2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Conformal Off-Policy Prediction for Multi-Agent Systems

Tom Kuipers*, Renukanandan Tumu*, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti

Abstract— **Off-Policy Prediction (OPP), i.e., predicting the outcomes of a target policy using only data collected under a nominal (behavioural) policy, is a paramount problem in data-driven analysis of safety-critical systems where the deployment of a new policy may be unsafe. To achieve dependable off-policy predictions, recent work on Conformal Off-Policy Prediction (COPP) leverage the conformal prediction framework to derive prediction regions with probabilistic guarantees under the target process. Existing COPP methods can account for the distribution shifts induced by policy switching, but are limited to single-agent systems and scalar outcomes (e.g., rewards). In this work, we introduce *MA-COPP*, the first conformal prediction method to solve OPP problems involving multi-agent systems, deriving joint prediction regions for all agents' trajectories when one or more "ego" agents change their policies. Unlike the single-agent scenario, this setting introduces higher complexity as the distribution shifts affect predictions for all agents, not just the ego agents, and the prediction task involves full multi-dimensional trajectories, not just reward values. A key contribution of MA-COPP is to avoid enumeration or exhaustive search of the output space of agent trajectories, which is instead required by existing COPP methods to construct the prediction region. We achieve this by showing that an over-approximation of the true *joint prediction region* (JPR) can be constructed, without enumeration, from the maximum density ratio of the JPR trajectories. We evaluate the effectiveness of MA-COPP in multi-agent systems from the PettingZoo library and the F1TENTH autonomous racing environment, achieving nominal coverage in higher dimensions and various shift settings.**

I. Introduction

In reinforcement learning, off-policy prediction (OPP) [1] is the problem of predicting some outcome (e.g., reward, performance) of a given policy—the *target policy*—using only data collected under a different policy—the *behavioural policy*. Such a problem is relevant in data-driven analysis of cyber-physical systems, wherein we leverage observations/executions of the system rather than a mechanistic model, where such a model may be unavailable or just unreliable. OPP is motivated by safety-critical applications such as robotics [2] and healthcare [3], where evaluating the target policy on the real system may be too risky or unethical.

A naïve approach to solving the OPP problem consists of learning an (input-conditional) model of the system from behavioural data and plugging the target policy into the learned model [4]. However, such an approach does not take into account a fundamental issue of OPP: switching policies *induces a distribution shift*, and so, the model inferred under the behavioural distribution cannot offer, in general, reliable predictions under the target distribution.

In this paper, we focus on the OPP problem in systems comprising multiple interacting agents that evolve over time according to stochastic policies and stochastic dynamics. Here, we deal with a situation where one or more *ego* agents change their policies. This setting is considerably more challenging than the single-agent case because the distribution shift involves the predictions of all agents, not just the ego agents: even if the other agents remain with their behavioural (observational) policies, their actions will change in response to the shift in the ego agents' behaviours.

To obtain reliable predictions, we leverage the framework of *conformal prediction (CP)* [5], [6], a technique to derive prediction regions guaranteed to cover the true (unknown) output with arbitrary probability. Crucially, these probabilistic guarantees are finite-sample (non-asymptotic) and distribution-free, i.e., they don't rely on priors or parametric assumptions about the data distribution. For these properties, it is unsurprising that CP has become, in recent years, the go-to method for uncertainty quantification, especially in safety-critical applications, see, [7]–[10].

CP uses a set of calibration points to derive a distribution of model residuals, a.k.a. *scores*. The prediction region for a test point is then obtained by including all outputs whose scores appear sufficiently likely w.r.t. such calibration distribution. This procedure yields the desired probabilistic guarantees as long as the calibration and test data are exchangeable (a weaker assumption than i.i.d.). Exchangeability, however, is violated in the presence of distribution shifts, which are inherent to OPP. The framework of *weighted exchangeability* [11] extends CP to handle distribution shifts by reweighting the calibration points "as if" they were sampled under the target distribution. This is achieved by using estimates of the density ratios (DRs) between target and behavioural distributions. Recent works on *Conformal Off-Policy Prediction (COPP)* [12]–[14] leverage CP and weighted exchangeability to provide valid prediction regions for OPP problems. These methods, however, only support single-agent systems and construct regions for scalar outcomes (e.g., reward).

In this paper, we present *MA-COPP*, the first conformal prediction algorithm to solve OPP problems involving multi-agent systems. Crucially, our approach derives joint prediction regions (JPRs, akin to reach-tubes) for the future (multi-dimensional) trajectories of all agents, making it more comprehensive than existing COPP methods which are limited to scalar outcomes.

Indeed, constructing a valid JPR under generic distribution shifts using CP and weighted exchangeability normally requires, for each test input, an exhaustive search over the output trajectory space, which is infeasible in our high-dimensional settings. To overcome this problem, we demonstrate that we can derive a conservative over-approximation of the true (unknown) JPR without resorting to exhaustive search, if the

T. Kuipers, M. Kazemi and N. Paoletti are with the Department of Informatics, King's College London, UK. Email: {tom.kuipers, milad.kazemi, nicola.paoletti}@kcl.ac.uk

R. Tumu, S. Yang and R. Mangharam are with the Department of Electrical and Systems Engineering, University of Pennsylvania, USA. Email: {nandant, yangs1, rahulm}@seas.upenn.edu

*Authors contributed equally.

979-8-3503-1632-2/24/$31.00 ©2024 IEEE

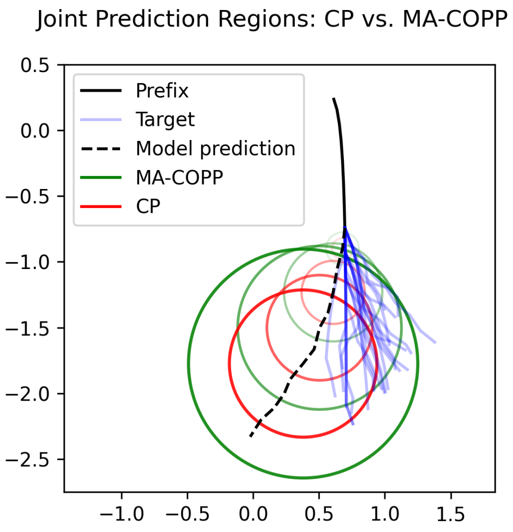


Fig. 1: 2D visualisation of actual JPRs constructed over the position of a single ego agent in the MPE environment defined in § V-B.1

maximum density ratio (DR) over all the JPR's trajectories is known. Building on this result, in MA-COPP, we pivot the search task over the maximum DR, which is significantly more effective than an exhaustive output search.

We evaluate MA-COPP on two multi-agent case studies: a multi-particle environment from the PettingZoo library for multi-agent reinforcement learning problems [15], and a continuous control racing model from the F1TENTH autonomous racing environment [16]. We find that our approach consistently achieves (close to) target coverage under a variety of distribution shifts and for output spaces of up to 72 dimensions, in settings where standard CP (which disregards the distribution shift) results in a coverage drop of up to 20%. Moreover, we find that MA-COPP yields over-conservative regions only for a very small proportion of instances.

We provide a practical example of the regions produced by the MA-COPP method using real data in Figure 1, where the effect of the miscoverage error of standard CP can be clearly seen. In contrast, the MA-COPP method produces a JPR which sufficiently covers the target trajectories.

