UNCERTAINTY QUANTIFICATION IN NON-STATIONARY TIME SERIES VIA ADAPTIVE CONFORMAL REGRESSION AND REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction (CP) provides rigorous distribution-free uncertainty quantification, yet its effectiveness in time-series forecasting is hindered by non-stationarity and data dependencies. Recent adaptive CP methods mitigate distributional shifts by dynamically calibrating intervals based on recent residuals and adaptive weighting strategies. However, they remain limited by outlier sensitivity, undetected systematic biases, and the decoupling of calibration from model learning. In this work, we introduce `CORE` that establishes a mutual feedback loop between reinforcement learning (RL) and conformal prediction. By reformulating conformal regression as a sequential decision-making process, `CORE` utilizes RL exploration to better cover uncertain or outlier regions, adapts calibration through exploration feedback, and designs uncertainty-guided rewards, enabling dynamically improved interval quality through policy interaction. We provide theoretical analysis and conduct extensive empirical experiments to validate its effectiveness across 9 time-series standard datasets. The results demonstrate that `CORE` achieves superior validity and adaptivity, consistently outperforming 6 state-of-the-art baselines, with solid coverage rates and averaged 2.34% improvement in interval length.

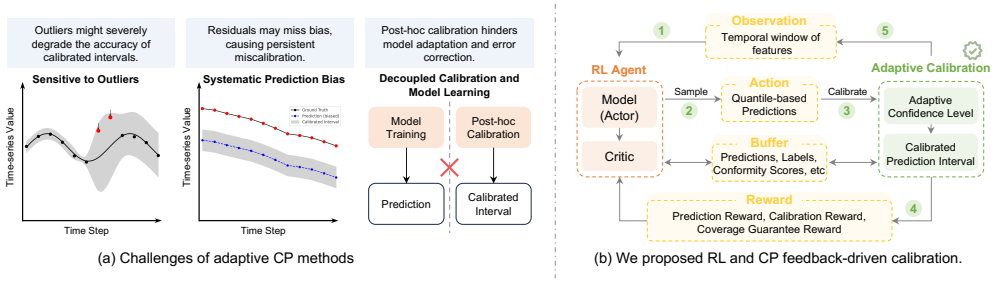
1 INTRODUCTION

Conformal prediction (CP) is a model-agnostic framework for distribution-free uncertainty quantification that provides rigorous finite-sample coverage guarantees under data exchangeability (De Finetti, 1937; Shafer & Vovk, 2008). In regression task, let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ denote the covariate-response pair. Given exchangeable samples $\{(X_i, Y_i)\}_{i=1}^n$ and a target miscoverage level $\alpha \in (0, 1)$, conformal regression constructs a prediction interval $\hat{C}(X_{n+1}) \in \mathbb{R}$ that attains marginal coverage at least $1 - \alpha$: $\Pr(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$ (Lei et al., 2018). However, in sequential time-series forecasting, temporal dependence breaks exchangeability assumption since the joint distribution is not invariant under reordering (Gibbs & Candes, 2021). This limits classical conformal regression (Lei et al., 2018; Romano et al., 2019; Xu & Xie, 2021) in high-stakes domains like finance, healthcare and industrial control, where reliable intervals are needed for risk-aware decisions.

Recent advances have introduced adaptive or online CP methods (Gibbs & Candes, 2021; Zaffran et al., 2022), which rely on approximate independence assumptions and update intervals through sliding or weighted calibration, often with dynamic adjustments of conformity scores, thereby enabling responsive uncertainty quantification under distribution shifts (Xu & Xie, 2023a). However, despite the flexibility of adaptive CP methods, they face some fundamental limitations in dynamic environments, as depicted in Figure 1(a). Temporally, sensitivity to outliers and shifting biases (Cheung et al., 2024) under asymptotic coverage mandates triggers a lagging compensation cycle. Instead, the methods merely offset past errors via subsequent over- or under-coverage, resulting in persistently distorted intervals (Sun & Yu, 2025). Architecturally, post-hoc calibration operates independently from model training, blocking feedback that could proactively mitigate biases (Huang et al., 2025). These limitations necessitate a unified approach that integrates calibration with training to enable feedback-driven, uncertainty-aware learning.

However, such integration is fundamentally different from simply adding another supervised objective, because prediction and calibration form a sequential feedback system whose long-term utility is not

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Challenges of adaptive CP methods

(b) We proposed RL and CP feedback-driven calibration.

Figure 1: Challenges of adaptive CP methods and our proposed CORE architecture.

fully captured by prediction loss alone. This motivates the use of RL, which allows such calibration feedback to directly influence model learning and optimize predictions according to their downstream uncertainty quality. In this way, calibration is no longer merely a post-hoc correction mechanism, but becomes an explicit learning signal for improving the predictor.

To this end, we propose CORE, standing for **CO**nformal regression with **RE**inforcement learning (RL), which reformulates conformal regression as a sequential decision-making process. As illustrated in Figure 1(b), CORE replaces static post-hoc correction with a dynamic feedback loop among the agent, environment, and calibration mechanism. Central to our approach is a quantile-guided actor-critic framework that generates adaptive prediction intervals, coupled with an adaptation-aware calibration module that dynamically updates conformity scores via a sliding buffer. By incorporating uncertainty-guided rewards, CORE synchronizes prediction accuracy, coverage validity, and interval efficiency into a single RL objective, enabling the policy to proactively refine both forecasting and calibration through environmental interaction.

We evaluate CORE on 9 standard time-series datasets and further analyze, from both theoretical and empirical perspectives, how its main components affect the validity and efficiency trade-off. Our theoretical analysis decomposes the coverage error into three sources, finite-window calibration error governed by the calibration window w_c, prediction-induced error associated with policy learning, and tracking error introduced by the adaptive update of α_t. Specifically, we establish the coverage guarantee $\sup_{t\le T}\left|\Pr\{Y_t \in \hat{C}_{1-\alpha_t}^{(w_c,t)}\} - (1-\alpha)\right| = \mathcal{O}\left(\left(\frac{\log w_c}{w_c}\right)^{1/3} + T^{-p/3} + \gamma^{1/2}\right)$, where the three terms respectively characterize finite-window calibration error, prediction-induced approximation error, and adaptive tracking error. This decomposition directly connects the theoretical analysis to the main components of CORE and motivates our empirical studies of their effects. In particular, we examine the sensitivity to the calibration window w_c and adjustment rate γ, while complementary ablation studies isolate the empirical roles of RL-based policy learning, adaptive confidence adjustment, and quantile-based prediction. The main contributions of this work are summarized as follows:

- We propose a mutual feedback loop between RL and CP, enabling the agent to learn uncertainty-aware policies while simultaneously improving calibration through interaction.
- Our empirical results show that CORE achieves strong performance across all tasks, with an average improvement of 2.34% in interval length.
- Our theoretical analysis decomposes coverage error into finite-window calibration, prediction-induced, and adaptive tracking terms, establishing validity and interval-efficiency guarantees under temporal dependence and distributional drift.

2 Conformal Regression

Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ with $Y = f(X) + \epsilon$, where ϵ is independent of X. Here f denotes a generic predictive model, and conformal prediction is *model-agnostic*, independent of the specific form of f. Given samples $(\mathbf{x}_i, y_i)_{i=1}^n$, a conformity score is introduced to measure how well a candidate pair (x, y) aligns with the predictive model. A common choice in regression is the absolute residual, i.e., $S(x, y) = |\hat{f}(x) - y|$. Regardless of the specific form of the conformity score, the key issue is to determine how small the score must be for $\tilde{f}(x)$ to be accepted as a reasonable prediction (Gibbs & Candes, 2021). For a specified index set $\mathcal{I}_{\text{cal}}$, a held-out calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, y_i)\}_{i\in\mathcal{I}_{\text{cal}}}$ disjoint

from the data is used to fit the regression model (Romano et al., 2019), establishes empirical quantile thresholds on conformity scores for acceptance. Under exchangeability, conformal regression then constructs a valid $(1-\alpha)$ coverage interval via the calibration set, and computes the predicted interval according to the quantiles:

$$\hat{C}_{1-\alpha}(x) := \hat{f}(x) \pm (1-\alpha) \text{ quantile of } \{S(\mathbf{x}_i, y_i)\}_{i \in \mathcal{I}_{\text{cal}}}. \tag{1}$$

However, in non-stationary settings such as time-series forecasting, the exchangeability assumption fails. A common remedy is a sliding calibration window using recent residuals for threshold estimation, resembling local exchangeability (Campbell et al., 2019). While theoretical coverage guarantees are more difficult to establish, these adaptive methods (Xu & Xie, 2021; Gibbs & Candes, 2021; Xu & Xie, 2023b) have been shown to maintain approximate validity under mild assumptions. Formally, at each time t, given a stream of observations $\{(\mathbf{x}_i, y_i)\}_{i=t-w+1}^{t}$ with $\mathbf{x}_t \in \mathbb{R}^{d \times w}$, we construct a sliding calibration window of size w_c ($\leq t$), and estimates conformity threshold $\hat{q}_{1-\alpha}^{(w_c,t)}$ from the calibration window and form the interval $\hat{C}_{1-\alpha}^{(w_c,t)}$.

3 METHOD

Uncertainty quantification is essential for developing robust and trustworthy machine learning systems. CP provides rigorous finite-sample coverage guarantees, but its static calibration often falters under non-stationarity (Gibbs & Candes, 2021). RL excels in dynamic exploration but is prone to unstable behaviors (Fox et al., 2015). In this work, we propose `CORE`, a unified framework that integrates CP and RL via feedback-driven calibration and enable uncertainty-aware learning in dynamic environments.

We begin by formulating the problem as a sequential decision-making process within a standard RL framework, structured into three functional aspects: conformal exploration, where the agent samples quantile-distributed actions to guide exploration under uncertainty; adaptive-aware calibration, which dynamically adjusts nonconformity scores based on exploration results; and uncertainty-guided rewards, which jointly regulate both exploration and calibration to enhance predictive reliability.

3.1 RL FORMULATION

We cast adaptive conformal regression as a RL problem, where the agent learns to construct valid and efficient prediction intervals through interaction with the data-generating environment. Formally, we model this process as a Markov Decision Process (MDP), represented by a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \nu)$. State $\mathcal{S}$: at time t, the state $\mathbf{s}_t$ encodes a temporal feature window $\mathbf{x}_t \in \mathbb{R}^{d \times w}$, where w is the window size. Action $\mathcal{A}$: the agent outputs quantile-based predictions $\mathbf{a}_t$, which specify a prediction interval $\hat{C}_{1-\alpha_t}^{(w_c,t)} \subseteq \mathbb{R}$ for the next response y_t. Reward $\mathcal{R}$: feedback is provided through conformity scores computed from recent samples, which balances calibration validity (whether $y_t \in \hat{C}_{1-\alpha}^{(w_c,t)}$) with efficiency (interval width). Transition $\mathcal{T}$: the next state s_{t+1} arises from the temporal evolution of the sequence, reflecting potential distribution shifts in the data. The agent thus learns a policy $\pi_\theta : \mathcal{S} \to \mathcal{A}$ that maximizes the expected discounted return $R_t = \mathbb{E}\big[\sum_{i=0}^{\infty} \beta^i r_{t+i}\big]$, where calibration feedback is integrated into the reward signal. We adopt Proximal Policy Optimization (PPO) (Schulman et al., 2017) for stable and efficient learning in the continuous action space.

3.2 CONFORMAL EXPLORATION

We consider the usual setting of an MDP, in which an agent interacts with the environment $\mathcal{E}$. At each time step t, the observation $\mathbf{s}_t$ is defined as a temporal window of the past features, i.e., $\mathbf{s}_t = \mathbf{x}_{t-w+1:t} \in \mathbb{R}^{d \times w}$, where w is the window size and d is the feature dimension.

