# Conformal Off-Policy Evaluation in Markov Decision Processes

Daniele Foffano*, Alessio Russo* and Alexandre Proutiere

***Abstract*— Reinforcement Learning aims at identifying and evaluating efficient control policies from data. In many real-world applications, the learner is not allowed to experiment and cannot gather data in an online manner (this is the case when experimenting is expensive, risky or unethical). For such applications, the reward of a given policy (the *target* policy) must be estimated using historical data gathered under a different policy (the *behavior* policy). Most methods for this learning task, referred to as Off-Policy Evaluation (OPE), do not come with accuracy and certainty guarantees. We present a novel OPE method based on Conformal Prediction that outputs an interval containing the true reward of the target policy with a prescribed level of certainty. The main challenge in OPE stems from the distribution shift due to the discrepancies between the target and the behavior policies. We propose and empirically evaluate different ways to deal with this shift. Some of these methods yield conformalized intervals with reduced length compared to existing approaches, while maintaining the same certainty level.**

## I. Introduction

In this work, we consider the problem of off-policy evaluation (OPE) in finite time-horizon Markov Decision Processes (MDPs). This problem is concerned with the task of learning the expected cumulative reward of a *target* policy from data gathered under a different *behavior* policy. In fact, OPE has attracted a lot of attention recently [10], [23], [19], [25], [5], [15] since it is particularly relevant in real-world scenarios where the learner is not allowed to experiment and deploy the target policy to infer its value. In these scenarios, testing a new policy in an online manner can be indeed too risky or unethical (e.g., in finance or healthcare).

The main challenge in OPE algorithms stems from the distribution shift of the target and behavior policies. To address this issue, researchers have developed various solutions, often based on Importance Sampling methods (refer to §II and to [29] for a recent survey). Lastly, while existing OPE algorithms sometimes enjoy asymptotic convergence properties, most of them do not come with accuracy and certainty guarantees [25], [26], [7].

To that aim, we are concerned with devising OPE estimators that enjoy non-asymptotic performance guarantees. We leverage techniques from Conformal Prediction (CP) [30], [28], [21], which, directly from the data, allow to build *conformalized* sets that provably includes the true value of the quantity to be estimated with a prescribed level of certainty. Furthermore, CP is a distribution-free method, thus circumventing the burden of estimating a model while providing non-asymptotic guarantees. Due to these desirable properties, CP has been applied with success in many fields, including medicine [14], [33], [16], aerospace engineering [32], finance [31] and safe motion planning [13].

Nevertheless, standard CP assumes to be trained on i.i.d. data, and that at test time the data comes from the same distribution from which the training data was drawn (a.k.a. as *distribution/covariates shift*). This latter assumption is violated in OPE problems, since the training data is gathered using a policy than is different from the target policy to be evaluated. A solution to address the distribution shift is to leverage the concept of *weighted exchangeability* [28], [12].

By exploiting the concept of weighted exchengeability, we study the *conformalized OPE* problem for Markov Decision Processes (MDPs). Our method builds on top of the technique described in [24], which introduces conformalized OPE for contextual bandit models (which can be seen as MDPs with i.i.d. states). Compared to [24], we have to handle additional difficulties, including the inherent dependence in the data (which consists of trajectories of a controlled Markov chain) and the statistical hardness of dealing with the distribution shift when the time horizon grows large.

*Contribution-wise,* we present and empirically evaluate CP algorithms that yield conformalized intervals with reduced length compared to existing approaches, while maintaining the same certainty level. These algorithms are based on the two following new components. (i) Asymmetric score functions: existing CP approaches use symmetric score functions and hence, for our problem, would output conformalized intervals centered on the value of the behavior policy. We introduce asymmetric score functions, so that the CP algorithm yields an interval that efficiently moves its center to follow the distribution shift. In turn, CP with asymmetric score functions results in intervals of smaller size. (ii) We propose methods to address the distribution shift in MDPs.

We finally illustrate the performance of our algorithms numerically on the classical inventory control problem [20]. The experiments demonstrate that indeed our algorithms achieve smaller interval lengths than existing approaches, while retaining the same certainty guarantees.

## II. Related work

### A. Off-Policy Evaluation (OPE)

There are mainly three classes of OPE algorithms in the literature: Direct, Importance Sampling and Doubly Robust Methods. Direct Methods (DMs) learn a model of the system [10], [23] and then evaluate the policy against it. DMs can lead to biased estimators due to a mismatch between the

∗ Equal contribution

Daniele Foffano, Alessio Russo and Alexandre Proutiere are in the Division of Decision and Control Systems of the EECS School at KTH Royal Institute of Technology, Stockholm, Sweden. {foffano,alessior,alepro}@kth.se

model and the true system. Importance Sampling (IS) is a well-known method [19], [25], [5], [15] used to correct the distribution mismatch caused by the discrepancies between the target and the behavior policies by re-weighting the sampled rewards. Still, IS-based algorithms suffer from high variance in long-horizon problems. Doubly Robust (DR) methods combine DMs and IS to obtain more robust estimators [7], [6]. [15] introduce Marginalized Importance Sampling, reducing the variance by applying IS directly on the stationary state-visitation distribution.

The aforementioned approaches only provide an accurate point-wise estimate of the policy value, without quantifying its uncertainty. [1] derived confidence intervals (CIs) using the Central Limit Theorem. In [25], [9], the authors leveraged concentration inequalities to estimate good CIs, which, however, tend to be overly-conservative. For short-horizon problems, [26], [5] approximate CIs for OPE can also be found by means of bootstrapping. [22] derives a non-asymptotic CI using concentration bounds on a kernel-based Q-function.

In [8], the authors derive an asymptotic CI using Double Reinforcement Learning (DRL), also addressing the curse of the horizon. However, the DRL method might not converge in high-dimensional RL tasks, resulting in an asymptotically biased estimator. [3], [23] derive non-asymptotic and asymptotic CIs by approximating the value function with linear functions, but their approaches might lead to a biased estimator if the model assumption is incorrect. [7] derived a CI that involves solving a linear program, but they assume the observations to be i.i.d., whereas transitions are time-dependent in many RL problems.

### B. Conformal Prediction (CP)

CP is a frequentist technique to derive CIs with a specified coverage (*i.e.*, confidence) and a finite number of i.i.d. samples (we refer the reader to [17] for a comprehensive list of CP-related papers). The advantage of CP with respect to other methods is that the provided coverage guarantees are distribution-free and non-asymptotic.

CP for off-policy evaluation has been recently applied to the contextual bandit setting [24], which, in contrast to our work, has no dynamics and no time-dependent data. To address the distribution shift, the authors in [24] use of the weighted exchangeability property, which was previously introduced in [28]. In [2], the authors apply CP to predict the expected value of MDPs trajectories. They consider an online setting where they do not have to deal with the distribution shift.

## III. Preliminaries

### A. Off-policy evaluation in Markov Decision Processes

We consider finite-time horizon MDPs [20]. Such an MDP is defined by a tuple $M = \langle \mathcal{X}, \mathcal{A}, T, q, p, H \rangle$, where $\mathcal{X}$ and $\mathcal{A}$ are the (finite) state and action spaces, respectively. For all $(x, a) \in \mathcal{X} \times \mathcal{A}$, $T(\cdot|x, a)$ and $q(\cdot|x, a)$ denote the distributions of the next state and of the instantaneous reward given that the current state is $x$ and that the decision maker selects action $a$ (for simplicity, we assume that the transition probabilities and the reward distributions are stationary; our results can be easily generalized to non-stationary dynamics and rewards). Finally, $p \in \Delta(\mathcal{S})$ denotes the distribution of the initial state, and $H$ the time horizon.

