Universal Off-Policy Evaluation

Yash Chandak
University of Massachusetts

Scott Niekum
University of Texas Austin

Bruno Castro da Silva
University of Massachusetts

Erik Learned-Miller
University of Massachusetts

Emma Brunskill
Stanford University

Philip S. Thomas
University of Massachusetts

Abstract

When faced with sequential decision-making problems, it is often useful to be able to predict what would happen if decisions were made using a new policy. Those predictions must often be based on data collected under some previously used decision-making rule. Many previous methods enable such *off-policy* (or counterfactual) estimation of the *expected* value of a performance measure called the *return*. In this paper, we take the first steps towards a *universal off-policy estimator* (UnO)—one that provides off-policy estimates and high-confidence bounds for *any* parameter of the return distribution. We use UnO for estimating and simultaneously bounding the mean, variance, quantiles/median, inter-quantile range, CVaR, and the entire cumulative distribution of returns. Finally, we also discuss UnO's applicability in various settings, including fully observable, partially observable (i.e., with unobserved confounders), Markovian, non-Markovian, stationary, smoothly non-stationary, and discrete distribution shifts.

1 Introduction

Problems requiring sequential decision-making are ubiquitous [5]. When online experimentation is costly or dangerous, it is essential to conduct off-policy evaluation before deploying a new policy; that is, one must leverage existing data collected using some policy β (called a behavior policy) to evaluate a performance metric of another policy π (called the evaluation policy). For problems with high stakes, such as in terms of health [56] or financial assets [86], it is also crucial to provide high-confidence bounds on the desired performance metric to ensure reliability and safety.

Perhaps the most widely studied performance metric in the off-policy setting is the expected return [83]. However, this metric can be limiting for many problems of interest. Safety-critical applications, such as automated healthcare, require minimizing the chances of risk-prone outcomes, and so performance metrics such as value at risk (VaR) or conditional value at risk (CVaR) are more appropriate [49, 14]. By contrast, applications like online recommendations are subject to noisy data and call for robust metrics like the median and other quantiles [2]. In order to improve user experiences, applications involving direct human-machine interaction, such as robotics and autonomous driving, focus on minimizing uncertainty in their outcomes and thus use metrics like variance and entropy [52, 84]. Recent work in distributional reinforcement learning (RL) have also investigated estimating the cumulative distribution of returns [7, 24] and its various statistical functionals [76]. While it may even be beneficial to use all of these different metrics simultaneously to inform better decision-making, even individually estimating and bounding any performance metric, other than mean and variance, in the *off-policy setting* has remained an open problem.

This raises the main question of interest: *How do we develop a universal off-policy method—one that can estimate any desired performance metrics and can also provide finite-sample confidence bounds that hold simultaneously with high probability for those metrics?*

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Prior Work: Off-policy methods can be broadly categorized as model-based or model-free [83]. Model-based methods typically require strong assumptions on the parametric model when statistical guarantees are needed. Further, using model-based approaches to estimate parameters other than the mean can also require estimating the *distribution* of rewards for *every* state-action pair in order to obtain the complete return distribution for any policy.

By contrast, model-free methods are applicable to a wider variety of settings. Unfortunately, the popular technique of using *importance-weighted returns* [71] only corrects for the *mean* under the off-policy distribution. Recent work by Chandak et al. [18] provides a specialized extension to only correct for the variance. Outside RL, works in the econometrics and causal inference literature have also considered quantile treatments [29, 99] and inferences on counterfactual distributions [28, 20, 36], but these methods are not developed for sequential decisions and do not provide any high-confidence bounds with guaranteed coverage. Further, they often mandate stationarity, identically distributed data, and full observability (i.e., no confounding).

Existing frequentist high-confidence bounds are not only specifically designed for either the mean or variance, but also hold only *individually* [92, 45, 18]. Instead of frequentist intervals, a Bayesian posterior distribution over the mean return and various statistics of that distribution can also be obtained [105]. We are not aware of any method that provides off-policy bounds or even estimates for *any* parameter of the return, while also handling different domain settings that are crucial for RL related tasks. Therefore, a detailed discussion of existing work is deferred to Appendix C.

Contributions: We take the first steps towards a *universal off-policy estimator* (UnO) that estimates and bounds the *entire distribution* of returns, and then derives estimates and simultaneous bounds for all parameters of interest. With UnO, we make the following contributions:

A. For *any* distributional parameter (mean, variance, quantiles, entropy, CVaR, CDF, etc.), we provide an off-policy method to obtain **(A.1)** model-free estimators; **(A.2)** high-confidence bounds that have guaranteed coverage *simultaneously* for all parameters and that, perhaps surprisingly, often nearly match or outperform prior bounds specifically designed for the mean and the variance; and **(A.3)** approximate bounds using statistical bootstrapping that can often be significantly tighter.

B. The above advantages hold for **(B.1)** fully observable and partially observable (i.e., with unobserved confounders) settings, **(B.2)** Markovian and non-Markovian settings, and **(B.3)** settings with stationary, smoothly non-stationary, and discrete distribution shifts in a policy's performance.

Limitations: Our method uses importance sampling and thus **(1)** Requires knowledge of action probabilities under the behavior policy β, **(2)** Any outcome under the evaluation policy should have a sufficient probability of occurring under β, and **(3)** Variance of our estimators scales exponentially with the horizon length [39, 57], which may be unavoidable in non-Markovian domains [46].

Notation: For brevity, we first restrict our focus to the stationary setting. In Section 5, we discuss how to tackle non-stationarity and distribution shifts. A *partially observable Markov decision process* (POMDP) is a tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \Omega, \mathcal{R}, \gamma, d_0)$, where $\mathcal{S}$ is the set of states, $\mathcal{O}$ is the set of observations, $\mathcal{A}$ is the set of actions, $\mathcal{P}$ is the transition function, Ω is the observation function, $\mathcal{R}$ is the reward function, $\gamma \in [0, 1]$ is the discount factor, and d_0 is the starting state distribution. Although our results extend to the continuous setting, for notational ease, we consider $\mathcal{S}, \mathcal{A}, \mathcal{O}$, and the set of rewards to be finite. Since the true underlying states are only partially observable, the resulting rewards and transitions from one partially observed state to another are therefore also potentially non-Markovian [80]. We write S_t, O_t, A_t, and R_t to denote random variables for state, observation, action, and reward respectively at time t. Let $\mathcal{D}$ be a data set $(H_i)_{i=1}^n$ collected using *behavior* policies $(\beta_i)_{i=1}^n$, where each H_i denotes the *observed trajectory* $(O_0, A_0, \beta(A_0|O_0), R_0, O_1, ...)$. Notice that an observed trajectory contains $\beta(A_t|O_t)$ and does not contain the states S_t, for all t. Let $G_i := \sum_{j=0}^{T} \gamma^j R_j$ be the *return* of H_i, where $\forall i,\ G_{\min} < G_i < G_{\max}$ for some finite constants $G_{\min}$ and $G_{\max}$, and T is a finite horizon length. Let G_π and H_π be the random variables for returns and complete trajectories under any policy π, respectively. Since the set of observations, actions, and rewards are finite, and T is finite, the total number of possible trajectories is finite. Let $\mathcal{X}$ be the finite set of returns corresponding to these trajectories. Let $\mathscr{H}_\pi$ be the set of all possible trajectories for any policy π. Sometimes, to make the dependence explicit, we write $g(h)$ to denote the return of trajectory h. Further, to ensure that samples in $\mathcal{D}$ are informative, we make a standard assumption that any outcome under π has sufficient probability of occurring under β (see Appendix B.1 for further discussion of assumptions in general),

Assumption 1. *The set $\mathcal{D}$ contains independent (not necessarily identically distributed) observed trajectories generated using $(\beta_i)_{i=1}^n$, such that for some (unknown) $\varepsilon > 0$, $(\beta_i(a|o) < \varepsilon) \implies (\pi(a|o) = 0)$, for all $o \in \mathcal{O}, a \in \mathcal{A}$, and $i \in \{1, 2, ..., n\}$.*

2 Idea Summary

For the desired universal method, instead of considering each parameter individually, we suggest estimating the entire *cumulative distribution function* (CDF) of returns first:

$$\forall \nu \in \mathbb{R}, \qquad F_\pi(\nu) \coloneqq \Pr\Big(G_\pi \leq \nu\Big).$$

Any distributional parameter, $\psi(F_\pi)$, can then be estimated from the estimate of F_π. However, we only have off-policy data from a behavior policy β, and the typical use of importance sampling [71] only corrects for the mean return. To overcome this, we propose an estimator $\hat{F}_n$ that uses importance sampling from the *perspective of the CDF* to correct for the *entire* distribution of returns. The CDF estimate, $\hat{F}_n$, is then used to obtain a plug-in estimator $\psi(\hat{F}_n)$ for any distributional parameter $\psi(F_\pi)$.

Next, we show that this CDF-centric perspective provides the additional advantage that, if we can compute a $1-\delta$ *confidence band* $\mathcal{F} : \mathbb{R} \to 2^{\mathbb{R}}$ such that

$$\Pr\Big(\forall \nu \in \mathbb{R},\ \Pr\left(G_\pi \leq \nu\right) \in \mathcal{F}(\nu)\Big) \geq 1 - \delta,$$

then a $1-\delta$ upper (or lower) high-confidence bound on any parameter, $\psi(F_\pi)$, can be obtained by searching for a function F that maximizes (or minimizes) $\psi(F)$ and $\forall \nu \in \mathbb{R}$ has $F(\nu) \in \mathcal{F}(\nu)$.

3 UnO: Universal Off-Policy Estimator

In the *on-policy* setting, one approach for estimating any parameter of returns, G_π, might be to first estimate its *cumulative distribution* F_π and then use that to estimate its parameter $\psi(F_\pi)$. However, doing this in the off-policy setting requires additional consideration as the *entire* distribution of the observed returns needs to be adjusted to estimate F_π since the data is collected using behavior policies that can be different from the evaluation policy π.

We begin by observing that $\forall \nu \in \mathbb{R}$, $F_\pi(\nu)$ can be expanded using the fact that the probability that the return G_π equals x is the sum of the probabilities of the trajectories H_π whose return equals x,

$$F_\pi(\nu) = \Pr(G_\pi \leq \nu) = \sum_{x \in \mathcal{X}, x \leq \nu} \Pr(G_\pi = x) = \sum_{x \in \mathcal{X}, x \leq \nu} \left(\sum_{h \in \mathscr{H}_\pi} \Pr(H_\pi = h) \mathbb{1}_{\{g(h) = x\}} \right), \quad (1)$$

where $\mathbb{1}_A = 1$ if A is true and 0 otherwise. Now, observing that the indicator function can be one for at most a single value less than ν as $g(h)$ is a deterministic scalar given h, (1) can be expressed as,

$$F_\pi(\nu) = \sum_{h \in \mathscr{H}_\pi} \Pr(H_\pi = h) \sum_{x \in \mathcal{X}, x \leq \nu} \mathbb{1}_{\{g(h) = x\}} = \sum_{h \in \mathscr{H}_\pi} \Pr(H_\pi = h) \Big(\mathbb{1}_{\{g(h) \leq \nu\}} \Big),$$

where the red color is used to highlight changes. Now, from Assumption 1 as $\forall \beta,\ \mathscr{H}_\pi \subseteq \mathscr{H}_\beta$,[1]

$$F_\pi(\nu) = \sum_{h \in \mathscr{H}_\beta} \Pr(H_\pi = h) \Big(\mathbb{1}_{\{g(h) \leq \nu\}} \Big) = \sum_{h \in \mathscr{H}_\beta} \Pr(H_\beta = h) \frac{\Pr(H_\pi = h)}{\Pr(H_\beta = h)} \Big(\mathbb{1}_{\{g(h) \leq \nu\}} \Big). \quad (2)$$

The form of $F_\pi(\nu)$ in (2) is beneficial as it suggests a way to not only perform off-policy corrections for one specific parameter, as in prior works [71, 18], but for the *entire cumulative distribution function* (CDF) of return G_π. Formally, let $\rho_i \coloneqq \prod_{j=0}^{T} \frac{\pi(A_j|O_j)}{\beta_i(A_j|O_j)}$ denote the importance ratio for H_i, which is equal to $\Pr(H_\pi = h) / \Pr(H_\beta = h)$ (see Appendix D).

Then, based on (2), we propose the following non-parametric and model-free estimator for F_π.

$$\forall \nu \in \mathbb{R}, \quad \hat{F}_n(\nu) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \rho_i \mathbb{1}_{\{G_i \leq \nu\}}. \quad (3)$$

[1] Results can be extended to hybrid probability measures using Radon-Nikodym derivatives.

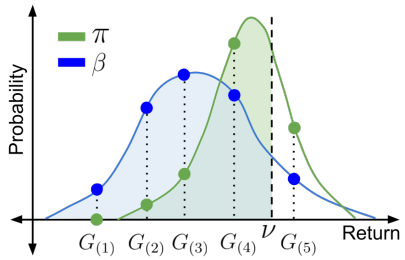

Figure 1: An illustration of return distributions for π and β. The CDF at any point ν corresponds to the area under the probability distribution up until ν. Having order statistics $(G_{(i)})_{i=1}^5$ of samples $(G_i)_{i=1}^5$ drawn using β, (3) constructs an empirical estimate of the CDF for π (*green* shaded region) by correcting for the probability of observing each G_i using the *importance-sampled counts* of $G_i \leq \nu$. Additionally, weighted-IS (WIS) can be used as in (27) for a variance-reduced estimator for F_π.

Figure 1 provides intuition for (3). In the following theorem, we establish that this estimator, $\hat{F}_n$, is unbiased and not only pointwise consistent, but also a uniformly consistent estimator of F_π, even when the data $\mathcal{D}$ is collected using multiple behavior policies $(\beta_i)_{i=1}^n$. The proof (deferred to Appendix D) also illustrates that by using knowledge of action probabilities under the behavior policies, no additional adjustments (e.g., front-door or backdoor [70]) are required by $\hat{F}_n$ to estimate F_π, even when the domain is non-Markovian or has partial observability (confounders).

