Conformal Prediction

Junyi Liao

Conformal prediction is a popular, modern technique for providing valid predictive inference for arbitrary machine learning models. It deals with a contemporary challenge: when working with a "black box" algorithm that constructs a predictive model from training data, how do we establish calibrated prediction intervals around the model's output, ensuring they reliably achieve a desired coverage level?

1 Adjusted Quantiles

1.1 General setting

Let $(X_i, Y_i) \sim P,\ i = 1, 2, \cdots, n$ be i.i.d. feature and response pairs from a distribution P on $\mathcal{X} \times \mathcal{Y}$. Let $\alpha \in (0, 1)$ be a small error level. We are possibly interested in finding a prediction band $\widehat{C}_n : \mathcal{X} \to \mathscr{Y}$, where $\mathscr{Y}$ is the class of measurable subsets of $\mathcal{Y}$. Moreover, for a new pair $(X_{n+1}, Y_{n+1}) \sim P$, we hope that our prediction band covers the true response with high probability:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n(X_{n+1})\right) \geq 1 - \alpha. \tag{1.1}$$

Intuitively, a narrower band yields less uncertainty in our prediction at a constant error level α.

We consider a simpler context where there are no features at all and $\mathcal{Y} = \mathbb{R}$. Let $q = \text{Quantile}(1 - \alpha; P)$,

$$\mathbb{P}\left(Y_{n+1} \in (-\infty, q]\right) = 1 - \alpha \tag{1.2}$$

gives a natural prediction band. Given $Y_1, \cdots, Y_n \sim P$, to approximate quantile q, we can use the empirical distribution $n^{-1} \sum_{i=1}^{n} \delta_{Y_i}$:

$$\widehat{q}_n = \text{Quantile}\left(1 - \alpha; \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}\right). \tag{1.3}$$

However, our prediction band $\widehat{C}_n = (-\infty, \widehat{q}_n]$ is not an exact confidence set, since (1.2) only holds asymptotically when $n \to \infty$ under some regular conditions when q is replaced by $\widehat{q}_n$. To address this problem, we introduce the adjusted quantile.

1.2 Adjusted quantiles

We first introduce a useful tool which allows us to generate random variables distributed according to arbitrary cumulative distribution function F from a uniform variable $U \sim \text{Unif}(0, 1)$.

Lemma 1.1 (Galois inequality). Let F be a cumulative distribution function (c.d.f.) and Z be an $\mathbb{R}$-valued random variable such that $\mathbb{P}(Z \leq z) = F(z), z \in \mathbb{R}$. Define the corresponding quantile function Q_F as follows:

$$Q_F(t) = \inf\{z : F(z) \geq t\}, \quad t \in [0, 1]. \tag{1.4}$$

Then for any $t \in [0,1]$ and $z \in \mathbb{R}$, we have

$$F(z) \geq t \quad \Leftrightarrow \quad Q_F(t) \leq z. \tag{1.5}$$

Moreover, if $U \sim \text{Unif}(0,1)$, then $Q_F(U) \sim F$.

Proof. By definition, $F(z) \geq t$ implies $z \geq Q_F(t)$. Now suppose $z \geq Q_F(t)$. By definition, $\forall \epsilon > 0$, we have $F(z+\epsilon) \geq t$. Since F as a c.d.f. is right-continuous, it holds

$$F(z) = \lim_{\epsilon \to 0^+} F(z+\epsilon) \geq t. \tag{1.6}$$

Then we conclude the proof of (1.5). Moreover, if $U \sim \text{Unif}(0,1)$, then for any $t \in [0,1]$,

$$\mathbb{P}(Q_F(U) \leq t) = \mathbb{P}(F(t) \geq U) = F(t). \tag{1.7}$$

Hence $Q_F(U)$ is distributed according to F. □

Corollary 1.2. Fix $t \in [0,1]$. By setting $z = Q_F(t)$ in (1.5), we have

$$\mathbb{P}(Z \leq Q_F(t)) = F(Q_F(t)) \geq t. \tag{1.8}$$

Namely, with probability at least t, Z is not greater than its t quantile.

