on-policy estimate, which is typically unavailable in real-world settings. As a result, external information may be essential for further bias reduction.

Learned concepts yield improvements across all metrics besides bias over known concept estimators. From Fig 3a,3b, we observe that our methodology improves variance, MSE, and ESS by 1–2 orders of magnitude compared to known concepts. This suggests our algorithm can learn concepts that outperform human-defined ones in OPE metrics. Fig 3c further supports this, showing a lower frequency of high IPS scores for learned concepts than for known ones. This indicates our algorithm discovers novel concepts that satisfy concept desiderata in Section A while enhancing OPE characteristics, particularly variance. Such capability is valuable in domains with imperfect experts or complex real-world settings where perfect expertise is unattainable. However, these learned concepts introduce higher bias, as the training algorithm prioritizes variance reduction over bias minimization. This could be mitigated by regularizing variance during training.

Learned concepts are interpretable, show conciseness and diversity. We list the optimized concepts in Supplementary material H. These concepts exhibit sparse weights, enhancing their conciseness, with significant variation in weights across different dimensions of the concepts, reflecting diversity. This work focuses on linearly varying concepts, but more complex concepts, such as symbolic representations (Majumdar et al., 2023), could better model intricate environments.

7 Interventions on Concepts for Insights on Evaluation

Concepts provide interpretations, allowing practitioners to identify sources of variance—an advantage over traditional state abstractions like Pavse & Hanna (2022a). Concepts also clarify reasons behind OPE characteristics, such as high variance, enabling corrective interventions based on domain knowledge or human evaluation. We outline the details of performing interventions next.

7.1 Methodology

Given trajectory history h_t and concept c_t, we define c_t^{int} as the intervention (alternative) concept an expert proposes at time t. We define criteria $\kappa : (h_t, c_t) \rightarrow \{0, 1\}$ as a function constructed from domain expertise that takes in (h_t, c_t) as input and outputs a boolean value. This criteria function determines whether an intervention needs to be conducted over the current concept c_t or not. For e.g., if a practitioner has access to true on-policy values, he/she can estimate which concepts suffer from bias. If a concept doesn't suffer from bias, the criteria $\kappa(h_t, c_t) = 1$ is satisfied and the concept is not intervened upon, else $\kappa(h_t, c_t) = 0$ and the intervened concept c_t^{int} is used instead. The final concept $\tilde{c}_t$ is then defined as: $\tilde{c}_t = \kappa(h_t, c_t) \cdot c_t + (1 - \kappa(h_t, c_t)) \cdot c_t^{\text{int}}$. Under the absence of true on-policy values, the practitioner may chose to intervene using a different criteria instead.

We define criteria κ for our experiments as follows. In gridworld, we assume access to oracle concepts, listed in Supp. Material G. When the learned concept c_t matches the true concept, $\kappa(h_t, c_t) = 1$, otherwise 0. In MIMIC, the interventions are based on a patient's urine output at a specific timestep with $\kappa(h_t, c_t) = 1$ when urine output > 30 ml/hr, and 0 otherwise. Performing interventions based on urine output enables us to assess the role of kidney function in hypotension management. In this work, we consider 3 possible intervention strategies either based on states or domain knowledge.

Interventions that replace concepts with state representations and state-based policies. We intervene on the concept with the state and use policies dependent on state to perform OPE, i.e $c_t^{\text{int}} = s_t$, $\pi_e^c(a_t|\tilde{c}_t) = \pi_e(a_t|s_t)$, $\pi_b^c(a_t|\tilde{c}_t) = \pi_b(a_t|s_t)$. This can be thought of as a comparative measure a practitioner can look for between the concept and the state representations.

