

Conformal Prediction Regions for Time Series Using Linear Complementarity Programming

Matthew Cleaveland¹, Insup Lee², George J. Pappas¹, Lars Lindemann³

¹Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

²Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA

³Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA
mcleav@seas.upenn.edu, lee@seas.upenn.edu, pappasg@seas.upenn.edu, llindema@usc.edu

Abstract

Conformal prediction is a statistical tool for producing prediction regions of machine learning models that are valid with high probability. However, applying conformal prediction to time series data leads to conservative prediction regions. In fact, to obtain prediction regions over T time steps with confidence $1-\delta$, previous works require that each individual prediction region is valid with confidence $1-\delta/T$. We propose an optimization-based method for reducing this conservatism to enable long horizon planning and verification when using learning-enabled time series predictors. Instead of considering prediction errors individually at each time step, we consider a parameterized prediction error over multiple time steps. By optimizing the parameters over an additional dataset, we find prediction regions that are not conservative. We show that this problem can be cast as a mixed integer linear complementarity program (MILCP), which we then relax into a linear complementarity program (LCP). Additionally, we prove that the relaxed LP has the same optimal cost as the original MILCP. Finally, we demonstrate the efficacy of our method on case studies using pedestrian trajectory predictors and F16 fighter jet altitude predictors.

1 Introduction

Autonomous systems perform safety-critical tasks in dynamic environments where system errors can be dangerous and costly, e.g., a mistake of a self-driving car navigating in urban traffic. In such scenarios, accurate uncertainty quantification is vital for ensuring safety of learning-enabled systems. In this work, we focus on quantifying uncertainty of time series data. For instance, for a given time series predictor that predicts the future trajectory of a pedestrian, how can we quantify the uncertainty and accuracy of this predictor?

Conformal prediction (CP) has emerged as a popular method for statistical uncertainty quantification (Shafer and Vovk 2008; Vovk, Gammerman, and Shafer 2005). It aims to construct prediction regions that contain the true quantity of interest with a user-defined probability. CP does not require any assumptions about the underlying distribution or the predictor itself. Instead, one only needs calibration data that is exchangeable or independent and identically distributed. That means that CP can seamlessly be applied to learning-enabled predictors like neural networks, without the need to

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

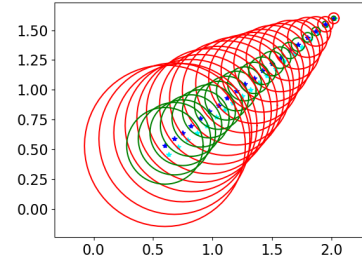


Figure 1: Sample trajectory of a pedestrian (teal stars), the pedestrian predictions using a social LSTM from (Alahi et al. 2016) (blue stars), the conformal prediction regions for our approach (green circles), and the conformal prediction regions for the approach from (Lindemann et al. 2023a) (red circles). Notably, our approach is less conservative.

analyze the underlying architecture or change the training procedure (Angelopoulos and Bates 2023).

One challenge of CP is that its standard variant can not directly construct valid prediction regions for time series. This is because datapoints at different time steps are not exchangeable. Variants of CP have been developed for time series, such as adaptive conformal inference (Gibbs and Candès 2021; Zaffran et al. 2022). However, their coverage guarantees are weaker and only asymptotic. While these types of guarantees are useful in many contexts, they are insufficient in safety-critical applications such as self-driving cars.

In previous works (Stankeviciute, M Alaa, and van der Schaar 2021; Lindemann et al. 2023a,b), this issue was mitigated by instead considering a calibration dataset that consists of full rollouts of the time series, which do not violate the exchangeability assumption. At design time, one runs a separate instance of CP for each of the T time steps of the time series predictor (also referred to as a trajectory predictor). Then at runtime, these CP instances are combined with the predictor’s T predictions to compute prediction regions. However, to obtain prediction regions that are valid over all T time steps with confidence $1-\delta$, each individual prediction region has to be valid with confidence $1-\delta/T$.¹ This results in conservative prediction regions.

Contributions: To address this conservatism, this paper

¹A union bound argument can be used to show this result.

proposes constructing prediction regions for time series using linear complementary programming. Our main idea is based on i) mapping trajectory prediction errors to a single parameterized conformal scoring function to avoid union bounding, and ii) finding the optimal parameters from an additional calibration dataset. As opposed to existing works, our method enables non-conservative long horizon planning and verification. Our contributions are as follows:

- We propose the conformal scoring function $R := \max(\alpha_1 R_1, \dots, \alpha_T R_T)$ for time horizon T , prediction errors R_t , and parameters α_t . Using R , we show that we can obtain non-conservative prediction regions with $1 - \delta$ confidence using conformal prediction.
- We formulate the problem of finding the parameters α_t that minimize R from an additional calibration dataset as a mixed integer linear complementarity program, and we show how we can derive a linear complementarity program that has the same optimal value.
- On two case studies, we demonstrate that our method produces much smaller valid conformal prediction regions compared to (Stankeviciute, M Alaa, and van der Schaar 2021; Lindemann et al. 2023a,b), see Figure 1.

1.1 Related Work

Conformal prediction was originally proposed by Vovk in (Vovk, Gammernan, and Shafer 2005; Shafer and Vovk 2008) to quantify uncertainty of prediction models. The original method, however, required training a prediction model for each calibration datapoint, which is computationally intractable for learning-enabled predictors. Inductive conformal prediction, also referred to as split conformal prediction, addresses this issue by splitting the calibration data into two sets, one used for training the prediction model and one for applying conformal prediction (Papadopoulos 2008). Split conformal prediction has been extended to provide conditional probabilistic guarantees (Vovk 2012), to handle distribution shifts (Tibshirani et al. 2019; Fannjiang et al. 2022), and to allow for quantile regression (Romano, Patterson, and Candes 2019). Further, split conformal prediction has been used to construct probably approximately correct prediction sets for machine learning models (Park et al. 2020; Angelopoulos et al. 2022), to perform out-of-distribution detection (Kaur et al. 2022, 2023), to guarantee safety in autonomous systems (Luo et al. 2022), and to quantify uncertainty for F1/10 car motion predictions (Tumu et al. 2023). Additionally, in (Stutz et al. 2022) the authors encode the width of the generated prediction sets directly into the loss function of a neural network during training.