Lemma 1.3 (Order statistics). Let F be a c.d.f. and $Y_1, \cdots, Y_{n+1}$ be i.i.d. random variables drawn from F. Let $Y_{(1)}, \cdots, Y_{(n)}$ be the order statistics of $Y_1, \cdots, Y_n$. Then

$$\mathbb{P}(Y_{n+1} \leq Y_{(k)}) \geq \frac{k}{n+1}, \quad k = 1, \cdots, n. \tag{1.9}$$

Proof. We first prove that Q_F is non-decreasing. Since F is non-decreasing, we have

$$t_1 \leq t_2 \;\Rightarrow\; \{z : F(z) \geq t_1\} \supseteq \{z : F(z) \geq t_2\} \;\Leftrightarrow\; \inf\{z : F(z) \geq t_1\} \leq \inf\{z : F(z) \geq t_2\}. \tag{1.10}$$

Now we show (1.9). Let $U_1, \cdots, U_{n+1} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$. By Lemma 1.1, we have the representation $Y_i = Q_F(U_i)$ for $i = 1, \cdots, n+1$. Since Q_F is non-decreasing, $Y_{(k)} = Q_F(U_{(k)})$ holds, where $U_{(1)}, \cdots, U_{(n)}$ are order statistics of $U_1, \cdots, U_n$. By (1.10), we have

$$\mathbb{P}(Y_{n+1} \leq Y_{(k)}) \geq \mathbb{P}(U_{n+1} \leq U_{(k)}) = \mathbb{E}\left[\mathbb{P}(U_{n+1} \leq U_{(k)} | U_1, \cdots, U_n)\right] = \mathbb{E}[U_{(k)}] = \frac{k}{n+1}. \tag{1.11}$$

Then we conclude the proof. □

Remark. If F is continuous on $\mathbb{R}$, then (1.9) becomes an equality: $\mathbb{P}(Y_{n+1} \leq Y_{(k)}) = \frac{k}{n+1}$.

Now we are prepared to introduce the adjusted quantile for empirical distributions.

Definition 1.4 (Adjusted quantile). Given i.i.d. samples $Y_1, \cdots, Y_n$, the adjusted quantile of their empirical distribution is defined as

$$\widehat{q}_n = \text{Quantile}\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}; \sum_{i=1}^{n} \delta_{Y_i}\right) = Y_{(\lceil (1-\alpha)(n+1) \rceil)}. \tag{1.12}$$

Using this definition, we can achieve (1.1) exactly by setting $\widehat{C}_n = (-\infty, \widehat{q}_n]$. Moreover, if $Y_1, \cdots, Y_{n+1}$ are drawn from a continuous distribution, we can bound the coverage rate of $\widehat{C}_n$ as follows:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n\right) = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \in \left[1-\alpha, 1-\alpha+\frac{1}{n+1}\right). \tag{1.13}$$

An alternative formulation. Parallel to Lemma 1.3, we can prove $\mathbb{P}(Y_{n+1} \geq Y_{(k)}) \geq 1 - \frac{k}{n+1},\ k = 1, \cdots, n$. Then we define

$$\widetilde{q}_n = \text{Quantile}\left(\frac{\lfloor \alpha(n+1) \rfloor}{n}; \sum_{i=1}^{n} \delta_{Y_i}\right) = Y_{(\lfloor \alpha(n+1) \rfloor)}, \tag{1.14}$$

and we can also achieve (1.1) exactly by setting $\widehat{C}_n = [\widetilde{q}_n, \infty)$. Similarly, the bound (1.13) holds in a continuous setting.

2 Split Conformal Prediction

Regression Setting. To address the case where both features $X_i \in \mathcal{X}$ and responses $Y_i \in \mathbb{R}$ are observed, we can find a point estimator $\widehat{f}_n : \mathcal{X} \to \mathbb{R}$ that predict the value of Y_i based on X_i. Then, define the residuals the on training set:

$$R_i = |Y_i - \widehat{f}_n(X_i)|,\ i = 1, \cdots, n. \tag{2.1}$$

Let $\widehat{q}_n$ be the adjusted $1-\alpha$ quantile of $R_1, \cdots, R_n$, we can immediately construct a prediction band:

$$\widehat{C}_n(x) = \{y : |y - \widehat{f}_n(x)| \leq \widehat{q}_n\} \quad \Rightarrow \quad \widehat{C}_n(x) = \left[\widehat{f}_n(x) - \widehat{q}_n, \widehat{f}_n(x) + \widehat{q}_n\right]. \tag{2.2}$$

However, this prediction band may undercover because $R_{n+1} = |Y_{n+1} - \widehat{f}_n(X_{n+1})|$ is not exchangeable with $R_1, \cdots, R_n$, since $\widehat{f}_n$ is only trained on $\{(X_i, Y_i)\}_{i=1}^n$. (Generally, R_{n+1} are greater than expected.)