II. The Multi-Agent OPP Problem

a) Notation: We use capital letters to denote random variables, lowercase for concrete values of those variables, and bold letters for the corresponding vectors (of random variables or concrete values) for all agents. Also, we will often use the notation $y_{1...T}$ as a shortcut for the sequence $(y_1,...,y_T)$.

b) Behavioural process: We consider a multi-agent system consisting of K agents and described, for each agent $k=1,...,K$ and time $t=1,...,T$, by the following process:

$$\begin{aligned} \mathbf{X}_1 &\sim P_{init}(\cdot); \\ A_{k,t} &\sim \pi_k^b(\cdot\,|\,\mathbf{X}_t); \\ X_{k,t+1} &\sim P_k(\cdot\,|\,\mathbf{X}_t, A_{k,t}) \end{aligned} \tag{1}$$

where $X_{k,t} \in \mathbb{R}^n$ is the agent's state, $\mathbf{X}_t = (X_{k,t})_{k=1}^K \in \mathbb{R}^{n\times K}$ is the environment/global state; $A_{k,t} \in \mathcal{A}_k$ is the performed action, and $\mathcal{A}_k$ is the discrete or continuous action space of the agent; $P_{init} : \mathbb{R}^{n\times K} \to [0,1]$ is the distribution of initial environment states; $\pi_k^b : (\mathbb{R}^{n\times K} \times \mathcal{A}_k) \to [0,1]$ is the stochastic behavioural policy of the agent; and $P_k : (\mathbb{R}^{n\times K} \times \mathcal{A}_k \times \mathbb{R}^n) \to [0,1]$ is the transition probability function determining the distribution of the next agent's states.

Each agent has access to the global environment state but does not know the actions of the other agents. Furthermore, we assume a subset of *ego agents* $E \subseteq \{1,\dots,K\}$ as the agents that are under our direct control and allowed to switch policies in the target process.

In our OPP settings, we do not have access to the process defined in (1), that is, we do not know the initial probabilities P_{init}, the transition probabilities P_k and the non-ego behavioural policies $\{\pi_k^b\}_{k\notin E}$. We only know the ego behavioural policy $\{\pi_e^b\}_{e\in E}$ and have access to the observational data defined below. For simplicity, we will now refer to terms pertaining to ego agents with subscript e and the non-ego agents with subscript k.

c) Observational data: We can observe a set of N global state trajectories of length T, and for ego agents only, we can also observe the actions they perform at each step of the trajectory. More formally, we have access to the following dataset of realisations of the process $(\mathbf{X}_1, (A_{e,1})_{e\in E}, \mathbf{X}_2, (A_{e,2})_{e\in E}, ..., \mathbf{X}_T)$ induced by (1):

$$\mathcal{D} = \left\{ \left(\mathbf{x}_t^{(i)}, \left(a_{e,t}^{(i)} \right)_{e\in E} \right)_{t=1}^{T} \right\}_{i=1}^{N}. \tag{2}$$

d) Problem definition: We focus on the problem of deriving a *joint prediction region* for the future agent trajectories $\mathbf{X}_{H+1...T}$, given a prefix $\mathbf{X}_{1...H}$, where H is the prefix length and $T > H$ is the total trajectory length. In particular, this problem is one of *off-policy prediction*: using observational data only (i.e., realisations of the behavioural process), we seek to construct a prediction region for $\mathbf{X}_{H+1...T}$ under a target policy different from the behavioural one. We consider the case where the policies of the ego agents $e \in E$ change after time H and the policies of all the other agents remain fixed, described by the *target process* in (3), where only $A_{k,t}$ differs from the behavioural process.

$$A_{k,t} \sim \pi_{k,t}(\cdot\,|\,\mathbf{X}_t); \tag{3}$$

where all non-ego agents always follow their behavioural policy ($\pi_{k,t} = \pi_k^b$ if $k \notin E$) and each ego agent initially follows their behavioural policy ($\pi_{k,t} = \pi_k^b$ if $t \leqslant H$) and switches their policy to a *known* target policy thereafter ($\pi_{e,t} = \pi_e^*$ if $t > H$). Below, we denote with P^{π^b} and P^{π^*} the distributions under the behavioural process (1) and the target process (3), respectively.

Problem 1 (Multi-Agent Off-Policy Prediction)**.** *Given observations $\mathcal{D} \sim_{iid} P^{\pi^b}$ of length T and of the form* (2) *generated by the behavioural process* (1)*, construct a* joint prediction region (JPR) $C_\alpha(\cdot)$ *with coverage $1-\alpha$ for i.i.d. test trajectories $\mathbf{X}_{1...T} \sim P^{\pi^*}$, where α is the desired miscoverage rate. Formally, the JPR must satisfy the following:*

$$\mathbb{P}_{\mathbf{X}_{1...T}\sim P^{\pi^*}}\left(\mathbf{X}_{H+1...T} \in C_\alpha(\mathbf{X}_{1...H})\right) \geqslant 1-\alpha. \tag{4}$$

where $\mathbf{X}_{1\dots T} \sim P^{\pi^*}$ is generated according to the target process (3).

Remark 1. *Although process* (1) *is Markovian, our predictions consider, in input, a sequence of past states to accommodate models (e.g., autoregressive, recurrent) that make use of a sequence of past states.*

III. Background

Conformal Prediction: an uncertainty quantification framework that can be applied on top of any supervised learning task for constructing distribution-free prediction regions with guaranteed marginal coverage. We now introduce the method using a standard regression example. Starting from a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_i$ of (x, y) pairs sampled i.i.d. from an unknown distribution $P_{X,Y}$. CP performs the following steps:

1) Split $\mathcal{D}$ into disjoint training and calibration datasets, $\mathcal{D}_t$ and $\mathcal{D}_c$, with $|\mathcal{D}_c|$ being the number of samples in $\mathcal{D}_c$;
2) Train a predictor $\hat{T} : X \to Y$ using $\mathcal{D}_t$;
3) Define a *non-conformity score function* $S : X \times Y \to \mathbb{R}$, such that $S(x, y)$ quantifies the discrepancy/residual between y and the prediction $\hat{T}(x)$;
4) Use $\mathcal{D}_c$ to define the calibration distribution
$$\widehat{F} = \sum_{i=1}^{|\mathcal{D}_c|} \frac{1}{|\mathcal{D}_c|+1}\delta_{s_i} + \frac{1}{|\mathcal{D}_c|+1}\delta_{\infty}, \tag{5}$$
where δ_s is the Dirac distribution with parameter s, $s_i = S(x^{(i)}, y^{(i)})$ is the score of the i-th calibration point, and δ_∞ represents the unknown score of the test point[1];
5) For a given test point x and failure rate α, construct the prediction region as $C_\alpha(x) = \{y : S(x,y) \leqslant Q_{1-\alpha}(\widehat{F})\}$, where $Q_{1-\alpha}(\widehat{F})$ is the $1-\alpha$ quantile of $\widehat{F}$. Such a prediction region satisfies the following coverage guarantee w.r.t. unseen test data $(x,y) \sim P_{X,Y}$:
$$\mathbb{P}_{(x,y)\sim P_{X,Y}}(y \in C_\alpha(x)) \geqslant 1-\alpha. \tag{6}$$
Note that the above holds in the more general case when (x,y) is exchangeable w.r.t. calibration data, i.e., when the joint probability of $(x^1, y^1), \dots, (x^{|\mathcal{D}_c|}, y^{|\mathcal{D}_c|}), (x, y)$ remains the same for any permutation of the data points.

