Efficient Quantification of Time-Series Prediction Error: Optimal Selection Conformal Prediction

Boyu Pang and Kostas Margellos

***Abstract*— Uncertainty is almost ubiquitous in safety-critical autonomous systems due to dynamic environments and the integration of learning-based components. Quantifying this uncertainty–particularly for time-series predictions in multi-stage optimization–is essential for safe control and verification tasks. Conformal Prediction (CP) is a distribution-free uncertainty quantification tool with rigorous finite-sample guarantees, but its performance relies on the design of the nonconformity measure, which remains challenging for time-series data. Existing methods either overfit on small datasets, or are computationally intensive on long-time-horizon problems and/or large datasets. To overcome these issues, we propose a new parameterization of the score functions and formulate an optimization program to compute the associated parameters. The optimal parameters directly lead to norm-ball regions that constitute minimal-average-radius conformal sets. We then provide a reformulation of the underlying optimization program to enable faster computation. We provide theoretical proofs on both the validity and efficiency of predictors constructed based on the proposed approach. Numerical results on various case studies demonstrate that our method outperforms state-of-the-art methods in terms of efficiency, with much lower computational requirements.**

I. Introduction

Uncertainty is almost ubiquitous in safety-critical autonomous systems. The dynamic nature of external environments (e.g., autonomous driving) and the incorporation of learning-based methods (e.g., neural-networks) introduce uncertainties into the systems, which pose new challenges to safe controller design and verification. To address this issue, one way is to quantify uncertainty, in particular, for the time-series predictions that arise in multi-stage optimization problems. We classify uncertainty quantification methods into two main streams:

a) Bayesian methods and concentration-bounds: Uncertainty quantification using Bayesian methods include Bayesian Inference, Bayesian Neural Network and other variants [1]–[3]. Alternative approaches include concentrations bounds, such as Chernoff-Hoeffding (e.g., [4]), Clopper-Pearson (e.g., [5]). A more detailed review can be found in Section 2.3 of the survey paper [6]. However, Bayesian methods do not have finite-sample guarantees and become computationally intractable for large-scale problems; concentration bounds can be conservative in uncertainty quantification. An alternative to these methods is conformal prediction, which are presented next.

b) Conformal Prediction: Conformal Prediction (CP) constitutes a statistical framework that has been introduced [7] as a distribution-free and statistically rigorous tool to predict a $(1-\alpha)$-confidence region for the trained prediction model, with $\alpha \in (0,1)$. Under mild assumptions on a calibration dataset, CP provides tight finite-sample guarantees on both marginal coverage and conditional coverage. Although closely related with scenario optimization [8], a tool that has also been used for tight uncertainty quantification [9], CP focuses on a different aspect and thus complements the scenario approach (see [10] for some connections between CP and the scenario approach). Thanks to its simplicity, flexibility, and computational-efficiency, CP has been applied in probabilistic safe control synthesis and verification, such as moving-objects avoidance control [11]–[14], probabilistic reachability analysis [15]–[17], probabilistic reachable sets construction [18]–[20].

However, while the confidence region produced by CP comes with rigorous probabilistic guarantees, its performance (e.g., size and shape of the region) still depends critically on one of its core components—the nonconformity measure (also called score function). Designing an appropriate score function for time-series uncertainty quantification is non-trivial. Existing works [21]–[24] either suffer from 1) overly conservative confidence regions [21], 2) overfitting or fitting-errors [22], [23], or 3) computational intractability issues [23], [24]. To the best of our knowledge, no work seems to alleviate these issues at the same time.

In this paper, we aim to overcome these challenges when using CP for time series, thus opening the road for its use in multi-stage optimization and safety problems that require uncertainty quantification over entire trajectories rather than single time-steps. We propose a new parameterized score function that can be optimized to provide minimal-average-radius CP regions. Our CP method generates norm-ball regions, that are convex and as we will show also tight, for multi-dimensional time series and exhibits lower computational requirements compared to other algorithmic alternatives. Our proposed approach is directly applicable to control problems such as safe learning-based MPC [11]–[14] and multi-stage safety verification [25].

Our main contributions can be summarized as:

1) We propose a new parameterized non-conformity measure for calibrating multi-dimensional time-series data in CP, and a mixed-integer linear programming (MILP) problem to determine optimal parameter solutions. We then provide a re-formulation of this MILP with fewer

The authors are with the Department of Engineering Science, University of Oxford, Oxford, United Kingdom. E-mails: {boyu.pang, kostas.margellos}@eng.ox.ac.uk

constraints to enable faster computation times.
2) We prove that our method is *valid* (concept at the core of CP); we also prove that the optimal parameters result in determining the minimum average-radius conformal set for any pre-specified normed-ball region.
3) We evaluate the efficacy of our approach numerically on 4 case studies and show that it produces valid conformal regions with the smallest size among baselines [21]–[24], [26]. Specifically, the results suggest that our proposed approach reduces the conformal set size by 16.03%, 14.32%, 14.01%, 16.93% on the 4 case studies, respectively, compared to the previous State-of-the-Art (SOTA) method [24].
4) Our optimization program runtime requirements are mild; compared to previous SOTA [24], it leads to 8812.0, 78622.0, 14.4, 22.1 times faster computation on the 4 benchmark studies we have investigated numerically, respectively.

The remainder of this paper is organized as follows: Section II introduces the problem setting and the conformal prediction. In Section III we formally propose our approach, which we term Optimal Selection Conformal Prediction (OSCP). Then Section IV compares our method with 5 baseline methods via numerical experiments on 3 synthetic datasets and 1 real dataset. Finally, Section V concludes the study.

II. Problem Setting and Conformal Prediction Preliminaries

A. Problem Setting

In a discrete-time control system, let $\hat{\mathbf{Y}}_{0:T-1} = (\hat{Y}_0, ..., \hat{Y}_{T-1}) \in \mathcal{Y} \subseteq \mathbb{R}^{d\times T}$ and $\mathbf{Y}_{0:T-1} = (Y_0, ..., Y_{T-1}) \in \mathcal{Y} \subseteq \mathbb{R}^{d\times T}$ denote the nominal (predicted) and true trajectories of a parameter Y_t evolving over a horizon of T time-steps. For example, $\mathbf{Y}_{0:T-1}$ can be the trajectory of a moving obstacle to be avoided, while $\hat{\mathbf{Y}}_{0:T-1}$ is the predicted trajectory given by a neural-network. As another example, $\hat{\mathbf{Y}}_{0:T-1}$ can be the system state trajectory provided by the nominal system model which does not account for noise/disturbance, and thus different from the real trajectory $\mathbf{Y}_{0:T-1}$. Let $\tilde{\mathbf{Y}}_{0:T-1} := (\tilde{Y}_0, ..., \tilde{Y}_{T-1})$ be the residual sequence capturing the error between the nominal and the true trajectory, i.e., $\tilde{Y}_t := Y_t - \hat{Y}_t$.

We stipulate that the residual sequence $\tilde{\mathbf{Y}}_{0:T-1}$ is a random quantity distributed according to a probability measure $\mathbb{P}$. We assume that the corresponding probability space is defined as appropriate.

