Universal Off-Policy Evaluation

Anonymous Author(s)
Affiliation
Address
email

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 do 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 [55] or financial assets [87], 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 [84]. 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, thus performance metrics such as value at risk (VaR) or conditional value at risk (CVaR) are more appropriate [48, 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 experience, 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 [51, 85]. Recent works in distributional reinforcement learning (RL) have also investigated estimating the cumulative distribution of returns [7, 24] and its various statistical functionals [77]. 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?*

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Prior Work: Off-policy methods can be broadly categorized as model-based or model-free [84]. Model-based methods typically require strong assumptions when statistical guarantees are needed on the function approximation estimates. 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.

By contrast, model-free methods are applicable to a wider variety of settings. Unfortunately, the popular technique of using *importance-weighted returns* [72] 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, 100] 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* [93, 45, 18]. Instead of the frequentist intervals, Bayesian posterior distribution over the mean return and various statistics of that distribution can also be obtained [106]. 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.

Notation: For brevity, we first restrict our focus to the stationary setting. In Section 4, 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 maps states to observations, $\mathcal{R}$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and d_0 is the starting state distribution. 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 [81]. 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 history $(O_0, A_0, \beta(A_0|O_0), R_0, O_1, ...)$. Let $G_i := \sum_{j=0}^{T} \gamma^j R_j$ be the *return* for 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 denote $g(h)$ as the return for trajectory h. Further, to ensure that trajectories in $\mathcal{D}$ are informative, we make a standard assumption that any outcome under π has sufficient probability of occurring under β,

Assumption 1. *The set $\mathcal{D}$ contains independent (not necessarily identically distributed) histories generated using $(\beta_i)_{i=1}^n$, along with the probability of the actions chosen, such that for some $\varepsilon > 0$, $(\beta_i(a|o) < \varepsilon) \implies (\pi(a|o) = 0)$, for all $o \in \mathcal{O}, a \in \mathcal{A}$, and $i \in (1, \ldots, n)$.*

2 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 the same in the off-policy setting requires additional consideration as **(a)** the *entire* distribution of the observed returns needs to be adjusted to estimate F_π, even when **(b)** data might be collected using multiple behavior policies and **(c)** data collection might have been influenced by unobserved confounders.

We begin by observing that F_π for any $\nu \in \mathbb{R}$ 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 [72, 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 the trajectory 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.

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

Figure 1 provides an 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 [71]) 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,*

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

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 $\rho > 1$.*

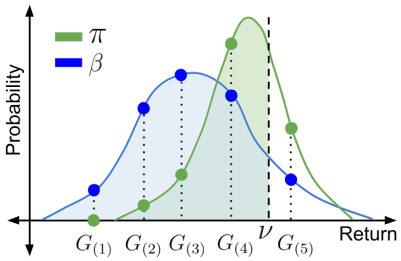

Figure 1: An illustration of return distributions for π and β. The CDF at any point ν corresponds to the area under the probability distribution up until ν. Having order statistics $(G_{(i)})_{i=1}^5$ of samples $(G_i)_{i=1}^5$ drawn using β, (3) constructs an empirical estimate of the CDF for π (*green*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

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

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,

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

where $\mathrm{d}\hat{F}_n(\nu) \coloneqq 0$ if $\nu \neq G_{(i)}$ for any $i \in (1, \dots, n)$. 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.

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

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

Remark 2. *Let H_i be the 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 [72].*

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 $\psi(F_\pi)$.

3 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 [95] 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_π is within this set 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 **(c)** DKW assumes samples are identically distributed, however, off-policy data $\mathcal{D}$ might be collected using multiple different β's. This raises the question: *How do we obtain $\mathcal{F}$ in the off-policy setting?*

Before constructing a confidence band $\mathcal{F}$, let us focus on obtaining bounds for a single point, $F_\pi(\kappa)$, first. 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. [92] present tight bounds for the mean of IS-based random variables by mitigating the variance resulting from the heavy tails associated with IS; their method can thus be used with $\hat{F}_n(\kappa)$ to bound $F_\pi(\kappa)$. Alternatively, recent work by Kuzborskij et al. [52] can potentially be used with a WIS-based F_π estimate.

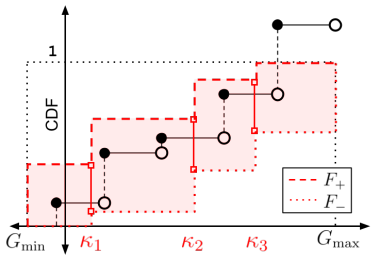

Figure 2: An illustration of $\hat{F}_n$ (in black) using five return samples and the confidence band $\mathcal{F}$ (red shaded region) computed using (5) with confidence intervals (red lines) at three key points $(\kappa_i)_{i=1}^3$. Notice that the vertical "steps" in $\hat{F}_n$ can be of different 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.