In off-policy evaluation, we gather data using a behavior policy $\pi^b$, and we wish to estimate the value function of different policy $\pi$. Here again for simplicity, we consider stationary policies: both $\pi^b$ and $\pi$ are mappings between the state space and the set $\Delta(\mathcal{A})$ of distributions over actions. The value function of $\pi$ maps the initial state $x$ to the expected reward gathered under $\pi$ when starting in $x$: $V_H^\pi(x) = \mathbb{E}_\pi[\sum_{t=1}^H r_t | x_1 = x]$, where $r_t \sim q(\cdot|x_t, a_t)$, $a_t \sim \pi(\cdot|x_t)$, and $x_{t+1} \sim T(\cdot|x_t, a_t)$ for $t = 1, \dots, H$.

### B. Standard Conformal Prediction

Conformal Prediction (CP) is a method for distribution-free uncertainty quantification of learning methods, see e.g. [30], [18], [11]. To illustrate how CP works, we consider classical supervised learning tasks and restrict our attention to split CP where the pre-training and the calibration phases are conducted on different datasets. The learner starts with a pre-trained model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ that maps inputs to predicted labels (this model may also consist of upper and lower estimated quantiles if the pre-training procedure corresponds to quantile regression). She also has i.i.d. calibration data $\mathcal{D}_{cal} = \{X_i, Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{X,Y}$. From $\hat{f}$ and $\mathcal{D}_{cal}$, CP constructs for each possible input $x$ a subset $\hat{C}_n(x)$ of possible labels. More precisely, the method proceeds as follows: (i) first a score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is constructed from the model $\hat{f}$ (e.g., it could be the residuals $|y - \hat{f}(x)|$ if $\mathcal{Y} \subset \mathbb{R}$); (ii) the scores of the various calibration samples are computed $V_i = s(X_i, Y_i)$, and (iii) the confidence set is built according to $\hat{C}_n(x) = \{y \in \mathcal{Y} : s(x, y) \leq \eta\}$, where $\eta = \text{Quantile}_{1-\alpha}\left(\frac{1}{n+1}\left(\sum_{i=1}^n \delta_{V_i} + \delta_{\{\infty\}}\right)\right)$. If $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$ are exchangeable, this construction ensures coverage with certainty level $1 - \alpha$:

$$1 - \alpha \leq \mathbb{P}(Y \in \hat{C}_n(X)) \leq 1 - \alpha + \frac{1}{n+1}. \tag{1}$$

## IV. Conformalized Off-Policy Evaluation

Our objective is to get conformalized predictions for the value function of a policy $\pi$, based on training and calibration data gathered under a different behavior policy $\pi_b$. We address this *distribution shift* by extending and improving the techniques developed in [28], [24]. We apply the CP formalism where the input $X$ corresponds to the initial state, and the output $Y$ to $V_H^\pi(X)$. Our method is illustrated in Figure 1. Next, we describe its components in detail. Specifically, (i) we explain how the aforementioned distribution shift can be addressed by weighing scores; (ii) we then discuss the important choice of the score function.

### A. Weighted conformal prediction

As suggested [28], [24], we can handle the distribution shift by weighing the scores using estimates of the likelihood ratio

$$w(x, y) \coloneqq \frac{\mathrm{d}P_{X,Y}^\pi}{\mathrm{d}P_{X,Y}^{\pi^b}}(x, y) = \frac{\mathrm{d}P_{Y|X}^\pi}{\mathrm{d}P_{Y|X}^{\pi^b}}(y|x),$$

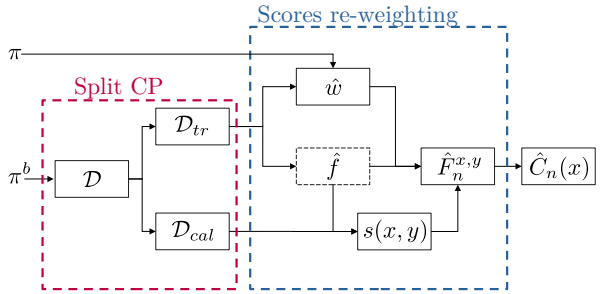


Fig. 1. Conformal prediction for off-policy evaluation. The dataset $\mathcal{D}$ is collected using a behavior policy $\pi^b$, which is then split into the *training* $\mathcal{D}_{tr}$ and *calibration* $\mathcal{D}_{cal}$ datasets. When evaluating a different policy $\pi$, there is a shift in the data distribution, and we need to learn a likelihood ratios $\hat{w}$ to compensate for this shift. The training data is used to learn estimates of the weights $\hat{w}$ and a model $\hat{f}$ used in the computation of the scores. The estimated weights are used as plug-in estimates to re-weight the cumulative distribution function of the scores $\hat{F}_n^{x,y}$, which is then used to compute the conformalized intervals $\hat{C}_n(x)$.

where for any policy $\pi' \in \{\pi, \pi^b\}$, $P^{\pi'}_{X,Y}(x,y) = P^{\pi'}_{Y|X}(y|x)p(x)$ denotes the distribution of the observation $(X,Y)$ under $\pi'$ ($P^{\pi'}_{Y|X}$ is this distribution given $X$), and $p(x)$ is the initial state distribution, which is the same in both cases. The value of a given trajectory $\tau = \{x_1, a_1, r_1, \ldots, x_H, a_H, r_H\}$ is $y = \sum_{t=1}^H r_t$. For any policy $\pi' \in \{\pi, \pi_b\}$, the probability of observing $\tau$ under $\pi'$ given the initial state $x_1 = x$ is:

$$P^{\pi'}(\tau|x) = \pi'(a_1|x)q(r_1|x,a_1)\prod_{t=2}^{H}\pi'(a_t|x_t) \times T(x_t|x_{t-1},a_{t-1})q(r_t|x_t,a_t).$$

Hence the weights can be written as:

$$w(x,y) = \frac{\int \mathbf{1}_{\{y=\sum_{t=1}^H r_t\}} P^{\pi}(\tau|x)\mathrm{d}\tau}{\int \mathbf{1}_{\{y=\sum_{t=1}^H r_t\}} P^{\pi^b}(\tau|x)\mathrm{d}\tau}.$$

We make the following assumption to guarantee that the above weights are always well defined, and that the calibration data is i.i.d.

*Assumption 1: We assume throughout the paper that $P^{\pi}(\cdot|x)$ is absolutely continuous w.r.t. $P^{\pi^b}(\cdot|x)$ for all $x \in \mathcal{X}$. We further assume that calibration data $\mathcal{D}_{cal}$ provides $n$ i.i.d. samples $(X_i, Y_i) \sim P^{\pi^b}_{X,Y}$.*

Then, we can compute the scores $V_i = s(X_i, Y_i)$. For each possible pair $(x,y)$, using the normalized weights, we form the distribution $\hat{F}_n^{x,y} := \sum_{i=1}^n p_i^w(x,y)\delta_{V_i} + p_{n+1}^w(x,y)\delta_\infty$, with

$$p_i^w(x,y) = \begin{cases} \dfrac{w(X_i,Y_i)}{\sum_{j=1}^n w(X_j,Y_j) + w(x,y)} & \text{if } i \le n, \\ \dfrac{w(x,y)}{\sum_{j=1}^n w(X_j,Y_j) + w(x,y)} & \text{if } i = n+1, \end{cases} \tag{2}$$

and the conformalized set

$$\hat{C}_n(x) := \left\{ y \in \mathbb{R} : s(x,y) \le \text{Quantile}_{1-\alpha}\left(\hat{F}_n^{x,y}\right)\right\}. \tag{3}$$

*Proposition 1: Under Assumption 1, for any score function $s$ and any $\alpha \in (0,1)$,*

$$\mathbb{P}^{\pi^b,\pi}\left[Y \in \hat{C}_n(X)\right] \ge 1-\alpha, \tag{4}$$

*where $\mathbb{P}^{\pi^b,\pi}$ accounts for the randomness of $(X,Y) \sim P^{\pi}_{X,Y}$ and that of the data $\mathcal{D}_{cal} = \{X_i, Y_i\}_{i=1}^n$ (with for all $i \in [n]$, $(X_i, Y_i) \sim P^{\pi^b}_{X,Y}$).*