Remark 2. *In standard CP for regression, to construct $C_\alpha(x)$ it is not required to enumerate and check inclusion for all candidate outputs. For instance, a common choice for S is $S(x,y) = \|y - \hat{T}(x)\|_p$ for some $p \geqslant 1$, and so $C_\alpha(x)$ can be constructed implicitly (without enumeration) as the L_p-ball centered at $\hat{T}(x)$ with radius $Q_{1-\alpha}(\widehat{F})$. This is possible because the quantile $Q_{1-\alpha}(\widehat{F})$ is the same for all candidate outputs.*

CP under Distribution Shift: Standard CP provides marginal coverage guarantees in the case where data observed at test-time is sampled from an exchangeable distribution as that of the calibration set $\mathcal{D}_c$. Should the distribution of test data differ in this regard (i.e. $\mathcal{D}_{test} \sim P^*_{X,Y} \neq P_{X,Y}$), it induces a distribution shift which violates the exchangeability assumption and with that, the coverage guarantee (6). The weighted exchangeability notion of [11] extends CP to deal with such shifts. It does so by reweighting the probabilities of the calibration distribution $\widehat{F}$ by the *density ratio (DR)* $w(x,y) = \mathrm{d}P^*_{X,Y}(x,y)/\mathrm{d}P_{X,Y}(x,y)$: in this way, we transform $\widehat{F}$ as if the scores had been computed over the target distribution. For a test point x and candidate output y, the reweighted distribution becomes:
$$\widehat{F}(x,y) = \sum_{i=1}^{|\mathcal{D}_c|} \frac{w(x^{(i)},y^{(i)})}{W + w(x,y)}\delta_{s_i} + \frac{w(x,y)}{W + w(x,y)}\delta_\infty \tag{7}$$
where $W = \sum_{i=1}^{|\mathcal{D}_c|} w(x^{(i)}, y^{(i)})$. We denote with $p_i(x,y)$ the probability of the i-th calibration point reweighted as above. Note that each score s_i has a higher (lower) probability if the target distribution makes $(x^{(i)}, y^{(i)})$ more (less) likely. By using $\widehat{F}(x,y)$ to construct $C_\alpha(x)$, we recover guarantee (6) for when $(x,y) \sim P^*_{X,Y}$ [11, Theorem 2].

Remark 3. *The reweighted calibration distribution depends on the test point (x,y) because the probability $p_{|\mathcal{D}_c|+1}$ of the test score needs to be reweighted by $w(x,y)$. This implies that, to construct a prediction region $C_\alpha(x)$ for a given test input x, we need to reweight $\widehat{F}$ for every candidate output y to determine if $y \in C_\alpha(x)$ by checking $S(x,y) \leqslant Q_{1-\alpha}(\widehat{F}(x,y))$. Thus, for general shifts, $C_\alpha(x)$ needs to be constructed by* enumerating *(and checking) individual candidates y, because the quantile $Q_{1-\alpha}(\widehat{F}(x,y))$ changes with y, i.e., the region cannot be constructed implicitly as per Remark 2. This is not the case, however, with covariate shift, whereby P_X changes but $P_{Y|X}$ does not: the DR now depends only on the input, and so $\widehat{F}(x,y)$ remains the same for every y.*

Conformal Off-Policy Prediction: The work of [12] builds on weighted exchangeability for conformal off-policy prediction (COPP) in contextual bandits—a special (and simpler) case of our problem, restricted to one step (i.e., $T = 2$ and $H = 1$), one agent ($K = 1$), and scalar outcomes (as opposed to our multi-dimensional JPRs). In this setting, changing the behavioural policy π^b into the target policy π^* induces a shift in the distribution of $Y \mid X$, while the distribution of X stays the same. Hence, the DR can be derived as follows:
$$\begin{aligned} w(x,y) &= \frac{\mathrm{d}P^{\pi^*}_{X,Y}(x,y)}{\mathrm{d}P^{\pi^b}_{X,Y}(x,y)} = \frac{\mathrm{d}P^{\pi^*}_{Y|X}(x,y) \cdot \mathrm{d}P^{\pi^*}_{X}(x,y)}{\mathrm{d}P^{\pi^b}_{Y|X}(x,y) \cdot \mathrm{d}P^{\pi^b}_{X}(x,y)} \\ &= \frac{\mathrm{d}P^{\pi^*}_{Y|X}(x,y)}{\mathrm{d}P^{\pi^b}_{Y|X}(x,y)} = \frac{\int P(y \mid a,x) \cdot \pi^*(a \mid x)\mathrm{d}a}{\int P(y \mid a,x) \cdot \pi^b(a \mid x)\mathrm{d}a} \end{aligned} \tag{8}$$

Since the transition probabilities P are unknown, the COPP approach derives an estimation $\hat{w}$ of w by plugging in (8) a data-driven surrogate $\hat{P}$ learned from (behavioural) data and by approximating the integrals with Monte-Carlo sampling of the policies. To construct $C_\alpha(x)$, instead of enumerating all candidate outputs y (required as per Remark 3), COPP implements an exhaustive grid search over the output space and returns the interval closure of all ys that pass the CP inclusion test. This operation is costly but remains feasible because, in the COPP method, y is a scalar.

Using an estimation of the DR has limitations, in that it directly affects the accuracy of the reweighted distribution (7)

[1] Since we do not know the true output for the test input, a worst-case score of ∞ is assumed.

and consequently the validity guarantees. In [17], the authors show that the miscoverage gap induced by using $\hat{w}$ instead of w is bounded by $\frac{1}{2}\mathbb{E}_{X,Y\sim P^{\pi^b}_{X,Y}}|w(X,Y)-\hat{w}(X,Y)|$. However, the deviation $|w(X,Y)-\hat{w}(X,Y)|$ and the resulting miscoverage error are inevitably exacerbated when considering, like in our settings, sequential processes involving multiple steps.

IV. The MA-COPP approach

The MA-COPP method provides a solution to Problem 1 by constructing valid joint prediction regions (JPRs) for the agents' future trajectories under the target policy despite only having access to observational data under the behavioural policy. To do so, we extend weighted exchangeability and the COPP method to deal with high-dimensional output spaces and JPRs arising from multiple steps and multiple agents.

An overview of the algorithm is presented in Figure 2.

a) Lifting of prediction task: We start by slightly reformulating our prediction task. Instead of constructing a JPR C_α for sequences of future global states $\mathbf{X}_{H+1\dots T}$, as per Problem 1, we will construct a JPR C^+_α for sequences of future global states *and* ego agents actions $\left(\mathbf{X}_t,(A_{t,e})_{e\in E}\right)_{t=H+1}^{T}$. For simplicity, we denote this "augmented" sequence by $\mathbf{Y}_{H+1\dots T}$. As explained below, we can trivially use a JPR that is valid for $\mathbf{Y}_{H+1\dots T}$ to construct a JPR for $\mathbf{X}_{H+1\dots T}$ (i.e., for the original problem).