We assume throughout that we are given an *exchangeable* calibration dataset $D_{\text{cal}} = \{\tilde{\mathbf{Y}}^{(i)}_{0:T-1}\}_{i=1}^{N}$ containing the residual sequences of historical trajectories, where the term exchangeability is defined as follows:

Definition 1 (Exchangeability)**.** *A collection of N random variables is said to be exchangeable if the joint probability distribution of any permutation of these N random variables are the same.*[1]

[1]Note that exchangeability is a weaker condition compared to assuming that data are independent and identically distributed (i.i.d.).

It should be noted that we only assume that the multiple complete T-horizon sequences are exchangeable, without requiring exchangeability within time-horizons; i.e. we do not assume $\tilde{Y}_t$ and $\tilde{Y}_{t'}$ are exchangeable or i.i.d. for a given $0 \le t \ne t' < T$, as residuals at different time steps may exhibit temporal correlation. Another important remark is that if the nominal trajectory is generated from a data-driven model (e.g., neural-network), the calibration dataset D_{cal} must not involve any residual of training data, as we have assumed that all data come from the same distribution and such an operation would alter it.

For a pre-defined error level $\epsilon \in (0, 1)$, our goal is to use D_{cal} to construct a set-value predictor Γ^ϵ that predicts a closed and bounded abstraction region $\mathcal{C}_t$ (such as a norm-ball) around each $\hat{Y}_t$ such that with at least $(1-\epsilon)$ probability, the true trajectory $\mathbf{Y}_{0:T-1}$ is completely inside these abstraction regions simultaneously for each $t = 0, \dots, T-1$.

More formally, we want to use D_{cal} to construct a set-valued predictor

$$\Gamma^\epsilon : \hat{\mathbf{Y}}_{0:T-1} \mapsto \otimes_{t=0}^{T-1} \mathcal{C}_t \subset \mathcal{Y} \tag{1}$$

that produces T decoupled and *valid* (Def. 2) abstraction regions for $\mathbf{Y}_{0:T-1}$, where the term validity is defined as follows.

Definition 2 (Validity, [7])**.** *Given a desired error level $\epsilon \in (0, 1)$, a statistical abstraction predictor Γ^ϵ is said to be valid if for any new trajectory $\mathbf{Y}^{(\text{new})}_{0:T-1}$, we have*

$$\mathbb{P}\left(\mathbf{Y}^{(\text{new})}_{0:T-1} \in \Gamma^\epsilon(\hat{\mathbf{Y}}^{(\text{new})}_{0:T-1})\right) \ge 1 - \epsilon, \tag{2}$$

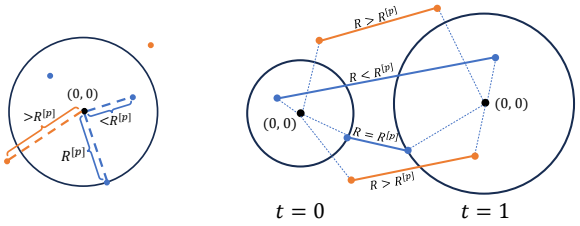
where $\mathbb{P}$ is the joint probability measure of the new residual sequence $\tilde{\mathbf{Y}}^{(\text{new})}_{0:T-1}$ and the calibration data D_{cal}. In the setting of this paper, (2) is equivalent to:

$$\mathbb{P}\left(Y^{(\text{new})}_t \in \mathcal{C}_t \text{ for all } t = 0, ..., T-1\right) \ge 1 - \epsilon. \tag{3}$$

B. Conformal Prediction and ICP Framework

Conformal Prediction (CP) is a model-agnostic and distribution-free tool that aims to quantify prediction uncertainty and produce valid prediction sets Γ^ϵ without assumptions on the data distribution. The most commonly used CP methods employ a computation-friendly framework called Inductive-Conformal-Prediction (ICP, see [7]), sometimes also called Split-Conformal-Prediction. Such framework assumes that the data in calibration set is exchangeable with the test data. Then we use a score function (non-conformity measure) $\mathcal{A}$ to assign each calibration data with a non-conformity score R_i, and find the pth smallest score with $p = \lceil (1-\epsilon)(N+1) \rceil$. Then any data point with score smaller or equal to this pth smallest score is in the *valid* conformal region.

One crucial challenge is how to define the non-conformity measure $\mathcal{A}$. This is in fact non-trivial, as using a different $\mathcal{A}$ will induce CP-regions with different shapes and sizes. A CP method that has a too large CP-region is conservative and does not provide meaningful results. To evaluate the performance of a CP method, we use the term *efficiency*:



(a) 2D-regression example (b) time-series data with $d = 2, T = 2$

Fig. 1: Data with non-conformity scores lower or equal to $R^{[p]}$ are drawn in blue, otherwise in orange. A valid CP contains at least p number of residuals inside CP regions. Motivated by simple regression case in (a), we minimize the average radius of normed-ball regions that containing p number of time-series residuals inside.

Definition 3 (Efficiency, [7])**.** *Given an efficiency metric $\mathcal{L}_{\text{eff}}$, an error level ϵ, and a fixed input $\hat{\mathbf{Y}}_{0:T-1}$, a CP method Γ_1^ϵ is said to be more efficient than another CP method Γ_2^ϵ if*

$$\mathcal{L}_{\text{eff}}\left(\Gamma_1^\epsilon(\hat{\mathbf{Y}}_{0:T-1})\right) < \mathcal{L}_{\text{eff}}\left(\Gamma_2^\epsilon(\hat{\mathbf{Y}}_{0:T-1})\right). \tag{4}$$

In the context of time-series setting where we need to produce a sequence of T decoupled regions, the efficiency metric $\mathcal{L}_{\text{eff}}$ is usually taken as the sum-of-widths (diameters) or sum-of-volumes (Lebesgue measure) of the T CP-regions.

III. Optimal Selection Conformal Prediction (OSCP)

In this paper, we propose a new CP method for time series using the Re-calibrate ICP framework [22]–[24], which splits the D_{cal} into 2 halves, one for learning the parameters in a parameterized score function and another for calibration (as in the standard CP). We propose a novel parameterization that provides very tight conformal regions in theory and achieves SOTA on the experiments in Section IV. We also show that the computation of this method is much more efficient than the previous SOTA method.

A. Motivation: Minimal-Average-Radius Regions Containing p Residuals

In the classical ICP framework for a simple regression task of predicting a 2-D vector $\vec{y} = (y_1, y_2)^\top$, the score function can be simply defined as the l_2-norm of the residuals $\tilde{y}$, i.e. $\mathcal{A}(\tilde{y}) := ||\tilde{y}||_2 = ||\vec{y}_{\text{real}} - \vec{y}_{\text{pred}}||_2$. Suppose with probability-one, there are no ties (i.e., the score of each data is distinct almost surely). Now, if we draw a circle centered at the origin and plot each residual vector $\tilde{y}$ of the calibration data on the graph (see Figure 1a), then we can see that constructing the CP-region is equivalent to constructing the smallest circle that contains exactly p number of residual vectors (either in its interior or on its boundary) with $p = \lceil (1-\epsilon)(N+1) \rceil$. This is because the radius of this circle is equal to the p-th smallest score, which we denote by $R^{[p]}$.