However, the aforementioned methods cannot directly be applied to time series data. Multiple works have adapted conformal prediction algorithms for the time series domain, including enbpi (Xu and Xie 2021), adaptive conformal inference (Gibbs and Candes 2021), fully adaptive conformal inference (Gibbs and Candès 2022; Bhatnagar et al. 2023), and copula conformal prediction (Sun and Yu 2022; Tonkens et al. 2023). However, they either only provide averaged coverage guarantees over long horizons or require a lot of extra calibration data. Regardless, these methods have been

used for analyzing financial market data (Wisniewski, Lindsay, and Lindsay 2020), synthesizing safe robot controllers (Dixit et al. 2022), and dynamically allocating compute resources (Cohen et al. 2023). For a more comprehensive overview of the conformal prediction landscape see (Angelopoulos and Bates 2023).

Our paper is motivated by (Stankeviciute, M Alaa, and van der Schaar 2021; Lindemann et al. 2023a,b). These works apply inductive conformal prediction in a time series setting, which requires constructing individual prediction regions with confidence $1 - \delta/T$ to obtain $1 - \delta$ coverage for all time steps. This results in unnecessarily conservative prediction regions that are impractical for downstream tasks such as motion planning. We overcome this limitation by considering a parameterized conformal scoring function that is defined over multiple time steps.

2 Background

In this section, we present background on time series predictors and split conformal prediction for time series.

2.1 Time Series Predictors

We want to predict future values Y_1, \dots, Y_T of a time series for a prediction horizon of T from past observed values $Y_{T_{obs}}, \dots, Y_0$ for an observation length of T_{obs} . Let \mathcal{D} denote an unknown distribution over a finite-horizon time series $Y := (Y_{T_{obs}}, \dots, Y_0, Y_1, \dots, Y_T)$, i.e., let

$$(Y_{T_{obs}}, \dots, Y_0, Y_1, \dots, Y_T) \sim \mathcal{D}$$

denote a random trajectory drawn from \mathcal{D} where $Y_t \in \mathbb{R}^m$ is the value at time t . We make no assumptions about the distribution \mathcal{D} , but we do assume availability of a calibration dataset in which each trajectory is drawn independently from \mathcal{D} . Particularly, define

$$D_{cal} := \{Y^{(1)}, \dots, Y^{(n)}\}$$

where the element $Y^{(i)} := (Y_{T_{obs}}^{(i)}, \dots, Y_0^{(i)}, Y_1^{(i)}, \dots, Y_T^{(i)})$ is independently drawn from \mathcal{D} , i.e., $Y^{(i)} \sim \mathcal{D}$.

A key challenge lies in constructing accurate predictions of future values Y_1, \dots, Y_T of the time series. We assume here that we are already given a time series predictor $h : \mathbb{R}^{m(T_{obs}+1)} \rightarrow \mathbb{R}^{mT}$ that maps $(Y_{T_{obs}}, \dots, Y_0)$ to estimates of the next T values (Y_1, \dots, Y_T) , denoted as

$$(\hat{Y}_1, \dots, \hat{Y}_T) := h(Y_{T_{obs}}, \dots, Y_0).$$

We make no assumptions about h . For example, it can be a long short-term memory network (LSTM) (Hochreiter and Schmidhuber 1997) or a sliding linear predictor with extended Kalman filters (Wei et al. 2023). Using the predictor h , we can obtain predictions for the calibration data in D_{cal} . Specifically, for each $Y^{(i)} \in D_{cal}$, we obtain

$$(\hat{Y}_1^{(i)}, \dots, \hat{Y}_T^{(i)}) := h(Y_{T_{obs}}^{(i)}, \dots, Y_0^{(i)}). \quad (1)$$

The predictor h may not be exact, and the inaccuracy of its predictions $(\hat{Y}_1, \dots, \hat{Y}_T)$ is unknown. We will quantify prediction uncertainty using the calibration dataset D_{cal} and split conformal prediction, which we introduce next.

2.2 Split Conformal Prediction

Conformal prediction was introduced in (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008) to obtain valid prediction regions for complex predictive models such as neural networks without making assumptions on the distribution of the underlying data. Split conformal prediction is a computationally tractable variant of conformal prediction (Papadopoulos 2008) where it is assumed that a calibration dataset is available that has not been used to train the predictor. Let $R^{(0)}, \dots, R^{(k)}$ be $k + 1$ exchangeable real-valued random variables.² The variable $R^{(i)}$ is usually referred to as the *nonconformity score*. In supervised learning, it is often defined as $R^{(i)} := \|Z^{(i)} - \mu(X^{(i)})\|$ where the predictor μ attempts to predict the output $Z^{(i)}$ based on the input $X^{(i)}$. Naturally, a large nonconformity score indicates a poor predictive model. Our goal is to obtain a prediction region for $R^{(0)}$ based on the calibration data $R^{(1)}, \dots, R^{(k)}$, i.e., the random variable $R^{(0)}$ should be contained within the prediction region with high probability.

Formally, given a failure probability $\bar{\delta} \in (0, 1)$, we want to construct a valid prediction region $C \in \mathbb{R}$ so that³

$$\text{Prob}(R^{(0)} \leq C) \geq 1 - \bar{\delta}. \quad (2)$$

We pick $C := \text{Quantile}(\{R^{(1)}, \dots, R^{(k)}, \infty\}, 1 - \bar{\delta})$ which is the $(1 - \bar{\delta})$ th quantile of the empirical distribution of the values $R^{(1)}, \dots, R^{(k)}$ and ∞ . Equivalently, by assuming that $R^{(1)}, \dots, R^{(k)}$ are sorted in non-decreasing order and by adding $R^{(k+1)} := \infty$, we can obtain $C := R^{(p)}$ where $p := \lceil (k + 1)(1 - \bar{\delta}) \rceil$ with $\lceil \cdot \rceil$ being the ceiling function, i.e., C is the p th smallest nonconformity score. By a quantile argument, see (Tibshirani et al. 2019, Lemma 1), one can prove that this choice of C satisfies Equation (2). We remark that $k \geq \lceil (k + 1)(1 - \bar{\delta}) \rceil$ is required to hold to obtain meaningful, i.e., bounded, prediction regions. It is known that the guarantees in (2) are marginal over the randomness in $R^{(0)}, R^{(1)}, \dots, R^{(k)}$ as opposed to being conditional on $R^{(1)}, \dots, R^{(k)}$. In fact, the probability $\text{Prob}(R^{(0)} \leq C | R^{(1)}, \dots, R^{(k)})$ is a random variable that follows a beta distribution centered around $1 - \bar{\delta}$ with decreasing variance as k increases, see (Angelopoulos and Bates 2023, Section 3.2) for details. As a result, larger calibration datasets reduce the variance of conditional coverage.

