Conformalized Online Learning: Online CalibrationWithout a Holdout Set

Shai Feldman, Stephen Bates, and Yaniv Romano

Kiho Park
parkkiho@uchicago.edu

1 Introduction

To confidently deploy learning models in high-stakes applications, we need both high predictive accuracy and reliable safeguards to handle unanticipated changes in the underlying data generating process. Reasonable accuracy on a fixed validation set is not enough; we must also quantify uncertainty to correctly handle hard input points and take into account shifting distributions. In this work, we encode uncertainty in a rigorous manner via prediction intervals/sets that augment point predictions. In this paper, we introduce a novel calibration framework that can wrap any online learning algorithm (e.g., an LSTM model trained online) to construct prediction sets with guaranteed validity.

Suppose we are given data stream $\{(X_t, Y_t)\}_{t\in\mathbb{N}}$, where $X_t \in \mathcal{X} = \mathbb{R}^p$ and $Y_t \in \mathcal{Y}$. At each time step $t \in \mathbb{N}$, we use $\{(X_i, Y_i)\}_{i=1}^{t-1}$ along with x_t to construct a prediction set $\widehat{C}_t \subseteq \mathcal{Y}$, guaranteed to attain any user-specified coverage frequency $1-\alpha$:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{Y_t\in\widehat{C}_t(X_t)\}} = 1-\alpha \tag{1}$$

Conformal inference is a generic approach for constructing prediction sets with valid coverage under exchangeability. Also, extensions to confromal inference were developed. However, these methods are not guaranteed to construct valid predcition sets for time-series data with less regular distributional shifts. By contrast, *adaptive conformal inference* (ACI) Gibbs and Candés (2021) is a groundbreaking calibration scheme for constructing prediction sets with valid coverage in the sense of (1) for any arbitrary time-varying data distribution. However, this method is computationally inefficient and the model will slowly adapt to changes in the distribution.

In this work, we introduce *rolling conformal inference* (`Rolling CI`): a novel calibration procedure that addresses the drawbacks of ACI by removing the requirement of a holdout set. Our scheme is guaranteed to attain any pre-specified coverage level without making any assumptions on the data distribution. It can be integrated with any online learning algorithm, and can utilize all the data for training.

1.1 A Synthetic Example

To illustrate the challenges in constructing prediction intervals with accurate coverage rate, we conduct a synthetic experiment using time-series data with drastic distributional shifts. We generate a response variable $Y_t \in \mathbb{R}$ at time t as $Y_t = \frac{1}{2}Y_{t-1} + |\beta_t^T X_t|\omega_t + \varepsilon_t$. A distributional shift occurs every 500 time steps.

Next, we fit a neural network as a base model and a construct prediction interval using ACI and our proposed method, where we set the target coverage level to 90%. In the figure, we can see that the proposed method not only constructs narrower intervals, but also accurately attains the desired coverage level for each group ($\omega_t = 1$ and otherwise), in contrast to ACI.

1.2 Our contribution

Valid coverage. Theorem 1 guarantees that our calibration method is guaranteed to satisfy the requirement in (1) for any time-varying data distribution and predictive model.

Faster reaction to distribution shift. In contrast to existing techniques, our proposal does not reserve a holdout set for calibrating the prediction interval at each inference step. This allows fitting the model on the entire data stream, and utilizing the most recent observations to better track the underlying distribution.

Compatibility with online learning models. Our method works together with any black-box online learning algorithm to adaptively control any parameter that encodes the size of the prediction set. Since we avoid costly model refitting, our procedure is more computationally efficient.

2 Background

In this section, we provide a short description of uncertainty quantification techniques that we build upon.

2.1 Split Conformal Prediction

Split conformal prediction (SCP) is a general tool for constructing prediction sets in regression or classification tasks. The idea is to split the observed labeled data $\{(X_i, Y_i)\}_{i=1}^n$ into training and calibration sets, fit a model on the training set, and evaluate the model goodness-of-fit on the reserved holdout calibration points. This goodness-of-fit information is then cleverly used to assess the uncertainty in future predictions. The coverage guarantee does not hold for time-varying data with arbitrary distributional shifts, as the i.i.d. assumption would not hold anymore.