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

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

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 $F_\pi(\kappa_i) \ge \texttt{CI}_-(\kappa_i, \delta_i)$ then for any $\nu \ge \kappa_i$, $F_\pi(\nu)$ must be no less than $\texttt{CI}_-(\kappa_i, \delta_i)$. Similarly, if $F_\pi(\kappa_i) \le \texttt{CI}_+(\kappa_i, \delta_i)$ then for any $\nu \le \kappa_i$, $F_\pi(\nu)$ 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,

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

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

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

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

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

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,*

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

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_+ [75]. 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, $\mathcal{F}$.*

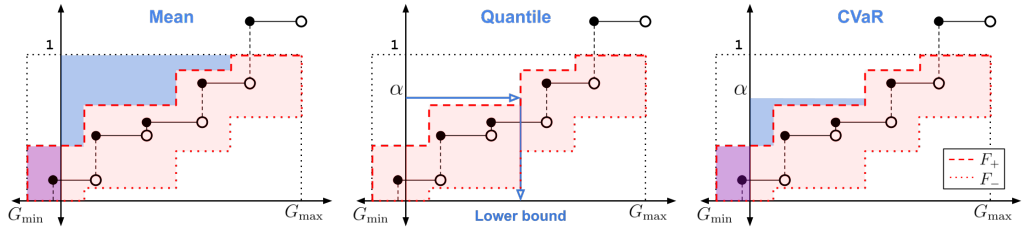

Figure 3: Given a confidence band $\mathcal{F}$, bounds for many parameters can be obtained using geometry. **(Left)** For a lower bound on the mean, we would want a CDF $F \in \mathcal{F}$ that assigns as high a probability as possible on lower G values, and F_+ is the CDF which does that. To obtain 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 [91] 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.

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 the 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.

4 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 [95], 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, loan approval) 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 the possible alternatives in Appendix B.2.

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 3. 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 [94, 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,* $\forall i \in [1, L + \ell], \quad F_\pi^{(i)}(\nu) = \phi(i)^\top w_\nu$.

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 `CI`s. In particular, we propose using *wild bootstrap* [58, 26] that provides *approximate* `CI`s 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 3.

5 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 [103] 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 domain descriptions and descriptions of the procedures for obtaining π and β are provided in Appendix F.1. 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 provides consistent estimates for the entire CDF and a wide range of its parameters: mean, variance, quantile, and CVaR.

(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 of Thomas et al. [92] 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.

