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*. Drawing inspiration from recent observations on animal learning that highlight the ability of dopaminergic neurons to encode the entire distribution (not just the *expectation*) of outcomes, we take the first steps towards a *universal off-policy estimator* (UnO)—one that provides off-policy estimates and high-confidence bounds for the entire distribution of returns and *any* of its parameters. 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.

Keywords: off-policy evaluation, counterfactuals, high-confidence bounds, distributional reinforcement learning, distributional statistics, risk measures.

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 NeurIPS 2021 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.

1 Introduction

Canonically, the reward prediction error theory of dopamine was based on reward predictions that were represented as a single scalar quantity and supported learning about the expectation, or mean, of stochastic outcomes. Based on surprising new observations of the dopaminergic neurons, Dabney et al. [3] instead hypothesized that the brain represents possible future rewards not as a single mean, but instead as a probability distribution, effectively representing multiple future outcomes simultaneously. This study was inspired by recent developments in *distributional reinforcement learning* [1] that demonstrated benefits of explicitly accounting for the return distribution in the *on-policy* setting for practical applications. However, bridging these connections to the *off-policy* setting had remained an open question so far.

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 or financial assets, 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. However, this metric can be limiting for many problems of interest.

As controlled experiments could be expensive in terms of time, effort, and money, one may want to analyze the shift in dopamine based reward distribution for different experimental variations (for example, under a different probability of liquid rewards for a given cue to the mice [3]), *without* conducting any new experiments. Similarly, leveraging historical data to design safety-critical treatments (e.g., Alzheimer's and other brain disorders) may require ensuring improvement not only in terms of the *expected* outcome but also minimizing the chances of risk-prone outcomes, and so performance metrics such as value at risk (VaR) or conditional value at risk (CVaR) may be more appropriate. Similarly, 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 may use metrics like variance and entropy. As deciding the right metric often requires careful analysis of ethical and moral concerns, it may be beneficial to have all of these different metrics simultaneously to inform better decision-making. However, 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 and also provide finite-sample confidence bounds that hold simultaneously with high probability for any desired performance metrics?*

Prior Work: Off-policy methods can be broadly categorized as model-based or model-free. 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* only corrects for the *mean* under the off-policy distribution. 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. A detailed discussion of existing work can be found in the work by Chandak et al. [2].

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)** exact high-confidence bounds that hold *simultaneously* for all parameters, and 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 restrict our focus to the stationary setting in this draft. We consider a *partially observable Markov decision process* (POMDP) and 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* $(O_0, A_0, \beta(A_0|O_0), R_0, O_1, ...)$. Let $G_i \coloneqq \sum_{j=0}^{T} \gamma^j R_j$ be the *return* of H_i, where $\forall i,\ G_{\min} < G_i < G_{\max}$ for some finite constants $G_{\min}$ and $G_{\max}$, $\gamma \in [0, 1]$ is a discounting factor 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. For notational simplicity we consider the set of observations, actions, and rewards to be finite, such that when T is finite, the total number of possible trajectories is also finite (although all our results extend to the continuous setting as well). 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.

2 UnO: Universal Off-Policy Estimator

With an aim to communicate the main idea across diverse disciplines at RLDM, we primarily focus on illustrative depictions of the key takeaway points and sketch out the initial steps towards the main theoretical results. We encourage the readers to check out the work by Chandak et al. [2] for complete details.

For the desired universal method, instead of considering each parameter individually, we suggest estimating (and bounding) the entire *cumulative distribution function* (CDF) of returns first, i.e., $\forall \nu \in \mathbb{R}, F_\pi(\nu) := \Pr(G_\pi \leq \nu)$. Any distributional parameter, $\psi(F_\pi)$, can then be estimated (and bounded) using the estimate of F_π.

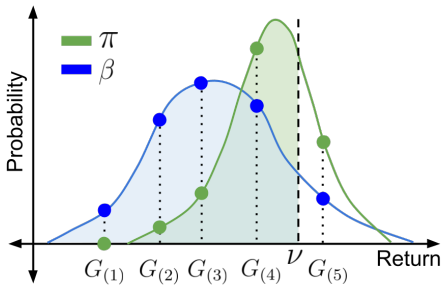

Figure 1: While we aim to evaluate π, we only have off-policy data from a behavior policy β, and the typical use of importance sampling only corrects for the *mean* return. To overcome this, we propose an estimator $\hat{F}_n$ that uses importance sampling from the *perspective of the CDF* to correct for the *entire* distribution of returns. Figure on left illustrates 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, (2) 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-weighted counts* of $G_i \leq \nu$.