2.1 Overcovering conformal sets

Symmetrization. From the above analysis, it is necessary to generate residuals satisfying the exchangeability condition if we want to use the adjust quantile method. In this context, the split conformal prediction will be helpful. The split conformal prediction divides the training set D into two disjoint subsets:

- proper training set $D_1 \subsetneq D$ with $|D_1| = n_1$, and
- calibration set $D_2 = D \backslash D_1$ with $|D_2| = n_2 = n - n_1$.

Then, fit a point estimator $\widehat{f}_{n_1}$ on the proper training set $\{(X_i, Y_i),\ i \in D_1\}$, and calculate the calibration residuals $\{R_j = |Y_j - \widehat{f}_{n_1}(X_j)|,\ j \in D_2\}$ and the conformal quantile:

$$\widehat{q}_{n_2} = \text{Quantile}\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n_2}; \frac{1}{n_2} \sum_{j \in D_2} \delta_{R_j}\right). \tag{2.3}$$

Use $\widehat{f}_{n_1}$ and $\widehat{q}_{n_2}$ to construct the conformal set

$$\widehat{C}_n = \left[\widehat{f}_{n_1}(x) - \widehat{q}_{n_2}, \widehat{f}_{n_1}(x) + \widehat{q}_{n_2}\right] \tag{2.4}$$

Conditioning on the proper training set $\{(X_i, Y_i), i \in D_1\}$, the calibration residuals $\{R_j, j \in D_2\}$ and the test residual $R_{n+1} = Y_{n+1} - \widehat{f}_{n_1}(X_{n+1})$ are i.i.d., hence our confidence set satisfies (1.1) exactly:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, (X_i, Y_i), i \in D_1\right) = \frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2+1} \in \left[1-\alpha, 1-\alpha+\frac{1}{n_2+1}\right). \tag{2.5}$$

Modification of residuals. Let $V(x, y)$ be a negatively-oriented score function that measures the conformity of point (x, y) (negatively-oriented meaning that a lower value indicates better conformity). For example, in the previous discussion, $V(x, y) := |y - \widehat{f}_{n_1}(x)|$.

Then we generalize the residuals by the conformity score:

$$\begin{cases} R_j := V(X_j, Y_j), \ j \in D_2, \\ R_{n+1} := V(X_{n+1}, Y_{n+1}). \end{cases} \tag{2.6}$$

And we can construct the conformal set:

$$\widehat{C}_n(x) = \left\{ y : V(x, y) \leq \text{Quantile}\left(\frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2}; \frac{1}{n_2} \sum_{j \in D_2} \delta_{R_j} \right) \right\}. \tag{2.7}$$

Relation to the prediction algorithm. It can be seen that the width of the prediction band is exactly the same at each point $x \in \mathcal{X}$. Any prediction algorithm (which fits or interpolates the proper training set D_1 by a point estimator) produces a conformal band with valid coverage, which protects the point estimator $\widehat{f}_{n_1}$ against overfitting. However, a good prediction algorithm often yields a smaller prediction sets, because the point estimator $\widehat{f}_{n_1}(X)$ falls in high density regions of our conditional distribution $P_{Y|X}$.

2.2 Auxiliary randomization*

The conformal set (2.7) can be rewritten as a c.d.f. form:

$$\widehat{C}_n(x) = \left\{ y : \frac{1}{n_2} \sum_{j \in D_2} \mathbb{1}_{\{R_j \leq V(x,y)\}} \leq \frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2} \right\}. \tag{2.8}$$

Denote by $\widehat{F}_{n_2+1}$ the empirical distribution of $\{R_j, j \in D_2\}$ and R_{n+1}. Then

$$Y_{n+1} \in \widehat{C}_n(X_{n+1}) \quad \Leftrightarrow \quad \widehat{F}_{n_2+1}(R_{n+1}) \leq \frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2+1}, \tag{2.9}$$

which occurs with probability at least $1 - \alpha$.