Given this fact, we now consider a simple time-series setting with $d = 2$ & $T = 2$ (Figure 1b). At each time step, we first fix a local "center" point, and plot residual vector $\tilde{Y}_t^{(i)} := Y_t^{(i)} - \hat{Y}_t^{(i)}$ with respect to this center point (think of it as the origin in x-y plane). The solid lines connecting two points denote the residuals at 2 time steps from the same time-series data, while the time series is inside the CP region if both ends of this line segment are inside the respective circles (connected via dashed lines). The first fact is that there are at least p residuals of calibration data inside the CP regions (in fact there are exactly p residuals when there are no "ties"). The question we seek to answer is whether we can perform a similar procedure with the single-stage regression case above. That is, construct CP-regions that give rise to norm-balls with the smallest average radius while containing at least p residuals inside. We provide a positive answer to this question, and refer to our proposed method to achieve this as Optimal Selection Conformal Prediction (OSCP).

B. OSCP: Algorithm Description, Validity and Efficiency

Considering the motivation of constructing smallest regions containing p calibration data, we now present the Optimal Selection Conformal Prediction (OSCP) method using the Re-calibrate ICP framework and the Empirical-Risk-Minimization principle. This method is compatible with the common convex shapes (hyper-ball, hyper-cube, hyper-ellipsoid, etc.) for CP regions, which depends on the specific norm the user is using to calculate the absolute residual at each time step. To make the statement clearer, we first define the normed-residual-series of a residual sequence between nominal and real time-series as follows:

Definition 4 (Normed-residual-series $\tilde{\boldsymbol{\epsilon}}$)**.** *Given any norm $||\cdot|| : \mathbb{R}^d \to \mathbb{R}$, the normed-residual-series $\tilde{\mathbf{e}}_{0:T-1}^{(i)} = (\tilde{e}_0^{(i)}, ..., \tilde{e}_{T-1}^{(i)})^\top$ for a residual sequence $\tilde{\mathbf{Y}}_{0:T-1}^{(i)}$ is defined as a time series*

$$\tilde{\mathbf{e}}_{0:T-1}^{(i)} := (||\tilde{Y}_0^{(i)}||, ..., ||\tilde{Y}_{T-1}^{(i)}||)^\top \in \mathbb{R}^T. \tag{5}$$

Using a different norm will induce different shapes of CP region for this method, which will be discussed later. Often, a natural choice is to use the l_2-norm.

a) Step 1: Split data: We conform to the Re-calibrate ICP framework, which requires to further split the calibration dataset. Suppose we have an *exchangeable* dataset D_{cal} drawn from $\mathbb{P}$, we split it into two disjoint subsets $D_{\text{cal},1}$, $D_{\text{cal},2}$ with size n_1, n_2. Although there is no requirement on how to split the dataset, in our numerical implementation we use $n_1 \approx n_2$.

b) Step 2: Determine the optimal score function: Given a norm to calculate *normed-residual-series* of the calibration data, the parameterized score function $\mathcal{A}$ is defined as

$$\mathcal{A}(\tilde{\mathbf{Y}}_{0:T-1}^{(i)}) := \max\left\{\tilde{e}_0^{(i)} - r_0, ..., \tilde{e}_{T-1}^{(i)} - r_{T-1}\right\}, \tag{6}$$

where $r_0, ..., r_{T-1}$ are parameters need to be determined. Specifically, we use $D_{\text{cal},1}$ to formulate a mixed-integer linear programming problem to find optimal parameters $r_0^*, ..., r_{T-1}^*$, and the detailed formulations are in Section III-C.

c) Step 3: Calibrate and construct the final CP regions: Once we determine the parameters r_t, $t = 0, \ldots, T-1$, based on $D_{\text{cal},1}$, we have a well-defined score function $\mathcal{A}$, and as a result we can calculate non-conformity scores $R_i = \mathcal{A}(\tilde{\mathbf{Y}}^{(i)}_{0:T-1})$ for each data in $D_{\text{cal},2}$.

Now, let $p_2 = \lceil (1-\epsilon)(n_2+1) \rceil$. Suppose $R^{[p_2]}$ is the p_2th smallest non-conformity score among scores of $D_{\text{cal},2}$, we can then construct the final CP regions $\Gamma^\epsilon : \hat{\mathbf{Y}}_{0:T-1} \mapsto \otimes_{t=0}^{T-1} \mathcal{C}_t$, where the region at each time step is:

$$\mathcal{C}_t = \{y \in \mathbb{R}^d : ||y - \hat{Y}_t|| \leq R^{[p_2]} + r_t\}. \tag{7}$$

Then these regions are *valid* conformal regions.

Theorem 1 (*Validity*)**.** *Suppose $\mathcal{C}_0, ..., \mathcal{C}_{T-1}$ are derived via the procedures above, then we have $\mathbb{P}\left(Y^{(\text{new})}_t \in \mathcal{C}_t, \forall t = 0, ..., T-1\right) \geq 1 - \epsilon$, where $\mathbb{P}$ is the joint probability measure of $D_{\text{cal},2}$ and $\tilde{\mathbf{Y}}^{(\text{new})}_{0:T-1}$.*

Proof: The construction of score function $\mathcal{A}$ depends only on $D_{\text{cal},1}$, and does not include information from $D_{\text{cal},2}$. Thus, the non-conformity score for each data in $D_{\text{cal},2}$ is exchangeable with that of a new data drawn from $\mathcal{D}$. Let $p_2 = \lceil (1-\epsilon)(n_2+1) \rceil$, by Lemma 1 in [27], the conformal regions defined as $\Gamma^\epsilon(\hat{\mathbf{Y}}^{(\text{new})}_{0:T-1}) := \left\{\mathbf{Y} \in \mathbb{R}^{d \times T} \mid \mathcal{A}(\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{new})}_{0:T-1}) \leq R^{[p_2]}\right\}$ has property that $\mathbb{P}\left(\mathbf{Y}^{(\text{new})}_{0:T-1} \in \Gamma^\epsilon(\hat{\mathbf{Y}}^{(\text{new})}_{0:T-1})\right) \geq 1 - \epsilon$. Now, for any $\mathbf{Y} \in \mathbb{R}^{d \times T}$,

$$\mathcal{A}(\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{new})}_{0:T-1}) = \max_t \left\{||Y_t - \hat{Y}^{(\text{new})}_t|| - r_t\right\} \leq R^{[p_2]}$$
$$\Leftrightarrow ||Y_t - \hat{Y}^{(\text{new})}_t|| \leq R^{[p_2]} + r_t, \quad \forall t = 0, \ldots, T-1.$$

Thus, if we define CP-region at each t as $\mathcal{C}_t = \left\{y \in \mathbb{R}^d : ||y - \hat{Y}_t|| \leq R^{[p_2]} + r_t\right\}$, we guarantee that $\mathbb{P}\left(Y^{(\text{new})}_t \in \mathcal{C}_t, \forall t = 0, ..., T-1\right) \geq 1 - \epsilon$. ■

d) Efficiency of this method: This method produces the smallest average radius (over T regions) regions that a valid CP method can achieve based on $D_{\text{cal},1}$, with respect to a user's predefined norm for calculating normed-residual-series, e.g., if a user uses l_2-norm to calculate absolute residuals, then this method produces the 2-norm balls with the minimum average radius any CP method can achieve. We formalize this in the theorem below.