Remark 1. Note that split conformal prediction assumes that the non-conformity scores $R^{(0)}, \dots, R^{(k)}$ are exchangeable. This complicates its use for time series data where Y_{t+1} generally depends on Y_t . However, with a calibration dataset D_{cal} of trajectories independently drawn from \mathcal{D} , the application of conformal prediction is possible (Stankeviciute, M Alaa, and van der Schaar 2021; Lindemann et al. 2023a,b). In summary, one defines T non-conformity scores

²Exchangeability is a slightly weaker form of independent and identically distributed (i.i.d.) random variables.

³More formally, we would have to write $C(R^{(1)}, \dots, R^{(k)})$ as the prediction region C is a function of $R^{(1)}, \dots, R^{(k)}$. For this reason, the probability measure P is defined over the product measure of $R^{(0)}, \dots, R^{(k)}$.

$R_t^{(i)} := \|\hat{Y}_t^{(i)} - Y_t^{(i)}\|$ for each time $t \in \{1, \dots, T\}$ and each calibration trajectory $Y^{(i)} \in D_{cal}$; $R_t^{(i)}$ can be interpreted as the t -step ahead prediction error. Using conformal prediction, for each time t we can then construct a prediction region C_t so that $\text{Prob}(\|\hat{Y}_t - Y_t\| \leq C_t) \geq 1 - \bar{\delta}$. To then obtain a prediction region over all time steps, i.e., $\text{Prob}(\|\hat{Y}_t - Y_t\| \leq C_t, \forall t \in \{1, \dots, T\}) \geq 1 - \delta$, we simply set $\bar{\delta} := \delta/T$. As previously mentioned, this results in conservative prediction regions that are not practical.

3 Problem Formulation and Main Idea

Our goal is to construct valid conformal prediction regions for time series. The main idea is to use a parameterized non-conformity score that considers prediction errors over all T time steps. We consider the non-conformity score

$$R := \max(\alpha_1 R_p(Y_1, \hat{Y}_1), \dots, \alpha_T R_p(Y_T, \hat{Y}_T)) \quad (3)$$

where $\alpha_1, \dots, \alpha_T \geq 0$ are parameters and where R_p is the prediction error for an individual time step. For the purpose of this paper, we use the Euclidean distance between the actual and the predicted value of the time series at time t as

$$R_p(Y_t, \hat{Y}_t) := \|Y_t - \hat{Y}_t\| \quad (4)$$

We note that R_p can in general be any real-valued function of Y_t and \hat{Y}_t . In this paper, however, we set R_p to be the Euclidean distance between Y_t and \hat{Y}_t as in (4).

Taking the maximum prediction error over all T time steps as in equation (4) will allow us to avoid the potentially conservative prediction regions from (Stankeviciute, M Alaa, and van der Schaar 2021; Lindemann et al. 2023a,b) as we, in our proof, do not need to union bound over the individual prediction errors. However, the choice of the parameters $\alpha_1, \dots, \alpha_T$ is now important. If we apply conformal prediction as described in Section 2.2 to the non-conformity score in (3) with $\alpha_1 = \dots = \alpha_T = 1$, this would typically result in large prediction regions for the first time step since prediction errors usually get larger over time, i.e., $R_p(Y_T, \hat{Y}_T)$ in (3) would dominate so that $R = R_p(Y_T, \hat{Y}_T)$. The $\alpha_1, \dots, \alpha_T$ parameters serve to weigh the different time steps, and we generally expect α_t for larger times to be smaller than α_t for smaller times.

Given the parameterized function R as in Equation (3) and a failure threshold of $\delta \in (0, 1)$, we are in this paper particularly interested in computing the parameters $\alpha_1, \dots, \alpha_T$ that minimize the constant C that satisfies

$$\text{Prob}(R \leq C) \geq 1 - \delta.$$

For given parameters $\alpha_1, \dots, \alpha_T$, note that the constant C can be found by applying conformal prediction to the non-conformity score R using the calibration set D_{cal} . When we have found this constant C , we know that

$$\text{Prob}(\|Y_t - \hat{Y}_t\| \leq C/\alpha_t, \forall t \in \{1, \dots, T\}) \geq 1 - \delta.$$

To solve the aforementioned problem, we split the calibration dataset D_{cal} into two sets $D_{cal,1}$ and $D_{cal,2}$. We will use the first calibration dataset $D_{cal,1}$ to compute the values

of $\alpha_1, \dots, \alpha_T$ that give the smallest prediction region C for R (see Section 4). In a next step, we use these parameters that now fully define the nonconformity score R along with the second calibration dataset $D_{cal,2}$ to construct valid prediction regions via conformal prediction (see Section 5).

4 Nonconformity Scores for Time-Series via Linear Programming

We first have to set up some notation. Let us denote the elements of the first calibration dataset as $D_{cal,1} := \{Y^{(1)}, \dots, Y^{(n_1)}\}$ where $n_1 > 0$ is the size of the first calibration dataset. Recall that each element $Y^{(i)}$ of $D_{cal,1}$ is defined as $Y^{(i)} = (Y_{T_{obs}}^{(i)}, \dots, Y_0^{(i)}, Y_1^{(i)}, \dots, Y_T^{(i)})$. Using the trajectory predictor h , we obtain predictions $(\hat{Y}_1^{(i)}, \dots, \hat{Y}_T^{(i)})$ for $(Y_1^{(i)}, \dots, Y_T^{(i)})$ as per equation (1).