Theorem 1. *Under Assumption 1, $\hat{F}_n$ is an unbiased and uniformly consistent estimator of F_π,*

$$\forall \nu \in \mathbb{R}, \quad \mathbb{E}_{\mathcal{D}}\left[\hat{F}_n(\nu)\right] = F_\pi(\nu), \qquad \sup_{\nu \in \mathbb{R}} \left|\hat{F}_n(\nu) - F_\pi(\nu)\right| \xrightarrow{a.s.} 0.$$

Remark 1. *Notice that the value of $\hat{F}_n(\nu)$ can be more than one, even though $F_\pi(\nu)$ cannot have a value greater than one for any $\nu \in \mathbb{R}$. This is an expected property of estimators based on importance sampling (IS). For example, the IS estimates of expected return during off-policy mean estimation can be smaller or larger than the smallest and largest possible return when $\rho > 1$.*

Having an estimator $\hat{F}_n$ of F_π, any parameter $\psi(F_\pi)$ can now be estimated using $\psi(\hat{F}_n)$. However, some parameters like the mean μ_π, variance σ_π^2, and entropy $\mathcal{H}_\pi$, are naturally defined using the probability distribution $\mathrm{d}F_\pi$ instead of the cumulative distribution F_π. Similarly, parameters like the α-quantile Q_π^α and inter-quantile range (which provide tail-robust measures for the mean and deviation from the mean) and conditional value at risk CVaR_π^α (which is a tail-sensitive measure) are defined using the inverse CDF $F_\pi^{-1}(\alpha)$. Therefore, let $(G_{(i)})_{i=1}^n$ be the *order statistics* for samples $(G_i)_{i=1}^n$ and $G_{(0)} \coloneqq G_{\min}$. Then, we define the off-policy estimator of the inverse CDF for all $\alpha \in [0, 1]$, and the probability distribution estimator $\mathrm{d}\hat{F}_n$ as,

$$\hat{F}_n^{-1}(\alpha) \coloneqq \min\left\{g \in (G_{(i)})_{i=1}^n \middle| \hat{F}_n(g) \geq \alpha\right\}, \qquad \mathrm{d}\hat{F}_n(G_{(i)}) \coloneqq \hat{F}_n(G_{(i)}) - \hat{F}_n(G_{(i-1)}), \tag{4}$$

where $\mathrm{d}\hat{F}_n(\nu) \coloneqq 0$ if $\nu \neq G_{(i)}$ for any $i \in (1, \ldots, n)$. Using (4), we now define off-policy estimators for parameters like the mean, variance, quantiles, and CVaR (see Appendix E.1 for more details on these). This procedure can be generalized to any other parameter of F_π for which a sample estimator $\psi(\hat{F}_n)$ can be directly created using $\hat{F}_n$ as a plug-in estimator for F_π.

$$\mu_\pi(\hat{F}_n) \coloneqq \sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)}) G_{(i)}, \qquad \sigma_\pi^2(\hat{F}_n) \coloneqq \sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)}) \left(G_{(i)} - \mu_\pi(\hat{F}_n)\right)^2,$$

$$Q_\pi^\alpha(\hat{F}_n) \coloneqq \hat{F}_n^{-1}(\alpha), \qquad \text{CVaR}_\pi^\alpha(\hat{F}_n) \coloneqq \frac{1}{\alpha} \sum_{i=1}^n \mathrm{d}\hat{F}_n(G_{(i)}) G_{(i)} \mathbb{1}_{\{G_{(i)} \leq Q_\pi^\alpha(\hat{F}_n)\}}.$$

Remark 2. *Let H_i be the observed trajectory for the G_i that gets mapped to $G_{(i)}$ when computing the order statistics. Note that $d\hat{F}_n(G_{(i)})$ equals ρ_i/n for this H_i. This implies that the estimator for the mean, $\mu_\pi(\hat{F}_n)$, reduces exactly to the existing full-trajectory-based IS estimator [71].*

Notice that many parameters and their sample estimates discussed above are nonlinear in F_π and $\hat{F}_n$, respectively (the mean is one exception). Therefore, even though $\hat{F}_n$ is an unbiased estimator of F_π, the sample estimator, $\psi(\hat{F}_n)$, may be a biased estimator of $\psi(F_\pi)$. This is expected behavior because even in the on-policy setting it is not possible to get unbiased estimates of some parameters (e.g., standard deviation), and UnO reduces to the on-policy setting when $\pi = \beta$. However, perhaps surprisingly, we establish in the following section that even when $\psi(\hat{F}_n)$ is a biased estimator of $\psi(F_\pi)$, high-confidence upper and lower bounds can still be computed for both F_π and $\psi(F_\pi)$.

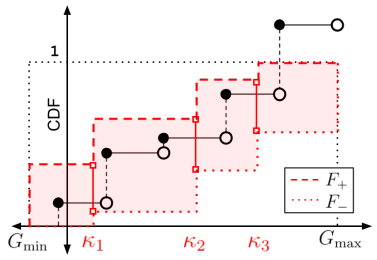

Figure 2: An illustration of $\hat{F}_n$ (in black) using five return samples and the confidence band $\mathcal{F}$ (red shaded region) computed using (5) with confidence intervals (red lines) at three key points $(\kappa_i)_{i=1}^3$. Notice that the vertical "steps" in $\hat{F}_n$ can be of different heights and their total can be greater than 1 due to importance weighting. However, since we know that F_π is never greater than 1, $\mathcal{F}$ can be clipped at 1.

4 High-Confidence Bounds for UnO

Off-policy estimators are typically prone to high variance, and when the domain can be non-Markovian, the curse of horizon might be unavoidable [46]. For critical applications, this might be troublesome [94] and thus necessitates obtaining confidence intervals to determine how much our estimates can be trusted. Therefore, in this section, we aim to construct a set of possible CDFs $\mathcal{F} : \mathbb{R} \to 2^{\mathbb{R}}$, called a *confidence band*, such that the true $F_\pi(\nu)$ is within the set $\mathcal{F}(\nu)$ with high probability, i.e., $\Pr(\forall \nu \in \mathbb{R},\ F_\pi(\nu) \in \mathcal{F}(\nu)) \geq 1 - \delta$, for any $\delta \in (0, 1]$. Subsequently, we develop finite-sample bounds for any parameter $\psi(F_\pi)$ using $\mathcal{F}$.

In the on-policy setting, $\mathcal{F}$ can be constructed using the DKW inequality [31] and its tight constants [60]. However, its applicability to the off-policy setting is unclear as **(a)** unlike the on-policy CDF estimate, the "steps" of an off-policy CDF estimate are not of equal heights, **(b)** the "steps" do not sum to one (see Figure 2) and the maximum height of the steps need not be known either, and **(c)** DKW assumes samples are identically distributed, however, off-policy data $\mathcal{D}$ might be collected using multiple different behavior policies. This raises the question: *How do we obtain $\mathcal{F}$ in the off-policy setting?*

Before constructing a confidence band $\mathcal{F}$, let us first focus on obtaining bounds for a single point, $F_\pi(\kappa)$. Let $X \coloneqq \rho(\mathbb{1}_{\{G \leq \kappa\}})$. Then, from Theorem 1, we have that $\mathbb{E}_{\mathcal{D}}[X] = F_\pi(\kappa)$. This implies that a confidence interval for the mean of X provides a confidence interval for $F_\pi(\kappa)$. Using this observation, existing confidence intervals for the mean of a bounded random variable can be directly applied to X to obtain a confidence interval for $F_\pi(\kappa)$. For example, Thomas et al. [91] present tight bounds for the mean of IS-based random variables by mitigating the variance resulting from the heavy tails associated with IS; we use their method on $\hat{F}_n(\kappa)$ to bound $F_\pi(\kappa)$. Alternatively, recent work by Kuzborskij et al. [53] can potentially be used with a WIS-based F_π estimate (27).

Before moving further, we introduce some additional notation. Let $(\kappa_i)_{i=1}^K$ be any K "key points" and let $\texttt{CI}_-(\kappa_i, \delta_i)$ and $\texttt{CI}_+(\kappa_i, \delta_i)$ be the lower and the upper confidence bounds on $F_\pi(\kappa_i)$ constructed at each key point using the observation made in the previous paragraph, such that

$$\forall i \in (1, ..., K), \quad \Pr\Big(\texttt{CI}_-(\kappa_i, \delta_i) \leq F_\pi(\kappa_i) \leq \texttt{CI}_+(\kappa_i, \delta_i)\Big) \geq 1 - \delta_i.$$

We now use the following observation to obtain a band, $\mathcal{F}$, that contains F_π with high confidence. Because F_π is a CDF, it is necessarily monotonically non-decreasing, and so if $F_\pi(\kappa_i) \geq \texttt{CI}_-(\kappa_i, \delta_i)$ then for any $\nu \geq \kappa_i$, $F_\pi(\nu)$ must be no less than $\texttt{CI}_-(\kappa_i, \delta_i)$. Similarly, if $F_\pi(\kappa_i) \leq \texttt{CI}_+(\kappa_i, \delta_i)$ then for any $\nu \leq \kappa_i$, $F_\pi(\nu)$ must also be no greater than $\texttt{CI}_+(\kappa_i, \delta_i)$. Let $\kappa_0 \coloneqq G_{\min}$, $\kappa_{K+1} \coloneqq G_{\max}$, $\texttt{CI}_-(\kappa_0, \delta_0) \coloneqq 0$, and $\texttt{CI}_+(\kappa_{K+1}, \delta_{K+1}) \coloneqq 1$; then, as illustrated in Figure 2, we can construct a lower function F_- and an upper function F_+ that encapsulate F_π with high probability,

$$F_-(\nu) \coloneqq \begin{cases} 1 & \text{if } \nu > G_{\max}, \\ \max\limits_{\kappa_i \leq \nu} \texttt{CI}_-(\kappa_i, \delta_i) & \text{otherwise.} \end{cases} \qquad F_+(\nu) \coloneqq \begin{cases} 0 & \text{if } \nu < G_{\min}, \\ \min\limits_{\kappa_i \geq \nu} \texttt{CI}_+(\kappa_i, \delta_i) & \text{otherwise.} \end{cases} \tag{5}$$

Theorem 2. *Under Assumption 1, for any $\delta \in (0, 1]$, if $\sum_{i=1}^K \delta_i \leq \delta$, then the confidence band defined by F_- and F_+ provides guaranteed coverage for F_π. That is,*

$$\Pr\Big(\forall \nu \in \mathbb{R},\ F_-(\nu) \leq F_\pi(\nu) \leq F_+(\nu)\Big) \geq 1 - \delta.$$

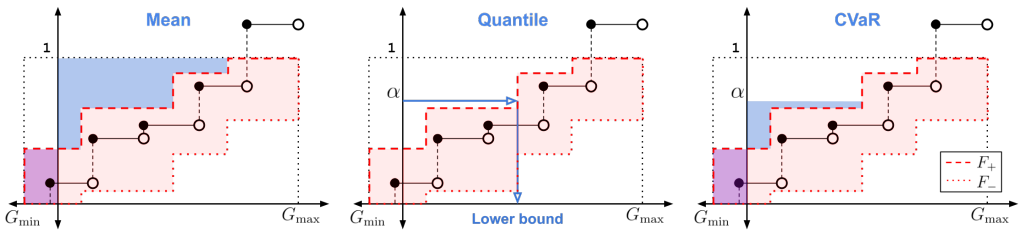

Figure 3: Given a confidence band $\mathcal{F}$, bounds for many parameters can be obtained using geometry. **(Left)** For a lower bound on the mean, we would want a CDF $F \in \mathcal{F}$ that assigns as high a probability as possible on lower G values, and F_+ is the CDF which does that. To obtain the mean of F_+, we use the property that the mean of a distribution is the area above the CDF on the positive x-axis minus the area below the CDF on the negative x-axis [3]. Hence, the mean of the distribution characterized by F_+ is the area of the shaded blue region minus the area of the shaded purple region, and this value is the high-confidence lower bound on the mean. **(Middle)** Similarly, within $\mathcal{F}$, F_+ characterizes the distribution with the smallest α-quantile. **(Right)** Building upon the lower bounds for the mean and the quantile, Thomas and Learned-Miller [90] showed that the lower bound for α-CVaR can be obtained using the area of the shaded blue region minus the area of the shaded purple region, normalized by α. To get the upper bounds on the mean, quantile, and CVaR, analogous arguments hold using the lower bound CDF F_-. See Appendix E.5 for discussions of variance, inter-quantile, entropy, and other parameters.

Remark 3. *Notice that any choice of $(\kappa_i)_{i=1}^K$ results in a valid band $\mathcal{F}$. However, $\mathcal{F}$ can be made tighter by optimizing over the choice of $(\kappa_i)_{i=1}^K$. In Appendix E.5, we present one such method using cross-validation to minimize the area enclosed within $\mathcal{F}$.*

Having obtained a high-confidence band for F_π, we now discuss how high-confidence bounds for any parameter $\psi(F_\pi)$ can be obtained using this band. Formally, with a slight overload of notation let $\mathcal{F}$ be the set of all possible CDFs bounded between F_- and F_+, that is,

$$\mathcal{F} \coloneqq \left\{ F \;\middle|\; \forall \nu \in \mathbb{R},\ F_-(\nu) \le F(\nu) \le F_+(\nu) \right\}.$$

This band $\mathcal{F}$ contains many possible CDFs, one of which is F_π with high probability. Therefore, to get a lower or upper bound, ψ_- or ψ_+, on $\psi(F_\pi)$, we propose deriving a CDF $F \in \mathcal{F}$ that minimizes or maximizes $\psi(F)$, respectively, and we show that these contain $\psi(F_\pi)$ with high probability:

$$\psi_- \coloneqq \inf_{F \in \mathcal{F}} \ \psi(F), \qquad \psi_+ \coloneqq \sup_{F \in \mathcal{F}} \ \psi(F). \tag{6}$$

Theorem 3. *Under Assumption 1, for any $1-\delta$ confidence band $\mathcal{F}$, the confidence interval defined by ψ_- and ψ_+ provides guaranteed coverage for $\psi(F_\pi)$. That is,*

$$\Pr\left(\psi_- \le \psi(F_\pi) \le \psi_+\right) \ge 1 - \delta.$$

While obtaining ψ_- might not look straightforward, one can obtain closed-form expressions for many popular parameters of interest. In other cases, simple algorithms exist for computing ψ_- and ψ_+ [74]. Figure 3 provides geometric depictions of the closed-form expressions for some parameters.

Remark 4. *Perhaps surprisingly, even though $\psi(\hat{F}_n)$ may be biased, we can obtain high-confidence bounds with guaranteed coverage on any $\psi(F_\pi)$ using the confidence band $\mathcal{F}$. In fact, confidence bounds for* all *parameters computed using (6) hold* simultaneously *with probability at least $1-\delta$ as they are all derived from the same confidence band, $\mathcal{F}$.*

3.1. Statistical Bootstrapping: An important advantage of having constructed an off-policy estimator of any $\psi(F_\pi)$ is that it opens up the possibility of using *resampling*-based methods, like statistical bootstrapping [32], to obtain *approximate* confidence intervals for $\psi(F_\pi)$. In particular, we can use the *bias-corrected and accelerated* (BCa) bootstrap procedure to obtain ψ_- and ψ_+ for $\psi(F_\pi)$. This procedure is outlined in Algorithm 1 in Appendix E.4.

Unlike the bounds from (6), BCa-based bounds do not offer guaranteed coverage and need to be computed individually for each parameter ψ. However, they can be combined with UnO to get significantly tighter bounds with less data, albeit without guaranteed coverage.

5 Confounding, Distributional Shifts, and Smooth Non-Stationarities

A particular advantage of UnO is the remarkable simplicity with which the estimates and bounds for F_π or $\psi(F_\pi)$ can be extended to account for confounding, distributional shifts, and smooth non-stationarities that are prevalent in real-world applications [30].

