Adaptive Conformal Inference Under Distribution Shift

Isaac Gibbs and Emmanuel J. Candès

Kiho Park

parkkiho@uchicago.edu

1 Introduction

As the popularity of black box methods and the cost of making wrong decisions grow it is crucial that we develop tools to quantify the uncertainty of their predictions. In this paper, we develop methods for constructing prediction sets that are guaranteed to contain the target label with high probability. We focus on an online learning setting. Perhaps the most powerful and flexible tools for solving this problem come from conformal inference. This paper develops *adaptive conformal inference* (ACI), a method for forming prediction sets that are robust to changes in the marginal distribution of the data.

1.1 Conformal inference

Suppose we are given a fitted regression model for predicting the value of Y from X. Let y be a candidate value for y_t. To determine if y is reasonable estimate of Y_t, we define a conformity score $S(X, Y)$. Assume we have a calibration set $\mathcal{D}_{\text{cal}} \subseteq \{(X_r, Y_r)\}_{1 \leq r \leq t-1}$. Using this calibration set we define the fitted quantiles of the conformity scores to be

$$\widehat{Q}(p) := \inf \left\{ s : \left(\frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{(X_r, Y_r) \in \mathcal{D}_{\text{cal}}} \mathbb{1}_{\{S(X_r, Y_r) \leq s\}} \right) \geq p \right\}$$

and define our prediction set to be $\widehat{C}_t := \{y : S(X_t, y) \leq \widehat{Q}(1-\alpha)\}$. Then, exchangeability gives the marginal coverage guarantee

$$\mathbb{P}(Y_t \in \widehat{C}_t) = \mathbb{P}(S(X_t, Y_t) \leq \widehat{Q}(1-\alpha)) = \frac{\lceil |\mathcal{D}_{\text{cal}}|(1-\alpha) \rceil}{|\mathcal{D}_{\text{cal}}| + 1}$$

We remark that the method outlined above is often referred to as split or inductive conformal inference. The adaptive conformal inference method developed in this article can also be easily adjusted to work with full conformal inference.

2 Adapting conformal inference to distribution shifts

In the general case where the distribution of the data is shifting over time, both the functions $S(\cdot)$ and $\widehat{Q}(\cdot)$ should be regularly re-estimated to align with the most recent observations. Therefore, we assume that at each time t we are given $S_t(\cdot)$ and $\widehat{Q}_t(\cdot)$. We define the realized miscoverage rate of the prediction set $\widehat{C}_t(\alpha) := \{y : S_t(X_t, y) \leq \widehat{Q}_t(1-\alpha)\}$ as

$$M_t(\alpha) := \mathbb{P}(S_t(X_t, Y_t) > \widehat{Q}_t(1-\alpha))$$

where the probability is over the test point (X_t, Y_t) as well as the data used to fit $S_t(\cdot)$ and $\widehat{Q}_t(\cdot)$.

Assume that with probability one, $\widehat{Q}_t(\cdot)$ is continuous, non-decreasing and $\widehat{Q}_t(0) = -\infty$ and $\widehat{Q}_t(1) = \infty$. Then, $M_t(\cdot)$ will be non-decreasing on $[0,1]$ with $M_t(0) = 0$ and $M_t(1) = 1$ and so we may define

$$\alpha_t^* := \sup\{\beta \in [0,1] : M_t(\beta) \leq \alpha\}$$

Moreover, if we additionally assume that

$$\mathbb{P}(S_t(X_t, Y_t) = \widehat{Q}_t(1 - \alpha_t^*)) = 0$$

then we will have that $M_t(\alpha_t^*) = \alpha$. So, we find that we can achieve either approximate or exact marginal coverage by correctly calibrating the argument to $\widehat{Q}_t(\cdot)$.

To perform this calibration we will use a simple online update. Let $\alpha_1 = \alpha$. Recursively define the sequence of miscoverage events

$$\text{err}_t := \begin{cases} 1 & \text{if } Y_t \notin \widehat{C}_t(\alpha_t) \\ 0 & \text{otherwise} \end{cases} \quad \text{where } \widehat{C}_t(\alpha_t) := \{y : S_t(X_t, y) \leq \widehat{Q}_t(1 - \alpha_t)\}$$

Then, fixing a step size parameter $\gamma > 0$ we consider the simple online update

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \text{err}_t)$$

We refer to this algorithm as adaptive conformal inference. A natural alternative to this is the update $\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \sum_{s=1}^t w_s \,\text{err}_s\right)$ where $\{w_s\}_{1 \leq s \leq t} \subseteq [0,1]$ is a sequence of increasing weights with $\sum_{s=1}^t w_s = 1$. In practice, we find that the two terms produce almost identical results.