The following proposition is useful in our analysis.

Proposition 2.1. Suppose an $\mathbb{R}$-valued random variable Z is distributed according to c.d.f. F. Then variable $F(Z)$ is sub-uniform, i.e. $\mathbb{P}(F(Z) \leq t) \leq t, \ \forall t \in [0, 1]$. Furthermore, $\mathbb{P}(F(Z) \leq t) = t$ if and only if $t \in \overline{\{F(z) : z \in \mathbb{R}\}}$.

Proof. If $t \in \overline{\{F(z) : z \in \mathbb{R}\}}$, let $z^* = \sup\{z : F(z) \leq t\}$. Then either $F(z^*) = t$ or $\lim_{\epsilon \to 0^+} F(z^* - \epsilon) = t$, and

$$\mathbb{P}(F(Z) \leq t) = \begin{cases} \mathbb{P}(Z < z^*) = \lim_{\epsilon \to 0^+} F(z^* - \epsilon) = t, \ t \notin \{F(z) : z \in \mathbb{R}\}, \\ \mathbb{P}(Z \leq z^*) = F(z^*) = t, \ t \in \{F(z) : z \in \mathbb{R}\}. \end{cases} \tag{2.10}$$

If $t \notin \overline{\{F(z) : z \in \mathbb{R}\}}$, then $\exists \epsilon > 0$ such that

$$\mathbb{P}(F(Z) \le t) = \mathbb{P}(F(Z) \le t - \epsilon) = \mathbb{P}(F(Q_F(U)) \le t - \epsilon) \le \mathbb{P}(U \le t - \epsilon) = t - \epsilon, \tag{2.11}$$

where $U \sim \text{Unif}(0,1)$, and the inequality follows by Corollary 1.2. Therefore $\mathbb{P}(F(Z) \le t) \le t$, and $\mathbb{P}(F(Z) \le t) = t$ if and only if $t \in \overline{\{F(z) : z \in \mathbb{R}\}}$. □

To deal with possible discontinuity of a c.d.f. F, we make a modification

$$F^*(z;u) = uF(z) + (1-u) \lim_{\epsilon \to 0^+} F(z - \epsilon), \quad u \in [0,1]. \tag{2.12}$$

Then if F is discontinuous at z, we can connect $F(z)$ and $\lim_{\epsilon \to z^+} F(z-\epsilon)$ by sliding u in F^* from 0 to 1.

Proposition 2.2. Suppose $Z \sim F$ and $U \sim \text{Unif}(0,1)$. Then $\mathbb{P}(F^*(Z;U) \le t) = t, \ \forall t \in [0,1]$.

Proof. Fix $t \in (0,1)$, and let $z^* = \sup\{z : F(z) \le t\}$. The case of $z^* \in \{z : F(z) \le t\}$ is easy. We show the case of $z^* \notin \{z : F(z) \le t\}$. By Proposition 2.1, we know that $F(z^*) > t$, and $F^-(z^*) := \lim_{\epsilon \to 0^+} F(z^* - \epsilon) \le t$. Note that

$$t = \frac{t - F^-(z^*)}{F(z^*) - F^-(z^*)} F(z^*) + \frac{F(z^*) - t}{F(z^*) - F^-(z^*)} F^-(z^*), \tag{2.13}$$

we have

$$\begin{aligned} \mathbb{P}(F^*(Z;U) \le t) &= \mathbb{P}\left(\{Z < z^*\} \cup \left\{ Z = z^*, U \le \frac{t - F^-(z^*)}{F(z^*) - F^-(z^*)} \right\} \right) \\ &= F^-(z^*) + \big(F(z^*) - F^-(z^*)\big) \frac{t - F^-(z^*)}{F(z^*) - F^-(z^*)} = t. \end{aligned} \tag{2.14}$$