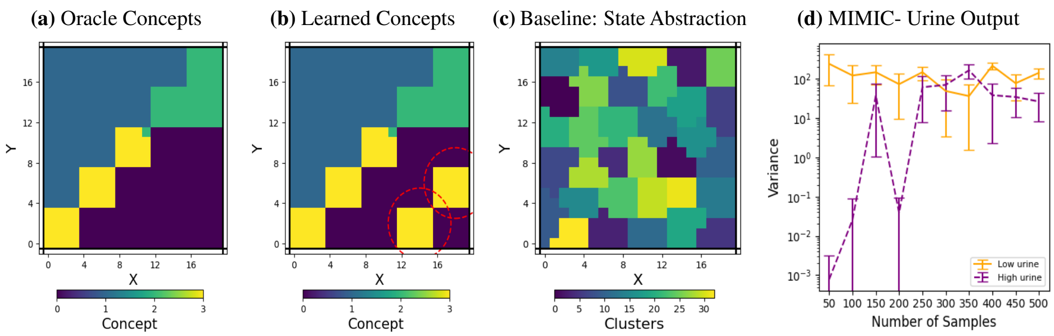


Figure 4: Interpretations of learned concepts. *Windy Gridworld:* Fig 4a and 4b compare true oracle concepts with learned concepts derived from the proposed methodology. We observe a deviation between learned and oracle concepts (circled in red), identifying potential interventions. We compare our learnt concepts with a state abstraction baseline in Fig 4c obtained using K-means clustering. We observe the clusters to significantly differ from both oracle and optimized concepts, underscoring the meaningfulness of learned concepts. *MIMIC:* From Fig 4d, we observe patients with low urine output exhibit greater variance in learned concepts compared to high-output patients, revealing potential intervention targets.

Interventions that replace concepts with state representations and maximum likelihood estimator (MLE) of state-based policies. We replace the erroneous concept with the corresponding state and use the MLE of the state conditioned policy to perform OPE, i.e $c_t^{\text{int}} = s_t$, $\pi_e^c(a_t|\tilde{c}_t) = MLE(\pi_e(a_t|s_t))$, $\pi_b^c(a_t|\tilde{c}_t) = MLE(\pi_b(a_t|s_t))$. This can be thought of as a comparative measure a practitioner can look for between the concept and states, while priortizing over the most confident action.

Interventions using a qualitative concept while retaining concept-based policies. In this approach, a human expert replaces the concept using external domain knowledge. This is similar to Tang & Wiens (2023), where the authors augment the dataset with counterfactual annotations to improve sample efficiency in regions where the coverage is low. However, while Tang & Wiens focus on quantitative counterfactual annotations in the state representation, we employ human interventions to qualitatively edit concepts. In case of gridworld, we consider the oracle concepts as our qualitative concept, while for MIMIC, we consider the learnt CPDIS estimator as qualitative concept while intervening on CIS estimator.

7.2 Results and Interpretations from Interventions on learned concepts

We interpret the optimized concepts in Fig.4. In the gridworld environment, we compare the ground-truth concepts with the optimized ones and observe two additional concepts predicted in the bottom-right region. This likely stems from overfitting to reduce variance in the OPE loss, suggesting a need for inspection and possible intervention. Additionally, we compare our clusters with state-abstraction baseline (clustering in the state-space), and observe the clusters to be widely different from the learnt concepts. In MIMIC, prior studies indicate that patients with urine output above 30 ml/hr are less susceptible to hypotension than those with lower output Kellum & Prowle (2018); Singer et al. (2016); Vincent & De Backer (2013). Using this, we analyze patient trajectories and find that lower urine output correlates with higher variance, while higher output corresponds to lower variance. This insight helps identify patients who may benefit from targeted interventions.

Interpretable concepts allow for targeted interventions that further enhance OPE estimates.

Using qualitative interventions, we observe a reduction in bias and, consequently, MSE in WindyGridworld. This occurs because replacing erroneous concepts with oracle concepts introduces previously missing on-policy information during the optimization of unknown concepts, while preserving the order of variance and ESS estimates. Similarly, in MIMIC, intervening on states with low urine output reduces variance by 1–2 orders of magnitude. This is further supported by the decreased frequency of high IPS scores after intervention, as shown in Figure 5c.