Proposition 1. *For $\alpha\in(0,1)$ and $H<T$, let C^+_α be a JPR valid for $\mathbf{Y}_{H+1\dots T}$, i.e., such that*

$$\mathbb{P}_{(\mathbf{X}_{1\dots H},\mathbf{Y}_{H+1\dots T})\sim P^{\pi^*}}\left(\mathbf{Y}_{H+1\dots T}\in C^+_\alpha(\mathbf{X}_{1\dots H})\right)\geqslant 1-\alpha.$$

Let $C'_\alpha(\mathbf{X}_{1\dots H})=\Pi_{\mathbf{x}_{H+1\dots T}}(C^+_\alpha(\mathbf{X}_{1\dots H}))$ be the projection of the JPR onto components $\mathbf{x}_{H+1\dots T}$. Then, C'_α provides a solution to Problem 1 in that

$$\mathbb{P}_{(\mathbf{X}_{1\dots H},\mathbf{X}_{H+1\dots T})\sim P^{\pi^*}}\left(\mathbf{X}_{H+1\dots T}\in C'_\alpha(\mathbf{X}_{1\dots H})\right)\geqslant 1-\alpha.$$

Proof. For simplicity, we denote $\mathbf{X}_{1\dots H}$ by X, $\mathbf{X}_{H+1\dots T}$ by X', and $\mathbf{A}_{H+1\dots T,e}$ by A. So, we can define the coverage of (X,X',A) (the augmented sequence) w.r.t. the prediction region $C^+_\alpha(X)$ by $\mathbb{E}_{X,X',A}[f(X,X',A)]$, where f is the indicator function telling if $(X',A)\in C^+_\alpha(X)$. By the premises of the proposition, we have $\mathbb{E}_{X,X',A}[f(X,X',A)]\geqslant 1-\alpha$. So, we need to prove that $\mathbb{E}_{X,X'}[g(X,\ X')]\ \geqslant\ \mathbb{E}_{X,X',A}[f(X,X',A)]$ where g is the indicator function telling if $X'\in\Pi_{x'}(C^+_\alpha(X))$. We have the following derivation:

$$\begin{aligned}
&\mathbb{E}_{X,X',A}[f(X,X',A)]\\
&=\int f(x,x',a)\cdot\mathbb{P}_{X,X',A}(x,x',a)\mathrm{d}x,\mathrm{d}x',\mathrm{d}a\\
&\leqslant\int g(x,x')\cdot\mathbb{P}_{X,X',A}(x,x',a)\mathrm{d}x,\mathrm{d}x',\mathrm{d}a\\
&=\int g(x,x')\cdot\left(\int\mathbb{P}_{X,X',A}(x,x',a)\mathrm{d}a\right)\mathrm{d}x,\mathrm{d}x'\\
&=\int g(x,x')\cdot\mathbb{P}_{X,X'}(x,x')\mathrm{d}x,\mathrm{d}x'=\mathbb{E}_{X,X'}[g(X,X')],
\end{aligned}$$

where the first inequality holds because for any x,x', and a, we have that $f(x,x',a)\leqslant g(x,x')$ (because, by definition, $(x',a)\in C^+_\alpha(x)\rightarrow x'\in\Pi_{x'}(C^+_\alpha(x))$). □

Moving our focus to a more complex prediction task may seem counter-intuitive, but it allows us to use a precise definition of DR, as we will see next.

b) Density ratio computation: Similarly to COPP, our change of policy causes a shift in the distribution of $\mathbf{Y}_{H+1\dots T}\,|\,\mathbf{X}_{1\dots H}$, while the distribution of $\mathbf{X}_{1\dots H}$ stays the same. Hence, we have the following derivation for the DR:

$$\begin{aligned}
w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})&=\frac{\mathrm{d}P^{\pi^*}_{\mathbf{Y}_{H+1\dots T}|\mathbf{X}_{1\dots H}}(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})}{\mathrm{d}P^{\pi^b}_{\mathbf{Y}_{H+1\dots T}|\mathbf{X}_{1\dots H}}(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})}\\
&=\left(\frac{\prod_{t=H}^{T-1}\prod_{e\in E}P_e(x_{e,t+1}\,|\,\mathbf{x}_t,a_{e,t})\pi^*_e(a_{e,t}\,|\,\mathbf{x}_t)}{\prod_{t=H}^{T-1}\prod_{e\in E}P_e(x_{e,t+1}\,|\,\mathbf{x}_t,a_{e,t})\pi^b_e(a_{e,t}\,|\,\mathbf{x}_t)}\times\right.\\
&\qquad\left.\frac{\prod_{t=H}^{T-1}\prod_{k\notin E}\int P_k(x_{k,t+1}\,|\,\mathbf{x}_t,a_k)\pi^b_k(a_k\,|\,\mathbf{x}_t)\mathrm{d}a_k}{\prod_{t=H}^{T-1}\prod_{k\notin E}\int P_k(x_{k,t+1}\,|\,\mathbf{x}_t,a_k)\pi^b_k(a_k\,|\,\mathbf{x}_t)\mathrm{d}a_k}\right)\\
&=\frac{\prod_{t=H}^{T-1}\prod_{e\in E}\pi^*_e(a_{e,t}\,|\,\mathbf{x}_t)}{\prod_{t=H}^{T-1}\prod_{e\in E}\pi^b_e(a_{e,t}\,|\,\mathbf{x}_t)}
\end{aligned}\tag{9}$$

Owing to the reformulation of the prediction task, we can now express the DR in terms of the ego agents' policies only—which are known. This means that we can compute the DR precisely. Without such a reformulation, we would need to estimate the agent actions and transition probabilities jointly, as in (8), leading to an approximate DR.

Note that, for w to be well-defined, the likelihood of $(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})$ cannot be zero in both behavioural and target distributions. This assumption holds, for instance, when for any state $\mathbf{x}$, $\pi^b_e(\cdot\,|\,\mathbf{x})$ has full support on $\mathcal{A}_e$ for any ego agent $e\in E$, and $P_k(\cdot\,|\,\mathbf{x},a_k)$ has full support on $\mathbb{R}^n$ for any agent k and action $a_k\in\mathcal{A}_k$[2].

c) Non-conformity score: To define the score function $S(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})$, we need to establish a predictor $\hat{T}$ first. In our case, $\hat{T}$ is a multivariate regression function $\hat{T}:\mathbf{x}_{1\dots H}\mapsto\hat{\mathbf{x}}_{H+1\dots T}$ mapping a state sequence $\mathbf{x}_{1\dots H}$ into an estimate of its continuation $\hat{\mathbf{x}}_{H+1\dots T}$[3]. We choose the score function recently proposed in [18], which is designed for time series prediction tasks. This score describes the deviation between two (multi-dimensional) trajectories in terms of the maximum deviation across all time points. In particular, for an arbitrary choice of $\gamma_{H+1},\dots,\gamma_T>0$, S is defined as follows:

$$\begin{aligned}
&S(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})=\\
&\max_{t=H+1\dots T}\left\{\gamma_t\left\|\sigma\circ\left(\hat{T}(\mathbf{x}_{1\dots H})-\mathbf{y}_{H+1\dots T}\right)\right\|_2\right\},
\end{aligned}\tag{10}$$

where $\hat{T}(\mathbf{x}_{1\dots H})$ (resp., $\mathbf{y}_{H+1\dots T}$) is the global state at time t according to the prediction (resp., the output $\mathbf{y}_{H+1\dots T}$), and σ includes normalisation constants for each dimension. The γ_t parameters determine how much the residuals at different time points contribute to the score. Since residuals tend to grow as the prediction horizon increases, without such parameters (i.e., if $\gamma_t=1$ for all t), the score would be dominated by prediction errors at time points far ahead in the horizon. Thus, a sensible choice for γ_t is any monotonically decreasing series – in our experiments we choose $\gamma_t=(t-H)^{-1}$ – which

[2] Full support can be ensured, for instance, by adding Gaussian noise to the outputs of π^b_e and P_k.