3.2.1 QUANTILE-GUIDED ACTOR-CRITIC

In regression, predicting y_t can be viewed as selecting an action in a continuous space. Instead of a point estimate, we model the predictive process as a distributional policy π_θ, which samples quantile-based predictions $\mathbf{a}_t = \{a_t^{(q)}\}_{q \in \mathcal{Q}}$ where $\mathcal{Q} := \{q^{(1)}, \ldots, q^{(K)}\}$ denotes K fixed quantiles. For example, with $\mathcal{Q} = \{0.05, 0.5, 0.95\}$, the actor samples three actions of 5^{th} percentile, median,

Algorithm 1: CORE: Adaptive Conformal Regression via Reinforcement Learning

Input: Data $\{(\mathbf{x}_t, y_t)\}_{t=1}^{T}$; PPO policy π_θ, critic V_ψ, learning rate η; Significance level α, quantile levels $\mathcal{Q} = \{q^{(1)}, \ldots, q^{(K)}\}$; Conformity score function $S(\cdot, \cdot)$, observation window size w, calibration window size w_c, adjustment rate γ

Output: Calibrated policy π_{θ^*} with uncertainty-aware policy; Prediction intervals $\hat{C}_{1-\alpha_t}^{(w_c,t)}(\mathbf{x}_t)$

Initialize conformity score buffer $\mathcal{S} \leftarrow \emptyset$, prediction interval buffer $\mathcal{C} \leftarrow \emptyset$, ground-truth buffer $\mathcal{Y} \leftarrow \emptyset$
for *each training iteration* **do**
  Initialize environment $\mathcal{E}$ and receive initial window $\mathbf{x}_{1:w}$;
  for $t = w$ **to** T **do**
    Observe state $\mathbf{s}_t = \mathbf{x}_{t-w+1:t}$;
    Draw actions $\{a_t^{(q)}\}_{q \in \mathcal{Q}} \sim \pi_\theta(\mathbf{s}_t)$ and let $\bar{\mathbf{a}}_t = a_t^{(0.5)}$;
    Update $\mathcal{S} \leftarrow \mathcal{S} \cup \{S(a_t^{0.5}, y_t)\}$;
    if $|\mathcal{S}| > w_c$ Remove oldest score from $\mathcal{S}$ to maintain $|\mathcal{S}| = w_c$;
    Update buffers $\mathcal{C} \leftarrow \mathcal{C} \cup \{a_t^{0.5}\}$, $\mathcal{Y} \leftarrow \mathcal{Y} \cup \{y_t\}$, then compute error rate $\hat{\epsilon}_t$ using $(\mathcal{C}, \mathcal{Y})$;
    $\alpha_{t+1} \leftarrow \alpha_t + \gamma(\alpha - \hat{\epsilon}_t)$; // ▷ adaptive confidence level
    if $|\mathcal{S}| < w_c$ **then** Set $\hat{q}_{1-\alpha_t}^{(w_c,t)} \leftarrow 0$;
    else $\hat{q}_{1-\alpha_t}^{(w_c,t)} \leftarrow \text{Quantile}_{1-\alpha_t}(\mathcal{S})$;
    $\hat{C}_{1-\alpha_t}^{(w_c,t)}(\mathbf{x}_t) \leftarrow [a_t^{(\text{low})} - \hat{q}_{1-\alpha_t}^{(w_c,t)},\ a_t^{(\text{up})} + \hat{q}_{1-\alpha_t}^{(w_c,t)}]$;
    Compute reward $r_t = \lambda_{\text{acc}} \cdot r_{\text{acc}} + \lambda_{\text{len}} \cdot r_{\text{len}} + \lambda_{\text{cov}} \cdot r_{\text{cov}}$; // ▷ refer to Eq.5-7
    Observe y_t and next state $\mathbf{s}_{t+1}$ Store transition $(\mathbf{s}_t, \bar{\mathbf{a}}_t, r_t, \mathbf{s}_{t+1})$ in PPO buffer $\mathcal{T}$;
  Update Critic V_ψ with trajectory $\mathcal{T}$ using generalized advantage estimation (GAE) loss;
  Update Actor policy π_θ with trajectory $\mathcal{T}$ using quantile loss $\mathcal{L}_{\text{quantile}}$ and behavior cloning loss $\mathcal{L}_{\text{BC}}$;

and 95^{th} percentile, where quantile regression captures local heteroscedasticity and yields adaptive prediction intervals that reflect varying uncertainty (Takeuchi et al., 2006; Romano et al., 2019).

The actor-critic architecture strengthens this formulation: the actor generates quantile predictions to represent uncertainty, while the critic provides value-based feedback to stabilize training and improve efficiency under noisy targets (Haarnoja et al., 2018). Thus, the task becomes a sequential decision process where the agent refines predictions through exploration and policy updates.

3.2.2 Training Objectives

To enable uncertainty-aware prediction, outputs must not only be diverse but also reflect the underlying distribution structure (Meinshausen & Ridgeway, 2006). We therefore adopt a dual-objective scheme enforcing both quantile correctness and pointwise accuracy.

Quantile Regression Loss. Each output $a_t^{(q)}$ is optimized to approximate the true conditional quantile at level q by minimizing the tilted (pinball) loss:

$$\mathcal{L}_{\text{quantile}} = \sum_{q \in \mathcal{Q}} \rho_q\big(y_t - a_t^{(q)}\big), \text{ where } \quad \rho_q(u) = u\big(q - \mathbb{I}[u < 0]\big). \tag{2}$$

This asymmetric loss penalizes over-estimation and under-estimation differently, guiding the model to align each quantile prediction with its statistical meaning. As a result, the agent learns to represent the spread and skewness of the conditional outcome distribution.

Trajectory Behavior Cloning Loss. Quantile regression enforces distributional consistency, but it does not explicitly ensure that the central prediction aligns with the ground-truth value. We therefore introduce a behavior cloning loss as a trajectory-level prior, anchoring the median prediction $a_t^{(0.5)}$ to the observed target y_t. Such structural imitation ensures that the generated trajectory remains physically grounded, and the RL component adaptively refines the interval widths to satisfy conformal coverage requirements.

3.3 Adaptive-aware Calibration

We adopt a trajectory-level calibration strategy that adaptively adjusts the conformal confidence to handle non-stationary dynamics. Unlike a fixed confidence level case $1 - \alpha$, α_t is updated online based on recent calibration error, maintaining both validity and tightness of prediction intervals.

At each timestep t, we compute a conformity score $S(\mathbf{x}_t, y_t) = |a_t^{(0.5)} - y_t|$, and append it to a sliding window buffer $\mathcal{S}_t$ of size w_c. To capture temporal dynamics, we use the trajectory of median actions $\mathcal{T}_{t-w_c+1:t} := \{a_\tau\}_{\tau=t-w_c+1}^{t}$ together with their conformity scores, estimate the $(1-\alpha_t)$ quantile $\hat{q}_{1-\alpha_t}^{(w_c,t)}$, and form prediction intervals. At each timestep t, the actor outputs quantile-based actions, with the smallest and largest denoted by $a_t^{(\text{low})} := a_t^{(q^{(1)})}$ and $a_t^{(\text{up})} := a_t^{(q^{(K)})}$, respectively. The prediction interval is therefore given by:

$$\hat{C}_{1-\alpha_t}^{(w_c,t)}(\mathbf{x}_t) = [a_t^{(\text{low})} - \hat{q}_{1-\alpha_t}^{(w_c,t)},\ a_t^{(\text{up})} + \hat{q}_{1-\alpha_t}^{(w_c,t)}]. \tag{3}$$

Let the calibration index set be $\mathcal{I}_t$, so that $\mathcal{S}_t = \{S(\mathbf{x}_i, y_i) : i \in \mathcal{I}_t\}$. Using the sliding window calibration set, we then estimate the empirical miscoverage and update the confidence level via exponential smoothing under the law:

$$\hat{\epsilon}_t = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \mathbf{1}\left\{ y_i \notin \hat{C}_{1-\alpha_t}^{(w_c,i)} \right\}, \quad \alpha_{t+1} = \alpha_t + \gamma(\alpha - \hat{\epsilon}_t), \tag{4}$$

where $\gamma \in \mathbb{R}^+$ controls the update rate and α is the target miscoverage. Calibrating over recent trajectories ensures that validity reflects cumulative performance rather than isolated steps, providing trajectory-aware uncertainty quantification that adapts to temporal patterns and avoids both under-confident and overly conservative intervals.

3.4 Uncertainty Rewards

To guide the learning of an uncertainty-aware policy, we design a composite reward that encourages predictive accuracy, calibrated coverage, and the construction of compact and informative prediction intervals. The overall reward at time t is defined as $r_t = \lambda_{\text{acc}} \cdot r_{\text{acc}} + \lambda_{\text{len}} \cdot r_{\text{len}} + \lambda_{\text{cov}} \cdot r_{\text{cov}}$, where each term reflects a different aspect of performance under uncertainty.

Accuracy-driven Signal. To encourage accurate point predictions, we introduce a reward signal based on the absolute deviation between the predicted median and the true target:

$$r_{\text{acc}} = \sigma\big(1 - \text{residual}(a_t^{(0.5)}, y_t)\big). \tag{5}$$

The ℓ_1 residual provides robustness to outliers and naturally soft-clips rewards by bounding gradient growth compared to quadratic loss (Haarnoja et al., 2018) To further stabilize training, we apply a $\sigma(\cdot)$ activation to clip extreme values and maintain a consistent reward scale. The reward design is task-adaptive: residuals are used for regression, while cross-entropy or margin-based terms apply for classification. Specifically, for classification task like anomaly detection, with label $y_t \in \{0, 1\}$ and predicted probability p_t, we define $r_{\text{acc}} = 1 - \ell_{\text{BCE}}(y_t, p_t) = 1 - \big(y_t \log p_t + (1 - y_t)\log(1 - y_t)\big)$.

Compactness-aware Interval Penalty. To balance coverage rate and interval length, we note that overly wide bands diminish the practical value of uncertainty estimates (Romano et al., 2019). We therefore design the following penalty:

$$r_{\text{len}} = 1 - \frac{\left|\hat{C}_{1-\alpha_t}^{(w_c,t)}(\mathbf{x}_t)\right|}{\tilde{y} + \delta} \quad \text{with} \quad \tilde{y} = \text{Median}(\{y_i\}_{i=t-w_c+1}^{t}), \tag{6}$$

where the numerator is the interval length and the denominator normalizes it by the recent target magnitude (using the local median as a robust reference); and a small δ avoids division-by-zero. This discourages intervals excessively wide relative to local scale, while α_t controls under-coverage.

Coverage-consistency Reward. To ensure that predictions remain consistent with the calibrated interval, we design a reward encouraging the true response and predicted median to be contained and lie inside the interval, respectively:

$$r_{\text{cov}} = \begin{cases} 1 & \text{if } y_t \in \hat{C} \wedge \bar{\mathbf{a}}_t \in \hat{C} \\ -1 & \text{if } y_t \notin \hat{C} \wedge \bar{\mathbf{a}}_t \in \hat{C} \\ -1 & \text{if } y_t \in \hat{C} \wedge \bar{\mathbf{a}}_t \notin \hat{C} \end{cases} \tag{7}$$

where $\hat{C}$ is the surrogate for $\hat{C}_{1-\alpha_t}^{(w_c,t)}(\mathbf{x}_t)$, applying graded penalties depending on whether the true value and/or the predicted median fall outside the interval. Notably, in practice, we fix $\lambda_{\text{acc}} = \lambda_{\text{len}} = \lambda_{\text{cov}} = 1.0$ across all datasets, without any dataset-specific tuning. This choice follows the heuristic practice of shaping rewards with task knowledge rather than extensive tuning (Cheng et al., 2021; Gupta et al., 2022).

4 THEORETICAL ANALYSIS

4.1 DATA ASSUMPTIONS

The theoretical results rely on several mild assumptions on the data-generating process, which ensure that the learning dynamics remain well-behaved over time.

Weak Dependence. The sequence $(X_t, Y_t)_{t\geq 1}$ is β-mixing, where $\beta(k)$ denotes the dependence coefficient between the past and the future separated by lag k. We assume the summable-β condition $\sum_{k=1}^{\infty}\beta(k) < \infty$. This ensures temporal correlations decay sufficiently for concentration, accommodating the dependence inherent in RL trajectories (Xu & Xie, 2023a).

Local Regularity. For each t, let $F_t(z) = \Pr(\varepsilon_t \leq z)$ be the cumulative distribution function (CDF) of the regression residual $\varepsilon_t := Y_t - f_t(X_t)$ and assume F_t is L-Lipschitz continuous: $|F_t(u) - F_t(v)| \leq L|u - v|$ for all $u, v \in \mathbb{R}$. This regularity ensures stable quantile estimation through a bounded probability density.

Bounded Drift. The temporal evolution of the regression function $f_t(x) = \mathbb{E}[Y_t \mid X_t = x]$ is controlled by $\xi_t := \sup_x |f_{t+1}(x) - f_t(x)|$. We assume the cumulative drift $\xi_{s:t} := \sum_{k=s}^{t}\xi_k$ remains bounded, allowing for controlled non-stationarity.

4.2 MAIN RESULTS

We establish concentration bounds for the empirical residual distribution, link policy optimization to prediction error decay, and ensure stability via adaptive calibration. Together, these yield a uniform validity and efficiency theorem. A complete proof is provided in Appendix D.

Lemma 1 (Rio-type DKW inequality)**.** *Let $\tilde{F}$ denote the empirical CDF of the true residuals within $[t - w_c + 1, t]$ over a calibration window of size w_c, and let F denote the true CDF. Under the summable-β weak dependence condition,*

$$\sup_x \left|\tilde{F}(x) - F(x)\right| \leq C_1(\log w_c/w_c)^{1/3}.$$

Lemma 2 (Quantile-error transfer)**.** *Let $\delta_T^2 = \frac{1}{T}\sum_{t=1}^{T}(\hat{\varepsilon}_t - \varepsilon_t)^2$. If F_t is L-Lipschitz, then*

$$\sup_x \left|\hat{F}(x) - \tilde{F}(x)\right| \leq C_2\delta_T^{2/3} + 2\sup_x \left|\tilde{F}(x) - F(x)\right|,$$

where $C_2 := L + 1$, with the additive constant serving as a technical slack.