We then compute the prediction error $R_t^{(i)}$ according to equation (4) for each calibration trajectory and for each time step, i.e., for each $t \in \{1, \dots, T\}$ and $i \in \{1, \dots, n_1\}$

$$R_t^{(i)} := R_p(Y_t^{(i)}, \hat{Y}_t^{(i)}). \quad (5)$$

We can now cast the problem of finding the parameters $\alpha_1, \dots, \alpha_T$ as the following optimization problem.

$$\begin{aligned} \min_{\alpha_1, \dots, \alpha_T} \quad & \text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, 1 - \delta) \quad (6a) \\ \text{s.t.} \quad & R^{(i)} = \max(\alpha_1 R_1^{(i)}, \dots, \alpha_T R_T^{(i)}), i = 1, \dots, n_1 \quad (6b) \end{aligned}$$

$$\sum_{j=1}^T \alpha_j = 1 \quad (6c)$$

$$\alpha_j \geq 0, j = 1, \dots, T \quad (6d)$$

where $\text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, 1 - \delta)$ denotes the empirical $1 - \delta$ quantile over $R^{(1)}, \dots, R^{(n_1)}$. We removed the value ∞ from Quantile , as required in Section 2.2. However, this only affects the computation of $\alpha_1, \dots, \alpha_T$, but not the correctness of our method as shown later in Section 5. Note also that one could alternatively use $\text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, (1 + 1/n_1)(1 - \delta))$ as it holds that $\text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}, \infty\}, 1 - \delta) = \text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, (1 + 1/n_1)(1 - \delta))$ if $(1 + 1/n_1)(1 - \delta) \in (0, 1)$. Finally, note that Constraint (6d) constrains $\alpha_1, \dots, \alpha_T$ to be non-negative. However, one of the α_t could be zero which would be a problem. Equation (6c) normalizes $\alpha_1, \dots, \alpha_T$ so that this does not happen. Particularly, we show next that the optimal values $\alpha_1, \dots, \alpha_T$ of (6) will be positive as long as all $R_t^{(i)}$ are non-zero.

Theorem 1. *Assume that $R_t^{(i)} > 0$ for all $t \in \{1, \dots, T\}$ and for all $i \in \{1, \dots, n_1\}$. Then, the optimal values $\alpha_1, \dots, \alpha_T$ of (6) are positive, i.e., $\alpha_1^*, \dots, \alpha_T^* > 0$.*

The proof of this theorem, along with subsequent theorems, is in the Appendix of (Cleaveland et al. 2023).

Note that the intuition behind this result is that Equation (6c) assigns a budget for the parameters $\alpha_1, \dots, \alpha_T$ that need to sum to one. If $\alpha_t = 0$, one can increase α_t to allow

another α_τ with $\tau \neq t$ to decrease which will result in lowering the cost in equation (6a). We note that the assumption of having non-zero values of $R_t^{(i)}$ is not limiting as in practice prediction errors are never zero.

Remark 2. *We note that the optimization problem in equation (6) is always feasible. This follows since any set of parameters $\alpha_1, \dots, \alpha_T$ that satisfy Equations (6c) and (6d) constitute a feasible solution to (6).*

To solve the optimization problem in equation (6), we transform it into a linear complementarity program (Cottle, Pang, and Stone 2009) that has the same optimal value as (6). To do so, we first reformulate the $\max()$ operator in equation (6b) as a set of mixed integer linear program (MILP) constraints, which we then transform into a set of linear program constraints. We show that this transformation preserves the optimal value of (6) (Section 4.1). Then, we reformulate the Quantile function from equation (6a) into a linear program (LP) that we reformulate using its KKT conditions (Section 4.2), resulting in a linear complementarity program.

4.1 Reformulating the Max Operator

We first use ideas from (Bemporad and Morari 1999; Raman et al. 2014) for encoding the max operator in (6b) as an MILP. Therefore, we introduce the binary variables $b_t^{(i)} \in \{0, 1\}$ for each time $t \in \{1, \dots, T\}$ and calibration trajectory $i \in \{1, \dots, n_1\}$. The equality constraint $R^{(i)} = \max(\alpha_1 R_1^{(i)}, \dots, \alpha_T R_T^{(i)})$ in equation (6b) is then equivalent to the following MILP constraints:

$$R^{(i)} \geq \alpha_t R_t^{(i)}, t = 1, \dots, T \quad (7a)$$

$$R^{(i)} \leq \alpha_t R_t^{(i)} + (1 - b_t^{(i)})M, t = 1, \dots, T \quad (7b)$$

$$\sum_{t=1}^T b_t^{(i)} = 1 \quad (7c)$$

$$b_t^{(i)} \in \{0, 1\}, t = 1, \dots, T \quad (7d)$$

where $M > 0$ is a sufficiently large and positive constant.⁴

By replacing the max operator in Equation (6b) with the MILP encoding in Equation (7), we arrive at the following optimization problem:

$$\min_{\alpha_t, b_t^{(i)}, R^{(i)}} \quad \text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, 1 - \delta) \quad (8a)$$

$$\text{s.t.} \quad R^{(i)} \geq \alpha_t R_t^{(i)}, i = 1, \dots, n_1 \wedge t = 1, \dots, T \quad (8b)$$

$$R^{(i)} \leq \alpha_t R_t^{(i)} + (1 - b_t^{(i)})M, i = 1, \dots, n_1 \wedge t = 1, \dots, T \quad (8c)$$

$$\sum_{t=1}^T b_t^{(i)} = 1 \quad (8d)$$

$$b_t^{(i)} \in \{0, 1\}, t = 1, \dots, T \quad (8e)$$

$$(6c), (6d)$$

⁴See (Bemporad and Morari 1999). In this case, we need M to satisfy $M \geq \max_{i \in \{1, \dots, n_1\}, t \in \{1, \dots, T\}} R_t^{(i)}$.