Theorem 2 (Empirical Average Radius Minimization)**.** *Suppose $\{r^*_0, \ldots, r^*_{T-1}\}$ are the optimal parameters computed in Step 2 (see also the optimization program in III-C), based on $D_{\text{cal},1}$. The minimum average radius of an empirical CP-region is then equal to $\frac{1}{T}\sum_{t=0}^{T-1} r^*_t$.*

Note that the term "empirical" refers to the Empirical-Risk-Minimization (ERM) principle. The regions generated by calibrating data in $D_{\text{cal},1}$ via score function $\mathcal{A}(\tilde{\mathbf{Y}}^{(i)}_{0:T-1}) := \max\left\{\tilde{e}^{(i)}_0 - r^*_0, ..., \tilde{e}^{(i)}_{T-1} - r^*_{T-1}\right\}$ are not *valid* CP regions, but this Theorem shows that OSCP is *efficient* in the sense of ERM principle. The proof can be found in Section VI at the end of the paper.

e) Shapes of CP regions: The shape of CP regions that this method produces depends on the user pre-defined norm for calculating the absolute residuals. This can be viewed as a hyper-parameter for the method. For instance, l_2-norm produces ball-shaped regions, l_1 or l_∞-norm induces hyper-rectangle regions, and positive-definite matrix A-norm results in ellipsoid-shaped regions (see [28]). Ellipsoidal regions is flexible, but may lead to over-fitting when the data is not enough to reflect its shape in high dimensional spaces. When the dataset is small and we don't have prior knowledge of the data distribution, we can assume the prediction residual follows a gaussian error and employ l_2-norm.

C. Optimal Parameter Computation

Suppose we have select the shape of CP region by choosing a specific norm $||\cdot||$, and we have calculated the normed-residual-series for data in $D_{\text{cal},1}$. Let $p_1 = \lceil (1-\epsilon)(n_1+1) \rceil$. Our goal is to determine optimal parameters $r_0, ..., r_{T-1}$ by using the first calibration dataset $D_{cal,1}$.

Recalling the motivating example in III-A, we seek to determine a series of norm-balls with radii $r_t, t = 0, \ldots, T-1$, that have the minimum average radius (equivalently radius sum, i.e., $\sum_{t=0}^{T-1} r_t$) and contain at least p_1 *normed-residual-series* $\tilde{e}^{(i)}_{0:T-1}$'s. We can achieve this by means of the following optimization problem:

$$\min_{\{r_t\},\{b_i\}} \quad \sum_{t=0}^{T-1} r_t \tag{MILP}$$
$$\text{subject to} \quad b_i \cdot (\tilde{e}^{(i)}_t - r_t) \leq 0, \quad t = 0, \ldots, T-1, \quad i = 1, \ldots, n_1 \tag{8}$$
$$\sum_{i=1}^{n_1} b_i = p_1 \tag{9}$$
$$b_i \in \{0, 1\}, \quad i = 1, ..., n_1. \tag{10}$$

This is a mixed-integer linear programming problem that is always feasible, and in fact we can remove a large fraction of redundant constraints to enable faster computation while keeping the optimal solutions of $r_0, ..., r_{T-1}$ unchanged. To reduce the size of this program and improve the associated computational efficiency, two redundant constraint sets, can be identified and removed. To this end, suppose that for each t, $\tilde{e}^{[p_1]}_t$ is the p_1th smallest value among $\tilde{e}^{(1)}_t, ..., \tilde{e}^{(n_1)}_t$. Then the first redundant constraint set is defined by

$$S_1 := \left\{i \in \{1, \ldots, n_1\} \mid \tilde{e}^{(i)}_t \leq \tilde{e}^{[p_1]}_t, \ \forall t = 0, ..., T-1\right\}. \tag{11}$$

This set denotes the indices of all inactive constraints. In the case we are provided (or often it is easy to identify) a feasible solution $\{r^{(\text{feas})}_0, ..., r^{(\text{feas})}_{T-1}, b^{(\text{feas})}_1, ..., b^{(\text{feas})}_{n_1}\}$ to the (MILP), then we can neglect a second redundant set

$$S_2 := \left\{i \in \{1, ..., n_1\} \mid \tilde{e}^{(i)}_t > r^{(\text{feas})}_t \text{ for } \forall t = 0, ..., T-1\right\}, \tag{12}$$

which includes all solutions that would lead to a cost (sum of radii) greater than that of the available feasible solution.

Although the method to find such feasible solutions is not unique, one fast and easy-to-implement heuristic procedure is as follows: for each $i = 1, ..., n_1$, we calculate the sum of normed residuals $\text{TotalRes}(i) := \sum_{t=0}^{T-1} \tilde{e}_t^{(i)}$ and sort $\text{TotalRes}(i)$'s in non-decreasing order. Then we pick the first p_1 indices, $i_1, ..., i_{p_1}$ and let $b_i^{(\text{feas})} = 1$ for these indices, $b_i^{(\text{feas})} = 0$ otherwise. Let $r_t^{(\text{feas})} = \max_{i=i_1,...i_{p_1}} \tilde{e}_t^{(i)}$. Then we have a feasible solution to (MILP).

Once S_1 & S_2 are identified, we can set up a modified optimization program as follows.

$$\min_{r_t, b_i} \quad \sum_{t=0}^{T-1} r_t \tag{MILP-fast}$$

$$\text{subject to} \quad b_i \cdot (\tilde{e}_t^{(i)} - r_t) \le 0, \quad t = 0, \dots, T-1, \ i \in S \tag{13}$$

$$\max_{i \in S_1} \{\tilde{e}_t^{(i)}\} \le r_t, \quad t = 0, \dots, T-1 \tag{14}$$

$$\sum_{i \in S} b_i = p_1 - |S_1| \tag{15}$$

$$b_i \in \{0, 1\}, \quad i \in S \tag{16}$$

$$\text{where } S := \{1, 2, ..., n_1\} \setminus (S_1 \cup S_2).$$

Theorem 3 (Equivalence of (MILP) & (MILP-fast))**.** *When* $|S_1| < p_1$, (MILP-fast) *is always feasible, and its set of optimal solutions of* $\{r_0^*, ..., r_{T-1}^*\}$ *coincides with that of* (MILP)*. Otherwise, if* $|S_1| \ge p_1$*, the optimal solution to* (MILP) *is* $r_t = \tilde{e}_t^{[p_1]}, \ t = 0, ..., T-1$.

The corresponding proof is in Section VII at the end of the paper. Theorem 3 implies that when $|S_1| < p_1$, we can solve (MILP-fast) to find optimal parameters $r_0^*, ..., r_{T-1}^*$. The rough idea is that to make the choice of r_t's optimal, we must always consider containing residual-time-series from S_1 but not considering containing those from S_2. Thus, when $S_1 < p_1$, solving (MILP) is equivalent to solve (MILP-fast). On the other hand, when $|S_1| \ge p_1$ (although not common in practice), there are already more than p_1 residual-time-series to be contained inside the norm-balls, so we can simply choose $r_t = \tilde{e}_t^{[p_1]}$, $t = 0, \dots, T-1$, as the optimal parameters.