To complete the coverage bound, it remains to control the error induced by policy optimization.

Proposition 1 (Prediction-error decay under PPO)**.** *Suppose the policy is updated by PPO with step sizes satisfying $\sum_k \eta_k = \infty$ and $\sum_k \eta_k^2 < \infty$. Then, for some rate exponent $p \in [1, 2]$, the median prediction error decays as*

$$\delta_T^2 := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[(a_t^{(0.5)} - f_t(X_t))^2\right] = \mathcal{O}(T^{-p/2}).$$

In addition, the adaptive update of the miscoverage level stabilizes empirical coverage.

Proposition 2 (Adaptive α stability)**.** *For $\alpha_{t+1} = \alpha_t + \gamma(\alpha - \hat{\epsilon}_t)$, let $\hat{\epsilon}_t$ be the empirical miscoverage $\hat{\epsilon}_t = \frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\mathbf{1}\{y_i \notin \hat{C}_{1-\alpha_t}^{(w_c,i)}\}$. Then the gap between the running level α_t and the true miscoverage ϵ_t satisfies $|\alpha_t - \epsilon_t| = \mathcal{O}(\gamma^{1/2})$.*

Theorem 1 (Master Theorem: validity and efficiency)**.** *Combining Lemmas 1 and 2, Proposition 1, and 2, the* `CORE` *intervals satisfy:*

(Validity) *For any horizon T, the coverage error satisfies:*

$$\sup_{t\leq T}\left|\Pr\{Y_t \in \hat{C}_{1-\alpha_t}^{(w_c,t)}\} - (1-\alpha)\right| = \mathcal{O}\left(\left(\frac{\log w_c}{w_c}\right)^{1/3} + T^{-p/3} + \gamma^{1/2}\right).$$

(Efficiency) *Let $C_t^\star$ denote the oracle shortest valid interval. Then:*

$$\sum_{t=1}^{T}\left(|\hat{C}_t| - |C_t^\star|\right) = \tilde{\mathcal{O}}(T^{1-p/3}).$$

Table 1: Performance comparison for uncertainty quantification and prediction accuracy.

Dataset	Weather		Traffic[1]		Electricity		Illness		ETTh1		ETTh2		ETTm1		ETTm2		TSLA	
Metric	COV	LEN	COV	LEN($\times 10^{-2}$)	COV	LEN	COV	LEN($\times 10^{6}$)	COV	LEN	COV	LEN	COV	LEN	COV	LEN	COV	LEN
Target Coverage: 90% ($\alpha = 0.1$)																		
EnbPI	0.901	7.913	0.899	7.451	0.897	1691.524	0.905	1.904	0.899	12.211	0.897	38.924	0.900	12.241	0.898	38.337	0.911	12.888
	(0.003)	(2.789)	(0.002)	(0.231)	(0.000)	(21.181)	(0.005)	(0.005)	(0.001)	(0.347)	(0.003)	(1.519)	(0.001)	(0.464)	(0.001)	(0.623)	(0.002)	(0.444)
ACI	0.901	7.124[3]	0.903	7.243	0.901	1776.118	0.270	1.621	0.901	12.338	0.901	37.844	0.901	12.299	0.901	38.361	0.898	11.705
	(0.000)	(0.015)	(0.000)	(0.052)	(0.000)	(3.849)	(0.064)	(0.156)	(0.000)	(0.005)	(0.000)	(0.057)	(0.000)	(0.009)	(0.000)	(0.012)	(0.002)	(0.112)
SPCI	0.900	7.468	0.899	7.000	0.897	1686.365	0.899	1.879	0.900	12.307	0.896	38.665	0.900	12.181	0.898	38.740	0.916	13.066
	(0.001)	(1.264)	(0.001)	(0.001)	(0.002)	(15.594)	(0.011)	(0.076)	(0.002)	(0.284)	(0.100)	(0.623)	(0.001)	(0.413)	(0.001)	(0.926)	(0.004)	(0.440)
PID	0.896	7.713	0.909	7.300	0.897	1572.986	0.904	1.979	0.902	11.963	0.899	36.481	0.894	11.764	0.903	36.787	0.901	11.478
	(0.073)	(1.714)	(0.037)	(1.110)	(0.044)	(91.360)	(0.003)	(0.371)	(0.000)	(0.960)	(0.062)	(1.296)	(0.005)	(0.533)	(0.011)	(1.825)	(0.006)	(3.231)
ECI	0.902	7.759	0.884	7.111	0.901	1578.492	0.894	1.730	0.901	11.782	0.903	35.161	0.890	11.915	0.902	37.851	0.897	11.534
	(0.063)	(0.081)	(0.003)	(0.386)	(0.002)	(40.259)	(0.081)	(0.224)	(0.003)	(0.657)	(0.003)	(3.054)	(0.008)	(1.773)	(0.004)	(1.242)	(0.003)	(1.024)
NCC	0.903	7.754	0.905	7.347	0.912	1907.435	0.901	1.807	0.902	11.780	0.902	36.357	0.903	12.374	0.904	37.452	0.902	12.074
	(0.001)	(0.864)	(0.005)	(0.248)	(0.003)	(0.682)	(0.003)	(0.175)	(0.012)	(1.372)	(0.003)	(1.573)	(0.005)	(0.247)	(0.003)	(1.037)	(0.005)	(0.248)
CORE	0.899	7.451	0.899	**6.863**[3]	0.902	**1509.209**	0.906	**1.520**	0.903	**11.548**	0.902	**32.060**	0.905	11.866	0.900	**36.626**	0.898	**11.416**
	(0.008)	(1.035)	(0.024)	(0.938)	(0.036)	(87.433)	(0.001)	(0.138)	(0.012)	(0.312)	(0.012)	(0.900)	(0.014)	(0.431)	(0.012)	(1.767)	(0.023)	(1.397)
-w/o-updated α	0.909	7.696	0.931	8.442	0.905	1636.115	0.882	1.533	0.895	12.148	0.893	37.115	0.909	14.072	0.901	39.741	0.906	12.015
	(0.024)	(0.552)	(0.054)	(0.762)	(0.033)	(187.355)	(0.067)	(0.104)	(0.020)	(1.443)	(0.018)	(3.855)	(0.020)	(1.731)	(0.007)	(3.622)	(0.047)	(2.007)
-w/o-RL	0.868	8.294	0.875	9.655	0.881	1820.848	0.891	1.686	0.891	15.565	0.872	35.064	0.845	17.788	0.840	54.393	0.828	11.431
	(0.032)	(0.352)	(0.025)	(1.140)	(0.060)	(231.058)	(0.038)	(0.114)	(0.023)	(2.147)	(0.044)	(4.778)	(0.056)	(3.210)	(0.078)	(5.982)	(0.120)	(1.321)
-w/o-quantiles	0.898	9.301	0.863	7.455	0.901	1700.582	0.875	1.553	0.890	13.934	0.899	44.455	0.886	16.982	0.894	44.395	0.895	12.748
	0.021	(1.536)	(0.037)	(0.259)	(0.042)	(184.252)	(0.042)	(0.072)	(0.042)	(2.581)	(0.030)	(3.398)	(0.025)	(1.679)	(0.046)	(4.922)	0.028	0.938
Target Coverage: 95% ($\alpha = 0.05$)																		
EnbPI	0.949	8.714	0.944	8.180	0.950	2092.743	0.939	2.413	0.946	14.706	0.950	44.577	0.948	14.733	0.949	43.675	0.961	16.195
	(0.002)	(2.250)	(0.001)	(0.623)	(0.001)	(37.576)	(0.002)	(0.056)	(0.001)	(0.095)	(0.002)	(1.344)	(0.001)	(0.137)	(0.001)	(0.780)	(0.006)	(0.615)
ACI	0.951	8.682	0.952	8.292	0.951	2127.738	0.948	3.087	0.951	14.852	0.950	43.582	0.951	15.120	0.951	44.450	0.953	15.449
	(0.000)	(0.037)	(0.000)	(0.025)	(0.000)	(2.812)	(0.011)	(2.134)	(0.000)	(0.026)	(0.000)	(0.029)	(0.000)	(0.026)	(0.000)	(0.055)	(0.002)	(0.171)
SPCI	0.948	6.840	0.944	8.027	0.951	2055.694	0.942	2.426	0.947	14.663	0.951	44.157	0.948	14.626	0.948	44.019	0.960	15.761
	(0.001)	(0.258)	(0.001)	(0.119)	(0.000)	(19.649)	(0.008)	(0.101)	(0.001)	(0.126)	(0.000)	(0.808)	(0.001)	(0.061)	(0.001)	(0.676)	(0.005)	(0.272)
PID	0.934	7.370	0.948	7.980	0.962	1992.618	0.942	2.338	0.945	14.075	0.949	42.738	0.944	13.832	0.964	44.231	0.953	15.769
	(0.031)	(0.658)	(0.023)	(1.485)	(0.127)	(24.317)	(0.002)	(0.155)	(0.007)	(1.377)	(0.001)	(0.973)	(0.001)	(0.792)	(0.018)	(2.071)	(0.013)	(1.541)
ECI	0.964	8.276	0.940	8.167	0.942	1846.243	0.947	2.397	0.948	14.228	0.954	44.018	0.963	14.395	0.952	41.249	0.932	15.957
	(0.006)	(0.735)	(0.017)	(0.128)	(0.011)	(13.274)	(0.005)	(0.204)	(0.000)	(2.020)	(0.006)	(2.311)	(0.008)	(1.089)	(0.004)	(0.336)	(0.017)	(0.592)
NCC	0.951	7.309	0.945	8.051	0.948	1814.532	0.950	2.402	0.943	14.782	0.941	43.570	0.948	13.785	0.945	43.572	0.942	15.521
	(0.004)	(0.785)	(0.005)	(0.137)	(0.003)	(24.671)	(0.010)	(0.114)	(0.003)	(0.978)	(0.006)	(1.355)	(0.007)	(1.326)	(0.005)	(1.535)	(0.011)	(1.349)
CORE	0.951	7.872	0.949	**7.771**	0.951	**1778.566**	0.952	**1.692**	0.947	**13.972**	0.947	**42.474**	0.949	14.039	0.950	41.254	0.947	**14.560**
	(0.012)	(1.010)	(0.016)	(0.296)	(0.004)	(34.036)	(0.011)	(0.147)	(0.016)	(0.855)	(0.007)	(1.876)	(0.012)	(0.940)	(0.007)	(2.851)	(0.006)	(1.136)
-w/o-updated α	0.918	10.602	0.948	13.225	0.964	1830.541	0.895	1.709	0.924	15.558	0.967	47.534	0.937	16.440	0.924	43.656	0.950	15.333
	(0.050)	(1.873)	(0.013)	(0.552)	(0.013)	(161.974)	(0.076)	(0.131)	(0.022)	(1.385)	(0.011)	(2.162)	(0.024)	(1.593)	(0.025)	(3.126)	(0.013)	(0.291)
-w/o-RL	0.935	11.415	0.934	10.451	0.939	1904.924	0.930	1.838	0.934	18.485	0.942	51.444	0.937	17.964	0.940	48.380	0.943	15.055
	(0.021)	(1.394)	(0.018)	(1.000)	(0.019)	(281.395)	(0.055)	(0.127)	(0.036)	(1.809)	(0.017)	(5.002)	(0.040)	(3.693)	(0.011)	(2.998)	(0.009)	(2.039)
-w/o-quantiles	0.935	8.691	0.945	8.135	0.946	1911.147	0.923	1.697	0.945	16.599	0.946	49.156	0.954	18.194	0.954	47.131	0.952	16.466
	(0.028)	(2.150)	(0.008)	(0.678)	(0.010)	(89.761)	(0.110)	(0.020)	(0.023)	(2.120)	(0.024)	(3.519)	(0.010)	(4.281)	(0.007)	(6.444)	(0.006)	(0.435)

[1] For clarity, LEN values are rescaled by 10^{-2} for Traffic and 10^{6} for Illness to account for the differing magnitudes of ground-truth scales.
[2] Underlined marks the best baseline.
[3] Bold indicates CORE outperforms all baselines.

4.3 Discussion

(1) **Adaptive calibration stabilizes miscoverage.** Proposition 2 shows that the adaptive update of α_t contributes the term $\gamma^{1/2}$, characterizing the stability cost of online error-budget adjustment: increasing γ improves responsiveness to coverage violations at the cost of larger tracking fluctuations. This guides the tuning of γ in Section 5.3.