The constraints in (8) are linear in their parameters, i.e., in α_t , $R^{(i)}$, and $b_t^{(i)}$. The cost function in (8a) can be reformulated as a linear complementarity program (LCP), as shown in the next section. Thus, (8) is an MILCP since the $b_t^{(i)}$ variables are binary due to equation (8e). Commercial solvers can solve MILCPs, but they present scalability issues for larger time horizons T and calibration data n_1 . Therefore, we present an LCP reformulation of (8) which will achieve the same optimality in terms of the constraint function (8a) while providing better scalability.

Linear complementarity relaxation of the MILCP in (8)
We first remove the upper bound on $R^{(i)}$ in equation (8c) and the binary variables $b_t^{(i)}$ in equations (8d) and (8e) from the MILCP in (8). This results in the optimization problem:

$$\min_{\alpha_t, R^{(i)}} \text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, 1 - \delta) \quad (9a)$$

$$\text{s.t. } R^{(i)} \geq \alpha_t R_t^{(i)}, i = 1, \dots, n_1 \wedge t = 1, \dots, T \quad (9b)$$

$$(6c), (6d)$$

Note that the above optimization problem is now a linear complementarity program due to the removal of the binary variables $b_t^{(i)}$. While the LCP in (9) is clearly computationally more tractable than the MILCP in (8), one may ask in what way the optimal solutions of these two optimization problems are related. We next prove that the optimal values (in terms of the cost functions in (8a) and (9a)) of the two optimization problems (8) and (9) are the same.

Theorem 2. *Let $\delta \in (0, 1)$ be a failure probability and $R_t^{(i)}$ be the prediction errors as in (5) for times $t \in \{1, \dots, T\}$ and calibration trajectories $i \in \{1, \dots, n_1\}$ from $D_{cal,1}$. Then, the optimal cost values of the optimization problems (8) and (9) are equivalent.*

Note that Theorem 2 guarantees that the optimal values of the optimization problems in (8) and (9) are equivalent. However, we do not guarantee that the optimal parameters $\alpha_1^*, \dots, \alpha_T^*$ obtained by solving (8) and (9) will be the same as the optimization problem in (8) may have multiple optimal solutions $\alpha_1^*, \dots, \alpha_T^*$. Nonetheless, the linear reformulation in (9) is computationally tractable and cost optimal. In our experiments in Section 6.1, we also show that we obtain parameters from (8) and (9) that are almost equivalent.

4.2 Reformulating the Quantile Function

We note that the Quantile function in (6a) can be written as the following linear program (Koenker and Bassett 1978):

$$\text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, 1 - \delta)$$

$$= \text{argmin}_q \sum_{i=1}^{n_1} ((1 - \delta)e_i^+ + \delta e_i^-) \quad (10a)$$

$$\text{s.t. } e_i^+ - e_i^- = R^{(i)} - q, i = 1, \dots, n_1 \quad (10b)$$

$$e_i^+, e_i^- \geq 0, i = 1, \dots, n_1 \quad (10c)$$

where the optimal solution q of (10) is equivalent to $\text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, 1 - \delta)$. However, we cannot di-

rectly replace $\text{Quantile}(\{R^{(1)}, \dots, R^{(n_1)}\}, 1 - \delta)$ in the optimization problem in (6a) with the linear program from equation (10) as that would result in a problem with an objective function that is itself an optimization problem. To address this issue, we replace the optimization problem in (10) with its KKT conditions, which we provide in equation (17) in Appendix C due to space limitations.

The KKT conditions form an LCP because every constraint is either linear ((17a)-(17c) and (17f)-(17j)) or an equality constraint stating the multiple of two positive variables equals 0 ((17d) and (17e)).⁵ Despite the slight non-linearity, LCPs can be efficiently solved with existing tools (Yu, Mitchell, and Pang 2019). Because feasible linear optimization problems have zero duality gap (Bradley, Hax, and Magnanti 1977), any solution to the KKT conditions in (17) is also an optimal solution to Problem (10).

4.3 Summarizing the Optimization Programs

Based on the results from the previous subsections, we can formulate two optimization problems that solve (6).

First, we can exactly solve (6) by replacing the max operator in (6b) with the MILP in (7) and the quantile function in (6a) with its KKT conditions in (17). The following result follows by construction (for sufficiently large M).

Proposition 1. *The MILCP*

$$\min_{q, \alpha_t, R^{(i)}, b_t^{(i)}} q \quad (11a)$$

$$\text{s.t. } (8b), (8c), (8d), (8e)$$

$$(17a), (17b), (17c), (17d), (17e)$$

$$(17f), (17g), (17h), (17i), (17j)$$

$$(6c), (6d)$$

is equivalent to the optimization problem in (6).

Second, by further removing the upper bound from the max operator as described in Section 4.1, we can solve a linear complementarity program that has the same optimal cost value as (6). This result follows from Theorem 3.

Proposition 2. *The LCP*

$$\min_{q, \alpha_t, R^{(i)}} q \quad (12a)$$

$$\text{s.t. } (8b)$$

$$(17a), (17b), (17c), (17d), (17e)$$

$$(17f), (17g), (17h), (17i), (17j)$$

$$(6c), (6d)$$

has the same optimal value as the optimization problem (6).

5 Conformal Prediction Regions for Time-Series

Finally, after obtaining the parameters $\alpha_1, \dots, \alpha_T$ from the first calibration dataset $D_{cal,1}$ by solving (11) or (12), we can apply conformal prediction with the non-conformity score R as defined in (3) to obtain prediction regions for

⁵These constraints ensure that one of the variables equals 0.

time series. We do so by following Section 2.2 and using the second calibration dataset $D_{cal,2}$.

More formally, let us denote the elements of the second calibration dataset as $D_{cal,2} := \{Y^{(n_1+1)}, \dots, Y^{(n)}\}$ where $n - n_1 > 0$ is the size of the second calibration dataset.⁶ For each calibration trajectory $i \in \{n_1 + 1, \dots, n\}$, we compute the nonconformity score as

$$R^{(i)} := \max \left(\alpha_1 R_p(Y_1^{(i)}, \hat{Y}_1^{(i)}), \dots, \alpha_T R_p(Y_T^{(i)}, \hat{Y}_T^{(i)}) \right). \quad (13)$$

We are now ready to state the final result of our paper in which we obtain prediction regions for $R := \max(\alpha_1 R_p(y_1, \hat{y}_1), \dots, \alpha_T R_p(y_T, \hat{y}_T))$.