Then we conclude the proof. □

Back to our discussion of conformal sets. Note that

$$\widehat{F}_{n_2+1}(R_{n+1}) = \frac{1}{n_2+1} \left(\sum_{j \in D_2} \mathbb{1}_{\{R_j \le R_{n+1}\}} + 1 \right), \quad \lim_{\epsilon \to 0^+} \widehat{F}_{n_2+1}(R_{n+1} - \epsilon) = \frac{1}{n_2+1} \sum_{j \in D_2} \mathbb{1}_{\{R_j < R_{n+1}\}}, \tag{2.15}$$

we can calculate the modified empirical distribution function $\widehat{F}^*_{n_2+1}$:

$$\widehat{F}^*_{n_2+1}(R_{n+1};u) = \frac{1}{n_2+1} \sum_{j \in D_2} \mathbb{1}_{\{R_j < R_{n+1}\}} + \frac{u}{n_2+1} \left(\sum_{j \in D_2} \mathbb{1}_{\{R_j = R_{n+1}\}} + 1 \right). \tag{2.16}$$

By defining the randomized confidence set

$$\widehat{C}^*_n(x;U) = \left\{ y : \frac{1}{n_2+1} \sum_{j \in D_2} \mathbb{1}_{\{R_j < V(x,y)\}} + \frac{U}{n_2+1} \left(\sum_{j \in D_2} \mathbb{1}_{\{R_j = V(x,y)\}} + 1 \right) \le 1 - \alpha \right\}, \tag{2.17}$$

we have

$$\widehat{F}^*_{n_2+1}(R_{n+1};U) \le 1 - \alpha \quad \Leftrightarrow \quad Y_{n+1} \in \widehat{C}^*_n(X_{n+1};U). \tag{2.18}$$

Applying Proposition 2.2:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n^*(X_{n+1}; U) \,\middle|\, (X_i, Y_i), i \in D_1\right) = \mathbb{P}\left(\widehat{F}_{n_2+1}^*(R_{n+1}; U) \leq 1-\alpha \,\middle|\, (X_i, Y_i), i \in D_1\right) = 1-\alpha. \quad (2.19)$$

It can be seen that our auxiliary randomization achieves an exact coverage in our prediction sets.

2.3 Conditional coverage

From (2.5), we can see that split conformal prediction comes with the strong, distribution-free coverage guarantee. By marginalizing over the proper training set, we get the unconditional coverage property:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n(X_{n+1})\right) \in \left[1-\alpha, 1-\alpha+\frac{1}{n_2+1}\right). \quad (2.20)$$

2.3.1 Conditioning on entire training set

Lemma 2.3. Let $X_1, \cdots, X_{n+1} \overset{\text{i.i.d.}}{\sim} F$, where F is continuous on $\mathbb{R}$. Let $X_{(1)}, \cdots, X_{(n)}$ be the order statistics of $X_1, \cdots, X_n$. Then

$$\mathbb{P}(X_{n+1} \leq X_{(k)} | X_1, \cdots, X_n) = \text{Beta}(k, n+1-k). \quad (2.21)$$

Proof. Let $U_1, \cdots, U_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$, and $X_i = Q_F(U_i)$. Let $U_{(i)}, i = 1, \cdots, n$ be the order statistics of $U_1, \cdots, U_n$. Since F is non-decreasing, we have $X_{(i)} = Q(U_{(i)}), i = 1, \cdots, n$. Moreover, the continuity of F implies $F(Q_F(u)) = u,\ \forall u \in [0,1]$, because $Q_F(u) = \inf\{z : F(z) \geq u\} \in \{z : F(z) = u\}$. Then

$$\begin{aligned}\mathbb{P}(X_{n+1} \leq X_{(k)} | X_1, \cdots, X_n) &= \mathbb{P}(X_{n+1} \leq Q_F(U_{(k)}) | U_1, \cdots, U_n) \\ &= F(Q_F(U_{(k)})) = U_{(k)} \sim \text{Beta}(k, n-k+1), \end{aligned} \quad (2.22)$$

which concludes the proof. □

The following proposition is an immediate corollary of Lemma 2.3.

Proposition 2.4. Consider the form of conformal set given in (2.7). Provided the residuals are almost surely distinct, we have

$$\mathbb{P}\left(Y_n \in \widehat{C}_n(X_{n+1}) \,\middle|\, (X_i, Y_i), i = 1, \cdots, n\right) \sim \text{Beta}\left(k_\alpha, n_2 + 1 - k_\alpha\right), \quad (2.23)$$

where $k_\alpha = \lceil (1-\alpha)(n_2+1) \rceil$.