[3] Note that we do not require $\hat{T}$ to predict actions but only states. Our final aim is to construct JPRs for state sequences and so actions should not affect the score $S(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})$, that is, the deviation between predictions and ground truth. Actions are only involved in the reweighting of $\hat{F}$ (through the DR (9)).

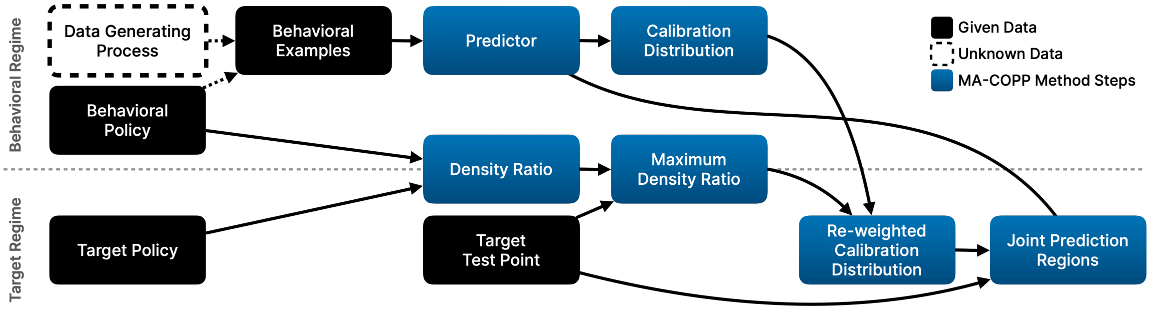


Fig. 2: Overview of the MA-COPP method. A calibration distribution is first derived from behavioural data and a predictor (see § IV-.0.c). Density ratios are computed as described in § IV-.0.b. To construct the JPR for a given test point, we estimate the maximum DR over all the outputs that pass the CP test, and use this estimate to reweight the calibration distribution, see § IV-.0.d.

mitigates the above issue by assigning more importance to prediction errors in the immediate future. Finally, given a test input $\mathbf{x}_{1\dots H}$, we define the JPR C_α^+ as follows:

$$C_\alpha^+(\mathbf{x}_{1\dots H}) = \{\mathbf{y}_{H+1\dots T} \mid S(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T}) \leqslant Q_{1-\alpha}(\hat{F}(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T}))\} \tag{11}$$

where $\hat{F}(\mathbf{x}_{1\dots H}, \mathbf{y}_{H+1\dots T})$ is the calibration distribution constructed with behavioural data and reweighted according to the DR $w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})$ (see Eq. 7).

d) Construction of Prediction Regions: As discussed in Remark 3, with distribution shifts (other than covariate shifts), the calibration distribution needs to be reweighted for every candidate output. However, enumerating the outputs as in [12] quickly becomes infeasible as the size of the search space is exponential in the number of output dimensions.

To alleviate this problem, various search strategies could be applied. Then, the prediction region would be constructed by taking some closure (to avoid zero-volume regions) of those visited trajectories that pass the CP test. However, such strategies are incomplete, and the resulting region will necessarily undercover.

Crucially, our MA-COPP approach overcomes the issue of enumerating or searching the output space, based on the following intuition. Due to the above limitations, we cannot compute precisely $C_\alpha^+(\mathbf{x}_{1\dots H})$, i.e., the true JPR (11). However, if we knew the value $w^\top(C_\alpha^+(\mathbf{x}_{1\dots H}))$ of the maximum DR among all trajectories in $C_\alpha^+(\mathbf{x}_{1\dots H})$, then we could use the same $w^\top(C_\alpha^+(\mathbf{x}_{1\dots H}))$ for all candidate outputs $\mathbf{y}_{H+1\dots T}$ when reweighting the calibration distribution. The resulting region would meet the coverage guarantees because it is a conservative approximation of $C_\alpha^+(\mathbf{x}_{1\dots H})$ and, importantly, can be constructed implicitly without enumerating the output space because all outputs have now the same critical value, as per Remark 2. Before introducing this statement more formally, let us denote with $\hat{F}(w)$ the calibration distribution $\hat{F}$ reweighted by $w \in (0,\infty)$.

Proposition 2 (Max-DR region)**.** *Let* $\mathcal{Y} \subseteq C_\alpha^+(\mathbf{x}_{1\dots H})$ *be a subset of the JPR* (11)*. Let* $w^\top(\mathcal{Y}) = \max_{\mathbf{y}_{H+1\dots T} \in \mathcal{Y}} w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})$ *be the maximum DR within* $\mathcal{Y}$*. Define the* max-DR region *as* $C_\alpha^+(\mathbf{x}_{1\dots H}, w^\top(\mathcal{Y})) = \{\mathbf{y}_{H+1\dots T} \mid S(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T}) \leqslant Q_{1-\alpha}(\hat{F}(w^\top(\mathcal{Y})))\}$*. Then,* $\mathcal{Y} \subseteq C_\alpha^+(\mathbf{x}_{1\dots H}, w^\top(\mathcal{Y}))$.

Proof. To prove that $\mathcal{Y} \subseteq C_\alpha^+(\mathbf{x}_{1\dots H}, w^\top(\mathcal{Y}))$, we show that for any $\mathbf{y}_{H+1\dots T} \in \mathcal{Y}$, $Q_{1-\alpha}(\hat{F}(\mathbf{x}_{1\dots H}, \mathbf{y}_{H+1\dots T})) \leqslant Q_{1-\alpha}(\hat{F}(w^\top(\mathcal{Y})))$. This follows from (7), since the reweighted probability of the test point (which has ∞ score) is always higher in $\hat{F}(w^\top(\mathcal{Y}))$, i.e., $\dfrac{w^\top(\mathcal{Y})}{W + w^\top(\mathcal{Y})} \geqslant \dfrac{w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})}{W + w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})}$ for any $\mathbf{y}_{H+1\dots T} \in \mathcal{Y}$. □

A consequence of Proposition 2 is that, given the true maximum DR $w^\top(C_\alpha^+(\mathbf{x}_{1\dots H}))$, the corresponding max-DR region is valid (i.e., it has coverage of at least $1-\alpha$) because it contains the true JPR $C_\alpha^+(\mathbf{x}_{1\dots H})$. However, $w^\top(C_\alpha^+(\mathbf{x}_{1\dots H}))$ is not known a priori and needs to be estimated.