*Proof:* The proof follows that in [24, Proposition 4.1]. The idea relies on the fact that $\{(X_i,Y_i)\}_{i=1}^n \cup (X_{n+1}, Y_{n+1})$ are weighted exchangeable (see Lemma 1 in the appendix), where $(X_{n+1}, Y_{n+1})$ is sampled according to $P^{\pi}_{X,Y}$ and $\{(X_i,Y_i)\}_{i=1}^n$ according to $P^{\pi^b}_{X,Y}$. Then, assume for simplicity that $V_1, \ldots, V_{n+1}$ are distinct almost surely. We define $f$ as the joint distribution of the random variables $\{X_i, Y_i\}_{i=1}^{n+1}$. We also denote $E_z$ as the event of $\{Z_1, \ldots, Z_{n+1}\} = \{z_1, \ldots, z_{n+1}\}$ (where the equality refers to the equality between sets) and let $v_i = s(z_i) = s(x_i, y_i)$. Then, for each $i$ :

$$\mathbb{P}\left[V_{n+1} = v_i \mid E_z\right] = \mathbb{P}\left[Z_{n+1} = z_i \mid E_z\right], = \frac{\sum_{\sigma:\sigma(n+1)=i} f\left(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)}\right)}{\sum_{\sigma} f\left(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)}\right)}$$

Now using the fact that $Z_1, \ldots, Z_{n+1}$ are weighted exchangeable, as in [24] we find that $\mathbb{P}\left[Z_{n+1} = z_i \mid E_z\right] = p_i^w(z_{n+1})$.

Next, just as in [28] we can view:

$$\{V_{n+1} = v_i \mid E_z\} \sim \sum_{i=1}^{n+1} p_i^w(z_{n+1})\,\delta_{v_i}$$

which implies that:

$$\mathbb{P}\left[V_{n+1} \le \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^{n+1} p_i^w(X_{n+1})\,\delta_{v_i}\right) \mid E_z\right] \ge 1-\alpha.$$

Marginalizing over $E_z$ concludes the proof. ∎

Proposition 1 shows that, in absence of data from the target policy, we can still use a shifted CDF of the scores to assess the target policy. The result however relies on the assumption that the weights $w(x,y)$ are known. In practice, we could use the training data to learn these weights, refer to Section V for details. The next proposition quantifies the impact of the error in this estimation procedure on the coverage. Its proof follows the same arguments as those in [24].

*Proposition 2: Assume that the conformalized sets (3) are defined using estimated the weights $\hat{w}(x,y)$ satisfying $\mathbb{E}^{\pi^b}[\hat{w}(X,Y)^r] \le M_r^r < \infty$ for some $r \ge 2$. Define $\Delta_w = \frac{1}{2}\mathbb{E}^{\pi^b}|\hat{w}(X,Y) - w(X,Y)|$. Then*

$$\mathbb{P}^{\pi^b,\pi}\left[Y \in \hat{C}_n(X)\right] \ge 1-\alpha-\Delta_w, \tag{5}$$

*If, in addition, the non-conformity scores $\{V_i\}_{i=1}^n$ have no ties almost surely, then we also have*

$$\mathbb{P}^{\pi^b,\pi}\left[Y \in \hat{C}_n(X)\right] \le 1-\alpha+\Delta_w + cn^{1/r-1},$$

*for some positive constant $c$ depending on $M_r$ and $r$ only.*

*Proof:* The proof is omitted for brevity, since it follows *mutatis mutandis* from that in [24, Proposition 4.2]. ∎

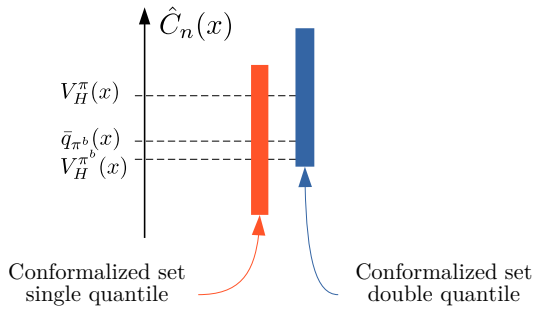


Fig. 2. Symmetry problem. For the original confidence set with one single quantile, and score function $s(x,y) = \max(q_{\alpha_{\text{lo}}}(x) - y, y - q_{\alpha_{\text{hi}}}(x))$, we obtain a set that is symmetric around its middle point $(q_{\alpha_{\text{lo}}}(x) + q_{\alpha_{\text{hi}}}(x))/2$. We can break this symmetry by considering two different score quantiles, one for $q_{\alpha_{\text{lo}}}(x) - y$ and one for $y - q_{\alpha_{\text{hi}}}(x)$, thus leading to a less conservative conformalized set.

### B. Selecting the score function

The choice of the score function critically impacts the size and center of the conformalized sets $\hat{C}_n(x)$. In previous work [21], [24], the pre-training procedure outputs some estimated quantiles $q_{\alpha_{\text{lo}}}(x)$ and $q_{\alpha_{\text{hi}}}(x)$ for the value of the behavior policy with initial state $x$, and the use of the symmetric score function

$$s(x,y) = \max(q_{\alpha_{\text{lo}}}(x) - y, y - q_{\alpha_{\text{hi}}}(x)), \tag{6}$$

is advocated. This choice yields a set $\hat{C}_n(x)$ centered $\bar{q}_{\pi^b}(x) = (q_{\alpha_{\text{lo}}}(x) + q_{\alpha_{\text{hi}}}(x))/2$. Indeed, in view of (3) and (6), there is $\eta(x) \in \mathbb{R}$ such that $\hat{C}_n(x) = [\bar{q}_{\pi^b}(x) - \eta(x), \bar{q}_{\pi^b}(x) + \eta(x)]$ (note that when $n$ grows large, $\eta(x)$ becomes independent of $x$). Having $\hat{C}_n(x)$ centered on the estimated median value for $\pi^b$ is of course very problematic when the values of $\pi^b$ and $\pi$ significantly differ. In this case, the length of $\hat{C}_n(x)$ becomes unnecessarily large. Next we propose methods and score functions that efficiently re-center $\hat{C}_n(x)$ around the value of $\pi$ (instead of $\pi_b$), and that in turn yield much smaller conformalized sets.

*1) Double-quantile score:* a first idea is to break the symmetry of the score function used in [24] by considering the following confidence set

$$\begin{aligned}\hat{C}_n(x) \coloneqq &\left\{y \in \mathbb{R} : q_{\alpha_{\text{lo}}}(x) - y \le \text{Quantile}_{1-\alpha/2}\left(\hat{F}^{x,y}_{n,0}\right)\right\} \\ &\cap \left\{y \in \mathbb{R} : y - q_{\alpha_{\text{hi}}}(x) \le \text{Quantile}_{1-\alpha/2}\left(\hat{F}^{x,y}_{n,1}\right)\right\},\end{aligned} \tag{7}$$

where $\hat{F}^{x,y}_{n,0} \coloneqq \sum_{i=1}^n p_i^w(x,y)\delta_{V_{i,0}} + p_{n+1}^w(x,y)\delta_\infty$ and $\hat{F}^{x,y}_{n,1} \coloneqq \sum_{i=1}^n p_i^w(x,y)\delta_{V_{i,1}} + p_{n+1}^w(x,y)\delta_\infty$, with $V_{i,0} = q_{\alpha_{\text{lo}}}(X_i) - Y_i$ and $V_{i,1} = Y_i - q_{\alpha_{\text{hi}}}(X_i)$. In essence, we separately look at the lower and upper quantiles of the shifted distribution of the scores. A graphical illustration is provided in Fig. 2. The new construction of $\hat{C}_n(x)$ does not affect coverage guarantees:

*Proposition 3: Under Assumption 1, for $\alpha \in (0,1)$ the sets $\hat{C}_n(x)$ in (7) satisfies*

$$\mathbb{P}^{\pi^b,\pi}\left[Y \in \hat{C}_n(X)\right] \ge 1 - \alpha. \tag{8}$$

*Proof:* The proof follows from that of Proposition 1. Assume for simplicity that for a fixed $j \in \{0,1\}$ the values $\{V_{i,j}\}_{i=1}^n$ are distinct almost surely and let $s_0(x,y) = q_{\alpha_{\text{lo}}}(x) - y, s_1(x,y) = y - q_{\alpha_{\text{hi}}}(x)$. As before, we define $f$ as the joint distribution of the random variables $\{X_i, Y_i\}_{i=1}^{n+1}$. Recall that $Z_i = (X_i, Y_i)$, then, we denote by $E_z$ the event that $\{Z_1, \ldots, Z_{n+1}\} = \{z_1, \ldots, z_{n+1}\}$ and let $v_{i,j} = s_j(z_i) = s_j(x_i, y_i)$ for $j \in \{0,1\}$. First, following the previous proposition, we observe that

$$\{V_{n+1,j} = v_{i,j} \mid E_z\} \sim \sum_{i=1}^{n+1} p_i^w(z_{n+1})\,\delta_{v_{i,j}}, \quad j \in \{0,1\}, \tag{9}$$

and, similarly, $\mathbb{P}\left[V_{n+1,j} \le \eta_{1-\alpha/2}(Z_{n+1}) \mid E_z\right] \ge 1 - \alpha/2, \quad j \in \{0,1\}$, where

$$\eta_{1-\alpha/2}(Z_{n+1}) = \text{Quantile}_{1-\alpha/2}\left(\sum_{i=1}^{n+1} p_i^w(Z_{n+1})\,\delta_{v_{i,j}}\right).$$

Let $\hat{C}_{n,j} = \left\{y \in \mathbb{R} : s_j(x,y) \le \text{Quantile}_{1-\alpha/2}\left(\hat{F}^{x,y}_{\eta,j}\right)\right\}$, then $\mathbb{P}^{\pi^b,\pi}\left[Y \notin \hat{C}_{n,j}(X) | E_z\right] \le \alpha/2, j \in \{0,1\}$, from which follows (through a union bound) $\mathbb{P}^{\pi^b,\pi}\left[Y \notin \hat{C}_n(X) | E_z\right] \le \alpha$. We get the claim after marginalizing over $E_z$. ∎

We also obtain the following guarantees in case $w(x,y)$ is replaced by $\hat{w}(x,y)$.

*Proposition 4: Let $\hat{C}_n(x)$ be as in (7) with weights $w(x,y)$ replaced by $\hat{w}(x,y)$. Under the same assumptions as in Proposition 2, we have*

$$\mathbb{P}^{\pi^b,\pi}\left[Y \in \hat{C}_n(X)\right] \ge 1 - \alpha - \Delta_w.$$

*If, in addition, non-conformity scores $\{V_{i,0}\}_{i=1}^n$ and $\{V_{i,1}\}_{i=1}^n$ have no ties almost surely, then we also have*

$$\mathbb{P}^{\pi^b,\pi}\left[Y \in \hat{C}_n(X)\right] \le 1 - \alpha + \Delta_w + cn^{1/r-1},$$

*for some positive constant $c$ depending only on $M_r$ and $r$.*

*Proof:* We take inspiration from [24, Proposition 4.2]. For the sake of notation, we denote the test point by $(X', Y')$ instead of $(X_{n+1}, Y_{n+1})$. We also denote by $\hat{C}_n(x)$ the confidence set $\hat{C}_n(x) \coloneqq C_{0,n}(x) \cap C_{1,n}(x)$, with

$$\begin{aligned}\hat{C}_{0,n}(x) &\coloneqq \left\{y \in \mathbb{R} : q_{\alpha_{\text{lo}}}(x) - y \le \text{Quantile}_{1-\alpha/2}\left(\hat{F}^{x,y}_{\eta,0}\right)\right\}, \\ \hat{C}_{1,n}(x) &\coloneqq \left\{y \in \mathbb{R} : y - q_{\alpha_{\text{hi}}}(x) \le \text{Quantile}_{1-\alpha/2}\left(\hat{F}^{x,y}_{\eta,1}\right)\right\}\end{aligned}$$

where $\hat{F}^{x,y}_{n,0} \coloneqq \sum_{i=1}^n p_i^{\hat{w}}(x,y)\delta_{V_{i,0}} + p_{n+1}^{\hat{w}}(x,y)\delta_\infty$ and $\hat{F}^{x,y}_{n,1} \coloneqq \sum_{i=1}^n p_i^w(x,y)\delta_{V_{i,1}} + p_{n+1}^w(x,y)\delta_\infty$, with $V_{i,0}$ and $V_{i,1}$ are as before.

We first prove the lower bound. Let $\tilde{P}^{\pi}_{X,Y}$ be a probability measure with $\mathrm{d}\tilde{P}^{\pi}_{X,Y}(x,y) = \hat{w}(x,y)\mathrm{d}P^{\pi^b}_{X,Y}(x,y)$. Further, let $\text{TV}(P,Q)$ be the total variation distance between two distributions $P, Q$, and observe that

$$\text{TV}(\tilde{P}^{\pi}, P^{\pi}) = \frac{1}{2}\int |\hat{w}(x,y) - w(x,y)|\mathrm{d}P^{\pi^b}(x,y) = \Delta_w.$$

First, note that from an application of Proposition 3 we have that $\mathbb{P}_{(X,Y)\sim\tilde{P}^{\pi}_{X,Y}}[(Y \in \hat{C}_n(X)] \geq 1-\alpha$. Then, we can use the total variation to bound the difference in probability

$$|\mathbb{P}_{(X,Y)\sim\tilde{P}^{\pi}_{X,Y}}[Y \in \hat{C}_n(X)] - \mathbb{P}_{(X,Y)\sim P^{\pi}_{X,Y}}[Y \in \hat{C}_n(X)]| \leq \Delta_w.$$

Using the triangle inequality we find the lower bound:

$$\mathbb{P}_{(X,Y)\sim P^{\pi}_{X,Y}}[Y \in \hat{C}_n(X)] \geq 1-\alpha-\Delta_w.$$

We now prove the upper bound. The moment assumption on $\hat{w}(x,y)$ guarantees that $\hat{w}$ is bounded a.s. under $P^{\pi^b}$. Then, W.l.o.g., assume $\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)] = 1$. As shown in [24, Proposition 4.2], we have $\mathbb{E}_{(X,Y)\sim\tilde{P}^{\pi}_{X,Y}}[\hat{w}(X,Y)] \leq M_r^2$. Then

$$\begin{aligned}\mathbb{P}[Y' \in &\hat{C}_n(X')] = \mathbb{P}[Y' \in \hat{C}_{0,n}(X') \cap Y' \in \hat{C}_{1,n}(X')],\\ &\leq \min\left[\mathbb{P}(Y' \in \hat{C}_{0,n}(X')), \mathbb{P}(Y' \in \hat{C}_{1,n}(X'))\right].\end{aligned}$$

The rest of the proof follows as in [24, Proposition 4.2], where we note that $\mathbb{P}[Y' \in \hat{C}_{i,n}(X')] \leq 1-\alpha+cn^{1/r-1}, i \in \{1,2\}$, and thus $\mathbb{P}_{(X,Y)\sim\tilde{P}^{\pi}_{X,Y}}[Y' \in \hat{C}_n(X')] \leq 1-\alpha+cn^{1/r-1}$. Using the triangle inequality on

$$|\mathbb{P}_{(X,Y)\sim\tilde{P}^{\pi}_{X,Y}}[Y \in \hat{C}_n(X)] - \mathbb{P}_{(X,Y)\sim P^{\pi}_{X,Y}}[Y \in \hat{C}_n(X)]| \leq \Delta_w.$$

we conclude that $\mathbb{P}_{(X,Y)\sim P^{\pi}_{X,Y}}[Y' \in \hat{C}_n(X')] \leq 1-\alpha+\Delta_w+cn^{1/r-1}$. ∎

*2) Shifted values:* a second idea is to simply shift the values of the behavior policy $\pi^b$ using the likelihood ratios $w(x,y)$, as one would in important sampling methods. This can be done by simply using $s(x,y) = y$. This choice of score function makes sense intuitively: if we are interested in the value of the target policy $\pi$, then we may look at the shifted distribution of the values of the behavior policy.