When the error level ϵ is small, (MILP-fast) usually can remove a large number of mixed-integer constraints in (8) and integer variables b_i's, which makes the computation much faster. The detailed results of increased running speed can be seen in Section IV-C.

The pseudo-code of this faster algorithm for computing optimal parameters is in [29].

IV. Numerical Experiments

To demonstrate the performance of our method, we test on both simulated and real time-series with different time horizons T, dimensions d, and calibration dataset sizes N, taken from [22]. We compare our method with 5 baseline uncertainty-quantification (UQ) methods, and the results show that among all *valid* alternatives, the *efficiency* of our proposed approach outperforms the state-of-the-art method on all case studies.

a) Baseline Approaches: We selected MC-dropout [26] and CF-RNN [21] as the baseline approaches for Bayesian UQ method and CP for time series, respectively; as well as three recent approaches in CP for time series, namely, CopulaCPTS [22] and LCP [24] as parameter optimization methodologies, and CRD [23] as a convex CP baseline. Since both CRD and our method allow users to specify shapes, we test hyper-rectangle & ellipsoid shapes of CRD (denoted as CRD-Rect & CRD-Ell, respectively) and ball & ellipsoid shapes of our method (OSCP-l_2 & OSCP-Ell).

A. Synthetic Datasets

a) Particle trajectory: According to [22], the first two datasets are generated from the Interacting Particle System [30], and extra Gaussian noises with standard deviation $\sigma = 0.01$ and 0.05 are added to the dynamics of particle simulations in two datasets, respectively. For each case, data is split into 2000/500/500 for training, calibration, testing, respectively. A prediction model is trained to predict the future dynamics $\mathbf{Y} \in \mathbb{R}^{d \times T}$ given the past observations $\mathbf{X} \in \mathbb{R}^{d \times \tau}$ of the particle simulation with $\tau = 35, \ T = 25, \ d = 2$. Then all UQ methods are evaluated according to the procedures stated in [29]. For UQ methods that require a further split of calibration dataset, the split ratio is set to be $0.5/0.5$ for $D_{\text{cal},1}$ & $D_{\text{cal},2}$ except LCP. We note that the optimization program in LCP becomes computationally intractable for this split, so we adopt $0.1/0.9$ split-ratio for LCP.

b) Drone trajectory: The drone trajectory dataset is generated from [31] with added Gaussian noise of $\sigma = 0.02$. The data-split is $600/200/200$ for training/calibration/testing. The prediction model forecasts a drone's future trajectory given its past observations ($\tau = 60, \ T = 10, \ d = 3$). After training the prediction model, all UQ methods are evaluated in a similar manner. For methods requiring a split of the calibration dataset, a split-ratio of $0.5/0.5$ is adopted (including LCP).

B. Real Dataset: Covid-19 Daily Cases

We also conduct a case study on the real dataset, UK Covid-19 Daily Cases. We used the preprocessed dataset from [22].[2] Each time-series data in the preprocessed Covid-19 dataset corresponds to 150-day daily cases from mid-September 2020 to mid-February 2021 at a region in UK. Then, $500/160/80$ time-series data are used for training/calibration/testing. The prediction model takes 100 days of data as input, and outputs the subsequent 50 days, i.e., $\tau = 100, T = 50, d = 1$. For UQ methods that require a further split of the calibration data, the split ratio was set to $0.5/0.5$.

C. Numerical Results

For target confidence levels from 0.5 to 0.95 (10 values), we tested each UQ method with 50 runs (random splits of calibration and test set but with the same proportion) on each dataset. OSCP aims at producing minimal radius norm-balls; it is thus not direct how to compare that radius with

[2]The original Covid-19 dataset can be download at https://coronavirus.data.gov.uk/

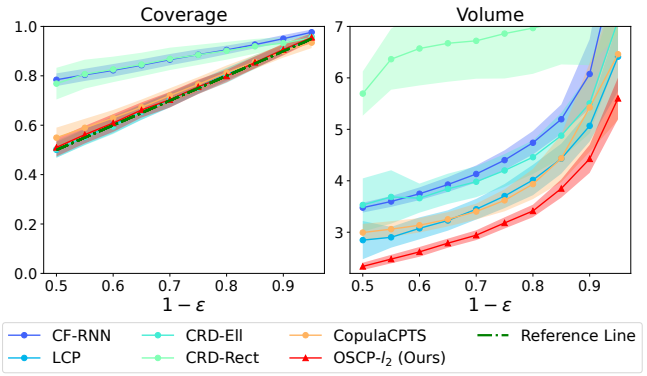


Fig. 2: Performance visualization on the Particle Dataset ($\sigma = 0.05$). The dashed reference line in the Coverage graph denotes the target confidences, and only methods with coverage curves at or above this line achieve the target coverages. In the Volume graph, curves closer to the bottom indicate better performances (less conservative).

some UQ methods whose outcome is not a norm-ball one. Thus, we consider comparing the total volume (area/length depending on the dimension) of confidence sets. A more detailed description of the comparison setup & performance evaluation of all methods is provided in [29].

Our method outperforms all the baselines on 4 case studies for all 10 confidence levels (0.5 to 0.95). Specifically, compared to previous SOTA method LCP [24] (the one with the smallest total volume among baselines that have empirical coverage no smaller than target ones), our method with l_2-norm, OSCP-l_2, reduces the total confidence region size (on average) of **16.03%, 14.32%, 14.01%, 16.93%** on dataset **Particle** ($\sigma = 0.01$), **Particle** ($\sigma = 0.05$), **Drone**, **Covid-19**, respectively. When using an ellipsoidal confidence region, our method achieves further reductions in the region size (see [29] for details).

Part of the experiment results (with standard error) are shown in Figure 2, and the complete results & visualizations are in [29]. From the results in Figure 2, it can be seen that our method returns a confidence region with the smallest volume among all alternatives, while achieving the target confidence levels.

a) Runtime comparison: We also compare the runtime used in solving the optimization problem between our method and the previous SOTA (LCP, [24]). For the Particle ($\sigma = 0.05$) dataset, LCP sometimes reach the time limit and terminate the simulation at 10000s, so the actual computing time is higher than the reported result. In Table I, we can see that when target confidence is set as 0.9, our method is **8812.0**, **78622.0**, **14.4**, **22.1** times faster than LCP on the four datasets, respectively.

TABLE I: Comparison of optimization runtime (in sec) for target confidence $1 - \epsilon = 0.9$

Case Study	Previous SOTA [24]	OSCP-l_2
Particle ($\sigma = 0.01$)	1215.002 $\pm$ 1450.119	0.137 $\pm$ 0.057
Particle ($\sigma = 0.05$)	>9392.66 $\pm$ 1358.109	0.119 $\pm$ 0.034
Drone	0.151 $\pm$ 0.033	0.010 $\pm$ 0.007
Covid-19	0.549 $\pm$ 0.166	0.025 $\pm$ 0.011

V. Conclusion

In this work, we propose a new parameterized score function for conformal prediction in multi-dimensional time series and an optimization program to determine an optimal parameter set. We prove validity and efficiency of our method, showing that optimizing these parameters is equivalent to determining the minimum-average-radius CP regions with a pre-specified norm-ball description. Numerical results on four different datasets (synthetic and actual data) demonstrate that our method outperforms alternative approaches, while having much lower computational requirements.