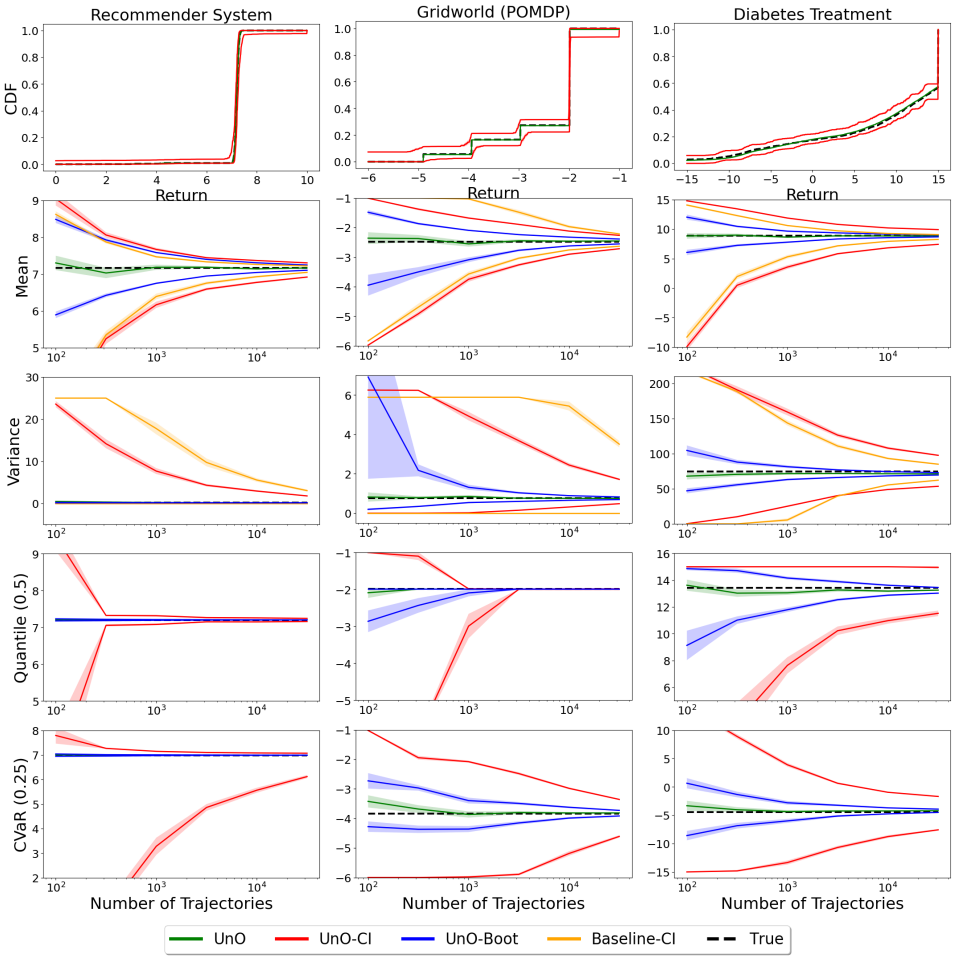

Figure 4: Performance trend of the proposed estimators and bounds on three domains. The black dashed line is the true value of F_π or $\psi(F_\pi)$, green is our UnO estimator, red is our `CI`-based UnO bound, blue is the bootstrap version of our UnO bound, and yellow is the baseline bound for the mean [92] or variance [18]. Each bound has two lines (upper and lower); however, some of those 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).

(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: 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

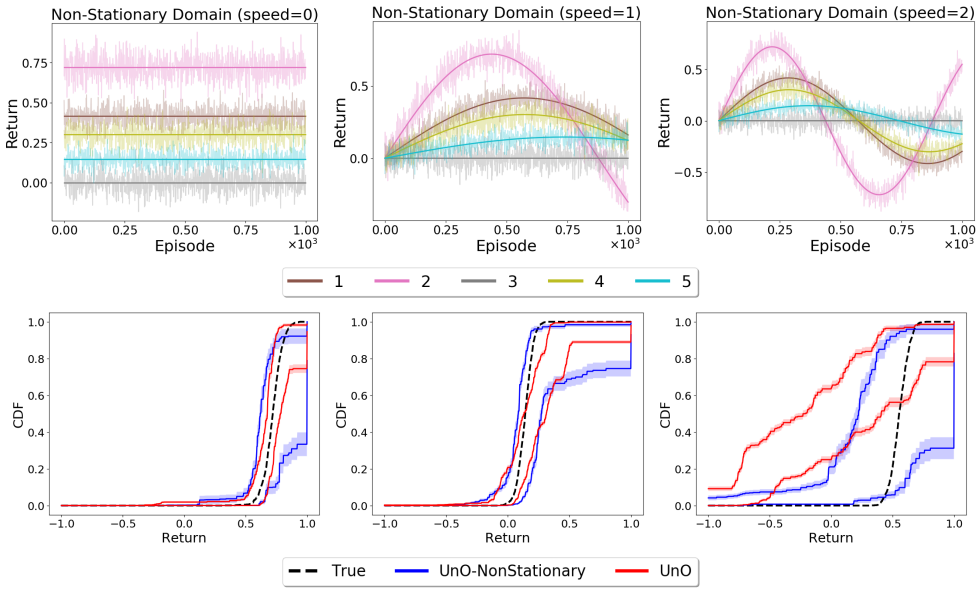

Figure 5: **(Top row)** True rewards (unknown to the RL agent) associated with each of the five items over the past 1000 episodes for different *speeds* of non-stationarity. Speed of 0 indicates stationary setting and higher speeds indicates greater degrees of non-stationarity. **(Bottom row)**. The 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 stderr 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, 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.

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: $1000 + 1$. Results for this domain are presented in Figure 5.

6 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 now fully characterize the (potentially dangerous or costly) behavior of a policy without having to deploy it. There are many new question regarding how UnO can be improved by further reducing data requirements or weakening assumptions. Subsequent to this work, Huang et al. [43] showed how models can be used to obtain UnO-style doubly robust estimators along with uniform CDF bands in the contextual bandit setting. This allows their method to also provide finite-sample bounds for a broad class of Lipschitz risk functionals.

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 A. Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *arXiv preprint arXiv:2002.04518*, 2020.

[48] 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.

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

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

[51] 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.

[52] 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.

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

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

[55] 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.

[56] 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.

[57] Y. Liu, P.-L. Bacon, and E. Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *International Conference on Machine Learning*, pages 6184–6193. PMLR, 2020.

[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. URL http://jmlr.org/papers/v21/20-124.html.

[62] 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.

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

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

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

[66] 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.

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

[68] 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.

[69] 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.

[70] 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.

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

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

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

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

[75] 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.

[76] 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.

[77] 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.

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

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

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

[81] 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.

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

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

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

[85] 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.

[86] 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.

[87] 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.

[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. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

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

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

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

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

[94] 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.

[95] 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.

[96] 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.

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

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

[99] 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.

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

[101] 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.

[102] 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.

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

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

[105] 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.

[106] 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] See Appendix B.1
   (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 included with the submission.

(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 [98], and PyTorch [69].

(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? [Yes] Our entire code is included in the supplemental.

(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]