We may also combine this choice with the double-quantile idea and construct $\hat{C}_n(x)$ as

$$\hat{C}_n(x) = \hat{C}_{n,0}(x) \cap \hat{C}_{n,1}(x), \tag{10}$$

where $\hat{C}_{n,0} = \left\{y \in \mathbb{R} : y \geq \text{Quantile}_{\alpha/2}\left(\hat{F}_n^{x,y}\right)\right\}$ and $\hat{C}_{n,1} = \left\{y \in \mathbb{R} : y \leq \text{Quantile}_{1-\alpha/2}\left(\hat{F}_n^{x,y}\right)\right\}$. Propositions 3 and 4 also hold for this choice.

## V. Offline estimation of the likelihood ratios

In this section, we present various ways to estimate the likelihood ratios $w(x,y)$, and discuss their pros and cons.

### A. Monte-Carlo method

To estimate $w(x,y)$, we need to compute $P^{\pi}_{X,Y}(x,y)$ and $P^{\pi^b}_{X,Y}(x,y)$. Recall that the likelihood ratio is equal to

$$w(x,y) = \frac{\int \mathbf{1}_{\{y=\sum_{t=1}^H r_t\}} P^{\pi}(\tau|x)\mathrm{d}\tau}{\int \mathbf{1}_{\{y=\sum_{t=1}^H r_t\}} P^{\pi^b}(\tau|x)\mathrm{d}\tau},$$

where $\tau$ is a trajectory of length $H$. Since $P^{\pi}(\tau|x)$ (sim. $P^{\pi^b}(\tau|x)$) depends on the transition kernel $T$ and the reward distribution $q$, one needs to estimate these distributions from the data. We may proceed as follows:

1) We use the training data $\mathcal{D}_{tr}$ to compute an estimate $(\hat{T},\hat{q})$ of $(T,q)$ (through maximum likelihood).
2) Compute an estimate of $\hat{w}(x,y)$ through Monte-Carlo sampling:

$$\hat{w}(x,y) = \frac{(1/h)\sum_{k=1}^h \mathbf{1}_{\{y=\sum_{t=1}^H r_t^{(k)}\}}}{(1/h)\sum_{k=1}^h \mathbf{1}_{\{y=\sum_{t=1}^H r_t^{(k)'}\}}}, \tag{11}$$

where $r_t^{(k)}$ and $r_t^{(k)'}$ are sequences of rewards generated, respectively, by starting in $x$ and following $\pi$ and $\pi^b$, and $h$ is the number of Monte Carlo samples. These trajectories are generated using $\hat{T}$ and $\hat{q}$, estimated in the previous step.

This approach has various shortcomings. First it requires us to estimate the model $(T,q)$. Then it forces us to generate a large number of trajectories, which is heavy computationally. Finally, the term $\mathbf{1}_{\{y=\sum_{t=1}^H r_t\}}$ is going to be 0 most of the times. A possible way to alleviate this issue consists in not including the last reward in the trajectory $\tau$. This implies that we replace $\mathbf{1}_{\{y=\sum_{t=1}^H r_t\}}$ by $\hat{q}(y-\sum_{n=1}^{H-1} r_n|x_H,a_H)$. As it turns out, this naive Monte-Carlo method, used with success in simple scenarios (contextual bandits [24]), does not work in MDPs.

### B. Empirical and gradient-based methods

Next we present an alternative and more scalable way to estimate the weights $w(x,y)$ from the training dataset $\mathcal{D}_{tr}$. We make use of the following simple rewriting of the likelihood ratio (also suggested in [24]):

$$\begin{aligned}w(x,y) &= \frac{P^{\pi}_{X,Y}(x,y)}{P^{\pi^b}_{X,Y}(x,y)},\\ &= \int \frac{P^{\pi}_{X,Y}(x,y)}{P^{\pi^b}_{X,Y}(x,y)} \frac{P^{\pi^b}_{\tau|X,Y}(\tau|x,y)}{P^{\pi^b}_{\tau|X,Y}(\tau|x,y)} P^{\pi}_{\tau|X,Y}(\tau|x,y)\mathrm{d}\tau,\\ &= \int \frac{P^{\pi}_{X,Y,\tau}(x,y,\tau)}{P^{\pi^b}_{X,Y,\tau}(x,y,\tau)} P^{\pi^b}_{\tau|X,Y}(\tau|x,y)\mathrm{d}\tau,\\ &= \mathbb{E}_{\tau\sim P^{\pi^b}_{\tau|X=x,Y=y}}\left[\frac{P^{\pi}_{X,Y,\tau}(x,y,\tau)}{P^{\pi^b}_{X,Y,\tau}(x,y,\tau)}\right].\end{aligned}$$

Next, observe that:

$$\frac{P^{\pi}_{X,Y,\tau}(x,y,\tau)}{P^{\pi^b}_{X,Y,\tau}(x,y,\tau)} = \frac{P(y|x,\tau)P^{\pi}(\tau|x)}{P(y|x,\tau)P^{\pi^b}(\tau|x)} = \frac{\prod_{t=1}^H \pi(a_t|x_t)}{\prod_{t=1}^H \pi^b(a_t|x_t)}.$$

Hence, learning $w$ amounts to learning the following expectation:

$$w(x,y) = \mathbb{E}_{\tau\sim P^{\pi^b}_{\tau|X=x,Y=y}}\left[\frac{\prod_{t=1}^H \pi(a_t|x_t)}{\prod_{t=1}^H \pi^b(a_t|x_t)}\right]. \tag{12}$$

To this aim, we propose the following two approaches.

*1) Empirical estimator:* this method applies to the case $x$ and $y$ belong to some finite spaces $\mathcal{X}$ and $\mathcal{Y}$ only. In this case, we can directly estimate $w(x,y)$ from the training data $\mathcal{D}_{tr}$ by simply computing

$$\hat{w}(x,y) = \frac{1}{N(x,y)} \sum_{\tau^i \in \mathcal{D}_{tr}(x,y)} \frac{\prod_{t=1}^{H} \pi(a_t^{(i)}|x_t^{(i)})}{\prod_{t=1}^{H} \pi^b(a_t^{(i)}|x_t^{(i)})}, \quad (13)$$

where the training data $\mathcal{D}_{tr}$ consists of $m$ trajectories generated under $\pi_b$, the $i$-th trajectory in this dataset is $\tau_i = (x_t^{(i)}, a_t^{(i)}, r_t^{(i)})_{t=1}^{H}$, $\mathcal{D}_{tr}(x,y)$ are trajectories with initial state and the accumulated reward $x$ and $y$, respectively, and $N(x,y) = |\mathcal{D}_{tr}(x,y)|$. When the likelihood ratios are bounded, we can quantify the accuracy of the above estimates using standard concentration results:

*Proposition 5: Let* $(\varepsilon, \delta) \in (0,1)$. *Assume the ratio* $\prod_{t=1}^{H} \pi(a_t|x_t) / \prod_{t=1}^{H} \pi^b(a_t|x_t)$ *to be bounded in* $[m, M]$ *for all possible trajectories of horizon* $H$ *generated under* $\pi^b$. *If* $\min_{x,y} N(x,y) \geq \frac{(M-m)^2}{2\varepsilon^2} \ln \frac{2|\mathcal{X}||\mathcal{Y}|}{\delta}$, *then*

$$\mathbb{P}^{\pi^b}[|\hat{w}(X,Y) - w(X,Y)| > \varepsilon] < \delta.$$

*Furthermore, we also have* $\Delta_w \leq \frac{(M-m)|\mathcal{X}||\mathcal{Y}|\sqrt{\pi}}{2\sqrt{2 \min_{x,y} N(x,y)}}$.