(2) **Prediction error drives coverage accuracy.** Proposition 1 shows that under PPO training with diminishing step sizes, the median prediction error δ_T^2 vanishes at a polynomial rate. Lemma 2 links this to quantile error, directly tightening coverage bounds.

(3) **The coverage theorem guarantees validity and efficiency.** The Master Theorem 1 establishes near-nominal coverage by combining the residual concentration term $(\log w_c/w_c)^{1/3}$, prediction-induced term $T^{-p/3}$, and adaptive-level error $\gamma^{1/2}$, while bounding cumulative excess width by $\tilde{O}(T^{1-p/3})$ under decreasing prediction error and mild drift.

5 Experiment

In this section, we evaluate CORE from three perspectives. Section 5.2 assesses interval validity and efficiency on standard time-series datasets. Section 5.3 studies key hyperparameters and provides tuning guidance by aligning theoretical analysis with empirical observations. Section 5.4 ablates the adaptive α update, RL-based policy, and quantile learning mechanism to clarify their individual roles.

5.1 Experimental Setup

Datasets, Metrics, Baselines. To evaluate the capabilities of CORE in *uncertainty quantification*, we conduct experiments on nine standard time-series forecasting datasets (Weather (Zhou et al., 2021), Traffic (Lai et al., 2018), Electricity (Lai et al., 2018), Illness (Lai et al., 2018), ETT (Zhou et al., 2021)), and TSLA stock dataset spanning from from January 1, 2021 to December 31, 2024, which characterizes significant real-world distribution shifts. For uncertainty quantification, we assess the model's ability to produce valid and informative prediction intervals, using coverage rate (**CVG**) and

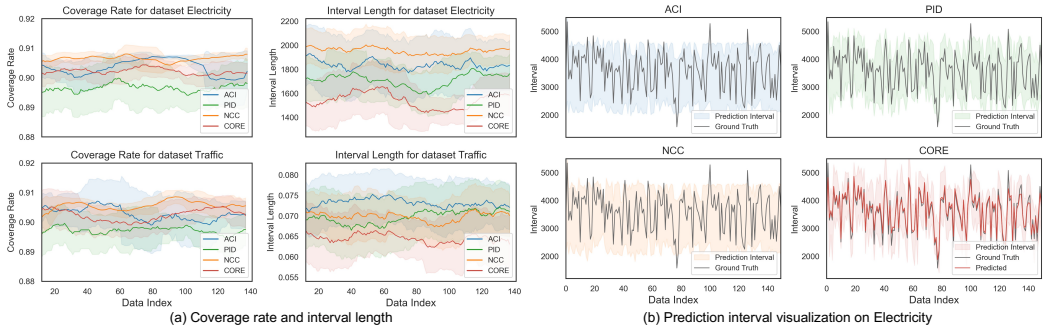


Figure 2: Prediction interval and coverage rate comparison between CORE and baselines on Electricity and Traffic datasets ($\alpha = 0.1$).

interval length (**LEN**), and compare against NCC (Li & Rodríguez, 2025), ECI (Wu et al., 2025), PID (Angelopoulos et al., 2023), SPCI (Xu & Xie, 2023b), EnbPI (Xu & Xie, 2021), and ACI (Gibbs & Candes, 2021).

Implementation Details. We implement CORE in PyTorch and train all models on a single NVIDIA A5000 GPU with 24GB memory, using 3-layer MLPs for both the base predictor and the policy network to ensure a fair, model-agnostic comparison. The policy is optimized via the PPO backbone with a learning rate of 5×10^{-5}. We set significance levels $\alpha \in \{0.05, 0.1\}$, and a temporal window $W = 50$. Further sensitivity analysis on the adjustment rate γ, action set Q, calibration window length w_c are in the section 5.3, and full details are deferred to Appendix B.1.4. Code is available at `https://anonymous.4open.science/r/CORE-3473`.

5.2 Interval Validity

To evaluate uncertainty quantification, we compare CORE with six adaptive conformal baselines on nine time-series datasets under two significance levels, $\alpha = 0.1$ and $\alpha = 0.05$. As shown in Table 1, CORE attains near-nominal coverage while producing shorter intervals on most datasets. Several baselines either over-cover with unnecessarily wide intervals or under-cover and fail to meet the target level, whereas CORE maintains a stable balance between coverage validity and interval efficiency.

We provide a closer view on Electricity and Traffic. In Figure 2(a), CORE tracks the target coverage over time while reducing interval length by an average of 3.01% compared with the strongest baseline. Figure 2(b) further compares the constructed intervals with the ground-truth values: CORE keeps intervals tight in stable regions and expands mainly near temporal shifts or extreme observations, whereas representative baselines tend to be either overly conservative or slow to react. Additional validation traces in Appendix C.1 show that the adaptive α_t update self-corrects coverage deviations and stabilizes around the target level, explaining the reliable calibration observed in the main results.

5.3 Hyperparameter Sensitivity

To evaluate the robustness of CORE under non-stationary conditions and provide practical guidelines for hyperparameter selection, we study three calibration-related hyperparameters: the adjustment rate γ for confidence adaptation, the calibration window size w_c, and quantile set granularity $\mathcal{Q}$. Our analysis uses these theory-motivated roles to guide tuning and examines their empirical effects in Figure 3 across datasets with varying temporal dynamics.

Impact of Adjustment Rate (γ). The adaptation rate γ controls how quickly α_t responds to coverage errors. Proposition 2 suggests increasing γ allows faster correction of miscoverage but also increases the tracking fluctuation of α_t through the $\mathcal{O}(\gamma^{1/2})$ term. This is consistent with the empirical results: small values such as $\gamma = 0.001$ react too slowly and lead to under-coverage, whereas large values such as $\gamma > 0.05$ introduce unstable oscillations around the target. A moderate range, $\gamma \in [0.01, 0.05]$, therefore provides a practical balance between adaptability and stability.

Impact of Calibration Window Size (w_c). The choice of window size w_c reflects the balance between historical perspective and responsiveness. For the large-scale Electricity dataset, a short window (e.g., $w_c = 25$) causes over-coverage and excessive interval lengths due to insufficient calibration horizon. Expanding w_c improves efficiency by capturing broader patterns, though coverage eventually declines as the window exceeds a certain threshold. In contrast, the smaller Illness dataset

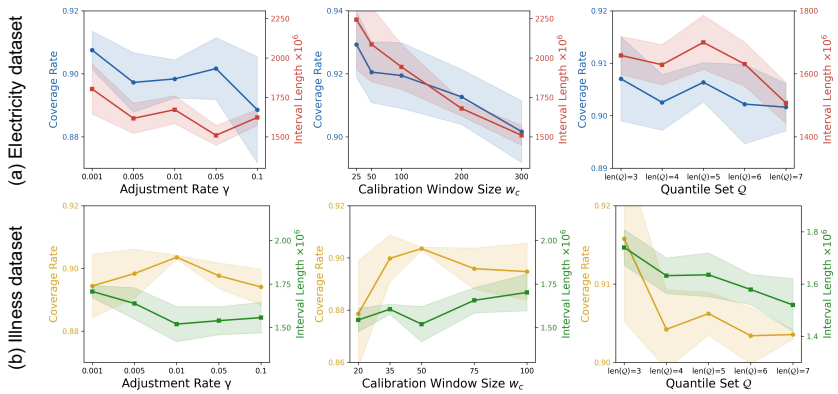


Figure 3: Hyperparameter sensitivity analysis on Electricity and Illness datasets ($\alpha = 0.1$).

achieves effective calibration with relatively shorter windows, as limited recent observations are sufficient to capture underlying distribution. These findings suggest that w_c is governed by the regimes rather than data volume. Beyond an optimal threshold, the benefits of a broader horizon are negated by the noise of stale data, prioritizing temporal relevance over sample size.

Impact of Quantile Set Granularity ($|\mathcal{Q}|$). We recommend a uniform and symmetric quantile grid for $\mathcal{Q}$. For instance, when $|\mathcal{Q}| = 3$, the set is typically defined as $\{\alpha/2, 0.5, 1-\alpha/2\}$. As illustrated in the results, increasing the granularity of $\mathcal{Q}$ can shorten intervals while keeping coverage near the target, but the improvement gradually saturates. A denser grid provides a finer quantile-action representation, enabling the RL agent to learn more accurate prediction trajectories, but also increases computation burden. Therefore, we set $|\mathcal{Q}| = 7$ as a practical trade-off.

5.4 Ablation Study

To rigorously evaluate the individual contributions of each component within CORE, we designed the following three variants, each removing one key component of the adaptation mechanism, the updated α, the RL-based policy learning structure, and the quantiles action learning:

- `CORE-w/o-updated` α: replaces the adaptive update of α_t with a fixed confidence level to isolate the role of the error-guided adjustment mechanism in maintaining coverage stability.
- `CORE-w/o-RL`: substitutes the RL-based policy with a supervised MLP-based quantile regressor to validate the advantages of dynamic reward optimization over myopic loss minimization.
- `CORE-w/o-quantiles`: replaces the quantile action set with a single median prediction to quantify the performance gain from our fine-grained, asymmetric quantile selection strategy.

Without the adaptive α_t update, coverage becomes less stable across datasets, indicating that a fixed target level cannot fully correct online miscoverage. Removing the RL policy causes the visible efficiency loss, suggesting that supervised quantile fitting alone is insufficient to actively trade off coverage and interval width. In contrast, replacing quantile actions with a median-centered interval mainly reduces the flexibility of the interval shape, leading to redundant width when the predictive uncertainty is asymmetric. These patterns indicate that CORE does not improve performance through a single component; rather, adaptive α_t corrects coverage drift, RL optimization controls interval efficiency, and quantile actions provide distributional flexibility.

6 Conclusion

We propose CORE, a unified framework that integrates reinforcement learning with conformal prediction through a mutual feedback loop. CORE enables adaptive calibration under distribution shifts, improving uncertainty quantification for time-series forecasting. Extensive empirical evaluations across nine datasets demonstrate that our method achieves valid coverage and superior interval efficiency compared to strong baselines. We also provide rigorous theoretical analysis that supports the validity and efficiency of CORE, supporting uncertainty quantification in dynamic environments.

AI USE STATEMENT

The authors confirm that large language models (LLMs), such as ChatGPT, were only used during the writing stage of this paper for editorial assistance, including grammar correction, spelling checks, and refinement of phrasing. No LLMs were employed in the design of the methodology, execution of experiments, analysis of data, or interpretation of results. All scientific and technical content was solely conceived and produced by the authors.

ETHICS STATEMENT

This work studies uncertainty quantification for time-series forecasting and does not involve human-subject experiments or personal data collection. Although reliable prediction intervals can support risk-aware decisions in high-stakes domains, conformal coverage guarantees depend on their underlying assumptions and should not be treated as absolute safety guarantees under arbitrary distribution shifts. Practical deployment should therefore include domain-specific validation and appropriate human oversight.

REPRODUCIBILITY STATEMENT

We have taken concrete steps to ensure the reproducibility of our work. The complete source code is provided in the supplementary materials. Detailed implementation settings, including model architectures, hyperparameters, and training procedures, are documented in Appendix B.1.4, while the full proofs of our theoretical results are presented in Appendix D. Together, these materials allow independent researchers to replicate both our empirical results and theoretical analysis.

REFERENCES

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36:23047–23074, 2023.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. Mondrian conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pp. 24–38. PMLR, 2021.

Trevor Campbell, Saifuddin Syed, Chiao-Yu Yang, Michael I Jordan, and Tamara Broderick. Local exchangeability. *arXiv preprint arXiv:1906.09507*, 2019.

Ching-An Cheng, Andrey Kolobov, and Adith Swaminathan. Heuristic-guided reinforcement learning. *Advances in Neural Information Processing Systems*, 34:13550–13563, 2021.

Matt Y Cheung, Tucker J Netherton, Laurence E Court, Ashok Veeraraghavan, and Guha Balakrishnan. Regression conformal prediction under bias. *arXiv preprint arXiv:2410.05263*, 2024.

Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pp. 1–68, 1937.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*, 2015.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, Sham Kakade, and Sergey Levine. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. *Advances in Neural Information Processing Systems*, 35:15281–15295, 2022.

Nikunj Gupta, Somjit Nath, and Samira Ebrahimi Kahou. Cammarl: Conformal action modeling in multi agent reinforcement learning. *arXiv preprint arXiv:2306.11128*, 2023.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. Pmlr, 2018.

Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Neural Information Processing Systems*, 2022.

Jianguo Huang, Jianqing Song, Xuanning Zhou, Bingyi Jing, and Hongxin Wei. Torchcp: A python library for conformal prediction. *Journal of Machine Learning Research*, 26(266):1–25, 2025.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 387–395, 2018.