(a) Windy Gridworld: Interventions

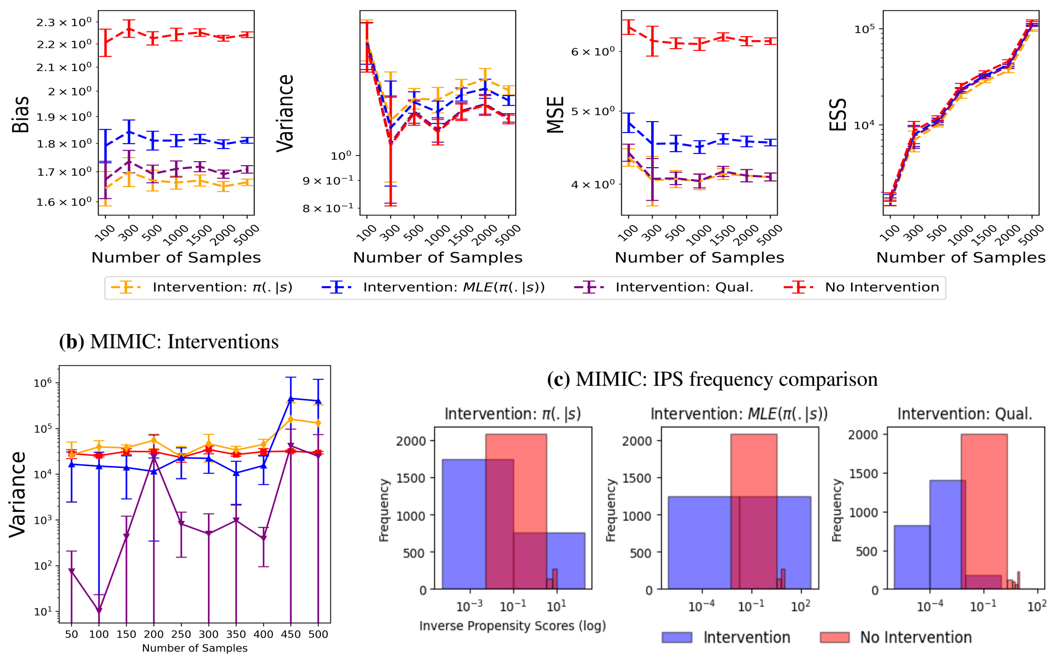


Figure 5: Interventions. (a-b) Qualitative interventions reduce bias in the learned concept estimator in Windy Gridworld and lower variance in MIMIC. In contrast, interventions based on traditional state-based policies $\pi(.|s)$ reduce bias and MSE compared to non-intervened concepts in Gridworld but are outperformed by qualitative interventions. (c) The frequency of high IPS scores decreases after applying qualitative interventions. Furthermore, $\pi(.|s)$-based interventions exhibit a higher frequency of IPS scores before intervention, indicating that not all strategies are equally effective.

Not all interventions improve Concept OPE characteristics and should be used at the practitioner's discretion. We analyze OPE after applying interventions using traditional state-based policies, $\pi(.|s)$. In gridworld, state-based interventions increase bias and MSE compared to qualitative ones, while in MIMIC, they lead to higher variance. This occurs because traditional state policies (π_b and π_e) fail to compensate for the lack of on-policy information, undermining the advantages of concept-based policies (π_b^c and π_e^c). In contrast, qualitative interventions—oracle concepts in WindyGridworld or urine output thresholds in MIMIC—retain these benefits and effectively address domain-specific challenges. Additionally, as shown in Figure 5c, $\pi(.|s)$-based interventions result in a higher frequency of both low and high IPS scores. However, since the effect of high IPS scores dominates, OPE variance increases compared to non-intervened OPE. Nevertheless, this framework allows practitioners to inspect and select among alternative interventions as needed.