References

[1] M. Fortunato, C. Blundell, and O. Vinyals, "Bayesian recurrent neural networks," 2017.

[2] R. M. Neal *et al.*, "Mcmc using hamiltonian dynamics," *Handbook of markov chain monte carlo*, vol. 2, no. 11, p. 2, 2011.

[3] D. P. Kingma, M. Welling, *et al.*, "Auto-encoding variational bayes," 2013.

[4] A. Legay, A. Lukina, L. M. Traonouez, J. Yang, S. A. Smolka, and R. Grosu, "Statistical model checking," in *Computing and software science: state of the art and perspectives*, pp. 478–504, Springer, 2019.

[5] Y. Wang, M. Zarei, B. Bonakdarpour, and M. Pajic, "Statistical verification of hyperproperties for cyber-physical systems," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 18, no. 5s, pp. 1–23, 2019.

[6] L. Lindemann, Y. Zhao, X. Yu, G. J. Pappas, and J. V. Deshmukh, "Formal verification and control with conformal prediction," *arXiv preprint arXiv:2409.00536*, 2024.

[7] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Berlin, Heidelberg: Springer-Verlag, 2005.

[8] M. C. Campi and S. Garatti, *Introduction to the scenario approach*. SIAM, 2018.

[9] K. Margellos, P. Goulart, and J. Lygeros, "On the road between robust optimization and the scenario approach for chance constrained optimization problems," *IEEE Transactions on Automatic Control*, vol. 59, no. 8, pp. 2258–2263, 2014.

[10] N. O'Sullivan, L. Romao, and K. Margellos, "Bridging conformal prediction and scenario optimization," *arXiv preprint arXiv:2503.23561*, 2025.

[11] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas, "Safe planning in dynamic environments using conformal prediction," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, 2023.

[12] C. Stamouli, L. Lindemann, and G. Pappas, "Recursively feasible shrinking-horizon mpc in dynamic environments with conformal prediction guarantees," in *6th Annual Learning for Dynamics & Control Conference*, pp. 1330–1342, PMLR, 2024.

[13] S. Tonkens, S. Sun, R. Yu, and S. Herbert, "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*, 2023.

[14] 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.

[15] L. Bortolussi, F. Cairoli, N. Paoletti, S. A. Smolka, and S. D. Stoller, "Neural predictive monitoring," in *International Conference on Runtime Verification*, pp. 129–147, Springer, 2019.

[16] F. Cairoli, L. Bortolussi, and N. Paoletti, "Neural predictive monitoring under partial observability," in *International Conference on Runtime Verification*, pp. 121–141, Springer, 2021.

[17] F. Cairoli, N. Paoletti, and L. Bortolussi, "Conformal quantitative predictive monitoring of stl requirements for stochastic processes," in *26th ACM International Conference on Hybrid Systems: Computation and Control (HSCC '23)*, 2023.

[18] N. Hashemi, X. Qin, L. Lindemann, and J. V. Deshmukh, "Data-driven reachability analysis of stochastic dynamical systems with conformal inference," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 3102–3109, IEEE, 2023.
[19] N. Hashemi, L. Lindemann, and J. V. Deshmukh, "Statistical reachability analysis of stochastic cyber-physical systems under distribution shift," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 11, pp. 4250–4261, 2024.
[20] A. Tebjou, G. Frehse, *et al.*, "Data-driven reachability using christoffel functions and conformal prediction," in *Conformal and Probabilistic Prediction with Applications*, pp. 194–213, PMLR, 2023.
[21] 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.
[22] S. H. Sun and R. Yu, "Copula conformal prediction for multi-step time series prediction," in *The Twelfth International Conference on Learning Representations*, 2024.
[23] R. Tumu, M. Cleaveland, R. Mangharam, G. Pappas, and L. Lindemann, "Multi-modal conformal prediction regions by optimizing convex shape templates," in *Proceedings of the 6th Annual Learning for Dynamics & Control Conference*, vol. 242, pp. 1343–1356, PMLR, 2024.
[24] M. Cleaveland, I. Lee, G. J. Pappas, and L. Lindemann, "Conformal prediction regions for time series using linear complementarity programming," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 20984–20992, 2024.
[25] 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)*, ICCPS '23, pp. 142–153, 2023.
[26] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, pp. 1050–1059, PMLR, 2016.
[27] R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[28] C. Xu, H. Jiang, and Y. Xie, "Conformal prediction for multi-dimensional time series by ellipsoidal sets," in *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.
[29] B. Pang and K. Margellos, "Efficient quantification of time-series prediction error: Optimal selection conformal prediction," *arXiv preprint arXiv:2511.02103*, 2025.
[30] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 2688–2697, PMLR, 2018.
[31] A. Sakai, D. Ingram, J. Dinius, K. Chawla, A. Raffin, and A. Paques, "Pythonrobotics: a python code collection of robotics algorithms," 2018.

VI. Proof of Theorem 2

Lemma 1. *Fix an error level $\epsilon \in (0,1)$, and a predefined score function $\mathcal{A}$. Fix also any norm, and let r_t^*, $t = 0,\dots,T-1$, be the optimal parameters of OSCP's score function. Any valid CP region (with same shape as OSCP's) constructed based on $D_{\text{cal},1}$ has average radius $\frac{1}{T}\sum_{t=0}^{T-1} r_t \geq \frac{1}{T}\sum_{t=0}^{T-1} r_t^*$.*

Proof of Lemma 1: We first show that if the set (not necessarily a CP-region) $\Gamma^\epsilon(\hat{\mathbf{Y}}_{0:T-1}) := \otimes_{t=0}^{T-1}\left\{y \in \mathbb{R}^d : ||y - \hat{Y}_t|| \leq r_t\right\}$ contains at least p_1 elements out of $\mathbf{Y}^{(i)}$, $i = 1,\dots,n_1$, then it has average radius no smaller than $\frac{1}{T}\sum_{t=0}^{T-1} r_t^*$.

Let $p_1 = \lceil (1-\epsilon)(n_1+1)\rceil$. Consider the set

$$\Gamma^\epsilon(\hat{\mathbf{Y}}_{0:T-1})^* := \otimes_{t=0}^{T-1}\left\{y \in \mathbb{R}^d : ||y - \hat{Y}_t|| \leq r_t^*\right\}, \quad (17)$$

where $\{r_t^*\}_{t=0}^{T-1}$ is the optimal solution to the (MILP). Since $r_0^*, r_2^*, ..., r_{T-1}^*$ is feasible to (MILP), there are at least p indices from $\{1,...,n_1\}$ such that $\tilde{e}_t^{(i)} \leq r_t^*, \forall t = 0,2,...,T-1$. Now, since $\{r_t^*\}$ is optimal to the objective of (MILP), then the average radius of $\Gamma^\epsilon(\hat{\mathbf{Y}}_{0:T-1})^*$ is the minimum among all norm-ball regions that contains at least p_1 elements out of $\mathbf{Y}^{(i)}$, $i = 1,\dots,n_1$.