Confounding / Partial Observability: Estimator $\hat{F}_n$ in (3) accounts for partial observability when both π and β have the same observation set. However, in systems like automated loan approval [94], data might have been collected using a behavior policy β dependent on sensitive attributes like race and gender that may no longer be allowable under modern laws. This can make the available observation, $\widetilde{O}$, for an evaluation policy π different from the observations, O, for β, which may also have been a partial observation of the underlying true state S.

However, an advantage of many such automated systems (e.g., online recommendation, automated healthcare, robotics) is the direct availability of behavior probabilities $\beta_i(A|O)$. In Appendix D, we provide generalized proofs for all the earlier results, showing that access to $\beta_i(A|O)$ allows UnO to handle various sources of confounding even when $\widetilde{O} \neq O$, without requiring any additional adjustments. When $\beta_i(A|O)$ is not available, we allude to possible alternatives in Appendix B.1.

Distribution Shifts: Many practical applications exhibit distribution shifts that might be discrete or abrupt. One example is when a medical treatment developed for one demographic is applied to another [37]. To tackle discrete distributional shifts, let $F_\pi^{(1)}$ and $F_\pi^{(2)}$ denote the CDFs of returns under policy π in the first and the second domain, respectively. To make the problem tractable, similar to prior work on characterizing distribution shifts [10], we assume that the Kolmogorov-Smirnov distance between $F_\pi^{(1)}$ and $F_\pi^{(2)}$ is bounded.

Assumption 2. *There exists $\epsilon \geq 0$, such that* $\sup_{\nu \in \mathbb{R}} \left| F_\pi^{(1)}(\nu) - F_\pi^{(2)}(\nu) \right| \leq \epsilon$.

Given data $\mathcal{D}$ collected in the first domain, one can obtain the bounds $F_-^{(1)}$ and $F_+^{(1)}$ on $F_\pi^{(1)}$ as in Section 4. Now since $F_\pi^{(2)}$ can differ from $F_\pi^{(1)}$ by at most ϵ at any point, we propose the following bounds for $F_\pi^{(2)}$ for all $\nu \in \mathbb{R}$ and show that they readily provide guaranteed coverage for $F_\pi^{(2)}$:

$$F_-^{(2)}(\nu) \coloneqq \max(0, F_-^{(1)}(\nu) - \epsilon), \qquad F_+^{(2)}(\nu) \coloneqq \min(1, F_+^{(1)}(\nu) + \epsilon). \tag{7}$$

Theorem 4. *Under Assumptions 1 and 2, $\forall \delta \in (0, 1]$, the confidence band defined by $F_-^{(2)}$ and $F_+^{(2)}$ provides guaranteed coverage for $F_\pi^{(2)}$. That is,* $\Pr(\forall \nu,\ F_-^{(2)}(\nu) \leq F_\pi^{(2)}(\nu) \leq F_+^{(2)}(\nu)) \geq 1 - \delta$.

Smooth Non-stationarity: The stationarity assumption is unreasonable for applications like online tutoring or recommendation systems, which must deal with drifts of students' interests or seasonal fluctuations of customers' interests [93, 88]. In the worst case, however, even a small change in the transition dynamics can result in a large fluctuation of a policy's performance and make the problem intractable. Therefore, similar to the work of Chandak et al. [16], we assume that the distribution of returns for any π changes smoothly over the past episodes 1 to L, and the ℓ episodes in the future. In particular, we assume that the trend of $F_\pi^{(i)}(\nu)$ for all ν can be modeled using least-squares regression using a nonlinear basis function $\phi : \mathbb{R} \to \mathbb{R}^d$ (e.g., the Fourier basis, which is popular for modeling non-stationary trends [12]).

Assumption 3. *For any ν, $\exists w_\nu \in \mathbb{R}^d$, such that, $\forall i \in [1, L + \ell]$,* $F_\pi^{(i)}(\nu) = \phi(i)^\top w_\nu$.

Estimating $F_\pi^{(L+\ell)}$ can now be seen as a time-series forecasting problem. Formally, for any key point κ, let $\hat{F}_n^{(i)}(\kappa)$ be the estimated CDF using H_i observed in episode i. From Theorem 1, we know that $\hat{F}_n^{(i)}(\kappa)$ is an unbiased estimator of $F_\pi^{(i)}(\kappa)$; therefore, $(\hat{F}_n^{(i)}(\kappa))_{i=1}^L$ is an unbiased estimate for the underlying time-varying sequence $(F_\pi^{(i)}(\kappa))_{i=1}^L$. Now, using methods from time-series literature, the trend of $(\hat{F}_n^{(i)}(\kappa))_{i=1}^L$ can be analyzed to forecast $F_\pi^{(L+\ell)}(\kappa)$, along with its `CIs`. In particular, we propose using *wild bootstrap* [58, 26], which provides *approximate* `CIs` with finite sample error of $O(L^{-1/2})$ while also handling non-normality and heteroskedasticity, which would occur when dealing with IS-based estimates resulting from different behavior polices [16]. See Appendix E.6 for more details. Finally, using the bounds obtained using wild bootstrap at multiple key points, an entire confidence band can be obtained as discussed in Section 4.

6 Empirical Studies

In this section, we provide empirical support for the established theoretical results for the proposed UnO estimator and high-confidence bounds. To do so, we use the following domains: **(1)** An open source implementation [102] of the FDA-approved type-1 diabetes treatment simulator [59], **(2)** A stationary and a non-stationary recommender system domain, and **(3)** A continuous-state Gridworld with partial observability, where data is collected using multiple behavior policies. Detailed description for domains and the procedures for obtaining π and β are provided in Appendix F.1; code is also publicly available here. In the following, we discuss four primary takeaway results.

(A) Characteristics of the UnO estimator: Figure 4 reinforces the universality of UnO. As can be seen, UnO can accurately estimate the entire CDF and a wide range of its parameters: mean, variance, quantile, and CVaR.

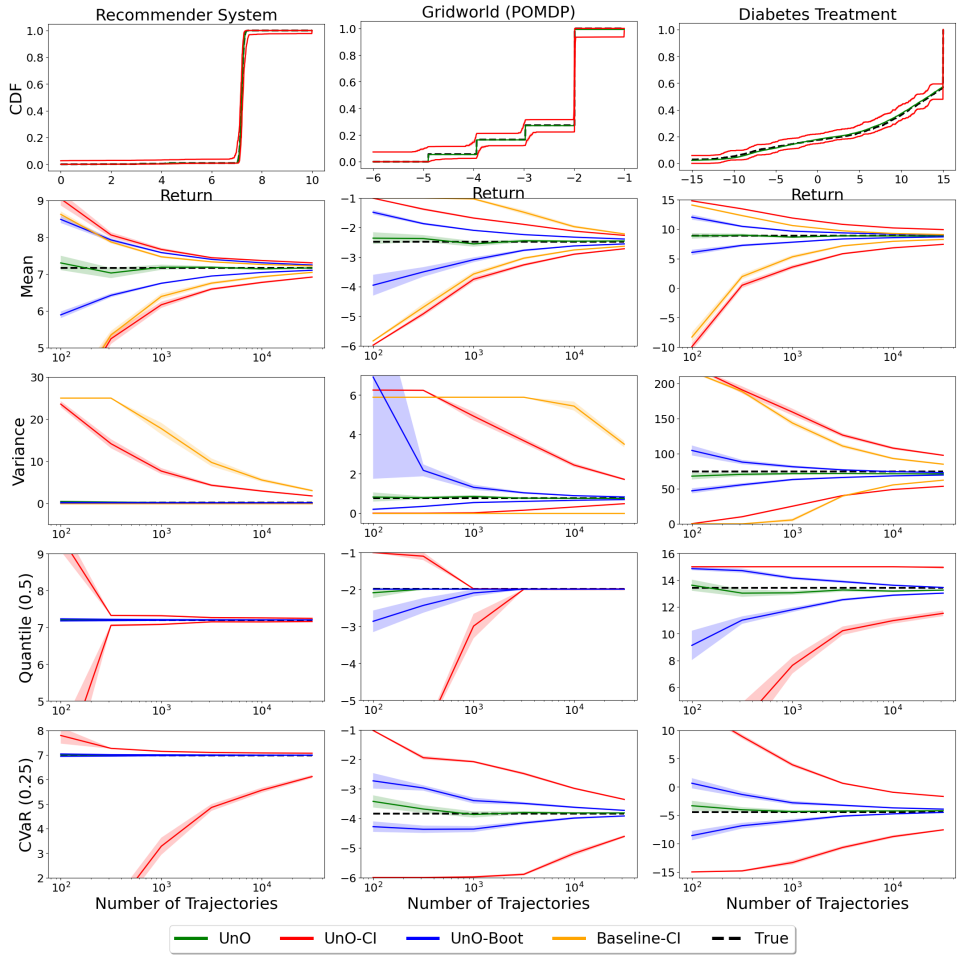

Figure 4: Performance trend of the proposed estimators and bounds on three domains. The black dashed line is the true value of F_π or $\psi(F_\pi)$, green is our UnO estimator, red is our `CI`-based UnO bound, blue is the bootstrap version of our UnO bound, and yellow is the baseline bound for the mean [91] or variance [18]. Each bound has two lines (upper and lower); however, some are not visible due to overlaps. The shaded regions are ± 2 standard error, computed using 30 trials. The plots in the top row are for CDFs obtained using $3 \times 10^{4.5}$ samples. The next four rows are for different parameters and share the same x-axis. Bounds were obtained for a failure rate $\delta = 0.05$. Since the UnO-Boot and Baseline-CI methods do not hold simultaneously for all the parameters, they were made to hold with failure rate of $\delta/4$ for a fair comparison (as there are 4 parameters in this plot).

(B) Comparison of UnO with prior work: Recent works for bounding the mean [45, 35] assume no confounding and Markovian structure. Therefore, for a fair comparison, we resort to the method

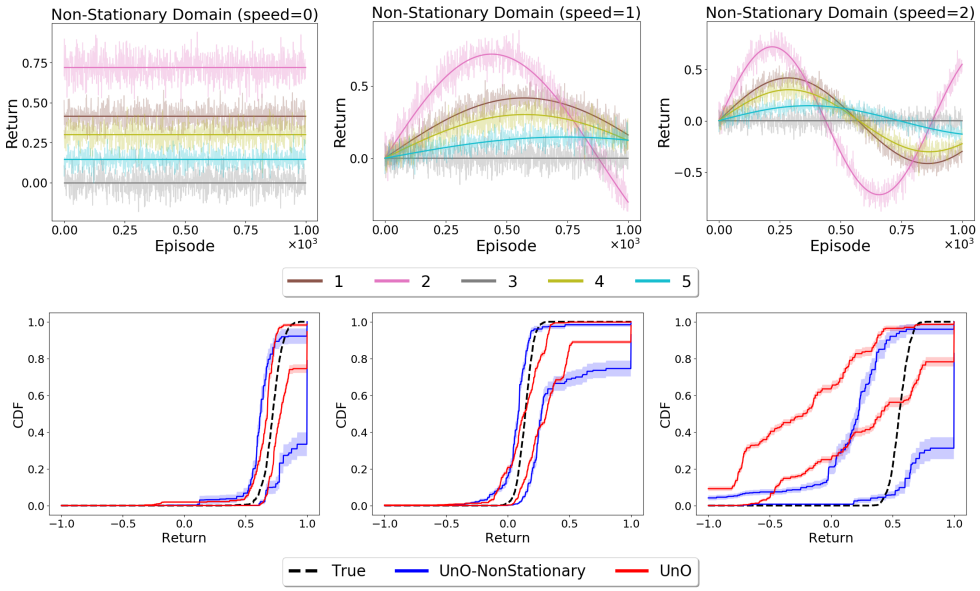

Figure 5: **(Top row)** True rewards (unknown to the RL agent) associated with each of the five items over the past 1000 episodes for different *speeds* of non-stationarity. Speed of 0 indicates stationary setting and higher speeds indicates greater degrees of non-stationarity. **(Bottom row)**. The black dashed line is the true value of the future distribution of returns under π: $F_\pi^{(L+\ell)}$, where $L = 1000$ and $\ell = 1$. In red is our UnO bound that does not account for non-stationarity, and in blue is the wild-bootstrap version of our UnO bound that accounts for non-stationarity. The shaded region corresponds to one standard error computed using 30 trials. Bounds were obtained for a failure rate $\delta = 0.05$. **(Left column)** In the stationary setting, both the variants of UnO bounds approximately contain the true future CDF $F_\pi^{(L+\ell)}$. In this setting, the UnO method designed only for stationary settings provides a tighter bound. **(Middle & Right columns)** As the domain becomes non-stationary, UnO bounds that do not account for non-stationarity fail to adequately bound the true future CDF $F_\pi^{(L+\ell)}$. When the degree of non-stationarity is high, not accounting for non-stationarity can lead to significantly inaccurate bounds. By comparison, UnO bounds that use wild bootstrap to tackle non-stationarity provide a more accurate bound throughout. As expected, when the fluctuations due to non-stationarity increase, the width of the confidence band increases as well. These results illustrate (a) the importance of accounting for non-stationarity, when applicable, and (b) the flexibility offered by our proposed universal off-policy estimator, UnO, to tackle such settings.

of Thomas et al. [91] that can provide tight bounds even when the domain is non-Markovian or has confounding (partial observability). Perhaps surprisingly, Figure 4 shows that the proposed guaranteed coverage bounds, termed *UnO-CI* here, can be competitive with this existing specialized bound, termed *Baseline-CI* here, for the mean. In fact, UnO-CI can often require an order of magnitude less data compared to the specialized bounds for variance [18]; we refer readers to Appendix F.2 for a discussion on potential reasons. This suggests that the universality of UnO can be beneficial even when only one specific parameter is of interest.

(C) Finite-sample confidence bounds for other parameters using UnO: Figure 4 demonstrates that UnO-CI also successfully addresses the open question of providing guaranteed coverage bounds for multiple parameters simultaneously without additional applications of the union bound. As expected, bounds for parameters like variance and CVaR that depend heavily on the distribution tails take more samples to shrink than bounds on other parameters (like the median [quantile(0.5)]). Additional discussion on the observed trends for the bounds is provided in Appendix F.2.

The proposed UnO-Boot bounds, as discussed in Section 3.1, are approximate and might not always hold with the specified probability. However, they stand out by providing *significantly* tighter, and thus more practicable, confidence intervals.

(D) Results for non-stationary settings: Results for this setting are presented in Figure 5. As discussed earlier, online recommendation systems for tutorials, movies, advertisements and other

products are ubiquitous. However, the popular assumption of stationarity is seldom applicable to these systems. In particular, personalizing for each user is challenging in such settings as interests of a user for different items among the recommendable products fluctuate over time. For an example, in the context of online shopping, interests of customers can vary based on seasonality or other unknown factors. To abstract such settings, in this domain the reward (interest of the user) associated with each item changes over time. See Figure 5 (top row) for visualization of the domain, for different "speeds" (degrees of non-stationarity).