8 Conclusions, Limitations and Future Work

We introduced a new family of concept-based OPE estimators, demonstrating that known-concept estimators can outperform traditional ones with greater accuracy and theoretical guarantees. For unknown concepts, we proposed an algorithm to learn interpretable concepts that improve OPE evaluations by identifying performance issues and enabling targeted interventions to reduce variance. These advancements benefit safety-critical fields like healthcare, education, and public policy by supporting reliable, interpretable policy evaluations. By reducing variance and providing policy insights, this approach enhances informed decision-making, facilitates personalized interventions, and refines policies before deployment for greater real-world effectiveness. A limitation of our work is trajectory distribution mismatch when learning unknown concepts, particularly in low-sample settings, which can lead to high-variance OPE. Targeted interventions help mitigate this issue. We also did not address hidden confounding variables or potential CBM concept leakage, focusing instead on evaluation. Future work will address these challenges and extend our approach to more general, partially observable environments.

References

Panagiotis Anagnostou, Petros T. Barmbas, Aristidis G. Vrahatis, and Sotiris K. Tasoulis. Approximate knn classification for biomedical data, 2020. URL https://arxiv.org/abs/2012.02149.

David M. Bossens and Philip S. Thomas. Low variance off-policy evaluation with state-based importance sampling, 2024. URL https://arxiv.org/abs/2212.03932.

Markus Böck, Julien Malle, Daniel Pasterk, Hrvoje Kukina, Ramin Hasani, and Clemens Heitzinger. Superhuman performance on sepsis mimic-iii data by distributional reinforcement learning. *PLOS ONE*, 17:e0275358, 11 2022. DOI: 10.1371/journal.pone.0275358.

Yinlam Chow, Marek Petrik, and Mohammad Ghavamzadeh. Robust policy optimization with baseline guarantees. *arXiv preprint arXiv:1506.04514*, 2015.

Shayan Doroudi, Philip S Thomas, and Emma Brunskill. Importance sampling for fair policy selection. *Grantee Submission*, 2017.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Raphael Fonteneau, Susan A Murphy, Louis Wehenkel, and Damien Ernst. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208:383–416, 2013.

Scott Fujimoto, David Meger, and Doina Precup. A deep reinforcement learning approach to marginalized importance sampling with the successor representation, 2023.

Ary L. Goldberger, Luis A. Nunes Amaral, L Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and Harry Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101 23:E215–20, 2000. URL https://api.semanticscholar.org/CorpusID:642375.

Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Anthony Celi, Emma Brunskill, and Finale Doshi-Velez. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions, 2020.

Deepak Gupta, Russell Loane, Soumya Gayen, and Dina Demner-Fushman. Medical image retrieval via nearest neighbor search on pre-trained image features, 2022. URL https://arxiv.org/abs/2210.02401.

Josiah Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pp. 604–613, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 0897919629. DOI: 10.1145/276698.276876. URL https://doi.org/10.1145/276698.276876.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, May 2016. ISSN 2052-4463. DOI: 10.1038/sdata.2016.35. URL https://europepmc.org/articles/PMC4878278.

Pulkit Katdare, Nan Jiang, and Katherine Rose Driggs-Campbell. Marginalized importance sampling for off-environment policy evaluation. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3778–3788. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/katdare23a.html.

John A. Kellum and John R. Prowle. Acute kidney injury in the critically ill: Clinical epidemiology and outcomes. *Nature Reviews Nephrology*, 14(10):641–656, 2018. DOI: 10.1038/s41581-018-0052-0.

Ramtin Keramati, Omer Gottesman, Leo Anthony Celi, Finale Doshi-Velez, and Emma Brunskill. Identification of subgroups with similar benefits in off-policy policy evaluation. *CoRR*, abs/2111.14272, 2021a. URL https://arxiv.org/abs/2111.14272.

Ramtin Keramati, Omer Gottesman, Leo Anthony Celi, Finale Doshi-Velez, and Emma Brunskill. Identification of subgroups with similar benefits in off-policy policy evaluation. *arXiv preprint arXiv:2111.14272*, 2021b.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716—1720, November 2018a. ISSN 1078-8956. DOI: 10.1038/s41591-018-0213-5. URL https://doi.org/10.1038/s41591-018-0213-5.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018b.