*Proof:* The proof is a simple application of Hoeffding's inequality: $\mathbb{P}[|\hat{w}(X,Y) - w(X,Y)| > \varepsilon] < \sum_{x,y} 2e^{\frac{-2N(x,y)\varepsilon^2}{(M-m)^2}} \leq 2|\mathcal{X}||\mathcal{Y}|e^{\frac{-2\varepsilon^2 \min_{x,y} N(x,y)}{(M-m)^2}}$, where we made use also of a union bound over $\mathcal{X} \times \mathcal{Y}$. Then, if we choose $\min_{x,y} N(x,y) \geq \frac{(M-m)^2}{2\varepsilon^2} \ln \frac{2|\mathcal{X}||\mathcal{Y}|}{\delta}$ then $\mathbb{P}[|\hat{w}(X,Y) - w(X,Y)| > \varepsilon] < \delta$. Regarding the inequality, we find that $\mathbb{E}[|\hat{w}(X,Y) - w(X,Y)|] \leq 2|\mathcal{X}||\mathcal{Y}| \int_0^\infty e^{\frac{-2\varepsilon^2 \min_{x,y} N(x,y)}{(M-m)^2}} \mathrm{d}\varepsilon = \frac{(M-m)|\mathcal{X}||\mathcal{Y}|\sqrt{\pi}}{\sqrt{2 \min_{x,y} N(x,y)}}$, and thus $\Delta_w \leq \frac{(M-m)|\mathcal{X}||\mathcal{Y}|\sqrt{\pi}}{2\sqrt{2 \min_{x,y} N(x,y)}}$. ■

The quantities $M$ and $m$ are usually function of the horizon $H$ (in general one can choose $m = 0$). For example, in case $\mathcal{A}$ is finite, we obtain:

- If $\pi^b$ is uniform over $\mathcal{A}$, then an upper bound $M$ is given by $|\mathcal{A}|^H$, and $m = (|\mathcal{A}| \min_{x,a} \pi(a|x))^H$.
- In case $\pi^b$ and $\pi$ are convex mixtures of a uniform distribution with another deterministic policy $\hat{\pi}$, for example $\pi(a|x) = \frac{\epsilon}{|\mathcal{A}|} + (1-\epsilon)\mathbf{1}_{\{a=\hat{\pi}(x)\}}$ (sim. $\pi^b$ with $\epsilon^b$), for some $\epsilon, \epsilon^b \geq 0$, then one can choose $M, m$ as

$$M^{1/H} = \max\left(\frac{\epsilon}{\epsilon^b}, \frac{(1-\epsilon)+\epsilon/|\mathcal{A}|}{(1-\epsilon^b)+\epsilon^b/|\mathcal{A}|}\right),$$
$$m^{1/H} = \min\left(\frac{\epsilon}{\epsilon^b}, \frac{(1-\epsilon)+\epsilon/|\mathcal{A}|}{(1-\epsilon^b)+\epsilon^b/|\mathcal{A}|}\right).$$

See also Fig. 3 for an example of the scaling of $M - m$.

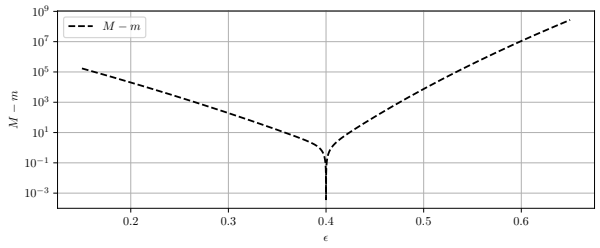


Fig. 3. An example of the difference $M - m$ for the case of a convex mixture, with $|\mathcal{A}| = 10, H = 40$ and $\epsilon^b = 0.4$.

**Algorithm 1** Conformal Off-Policy Evaluation in MDPs

**Require:** Datasets $\mathcal{D}_{tr}, \mathcal{D}_{cal}$; target coverage $\alpha$; policies $(\pi^b, \pi)$; score function $s$; test input $x^{\text{test}}$.

1: Use $\mathcal{D}_{tr}$ to learn the quantiles $q_{\alpha_{\text{lo}}}(x)$ and $q_{\alpha_{\text{hi}}}(x)$, as well as the weight $\hat{w}(x,y)$ using either the empirical estimator or the gradient-based method.
2: Compute $\hat{F}_n^{x,y}$ and the conformalized set $\hat{C}_n(x^{\text{test}})$ using $\hat{w}(x,y)$ and the scores derived from the dataset $\mathcal{D}_{cal}$ using either (3) or (7) or (10).

**Return** $\hat{C}_n(x^{\text{test}})$

In general, we see that the dependency on $H$ is mild when $\pi$ and $\pi_b$ that are somehow similar. As a future research direction, we could investigate possible ways to alleviate the impact of $H$ (for example, by looking at the stationary rewards of the MDP, as in [15]).

*2) Gradient method:* an alternative approach is to notice that $w$, as suggested in [24], can be seen as the solution of a MSE minimization problem. Indeed, $w$ solves the following problem:

$$\min_f \mathbb{E}_{(X,Y,\tau)\sim P_{X,Y,\tau}^{\pi^b}} \left[\left(\frac{\prod_{t=1}^{H} \pi(a_t|x_t)}{\prod_{t=1}^{H} \pi^b(a_t|x_t)} - f(X,Y)\right)^2\right]. \quad (14)$$

Therefore, given some function approximator $f_\theta$ parametrized by $\theta$, we can minimize over $\theta$ the following empirical risk:

$$\frac{1}{m} \sum_{\tau^i \in \mathcal{D}_{tr}} \left(\frac{\prod_{t=1}^{H} \pi(a_t^{(i)}|x_t^{(i)})}{\prod_{t=1}^{H} \pi^b(a_t^{(i)}|x_t^{(i)})} - f_\theta\left(x_1^{(i)}, \sum_{t=1}^{H} r_t^{(i)}\right)\right)^2.$$

As one would expect, this method still suffers from a large variance. For large horizons, it becomes quite difficult to learn the ratio of probabilities, especially when the two policies are extremely different. In fact for large $H$, in case the two policies are different, then it is likely that the ratio of action probabilities is 0 most of the time, with very few values different from 0 that tend to be extremely large. This makes the training procedure difficult, since most function approximators will just learn to output 0.

### C. Algorithm

To conclude this section, we present a generic sketch of our proposed algorithm, see Algorithm 1 for a pseudo-code. Following the split conformal prediction method, the algorithm first leverages the training data $\mathcal{D}_{tr}$ to estimate the quantiles of the value of $\pi_b$ and the weights $w$. It then uses the calibration data $\mathcal{D}_{cal}$ to compute the non-conformity scores. Using $\hat{w}$ as a plug-in estimate in the re-weighted

scores distribution $\hat{F}_n^{x,y}$, the algorithm can finally build the conformal prediction set $\hat{C}_n(x^{\text{test}})$.

## VI. Numerical Results

We evaluate our algorithms on the inventory problem [20], which can be modelled as an MDP with finite state and action spaces. We assume the behavior and target policies $(\pi, \pi^b)$ to be known, and to be $(\epsilon, \epsilon^b)-$greedy with respect to the optimal policy $\pi^\star$. For example, for $\pi$, this means that for all $(x, a)$,

$$\pi(a|x) = \frac{\epsilon}{|\mathcal{A}|} + (1-\epsilon)\mathbf{1}_{\{a=\pi^\star(x)\}},$$

and similarly for $\pi^b$ with $\epsilon^b$. The optimal policy $\pi^\star$ was computed by solving the infinite time-horizon discounted MDP, with discount factor $\gamma = 0.99$. For each method, we evaluate the prediction interval for the cumulative return of the target policy $\pi$ with different values of $\epsilon$, while the behavior policy $\pi^b$ has $\epsilon^b = 0.4$. By considering different values of $\epsilon$ for the target policy, we are able to observe how the coverage and interval length vary with respect to the distance between the target and the behavior policies.