Theorem 3. *Let Y be a trajectory drawn from \mathcal{D} and let $\hat{Y} = h(Y_{T_{obs}}, \dots, Y_0)$ be predictions from the time series predictor h . Let $\delta \in (0, 1)$ be a failure probability and $R_t^{(i)}$ be the prediction error as in (5) for times $t \in \{1, \dots, T\}$ and calibration trajectories $i \in \{1, \dots, n_1\}$ from $D_{cal,1}$. Let $\alpha_1, \dots, \alpha_T$ be obtained by solving (11) or (12). Further, let $R^{(i)}$ be the nonconformity score as in (13) for calibration trajectories $i \in \{n_1 + 1, \dots, n\}$ from $D_{cal,2}$. Finally, let $R = \max(\alpha_1 R_p(Y_1, \hat{Y}_1), \dots, \alpha_T R_p(Y_T, \hat{Y}_T))$. Then*

$$\text{Prob}(R \leq C) \geq 1 - \delta \quad (14)$$

where $C := \text{Quantile}(\{R^{(n_1+1)}, \dots, R^{(n)}, \infty\}, 1 - \delta)$.

We point out the necessity to split the dataset D_{cal} into the datasets $D_{cal,1}$ and $D_{cal,2}$ to conform to the exchangeability assumption in conformal prediction. To convert equation (14) into practical conformal prediction regions, note that

$$R = \max \left(\alpha_1 \|Y_1 - \hat{Y}_1\|, \dots, \alpha_T \|Y_T - \hat{Y}_T\| \right) \leq C \\ \Leftrightarrow \|Y_t - \hat{Y}_t\| \leq C/\alpha_t \quad \forall t \in \{1, \dots, T\}$$

Intuitively, the conformal prediction regions provide geometric balls of radius C/α_t around each prediction \hat{y}_t due to the use of the Euclidean norm. For an example, we refer to Figure 1. We remark that other choices of the prediction R_p may lead to different shapes. We summarize this result next.

Corollary 1. *Let all conditions of Theorem 3 hold. Then*

$$\text{Prob}(\|Y_t - \hat{Y}_t\| \leq C/\alpha_t \quad \forall t \in \{1, \dots, T\}) \geq 1 - \delta.$$

6 Case Study

We consider two case studies to evaluate our approach. The first case study is about predicting pedestrian paths, while the second case study is about predicting the altitude of an F16 fighter jet performing ground avoidance maneuvers. We compare our method to the approach from (Stankeviciute, M Alaa, and van der Schaar 2021; Lindemann et al. 2023a,b) that is described in Section 2.2, which requires conservative union bounding to obtain valid prediction regions. We colloquially refer to this method as the union bound approach. We use Gurobi (Gurobi Optimization, LLC 2023)

⁶Recall the n is the size of D_{cal} and n_1 is the size of $D_{cal,1}$.

to solve the optimization problems on a Windows machine running an Intel i7-8550U CPU with 4 cores and 16GB of RAM. The code for the case studies can be found at <https://github.com/earnedkibbles58/timeParamCPScores> and additional figures can be found in the Appendix of (Cleaveland et al. 2023). Throughout this section, we use (12) to compute the α values and we set $\delta := 0.05$.

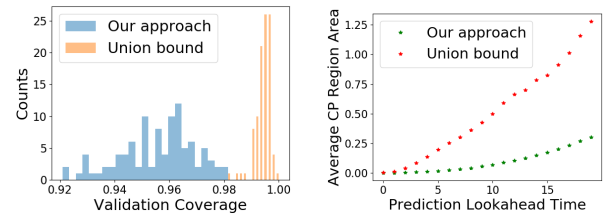
6.1 ORCA Pedestrian Simulator

For the first case study, we analyze pedestrian location predictions using data generated from the ORCA simulator (Van den Berg, Lin, and Manocha 2008). We use the social LSTM (Alahi et al. 2016) architecture to make predictions up to 2.5 seconds into the future at a rate of 8Hz (so that $T = 20$ predictions in total) with the previous 2.5 seconds as input (so that $T_{obs} = 20$) using the trajnet++ framework (Kothari, Kreiss, and Alahi 2021). For the exact details of training, we refer the reader to (Lindemann et al. 2023a).

We collect a dataset of 1291 trajectories. We randomly select 646 of these points for D_{cal} and use the rest to form D_{val} , which we use for validation. We randomly select 50 trajectories from D_{cal} for $D_{cal,1}$ (to compute the α values) and the remaining trajectories for $D_{cal,2}$ (to compute the conformal prediction regions). We first show the results for one realization of this data sampling process and then we run 100 trials to statistically analyze the results.

Statistical Analysis For one realization, trajectories of the actual and predicted pedestrian locations are shown alongside the conformal prediction regions in Figure 1, the α values are shown in Appendix E, and scatter plots of the prediction errors over D_{val} are shown in Appendix F.

Over 100 trials, the average (trajectory level) coverage for our approach was 0.9558 and the average coverage for the union bound approach was 0.9945. Histograms of the validation coverage for the two approaches are shown in Figure 2a. The union bound approach has larger coverage due to the conservatism introduced by union bounding. We also computed the average size of the conformal regions for each time horizon, which are shown in Figure 2b. This highlights that our method produces valid conformal prediction regions that are significantly smaller than previous approaches.



(a) Validation coverage histograms (b) Average area of the conformal prediction regions

Figure 2: Comparison of our approach and the union bound approach over 100 trials for the ORCA case study.

Original v. Relaxed Problem Difference For each trial we also computed the objective function and α values using

Equation (11) to compare to those computed using Equation (12). The average difference between the objective function values is 1.847×10^{-9} . The average l^2 norm difference between the α values is 3.495×10^{-7} and the largest l^2 norm difference is 3.453×10^{-5} . These differences are well within the solver’s default tolerance of 10^{-4} , which empirically validates the equivalence from Theorem 2. The average run-time for solving the original problem was 0.525 seconds and the relaxed problem was 0.207 seconds.