Now, consider CP regions constructed on $D_{\text{cal},1}$ via the score function $\mathcal{A}$ and a selected shape induced by $||\cdot||$:

$$\Gamma^\epsilon(\hat{\mathbf{Y}}_{0:T-1}) := \{\mathbf{Y} \in \mathbb{R}^{T\times d} : \mathcal{A}(\mathbf{Y} - \hat{\mathbf{Y}}_{0:T-1}) \leq R^{[p_1]}\}.$$

Since $R^{[p_1]}$ is the p_1th smallest nonconformity score, then there are at least p_1 number of i's satisfying

$$\mathcal{A}(\tilde{\mathbf{Y}}_{0:T-1}^{(i)}) \leq R^{[p_1]} \Rightarrow \mathbf{Y}_{0:T-1}^{(i)} \in \Gamma^\epsilon(\hat{\mathbf{Y}}_{0:T-1}^{(i)}).$$

Then the average radius of $\Gamma^\epsilon(\hat{\mathbf{Y}}_{0:T-1})$ cannot be smaller than that of (17), which is $\frac{1}{T}\sum_{t=0}^{T-1} r_t^*$. ∎

Proof of Theorem 2: Now with Lemma 1, we start proofing the Theorem 2. Given an optimal solution $\{r_0^*, ..., r_{T-1}^*, b_1^*, ..., b_{n_1}^*\}$ of (MILP), the non-conformity score R_i of each data is then calculated by:

$$R_i := \mathcal{A}(\tilde{\mathbf{Y}}_{0:T-1}^{(i)}) = \max\{\tilde{e}_0^{(i)} - r_0^*, ..., \tilde{e}_{T-1}^{(i)} - r_{T-1}^*\}.$$

To prove the result stated in the theorem, we first show that $R^{[p_1]} = 0$. Let $i_1, i_2, ..., i_{p_1}$ be the indices such that $b_i^* = 1$. Due to feasibility of r_t^* for any $t = 0,...,T-1$, we have $\tilde{e}_t^{(i)} \leq r_t^*, \forall i = i_1, ..., i_{p_1}$. Thus, for $i = i_1, ..., i_{p_1}$, we have $R_i \leq 0$. We then have $R^{[p_1]} \leq \max_{i=i_1,...,i_{p_1}} R_i \leq 0$. Now suppose for the sake of contradiction that $R^{[p_1]} < 0$. Then $\exists i_1', i_2', ..., i_{p_1}'$ such that $R_i < 0, \forall i = i_1', i_2', ..., i_{p_1}'$. Consequently, we have $\tilde{e}_t^{(i)} < r_t^*, \forall t = 0,...,T-1, \forall i = i_1', ..., i_{p_1}'$.

Consider a new solution candidate

$$r_t' = \max_{i=i_1',...,i_{p_1}'} \tilde{e}_t^{(i)}, \quad b_i' = \begin{cases} 1, & \text{if } i = i_1', ..., i_{p_1}'; \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to check that this new solution is feasible to (MILP) and $r_t' < r_t^*, \forall t = 0,...,T-1$. This contradicts to the fact that $\sum_{t=0}^{T-1} r_t^*$ is the minimum cost solution.

Thus, we can conclude that $R^{[p_1]} = 0$. Then for each t, the resulting CP-region is

$$\{y \in \mathbb{R}^d : ||y - \hat{Y}_t|| \leq R^{[p_1]} + r_t^*\} \iff \{y \in \mathbb{R}^d : ||y - \hat{Y}_t|| \leq r_t^*\}.$$

This completes the proof that the average radius of empirical CP-region of OSCP calculated from $D_{\text{cal},1}$ is equal to $\frac{1}{T}\sum_{t=0}^{T-1} r_t^*$, which is the minimum value that a valid CP-region with same shape can attain by Lemma 1. ∎

VII. Proof of Theorem 3

Proof: **Case 1**: $|S_1| < p_1$

Feasibility of (MILP-fast) is easy to check, as we can always randomly pick p_1 indices $i_1, ..., i_{p_1} \in S$, and then the solution $\{r_t = \max_i \tilde{e}_t^{(i)}\}_{t=0}^{T-1}$, $b_i = \begin{cases} 1, & \forall i = i_1, ..., i_{p_1} \\ 0, & \text{otherwise} \end{cases}$ is trivially feasible to the problem. We will now prove each direction seperately.

($\Rightarrow$) w.t.s. Any optimal parameters $\mathbf{r}^* = (r_0^*, ..., r_{T-1}^*)$ of (MILP), $\exists \mathbf{b}' = \{b_i',\ i \in S\}$ s.t. $(\mathbf{r}^*, \mathbf{b}')$ is an optimal solution to (MILP-fast).

We will first show that the feasibility region of (MILP-fast) is a subset of that of (MILP).

First of all, we augment the space of decision variables $(\mathbf{r}, \mathbf{b}) = \{r_0, ..., r_{T-1}\} \cup \{b_i\}_{i\in S}$ to $(\mathbf{r}, \bar{\mathbf{b}}) = \{r_0, ..., r_{T-1}, b_1, ..., b_{n_1}\}$, i.e. the feasibility region of (MILP-fast) now becomes (10), (13), (14) & (15).

Consider following constraints:

$$b_i = 1, \quad i \in S_1; \tag{18}$$

$$b_i = 0, \quad i \in S_2. \tag{19}$$

We add these two constraints on $(\mathbf{r}, \bar{\mathbf{b}})$, which has no effect on the feasible region of original decision variables $(\mathbf{r}, \mathbf{b})$. That being saying, the feasibility region of (MILP) is equivalent with that of augmented (MILP-fast), i.e.,

$$(8), (9), (10), (18), (19) \Leftrightarrow (10), (13), (14), (15), (18), (19). \tag{20}$$

This result is easy to check: for any solution r_t, b_i satisfying (18) & (19),

$$\begin{aligned}
&b_i \cdot (\tilde{e}_t^{(i)} - r_t) \le 0,\ t = 0, \dots, T-1,\ i = 1, \dots, n_1 \\
&\Leftrightarrow b_i \cdot (\tilde{e}_t^{(i)} - r_t) \le 0,\ t = 0, \dots, T-1,\ i \in \{1, \dots, n_1\} \setminus S_2 \\
&\Leftrightarrow \begin{cases} b_i \cdot (\tilde{e}_t^{(i)} - r_t) \le 0,\ t = 0, \dots, T-1,\ i \in S, \\ \max\limits_{i\in S_1}\{\tilde{e}_t^{(i)}\} \le r_t,\ t = 0, \dots, T-1. \end{cases}
\end{aligned}$$

Therefore, under (18) & (19), Constraint (8) $\Leftrightarrow$ Constraints (13), (14). Also, under (18) & (19),

$$\sum_{i=1}^{n_1} b_i = p_1 \Leftrightarrow \sum_{i\in S} b_i = p_1 - |S_1|,$$

hence Constraint (9) $\Leftrightarrow$ Constraint (15), and thus, we can conclude the result in (20). Consequently, we have shown that the feasible region of augmented (MILP-fast) is a subset of that of (MILP), which means the optimal objective value of (MILP-fast) is no smaller than that of (MILP).