2.2 Adaptive Conformal Inference

For $\{(X_t, Y_t)\}_{t=1}^{T-1}$ and X_T with unknown Y_T, ACI constructs a prediction set for T_Y by invoking SCP as a subroutine, with carefully tuned target coverage level $1-\alpha_T$ for the time step T, $C_T^{\text{ACI}}(X_T) = C_{1-\alpha_T}^{\text{SCP}}(X_T)$. Above, α_T is tuned recursively according to the following rule:

$$\alpha_T = \alpha_{T-1} + \gamma(\alpha - \text{err}_{T-1}), \quad \text{where} \quad \text{err}_T = \begin{cases} 1, & Y_T \notin C_T^{\text{ACI}}(X_T) \\ 0, & \text{otherwise} \end{cases}$$

which ensures that (1) will rigorously hold. It tracks the empirical long-range coverage rate, and adaptively increases α_T to compensate for systematic under coverage frequency.

To implement ACI, one needs to fit a new prediction model once a new sample (X_T, Y_T) is observed. This may be impractical. Also, splitting the data at random breaks the dependency structure across the time horizon. To address these issues, we combine ACI with online sequential split conformal prediction (OSSCP) Zaffran et al. (2022) to form a strong, feasible baseline method. We refer to it as `ACI-Online`. We will compare our proposed `Rolling CI` to this baseline method in the mnumberical experiments.

3 Proposed Method

3.1 General formulation

Algorithm 1 `Rolling CI`

Input: Data $\{(X_t, Y_t)\}_{t=1}^{T} \subseteq \mathcal{X} \times \mathcal{Y}$, given as a stream, miscoverage level $\alpha \in (0,1)$, a step size $\gamma > 0$, an interval/set constructing function $f : (\mathcal{X}, \mathbb{R}, \mathbb{M}) \to 2^{\mathcal{Y}}$ and an online learning model $\mathcal{M}$.

Process: Initialize $\theta_0 = 0$.

for $t = 1, \cdots, T$ **do**

Construct a prediction set for the new point X_t: $\widehat{C}_t^{\mathrm{RCI}}(X_t) = f(X_t, \theta_t, \mathcal{M}_t)$.

Obtain Y_t.

Compute $\mathrm{err}_t = \mathbb{1}\{Y_t \notin \widehat{C}_t^{\mathrm{RCI}}(X_t)\}$

Update $\theta_{t+1} = \theta_t + \gamma(\mathrm{err}_t - \alpha)$.

Fit the model $\mathcal{M}_t$ on (X_t, Y_t) and obtain the updated model $\mathcal{M}_{t+1}$.

end for

Output: Uncertainty sets $\widehat{C}_t^{\mathrm{RCI}}(X_t)$ for each time step $t \in \{1, \cdots, T\}$.

Theorem 1. *Suppose that $f : (\mathcal{X}, \mathbb{R}, \mathbb{M}) \to 2^{\mathcal{Y}}$ is an interval/set constructing function. In addition, suppose that there exist constants m and M such that for all X and $\mathcal{M}$, $f(X, \theta, \mathcal{M}) = \emptyset$ for all $\theta < m$ and $f(X, \theta, \mathcal{M}) = \mathcal{Y}$ for all $\theta > M$. Consider the following series of calibrated intervals: $\{\widehat{C}_t^{\mathrm{RCI}}(X_t)\}_{t=1}^{\infty}$. The calibrated intervals satisfy the coverage requirement in (1).*

Crucially, this theorem states that the coverage guarantee of `Rolling CI` holds for any distribution $\{P_{X_t,Y_t}\}_t$, any set0valued function f, and any sequence of online-updated predictive models $\{\mathcal{M}_t\}_t$. The empirical coverage falls within a C/T factor of the desired level $1 - \alpha$. Furthermore, our new formulation unlocks new design possibilities: we can define any prediction set function f while guaranteeing the validity of its output by calibrating any parameter θ_t.

3.2 `Rolling CI` for Regression

In this section, we introduce two concrete formulations for the function f for regression problems. They seek to construct prediction intervals that are short and highly adaptive.