2.1 Choosing the step size

The choice of γ gives a tradeoff between adaptability and stability. While raising the value of γ will make the method more adaptive to observed distribution shifts, it will also induce greater volatility in the value of α_t. In practice, large fluctuations in α_t may be undesirable as it allows the method to oscillate between outputting small conservative and large anti-conservative prediction sets. We give an upper bound on $(M_t(\alpha_t) - \alpha)^2$ that is optimized by choosing γ proportional to $\sqrt{|\alpha_{t+1}^* - \alpha_t^*|}$. In our experiments we will take $\gamma = 0.005$.

2.2 Real data example: predicting market volatility

We apply ACI to the prediction of market volatility. Let $\{P_t\}_{1 \leq t \leq T}$ denote a sequence of daily open prices for a stock. For all $t \geq 2$, define the return $R_t := (P_t - P_{t-1})/P_{t-1}$ and realized volatility $V_t = R_t^2$. Our goal is to use the previously observed returns $X_t := \{R_s\}_{1 \leq s \leq t-1}$ to form prediction sets for $Y_t := V_t$.

We start off by forming point predictions using a GARCH$(1,1)$ model. This method assumes that $R_t = \sigma_t \epsilon_t$ with $\epsilon_2, \cdots, \epsilon_T$ taken to be i.i.d. $\mathcal{N}(0,1)$ and σ_t satisfying the recursive update

$$\sigma_t^2 = \omega + \tau V_{t-1} + \beta \sigma_{t-1}^2$$

For all times $t > 1250$ we fit the coefficients $\widehat{\omega}_t, \widehat{\tau}_t, \widehat{\beta}_t$ as well as the sequence of variances $\widehat{\sigma}_s^t$ using only the data $\{R_r\}_{t-1250 \leq r < t}$. Then, our prediction for the realized volatility at time t is $(\widehat{\sigma}_t^t)^2 = \widehat{\omega}_t + \widehat{\tau}_t V_{t-1} + \widehat{\beta}_t (\widehat{\sigma}_{t-1}^t)^2$. To form prediction intervals we define the sequence of conformity scores

$$S_t := \frac{|V_t - (\widehat{\sigma}_t^t)^2|}{(\widehat{\sigma}_t^t)^2}$$

and the corresponding quantile function

$$\widehat{Q}_t(p) := \inf\left\{x : \frac{1}{1250}\sum_{r=t-1250}^{t-1} \mathbb{1}_{S_r \leq x} \geq p\right\}$$

Then, our prediction set at time t is

$$\widehat{C}_t(\alpha_t) := \left\{v : \frac{|v - (\widehat{\sigma}_t^t)^2|}{(\widehat{\sigma}_t^t)^2} \leq \widehat{Q}_t(1-\alpha_t)\right\}$$

where $\{\alpha_t\}$ is initialized with $\alpha_{1250} = \alpha = 0.1$ and then updated recursively.

We compare this algorithm to a non-adaptive alternative that takes $\alpha_t = \alpha$ fixed. To measure the performance of these methods across time we examine their local coverage frequencies defined as

$$\text{localCov}_t := 1 - \frac{1}{500}\sum_{r=t-250+1}^{t+250} \text{err}_t$$

If the methods perform well then we expect the local coverage frequency to stay near the target value $1-\alpha$ across all time points.

Daily open prices were obtained from publicly available datasets published by The Wall Street Journal. We see that the local coverage frequencies obtained by ACI always stay within the variation that would be expected from an i.i.d. Bernoulli sequence. On the other hand, the non-adaptive method undergoes large excursions away from the target level of $1-\alpha = 0.9$.

3 Related Work

Prior work on conformal inference has considered two different types of distribution shift. Tibshirani et al. (2019) showed that valid prediction sets can be obtained by re-weighting the calibration data using the likelihood ratio between P_1 and P_0. On the other hand, Cauchois et al. (2020) develop methods for forming prediction sets that are valid whenever P_1 and P_0 are close in f-divergence. Similar to our work, they show that if $D_f(P_1||P_0) \leq \rho$ then there exists a conservative value $\alpha_\rho \in (0,1)$ such that $M(\alpha_\rho) := \mathbb{P}(S(X_t,Y_t) > \widehat{Q}(1-\alpha_\rho)) \leq \alpha$.