Our approach uses search techniques to find an under-approximation $\tilde{w}^\top \leqslant w^\top(C_\alpha^+(\mathbf{x}_{1\dots H}))$ of the true maximum DR. While the resulting max-DR region $C_\alpha^+(\mathbf{x}_{1\dots H}, \tilde{w}^\top)$ may not achieve the target coverage, pivoting the search task over the maximum DR is substantially more effective than doing so over the output space directly – and this is confirmed by our experiments. The reason is that with our approach, as soon as we find *just one* output trajectory $\mathbf{y}_{H+1\dots T}$ of the true JPR (i.e., that passes the CP test), then we can use the corresponding DR $w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})$ to construct a max-DR region that is guaranteed to include *all* trajectories of the true JPR with a DR below or equal to $w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})$ (see Prop. 3 below). On the contrary, without our approach, we would need an extensive (and likely infeasible) search to cover all those trajectories sufficiently well.

Proposition 3. *For* $\tilde{w}^\top \in (0, \infty)$*, the corresponding max-DR region* $C_\alpha^+(\mathbf{x}_{1\dots H}, \tilde{w}^\top)$ *contains the JPR subset* $\{\mathbf{y}_{H+1\dots T} \in C_\alpha^+(\mathbf{x}_{1\dots H}) \mid w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T}) \leqslant \tilde{w}^\top\}$.

Proof. For any $\mathbf{y}_{H+1\dots T} \in C_\alpha^+(\mathbf{x}_{1\dots H})$, it holds that $S(\mathbf{x}_{H+1\dots T},\mathbf{y}_{H+1\dots T}) \leqslant Q_{1-\alpha}(\hat{F}(w(\mathbf{x}_{1\dots H},\mathbf{y}_{H+1\dots T})))$. As shown in the proof of Prop. 2, if $w(\mathbf{x}_{1\dots H},\mathbf{y}_{1\dots H}) \leqslant \tilde{w}^\top$, then we have $S(\mathbf{x}_{H+1\dots T},\mathbf{y}_{H+1\dots T}) \leqslant Q_{1-\alpha}(\hat{F}(\tilde{w}^\top))$, i.e., $\mathbf{y}_{H+1\dots T} \in C_\alpha^+(\mathbf{x}_{1\dots H},\tilde{w}^\top)$. □

To estimate the maximum DR, our search technique performs sampling of a *synthetic target process*, an approximation of the target process where the unknown transition probabilities P_k and the non-ego policies π_k^b are replaced by data-driven approximations $\hat{P}_k$ and $\hat{\pi}_k^b$ learned from behavioural data[4].

V. Results

A. Experimental Settings

We evaluate our method on two case studies, a collaborative environment with a continuous state space and a discrete action space, and an adversarial environment with a continuous state and action space. In both case studies, we compare the MA-COPP method against two configurations. 1) $T \rightarrow T$ is a standard CP approach using the true target distribution for both calibration and test data: the gold standard approach where everything is known. 2) $B \rightarrow T$ is standard CP under distribution shift whereby we use the behavioural distribution for the calibration data, and the target distribution at test-time: the standard approach which disregards distribution shifts. For all experiments, we consider the nominal coverage rate $\alpha = 0.95$

B. Case Studies

1) Multi-Particle Environment (MPE): We use an MPE environment based on the PettingZoo [15] library. In this collaborative 2D environment with discrete actions, there are k agents and m *landmarks*. Landmarks take the form of static circles in the state space. In our experiments, we set $k = m = 3$, with only one ego agent. The goal of the environment is for the agents to cooperatively cover all of the landmarks whilst avoiding collisions with each other.

The state space of the environment is given by the vector:

$$\mathbf{x}_t = \Big((x_{k,t}, v_{k,t})_{k=1}^{K}, l_1, ..., l_M\Big) \tag{12}$$

where $x_{k,t}$ and $v_{k,t}$ are the position and velocity vectors for each agent at time t. Each agent can only observe the position (with additive noise sampled from a Gaussian distribution) of the other agents and so its observation space is defined as follows:

$$y_{k,t} = \Big(x_{k,t}, v_{k,t}, (x_{j,t})_{j \neq k}, l_1, ..., l_M\Big) + w_k^{sensor} \tag{13}$$

At each time step, the agents sample their actions from a discrete space $\mathcal{A}$ = *{left, right, up, down, do nothing}* according to their own stochastic policies, which are parameterized by neural networks. Each action updates the agent's velocity by a constant vector—one for each of the cardinal directions w.r.t. the origin, as well as $[0,0]$. The transition kernel, P, for each agent can be defined as follows:

$$X_{k,t+1} = X_{k,t} + V_{k,t} + W_k^{act};\ V_{k,t+1} = V_{k,t} + U_{k,t} \tag{14}$$

where X_k and V_k are the position and velocity of agent k and W_k^{act} is a Gaussian noise term modelling actuation noise. The resultant control input $U_{k,t}$ that an agent receives is the unit vector corresponding to the cardinal direction (i.e. $[0,1]^T$ if $A_{k,t} = \text{up}$) multiplied a scalar representing the intensity of the acceleration.

We define the each behavioural policy as ϵ-greedy, that is, with probability $1 - \epsilon_{greedy}$ it selects the true action $A_{k,t}^{(i)}$ or a random action $A_{k,t}^{(j)}, i \neq j$ with probability $\epsilon_{greedy}/|\mathcal{A}| - 1$. The target policies for the ego agents are defined in the same way, with the exception of an additional *bias* term which we use to control the degree of distribution shift. This bias modifies the action probabilities further by selecting a fixed action (in our case, *down*) with probability ϵ_{bias}. In our experiments, we fix $\epsilon_{greedy}^{b} = 0.1$ and ϵ_{greedy}^{*}, and evaluate our method with various $\epsilon_{bias} \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$.

We generate 3,200 prefixes (1,600–800–800 for training–calibration–test respectively). The training set is used only for learning an LSTM, as required in IV-.0.c as well as learning the dynamics models for the synthetic process. For each prefix, we generate 25 Monte-Carlo continuations, both under the behavioural and true target—which we assume to be unknown, but use only for the sake of evaluation. We also sample 25 continuations from the synthetic process. For all of the datasets, we generate 9 steps and 12 steps for the prefix and continuations respectively. The prediction regions are formed over the continuation of test prefixes, and are of the dimension $k \times 2 \times (T - H + 1)$.

We find that the MA-COPP method performs very well with respect to standard CP, and results in useful and tight prediction regions. When evaluating marginal coverage rates in particular, we observe in both figures 3b and 3c, that with a relatively low epsilon bias of 0.1, the coverage of standard CP breaks down whilst the MA-COPP method continues to provide the 95% coverage guarantee until $\epsilon = 0.2$. Beyond this point, MA-COPP also experiences to drop in coverage below 95%. This is caused by an under-approximation of the true max-DR, leading to prediction regions that are too tight and therefore under-cover, albeit to a far lesser degree than standard CP.

In addition to the coverage results, we also observe a corresponding increase in region size as the JPRs grow to allow for more uncertainty. Since we assume that the true process is unknown and must rely on our synthetic process for the search over the max-DR (as described in IV-.0.d), we also evaluate a fourth setting in which we perform the reweighting with the true target distribution ($B^T \rightarrow T$). Crucially, we observe that in both figures 3b and 3d, we see that using the synthetic data-generating process (instead of the true process) proves to be sufficient for exploring the max-DR value necessary to increase the critical value sufficiently to cover the true target output space.