Byol Kim, Chen Xu, and Rina Foygel Barber. Predictive inference is free with the jackknife+-after-bootstrap. *arXiv preprint arXiv:2002.09025*, 2020.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Jonghyeok Lee, Chen Xu, and Yao Xie. Kernel-based optimally weighted conformal time-series prediction. In *International Conference on Learning Representations*, 2025.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Ruipu Li and Alexander Rodríguez. Neural conformal control for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18439–18447, 2025.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Aditya P Mathur and Nils Ole Tippenhauer. Swat: A water treatment testbed for research and training on ics security. In *2016 International Workshop on Cyber-physical Systems for Smart Water Networks*, pp. 31–36. IEEE, 2016.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021.

Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv preprint arXiv:1803.11347*, 2018.

Yaniv Romano, Evan Patterson, and Emmanuel J Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3543–3553, 2019.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pp. 4393–4402. PMLR, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Jaeuk Shin, Jungjin Lee, and Insoon Yang. Egocentric conformal prediction for safe and efficient navigation in dynamic cluttered environments. *arXiv preprint arXiv:2504.00447*, 2025.

Kegan J Strawn, Nora Ayanian, and Lars Lindemann. Conformal predictive safety filter for rl controllers in dynamic environments. *IEEE Robotics and Automation Letters*, 8(11):7833–7840, 2023.

Jiankai Sun, Yiqi Jiang, Jianing Qiu, Parth Nobel, Mykel J Kochenderfer, and Mac Schwager. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. *Advances in Neural Information Processing Systems*, 36:80324–80337, 2023.

Sophia Sun and Rose Yu. Conformal prediction for time-series forecasting with change points. *arXiv preprint arXiv:2509.02844*, 2025.

Ichiro Takeuchi, Quoc V Le, Timothy D Sears, Alexander J Smola, and Chris Williams. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(7), 2006.

Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Paolo Toccaceli and Alexander Gammerman. Combination of inductive mondrian conformal predictors. *Machine Learning*, 108(3):489–510, 2019.

Junxi Wu, Dongjian Hu, Yajie Bao, Shu-Tao Xia, and Changliang Zou. Error-quantified conformal inference for time series. *arXiv preprint arXiv:2502.00818*, 2025.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11559–11569. PMLR, 2021.

Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):11575–11587, 2023a.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pp. 38707–38727. PMLR, 2023b.

Chen Xu, Hanyang Jiang, and Yao Xie. Conformal prediction for multi-dimensional time series by ellipsoidal sets. *arXiv preprint arXiv:2403.03850*, 2024.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*, 2021.

Zhongwen Xu, Hado P van Hasselt, and David Silver. Meta-gradient reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Jianpeng Yao, Xiaopan Zhang, Yu Xia, Zejin Wang, Amit K Roy-Chowdhury, and Jiachen Li. Sonic: Safe social navigation with adaptive conformal inference and constrained reinforcement learning. *arXiv preprint arXiv:2407.17460*, 2024.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR, 2022.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Zhuangdi Zhu, Kaixiang Lin, Anil K Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13344–13362, 2023.

Appendices

A Related Work

A.1 Conformal Prediction

Conformal Prediction (CP) is a distribution-free framework for constructing prediction intervals with formal coverage guarantees under minimal assumptions Romano et al. (2019); Shafer & Vovk (2008); Xu & Xie (2021). By leveraging a nonconformity score function and calibration on held-out data Romano et al. (2019), CP methods can produce reliable uncertainty estimates. Over time, CP has evolved from inductive variants to stratified methods like Mondrian CP, and further to adaptive extensions for dynamic, non-stationary environments.

Inductive CP (ICP) forms prediction intervals by splitting data into training and calibration sets Lei et al. (2018). Extensions including importance-weighted calibration Tibshirani et al. (2019), ensemble resampling Kim et al. (2020), and quantile-based regression Romano et al. (2019) improve its robustness. Mondrian CP (MCP) Boström et al. (2021) further improves calibration by conditioning on predefined categories (e.g., class labels or tree leaves), providing group-conditional guarantees Boström et al. (2021); Toccaceli & Gammerman (2019). However, both ICP and MCP rely on fixed calibration sets and assume data exchangeability Xu & Xie (2023b), making them prone to miscalibration and poor adaptability in dynamic scenarios. Unlike ICP or MCP, online and adaptive CP methods dynamically update prediction intervals using recent residuals to adapt Xu & Xie (2021); Gibbs & Candes (2021). Techniques such as sliding-window residual regression Xu & Xie (2023b) and adaptive significance tuning Gibbs & Candes (2021); Zaffran et al. (2022) help enhance calibration. Building on these foundations, recent papers have introduced more sophisticated adaptation mechanisms. For instance, Conformal PID control Angelopoulos et al. (2023) and ECI Wu et al. (2025) reframe quantile updating through a control-theoretic lens to prospectively correct coverage errors or quantify error magnitudes via smooth feedback. Addressing abrupt or complex non-stationarity requires specialized logic; for instance, CPTC Sun & Yu (2025) enables fast adaptation and maintains validity during change-point occurrences, and kernel-based optimal weighting Lee et al. (2025) prioritizes historically relevant information. Among them, EnbPI Xu & Xie (2021) and SPCI Xu & Xie (2023b) maintain validity and efficient intervals under temporal dependence and distribution shifts commonly observed in time-series scenarios. Our work continues focusing on adaptive CP with the goal of providing valid coverage guarantees under non-exchangeable data.

Multivariate conformal prediction, as another line of research, is orthogonal to our current focus. The univariate CP naturally yields prediction intervals (i.e., line segments), the multivariate setting necessitates the construction of more complex geometric regions, such as copulas Messoudi et al. (2021) or ellipsoids Xu et al. (2024), to capture cross-dimensional dependencies. Therefore, comparing with these methods falls outside the scope of this work.

A.2 Conformal Reinforcement Learning

Reinforcement learning (RL) optimizes decision-making through interaction with environments and is widely used to learn policies under uncertainty and delayed feedback. Its ability to model dynamics and explore makes it well-suited for adaptive systems in non-stationary settings.

Despite its strengths, RL often suffers from training instability, policy divergence, and sensitivity to distributional shifts. To mitigate these limitations, recent work has explored integrating CP into RL. For instance, PlanCP Sun et al. (2023) employs CP for calibrated offline planning, while SoNIC Yao et al. (2024) and Egocentric CP Shin et al. (2025) improve inference and safety in navigation. In multi-agent and safety-critical scenarios, CP-based confidence sets and safety filters Gupta et al. (2023); Strawn et al. (2023) support risk-aware coordination. Despite their effectiveness, these methods typically treat CP as an external or post-hoc module, limiting the impact of uncertainty feedback on policy learning.

It is worth noting that distribution shift remains a challenge in RL. Although meta-RL Nagabandi et al. (2018); Xu et al. (2018) and transfer learning Taylor & Stone (2009); Zhu et al. (2023) improve adaptation through experience sharing, they often lack formal uncertainty quantification. To address

this gap, we propose integrating CP directly into RL learning process, which enables online detection of unreliable predictions with formal coverage guarantees, thus enhancing the safety and robustness of policy learning under distributional shifts. Unlike prior approaches, our goal is to promote **mutual adaptation**, which not only improves policy reliability in non-stationary environments, but also dynamically refines the CP calibration process based on the agent's evolving interaction history.

B Experimental Setup

B.1 Uncertainty Quantification

B.1.1 Dataset

Our evaluation spans nine diverse real-world datasets to ensure a thorough assessment of `CORE`. This exhaustive empirical study aims to demonstrate the versatility of our approach, confirming that the integrated feedback-driven calibration mechanism remains effective under a wide array of practical constraints and evolving data distributions.

- Weather Zhou et al. (2021): Collected at 10-minute intervals in 2020, this dataset includes 21 meteorological variables, with Wet Bulb temperature as the prediction target and the others as auxiliary features.
- Traffic Lai et al. (2018): This dataset contains hourly occupancy rates from 862 freeway sensors in the San Francisco Bay Area, using the last sensor's data as the target and the others as contextual inputs.
- Electricity Lai et al. (2018): This dataset contains hourly electricity consumption data from 321 residential clients (2012–2014). Typically, the last client's usage is the prediction target, with the others as exogenous inputs.
- Illness Lai et al. (2018): This dataset contains weekly influenza-like illness rates across multiple U.S. regions, with the goal of forecasting one region's trend using its own and others' historical data as inputs.
- ETT Zhou et al. (2021): The ETT dataset includes four subsets, ETTh1/ETTh2 (hourly) and ETTm1/ETTm2 (15-minute), with oil temperature as the target and six power metrics as features, spanning July 2016 to July 2018 across two Chinese counties.
- TSLA is a representative stock dataset spanning from from January 1, 2021 to December 31, 2024, which exhibits extreme price fluctuation and characterizes significant real-world distribution shifts.

B.1.2 Metrics

- **Coverage Rate (CVG)**: Measures the proportion of true values covered by the prediction intervals.

$$\text{CVG} = \frac{1}{|D_{\text{test}}|} \sum_{t \in D_{\text{test}}} \mathbb{I}[y_t \in \hat{C}_{1-\alpha}^{(w,t)}], \tag{B.8}$$

  where D_{test} represents the set of test samples, y_t is the ground-truth value, $\hat{C}_{1-\alpha}^{(w,t)}$ is the prediction interval, and $\mathbb{I}[\cdot]$ is the indicator function that returns 1 if the condition holds and 0 otherwise

- **Interval Length (LEN)**: Measures the average length of the prediction intervals.

$$\text{LEN} = \frac{1}{|D_{\text{test}}|} \sum_{t \in D_{\text{test}}} (\hat{u}_t - \hat{l}_t), \tag{B.9}$$

  where $\hat{u}_t$ and $\hat{l}_t$ are the upper and lower bounds of the interval for instance t.

B.1.3 Baselines

- NCC (Li & Rodríguez, 2025) implements a neural-network-based control mechanism that treats calibration as a dynamic task, allowing the model to learn error-correction policies from multi-view data in an end-to-end fashion while preserving formal coverage guarantees.

- ECI (Wu et al., 2025) incorporates a continuous error-quantification term to update thresholds, which enables the model to react to the magnitude of coverage deviations and distribution shifts.
- PID (Angelopoulos et al., 2023) adopts a control-theoretic perspective to achieve long-run coverage by applying a proportional-integral-derivative law to the update of q_t, allowing the algorithm to prospectively adapt to systematic errors such as seasonality and trends.
- SPCI (Xu & Xie, 2023b) constructs prediction intervals by modeling the conditional quantile of residuals, explicitly accounting for temporal dependence without relying on bootstrap ensembles.
- EnbPI (Xu & Xie, 2021) constructs prediction intervals from bootstrap ensembles, achieving approximate coverage without data exchangeability or retraining.
- ACI (Gibbs & Candes, 2021) adapts conformal inference to distribution shifts by recalibrating prediction sets over time to maintain target coverage under non-stationary data.

B.1.4 IMPLEMENT DETAILS

We implement our method in PyTorch and train all models on a single NVIDIA A5000 GPU with 24GB memory. All code is available in the repository `https://anonymous.4open.science/r/CORE-3473`. Each training sample is organized into batches with shape (B, W, D), where $B = 128$ denotes the batch size, $W = 50$ is the input temporal window length, and D is the feature dimension, which varies across datasets. We consider a candidate set $w_c \in \{25, 50, 100, 200, 300\}$, and evaluate the impact of this choice through extensive sensitivity analysis in the section 5.3. The adjustment rate γ for confidence level α_t is set to 0.05, and the initial significance level α is fixed at 0.1 and 0.05. We select $K = 7$ actions from the quantile set $Q = \{q^{(1)}, \ldots, q^{(K)}\}$ to estimate the median actions. The policy is trained using a PPO backbone with separate actor and critic networks, both implemented with 3-layer MLP with ReLU activations. The encoder and decoder modules are flexible and can also be instantiated with alternative architectures such as Transformers. For fairness and model-agnostic comparison, all baseline adaptive CP methods and `CORE` use the same base predictor, 3-layer nonlinear MLP, which ensures that any difference in interval length or coverage stems from the calibration or adaptation mechanism, rather than the capacity of the underlying predictor.

For optimization, the policy and value networks are optimized via two separate Adam optimizers with a learning rate of 5×10^{-5}. All experiments are run for 1000 training steps per episode, and the model is evaluated every 50 steps using average return and empirical coverage metrics.

B.1.5 COMPUTATIONAL COST

Although `CORE` introduces an RL-based policy, its computational overhead remains moderate in our implementation. As we mentioned before, the actor and critic network are both lightweight 3-layer MLPs, and online inference only requires a single forward pass together with sliding-window quantile calibration. In terms of *training and inference*, `CORE` takes 115.29s per epoch and 0.94ms per inference step on Weather, compared with 70.05s and 0.73ms for ACI. On Traffic, `CORE` takes 120.12s per epoch and 0.84ms per step, comparable to ACI's 94.18s and 0.86ms. In terms of *memory and model size*, `CORE` uses 1200 MiB peak GPU memory and 447,561 parameters, compared with 904 MiB and 67,650 parameters for ACI, and 720 MiB and 67,585 parameters for EnbPI. These results indicate that `CORE` increases model size due to the actor-critic components, but keeps online inference cost moderate.