3.2.1 Calibrating in $\mathcal{Y}$ Scale

Consider a quantile regression model $\mathcal{M}_t$ that produces estimates for the lower $\widehat{q}_t(X_t, \alpha/2)$ and upper $\widehat{q}_t(x_t, 1 - \alpha/2)$ conditional quantiles of the distribution of $Y_t \mid X_t$. In practice, $\mathcal{M}_t$ may not be accurate, and thus may result in a coverage rate lower or higher than the desired one. Taking inspiration from the method of conformalized quantile regression (CQR), we suggest the following interval construction function for `Rolling CI`:

$$f(X_t, \theta_t, \mathcal{M}_t) = [\widehat{q}_t(X_t, \alpha/2) - \varphi(\theta_t), \widehat{q}_t(X_t, 1 - \alpha/2) + \varphi(\theta_t)]$$

We present two design options for the stretching function φ.

Linear. The most natural option is $\varphi(x) = x$, but it might be sub-optimal when an aggressive and fast calibration is required. The calibration parameter θ_t might be updated too slowly, resulting in an unnecessary delay in the interval's adjustment.

Exponential. The exponential stretching function, defined as $\varphi(x) = e^x - 1$ for $x > 0$ and $\varphi(x) = -e^{-x} + 1$ for $x \leq 0$. It updates $\varphi(\theta_t)$ gently when the calibration is mild, and faster as the calibration is more aggressive ($\varphi(\theta_t)$ is away from zero).

3.2.2 Calibrating on the Quantile Scale

As an alternative, one can modify the interval's length by tuning the raw miscoverage level $\tau_t = \varphi(\theta_t)$ requested from the model:

$$f(X_t, \theta_t, \mathcal{M}_t) = [\widehat{q}_t(X_t, \tau_t/2), \widehat{q}_t(x_t, 1 - \tau_t/2)]$$

The straightforward option is to set $\varphi(\theta) = -\tau$, where $\theta_t \in [-1, 0]$.

3.3 `Rolling CI` for Classification

Consider a multi-class classification problem, where the target variable is discrete and unordered $t \in \mathcal{T} = \{1, 2, \cdots, K\}$. Suppose we are handed with a classifier that estimates $P_{Y_t|X_t}(Y_t = y \mid X_t = x)$ for each y, i.e. $\mathcal{M}_t(X_t, y) \in [0, 1]$ and $\sum_{y \in \mathcal{Y}} \mathcal{M}_t(X_t, y) = 1$. We define the prediction set constructing function as:

$$f(X_t, \theta_t, \mathcal{M}_t) = \{y : \mathcal{M}_t(X_t, y) \geq \varphi(\theta_t)\}$$

where one can choose $\varphi(x) = -x$. While this procedure is guaranteed to attain the pre-specified coverage level $1 - \alpha$, the function f may have unbalanced coverage across different sub-populations in the data.

3.4 `Rolling CI` with Calibration Set

We turn to present `Rolling CI with cal` that is closer to conformal prediction and ACI. `Rolling CI with cal` uses calibration points, but does not hold out a large block like `ACI-Online`. Rather, previous points are simultaneously used both for calibration and model fitting.

Denote by $S(\mathcal{M}_t(X_t), Y_t) \in \mathbb{R}$ a non-conformity score function. Define the prediction set constructing function as:

$$f(X_t, \theta_t, \mathcal{M}_t) = \{y \in \mathcal{Y} : S(\mathcal{M}_t(X_t), y) \leq Q_{1+\theta_t}(\mathcal{S}_{\text{cal}})\}$$

where $\mathcal{S}_{\text{cal}} = \{S(M_{t'}(X_{t'}), Y_{t'}) : t' = t - n, \cdots, t - 1\}$. The function $Q_{1+\theta_t}(\mathcal{S}_{\text{cal}})$ returns the $(1 + \theta_t)$-th empirical quantile of the scores in $\mathcal{S}_{\text{cal}}$. The size of the set gets smaller as $1 + \theta_t$ gets smaller. Our proposed sequential evaluation of the calibration scores is immune to over-fitting by design.

Corollary 1. *`Rolling CI with cal` satisfies the coverage requirement in (1).*

4 Time-Series Conditional Coverage Metrics

We describe two metrics to assess conditional/local coverage validity in the online learning regime.