6.2 F16 Fighter Jet

For the second case study, we analyze the F16 fighter jet ground avoidance maneuver from the open source simulator (Heidlauf et al. 2018). We again use an LSTM to predict the altitude of the F16 fighter jet up to 2.5 seconds into the future at a rate of 10Hz (so that is $T = 25$ predictions in total) with the altitude from the past 2.5 seconds as input. The trajectory predictor consists of an LSTM with two layers of width 25 neurons each followed by two Relu layers with 25 neurons each and a final linear layer. We used 1500 trajectories, each 5 seconds in duration, to train the network.

We collect a dataset of 1900 trajectories, all of length 5 seconds. We randomly select 1000 trajectories for D_{cal} and the rest for D_{val} . We randomly select either 100, 200 or 400 trajectories from D_{cal} for $D_{cal,1}$ and the remaining trajectories in D_{cal} for $D_{cal,2}$. We run 100 trials to statistically analyze the approach, just as in Section 6.1.

Statistical Analysis The average (trajectory level) coverage for our approach was 0.953 for each of $|D_{cal,1}| = 400, 200, 100$ and the average coverage for the union bound approach was 0.993. Histograms of the validation coverages for the union bound approach and our approach with $|D_{cal,1}| = 400$ are shown in Figure 3a. The union bound approach has larger coverage due to the conservatism introduced by union bounding. For one realization trajectories of the actual and predicted altitudes along with the conformal prediction regions for our approach using $|D_{cal,1}| = 400$ and the union bound approach are shown in Appendix G.

The variance of the coverage for our approach using $|D_{cal,1}| = 400, 200, 100$ was $9.04 \times 10^{-5}, 7.63 \times 10^{-5},$ and 5.89×10^{-5} , while the variance of the coverage for the union bound approach was 1.51×10^{-5} . For our approach, note that using larger $D_{cal,1}$ results in slightly higher variance for the coverage. That is because one is left with less data in $D_{cal,2}$ for running the conformal prediction step.

The average width of the conformal regions for each time horizon are shown in Figure 3b. This demonstrates that our approach results in smaller prediction regions. Additionally, note that the prediction regions get slightly smaller as one uses more data for $D_{cal,1}$. That is because more data allows for selecting better α values. This highlights that there is a tradeoff in how much data one uses in $D_{cal,1}$ as opposed to $D_{cal,2}$. Making $D_{cal,1}$ larger will result in slightly smaller prediction regions, but the coverage of the regions will have higher variance, due to the smaller amount of data available in $D_{cal,2}$ for running the conformal prediction step.

Original v. Relaxed Problem Difference We also computed the α values using Equation (11) to compare to those

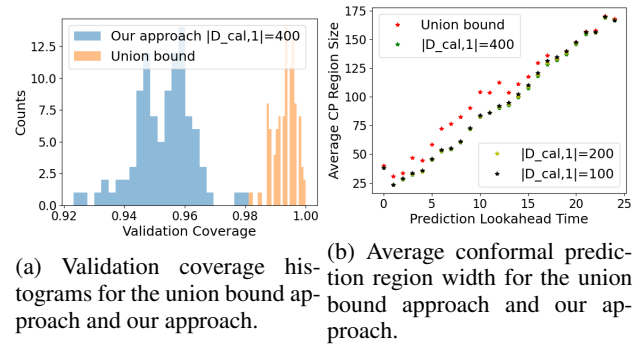


Figure 3: Comparison of our approach and the union bound approach over 100 trials for the F16 case study.

computed using Equation (12). In this case study, they were exactly the same, which empirically validates Theorem 2.

6.3 Case Study Comparison

Our method produced bigger relative improvements over the union bounding approach for the ORCA case study than for the F16 case study, see Figs. 2b and 3b. We hypothesize that this is due to the distributions of the non-conformity scores having longer tails in the ORCA case study, see histograms of individual prediction errors for each case study in Appendix H. This explains the larger difference in the ORCA case study between our approach, which looks at the $1 - \delta$ quantile, and the union bounding approach, which looks at the $1 - \delta/T$ quantile. The $1 - \delta$ and $1 - \delta/T$ quantiles for the two case studies are shown in Figure 4. The difference in quantile values is much larger for ORCA than the F16 case study, confirming our hypothesis.

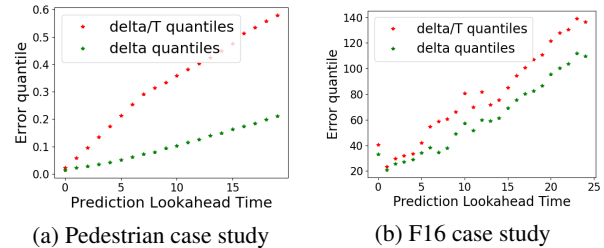


Figure 4: Non-conformity score quantiles for the calibration and testing data for the two case studies.

7 Conclusion

In this paper, we have presented a method for producing conformal prediction regions for time series data that are significantly tighter than previous approaches. To do this, we define a parameterized non-conformity score function and use optimization to fit the function to calibration data. This allows us to use standard inductive conformal prediction to get valid prediction regions. For future work, we plan to integrate our conformal prediction approach into planning and control frameworks and apply parameterized non-conformity score functions to the non-time series setting.

Acknowledgements

This work was generously supported by NSF award SLES-2331880.