Now we will show that the optimal objective are indeed equal by showing that any set of optimal parameters $\{r_0^*, ..., r_{T-1}^*\}$ of (MILP) is feasible in (MILP-fast).

Suppose $(\mathbf{r}^*, \mathbf{b}^*) = (r_0^*, ..., r_{T-1}^*, b_1^*, ..., b_{n_1}^*)$ is an optimal solution of (MILP).

First, we show that $b_i^* = 0$ for $i \in S_2$. For $\forall i \in S_2$, $\tilde{e}_t^{(i)} > r_t^{(\text{feas})}$ for $\forall t = 0, ..., T-1$. Since r_t^* is optimal, $\sum_{t=0}^{T-1} r_t^* \le \sum_{t=0}^{T-1} r_t^{(\text{feas})}$. Then there must be at least one $\hat{t}$ s.t. $r_{\hat{t}}^* \le r_{\hat{t}}^{(\text{feas})}$. This means that $\tilde{e}_{\hat{t}}^{(i)} > r_{\hat{t}}^{(\text{feas})} \ge r_{\hat{t}}^*$ Thus, to satisfy the constraint (8) in (MILP), it must be that $b_i^* = 0$ for $\forall i \in S_2$.

Next, we will show that $\exists \{b_i'\}_{i\in S}$ s.t. $(\mathbf{r}^*, \mathbf{b}')$ is feasible to (MILP-fast). Let $i_1, ..., i_{p_1}$ be the indices s.t. $b_i^* = 1$. Then by the above result we know that $i_1, ..., i_{p_1} \in \{1, ..., n_1\} \setminus S_2 = S \cup S_1$. Select $p_1 - |S_1|$ indices from $\{i_1, ..., i_{p_1}\} \setminus S_1$ (this is always possible as $|S_1| < p_1$) and set $b_i' = 1$ for these indices, and set $b_i' = 0$ for the rest of the indices in S.

Then we have $p_1 - |S_1|$ indices i's s.t. $i \in S$ and $\tilde{e}_t^{(i)} \le r_t^*$ at each t. So constraint (13) & (15) are satisfied. Now, since for $i_1, ..., i_{p_1}$, $b_i^* = 1$, it means that we have p_1 indices i's s.t. $\tilde{e}_t^{(i)} \le r_t^*$ at each t. Recall that $\tilde{e}_t^{[p_1]}$ is the p_1th smallest element in the sorted non-descending sequence $\{\tilde{e}_t^{(i)}\}_{i=1}^{n_1}$. This means that $\tilde{e}_t^{[p_1]} \le r_t^*, \quad \forall t = 0, ..., T-1$. Consequently, for $\forall i \in S_1$,

$$\tilde{e}_t^{(i)} \le \tilde{e}_t^{[p]} \le r_t^*,\ \forall t = 0, ..., T-1 \ \Leftrightarrow \max_{i\in S_1}\{\tilde{e}_t^{(i)}\} \le r_t^*,$$

so constraint (14) is also satisfied and we can therefore conclude that $(\mathbf{r}^*, \mathbf{b}')$ is feasible to (MILP-fast). As optimal value of (MILP) is always larger or equal to that of (MILP-fast), we can conclude that $(\mathbf{r}^*, \mathbf{b}')$ is optimal solution to (MILP-fast) and the optimal value of (MILP) and (MILP-fast) are indeed equal.

($\Leftarrow$) w.t.s. For any optimal parameters $\mathbf{r}^* = (r_0^*, ..., r_{T-1}^*)$ of (MILP-fast), $\exists \mathbf{b}' = (b_1', b_2', ..., b_{n_1}')$ s.t. $(\mathbf{r}^*, \mathbf{b}')$ is an optimal solution to (MILP).

Suppose $(\mathbf{r}^*, \mathbf{b}^*)$ is an optimal solution of (MILP-fast). Consider solution $(\mathbf{r}^*, \mathbf{b}') = (r_0^*, ..., r_{T-1}^*, b_1', ..., b_{n_1}')$, where

$$b_i' = \begin{cases} b_i^*, & i \in S; \\ 1, & i \in S_1; \\ 0, & i \in S_2. \end{cases}$$

Then we have

$$\sum_{i=1}^{n_1} b_i' = \sum_{i\in S} b_i^* + |S_1| \overset{\text{Constraint (15)}}{=} p_1 - |S_1| + |S_1| = p_1,$$

so constraint (9) is satisfied. For constraint (8), let's first consider the case $i \in S \cup S_2$, the constraints $b_i' \cdot (\tilde{e}_t^{(i)} - r_t^*) \le 0,\ t = 0, \dots, T-1$ are trivially satisfied. For case of $i \in S_1$, constraint (14) in (MILP-fast) says that $b_i' \cdot (\tilde{e}_t^{(i)} - r_t^*) = \tilde{e}_t^{(i)} - r_t^* \le 0,\ t = 0, ..., T-1$. Combining these results, constraint (8) is satisfied. As a result, the solution $(\mathbf{r}^*, \mathbf{b}')$ is feasible to (MILP). Since in the proof of ($\Rightarrow$) we have already shown that the optimal value of (MILP-fast) is equal to (MILP), we can conclude that $(\mathbf{r}^*, \mathbf{b}')$ is optimal to (MILP).

Case 2: $|S_1| \ge p_1$

Consider the solution candidate $r_t^* = \tilde{e}_t^{[p_1]}, t = 0, ...T-1$. We will first show it is feasible to (MILP). Pick arbitrary p_1 indices $i_1, i_2, ..., i_{p_1}$ from S_1, then let $b_i = 1, \forall i = i_1, ..., i_{p_1}$ and let $b_i = 0$ otherwise. This guarantees constraints (9) & (10) satisfied. Since $\forall i \in S_1$, $\tilde{e}_t^{(i)} \le \tilde{e}_t^{[p_1]}$ for $\forall t = 0, ..., T-1$, constraint (8) is also satisfied. Thus, the solution $r_t^* = \tilde{e}_t^{[p_1]}, t = 0, ..., T-1$ is feasible to (MILP).

Now, we will show that it is also optimal to (MILP). Suppose, for contradiction, $\exists$ a feasible solution $\{r_0', ..., r_{T-1}'\}$ of (MILP) such that the objective value $\sum_{t=0}^{T-1} r_t' < \sum_{t=0}^{T-1} r_t^*$. Then $\exists r_{\hat{t}}' < r_{\hat{t}}^* = \tilde{e}_{\hat{t}}^{[p_1]}$ for some $\hat{t}$. This means there are less than p_1 i's s.t. $\tilde{e}_{\hat{t}}^{(i)} \le r_{\hat{t}}'$. However, constraints (8) & (9) together imply that there $\exists$ at least p_1 i's such that $\tilde{e}_{\hat{t}}^{(i)} \le r_{\hat{t}}', \forall t = 0, ..., T-1$. Contradiction! Thus, the solution $r_t^* = \tilde{e}_t^{[p_1]}, t = 0, ..., T-1$ is optimal to (MILP). ■