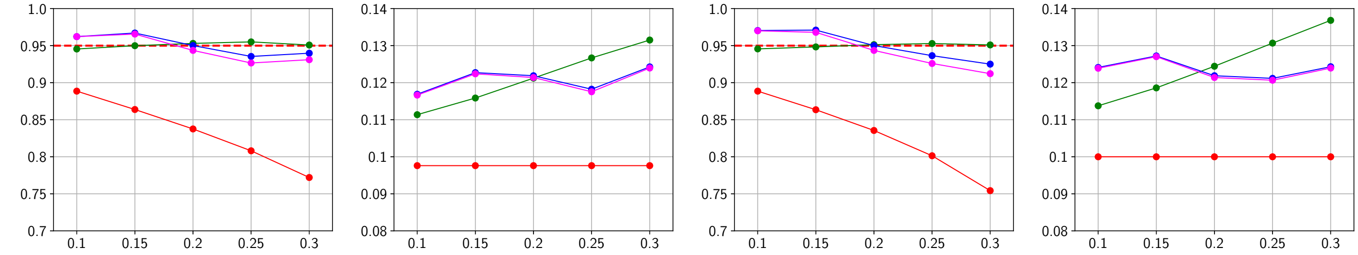
As the degree of policy shift increases, the max-DR often becomes very large. This results in test points where the critical value after reweighting is ∞. These regions are unavoidable in cases with large shift, however, as shown in Table V-B.1, only a small proportion of test points have such regions in our MPE experiments.

2) F1TENTH: We also evaluated MA-COPP in a competitive, continuous control racing environment, presented in

ϵ-bias T	0.1	0.15	0.2	0.25	0.3
8	0	0	0.06625	0.0475	0.0575
12	0.01	0.05	0.06625	0.06375	0.05125

TABLE I: Proportion of test points with unbounded prediction regions

[4] For efficiency, one may choose to learn an end-to-end approximation of the non-ego dynamics, i.e., a single predictor mapping $\mathbf{X}_t$ directly into $(X_{k,t+1})_{k \notin E}$ (thus avoiding to learn the non-ego agents' policies). We define $\hat{\pi}_k^b$ and $\hat{P}_k$ as isotropic multi-variate Gaussians with parameters predicted by a neural network model.



(a) Marginal coverage rates with a prediction length of 8 timesteps (b) Critical values of JPRs with a prediction length of 8 timesteps (c) Marginal coverage rates with a prediction length of 12 timesteps (d) Critical values of JPRs with a prediction length of 12 timesteps

Fig. 3: x-axis: ϵ_{bias}; Dashed red: nominal coverage level; Red: standard CP $B \rightarrow T$; Green: standard CP $T \rightarrow T$; Blue: MA-COPP with synthetic process $B^S \rightarrow T$; Magenta: MA-COPP with true process $B^T \rightarrow T$

[16]. In the F1TENTH environment, agents are in control of racecars, with states $s = \left(x,y,\delta,v,\Psi,\dot{\Psi},\beta\right) \in \mathbb{R}^7$. x,y represent the x and y position of the vehicle center. δ represents the steering angle, and v the forward velocity of the car. $\Psi,\dot{\Psi}$ denote the vehicle's yaw angle and yaw rate, and β gives the vehicle's slip angle at its centre. The actions of these vehicles are also continuous, $a = (v,\delta) \in \mathbb{R}^2$. $v \in [0,6]$ represents the velocity, and $\delta \in [-0.4,0.4]$ the steering angle. The full model is available in [19]. The simulator is run at a frequency of 100Hz.

Three agents, one ego, and two adversaries compete in a race beginning at various locations along a track. The opponent agents operate at 95% of the speed of the ego agent and follow pre-specified racelines that are centerline offsets. The ego agent's nominal trajectory will bring it into collision with the opponents. A collision avoidance algorithm is used to avoid this. We define the composition of these two components as our nominal policy. The behavioural policy is an $\epsilon^b_{greedy} = 0.2$ policy as described in Section V-B.1. The target policies replicate this approach with probabilities $\epsilon^t_{greedy} = \{0.3,0.4,0.5,0.6,0.7\}$, representing a shift away from the nominal policy. In the F1TENTH experiments, the data-generating process and the synthetic process used to search for the maximum DR are the same.

We predict agent trajectories four timesteps into the future. 5,077 prefixes (2538–1270–1269 for training–calibration–test respectively) were generated. For each of the 5,077 prefixes, 25 behavioural and target suffixes, as well as 50 synthetic suffixes each were generated. The prediction is done for each agent position, resulting in predictions with dimension (3,2,4), for 3 agents over 4 timesteps.

Results are presented in Table II. First, we note that standard CP ($B \rightarrow T$) performs worse as ϵ grows, losing more than 2% coverage at $\epsilon = 0.7$. We expect this behaviour as the target distribution shifts away from the behavioural. In contrast, the MA-COPP approach compensates for this shift and consistently provides coverage in line with the nominal 95% target. We also examine the gold-standard CP approach ($T \rightarrow T$), which gives us a benchmark for the critical value required to achieve the desired coverage. In all examples, we see that the MA-COPP approach has a larger critical value and, therefore, larger regions than the $T \rightarrow T$ approach. While MA-COPP does result in larger regions, it also provides coverage closer to the target coverage. We can see that in the continuous control example presented, MA-COPP succeeds in compensating for policy shift when standard CP approaches experience coverage gaps and that it does so without very conservative regions. When the distribution shift is too great, MA-COPP can provide trivial guarantees, as the density ratio can quickly explode. These results have been omitted for space.

Approach		T → T	B → T	MA-COPP
ϵ	Metric			
0.3	Coverage	94.22%	94.26%	**95.02%**
	Avg. CV	1.689	1.692	1.742
0.4	Coverage	94.45%	94.32%	**94.94%**
	Avg. CV	1.701	1.692	1.738
0.5	Coverage	94.24%	93.92%	**94.78%**
	Avg. CV	1.714	1.692	1.754
0.6	Coverage	94.39%	93.79%	**95.23%**
	Avg. CV	1.731	1.692	1.796
0.7	Coverage	94.16%	92.99%	**95.51%**
	Avg. CV	1.766	1.692	1.867

TABLE II: Results from the F1TENTH experiments, with a suffix length of 4 timesteps. The notation X → Y is used as shorthand for standard split conformal prediction, fit on the precedent, and evaluated on the consequent.

VI. Related Work

Conformal prediction (CP) [5], [6] is a popular uncertainty quantification method, with numerous applications including computer vision [20], language models [21], and system verification [8], [22]–[24]. Recently, conformal prediction has also been used for safe robotic planning and control, see, e.g., [9], [10], [25]–[28]. Our work is also related to time-series forecasting, for which we find a growing body of CP-based approaches such as [29], [30].

Past COPP work and our work leverage the weighted exchangeability method of [11] (discussed in more detail in Section III), but other CP-based approaches exist that address the distribution shift problem, such as *adaptive CP* [31], an approach that dynamically adjusts the coverage level in an online manner to compensate for observed under/over-coverage under unknown distribution shifts, or *robust CP* [32], which extends CP to handle bounded adversarial input perturbations.