B.2 ANOMALY DETECTION

To evaluate whether `CORE` can be generalized to other conformal inference tasks beyond regression, we adopt the `ECAD` (Xu & Xie, 2021) setup and introduce a variant, `CORE-AD`, which wraps around the original framework to enable unsupervised anomaly detection. Specifically, it leverages calibrated prediction intervals as decision signals by first computing anomaly scores in a sliding calibration window, where at each τ, $s_\tau = \max\{a_\tau^{(\text{low})} - y_\tau, y_\tau - a_\tau^{(\text{up})}, 0\}$. Then calculate p-value for the current test point as $p_t = \frac{1}{w_c} \sum_{i=1}^{w_c} \mathbf{1}(s_{t-i} \geq s_t)$, which quantifies how atypical the current observation is compared to the calibration set. A point is flagged as anomalous if $p_t \leq \alpha$, ensuring control of the false positive rate. We compare `CORE-AD` against a set of unsupervised and supervised anomaly detectors (Han et al., 2022) on standard time-series anomaly detection datasets (SMD (Xu et al., 2021),

SMAP (Hundman et al., 2018), and SWaT (Mathur & Tippenhauer, 2016)). In this task, we evaluate the model's ability to distinguish abnormal from normal time points using **ROC** curves and the area under the curve (**AUC**) as the performance metric. We compare the proposed variant `CORE-AD` against representative unsupervised and semi-supervised baselines, including LSTM (Hundman et al., 2018), IForest (Liu et al., 2008), and DeepSVDD (Ruff et al., 2018).

B.2.1 DATASET

To assess the ablility of variant model, `CORE-AD`, we conduct we experiments on some standard time-series anomaly detection datasets:

- SMAP Hundman et al. (2018) Soil moisture active passive dataset is collected from NASA's Earth observation satellite, containing 25-dimensional data used for anomaly detection in spacecraft systems.
- SMD Xu et al. (2021) Server machine dataset is collected from a large internet company, consisting of 38-dimensional sensor readings over 5 weeks.
- SWaT Mathur & Tippenhauer (2016) This dataset was collected from a real-world secure water treatment testbed, containing 51 tagged sensor and actuator signals recorded at 1-second intervals over 11 days.

B.2.2 METRICS

For anomaly detection task, we evaluate the model's ability to distinguish abnormal from normal time points using receiver operating characteristic(**ROC**) curves, which plot the true positive rate against the false positive rate at varying thresholds. The area under the curve (**AUC**) summarizes this trade-off, with higher values indicating better overall discrimination performance.

- **True Positive Rate (TPR)**: Measures the proportion of actual anomalies that are correctly detected by the model.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{B.10}$$

  where TP is the number of correctly identified anomalies, and FN is the number of missed anomalies.

- **False Positive Rate (FPR)**: Measures the proportion of normal points that are incorrectly classified as anomalies.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{B.11}$$

  where FP is the number of normal points incorrectly flagged as anomalies, and TN is the number of correctly identified normal points.

By integrating the ROC curve, AUROC summarizes the trade-off between TPR and FPR across all thresholds, where a higher AUROC value (closer to 1) indicates better overall detection performance.

B.2.3 BASELINES

- DeepSVDD Ruff et al. (2018) trains a neural network with an anomaly detection-specific objective, aiming to enclose normal data within a minimal-volume hypersphere in the feature space.
- IForest Liu et al. (2008) is a tree-based algorithm that detects anomalies by recursively partitioning the feature space and measuring the path length needed to isolate each point.
- LSTM Hundman et al. (2018) is an unsupervised method, which leverages the long short-term memory recurrent neural networks to achieve prediction performance while maintaining interpretability.

C Experimental Results

C.1 Supplementary Experimental Details on Interval Validity

The performance of our proposed framework is evaluated through two key metrics: prediction interval length (LEN) and empirical coverage rate (COV). We also illustrate the evolution of the adaptive α_t to demonstrate the agent's capability in dynamically regulating the error budget. As shown in Figure C.4, the validation process on the TSLA dataset reveals a distinct transition: during the initial **cold-start** phase, the COV remains near 1.0 due to limited calibration samples. However, as the online adaptation continues, the interval length plateaus toward an optimal efficiency, and the averaged coverage rate stabilizes around the predefined target $(1 - \alpha)$. This suggests that the early conservative behavior is a controlled adaptation process rather than uncontrolled instability, and highlights the role of adaptive α_t in maintaining calibration over time.

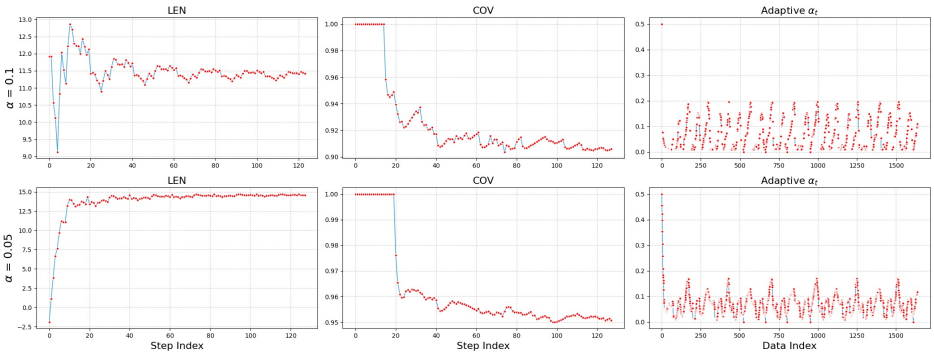


Figure C.4: Validation process on TSLA dataset for $\alpha = 0.1$ (top) and $\alpha = 0.05$ (bottom). Metrics (LEN, COV) and the adaptive parameter α_t are plotted against step/data indices to show system convergence and stability.

C.2 Extension to Anomaly Detection

We visualize the ROC curves of four methods across three datasets to assess anomaly detection performance. As shown in Figure C.5, `CORE-AD` exhibits clear separation from competing methods across nearly the entire ROC curve, indicating stronger robustness under different false-positive rate thresholds. These improvements highlight the advantage of leveraging calibrated prediction intervals as decision signals, providing principled uncertainty quantification while preserving high detection accuracy. Overall, the results confirm that `CORE-AD` is not only effective for anomaly detection but also demonstrates its ability to generalize beyond standard regression tasks as a reliable, uncertainty-aware alternative to traditional detectors.

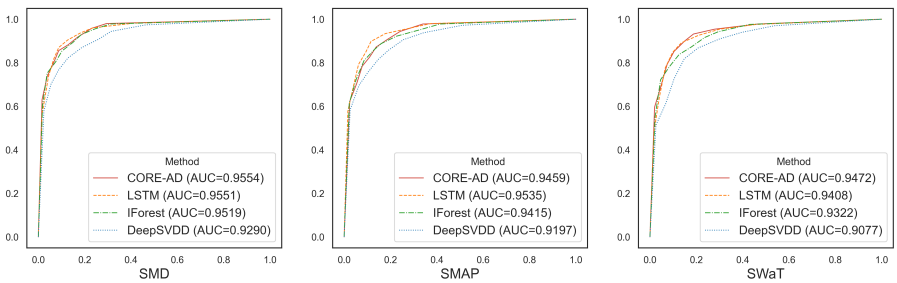


Figure C.5: ROC curves (horizontal-axis: false-positive rate; vertical-axis: true-positive rate)

D Theoretical Proof

For clarity and consistency, the primary notations employed in our formal analysis are consolidated in Table D.2 to distinguish between closely related variables.

Table D.2: Summary of mathematical notations in the method and derivation.

Symbol	Description	Scope
$X \in \mathbb{R}^d$	d-dimensional covariates (features) observed.	Section 2
$Y \in \mathbb{R}$	Univariate ground-truth response associated with X.	Section 2
$(\mathbf{x_t}, y_t) \in \mathbb{R}^{d\times w} \times \mathbb{R}$	The observation and its corresponding label at time step t.	Section 3
w, w_c	The window size of observations $\mathbf{x}_t$ and length of calibration set.	Section 3
$\mathcal{D}_{\text{cal}}, \mathcal{I}_{\text{cal}}$	Calibration set and the corresponding index set	Section 2
$f(\cdot), \hat{f}(\cdot), \tilde{f}(\cdot)$	True, learned, and conformal-accepted predictive functions.	Section 2
$S(x, y)$	Score function used to measure the (non-)conformity.	Sections 2, 3
α, α_t	Target miscoverage level and its time-adaptive counterpart at step t.	Sections 2, 3, 4
$q_{1-\alpha_t}^{(w_c,t)}$	$(1-\alpha_t)$-quantile threshold within w_c of calibration set at t.	Sections 2, 3
$\hat{C}_{1-\alpha_t}^{(w_c,t)}$	The constructed prediction interval within w_c of calibration set at t.	Sections 2, 3, 4
$(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \nu)$	The MDP tuple consisting of states, actions, transitions, rewards, and discount factor ν.	Section 3.1
π_θ	The decision-making policy parametrized by θ.	Section 3.1
γ	The coefficient for the dynamic adjustment of α_t.	Section 3, 4
$\mathcal{Q} := \{q^{(1)}, \ldots, q^{(K)}\}$	Quantile set with K fixed quantiles.	Section 3
$\mathbf{a}_t = \{a_t^{(q)}\}_{q\in\mathcal{Q}}$	The actor sampled quantile-based actions (predictions).	Section 3
$a_t^\star$	Optimal prediction based on the true distribution	Appendix D.2
$\tilde{a}_t$	Empirical prediction actually generated by the model at time t.	Appendix D.2
$F_t(\cdot)$	True cumulative distribution function (CDF) at time t.	Section 4, Appendix D.1, D.2
$\hat{F}_t(\cdot)$	Empirical CDF calculated from calibration samples.	Section 4, Appendix D.1, D.2
$\tilde{F}_t(\cdot)$	Estimated CDF produced by the predictive model $\hat{f}$.	Section 4, Appendix D.1, D.2
$\hat{C}_t$	The empirical prediction interval constructed by the method to achieve the target coverage $1-\alpha_t$.	Section 4, Appendix D.4
$C_t^\star$	The oracle prediction interval derived from the true residual distribution F_t to satisfy the $1-\alpha$ coverage.	Section 4, Appendix D.4
$\hat{e}_t$	The instantaneous miscoverage indicator, where $\hat{e}_t = \mathbf{1}\{Y_t \notin \hat{C}_t\}$.	Section 4, Appendix D.3
$\bar{e}_t$	The empirical miscoverage estimate, where $\bar{e}_t = \frac{1}{\lvert\mathcal{I}_t\rvert}\sum_{i\in\mathcal{I}_t} \hat{e}_i$.	Section 3, 4
$\varepsilon_t, \hat{\varepsilon}_t$	True and estimated (signed) residual at time t, where $\varepsilon_t := y_t - a_t^\star$ and $\hat{\varepsilon}_t := y_t - \tilde{a}_t$.	Appendix D.1, D.2
$\xi_t, \xi_{s:t}$	The instantaneous drift of regression function at time t and cumulative drift over time interval $[s,t]$.	Section 4
δ_t, δ_T^2	The instantaneous approximation error at time t and the expected mean squared approximation error over a horizon T, where $\delta_t := \hat{\varepsilon}_t - \varepsilon_t = a_t^\star - \tilde{a}_t$, and $\delta_T^2 := \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\delta_t^2]$.	Section 4, Appendix D.2

We recall the Theorem 1 here.

Theorem 1 (Master Theorem)**.** *The* CORE *intervals satisfy:*

(Validity) *For any horizon T,*

$$\sup_{t\le T}\left|\Pr\{Y_t \in \hat{C}_{1-\alpha}^{(w_c,t)}\} - (1-\alpha)\right| = \mathcal{O}\left((\log w_c/w_c)^{1/3} + T^{-p/3} + \gamma^{1/2}\right).$$

(Efficiency) *Let $C_t^\star$ denote the oracle shortest valid interval. Then*

$$\sum_{t=1}^{T}\left(|\hat{C}_t| - |C_t^\star|\right) = \tilde{\mathcal{O}}(T^{1-p/3}).$$

We first establish an auxiliary proposition which plays a key role in the proof of master theorem.

Proposition 3 (Marginal coverage of CORE)**.** *Under weak dependence, local regularity, and bounded drift, the conformalized predictor satisfies*

$$\Pr\{Y_t \in \hat{C}_{1-\alpha}^{(w_c,t)}\} \ge 1 - \alpha - \mathcal{O}\left((\log w_c/w_c)^{1/3} + \delta_T^{2/3}\right).$$

Proposition 3 shows that CORE attains near-nominal coverage, with deviations controlled by the prediction error δ_T^2.