4 Coverage guarantees

4.1 Distribution-free results

We outline the theoretical coverage guarantees of adaptive conformal inference. We will assume throughout that with probability one, $\alpha_1 \in [0,1]$ and $\widehat{Q}_t$ is non-decreasing with $\widehat{Q}_t(x) = -\infty$ for all $x < 0$ and $\widehat{Q}_t(x) = \infty$ for all $x > 1$. Our first result shows that over long time intervals ACI obtains the correct coverage frequency irrespective of any assumptions on the data-generating distribution.

Lemma 4.1. *With probability one we have that $\forall t \in \mathbb{N}$, $\alpha_t \in [-\gamma, 1+\gamma]$.*

Proposition 4.1. *With probability one we have that for all $T \in \mathbb{N}$,*

$$\left|\frac{1}{T}\sum_{t=1}^{T} \text{err}_t - \alpha\right| \leq \frac{\max\{\alpha_1, 1-\alpha_1\} + \gamma}{T\gamma}$$

In particular, $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \text{err}_t \overset{a.s.}{=} \alpha$.

Proposition 4.1 puts no constraints on the data generating distribution. We argue that without assumptions on the quality of the initialization, these results cannot be improved. So, we view Proposition 4.1 as both an agnostic guarantee and approximately tight bound on the worst-case behaviour immediately after initialization.

4.2 Performance in a hidden Markov model

We show that if α_1 is initialized appropriately and the distribution shift is small, then tighter coverage guarantees can be given. In order to obtain useful results, we will make some simplifying assumptions about the data generating process.

4.2.1 Setting

We model the data as coming from a hidden Markov model. Let $\{A_t\}_{t\in\mathbb{N}} \subseteq \mathcal{A}$ denote the underlying Markov chain for the environment. Assume that conditional on $\{A_t\}_{t\in\mathbb{N}}, \{(X_t, Y_t)\}_{t\in\mathbb{N}}$ is an independent sequence with $(X_t, Y_t) \sim P_{A_t}$ for some collection of distributions $\{P_a : a \in \mathcal{A}\}$. We assume additionally that $\widehat{Q}_t(\cdot)$ and $S_t(\cdot)$ do not depend on t and we denote them by $\widehat{Q}(\cdot)$ and $S(\cdot)$. In this setting, $\{(\alpha_t, A_t)\}_{t\in\mathbb{N}}$ forms a Markov chain on $[-\gamma, 1+\gamma] \times \mathcal{A}$. We assume that this chain has a unique stationary distribution π and that $(\alpha_1, A_1) \sim \pi$.

4.2.2 Large deviation bound for the errors

By Proposition 4.1 and since err_t is stationary, $\mathbb{E}[\mathrm{err}_t] = \alpha$. The following Theorem 4.1 gives a large deviation bound on $|T^{-1}\sum_{t=1}^T \mathrm{err}_t - \alpha|$. The idea behind this result is to decompose the dependence in $\{\mathrm{err}_t\}_{t\in\mathbb{N}}$ into two parts. First, there is dependence due to the fact that α_t is a function of $\{\mathrm{err}_r\}_{1\le r\le t-1}$. This gives rise to the first term by Hoeffding's inequality. Second, there is dependence due to the fact that A_t depends on A_{t-1}. This gives rise to the second term by Bernstein inequality.

Theorem 4.1. *Assume that $\{A_t\}_{t\in\mathbb{N}}$ has non-zero absolute spectral gap $1-\eta > 0$. Let*

$$B := \sup_{a\in\mathcal{A}} |\mathbb{E}[\mathrm{err}_t | A_t = a] - \alpha| \quad \textit{and} \quad \sigma_B^2 := \mathbb{E}[(\mathbb{E}[\mathrm{err}_t | A_t] - \alpha)^2]$$

Then,

$$\mathbb{P}\left(\left| \frac{1}{T}\sum_{t=1}^T \mathrm{err}_t - \alpha \right| \ge \epsilon \right) \le 2\exp\left(-\frac{T\epsilon^2}{8}\right) + 2\exp\left(-\frac{T(1-\eta)\epsilon^2}{8(1+\eta)\sigma_B^2 + 20B\epsilon}\right)$$

Assume $M(p|a) := \mathbb{P}(S(X_t, Y_t) > \widehat{Q}(1-p)|A_t = a)$ is continuous. Then, there exists an optimal value α_a^* such that $M(\alpha_a^*|a) = \alpha$. We shows that if $M(\cdot|a)$ admits a second order Taylor expansion, then

$$B \le C\left(\gamma + \gamma^{-1} \sup_{a\in\mathcal{A}} \sup_{k\in\mathbb{N}} \mathbb{E}\left[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}| \,\middle|\, A_{t+k} = a \right]\right) \quad \text{and} \quad \sigma_B^2 \le B^2$$