Remark. This distribution has mean $\frac{k_\alpha}{n_2+1} = \frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2+1}$, which is the same as the marginal version. Moreover, the variance $\frac{k_\alpha(n_2+1-k_\alpha)}{(n_2+1)^2(n_2+2)} \approx \frac{\alpha(1-\alpha)}{n_2+2}$ decreases as the calibration set D_2 expands.

2.3.2 X-conditional coverage

Our prediction band has the same width and coverage at each test location $x \in \mathcal{X}$. Therefore the coverage conditional on X_{n+1} is obtained easily:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n(x) \,\middle|\, (X_i, Y_i), i \in D_1,\ X_{n+1} = x\right) = \frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2+1} \geq 1-\alpha, \quad \forall x \in \mathcal{X}. \quad (2.24)$$

3 Full Conformal Prediction

Another idea of generating exchangeable residuals is to include the test data in the regression stage. We do this in a subtle approach: fix any test location $x \in \mathcal{X}$, we evaluate how possibly a given response value $y \in \mathbb{R}$ falls in our prediction band $\widehat{C}_n(x)$. The value y is called a trial or query. The procedure of full conformal prediction is summarized below:

- Fit a point estimator $\widehat{f}_{n,(x,y)}$ on an augmented training set: $(X_1, Y_1), \cdots, (X_n, Y_n), (x, y)$;
- Define residuals:

$$\begin{cases} R_i^{(x,y)} &= |Y_i - \widehat{f}_{n,(x,y)}(X_i)|, \quad i = 1, \cdots, n, \\ R_{n+1}^{(x,y)} &= |y - \widehat{f}_{n,(x,y)}(x)|. \end{cases} \tag{3.1}$$

- Define the conformal set:

$$\widehat{C}_n(x) = \left\{ y : R_{n+1}^{(x,y)} \leq \text{Quantile}\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}; \frac{1}{n} \sum_{i=1}^{n} \delta_{R_i^{(x,y)}} \right) \right\}. \tag{3.2}$$

By plugging in $(x, y) = (X_{n+1}, Y_{n+1})$ to equation (3.1), we can produce residuals $\{R_i := R_i^{(X_{n+1}, Y_{n+1})}\}$ that are exchangeable if our prediction algorithm treats all training point indiscriminately. Hence

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n(X_{n+1}) \right) \geq 1 - \alpha, \tag{3.3}$$

and an upper bound $1 - \alpha + \frac{1}{n+1}$ holds if all residuals are almost surely distinct. This approach has adaptivity to the test location, meaning that the band width is not a constant over the whole region. Meanwhile, it is far more computationally expensive than split conformal prediction. Theoretically we need to fit a point estimator at each location $y \in \mathcal{Y} \subseteq \mathbb{R}$, which is intractable if $\mathcal{Y}$ is continuous. Practically, we do this over a finite grid of y values, which still requires large computation cost.

All of the extensions mentioned in the split conformal section carry over to full conformal prediction.

Modification of residuals. Any symmetric negatively-oriented score function can replace the absolute residual score:

$$\begin{cases} R_i^{(x,y)} &= V\big((X_i, Y_i); (X_1, Y_1), \cdots, (X_n, Y_n), (x, y)\big), \quad i = 1, \cdots, n, \\ R_{n+1}^{(x,y)} &= V\big((x, y); (X_1, Y_1), \cdots, (X_n, Y_n), (x, y)\big). \end{cases} \tag{3.4}$$

Here V is symmetric in its last $n + 1$ arguments, which ensures the exchangeability of residuals.

Auxiliary randomization. Inject auxiliary randomness in the conformal set:

$$\widehat{C}_n^*(x) = \left\{ y : \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}_{\left\{ R_i^{(x,y)} < R_{n+1}^{(x,y)} \right\}} + \frac{U}{n+1} \left(\sum_{i=1}^{n} \mathbb{1}_{\left\{ R_i^{(x,y)} = R_{n+1}^{(x,y)} \right\}} + 1 \right) \leq 1 - \alpha \right\}. \tag{3.5}$$

This conformal set possesses an exact coverage of $1 - \alpha$.