In all the settings with different speeds, a uniformly random policy was used as a behavior policy β to collect data for 1000 episodes. To test the efficacy of UnO, when the future domain can be different from the past domains, the evaluation policy was chosen to be a near-optimal policy for the future episode: $1000 + 1$.

7 Conclusion

We have taken the first steps towards developing a *universal off-policy estimator* (UnO), closing the open question of whether it is possible to estimate and provide finite-sample bounds (that hold with high probability) for *any* parameter of the return distribution in the *off-policy* setting, with minimal assumptions on the domain. Now, without being restricted to the most common and basic parameters, researchers and practitioners can fully characterize the (potentially dangerous or costly) behavior of a policy without having to deploy it.

There are many new questions regarding how UnO can be improved for policy *evaluation* by further reducing data requirements or weakening assumptions. Using UnO for policy *improvement* also remains an interesting future direction. Subsequent to this work, Huang et al. [43] showed how models can be used to obtain UnO-style doubly robust estimators along with its convergence rates in the contextual bandit setting. This allows their method to also provide finite-sample uniform CDF bounds for a broad class of Lipschitz risk functionals.

8 Acknowledgements

We thank Shiv Shankar, Scott Jordan, Wes Cowley, and Nan Jiang for the feedback, corrections, and other contributions to this work. We would also like to thank Bo Liu, Akshay Krishnamurthy, Marc Bellemare, Ronald Parr, Josiah Hannah, Sergey Levine, Jared Yeager, and the anonymous reviewers for their feedback on this work.

Research reported in this paper was sponsored in part by a gift from Adobe, NSF award #2018372, and the DEVCOM Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 (ARL IoBT CRA). Research in this paper is also supported in part by NSF (IIS-1724157, IIS-1638107, IIS-1749204, IIS-1925082), ONR (N00014-18-2243), AFOSR (FA9550-20-1-0077), and ARO (78372-CS, W911NF-19-2-0333). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

[2] J. Altschuler, V.-E. Brunel, and A. Malek. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91):1–39, 2019.

[3] T. W. Anderson. Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function. Technical report, Stanford University, Department of Statistics, 1969.

[4] M. G. Azar, R. Munos, and H. J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

[5] A. G. Barto, P. S. Thomas, and R. S. Sutton. Some recent applications of reinforcement learning. In *Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems*, 2017.

[6] M. Bastani. Model-free intelligent diabetes management using machine learning. Master's thesis, University of Alberta, 2014.

[7] M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.

[8] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[9] A. Bennett, N. Kallus, L. Li, and A. Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. *arXiv preprint arXiv:2007.13893*, 2020.

[10] V. W. Berger and Y. Zhou. Kolmogorov–Smirnov test: Overview. *Wiley Statsref: Statistics Reference Online*, 2014.

[11] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[12] P. Bloomfield. *Fourier Analysis of Time Series: An Introduction*. John Wiley & Sons, 2004.

[13] D. B. Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

[14] D. S. Brown, S. Niekum, and M. Petrik. Bayesian robust optimization for imitation learning. *arXiv preprint arXiv:2007.12315*, 2020.

[15] F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4(421-424), 1933.

[16] Y. Chandak, S. M. Jordan, G. Theocharous, M. White, and P. S. Thomas. Towards safe policy improvement for non-stationary MDPs. *Neural Information Processing Systems*, 2020.

[17] Y. Chandak, G. Theocharous, S. Shankar, S. Mahadevan, M. White, and P. S. Thomas. Optimizing for the future in non-stationary MDPs. *International Conference on Machine Learning*, 2020.

[18] Y. Chandak, S. Shankar, and P. S. Thomas. High confidence off-policy (or counterfactual) variance estimation. *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

[19] S. X. Chen, W. Härdle, and M. Li. An empirical likelihood goodness-of-fit test for time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):663–678, 2003.

[20] V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[21] K.-J. Chung and M. J. Sobel. Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

[22] W. Dabney, G. Ostrovski, D. Silver, and R. Munos. Implicit quantile networks for distributional reinforcement learning. *arXiv preprint arXiv:1806.06923*, 2018.

[23] W. Dabney, M. Rowland, M. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[24] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.

[25] B. Dai, O. Nachum, Y. Chow, L. Li, C. Szepesvári, and D. Schuurmans. Coindice: Off-policy confidence interval estimation. *arXiv preprint arXiv:2010.11652*, 2020.

[26] R. Davidson and E. Flachaire. The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169, 2008.

[27] R. Dearden, N. Friedman, and S. Russell. Bayesian q-learning. In *AAAI/IAAI*, pages 761–768, 1998.

[28] J. DiNardo, N. M. Fortin, and T. Lemieux. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Technical report, National Bureau of Economic Research, 1995.

[29] S. G. Donald and Y.-C. Hsu. Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics*, 178:383–397, 2014.

[30] G. Dulac-Arnold, D. Mankowitz, and T. Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

[31] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[32] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

[33] F. Faccio, L. Kirsch, and J. Schmidhuber. Parameter-based value functions. *arXiv preprint arXiv:2006.09226*, 2020.

[34] R. Feldt and A. Stukalov. Blackboxoptim. jl, 2019.

[35] Y. Feng, Z. Tang, na zhang, and qiang liu. Non-asymptotic confidence intervals of off-policy evaluation: Primal and dual bounds. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dKg5D1Z1Lm.

[36] S. Firpo and C. Pinto. Identification and estimation of distributional impacts of interventions using changes in inequality measures. *Journal of Applied Econometrics*, 31(3):457–486, 2016.

[37] Y. Gao and Y. Cui. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nature Communications*, 11(1):1–8, 2020.

[38] V. Glivenko. Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, 4:92–99, 1933.

[39] Z. Guo, P. S. Thomas, and E. Brunskill. Using options and covariance testing for long horizon off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2492–2501, 2017.

[40] J. Hanna, P. Stone, and S. Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2017.

[41] J. Hanna, S. Niekum, and P. Stone. Importance sampling policy evaluation with an estimated behavior policy. In *International Conference on Machine Learning*, pages 2605–2613. PMLR, 2019.

[42] J. Harb, T. Schaul, D. Precup, and P.-L. Bacon. Policy evaluation networks. *arXiv preprint arXiv:2002.11833*, 2020.

[43] A. Huang, L. Leqi, Z. C. Lipton, and K. Azizzadenesheli. Off-policy risk assessment in contextual bandits. *arXiv preprint arXiv:2104.08977*, 2021.

[44] S. C. Jaquette. Markov decision processes with a new optimality criterion: Discrete time. *The Annals of Statistics*, pages 496–505, 1973.

[45] N. Jiang and J. Huang. Minimax confidence interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*, 2020.

[46] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[47] N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Mach. Learn. Res.*, 21:167–1, 2020.

[48] N. Kallus and A. Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *arXiv preprint arXiv:2002.04518*, 2020.

[49] R. Keramati, C. Dann, A. Tamkin, and E. Brunskill. Being optimistic to be conservative: Quickly learning a CVaR policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

[50] K. Khetarpal, M. Riemer, I. Rish, and D. Precup. Towards continual reinforcement learning: A review and perspectives. *arXiv preprint arXiv:2012.13490*, 2020.

[51] I. Kostrikov and O. Nachum. Statistical bootstrapping for uncertainty estimation in off-policy evaluation. *arXiv preprint arXiv:2007.13609*, 2020.

[52] S. R. Kuindersma, R. A. Grupen, and A. G. Barto. Variable risk control via stochastic optimization. *The International Journal of Robotics Research*, 32(7):806–825, 2013.

[53] I. Kuzborskij, C. Vernade, A. György, and C. Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. *arXiv preprint arXiv:2006.10460*, 2020.

[54] T. Lattimore and M. Hutter. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

[55] E. Learned-Miller and J. DeStefano. A probabilistic upper bound on differential entropy. *IEEE Transactions on Information Theory*, 54(11):5223–5230, 2008.

[56] P. Liao, P. Klasnja, and S. Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, pages 1–10, 2020.

[57] Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

[58] E. Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285, 1993.

[59] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology*, 8(1): 26–34, 2014.

[60] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[61] A. M. Metelli, M. Papini, N. Montali, and M. Restelli. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 21(141):1–75, 2020.

[62] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *ICML*, 2010.

[63] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

[64] O. Nachum and B. Dai. Reinforcement learning via Fenchel-Rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

[65] H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *arXiv preprint arXiv:2003.05623*, 2020.

[66] S. Padakandla. A survey of reinforcement learning algorithms for dynamically varying environments. *arXiv preprint arXiv:2005.10619*, 2020.

[67] M. Papini, A. M. Metelli, L. Lupo, and M. Restelli. Optimistic policy optimization via multiple importance sampling. In *International Conference on Machine Learning*, pages 4989–4999. PMLR, 2019.

[68] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[69] B. Pavse, I. Durugkar, J. P. Hanna, and P. Stone. Reducing sampling error in batch temporal difference learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, July 2020.

[70] J. Pearl. *Causality*. Cambridge university press, 2009.

[71] D. Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[72] M. L. Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

[73] J. P. Romano and C. DiCiccio. *Multiple data splitting for testing*. Department of Statistics, Stanford University, 2019.

[74] J. P. Romano and M. Wolf. Explicit nonparametric confidence intervals for the variance with guaranteed coverage. *Communications in Statistics-Theory and Methods*, 31(8):1231–1250, 2002.

[75] M. Rowland, M. G. Bellemare, W. Dabney, R. Munos, and Y. W. Teh. An analysis of categorical distributional reinforcement learning. *arXiv preprint arXiv:1802.08163*, 2018.

[76] M. Rowland, R. Dadashi, S. Kumar, R. Munos, M. G. Bellemare, and W. Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR, 2019.

[77] T. Schaul, D. Horgan, K. Gregor, and D. Silver. Universal value function approximators. In *International Conference on Machine Learning*, pages 1312–1320. PMLR, 2015.

[78] P. K. Sen and J. M. Singer. *Large Sample Methods in Statistics An Introduction With Applications*. Chapman & Hall, 1993.

[79] A. M. Shaikh. The Glivenko-Cantelli Theorem. http://home.uchicago.edu/~amshaikh/webfiles/glivenko-cantelli.pdf. Accessed: 2010-09-30.

[80] S. P. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable Markovian decision processes. In *Machine Learning Proceedings 1994*, pages 284–292. Elsevier, 1994.

[81] M. J. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

[82] D. A. Stephens. The Glivenko-Cantelli Lemma. http://wwwf.imperial.ac.uk/~das01/MyWeb/M3S3/Handouts/GlivenkoCantelli.pdf, 2006.

[83] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

[84] A. Tamar, D. Di Castro, and S. Mannor. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research*, 17(1):361–396, 2016.

[85] G. Tennenholtz, U. Shalit, and S. Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

[86] G. Theocharous, P. S. Thomas, and M. Ghavamzadeh. Ad recommendation systems for life-time value optimization. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1305–1310, 2015.

[87] G. Theocharous, P. S. Thomas, and M. Ghavamzadeh. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[88] G. Theocharous, Y. Chandak, P. S. Thomas, and F. de Nijs. Reinforcement learning for strategic recommendations. *arXiv preprint arXiv:2009.07346*, 2020.

[89] P. Thomas and E. Brunskill. Importance sampling with unequal support. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[90] P. Thomas and E. Learned-Miller. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, pages 6225–6233, 2019.

[91] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[92] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388, 2015.

[93] P. S. Thomas, G. Theocharous, M. Ghavamzadeh, I. Durugkar, and E. Brunskill. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *AAAI*, pages 4740–4745, 2017.

[94] P. S. Thomas, B. C. da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

[95] M. Uehara, J. Huang, and N. Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

[96] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[97] G. Van Rossum and F. L. Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[98] E. Veach and L. J. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pages 419–428, 1995.

[99] L. Wang, Y. Zhou, R. Song, and B. Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018.

[100] T. Wang, M. Bowling, and D. Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 44–51. IEEE, 2007.

[101] D. White. Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *Journal of Optimization Theory and Applications*, 56(1):1–29, 1988.

[102] A. Xie, J. Harrison, and C. Finn. Deep reinforcement learning amidst lifelong non-stationarity. *arXiv preprint arXiv:2006.10701*, 2020.

[103] J. Xie. *Simglucose v0.2.1 (2018)*, 2019. URL https://github.com/jxx123/simglucose.

[104] T. Xie, Y. Ma, and Y.-X. Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *arXiv preprint arXiv:1906.03393*, 2019.

[105] M. Yang, B. Dai, O. Nachum, G. Tucker, and D. Schuurmans. Offline policy selection under uncertainty. *arXiv preprint arXiv:2012.06919*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix B
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumptions 1, 2, and 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix D.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] code for the proposed UnO method(s) and the domains used for empirical studies are available https://github.com/yashchandak/UnO.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section F.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All plots have standard error bars computed using 30 trials.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] All the experiments were conducted on a personal computer with 32 GiB of memory and an Intel Core i7 CPU with 12 threads. Total runtime for all the experiments combined was less than a day.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] Our code uses Julia [11], Blackboxoptim library [34], Python [97], and PyTorch [68].

(b) Did you mention the license of the assets? [N/A] Only open-source assets have been used.

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Notation

Symbol	Meaning
$\mathcal{D}$	Data set of the observed trajectories
n	Total number of observed trajectories in $\mathcal{D}$
π	Evaluation policy
β_i	Behavior policy for the i^{th} trajectory
ρ_i	Importance ratio for the observed trajectory H_i
$\mathcal{S}$	State set
$\mathcal{O}, \widetilde{\mathcal{O}}$	Observation set for the behavior policy and the evaluation policy, respectively
$\mathcal{A}$	Action set
$\mathcal{P}$	Transition dynamics, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
$\mathcal{R}$	Reward function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$
Ω	Observation function for behavior policy, $\Omega : \mathcal{S} \to \Delta(\mathcal{O})$
Ω_2	Observation function for the evaluation policy, $\Omega_2 : \mathcal{S} \times \mathcal{O} \to \Delta(\widetilde{\mathcal{O}})$
γ	Discounting factor
d_0	Starting state distribution
T	Finite horizon length
H_i, H_π	i^{th} observed trajectory in the dataset and complete trajectory under policy π, respectively
G_i, G_π	Return observed in the i^{th} trajectory in the dataset and return under any policy π, respectively
$G_{\min}, G_{\max}$	Minimum and maximum value of a return, respectively
$F_\pi, \mathrm{d}F_\pi$	True CDF of returns under policy π and its associated probability distribution, respectively
$\hat{F}_n$, $\bar{F}_n$	Off-policy CDF estimator and weighted off-policy CDF estimator using n samples, respectively
F_-, F_+	Lower and upper bound on the CDF
$\mathcal{F}$	The set of all CDFs between the upper bound and the lower bounds
κ_i, K	i^{th} key point and total number of key points, respectively
α	Value for defining inverse CDF-based statistics
ψ	Generic functional for a distributional parameter/statistic
ψ_-, ψ_+	Lower and upper bounds for $\psi(F_\pi)$
δ	Failure rate for the bounds
$\mathcal{D}_{\text{eval}}, \mathcal{D}_{\text{train}}$	Evaluation and training split of the dataset $\mathcal{D}$
$\texttt{CI}_-, \texttt{CI}_+$	Lower and upper confidence bounds for a given random variable
θ	Parameters that are used to construct $\mathcal{F}$
$\mathscr{A}$	Euclidean area enclosed within $\mathcal{F}$
X_i^*	i^{th} bootstrap resampled value for any random variable X
ε, ϵ	Some small value in Assumption 1 and Assumption 2, respectively
w_ν, ϕ	Regression weights and basis function for the assumption on smooth non-stationarity
L, ℓ	Number of past and future episodes being considered in the smooth non-stationary setting

Table 1: List of symbols used in the main paper and their associated meanings.