To formalize the idea, 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,

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

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, under the support assumption [2] as $\forall \beta,\ \mathscr{H}_\pi \subseteq \mathscr{H}_\beta$,

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

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, but for the *entire cumulative distribution function* (CDF) of return G_π. Formally, let $\rho_i := \prod_{j=0}^{T} \frac{\pi(A_j|O_j)}{\beta_i(A_j|O_j)}$ denote the importance ratio for H_i, which is equal to $\Pr(H_\pi = h)/\Pr(H_\beta = h)$. Then, based on (2), we propose the following non-parametric and model-free estimator for F_π,

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

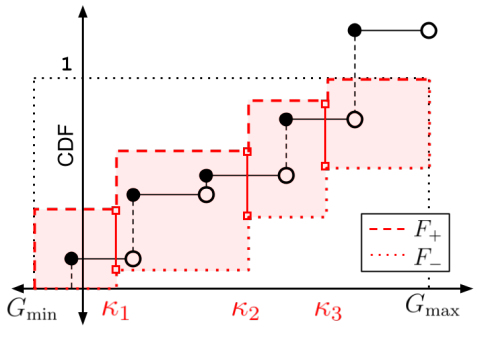

Figure 2: An illustration of $\hat{F}_n$ (in black) using five return samples. To construct a *confidence band* for $F(\pi)$, we show that estimating $F_\pi(\nu)$ for any *specific* ν can be reduced to mean estimation. Therefore, we can obtain confidence intervals for that specific $F_\pi(\nu)$ using existing bounds for the mean. Now using confidence intervals for multiple such points (e.g., solid red lines at three points $(\kappa_i)_{i=1}^3$ in the figure) a confidence band $\mathcal{F}$ (red shaded region) can be computed that contains the entire CDF F_π with high probability. 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 (this is an expected property of estimators based on importance sampling). However, since we know that F_π is never greater than 1, $\mathcal{F}$ can be clipped at 1.

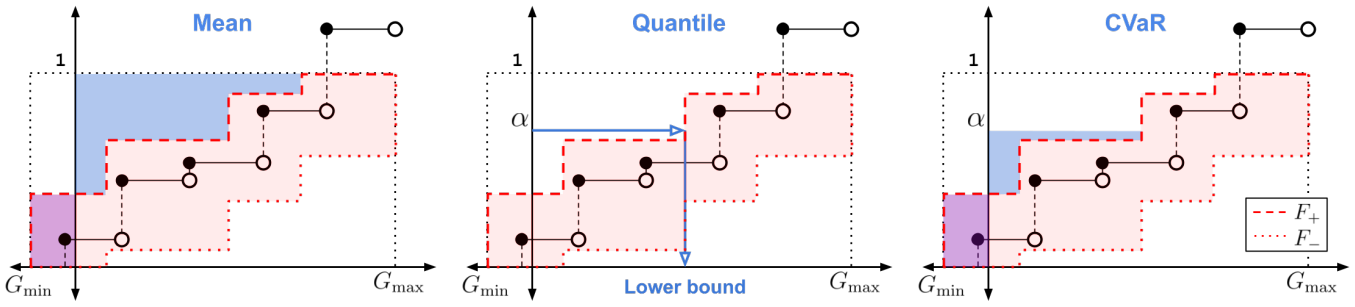

Figure 3: Given a confidence band $\mathcal{F}$, bounds for many parameters can be obtained using geometry. **(Left)** For a lower bound on the mean, we would want a CDF $F \in \mathcal{F}$ that assigns as high a probability as possible on lower G values, and F_+ is the CDF which does that. To obtain the mean of F_+, we use a known 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. 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, 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_-. Similar insights for variance, inter-quantile, entropy, etc., are also available [2].

Benefits of $\hat{F}_n$: Due to space constraints, we briefly summarize several benefits of $\hat{F}_n$.