References

- Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 961–971.
- Angelopoulos, A. N.; and Bates, S. 2023. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.*, 16(4): 494–591.
- Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2022. Conformal risk control. *arXiv preprint arXiv:2208.02814*.
- Bemporad, A.; and Morari, M. 1999. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3): 407–427.
- Bhatnagar, A.; Wang, H.; Xiong, C.; and Bai, Y. 2023. Improved Online Conformal Prediction via Strongly Adaptive Online Learning. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Bradley, S.; Hax, A.; and Magnanti, T. 1977. *Applied Mathematical Programming*. Addison-Wesley Publishing Company. ISBN 9780201004649.
- Cleaveland, M.; Lee, I.; Pappas, G. J.; and Lindemann, L. 2023. Conformal Prediction Regions for Time Series using Linear Complementarity Programming. *arXiv preprint arXiv:2304.01075*.
- Cohen, K. M.; Park, S.; Simeone, O.; Popovski, P.; and Shamai, S. 2023. Guaranteed Dynamic Scheduling of Ultra-Reliable Low-Latency Traffic via Conformal Prediction. *IEEE Signal Processing Letters*.
- Cottle, R. W.; Pang, J.-S.; and Stone, R. E. 2009. *The linear complementarity problem*. SIAM.
- Dixit, A.; Lindemann, L.; Wei, S.; Cleaveland, M.; Pappas, G. J.; and Burdick, J. W. 2022. Adaptive Conformal Prediction for Motion Planning among Dynamic Agents.
- Fannjiang, C.; Bates, S.; Angelopoulos, A. N.; Listgarten, J.; and Jordan, M. I. 2022. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43).
- Gibbs, I.; and Candes, E. 2021. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34: 1660–1672.
- Gibbs, I.; and Candès, E. 2022. Conformal inference for online prediction with arbitrary distribution shifts. *arXiv preprint arXiv:2208.08401*.
- Gurobi Optimization, LLC. 2023. Gurobi Optimizer Reference Manual.
- Heidlauf, P.; Collins, A.; Bolender, M.; and Bak, S. 2018. Reliable Prediction Intervals with Directly Optimized Inductive Conformal Regression for Deep Learning. *5th International Workshop on Applied Verification for Continuous and Hybrid Systems (ARCH 2018)*.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.
- Kaur, R.; Jha, S.; Roy, A.; Park, S.; Dobriban, E.; Sokolsky, O.; and Lee, I. 2022. iDECODe: In-Distribution Equivariance for Conformal Out-of-Distribution Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7): 7104–7114.
- Kaur, R.; Sridhar, K.; Park, S.; Yang, Y.; Jha, S.; Roy, A.; Sokolsky, O.; and Lee, I. 2023. CODiT: Conformal Out-of-Distribution Detection in Time-Series Data for Cyber-Physical Systems. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, 120–131.
- Koenker, R.; and Bassett, G. 1978. Regression Quantiles. *Econometrica*, 46(1): 33–50.
- Kothari, P.; Kreiss, S.; and Alahi, A. 2021. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*.
- Lindemann, L.; Cleaveland, M.; Shim, G.; and Pappas, G. J. 2023a. Safe Planning in Dynamic Environments Using Conformal Prediction. *IEEE Robotics and Automation Letters*, 8(8): 5116–5123.
- Lindemann, L.; Qin, X.; Deshmukh, J. V.; and Pappas, G. J. 2023b. Conformal prediction for stl runtime verification. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, 142–153.
- Luo, R.; Zhao, S.; Kuck, J.; Ivanovic, B.; Savarese, S.; Schmerling, E.; and Pavone, M. 2022. Sample-efficient safety assurances using conformal prediction. In *Algorithmic Foundations of Robotics XV: Proceedings of the Fifteenth Workshop on the Algorithmic Foundations of Robotics*, 149–169. Springer.
- Papadopoulos, H. 2008. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer.
- Park, S.; Bastani, O.; Matni, N.; and Lee, I. 2020. PAC Confidence Sets for Deep Neural Networks via Calibrated Prediction. In *International Conference on Learning Representations*.
- Raman, V.; Donzé, A.; Maasoumy, M.; Murray, R. M.; Sangiovanni-Vincentelli, A.; and Seshia, S. A. 2014. Model predictive control with signal temporal logic specifications. In *53rd IEEE Conference on Decision and Control*, 81–87. IEEE.
- Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, 32.
- Shafer, G.; and Vovk, V. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3).
- Stankeviciute, K.; M Alaa, A.; and van der Schaar, M. 2021. Conformal time-series forecasting. *Advances in Neural Information Processing Systems*, 34: 6216–6228.
- Stutz, D.; Dvijotham, K. D.; Cemgil, A. T.; and Doucet, A. 2022. Learning Optimal Conformal Classifiers. In *International Conference on Learning Representations*.

- Sun, S.; and Yu, R. 2022. Copula Conformal Prediction for Multi-step Time Series Forecasting. arXiv:2212.03281.
- Tibshirani, R. J.; Foygel Barber, R.; Candès, E.; and Ramdas, A. 2019. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.
- Tonkens, S.; Sun, S.; Yu, R.; and Herbert, S. 2023. Scalable Safe Long-Horizon Planning in Dynamic Environments Leveraging Conformal Prediction and Temporal Correlations. In *Long-Term Human Motion Prediction Workshop, International Conference on Robotics and Automation*.
- Tumu, R.; Lindemann, L.; Nghiem, T.; and Mangharam, R. 2023. Physics Constrained Motion Prediction with Uncertainty Quantification. *arXiv preprint arXiv:2302.01060*.
- Van den Berg, J.; Lin, M.; and Manocha, D. 2008. Reciprocal velocity obstacles for real-time multi-agent navigation. In *2008 IEEE international conference on robotics and automation*, 1928–1935. Ieee.
- Vovk, V. 2012. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, 475–490. PMLR.
- Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.
- Wei, S. X.; Dixit, A.; Tomar, S.; and Burdick, J. W. 2023. Moving Obstacle Avoidance: A Data-Driven Risk-Aware Approach. *IEEE Control Systems Letters*, 7: 289–294.
- Wisniewski, W.; Lindsay, D.; and Lindsay, S. 2020. Application of conformal prediction interval estimations to market makers’ net positions. In *International Symposium on Conformal and Probabilistic Prediction with Applications*.
- Xu, C.; and Xie, Y. 2021. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, 11559–11569. PMLR.
- Yu, B.; Mitchell, J. E.; and Pang, J.-S. 2019. Solving linear programs with complementarity constraints using branch-and-cut. *Mathematical Programming Computation*, 11(2): 267–310.
- Zaffran, M.; Féron, O.; Goude, Y.; Josse, J.; and Dieuleveut, A. 2022. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, 25834–25866. PMLR.