B Broader Impact

While our estimators and bounds are both theoretically sound and intuitively simple, it is important for a broader audience to understand the limitations of our method, assumptions being made, and what can be done when these assumptions do not hold. Understanding these assumptions can also help in mitigating any undesired biases in applications built around UnO and can thus avoid any potential negative societal impacts. In the following, we briefly allude to possible alternatives when the required assumptions are violated.

B.1 Discussion of Assumptions and Requirements of UnO

Knowledge of Subset Support: Through Assumption 1, UnO requires that all the behavior policies $(\beta_i)_{i=1}^n$ have sufficient support for actions that have non-zero probability under π. Particularly, it requires that the $\beta(a|o)$ is bounded below by (an unknown) ϵ when $\pi(a|o) > 0$. This ensures

that importance ratios are bounded and thus simplifies analysis for UnO's consistency results and constructing confidence intervals. This assumption is common both in the off-policy literature [47, 104, 105] and in real applications [87]

The above assumption is also equivalent to assuming bounded exponentiated-Renyi-divergence (for $\alpha = \infty$) between the probability distributions of trajectories under the behavior and the evaluation policies [61]. As the UnO's bound for the CDF uses CIs for the mean as a sub-routine, the above assumption can be relaxed by using CIs for the mean that depend on Renyi-divergence for other values of α [61]. Similarly, consistency results for UnO rely upon finite variance, which can also be achieved by instead assuming that the Renyi-divergence is bounded for $\alpha = 2$.

Alternatively, Assumption 1 can be relaxed to only absolute continuity by using methods that provide valid CIs for the mean by clipping the importance weights. (See the work by Thomas et al. [91, Theorem 1] for removal of the upper bound on the importance weights when lower-bounding the mean, and the work by Chandak et al. [18, Theorem 5] for removal of the upper bound on the importance weights when upper-bounding the mean). Furthermore, prior work has also shown how even the assumption of absolute continuity can in some cases be removed (See discussion around Eqn 8 in the appendix of the work by Thomas et al. [91]). If the supports for the behavior and the evaluation policies are unequal, Thomas and Brunskill [89] also present a technique to reduce variance resulting from IS.

Further, WIS might also be helpful in relaxing the assumptions on the IS ratios. Specifically, WIS-based mean bounds [53] can also be used along with the WIS-based UnO estimator (27) to get a valid confidence band for the entire CDF.

Using multi-importance sampling (MIS), the subset support requirement for *all* $(\beta_i)_{i=1}^n$ can be relaxed to the requirement that the *union of supports* under the behavior policies $(\beta_i)_{i=1}^n$ has sufficient support [98, 67, 61]. MIS can also help in substantially reducing variance. However, this relaxation requires an alternate assumption that a complete knowledge of all the behavior policies $(\beta_i)_{i=1}^n$, not just the probabilities of the action executed using them, is available.

Knowledge of Action Probabilities under Behavior Policies $(\beta_i)_{i=1}^n$: UnO requires access to the probability $\beta(a|o)$ (only the scalar probability value and not the entire policy β) of the actions available in the data set, $\mathcal{D}$, to compute the importance sampling ratios in (3). Access to the probability $\beta(a|o)$ is often available when $\mathcal{D}$ is collected using an automated policy; however, it might not be available in some cases, such as when decisions were previously made by humans.

When the probability $\beta(a|o)$ is not available, one natural alternative is to estimate it from the data and use this estimate of $\beta(a|o)$ in the denominator of the importance ratios. This technique is also known as regression importance sampling (RIS) and is known to provide biased but consistent estimates for the mean [41, 69] in the Markov decision process setting (MDP) setting. For UnO, $\hat{F}_n(\nu)$ is analogous to mean estimation of $X \coloneqq \rho\big(\mathbb{1}_{\{G \leq \nu\}}\big)$, for any ν. Therefore, the findings of RIS can be directly extended to UnO in the MDP setting, where $\widetilde{O} = O = S$. In the following, we provide a high-level discussion for the setting when $\beta(a|o)$ is *not* available and the states are partially observed,

- **Partial observability with $\widetilde{O} = O$:** In this setting, as $\beta(a|o) = \beta(a|\tilde{o})$, one can use density estimation on the available data, $\mathcal{D}$, to construct an estimator $\hat{\beta}(a|o)$ of $\Pr(a|\tilde{o}) = \beta(a|\tilde{o})$ and use RIS to get a biased but consistent estimator for F_π. Here, bias results from the estimation error in $\hat{\beta}(a|o)$ but consistency follows as the true $\beta(a|o)$ can be recovered in the limit when $n \to \infty$.

  In context of UnO, using $\hat{\beta}(a|o)$ instead of $\beta(a|o)$ violates the unbiased condition for $\hat{F}_n$, which was necessary to obtain the `CIs` and construct $\mathcal{F}$. Therefore, high-confidence bounds with guaranteed coverage cannot be obtained using UnO in this setting. However, point estimates and approximate bootstrap bounds can still be obtained.

- **Partial observability with $\widetilde{O} \neq O$:** In this setting, using RIS will produce neither an unbiased nor a consistent estimator for F_π. As $\mathcal{D}$ only has $\tilde{o}$ and not o, at best it is only possible to estimate $\Pr(a|\tilde{o}) = \sum_{x \in \mathcal{O}} \beta(a|x) \Pr(x|\tilde{o})$ through density estimation using data $\mathcal{D}$. However, in general, since $\beta(a|o) = \Pr(a|o) \neq \Pr(a|\tilde{o})$ we cannot even consistently estimate the denominator for importance sampling unless some other stronger assumptions are made. See work by Namkoong et al. [65], Tennenholtz et al. [85], Bennett et al. [9] and Kallus and Zhou [48] for possible alternative assumptions and approaches to tackle this setting.

Knowledge of $G_{\min}, G_{\max}$: To construct the CDF band $\mathcal{F}$, UnO requires knowledge of $G_{\min}$ and $G_{\max}$ in (5). Notice from Figure 2 that knowing $G_{\max}$ helps in clipping the *lower bound* for the *upper tail* (LBUT) of $\mathcal{F}$, which otherwise would have extended to $+\infty$. Similarly, knowing $G_{\min}$ helps in clipping the *upper bound* for the *lower tail* (UBLT) of $\mathcal{F}$, which otherwise would have extended to $-\infty$.

Typically, even if $G_{\min}$ or $G_{\max}$ is not known, they can be obtained as $R_{\min}/(1-\gamma)$ or $R_{\max}/(1-\gamma)$, respectively, where $R_{\min}$ and $R_{\max}$ are known finite lower and upper bounds for any individual reward. Otherwise, knowledge of $G_{\min}$ or $G_{\max}$ can be relaxed if the desired bound on ψ does not depend on UBLT or LBUT, respectively. For example, observe from Figure 3 that (a) The lower bound for the mean or quantile does not depend on LBUT. Analogously, if only an upper bound for the mean or quantile is required, then UBLT is not needed. (b) The lower bound on CVaR depends on UBLT, however, (for small values of α) the upper bound on CVaR neither depends on LBUT nor UBLT. (c) For an upper bound on variance, both LBUT and UBLT are required. However, for the variance's lower bound, neither LBUT nor UBLT are required. See Figure 6 for intuition.

Knowledge of Function Class ϕ: For the smoothly non-stationary setting, through Assumption 3, UnO requires access to the basis functions ϕ that can be used with least-squares regression to analyze the trend in the distributions of returns $(F_\pi^{(i)}(\nu))_{i=1}^L$ for any $\nu \in \mathbb{R}$. In practice, one can use sufficiently flexible basis functions to model time-series trends (e.g., Fourier basis [12]). To avoid overfitting or underfitting, one could also use goodness-of-fit tests to select the functional class ϕ for the trend [19].

Knowledge of Bound ϵ on the Distribution Shift: Unlike the smoothly non-stationary setting, if the underlying shift can be discrete and arbitrary, prior data may not contain any useful information towards characterizing the shift. Therefore, avoiding domain knowledge may be inevitable when setting the value for ϵ unless some other stronger assumptions are made.

C Extended Discussion on Related Work

In the on-policy RL literature, parameters other than the mean have also been explored [44, 81, 21, 101, 27, 54, 4], and recent distributional RL methods extend this direction by estimating the entire distribution of returns [62, 63, 7, 22, 23, 24, 75]. Our work builds upon many of these ideas and extends them to the off-policy setting.

In the off-policy RL setup, there is a large body of literature that tackles the off-policy mean estimation problem [71, 83]. Some works also aim at providing high-confidence off-policy mean estimation using concentration inequalities [91, 53] or bootstrapping [92, 40, 51]. Several recent approaches build upon a dual perspective for dynamic programming [72, 100, 64] for both estimating and bounding the mean [57, 104, 45, 95, 25, 35]. However, these methods are restricted to domains with Markovian dynamics and full observability. Some works have also focused on estimating the mean return in the setting where states are partially observed [65, 85, 48] or when there is non-stationarity [16, 17, 50, 66]. Recent work by Chandak et al. [18] also looks at (high-confidence) off-policy variance estimation. Our work extends these research directions by tackling these settings simultaneously, while also providing a general procedure to estimate and obtain high-confidence bounds for *any* parameter of the distribution of returns. Particularly, UnO is a single, unified, and universal procedure that can be used to mitigate the complexity associated with estimating different parameters for different domain settings.

A popular RL method that has similar name to UnO is the *Universal value function approximator* (UVFA) by Schaul et al. [77]. However, UVFA is fundamentally different from UnO: UVFA estimates *expected* return $\mathbb{E}[G_\pi]$ from a state given any desired goal. By comparison, UnO estimates any parameter of the return G_π for a single "goal". Recent work by Harb et al. [42] and Faccio et al. [33] propose using supervised learning to estimate parametric models that can map a *representation* of a policy π to the corresponding distribution of G_π. By training over a given distribution of policies, new policies in the test set can be evaluated without using new data. By comparison, UnO does not requires any parametric assumptions or any train-test distribution. Further, UnO also provides high-confidence bounds for all the parameters of the return distribution.

D Proofs for Theoretical Results

The main results in this paper are for the setting where both the evaluation and the behavior policies have the same observation set. In the following, we present generalized results where the available observations, $\widetilde{O}$, for the evaluation policy can be different from the behavior policy's observations, O. Further, for notational ease, in the main paper we had focused only on finite sets. In the following, we present a more general setting where states, actions, observations, and rewards are all continuous. Let $\Omega_2 : \mathcal{S} \times \mathcal{O} \to \Delta(\widetilde{\mathcal{O}})$ be the distribution over $\widetilde{\mathcal{O}}$, conditioned on state $s \in \mathcal{S}$ and observation $o \in \mathcal{O}$, which determines how the observations $\widetilde{O}$ are generated.

Let $\mathcal{D} = (H_i)_{i=1}^n$ be the available observed trajectories, where each H contains $(\widetilde{O}_0, A_0, \beta(A_0|O_0), R_0, \widetilde{O}_1, ...)$. Note that when the random variables $\widetilde{O} = O = S$, we recover a standard fully observable MDP setting. By comparison, H_π is the random variable corresponding to the complete trajectory $(S_0, O_0, \widetilde{O}_0, A_0, R_0, S_1, O_1, \widetilde{O}_1, ...)$ under any policy π. Of course, H_π is unknown. To make the dependence between a trajectory $h \in \mathscr{H}_\pi$ and its associated return G and importance ratios ρ explicit, we use the shorthand $g(h)$ and $\rho(h)$ to denote the return and importance ratios for the full trajectory h, respectively. To tackle this generalized setting, we also generalize the support assumption introduced earlier,

Assumption 1. *The set $\mathcal{D}$ contains independent (not necessarily identically distributed) observed trajectories generated using $(\beta_i)_{i=1}^n$, such that for some (unknown) $\varepsilon > 0$, $(\beta(a|o) < \varepsilon) \implies (\pi(a|\tilde{o}) = 0)$, for all $s \in \mathcal{S}, o \in$ supp$(\Omega(s)), \tilde{o} \in$ supp$(\Omega_2(s, o)), a \in \mathcal{A}$, and $i \in \{1, \dots, n\}$.*

Theorem 1. *Under Assumption 1, $\hat{F}_n$ is an unbiased and uniformly consistent estimator of F_π. That is,*

$$\forall \nu \in \mathbb{R}, \quad \mathbb{E}_{\mathcal{D}}\left[\hat{F}_n(\nu)\right] = F_\pi(\nu), \qquad \sup_{\nu \in \mathbb{R}} \left|\hat{F}_n(\nu) - F_\pi(\nu)\right| \xrightarrow{a.s.} 0.$$

Proof. This theorem has two results: unbiasedness and consistency of $\hat{F}_n$. Therefore, we break the proof into two parts.