- $\hat{F}_n$ provides an unbiased and an *uniformly* consistent estimator for F_π, i.e., $\forall \nu \in \mathbb{R}, \quad \mathbb{E}_{\mathcal{D}}[\hat{F}_n(\nu)] = F_\pi(\nu)$, and $\sup_{\nu \in \mathbb{R}} |\hat{F}_n(\nu) - F_\pi(\nu)| \xrightarrow{\text{a.s.}} 0$. The results holds even when the data $\mathcal{D}$ is collected using multiple behavior policies $(\beta_i)_{i=1}^n$, the domain is non-Markovian or has partial observability (confounders).
- Using $\hat{F}_n$, we can construct off-policy estimators of the inverse CDF as $\hat{F}_n^{-1}(\alpha) \coloneqq \min\{g \in (G_{(i)})_{i=1}^n | \hat{F}_n(g) \geq \alpha\}$ for all $\alpha \in [0, 1]$, and for the probability distribution estimator as $\mathrm{d}\hat{F}_n(G_{(i)}) \coloneqq \hat{F}_n(G_{(i)}) - \hat{F}_n(G_{(i-1)})$. Using these, any parameter $\psi(F_\pi)$ (e.g., variance, quantiles, risk measures, etc.) can now be directly estimated.
- As depicted in Figures 2 and 3, $\hat{F}_n$ can be used to construct a *confidence band* $\mathcal{F} : \mathbb{R} \to 2^{\mathbb{R}}$, such that the true $F_\pi(\nu)$ is within the set $\mathcal{F}(\nu)$ with high probability for any given sample-size $n > 1$, i.e., $\Pr(\forall \nu \in \mathbb{R},\ F_\pi(\nu) \in \mathcal{F}(\nu)) \geq 1 - \delta$, for any $\delta \in (0, 1]$. Further, using $\mathcal{F}$, lower and upper bounds for any parameter $\psi(F_\pi)$ (e.g., variance, quantiles, risk measures) can be constructed as $\psi_- \coloneqq \inf_{F \in \mathcal{F}} \psi(F)$ $\psi_+ \coloneqq \sup_{F \in \mathcal{F}} \psi(F)$, respectively.
- Having an off-policy estimator of any $\psi(F_\pi)$ opens up the possibility of using *resampling*-based methods, like statistical bootstrapping, to obtain *approximate* confidence intervals for $\psi(F_\pi)$. We show how they can be combined with UnO to get significantly tighter bounds with less data, albeit they may not be valid for finite sample sizes.
- Often historical data might be collected over an extended period of time, and thus may be subject to factors influenced by external changes. We also provide variants of UnO to handle such discrete distribution shifts and smooth non-stationarities resulting from changes in the environment [2].

Results and Discussion: We provide empirical results for the following domains: **(1)** An open source implementation of the FDA-approved type-1 diabetes treatment simulator, **(2)** A recommender system domain, and **(3)** A continuous-state Gridworld with partial observability. Additional results for non-stationary settings and other empirical details are available in the work by Chandak et al. [2].

Figure 4 reinforces the universality of UnO. UnO accurately estimates the entire CDF and a wide range of its parameters: mean, variance, quantile, and CVaR. Perhaps surprisingly, Figure 4 shows that the proposed guaranteed coverage bounds, termed *UnO-CI* here, can be competitive with existing specialized bound, termed *Baseline-CI* here, for the mean and variance. In fact, UnO-CI can often require an order of magnitude less data compared to the specialized bounds for variance; This suggests that the universality of UnO can be beneficial even when only one specific parameter is of interest. Recall that the proposed bootstrap based bounds 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.

Now, without being restricted to the most common and basic parameters, researchers and practitioners can fully characterize the behavior of a policy without having to deploy it. For example, practical experiments in neuroscience where designing new controlled studies could be expensive in terms of time, effort, and money, or for safety-critical applications of RL, like in healthcare, where it is important to assess the potential human-life risk *before* deploying the policy. Further, we have shown how algorithms can perform distributional RL in the completely off-policy setting (without sampling any new data), thereby bringing methods closer to resembling animal learning. Leveraging UnO in the *online* off-policy setting can provide a closer resemblance to animal learning and remains an interesting future direction.