While our work is the first to support multiple agents and prediction of multi-dimensional trajectories, it was inspired by prior COPP methods for single-agent systems such as [12] (see Section III), the work of [13] which considers dynamic models (Markov Decision Processes) but predicts scalar outcomes (the value of the MDP trajectories), and the subsampling-based approach of [14] which only supports discrete actions and becomes prohibitive for long trajectories.

VII. Conclusion

We presented MA-COPP, the first conformal prediction method for reliable off-policy prediction in multi-agent systems. Our approach avoids the output space enumeration that frustrates existing COPP approaches by reweighting (for every test input) the calibration distribution only once, using an estimate of the maximum density ratio. We evaluated our method on two case studies, respectively involving discrete and continuous action spaces, demonstrating that MA-COPP succeeds in adjusting the coverage without generating excessively conservative regions, even in cases where standard CP massively undercovers.

VIII. Disclosure of Funding

This work is supported by the *REXASI-PRO* H-EU project, call HORIZON-CL4-2021-HUMAN- 01-01, grant agreement ID: 101070028; the Engineering and Physical Sciences Research Council (EPSRC) under Award EP/W014785/2; NSF CCRI 1925587; NSF GRFP award DGE-2236662; and US DoT Safety21 National University Transportation Center.

References

[1] M. Uehara, C. Shi, and N. Kallus, "A review of off-policy evaluation in reinforcement learning," *arXiv preprint arXiv:2212.06355*, 2022.

[2] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[3] S. A. Murphy, M. J. van der Laan, J. M. Robins, and C. P. P. R. Group, "Marginal mean models for dynamic regimes," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1410–1423, 2001.

[4] H. Le, C. Voloshin, and Y. Yue, "Batch policy learning under constraints," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3703–3712.

[5] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.

[6] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint arXiv:2107.07511*, 2021.

[7] L. Bortolussi, F. Cairoli, N. Paoletti, S. A. Smolka, and S. D. Stoller, "Neural predictive monitoring," in *Runtime Verification: 19th International Conference, RV 2019, Porto, Portugal, October 8–11, 2019, Proceedings 19*. Springer, 2019, pp. 129–147.

[8] F. Cairoli, N. Paoletti, and L. Bortolussi, "Conformal quantitative predictive monitoring of stl requirements for stochastic processes," in *Proceedings of the 26th ACM International Conference on Hybrid Systems: Computation and Control*, 2023, pp. 1–11.

[9] X. Yu, Y. Zhao, X. Yin, and L. Lindemann, "Signal temporal logic control synthesis among uncontrollable dynamic agents with conformal prediction," *arXiv preprint arXiv:2312.04242*, 2023.

[10] S. Yang, G. J. Pappas, R. Mangharam, and L. Lindemann, "Safe perception-based control under stochastic sensor uncertainty using conformal prediction," *arXiv preprint arXiv:2304.00194*, 2023.

[11] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," *Advances in neural information processing systems*, vol. 32, 2019.

[12] M. F. Taufiq, J.-F. Ton, R. Cornish, Y. W. Teh, and A. Doucet, "Conformal off-policy prediction in contextual bandits," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 31 512–31 524. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/cc84bfabe6389d8883fc2071c848f62a-Paper-Conference.pdf

[13] D. Foffano, A. Russo, and A. Proutiere, "Conformal off-policy evaluation in markov decision processes," *arXiv preprint arXiv:2304.02574*, 2023.

[14] Y. Zhang, C. Shi, and S. Luo, "Conformal off-policy prediction," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 2751–2768.

[15] J. Terry, B. Black, N. Grammel, M. Jayakumar, A. Hari, R. Sullivan, L. S. Santos, C. Dieffendahl, C. Horsch, R. Perez-Vicente *et al.*, "Pettingzoo: Gym for multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 032–15 043, 2021.

[16] M. O'Kelly, H. Zheng, D. Karthik, and R. Mangharam, "F1TENTH: An Open-source Evaluation Environment for Continuous Control and Reinforcement Learning," in *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*. PMLR, Aug. 2020, pp. 77–89, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v123/o-kelly20a.html

[17] L. Lei and E. J. Candès, "Conformal inference of counterfactuals and individual treatment effects," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 83, no. 5, pp. 911–938, 2021.

[18] M. Cleaveland, I. Lee, G. J. Pappas, and L. Lindemann, "Conformal prediction regions for time series using linear complementarity programming," *arXiv preprint arXiv:2304.01075*, 2023.

[19] M. Althoff, M. Koschi, and S. Manzinger, "CommonRoad: Composable benchmarks for motion planning on roads," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, Jun. 2017, pp. 719–726. [Online]. Available: https://ieeexplore.ieee.org/document/7995802

[20] A. N. Angelopoulos, A. P. Kohli, S. Bates, M. Jordan, J. Malik, T. Alshaabi, S. Upadhyayula, and Y. Romano, "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging," in *International Conference on Machine Learning*. PMLR, 2022, pp. 717–730.

[21] V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. S. Jaakkola, and R. Barzilay, "Conformal language modeling," *arXiv preprint arXiv:2306.10193*, 2023.

[22] L. Bortolussi, F. Cairoli, N. Paoletti, S. A. Smolka, and S. D. Stoller, "Neural predictive monitoring and a comparison of frequentist and bayesian approaches," *International Journal on Software Tools for Technology Transfer*, vol. 23, no. 4, pp. 615–640, 2021.

[23] F. Cairoli, L. Bortolussi, and N. Paoletti, "Neural predictive monitoring under partial observability," in *Runtime Verification: 21st International Conference, RV 2021, Virtual Event, October 11–14, 2021, Proceedings 21*. Springer, 2021, pp. 121–141.

[24] L. Lindemann, X. Qin, J. V. Deshmukh, and G. J. Pappas, "Conformal prediction for stl runtime verification," in *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, 2023, pp. 142–153.

[25] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas, "Safe planning in dynamic environments using conformal prediction," *IEEE Robotics and Automation Letters*, 2023.

[26] A. Dixit, L. Lindemann, S. X. Wei, M. Cleaveland, G. J. Pappas, and J. W. Burdick, "Adaptive conformal prediction for motion planning among dynamic agents," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 300–314.

[27] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," *arXiv preprint arXiv:2307.01928*, 2023.

[28] A. Muthali, H. Shen, S. Deglurkar, M. H. Lim, R. Roelofs, A. Faust, and C. Tomlin, "Multi-agent reachability calibration with conformal prediction," *arXiv preprint arXiv:2304.00432*, 2023.

[29] K. Stankeviciute, A. M Alaa, and M. van der Schaar, "Conformal time-series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 6216–6228, 2021.

[30] S. Sun and R. Yu, "Copula conformal prediction for multi-step time series forecasting," *arXiv preprint arXiv:2212.03281*, 2022.

[31] I. Gibbs and E. Candes, "Adaptive conformal inference under distribution shift," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1660–1672, 2021.

[32] A. Gendler, T.-W. Weng, L. Daniel, and Y. Romano, "Adversarially robust conformal prediction," in *International Conference on Learning Representations*, 2021.