To prove Theorem 1, we establish three propositions:

- Proposition 3: establishes marginal coverage with errors controlled by prediction error and distributional drift,
- Proposition 1: bounds the optimization error under PPO,
- Proposition 2: guarantees stability of the adaptive update for α_t.

Their combination yields the stated validity and efficiency guarantees.

D.1 Proof of Proposition 3

Lemma 3 (Rio-type DKW inequality)**.** *Let $\tilde{F}$ denote the empirical CDF of the true residuals within $[t-w_c+1, t]$ over a calibration window of size w_c, and let F denote the true CDF. Under the summable-β weak dependence condition,*

$$\sup_x\left|\tilde{F}(x) - F(x)\right| \le C_1(\log w_c/w_c)^{1/3}.$$

Lemma 4 (Quantile-error transfer)**.** *Let $\delta_T^2 = \frac{1}{T}\sum_{t=1}^{T}(\hat{\varepsilon}_t - \varepsilon_t)^2$. If F_t is L-Lipschitz, then*

$$\sup_x \left|\hat{F}(x) - \tilde{F}(x)\right| \le C_2 \delta_T^{2/3} + 2\sup_x \left|\tilde{F}(x) - F(x)\right|,$$

where $C_2 := L + 1$, with the additive constant serving as a technical slack.

Proof. Let F_t be the true residual CDF at time t, $\tilde{F}_t$ the empirical CDF of true residuals on $[t - w_c + 1, t]$, $\hat{F}_t$ the empirical CDF used by the method. We denote the prediction interval as $\hat{C}_{1-\alpha_t}^{(w_c,t)} := [a_t^{(\text{low})} - \hat{q}_{1-\alpha_t}^{(w_c,t)},\ a_t^{(\text{up})} + \hat{q}_{1-\alpha_t}^{(w_c,t)}]$, where the lower and upper bounds are adjusted by $\hat{q}_{1-\alpha_t}^{(w_c,t)}$. Here $\hat{q}_{1-\alpha_t}^{(w_c,t)}$ denotes the empirical $(1-\alpha_t)$-quantile of the residual distribution estimated from the calibration window size and can be defined as $\hat{q}_{1-\alpha_t}^{(w_c,t)} := \inf\{z : \hat{F}_t(z) \ge 1 - \alpha_t\}$. By construction of the interval, the event $\{y_t \in \hat{C}_{1-\alpha_t}^{(w_c,t)}\}$ is equivalent to the event that the absolute residual at time t is at most $\hat{q}_{1-\alpha_t}^{(w_c,t)} + \max\{a_t^{0.5} - \alpha_t^{(low)}, \alpha_t^{(up)} - a_t^{0.5}\}$. Hence, $F_t(\hat{q}_{1-\alpha_t}^{(w_c,t)} + \Delta_{\min}) \le \Pr\{y_t \in \hat{C}_{1-\alpha_t}^{(w_c,t)}\} \le F_t(\hat{q}_{1-\alpha_t}^{(w_c,t)} + \Delta_{\max})$, where $\Delta_{\min} = \min\{a_t^{0.5} - \alpha_t^{(low)}, \alpha_t^{(up)} - a_t^{0.5}\}$, and $\Delta_{\max} = \max\{a_t^{0.5} - \alpha_t^{(low)}, \alpha_t^{(up)} - a_t^{0.5}\}$. For simplicity, denote $\hat{q}_{1-\alpha_t}^{(w_c,t)}$ as $\hat{q}$ and we start from the indicator rewriting:

$$\begin{aligned}
\left|\Pr\{y_t \in \hat{C}_{1-\alpha_t}^{(w_c,t)}\} - (1-\alpha_t)\right| &\le \max_{z \in \{\hat{q}+\Delta_{\min}, \hat{q}+\Delta_{\max}\}} |F_t(z) - (1-\alpha_t)| \\
&= \max_z |F_t(z) - \hat{F}_t(z) + \hat{F}_t(z) - (1-\alpha_t)| \\
&\le \max_z |F_t(z) - \hat{F}_t(z)| + \max_z |\hat{F}_t(z) - (1-\alpha_t)| \\
&\le \sup_{u \in \mathbb{R}} |F_t(u) - \hat{F}_t(u)| + \max_z \left|\hat{F}_t(z) - \hat{F}_t(\hat{q}) + \hat{F}_t(\hat{q}) - (1-\alpha_t)\right| \\
&\le \|F_t - \hat{F}_t\|_\infty + \max_z \left|\hat{F}_t(z) - \hat{F}_t(\hat{q})\right| + \left|\hat{F}_t(\hat{q}) - (1-\alpha_t)\right| \\
&= \|F_t - \hat{F}_t\|_\infty + \left(\hat{F}_t(\hat{q} + \Delta_{\max}) - \hat{F}_t(\hat{q})\right) + \left|\hat{F}_t(\hat{q}) - (1-\alpha_t)\right| \\
&\le \|F_t - \hat{F}_t\|_\infty + \frac{1}{w_c}\sum_{i=t-w_c+1}^{t} \mathbf{1}\{\hat{q} \le \epsilon_i \le \hat{q} + \Delta_{\max}\} + \frac{1}{w_c}\sum_{i=t-w_c+1}^{t} \mathbf{1}\{\epsilon_i \le \hat{q}\} \\
&\le \|F_t - \hat{F}_t\|_\infty + \frac{2}{w_c}
\end{aligned}$$

The reason for the third inequality $\left|\hat{F}_t(\hat{q}_{1-\alpha_t}^{(w_c,t)}) - (1-\alpha_t)\right| \le \frac{1}{w_c}$ is as follows. Since the empirical CDF $\hat{F}_t$ is a step function with jump size at most $1/w_c$, as $\hat{F}_t(\hat{q}_{1-\alpha_t}^{(w_c,t)}) = \frac{1}{w_c}\sum_{i=1}^{w_c} \mathbf{1}\{\epsilon_i \le \hat{q}_{1-\alpha_t}^{(w_c,t)}\}$. Next, decompose the CDF error and insert window-wise bounds:

$$\begin{aligned}
\|\hat{F}_t - F_t\|_\infty &\le \|\hat{F}_t - \tilde{F}_t\|_\infty + \|\tilde{F}_t - F_t\|_\infty \\
&\le (L+1)\delta^{2/3} + 2\|\tilde{F}_t - F_t\|_\infty + \|\tilde{F}_t - F_t\|_\infty \\
&\le (L+1)\delta^{2/3} + 3\|\tilde{F}_t - F_t\|_\infty
\end{aligned}$$

According to Lemma 1 we have $\sup_x \left|\tilde{F}(x) - F(x)\right| \le C_1(\log w_c / w_c)^{1/3}$, and substituting this bound into the above inequality yields $\|\hat{F}_t - F_t\|_\infty \le (L+1)\delta^{2/3} + 3C_1(\log w_c / w_c)^{1/3}$.

$$\begin{aligned}\left|\Pr\{y_t \in \hat{C}^{(w_c,t)}_{1-\alpha_t}\} - (1-\alpha_t)\right| &\le \|\hat{F}_t - F_t\|_\infty + \frac{1}{w_c} \\ &\le (L+1)\delta^{2/3} + 3\|\tilde{F}_t - F_t\|_\infty + \frac{1}{w_c} \\ &\le (L+1)\delta^{2/3} + 3C_1\big(\log w_c/w_c\big)^{1/3} + \frac{1}{w_c} \\ &\le 3C_1\big(\log w_c/w_c\big)^{1/3} + (L+1)\delta^{2/3} + \frac{1}{w_c}\end{aligned}$$

Therefore, we can get $\Pr\{Y_t \in \hat{C}^{(w_c,t)}_{1-\alpha}\} \ge 1 - \alpha - \mathcal{O}\left((\log w_c/w_c)^{1/3} + \delta_T^{2/3}\right)$. □

D.2 Proof of Proposition 1

Proof. Let $\hat{\varepsilon}_t := y_t - \tilde{a}_t, \quad \varepsilon_t := y_t - a_t^\star, \quad \delta_t := \hat{\varepsilon}_t - \varepsilon_t = a_t^\star - \tilde{a}_t, \quad \delta_T^2 := \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\delta_t^2]$.

The learning objective of PPO is written as:

$$J(\theta) = \mathbb{E}\left[\lambda_{\text{acc}}(1 - |\delta_t|) + \lambda_{\text{len}} r_{\text{len}} + \lambda_{\text{cov}} r_{\text{cov}}\right], \quad \lambda_{\text{acc}} \ge c_0 > 0, \quad \lambda_{\text{len}}, \lambda_{\text{cov}} \le c_1.$$

Let $R(\theta) := \mathbb{E}[|\delta_t|]$ (MAE w.r.t. oracle median), $R_2(\theta) := \mathbb{E}[\delta_t^2] = \mathbb{E}\big[(a_t^\star - \tilde{a}_t)^2\big]$ (MSE).

Assume R is L_R-smooth, then for any step $\eta_k > 0$,

$$R(\theta_{k+1}) \le R(\theta_k) + \langle \nabla R(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L_R}{2}\|\theta_{k+1} - \theta_k\|^2, \tag{D.12}$$

and the policy gradient estimator is unbiased with bounded the second moment, and the non-accuracy terms include a bounded deterministic shift $b(\theta_k)$:

$$\zeta_k := \widehat{\nabla} J(\theta_k) - \nabla J(\theta_k), \quad \mathbb{E}[\zeta_k \mid \theta_k] = 0, \quad \mathbb{E}[\|\zeta_k\|^2 \mid \theta_k] \le \sigma^2,$$
$$\nabla J(\theta_k) = -\lambda_{\text{acc}} \nabla R(\theta_k) + b(\theta_k), \quad \|b(\theta_k)\| \le B_{\text{grad}}.$$

So that $\widehat{\nabla} J(\theta_k) = -\lambda_{\text{acc}} \nabla R(\theta_k) + b(\theta_k) + \zeta_k$. For $\theta_{k+1} = \theta_k + \eta_k \widehat{\nabla} J(\theta_k)$ and taking the expectation of Eq. D.12 and substituting it into the above equation, we can obtain

$$\mathbb{E}[R(\theta_{k+1})|\theta_k] \le R(\theta_k) + \eta_k \langle \nabla R(\theta_k), \mathbb{E}[\widehat{\nabla} J(\theta_k)|\theta_k] \rangle + \frac{L_R}{2}\eta_k^2 \mathbb{E}[\|\widehat{\nabla} J(\theta_k)\|^2 | \theta_k].$$

According to $\mathbb{E}[\zeta_k|\theta_k] = 0$, $\mathbb{E}[\widehat{\nabla} J(\theta_k)|\theta_k] = -\lambda_{\text{acc}} \nabla R(\theta_k) + b(\theta_k)$. Consequently,

$$\begin{aligned}\eta_k \langle \nabla R(\theta_k), \mathbb{E}[\widehat{\nabla} J(\theta_k)|\theta_k] \rangle &= \eta_k \langle \nabla R(\theta_k), -\lambda_{\text{acc}} \nabla R(\theta_k) + b(\theta_k) \rangle \\ &= -\eta_k \lambda_{\text{acc}} \|\nabla R(\theta_k)\|^2 + \eta_k \langle \nabla R(\theta_k), b(\theta_k) \rangle \\ &= -\eta_k \lambda_{\text{acc}} \|\nabla R(\theta_k)\|^2 + \frac{\eta_k \lambda_{\text{acc}}}{2}\|\nabla R(\theta_k)\|^2 + \frac{\eta_k}{2\lambda_{\text{acc}}}\|b(\theta_k)\|^2 \\ &= -\eta_k \lambda_{\text{acc}} \|\nabla R(\theta_k)\|^2 + \frac{\eta_k \lambda_{\text{acc}}}{2}\|\nabla R(\theta_k)\|^2 + \frac{\eta_k}{2\lambda_{\text{acc}}} B_{\text{grad}}^2\end{aligned}$$

By the Cauchy–Schwarz inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we have

$$\begin{aligned}\mathbb{E}[R(\theta_{k+1})|\theta_k] \le & R(\theta_k) - \frac{\eta_k \lambda_{\text{acc}}}{2}\|\nabla R(\theta_k)\|^2 \\ & + \frac{L_R}{2}\eta_k^2 (3\lambda_{\text{acc}}^2 \|\nabla R(\theta_k)\|^2 + 3B_{\text{grad}}^2 + 3\sigma^2) + \frac{\eta_k}{2\lambda_{\text{acc}}} B_{\text{grad}}^2\end{aligned}$$