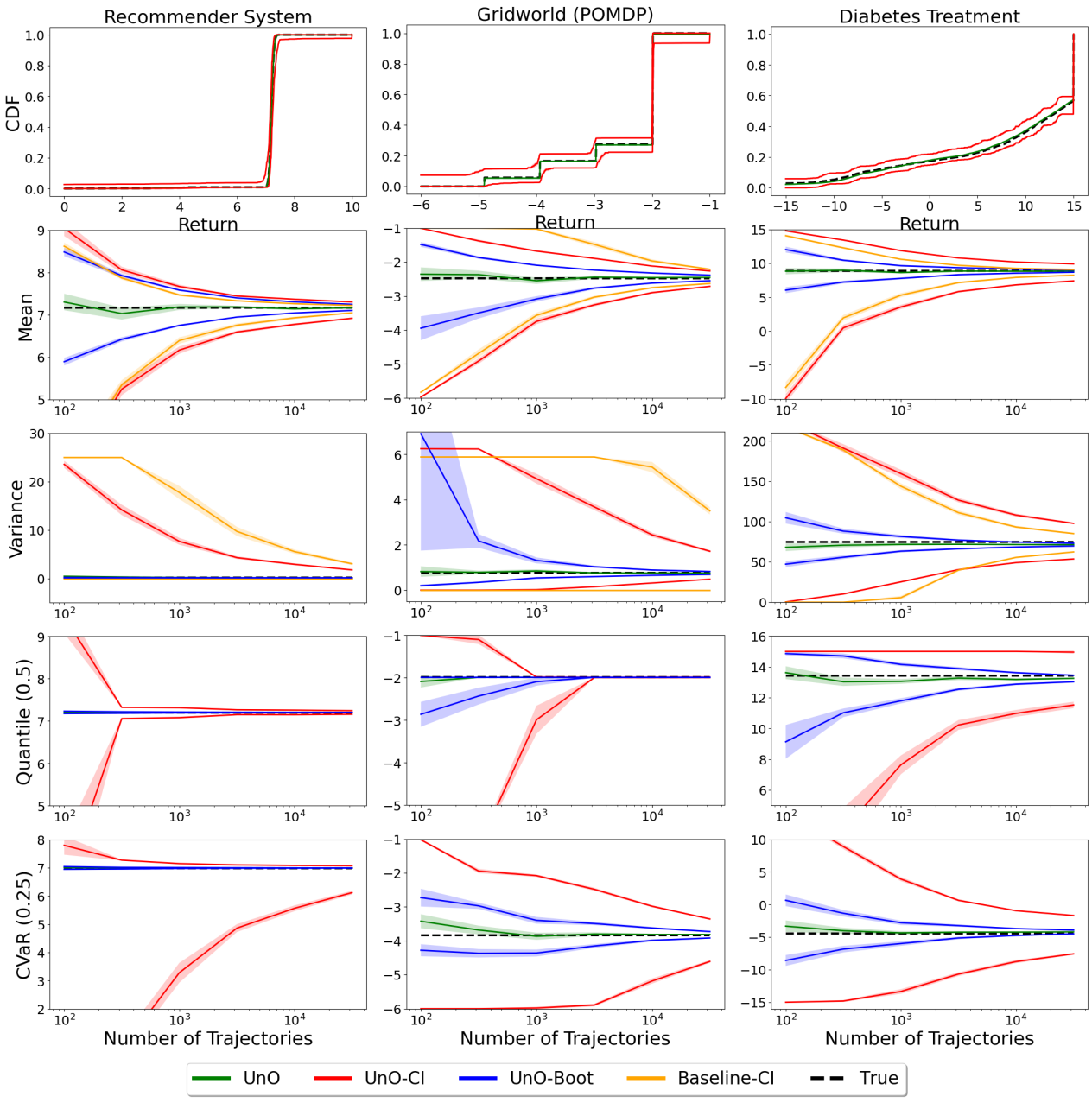

Figure 4: Performance trend of the proposed estimators and bounds on three domains. The black dashed line is the true value of F_π or $\psi(F_\pi)$, green is our UnO estimator, red is our `CI`-based UnO bound, blue is the bootstrap version of our UnO bound, and yellow is the baseline bound for the mean or variance. 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 $\delta = 0.05$.

References

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

[2] Y. Chandak, S. Niekum, B. da Silva, E. Learned-Miller, E. Brunskill, and P. S. Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34, 2021.

[3] 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.