### A. Environment

The inventory control problem is modelled as follows: an agent manages an inventory of size $N$ while facing a stochastic demand for what is stored in it. At each round, the agent must choose how many items to buy to meet the upcoming order for the next day. The action set is the same for every state, i.e. $\mathcal{A} = [0, N]$. We define the cost of buying $a$ items as $k\mathbf{1}_{\{a>0\}} + c(\min(N, x_t + a) - x_t)$, where $k > 0$ is the fixed cost for a single order and $c > 0$ is the cost of a single unit bought. At each round, the agent earns a quantity $pl$, where $p$ is the price of a single item and $l$ is the number of items sold. Finally, the agent has to pay a cost $zn$ for storing $n > 0$ items, with $z > 0$ and $p > z$. The order $o_t$ is sampled from a Poisson distribution with rate $\lambda$. The next state is computed according to $x_{t+1} = \max(0, \min(N, x_t + a_t) - o_{t+1})$, while the reward is the sum of the costs and earnings listed above, i.e., $r(x_t, a_t, x_{t+1}) = -k\mathbf{1}_{\{a_t>0\}} - zx_t - c(\min(N, x_t + a_t) - x_t) + p\max(0, \min(N, x_t + a_t) - x_{t+1})$. Note that here, the rewards are deterministic but depend on the next state – we can easily verify that all our results naturally extend to this setting. We considered two instances of the inventory environment when evaluating our algorithm. In the first one we chose the following parameters: $N = 10$, $k = 1$, $c = 2$, $z = 2$, $p = 4$, $\lambda = 10$. For the second one, we modified the parameters such that: $k = 3$, $\lambda = 6$. So in the second instance, the agent was penalized more for making a single order (i.e., higher $k$) and the demand rate $\lambda$ was decreased.

### B. Algorithm details

We consider three different implementations of our algorithm: the first using the classical pinball score function [21], [24], the second using the double quantile method and finally the shifted values method with double quantile.

*1) Pinball score function:* this method adapts the algorithm presented in [24] to our setting (which is described in Section IV-B). We use the training dataset $\mathcal{D}_{tr}$ to also learn two quantile networks $\hat{q}_{\alpha_{\text{lo}}}$ and $\hat{q}_{\alpha_{\text{hi}}}$, with $\alpha_{\text{lo}} = \alpha/2, \alpha_{\text{hi}} = 1 - \alpha/2$ (where $\alpha$ is the coverage parameter). The two functions are estimated using quantile regression and are modelled using two neural networks with two hidden layers of 64 nodes and ReLU activation functions. For this approach, the score function used is $s(x, y) = \max\left(y - \hat{q}_{\alpha_{\text{hi}}}, \hat{q}_{\alpha_{\text{lo}}} - y\right)$. Once we have computed the empirical CDF of the scores $\hat{F}_n^{x,y}$, the confidence set is obtained using (3).

*2) Double Quantile (DQ) method:* Here we apply the method in IV-B.1. In this method, we introduce two score functions

$$s_0(x, y) = \hat{q}_{\alpha_{\text{lo}}}(x) - y$$
$$s_1(x, y) = y - \hat{q}_{\alpha_{\text{hi}}}(x),$$

where $\hat{q}_{\alpha_{\text{lo}}}$ and $\hat{q}_{\alpha_{\text{hi}}}$ are the same networks as in the previous method. Lastly, the confidence set is computed using (7).

*3) Shifted Values (SV) with double quantile method:* Here we consider a score function that allows us to shift the values of the behavior policy $s(x, y) = y$, as explained in IV-B.2, and compute the confidence set according to (10).

### C. Baseline: Quantile Estimation through Importance Sampling with Bootstrap (QIS-Bootstrap)

We compare the conformal prediction method developed in this work to quantile estimation through importance sampling [4] with bootstrap. Importance sampling (IS) has been widely used as a variance reduction technique in statistical methods, but in our case it can be used to perform off-policy evaluation as in [19], [26]. However, compared to [19], [26] that try to estimate the mean value of the target policy $\pi$, we use the IS technique to estimate the $(\alpha_{\text{lo}}, \alpha_{\text{hi}})$-quantiles of the value of $\pi$. The key insight is that $q_\alpha^\pi(x)$, the $\alpha$-quantile of $\pi$ in $x$, can be estimated using the calibration data $\mathcal{D}_{cal}$ and the likelihood ratio $w(x, y)$ through the following expression

$$q_\alpha^\pi(x) = \text{Quantile}_\alpha\left(\sum_{y \in \mathcal{I}(x)} \frac{w\left(x, y\right)}{\sum_{y' \in \mathcal{I}(x)} w\left(x, y'\right)} \delta_y\right),$$

where $\mathcal{I}(x) = \{y \in \mathcal{D}_{cal} : x_1 = x\}$, *i.e.*, we only consider the cumulative rewards of the trajectories in $\mathcal{D}_{cal}$ that start in $x$.

The inner term can be seen as an empirical estimator of $F_x^\pi(y) = \mathbb{E}_{Y \sim P^\pi(Y|X=x)}[\mathbf{1}_{\{Y \le y\}}] = \mathbb{E}_{Y \sim P^{\pi^b}(Y|X=x)}[w(x, Y)\mathbf{1}_{\{Y \le y\}}]$, the CDF of the values of $\pi$ in $x$ (note that the normalization factor does not affect the outcome, see also [4]). Since $w(x, y)$ is unknown, we replace it by $\hat{w}(x, y)$.

Next, rather than using the estimate $q_\alpha^\pi$ directly, to obtain a better estimate we use bootstrapping [27] to estimate a confidence interval around the $\alpha$-quantile, obtaining a high-confidence interval $(q_{\alpha-}^\pi, q_{\alpha+}^\pi)$ and then taking the median point $\bar{q}_\alpha(x) \coloneqq (q_{\alpha-}^\pi + q_{\alpha+}^\pi)/2$. Finally, the confidence set for the value of $\pi$ is simply given by

$$\hat{C}_n(x) = [\bar{q}_{\alpha_{\text{lo}}}(x), \bar{q}_{\alpha_{\text{hi}}}(x)]. \tag{15}$$