Part 1 (Unbiasedness). We begin by expanding F_π for any $\nu \in \mathbb{R}$ using the definition of the CDF.

$$F_\pi(\nu) = \Pr(G_\pi \leq \nu) = \int_{-\infty}^{\nu} p(G_\pi = x)\mathrm{d}x = \int_{-\infty}^{\nu} \left(\int_{\mathscr{H}_\pi} p(H_\pi = h) \mathbb{1}_{\{g(h)=x\}} \mathrm{d}h \right) \mathrm{d}x, \quad (8)$$

where we used the fact that the probability density of the return G_π being x is the integral of the probability densities of the trajectories h whose return equals x. Therefore, as the integrands in (8) are finite and non-negative measurable functions, using Tonelli's theorem for interchanging the integrals, (8) can be expressed as,

$$F_\pi(\nu) = \int_{\mathscr{H}_\pi} p(H_\pi = h) \left(\int_{-\infty}^{\nu} \mathbb{1}_{\{g(h)=x\}} \mathrm{d}x \right) \mathrm{d}h = \int_{\mathscr{H}_\pi} p(H_\pi = h) \left(\mathbb{1}_{\{g(h) \leq \nu\}} \right) \mathrm{d}h, \quad (9)$$

where the last term follows because the output of $g(h)$ is a deterministic scalar given h and thus the indicator function can be one for at most a single value less than ν, and where the red color is used to highlight changes. Next, using Assumption 1 to change the support of the distribution in (9) and using importance weights we obtain,

$$F_\pi(\nu) = \int_{\mathscr{H}_\beta} p(H_\pi = h) \left(\mathbb{1}_{\{g(h) \leq \nu\}} \right) \mathrm{d}h = \int_{\mathscr{H}_\beta} p(H_\beta = h) \frac{p(H_\pi = h)}{p(H_\beta = h)} \left(\mathbb{1}_{\{g(h) \leq \nu\}} \right) \mathrm{d}h. \quad (10)$$

To simplify (10), we recursively use the fact that $p(X, Y) = p(X)p(Y|X)$ and note that under a given policy π the probability density of a trajectory with partial observations and non-Markovian structure is

$$\begin{aligned} p(H_\pi = h) =& p(s_0)p(o_0|s_0)p(\tilde{o}_0|o_0, s_0)p(a_0|s_0, o_0, \tilde{o}_0; \pi) \\ & \times \prod_{i=0}^{T-1} \Bigg(p(r_i|h_i)p(s_{i+1}|h_i)p(o_{i+1}|s_{i+1}, h_i)p(\tilde{o}_{i+1}|s_{i+1}, o_{i+1}, h_i) \\ & \quad \times p(a_{i+1}|s_{i+1}, o_{i+1}, \tilde{o}_{i+1}, h_i; \pi) \Bigg) p(r_T|h_T), \end{aligned} \quad (11)$$

where conditioning on π emphasizes that each action is sampled using π, and h_i represents the trajectory of all the states, partial observations, and actions up to time step i. Therefore, using (11), the ratio between $p(H_\pi = h)$ and $p(H_\beta = h)$ can be written as,

$$\begin{aligned}\frac{p(H_\pi = h)}{p(H_\beta = h)} &= \frac{p(a_0|s_0, o_0, \tilde{o}_0; \pi)}{p(a_0|s_0, o_0, \tilde{o}_0; \beta)} \prod_{i=0}^{T-1} \frac{p(a_{i+1}|s_{i+1}, o_{i+1}, \tilde{o}_{i+1}, h_i; \pi)}{p(a_{i+1}|s_{i+1}, o_{i+1}, \tilde{o}_{i+1}, h_i; \beta)} \\ &= \prod_{i=0}^{T} \frac{\pi(a_i|\widetilde{o}_i)}{\beta(a_i|o_i)} \\ &= \rho(h). \end{aligned} \tag{12}$$

Combining (10) and (12),

$$F_\pi(\nu) = \int_{\mathscr{H}_\beta} p(H_\beta = h)\rho(h)\Big(\mathbb{1}_{\{g(h)\leq\nu\}}\Big)\mathrm{d}h. \tag{13}$$

Finally, it can be shown that our proposed estimator $\hat{F}_n$ is an unbiased estimator of F_π by taking the expected value of $\hat{F}_n$,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\Big[\hat{F}_n(\nu)\Big] &= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\rho_i\Big(\mathbb{1}_{\{G_i\leq\nu\}}\Big)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}}\left[\rho_i\Big(\mathbb{1}_{\{G_i\leq\nu\}}\Big)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathscr{H}_{\beta_i}} p(H_{\beta_i} = h)\rho(h)\Big(\mathbb{1}_{\{g(h)\leq\nu\}}\Big)\mathrm{d}h \\ &\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n}F_\pi(\nu) \\ &= F_\pi(\nu), \end{aligned} \tag{14}$$

where (a) follows from (13), which holds for any behavior policy β that satisfies Assumption 1.

Note: H_π or H_β were invoked only for the purposes of the proof. Notice that the proposed estimator, $\hat{F}_n(\nu) = \frac{1}{n}\sum_{i=1}^{n}\rho_i\big(\mathbb{1}_{\{G_i\leq\nu\}}\big)$, only depends on the quantities available in the observed trajectory $(H_i)_{i=1}^n$ from $\mathcal{D}$.

Part 2 (Uniform Consistency). For this part, we will first show pointwise consistency, i.e., for any ν, $\hat{F}_n(\nu) \xrightarrow{\text{a.s.}} F_\pi(\nu)$, and then we will use this to establish *uniform* consistency, as required. To do so, let

$$X_i \coloneqq \rho_i\Big(\mathbb{1}_{\{G_i\leq\nu\}}\Big).$$

From Assumption 1, we know that trajectories are independent and that $\beta(a|o) \geq \varepsilon$ when $\pi(a|\tilde{o}) > 0$. This implies that the denominator in the IS ratio is bounded below when $\pi(a|\tilde{o}) \neq 0$, and hence the X_i's are bounded above and have a finite variance. Further, as established in (14), the expected value of X_i for all i equals $F_\pi(\nu)$. Therefore, using Kolmogorov's strong law of large numbers [78, Theorem 2.3.10 with Proposition 2.3.10],

$$\hat{F}_n(\nu) = \frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{\text{a.s.}} \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = F_\pi(\nu). \tag{15}$$

In the following, to obtain uniform consistency, we follow the proof for the Glivenko-Cantelli theorem [38, 15, 79, 82] using the pointwise consistency of the off-policy CDF estimator $\hat{F}_n$ established in (15). The proof relies upon the construction of K key points such that the difference in F_π at successive key points is bounded by a small ϵ_1. However, this would not be possible directly as there

can be discontinuties/jumps in F_π that are greater than ϵ_1. To tackle such discontinuties, we introduce some extra notation, Formally, let, $\forall \nu \in \mathbb{R}$,

$$F_\pi(\nu^-) \coloneqq \Pr(G_\pi < \nu) = F_\pi(\nu) - \Pr(G_\pi = \nu), \qquad \hat{F}_n(\nu^-) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \rho_i \Big(\mathbb{1}_{\{G_i < \nu\}} \Big). \quad (16)$$

Then, using arguments analogous to the ones used for (15), it can be observed that

$$\hat{F}_n(\nu^-) \xrightarrow{\text{a.s}} F_\pi(\nu^-). \quad (17)$$

Let $\epsilon_1 > 0$, and let K be any value more than $1/\epsilon_1$. Let $(\kappa_i)_{i=0}^K$ be K key points,

$$G_{\min} = \kappa_0 < \kappa_1 \leq \kappa_2 .... \leq \kappa_{K-1} < \kappa_K = G_{\max},$$

which create K intervals such that for all $i \in (1, ..., K-1)$,

$$F_\pi(\kappa_i^-) \leq \frac{i}{K} \leq F_\pi(\kappa_i).$$

Then by construction, if $\kappa_{i-1} < \kappa_i$,

$$F_\pi(\kappa_i^-) - F_\pi(\kappa_{i-1}) \leq \frac{i}{K} - \frac{i-1}{K} = \frac{1}{K} < \epsilon_1. \quad (18)$$

Intuitively, as F_π is monotonically non-decreasing, (18) restricts the intermediate values for any $F_\pi(\nu)$, to be within an ϵ_1 distance of the CDF values at its nearby key points. Notice the role of κ_i^- here: it would not have been possible to bound difference between $F_\pi(\kappa_i)$ and $F_\pi(\kappa_{i-1})$ by ϵ_1 as there could have been 'jumps' of value greater than ϵ_1 in F_π. However, κ^- and κ can be used to consider key points right before and after any jump in F_π, which ensures that we can always construct sequence of key points such that $F_\pi(\kappa_i^-) - F_\pi(\kappa_{i-1})$ is instead bounded by ϵ_1.

For the CDF estimates at the key points, let,

$$\Delta_n \coloneqq \max_{i \in (1...K-1)} \left\{ \left| \hat{F}_n(\kappa_i) - F_\pi(\kappa_i) \right|, \left| \hat{F}_n(\kappa_i^-) - F_\pi(\kappa_i^-) \right| \right\}. \quad (19)$$

From (15) and (17), as $\hat{F}_n(\nu)$ and $\hat{F}_n(\nu^-)$ are consistent estimators of $F_\pi(\nu)$ and $F_\pi(\nu^-)$, respectively, and since the maximum is over a finite set in (19), it follows that as $n \to \infty$,

$$\Delta_n \xrightarrow{\text{a.s.}} 0. \quad (20)$$

For any ν, let κ_{i-1} and κ_i be such that $\kappa_{i-1} \leq \nu < \kappa_i$. Then,

$$\begin{aligned} \hat{F}_n(\nu) - F_\pi(\nu) &\leq \hat{F}_n(\kappa_i^-) - F_\pi(\kappa_{i-1}) \\ &\leq \hat{F}_n(\kappa_i^-) - F_\pi(\kappa_i^-) + \epsilon_1, \end{aligned} \quad (21)$$

where the last step follows using (18). Similarly,

$$\begin{aligned} \hat{F}_n(\nu) - F_\pi(\nu) &\geq \hat{F}_n(\kappa_{i-1}) - F_\pi(\kappa_i^-) \\ &\geq \hat{F}_n(\kappa_{i-1}) - F_\pi(\kappa_{i-1}) - \epsilon_1. \end{aligned} \quad (22)$$

Then, using (21) and (22), $\forall \nu \in \mathbb{R}$,

$$\hat{F}_n(\kappa_{i-1}) - F_\pi(\kappa_{i-1}) - \epsilon_1 \leq \hat{F}_n(\nu) - F_\pi(\nu) \leq \hat{F}_n(\kappa_i^-) - F_\pi(\kappa_i^-) + \epsilon_1, \quad (23)$$

and thus using (19) and (23),

$$\left| \hat{F}_n(\nu) - F_\pi(\nu) \right| \leq \Delta_n + \epsilon_1. \quad (24)$$

Using (20), we obtain the following property of the upper bound in (24):

$$\Delta_n + \epsilon_1 \xrightarrow{\text{a.s}} \epsilon_1. \quad (25)$$

Finally, since (24) holds for $\forall \nu \in \mathbb{R}$ and (25) is valid for any $\epsilon_1 > 0$, making $\epsilon_1 \to 0$ gives the desired result,

$$\sup_{\nu \in \mathbb{R}} \left| \hat{F}_n(\nu) - F_\pi(\nu) \right| \xrightarrow{\text{a.s.}} 0. \quad (26)$$

□

Variance-reduced estimation: It is known that importance-sampling-based estimators are subject to high variance, which can often be limiting in practice [39]. A popular approach to mitigate variance is to use *weighted* importance sampling (WIS), which trades off variance for bias. Leveraging this approach, we propose the following variance-reduced estimator, $\bar{F}_n$, of F_π,

$$\forall \nu \in \mathbb{R}, \quad \bar{F}_n(\nu) \coloneqq \frac{1}{\sum_{j=1}^{n} \rho_j} \left(\sum_{i=1}^{n} \rho_i \left(\mathbb{1}_{\{G_i \leq \nu\}} \right) \right). \tag{27}$$

In the following theorem, we show that $\bar{F}_n$ is a biased estimator of F_π, though it preserves consistency.

Property 1. *Under Assumption 1, $\bar{F}_n$ may be biased but is a uniformly consistent estimator of F_π,*

$$\forall \nu \in \mathbb{R}, \quad \mathbb{E}_{\mathcal{D}}\left[\bar{F}_n(\nu)\right] \neq F_\pi, \qquad \sup_{\nu \in \mathbb{R}} \left| \bar{F}_n(\nu) - F_\pi(\nu) \right| \xrightarrow{a.s.} 0.$$

Proof. Similar to the proof for Theorem 1, we break this proof in two parts, one to establish bias and the other to establish consistency of $\hat{F}_n$.

Part 1 (Biased): We prove this using a counter-example. Let $n = 1$ and $\pi \neq \beta_1$, so

$$\begin{aligned}
\forall \nu \in \mathbb{R}, \quad \mathbb{E}_{\mathcal{D}}\left[\bar{F}_n(\nu)\right] &= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{\sum_{j=1}^{1} \rho_j} \left(\sum_{i=1}^{1} \rho_i \mathbb{1}_{\{G_i \leq \nu\}} \right) \right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}_{\{G_1 \leq \nu\}} \right] \\
&\overset{(a)}{=} \int_{\mathscr{H}_{\beta_1}} p(H_{\beta_1} = h) \left(\mathbb{1}_{\{g(h) \leq \nu\}} \right) \mathrm{d}h \\
&= F_{\beta_1}(\nu) \\
&\neq F_\pi(\nu),
\end{aligned}$$

where (a) follows analogously to (9).

Part 2 (Uniform Consistency): First, we will establish pointwise consistency, i.e., for any ν, $\bar{F}_n(\nu) \xrightarrow{\text{a.s.}} F_\pi(\nu)$, and then we will use this to establish *uniform* consistency, as required.

$$\begin{aligned}
\forall \nu \in \mathbb{R}, \quad \bar{F}_n(\nu) &= \frac{1}{\sum_{j=1}^{1} \rho_j} \left(\sum_{i=1}^{1} \rho_i \mathbb{1}_{\{G_i \leq \nu\}} \right) \\
&= \left(\frac{1}{n} \sum_{j=1}^{n} \rho_j \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \rho_i \mathbb{1}_{\{G_i \leq \nu\}} \right).
\end{aligned}$$

Let $X_n \coloneqq \frac{1}{n} \sum_{j=1}^{n} \rho_j$ and $Y_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} \rho_i \mathbb{1}_{\{G_i \leq \nu\}}$. Now, as $\bar{F}_n(\nu)$ is a continuous function of both X_n and Y_n, if both $(\lim_{n\to\infty} X_n)^{-1}$ and $(\lim_{n\to\infty} Y_n)$ exist then using the continuous mapping theorem [96, Theorem 2.3],

$$\forall \nu \in \mathbb{R}, \quad \lim_{n\to\infty} \bar{F}_n(\nu) = \left(\lim_{n\to\infty} X_n \right)^{-1} \left(\lim_{n\to\infty} Y_n \right). \tag{28}$$

Notice using Kolmogorov's strong law of large numbers [78, Theorem 2.3.10 with Proposition 2.3.10] that the term in the first parentheses will almost surely converge to the expected value of importance ratios, which equals one [71]. Similarly, we know from (15) that the term in the second parentheses will converge to $F_\pi(\nu)$ almost surely. Therefore, both parenthetical terms of (28) exist, and thus

$$\forall \nu \in \mathbb{R}, \quad \bar{F}_n(\nu) \xrightarrow{\text{a.s.}} (1)^{-1}(F_\pi(\nu)) = F_\pi(\nu). \tag{29}$$

Now, similar to the proof for Theorem 1, combining (29) with arguments from (16) to (26), it can be observed that

$$\sup_{\nu \in \mathbb{R}} \left| \bar{F}_n(\nu) - F_\pi(\nu) \right| \xrightarrow{\text{a.s.}} 0.$$