For simplicity, we denote $A := \frac{3}{2} L_R \lambda_{\text{acc}}^2$, $B := \frac{3}{2} L_R (B_{\text{grad}}^2 + \sigma^2)$, $C := \frac{1}{2\lambda_{\text{acc}}} B_{\text{grad}}^2$. At this point, the term C originates from the bias term $b(\theta)$ that comes from auxiliary rewards (e.g. length or

coverage penalties). For analysis, we absborb auxiliar reward terms into the objective and work with the total reward $R_{\text{total}} := \mathbb{E}[|\delta_t| - \frac{\lambda_{\text{len}}}{\lambda_{\text{acc}}} r_{\text{len}} - \frac{\lambda_{\text{cov}}}{\lambda_{\text{acc}}} r_{\text{cov}}]$, so that the update direction satisfies $\nabla J(\theta) = -\lambda R_{\text{total}}(\theta)$ and the bias disappears, i.e., $b(\theta) \equiv 0$. Thus we obtain:

$$\mathbb{E}[R(\theta_{k+1})|\theta_k] \leq R(\theta_k) - \eta_k(\frac{\lambda_{\text{acc}}}{2} - A\eta_k)\|\nabla R(\theta_k)\|^2 + B\eta_k^2. \tag{D.13}$$

Make step size small enough $\eta_k \leq \frac{\lambda_{\text{acc}}}{4A}$, then $\frac{\lambda_{\text{acc}}}{2} - A\eta_k \geq \frac{\lambda_{\text{acc}}}{4} \doteq c_0$. Then, we taking the expectation of Eq. D.13 and summing over $k = 1$ to T, we obtain

$$\sum_{k=1}^{T} c_0 \eta_k \mathbb{E}[\|\nabla R(\theta_k)\|^2] \leq \mathbb{E}[R(\theta_1)] - \mathbb{E}[R(\theta_{T+1})] + B\sum_{k=1}^{T} \eta_k^2.$$

$$\sum_{k=1}^{T} \eta_k \mathbb{E}[\|\nabla R(\theta_k)\|^2] \leq \frac{R(\theta_1) - R^\star}{c_0} + \frac{B}{c_0}\sum_{k=1}^{T} \eta_k^2,$$

where $R^\star$ represents the optimal value of the objective function $R^\star = \inf_\theta R(\theta)$. We choose a diminishing step size $\eta_k = \eta_0/\sqrt{k}$. The two sums that appear in the bound admit the standard estimates

$$\sum_{k=1}^{T} \eta_k = \eta_0 \sum_{k=1}^{T} \frac{1}{\sqrt{k}} \;\approx\; \eta_0 \int_1^T \frac{dx}{\sqrt{x}} = 2\eta_0(\sqrt{T} - 1) \;\asymp\; 2\eta_0\sqrt{T},$$

and

$$\sum_{k=1}^{T} \eta_k^2 = \eta_0^2 \sum_{k=1}^{T} \frac{1}{k} \;\approx\; \eta_0^2\,(\log T + \rho) \;\asymp\; \eta_0^2 \log T,$$

where ρ is a constant and $\asymp$ denotes equality up to absolute constants (i.e., same order). The first sum scales like $\sqrt{T}$, while the second scales like $\log T$. Dividing both sides of the previous inequality by $\sum_{k=1}^{T} \eta_k$ therefore yields an average-gradient bound that decays at rate $\tilde{\mathcal{O}}(T^{-1/2})$. So that

$$\frac{1}{\sum_{k=1}^{T} \eta_k} \sum_{k=1}^{T} \eta_k \mathbb{E}\|\nabla R(\theta_k)\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right)$$

Having established that the mean absolute error (MAE) decays at rate $\tilde{\mathcal{O}}(T^{-1/2})$, we next convert this guarantee into a bound for the mean squared error (MSE) δ_T^2. To this end, we consider two standard conditions under which MAE and squared error can be related: (i) bounded parameter error $|\delta_t| \leq B_{\text{err}}$ a.s., or (ii) a squared-loss descent/PL on $R_2(\theta) := \mathbb{E}[\delta_t^2]$.

Case (i) boundedness: If we assume the individual parameter error is uniformly bounded, i.e. $|\delta_t| \leq B_{\text{err}}$ almost surely, then the squared error can be controlled by the absolute error $\delta_t^2 \leq B_{\text{err}}|\delta_t|$. Taking expecations and averaging over t, we obtain:

$$\delta_T^2 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\delta_t^2] \leq B_{\text{err}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}|\delta_t| = B_{\text{err}} \frac{1}{T}\sum_{t=1}^{T} R(\theta_t) = \mathcal{O}\left(T^{-1/2}\right).$$

Case (ii) squared-loss PL (stronger): Alternatively, we assume the squared loss $R_2(\theta) := \mathbb{E}[\delta_t^2]$, then can directly analyze its descent under the update. Specifically, suppose for some constants $\tilde{\lambda}, \tilde{\mu} > 0$,

$$\mathbb{E}\left[R_2(\theta_{k+1}) \mid \theta_k\right] \leq R_2(\theta_k) - \tilde{\lambda}\eta_k \|\nabla R_2(\theta_k)\|^2 + \tilde{C}\eta_k^2, \quad R_2(\theta) - R_2^\star \leq \frac{1}{2\tilde{\mu}}\|\nabla R_2(\theta)\|^2,$$

then with $\eta_k = \eta_0/\sqrt{k}$,

$$\delta_T^2 = R_2(\theta_T) = \mathcal{O}\left(T^{-1}\right).$$

Unified rate: there exists $p \in [1, 2]$ such that

$$\delta_T^2 = \mathcal{O}\left(T^{-p/2}\right) \quad \Longrightarrow \quad \delta_T^{2/3} = \left(\delta_T^2\right)^{1/3} = \mathcal{O}\left(T^{-p/3}\right).$$

Therefore, the term entering Proposition 1 is $\mathcal{O}(T^{-p/3})$. Our implementation includes behavior cloning, so we adopt the square-loss descent/PL condition for $R_2(\theta) = \mathbb{E}[\delta_t^2]$ (Case (ii)) in our coverage bound. We also include a weaker, assumption-robust alternative (Case (i) with bounded or sub-Gaussian errors), which guarantees a valid but slower rate when the square-loss descent/PL condition may not hold. This provides a portable guarantee for ablations and external reproductions. □

D.3 Proof of Proposition 2

Proof. The adaptive update in practice is based on an empirical miscoverage estimate $\hat{\epsilon}_t$, aggregated from a sliding window. However, for theoretical analysis it is essential to work with the instantaneous indicator $\hat{e}_t = \mathbf{1}\{Y_t \notin \hat{C}_t\}$. Unlike $\hat{\epsilon}_t$, the variable $\hat{e}_t$ admits the decomposition

$$\hat{e}_t = e_t + \eta_t, \quad e_t := \mathbb{E}[\hat{e}_t \mid \mathcal{F}_{t-1}], \quad \eta_t := \hat{e}_t - e_t, \quad \mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}] = 0,$$

which ensures that the noise term has zero conditional mean. Hence, the recursion is written as $\alpha_{t+1} = \alpha_t + \gamma_t(\alpha - \hat{e}_t)$.

Let $e_t := \mathbb{E}[\hat{e}_t|\mathcal{F}_{t-1}]$, $\eta_t := \hat{e}_t - e_t$, $\Delta_t := \alpha_t - e_t$, $d_t := e_{t+1} - e_t$, $b_t := \alpha - e_t$, then one-step recursion is:

$$\begin{aligned}\Delta_{t+1} &= \alpha_{t+1} - e_{t+1} = \alpha_t + \gamma_t(\alpha - \hat{e}_t) - e_{t+1} \\ &= (\alpha_t - e_t) + \gamma_t(\alpha - \hat{e}_t) + e_t - e_{t+1} \\ &= \Delta_t + \gamma_t(b_t - \eta_t) - d_t\end{aligned}$$

To get contraction in Δ_t, we linearize the mapping $\alpha_t \mapsto e_t$ around equilibrium and assume a uniformly positive local slope:

$$b_t = \alpha - e_t = -\kappa_t\Delta_t + r_t, \quad \kappa_t \in [\kappa_-, \kappa_+], \quad |r_t| \le L_b|\Delta_t|^2,$$

where r_t is the higher-order remainder.

Then, the above one-step recursion becomes:

$$\begin{aligned}\Delta_{t+1} &= (1 - \gamma_t\kappa_t)\Delta_t - \gamma_t\eta_t + \gamma_t r_t - d_t \\ \Delta_{t+1}^2 &= (1 - \gamma_t\kappa_t)^2\Delta_t^2 + \gamma_t^2\eta_t^2 + \gamma_t^2 r_t^2 + d_t^2 \\ &\quad - 2(1 - \gamma_t\kappa_t)\gamma_t\Delta_t\eta_t + 2(1 - \gamma_t\kappa_t)\gamma_t\Delta_t r_t - 2(1 - \gamma_t\kappa_t)\Delta_t d_t \\ &\quad - 2\gamma_t^2\eta_t r_t + 2\gamma_t\eta_t d_t - 2\gamma_t r_t d_t.\end{aligned}$$

Take conditional expectation using $\mathbb{E}[\eta_t|\mathcal{F}_{t-1}] = 0$, $\mathbb{E}[\eta_t^2|\mathcal{F}_{t-1}] \le \sigma_\eta^2$ (Bernoulli $\implies \sigma_\eta^2 \le \frac{1}{4}$). Hence all cross terms containing η_t vanish in conditional expectation. The remaining cross terms are controlled by Young's inequality $2ab \le a^2 + b^2$, and absorbed into constants. For the leading factor:

$$(1 - \gamma_t\kappa_t)^2\Delta_t^2 = 1 - 2\kappa_-\gamma_t + \kappa_t^2\gamma_t^2 \le 1 - 2\kappa_-\gamma_t + c_1\gamma_t^2$$

$$\mathbb{E}[\Delta_{t+1}^2 \mid \mathcal{F}_{t-1}] \le (1 - 2\kappa_-\gamma_t + c_1\gamma_t^2)\Delta_t^2 + \sigma_\eta^2\gamma_t^2 + c_2(\gamma_t^2 r_t^2 + d_t^2),$$

where c_1 and c_2 are finite numerical constants.

Denote the mean-square tracking error by $V_t := \mathbb{E}[\Delta_t^2] = \mathbb{E}[(\alpha_t - e_t)^2]$. As we have constant step size γ. In this case, the leading term $(1 - 2\kappa_-\gamma + c_1\gamma^2)$ is strictly smaller than 1 when γ is sufficiently small, ensuring contraction in expectation. All other terms on the right-hand side are of order γ^2, leading to a recursion of the form:

$$\mathbb{E}[\Delta_{t+1}^2] \le (1 - c\gamma)\mathbb{E}[\Delta_t^2] + C\gamma^2,$$

for some constants $c, C > 0$. Solving this recursion yields:

$$V_t = \mathbb{E}[\Delta_t^2] = \mathcal{O}(\gamma) \quad \implies \quad \mathbb{E}[|\alpha_t - e_t|] = \mathcal{O}(\gamma^{1/2}).$$

Thus, the tracking error remains bounded and stable, though it does not vanish asymptotically. □

D.4 Proof of Theorem 1

Proof. According to the above results, we now combine the three propositions. Proposition 3 provides a marginal coverage guarantee:

$$\Pr\{Y_t \in \hat{C}_{1-\alpha_t}^{(w_c,t)}\} \geq 1 - \alpha - \mathcal{O}\Big((\log w_c/w_c)^{1/3} + \delta_T^{2/3}\Big).$$

The bound contains two error sources, the prediction error δ_T^2 and the adaptivity error in α_t. Proposition 1 shows that $\delta_T^2 = \mathcal{O}(T^{-p/2})$, which converts into a $\mathcal{O}(T^{-p/3})$ contribution in the coverage bound. Proposition 2 further shows that the adaptive update of α_t tracks the true miscoverage within $\mathcal{O}(\gamma^{1/2})$. Substituting these results into the bound of Proposition 3 yields the validity guarantee:

$$\sup_{t \leq T}\Big|\Pr\{Y_t \in \hat{C}_{1-\alpha_t}^{(w_c,t)}\} - (1-\alpha)\Big| = \mathcal{O}\Big((\log w_c/w_c)^{1/3} + T^{-p/3} + \gamma^{1/2}\Big).$$

Finally, we establish the efficiency result. Recall that the excess width of the constructed CORE intervals over the oracle benchmark $C_t^\star$ is governed by the prediction error δ_T^2. In particular, the deviation in interval width can be bounded by a function of $\delta_T^{2/3}$, reflecting the robustness of the quantile-based construction to estimation errors. Since we have already shown that $\delta_T^2 = O(T^{-p/2})$ for some $p \in [1, 2]$, it follows that each per-round excess width satisfies $|\hat{C}_t| - |C_t^\star|$. Summing over T rounds then yields

$$|\hat{C}_t| - |C_t^\star| \sim \delta_t^{2/3} = \tilde{O}(T^{-p/3}), \quad \sum_{t=1}^{T}\big(|\hat{C}_t| - |C_t^\star|\big) = \tilde{O}(T^{1-p/3}).$$

□