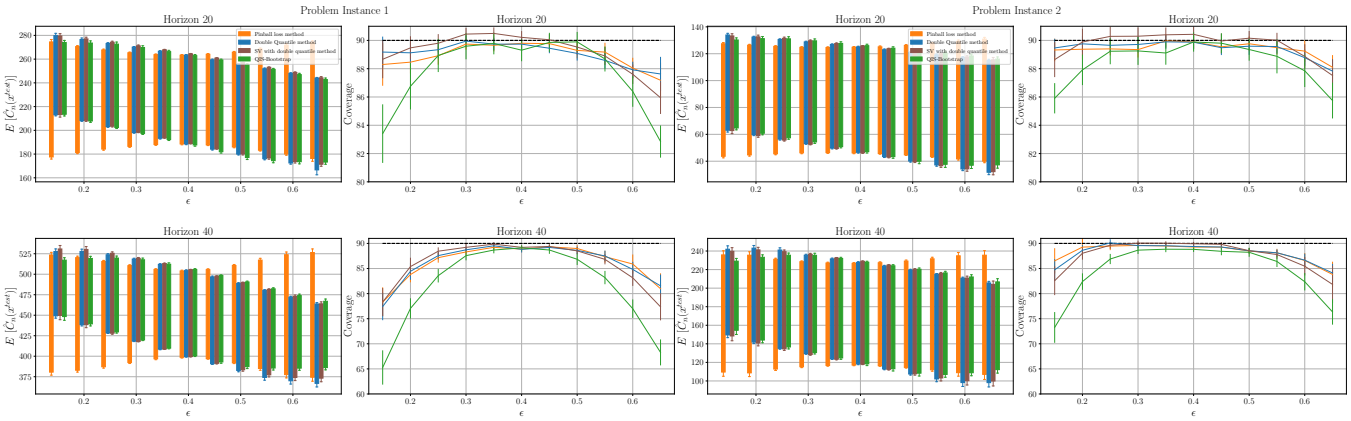


Fig. 4. Results for the inventory control problem for $H = 20, 40$, with target coverage of 90%. The policy $\pi^b$ is $\epsilon^b-$greedy w.r.t. $\pi^\star$ (an optimal discounted policy with discount factor $\gamma = 0.99$), with $\epsilon^b = 0.4$. We evaluated a target policy $\pi$ that is $\epsilon$-greedy w.r.t. $\pi^\star$, with varying $\epsilon$. The four plots on the left are the results corresponding to the first instance of the Inventory Problem, while on the right we present the results for the second instance (both described in section VI-A). The boxplots show average conformalized intervals for the various methods (whiskers indicate 95% confidence intervals for the minimum and the maximum). The line plots depict the obtained coverage level (bars indicate 95% confidence intervals).

It is important to remember that there is no coverage guarantees for this set $\hat{C}_n(x)$.

### D. Results and discussion

In Figure 4, we show the results of our methods in the Inventory Problem for horizons 20 and 40 (for both the problem instances defined in VI-A), where results are averaged over 30 runs. Recall that the policy $\pi^b$ is $\epsilon^b-$greedy w.r.t. $\pi^\star$, with $\epsilon^b = 0.4$, while $\pi$ is $\epsilon$-greedy w.r.t $\pi^\star$, with $\epsilon$ varying in $[0.15, 0.65]$. The target level of coverage was chosen as $1 - \alpha = 90\%$ (depicted as the dashed black line in the plots of the second column). We evaluated our algorithms using the empirical estimate of $\hat{w}$ (see V-B.1) against the QIS-Bootstrap baseline method in Section VI-C.

*1) Conformalized intervals:* the boxplots illustrate the conformalized interval obtained for each method. For each run, method, and value of $\epsilon$, we evaluated the confidence interval across 2000 tests-points $x^{\text{test}}$ sampled from $p(x)$, and averaged the corresponding minimum and maximum values of the confidence set $\hat{C}_n(x^{\text{test}})$. The whiskers indicate 95% confidence interval for the minimum and the maximum. As mentioned in Section IV-B, we observe that the pinball method yields an interval that enlarges/shrinks symmetrically around a fixed point. As a consequence, the interval becomes larger to maintain the desired coverage when the target policy $\pi$ becomes really different than $\pi^b$ (i.e., $\epsilon$ is different than $\epsilon^b = 0.4$). Instead, with the proposed double quantile method, the interval is shifted depending on how far the target policy $\pi$ is w.r.t. $\pi^b$, leading to smaller intervals even when the policies are far from each other. The intervals estimated by the QIS-Bootstrap method match the ones of our new score functions when $\pi$ is close to $\pi^b$. However, when the policies are far from each other, the estimated interval is too conservative (i.e., too small and off-centred), which reflects in the coverage level of the algorithm, quickly degrading as $\pi$ moves away from $\pi^b$, for both horizons.

*2) Coverage:* the line plots illustrate the achieved coverage, averaged over 30 runs (bars indicate 95% confidence interval). All the proposed conformalized methods achieved better levels of coverage than QIS-Bootstrap, as one would expect. For horizon $H = 20$, the pinball method can maintain the desired level of coverage for all the epsilons at the expense of the interval length, while the new methods achieve a better level of coverage with a smaller interval size. For a larger horizon ($H = 40$), we can see that the coverage of the QIS-Bootstrap method degrades very rapidly, maintaining the desired level only for $\pi \approx \pi^b$.

*3) Discussion and future work:* Some of the methods discussed to estimate the likelihood ratio $w(x, y)$ were not used in our numerical experiments. This is mostly due to computational challenges: as we previously mentioned, the computational complexity of the Monte-Carlo method vastly exceeds the complexity of the other methods (empirical estimator and gradient method), while the gradient method has several difficulties in learning the likelihood ratios for values of $(\epsilon, \epsilon^b)$ that greatly differ. We plan to investigate how to efficiently learn the likelihood ratios using neural networks. Finally, we note that one may try conformalize the QIS-Bootstrap method in Section VI-C to have a more fair comparison.

## VII. Conclusion

In this work, we considered the *offline off-policy evaluation* problem in finite time-horizon Markov Decision Processes. Using Conformal Prediction (CP) techniques, we developed methods to construct conformalized intervals that include the true reward of the target policy with a prescribed level of certainty. Some of the challenges addressed in this paper include dealing with time-dependent data, as well as addressing the distribution shift between the behavior policy and the target policy. Furthermore, we proposed improved CP methods that allow to obtain intervals with significantly reduced length when compared to existing CP methods,

while retaining the same certainty guarantees. We conclude with numerical results on the inventory control problem that demonstrated the efficiency of our methods. Several interesting research directions have been mentioned in the text, of which, the most significant, consists in improving the estimation of the likelihood ratio characterizing the distribution shift.

## REFERENCES


[1] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

[2] Thomas G Dietterich and Jesse Hostetler. Conformal prediction intervals for markov decision process trajectories. *arXiv preprint arXiv:2206.04860*, 2022.

[3] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[4] Peter W Glynn et al. Importance sampling for monte carlo estimation of quantiles. In *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*, pages 180–185. Citeseer, 1996.

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

[6] Nan Jiang and Jiawei Huang. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758, 2020.

[7] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *The Journal of Machine Learning Research*, 21(1):6742–6804, 2020.

[8] Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 2022.

[9] Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pages 640–648. PMLR, 2021.

[10] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[11] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

[12] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

[13] Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. Safe planning in dynamic environments using conformal prediction. *arXiv preprint arXiv:2210.10254*, 2022.

[14] Martin Lindh, Anders Karlén, and Ulf Norinder. Predicting the rate of skin penetration using an aggregated conformal prediction framework. *Molecular Pharmaceutics*, 14(5):1571–1576, 2017.

[15] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

[16] Charles Lu, Ken Chang, Praveer Singh, and Jayashree Kalpathy-Cramer. Three applications of conformal prediction for rating breast density in mammography. *arXiv preprint arXiv:2206.12008*, 2022.

[17] Valery Manokhin. Awesome conformal prediction, April 2022. "If you use Awesome Conformal Prediction, please cite it as below.".

[18] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

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

[20] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[21] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[22] Chengchun Shi, Runzhe Wan, Victor Chernozhukov, and Rui Song. Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pages 9580–9591. PMLR, 2021.

[23] Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):765–793, 2022.

[24] Muhammad Faaiz Taufiq, Jean-Francois Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. *arXiv preprint arXiv:2206.04405*, 2022.

[25] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[26] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388. PMLR, 2015.

[27] Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1), 1993.

[28] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[29] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning, 2022.

[30] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[31] Wojciech Wisniewski, David Lindsay, and Sian Lindsay. Application of conformal prediction interval estimations to market makers' net positions. In *Conformal and Probabilistic Prediction and Applications*, pages 285–301. PMLR, 2020.

[32] Zepu Xi, Xuebin Zhuang, and Hongbo Chen. Conformal prediction for hypersonic flight vehicle classification. In *Conformal and Probabilistic Prediction with Applications*, pages 118–206. PMLR, 2022.

[33] Xianghao Zhan, Zhan Wang, Meng Yang, Zhiyuan Luo, You Wang, and Guang Li. An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction. *Measurement*, 158:107588, 2020.