□

Theorem 2. *Under Assumption 1, for any $\delta \in (0,1]$, if $\sum_{i=1}^{K} \delta_i \leq \delta$, then the confidence band defined by F_- and F_+ provides guaranteed coverage for F_π. That is,*

$$\Pr\Big(\forall \nu,\ F_-(\nu) \leq F_\pi(\nu) \leq F_+(\nu)\Big) \geq 1-\delta.$$

Proof. Let A_i be the event that for the key point κ_i, $\texttt{CI}_-(\kappa_i,\delta_i) \leq F_\pi(\kappa_i) \leq \texttt{CI}_+(\kappa_i,\delta_i)$, for all $i \in (1,...,K)$. Let superscript c denote a complementary event; then by the union bound, the total probability of the bounds holding at each key point simultaneously is

$$\Pr\Big(\cap_{i=1}^{K} A_i\Big) = 1 - \Pr\Big((\cap_{i=1}^{K} A_i)^c\Big) = 1 - \Pr\Big(\cup_{i=1}^{K} A_i^c\Big) \geq 1 - \sum_{i=1}^{K} \Pr\Big(A_i^c\Big) \overset{(a)}{\geq} 1-\delta, \quad (30)$$

where (a) holds because the conditions of the theorem assert that the sum of probabilities of the bounds failing at each key point is at most δ. Therefore, using (30),

$$\Pr\left(\forall i \in (1,...,K),\ \texttt{CI}_-(\kappa_i,\delta_i) \leq F_\pi(\kappa_i) \leq \texttt{CI}_+(\kappa_i,\delta_i)\right) \geq 1-\delta. \quad (31)$$

Since by construction, at the key points $(\kappa_i)_{i=1}^{K}$, $F_-(\kappa_i) = \texttt{CI}_-(\kappa_i,\delta_i)$ and $F_+(\kappa_i) = \texttt{CI}_+(\kappa_i,\delta_i)$, it follows from (31) that

$$\Pr\left(\forall i \in (1,...,K),\ F_-(\kappa_i) \leq F_\pi(\kappa_i) \leq F_+(\kappa_i)\right) \geq 1-\delta. \quad (32)$$

Using the monotonically non-decreasing property of a CDF, at any point $\nu \in \mathbb{R}$ such that $\kappa_i \leq \nu \leq \kappa_{i+1}$, we know that $F_\pi(\kappa_i) \leq F_\pi(\nu) \leq F_\pi(\kappa_{i+1})$. Therefore, when the bounds at the key points hold, F_π at the key points can also be upper and lower bounded: $F_-(\kappa_i) \leq F_\pi(\nu) \leq F_+(\kappa_{i+1})$. Therefore, by (32) and the construct in (6), it immediately follows that

$$\Pr\Big(\forall \nu,\ F_-(\nu) \leq F_\pi(\nu) \leq F_+(\nu)\Big) \geq 1-\delta.$$

□

Theorem 3. *Under Assumption 1, for any $1-\delta$ confidence band $\mathcal{F}$, the confidence interval defined by ψ_- and ψ_+ provides guaranteed coverage for $\psi(F_\pi)$. That is,*

$$\Pr\Big(\psi_- \leq \psi(F_\pi) \leq \psi_+\Big) \geq 1-\delta.$$

Proof. Recall that the confidence band $\mathcal{F}$ is a random variable dependent on the data $\mathcal{D}$. Let $\mathbb{E}_{\mathcal{F}}[\cdot]$ represent expectation with respect to $\mathcal{F}$, then repeatedly using the law of total probability,

$$\begin{aligned}
\Pr\Big(\psi_- \leq \psi(F_\pi) \leq \psi_+\Big) &= \mathbb{E}_{\mathcal{F}}\Big[\Pr\Big(\psi_- \leq \psi(F_\pi) \leq \psi_+ \Big| \mathcal{F}\Big)\Big] \\
&= \mathbb{E}_{\mathcal{F}}\Big[\Pr\Big(\psi_- \leq \psi(F_\pi) \leq \psi_+ \Big| F_\pi \in \mathcal{F}, \mathcal{F}\Big) \Pr\Big(F_\pi \in \mathcal{F} \Big| \mathcal{F}\Big) \\
&\qquad + \Pr\Big(\psi_- \leq \psi(F_\pi) \leq \psi_+ \Big| F_\pi \notin \mathcal{F}, \mathcal{F}\Big) \Pr\Big(F_\pi \notin \mathcal{F} \Big| \mathcal{F}\Big)\Big] \\
&\geq \mathbb{E}_{\mathcal{F}}\Big[\Pr\Big(\psi_- \leq \psi(F_\pi) \leq \psi_+ \Big| F_\pi \in \mathcal{F}, \mathcal{F}\Big) \Pr\Big(F_\pi \in \mathcal{F} \Big| \mathcal{F}\Big)\Big] \\
&\overset{(a)}{=} \mathbb{E}_{\mathcal{F}}\Big[\Pr\Big(F_\pi \in \mathcal{F} \Big| \mathcal{F}\Big)\Big] \\
&= \Pr\Big(F_\pi \in \mathcal{F}\Big) \\
&\overset{(b)}{\geq} 1-\delta,
\end{aligned}$$

where (a) follows from that fact that $F_\pi \in \mathcal{F}$ implies $\psi_- \leq \psi(F_\pi) \leq \psi_+$. Step (b) follows from Theorem 2. □

Proof (Alternate). This proof is shorter but requires a theoretical construct of a *set of sets of functions*. That is, let $\mathbb{F}$ be any set of cumulative distribution functions and $\mathscr{F}$ be a set of such sets, such that

$$\mathscr{F} \coloneqq \Big\{\mathbb{F} \ \Big| F_\pi \in \mathbb{F}\Big\}.$$

In other words, $\mathbb{F}$ is the set of CDFs which contains the true CDF F_π, and $\mathscr{F}$ is the set of *all* such sets $\mathbb{F}$. From Theorem 2, we know that the confidence band $\mathcal{F}$ contains F_π with probability at least $1-\delta$. Therefore, it also holds that

$$\Pr(\mathcal{F} \in \mathscr{F}) \geq 1-\delta.$$

However, the event $(\mathcal{F} \in \mathscr{F})$ implies that $\psi_- \leq \psi(F_\pi) \leq \psi_+$ as F_π is contained in this specific $\mathcal{F}$ used to construct ψ_- and ψ_+. Therefore, it also holds that

$$\Pr(\psi_- \leq \psi(F_\pi) \leq \psi_+) \geq 1-\delta.$$

□

Theorem 4. *Under Assumptions 1 and 2, for any $\delta \in (0,1]$, the confidence band defined by $F_-^{(2)}$ and $F_+^{(2)}$ provides guaranteed coverage for $F_\pi^{(2)}$. That is,*

$$\Pr\left(\forall \nu,\ F_-^{(2)}(\nu) \leq F_\pi^{(2)}(\nu) \leq F_+^{(2)}(\nu)\right) \geq 1-\delta.$$

Proof. From Assumption 2, $\sup_{\nu\in\mathbb{R}} \left|F_\pi^{(1)}(\nu) - F_\pi^{(2)}(\nu)\right| \leq \epsilon$. Or equivalently,

$$\forall \nu \in \mathbb{R}, \quad F_\pi^{(1)}(\nu) - \epsilon \leq F_\pi^{(2)}(\nu) \leq F_\pi^{(1)}(\nu) + \epsilon. \tag{33}$$

Using Theorem 2 for the bound obtained on $F_\pi^{(1)}$ for the first domain,

$$\Pr\left(\forall \nu,\ F_-^{(1)}(\nu) \leq F_\pi^{(1)}(\nu) \leq F_+^{(1)}(\nu)\right) \geq 1-\delta. \tag{34}$$

Therefore, combining (33) and (34),

$$\Pr\left(\forall \nu,\ F_-^{(1)}(\nu) - \epsilon \leq F_\pi^{(2)}(\nu) \leq F_+^{(1)}(\nu) + \epsilon\right) \geq 1-\delta. \tag{35}$$

Then by the construct in (7), it follows from (35) that

$$\Pr\left(\forall \nu,\ F_-^{(2)}(\nu) \leq F_\pi^{(2)}(\nu) \leq F_+^{(2)}(\nu)\right) \geq 1-\delta.$$

□

E Extended Discussion for UnO

E.1 Nuances for CDF Inverse and CVaR

For brevity, some nuances for $\hat{F}_n^{-1}(\alpha)$ and $\text{CVaR}_\pi^\alpha(\hat{F}_n)$ were excluded from the main paper. We discuss them in this section.

As discussed earlier in Remark 1, it is possible that $\hat{F}_n(\nu) > 1$ for some $\nu \in \mathbb{R}$ due to the use of importance weighting. Similarly, it is also possible that $\hat{F}_n(\nu) < 1$ for all $\nu \in \mathbb{R}$. Specifically, if $\hat{F}_n(\nu) < \alpha$ for all ν, then it raises the question: how can one obtain an estimate of $F_\pi^{-1}(\alpha)$? To resolve this issue, we use the following estimator of $F_\pi^{-1}(\alpha)$ for UnO:

$$\hat{F}_n^{-1}(\alpha) \coloneqq \begin{cases} \min\left\{g \in (G_{(i)})_{i=1}^n \middle| \hat{F}_n(g) \geq \alpha\right\}, & \text{if} \quad \exists\, g \text{ s.t. } \hat{F}_n(g) \geq \alpha, \\ \max(G_{(i)})_{i=1}^n & \text{otherwise.} \end{cases}$$

However, it is known from Theorem 1 that $\hat{F}_n$ is a uniformly consistent estimator of F_π. Therefore, the edge case that $\hat{F}_n(\nu) < \alpha$ for all ν cannot occur in the limit as $n \to \infty$. Resolving this is required mostly when the sample size is small.

Regarding CVaR, it is known [1] that when the distribution of a random variable (which is G_π for UnO) is continuous, then CVaR can be expressed as,

$$\text{CVaR}_\pi^\alpha(F_\pi) = \mathbb{E}\left[G_\pi \middle| G_\pi \leq F_\pi^{-1}(\alpha)\right], \tag{36}$$

and thus an off-policy sample estimator for (36) can be constructed as,

$$\text{CVaR}_\pi^\alpha(\hat{F}_n) \coloneqq \frac{1}{\alpha}\sum_{i=1}^{n} \mathrm{d}\hat{F}_n(G_{(i)})G_{(i)}\mathbb{1}_{\{G_{(i)}\leq Q_\pi^\alpha(\hat{F}_n)\}}.$$

However, for distributions that are not continuous, a more generic definition for CVaR is [13],

$$\text{CVaR}_\pi^\alpha(F_\pi) = \inf_g \left\{ g - \frac{1}{\alpha}\mathbb{E}\Big[\max\left(0, g - G_\pi\right)\Big]\right\}. \tag{37}$$

We extend the sample estimator by Brown [13] for (37) and use the following off-policy estimator for UnO:

$$\text{CVaR}_\pi^\alpha(\hat{F}_n) \coloneqq \hat{F}_n^{-1}(\alpha) - \frac{1}{\alpha}\sum_{i=1}^{n} \mathrm{d}\hat{F}_n(G_{(i)}) \left(\max\left(0, \hat{F}_n^{-1}(\alpha) - G_{(i)}\right)\right)$$

E.2 Optimizing Confidence Bands for Tighter Bounds:

Constructing $\mathcal{F}$ requires selecting K key points for which `CI`s are computed. If too many key points are selected, then each δ_i has to be a very small positive value so that $\sum_{i=1}^{K}\delta_i \leq \delta$, as required by Theorem 2. This will make the confidence intervals wide at each key point. In contrast, if too few key points are selected, then the confidence intervals at the κ_i's will be relatively tighter, but this will not tighten the intervals *between* the κ_i's due to the way F_- and F_+ are constructed in (5). Further, the overall tightness of $\mathcal{F}$ is also affected by the location of each κ_i and its respective failure rate δ_i. Therefore, to get a tight $\mathcal{F}$, we propose searching for a $\theta \coloneqq \left(K, (\kappa_i)_{i=1}^K, (\delta_i)_{i=1}^K\right)$ that minimizes the area enclosed in $\mathcal{F}$. That is, let $\Delta_{i+1} \coloneqq \kappa_{i+1} - \kappa_i$, then the area enclosed in $\mathcal{F}$ is

$$\mathscr{A}(\theta) \coloneqq \sum_{i=0}^{K} \left(\texttt{CI}_+(\kappa_{i+1}, \delta_{i+1}) - \texttt{CI}_-(\kappa_i, \delta_i)\right)\Delta_{i+1}.$$

To avoid multiple comparisons [8], we first partition $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{eval}}$. Subsequently, $\mathcal{D}_{\text{train}}$ is used to search for θ^* as follows, and then θ^* is used with $\mathcal{D}_{\text{eval}}$ to obtain $\mathcal{F}$.

$$\theta^* \coloneqq \arg\min_\theta \ \mathscr{A}(\theta) \tag{38}$$

$$\text{s.t.} \qquad G_{\min} < \kappa_i < G_{\max}, \quad \sum_{i=1}^{K}\delta_i \leq \delta, \quad \delta_i \geq 0, \qquad \forall i \in (1, ..., K).$$

Remark 5. *A global optimum of* (38) *is not required—any feasible θ can be used with $\mathcal{D}_{eval}$ to obtain a confidence band $\mathcal{F}$. Optimization only helps by making the band tighter.*

For our experimental results, when searching θ^* for (38), we keep the number of key points, K, fixed to $\log(n)$, where n is the number of observed trajectory samples in $\mathcal{D}$. To search for the locations $(\kappa_i)_{i=1}^K$ and the failure rates $(\delta_i)_{i=1}^K$ at each key point, we use the BlackBoxOptim library[2] available in Julia [11]. To perform this optimization, we construct $\mathcal{D}_{\text{train}}$ using 5% of data from $\mathcal{D}$, and construct $\mathcal{D}_{\text{eval}}$ using the rest of the data. Following the idea by Thomas et al. [91], when searching for θ^* using $\mathcal{D}_{\text{train}}$, bounds for the key points $(\kappa_i)_{i=1}^K$ are obtained as if the number of samples are equal to the number of samples available in $\mathcal{D}_{\text{eval}}$ (see Equation 7 in the work by Thomas et al. [91] for more discussion on this). Instead of using a single split, one could potentially also leverage results by Romano and DiCiccio [73] to use multiple splits; we leave this for future work.

E.3 Bound Specialization

In (38), θ was searched to minimize the area $\mathscr{A}(\theta)$ enclosed within $\mathcal{F}(\theta)$, where $\mathcal{F}(\theta)$ represents the CDF band obtained using the parameter θ. This was done without any consideration of the downstream parameter ψ for which the bounds would be constructed using $\mathcal{F}(\theta)$. Therefore, the

[2] https://github.com/robertfeldt/BlackBoxOptim.jl

band $\mathcal{F}(\theta)$ is tight overall, but need not be the best possible if only a specific parameter ψ's bounds are required using $\mathcal{F}(\theta)$.