Matthieu Komorowski, Leo Anthony Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24:1716–1720, 2018c. DOI: 10.1038/s41591-018-0213-5. URL https://doi.org/10.1038/s41591-018-0213-5.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018a.

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation. *Advances in Neural Information Processing Systems*, 31, 2018b.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling, 2020. URL https://arxiv.org/abs/1910.06508.

Yao Liu, Yannis Flet-Berliac, and Emma Brunskill. Offline policy optimization with eligible actions, 2022. URL https://arxiv.org/abs/2207.00632.

Pedro Madeira, André Carreiro, Alex Gaudio, Luís Rosado, Filipe Soares, and Asim Smailagic. Zebra: Explaining rare cases through outlying interpretable concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3781–3787, 2023.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

Ritam Majumdar, Vishal Jadhav, Anirudh Deodhar, Shirish Karande, Lovekesh Vig, and Venkataramana Runkana. Symbolic regression for pdes using pruned differentiable programs, 2023. URL https://arxiv.org/abs/2303.07009.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

Ričards Marcinkevičs, Sonia Laguna, Moritz Vandenhirtz, and Julia E Vogt. Beyond concept bottleneck models: How to make black boxes intervenable? *arXiv preprint arXiv:2401.13544*, 2024.

Anton Matsson and Fredrik D. Johansson. Case-based off-policy policy evaluation using prototype learning, 2021. URL https://arxiv.org/abs/2111.11113.

S A Murphy, M J van der Laan, J M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96 (456):1410–1423, 2001. DOI: 10.1198/016214501753382327. URL https://doi.org/10.1198/016214501753382327. PMID: 20019887.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections, 2019.

Cosmin Paduraru. *Off-policy evaluation in Markov decision processes*. PhD thesis, 2013.

Konstantinos P Panousis, Dino Ienco, and Diego Marcos. Hierarchical concept discovery models: A concept pyramid scheme. *arXiv preprint arXiv:2310.02116*, 2023.

Brahma S. Pavse and Josiah P. Hanna. Scaling marginalized importance sampling to high-dimensional state-spaces via state abstraction, 2022a.

Brahma S. Pavse and Josiah P. Hanna. Scaling marginalized importance sampling to high-dimensional state-spaces via state abstraction, 2022b. URL https://arxiv.org/abs/2212.07486.

Achim Peine, Andreas Hallawa, Jan Bickenbach, Peter Sidler, Andreas Markewitz, Alexandre Levesque, Jeremy Levesque, and Nils Haake. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *npj Digital Medicine*, 4(1):32, 2021. DOI: 10.1038/s41746-021-00388-6.

Doina Precup, Richard S. Sutton, and Satinder P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Yoshihide Sawada and Keigo Nakamura. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10:41758–41765, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Simon P Shen, Yecheng Ma, Omer Gottesman, and Finale Doshi-Velez. State relevance for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9537–9546. PMLR, 2021.

Mervyn Singer, Clifford S. Deutschman, Christopher W. Seymour, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 2016. DOI: 10.1001/jama.2016.0287.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shengpu Tang and Jenna Wiens. Counterfactual-augmented importance sampling for semi-offline policy evaluation, 2023. URL https://arxiv.org/abs/2310.17146.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation, 2020.

Jean-Louis Vincent and Daniel De Backer. Circulatory shock. *New England Journal of Medicine*, 369(18):1726–1734, 2013. DOI: 10.1056/NEJMra1208943.

Carissa Wu, Sonali Parbhoo, Marton Havasi, and Finale Doshi-Velez. Learning optimal summaries of clinical time-series with concept bottleneck models. In *Machine Learning for Healthcare Conference*, pp. 648–672. PMLR, 2022.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in neural information processing systems*, 32, 2019.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values, 2020a.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values, 2020b.