The term $\mathbb{E}\left[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}| \,\middle|\, A_{t+k} = a \right]$ can be seen as a quantitative measurement of the size of the distribution shift in terms of the change in the critical value α_a^*. Thus, we interpret these results as showing that if the distribution shift is small and $\forall a \in \mathcal{A}$, $\widehat{Q}(\cdot)$ gives reasonable coverage of the distribution of $S(X_t, Y_t)|A_t = a$, then $T^{-1}\sum_{t=1}^T \mathrm{err}_t$ will concentrate well around α.

4.2.3 Achieving approximate marginal coverage

Theorem 4.1 bounds the distance between the average miscoverage rate and the target level over long stretches of time. On the other hand, it proves no information about the marginal coverage frequency at a single time step. The following result shows that if the distribution shift is small, the realized marginal coverage rate $M(\alpha_t|A_t)$ will be close to α on average.

Theorem 4.2. *Assume that there exists a constant $L > 0$ such that for all $a \in \mathcal{A}$ and all $\alpha_1, \alpha_2 \in \mathbb{R}$,*

$$|M(\alpha_2|a) - M(\alpha_1|a)| \leq L|\alpha_2 - \alpha_1|$$

Assume additionally that for all $a \in \mathcal{A}$ there exists $\alpha^*_a \in (0,1)$ such that $M(\alpha^*_a|a) = \alpha$. Then,

$$\mathbb{E}[(M(\alpha_t|A_t) - \alpha)^2] \leq \frac{L(1+\gamma)}{\gamma}\mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}|] + \frac{L}{2}\gamma$$

We remark that the right-hand side is minimized by choosing $\gamma = (2\mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}|])^{1/2}$, which gives the inequality

$$\mathbb{E}[(M(\alpha_t|A_t) - \alpha)^2] \leq L(\sqrt{2}+1)\sqrt{\mathbb{E}[|\alpha^*_{A_{t+1}} - \alpha^*_{A_t}|]}$$

This result can be interpreted as bounding the average difference between the realized and target marginal coverage in terms of the size of the underlying distribution shift. Note that the choice of γ formalizes our intuition that γ should be chosen to be larger in domains with greater distribution shift, while not being so large as to cause α_t to be overly volatile.

5 Impact of $S_t(\cdot)$ on the performance

The performance of all conformal inference methods depends heavily on the design of the conformity score. The lengths of conformal prediction sets depend directly on the quality of the fitted regression model. Thus, one should re-fit the model at each time step using the most recent data to build the most accurate perdictions.

In addition to this, the choice of $S_t(\cdot)$ can also have a direct effect on the coverage properties of adaptive conformal inference. They depend on how close $S_t(X_t, Y_t)$ is to being stationary. For example, the unnormalized score $\widetilde{S}_t := |V_t - \widehat{\sigma}^2_t|$ leads to much worse coverage properties. This issue can be partially mitigated by chosing a larger value of γ that gives greater adaptivity to the algorithm.

6 Real data example: election night predictions

During the 2020 US presidential election The Washington Post used conformalized quantile regression (CQR) to produce county level predictions of the vote total on election night. Here we replicate the core elements of this method using both fixed and adaptive quantiles. Let $\{Y_t\}_{1\leq t\leq T}$ denote the number of votes cast for $T = 3000$ counties in the U.S. Let X_t denote a set of demographic covariates associated to the tth county. The distribution of (X_t, Y_t) can be viewed as drifting throughout election night.

We apply CQR to predict the county-level vote totals. Figure shows the realized local coverage frequency over the most recent 300 counties for the non-adaptive and adaptive conformal methods. We find that the non-adaptive method fails to maintain the desired 90% coverage level, incurring large troughs in its coverage frequency during time zone changes. On the other hand, the adaptive method maintains approximate 90% coverage across all time points with deviations in its local coverage level comparable to what is observed in Bernoulli sequences.

7 Discussion

There are still many open problems in this area. There are many settings in which we receive the response in a delayed fashion or in large batches. It would be interesting to determine if similar results can be obtained in settings where $\widehat{Q}_t(\cdot)$ is fit in an online fashion on the most recent data. Another potential area for improvement is in the choice of the step size γ. Additional work is needed to determine if our methods can successfully protect against a wider variety of real-world distribution shifts.

References

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: Confident predictions even when distributions shift. *arXiv preprint*, 2020.

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

8 Question