For example, consider obtaining bounds for CVaR_π^α. As can be seen from the geometric insight in Figure 3, bounds for CVaR are mostly dependent on the tightness of $\mathcal{F}(\theta)$ near the lower tail. Therefore, if one can obtain $\mathcal{F}(\theta)$ that is tighter near the lower tail, albeit looser near the upper tail, that would provide a better bound for CVaR as opposed to a band $\mathcal{F}(\theta)$ that has uniform tightness throughout.

To get a tight $\mathcal{F}(\theta)$ in such cases where there is a single downstream parameter of interest, we propose searching for a $\theta := \left(K, (\kappa_i)_{i=1}^K, (\delta_i)_{i=1}^K\right)$ that directly optimizes for the final parameter of interest instead of the area enclosed in $\mathcal{F}(\theta)$. For example, if only the lower bound for $\psi(F_\pi)$ is required, then let

$$\psi_-(\theta) := \inf_{F \in \mathcal{F}(\theta)} \psi(F).$$

Next, the optimization using $\mathcal{D}_{\text{train}}$ can then be modeled as the following,

$$\begin{aligned} \theta^* := \arg\max_\theta \ & \psi_-(\theta) \\ \text{s.t.} \quad & G_{\min} < \kappa_i < G_{\max}, && \forall i \in (1, ..., K), \\ & \sum_{i=1}^K \delta_i \leq \delta, \quad \delta_i \geq 0, && \forall i \in (1, ..., K), \end{aligned}$$

This would result in θ^* that when used with $\mathcal{D}_{\text{eval}}$ can be expected to provide the CDF band which will yield the highest lower bound for $\psi(F_\pi)$.

E.4 Approximate Bounds for Any Parameter using Bootstrap

In Algorithm 1, we provide the pseudo code for obtaining bootstrap-based bounds for any parameter $\psi(F_\pi)$. In Line 1, B datasets $(\mathcal{D}_i^*)_{i=1}^B$ are generated from $\mathcal{D}$ using resampling, and for each of these resampled data sets, B (weighted IS-based) CDF estimates $(\bar{F}_{n,i}^*)_{i=1}^B$ are obtained. In Line 3, sample estimates $(\psi(\bar{F}_{n,i}^*))_{i=1}^B$ for the desired parameter $\psi(F_\pi)$ are constructed using the B estimated CDFs. In Line 4, these sample estimates for $\psi(F_\pi)$ can be subsequently passed to the bias-corrected and accelerated (BCa [32]) bootstrap procedure to obtain approximate lower and upper bounds (ψ_-, ψ_+).

Algorithm 1: Bootstrap Bounds for $\psi(F_\pi)$

1 **Input:** Dataset $\mathcal{D}$, Confidence level $1 - \delta$
2 Bootstrap B datasets $(\mathcal{D}_i^*)_{i=1}^B$ and create $(\bar{F}_{n,i}^*)_{i=1}^B$
3 Bootstrap estimates $(\psi(\bar{F}_{n,i}^*))_{i=1}^B$ using $(\bar{F}_{n,i}^*)_{i=1}^B$
4 Compute (ψ_-, ψ_+) using BCa$((\psi(\bar{F}_{n,i}^*))_{i=1}^B, \delta)$
5 **Return** (ψ_-, ψ_+)

E.5 Extended Discussion of High-Confidence Bounds for Any Parameter

Section 4 of the main paper discussed how high-confidence bounds ψ_- and ψ_+ can be obtained for any parameter $\psi(F_\pi)$ using the confidence band $\mathcal{F}$. Specifically, in Figure 3, geometric insights for obtaining the analytical form of the bounds for the mean, quantile, and CVaR were discussed. Extending that discussion, Figure 6 provides geometric insights for bounding other parameters, namely variance, inter-quantile ranges, and entropy, in the off-policy setting.

An advantage of having the CDF band $\mathcal{F}$ is that it can permit bounding other novel parameters that might be of interest. While analytical bounds using geometric insights, as discussed for a number of popular parameters, should also be the first attempt for the desired novel parameter, it may be the case that such geometric insight cannot be obtained. In such cases, a CDF F can be directly parameterized using a spline curve, or a piecewise non-decreasing function that is constrained to be within $\mathcal{F}$. Depending on how rich this parameterization is, it may be feasible to use a black-box optimization routine and obtain a globally optimal F that minimizes (maximizes) the desired parameter $\psi(F)$. If not feasible, an approximate bound can be achieved by using the best found local optima.

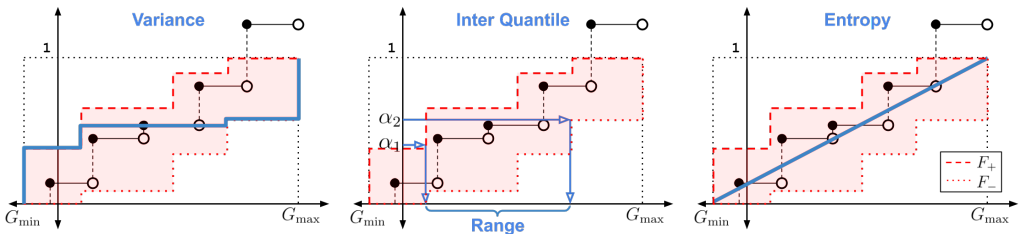

Figure 6: Similar to Figure 3, given a confidence band $\mathcal{F}$, lower and upper bounds for several other parameters can also be obtained using simple geometric insights. **(Left)** An upper bound for the variance can be obtained by observing that variance is maximized when the probability of events on either extreme are maximized. Therefore, the CDF $F \in \mathcal{F}$ for such a distribution will initially follow (from left to right) F_+ and then make a horizontal jump (at a specific jump point) to F_-, which it then follows until 1. The variance of the distribution with this CDF, F, will give the desired upper bound. Analogously, the CDF that initially follows F_- and then jumps vertically (at a specific jump point) to F_+, assigns highest probability to events near the mean and thus results in the lowest variance [74]. **(Middle)** An upper bound for the inter-quantile range can be obtained by maximizing the value of upper α_2-quantile and subtracting the minimum value for the lower α_1-quantile. This can be obtained by $F_-^{-1}(\alpha_2) - F_+^{-1}(\alpha_1)$. Analogously, a lower bound can be obtained using $\max(0, F_+^{-1}(\alpha_2) - F_-^{-1}(\alpha_1))$. **(Right)** An upper bound on the entropy can be obtained by what Learned-Miller and DeStefano [55] call a "string-tightening" algorithm. That is, if the ends of a tight string are held at the bottom-left and the upper-right corner of $\mathcal{F}$, and the entire string is constrained to be within $\mathcal{F}$, then the path of the string corresponds to the $F \in \mathcal{F}$ that has highest entropy. In our figure, such an F corresponds to the CDF of the uniform distribution, which is known to have maximum entropy. Unless some stronger assumptions are made, the lower bound on differential entropy is typically $-\infty$ if there is any possibility of a point mass.

E.6 Tackling Smooth Non-stationarity using Wild Bootstrap

From Theorem 1, it is known that the proposed estimator $\hat{F}_n(\kappa)$ provides unbiased estimates for $F_\pi(\kappa)$, even with a single observed trajectory. In the non-stationary setting, let the true underlying CDF of returns for π in the episode i be $F_\pi^{(i)}(\kappa)$, and the estimate of $F_\pi^{(i)}(\kappa)$ using the trajectory observed during the episode i be

$$\hat{F}_n^{(i)}(\kappa) \coloneqq \rho_i \mathbb{1}_{\{G_i \leq \kappa\}} \qquad \forall i \in \{1, 2, ..., L\}.$$

Next, the trend of the sequence $\left(\hat{F}_n^{(i)}(\kappa)\right)_{i=1}^L$ can be analyzed to forecast $\hat{F}_n^{(L+\ell)}(\kappa)$ for the future episode $L+\ell$ when the policy π will be executed. Particularly, under Assumption 3, $\exists w_\kappa$, such that, $\forall i \in (1, ..., L+\ell), \quad F_\pi^{(i)}(\kappa) = \phi(i)^\top w_\kappa$. Therefore, using the unbiased estimates $\left(\hat{F}_n^{(i)}(\kappa)\right)_{i=1}^L$ of $\left(F_\pi^{(i)}(\kappa)\right)_{i=1}^L$, we propose searching for w_κ using least-squares regression. Let $X \coloneqq [1, 2, ..., L]$ be the episode numbers in the past, then the predicates Φ_κ, the targets Y_κ, and the corresponding least-squares solution w_κ can be obtained as,

$$\begin{aligned}
\Phi_\kappa &\coloneqq [\phi(X_1), \phi(X_2), ..., \phi(X_L)] && \in \mathbb{R}^{L \times d}, \\
Y_\kappa &\coloneqq [\hat{F}_n^{(1)}(\kappa), \hat{F}_n^{(2)}(\kappa), ..., \hat{F}_n^{(L)}(\kappa)] && \in \mathbb{R}^{L \times 1}, \\
w_\kappa &\coloneqq \left(\Phi_\kappa^\top \Phi_\kappa\right)^{-1} \Phi_\kappa^\top Y_\kappa && \in \mathbb{R}^{d \times 1}.
\end{aligned}$$

Using w_κ, an unbiased estimate of $F_\pi^{(L+\ell)}(\kappa)$ can be obtained as,

$$\hat{F}_n^{(L+\ell)}(\kappa) \coloneqq \phi(L+\ell)^\top w_\kappa. \tag{39}$$

The point forecast $\hat{F}_n^{(L+\ell)}(\kappa)$ from (39) can then be combined with Algorithms 1 and 2 presented by Chandak et al. [16] to obtain wild-bootstrap-based confidence intervals for $F_\pi^{(L+\ell)}(\kappa)$. Once the confidence intervals are obtained at different key points, (5) can be used to construct an entire confidence band for $F_\pi^{(L+\ell)}$.

F Empirical Details

F.1 Domain Details

In this section, we discuss domain details and how π and β were selected for these domains. The code for the domains, baselines [91, 18], and the proposed UnO estimator can be found at https://github.com/yashchandak/UnO.

Recommender System: Systems for online recommendation of tutorials, movies, advertisements, etc., are ubiquitous [86, 88]. In these settings, it may be beneficial to fully characterize a customer's experience once the new system/policy is deployed. To abstract such settings, we created a simulated domain where the user's interest for a finite set of items is represented using the corresponding item's reward.

Using an actor-critic algorithm [83], we find a near-optimal policy π, which we use as the evaluation policy. Let $\pi^{\texttt{rand}}$ be a random policy with uniform distribution over the actions (items). Then for an $\alpha = 0.5$, we define the behavior policy $\beta(a|s) := \alpha\pi(a|s) + (1-\alpha)\pi^{\texttt{rand}}(a|s)$ for all states and actions.

Gridworld: We also consider a standard continuous-state Gridworld with partial observability (which also makes the domain non-Markovian in the observations), stochastic transitions, and eight discrete actions corresponding to up, down, left, right, and the four diagonal movements. The off-policy data was collected using two different behavior policies, β_1 and β_2, and the evaluation policies for this domain were obtained similarly as for the recommender system domain discussed above. Particularly, using $\alpha = 0.5$, we define $\beta_1(a|o) := \alpha\pi(a|0) + (1-\alpha)\pi^{\texttt{rand}}(a|o)$ for all states and actions. Similarly, β_2 was defined using $\alpha = 0.75$.

Diabetes Treatment: This domain is modeled using an open source implementation [103] of the U.S. Food and Drug Administration (FDA) approved Type-1 Diabetes Mellitus Simulator (T1DMS) [59] for the treatment of type-1 diabetes. An episode corresponds to a day, and each step of an episode corresponds to a minute in an *in silico* patient's body and is governed by a continuous time nonlinear ordinary differential equation (ODE) [59]. In such potentially critical medical applications, it is important to go beyond just the expected performance and to characterize the risk associated with it, *before deployment*.

To control the insulin injection, which is required for regulating the blood glucose level, we use a policy that controls the parameters of a *basal-bolus controller*. This controller is based on the amount of insulin that a person with diabetes is instructed to inject prior to eating a meal [6]:

$$\text{injection} = \frac{\text{current blood glucose} - \text{target blood glucose}}{CF} + \frac{\text{meal size}}{CR},$$

where "current blood glucose" is the estimate of the person's current blood glucose level, "target blood glucose" is the desired blood glucose, "meal size" is the estimate of the size of the meal the patient is about to eat, and $CR \in [CR_{\texttt{min}}, CR_{\texttt{max}}]$ and $CF \in [CF_{\texttt{min}}, CF_{\texttt{max}}]$ are two parameters of the controller that must be tuned based on the body parameters to make the treatment effective. We designed an RL policy that acts on the discretized space of the parameters, CR and CF, for the above basal-bolus controller. Behavior and evaluation policies were selected similarly as discussed for the recommender system domain.

F.2 Extended Discussion on Results for Stationary Settings

The main results for the stationary setting are provided in Figure 4 of the main body. In this section, we provide some additional discussion on the observed trends for the bounds.

Notice in Figure 4 that UnO-CI bounds for the variance can require up to an order of magnitude less data compared to the existing bound for the variance [18]. This can be attributed to the fact that Chandak et al. [18] construct the bounds using $\mathbb{E}[\rho G^2] - \mathbb{E}[\rho G]^2$, where it can be observed that the second term depends quadratically on ρ. This makes the variance of that term effectively "doubly exponential" in the horizon length. This does not happen in the CDF-based approach as the bounds at any key point κ depend on $\mathbb{E}[\rho \mathbb{1}_{G<\kappa}]$), which does not have any higher powers of ρ.

Another thing worth noting in Figure 4 is that not only the bounds for different parameters, but even the upper and lower bounds for the same parameter converge at different rates (especially for smaller values of n). Therefore, there are two particular trends to observe: (a) how close the bounds are to the true value at the beginning, and (b) how quickly they improve. Both of these depend on the direction for which clipping plays a major role and also how the bounds depend on the tails. For example, for the mean, as the distributions are right skewed (because evaluating policy π is a near-optimal policy), the bounds on the CDF are clipped more from the lower end (so that $F(\nu) >= 0$ always). Therefore, since the upper bound on the mean depends on the lower CDF bound (see Figure 3), it starts close to the estimate itself but the progress actually seems slow because shrinking CDFs bounds at any specific $F(\nu)$ from the lower end does not impact the bound until the point where clipping is not required anymore.

For variance, the upper bound depends on both the upper bound on the lower tail and the lower bound on the upper tail (see Figure 6), and these two benefit from clipping the least and also converge the slowest. In contrast, the lower bound for variance depends on the upper bound on the upper tail and the lower bound on the lower tail, which are clipped immediately to be below 1 and above 0, respectively. Appendix B.1 (knowledge of $G_{\min}$, $G_{\max}$) and Fig 6 provide more intuition on this.