Miscoverage Streak Length (MSL): Given a series of intervals $\{\widehat{C}_t(X_t)\}_{t=T_0}^{T_1} \subseteq 2^{\mathcal{Y}}$ and response variables $\{Y_t\}_{t=T_0}^{T_1} \subseteq \mathcal{Y}$, the MSL metric is defined as

$$\text{MSL} := \frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \min\{i : Y_{t+i} \in \widehat{C}_{t+i}(X_{t+i}) \text{ or } t+i = T_1\}$$

where $\mathcal{I}$ is a set containing the starting times of all miscoverage streaks. An ideal model that has access to the true conditional quantiles attains an MSL of $\frac{1}{1-\alpha}$. Therefore, we seek to produce the narrowest intervals having an MSL close to this value.

ΔCoverage: The time-series data sets we use in the experiments in Section 5 include the day of the week as an element in the feature vector. Therefore, we assess the violation of day-stratified coverage, as a proxy for conditional coverage. That is, we measure the average deviation of the coverage in each day of the week from the nominal coverage level. Since a lower value of this metric indicates for a better conditional coverage, we desire to have a minimal ΔCoverage.

5 Experiments

5.1 Classic, Single-Output Quantile Regression

We analyze the effectiveness of our two proposed calibration schemes, `Rolling CI` and `Rolling CI with cal`, on five real-world benchmark data sets: power, energy, traffic, wind, and prices. We commence by fitting an initial quantile regression model on the first 5000 data points, to obtain a reasonable predictive system. Then, for step 5001 to 10000, we apply the calibration procedure while continuing to fit the model in an online fashion. Lastly, we measure the performance of the deployed calibration method on data points corresponding to time steps 8001 to 10000. In all experiments, we fit an LSTM predictive model in an online fashion, minimizing the pinball loss to estimate the 0.05 and 0.95 conditional quantiles of $Y_t \mid X_t$. For both `ACI-Online` and `Rolling CI with cal`, we use the CQR scores to calibrate the prediction intervals with a calibration set of size 300. As for `Rolling CI`, we calibrate the intervals in y scale with an exponential stretching. We repeat each experiment for 20 random initializations of the neural network model.

Figure summarizes the performance metrics, showing that all methods attain the desired coverage level. That figure also shows that our `Rolling CI` constructs the narrowest intervals while attaining the best conditional coverage metrics. Additionally, one can see that `RCI with cal` tends to construct intervals of better conditional coverage than `ACI-Online`, as indicated by the MSL and ΔCoverage metrics. This, however, comes at the cost of widening the intervals' length.

5.2 Multiple-Output Quantile Regression

We now introduce an application of our calibration scheme for which the response is a multivariate vector ($\mathcal{Y} = \mathbb{R}^d$) rather than a one-dimensional scalar. Our goal is to construct a predictive region. For a multivariate-response version of the power data, calibration-set-based methods, such as `ACI-Online`, cannot be deployed because of computation complexity. Therefore, our proposal has a fractional speed up of 12,500, being the only calibration scheme that can be applied in this multivariate-response setting.

We assess the coverage validity in local windows in time via the metric of local coverage frequencies proposed in Gibbs and Candés (2021). Figure shows that our technique leads to the desired coverage level,

guaranteed by Theorem 1. Importantly, the local coverage is close to the nominal level throughout the entire process, indicating for good conditional coverage.

6 Discussion

Table: A summary of all sequential calibration methods described in this paper.

Calibration method	Is online	Has a coverage guarantee	The model uses all data
ACI	X	O	X
OSSCP	O	X	X
ACI-Online	O	O	X
Rolling CI with cal	O	O	O
Rolling CI	O	O	O

While valid calibration is an important step toward making learning systems safer, fairer, and more robust, we emphasize that calibration is not a panacea and one must always treat data with care. We offer initial guidelines herein, acknowledging that of course, they do not anticipate the consequences in all possible use cases. Nonetheless, we highlight that uncertainty quantification is an important component to designing learning systems to have a positive impact in the real world, where the data are complex and distributions are continually shifting. Our work makes one step toward this goal.

References

Isaac Gibbs and Emmanuel Candés. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

Margaux Zaffran, Aymeric Dieuleveut, Olivier Féron, Yannig Goude, and Julie Josse. Adaptive conformal predictions for time series. *arXiv preprint*, 2022.

7 Question