Connection to hypothesis testing. The conformal set in (3.2) can be rewritten as

$$\begin{aligned}\widehat{C}_n(x) &= \left\{ y : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{R_{n+1}^{(x,y)} \geq R_i^{(x,y)}\right\}} \leq \frac{\lceil (1-\alpha)(n+1) \rceil}{n} \right\} \\ &= \left\{ y : \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{R_{n+1}^{(x,y)} < R_i^{(x,y)}\right\}}}_{p\text{-value for } H_0: Y_{n+1}=y} \geq \frac{\lfloor \alpha(n+1) - 1 \rfloor}{n} \right\}. \end{aligned} \tag{3.6}$$

Hence it is equivalent to a hypothesis testing where the null hypothesis is $H_0 : Y_{n+1} = y$, and we compare the p-value to an adjusted significance level $\frac{\lfloor \alpha(n+1)-1 \rfloor}{n}$.

3.1 Impossibility of X-conditional coverage

In this subsection, we investigate the X-conditional coverage property of distribution-free conformal sets over the feature space $\mathcal{X}$. We first introduce the following tensorization inequality of total variation.

Lemma 3.1 (Le Cam). Let P, Q be two probability measures on $\mathcal{X} \subseteq \mathbb{R}^d$, and $P^{\otimes n}, Q^{\otimes n}$ the corresponding product measures on $\mathcal{X}^n$. Fix $\epsilon > 0$. If $d_{\mathrm{TV}}(P, Q) \leq \epsilon_n := 1 - (1 - \epsilon^2/2)^{1/n}$, then we have $d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n}) \leq \epsilon$. Here d_{TV} stands for the total variation between two probability measures.

Proof. Recall the definition of squared Hellinger distance: $H^2(P, Q) = 2 - 2\mathbb{E}_Q \left[\sqrt{\frac{\mathrm{d}P}{\mathrm{d}Q}} \right]$, we have the tensorization property of Hellinger distance:

$$H^2 \left(\prod_{i=1}^{n} P_i, \prod_{i=1}^{n} Q_i \right) = 2 - 2 \prod_{i=1}^{n} \left[1 - \frac{H^2(P_i, Q_i)}{2} \right]. \tag{3.7}$$

Use the sandwich bound for total variation:

$$\frac{1}{2} H^2 \leq d_{\mathrm{TV}} \leq H \sqrt{1 - \frac{H^2}{4}} \leq H, \tag{3.8}$$

we have

$$d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n}) \leq \sqrt{H^2(P^{\otimes n}, Q^{\otimes n})} = \sqrt{2 - 2\left(1 - \frac{H^2(P, Q)}{2}\right)^n} \leq \sqrt{2 - 2(1 - d_{\mathrm{TV}}(P, Q))^n}. \tag{3.9}$$

Hence $d_{\mathrm{TV}}(P, Q) \leq \epsilon_n = 1 - \left(1 - \frac{\epsilon^2}{2}\right)^{1/n}$ implies $d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n}) \leq \epsilon$. □

The following theorem is drawn from Lei and Wasserman (2014), which reveals the impossibility of construct uniform X-conditional coverage in a distribution-free setting.

Theorem 3.2 (Impossibility of finite sample conditional validity). Suppose that $\widehat{C}_n$ is a prediction band produced by i.i.d. $(X_i, Y_i) \sim P,\ i = 1, \cdots, n$, and $\widehat{C}_n$ satisfies

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n(x) \,\middle|\, X_{n+1} = x \right) \geq 1 - \alpha \tag{3.10}$$

for any distribution P and P_X-almost every x, where P_X is the marginal of X. Then for any P and any $x_0 \in \mathcal{N}(P_X) := \{x \in \mathcal{X} : \lim_{\delta \to 0} P_X(B(x_0, \delta)) = 0\}$, we have

$$\mathbb{P}\left(\lim_{\delta \to 0} \operatorname*{ess\,sup}_{x \in B(x_0, \delta)} \mu\left\{ \widehat{C}_n(x) \right\} = \infty \right) = 1, \tag{3.11}$$

where μ is the Lebesgue measure, and $B(x_0, \delta) := \{x \in \mathcal{X} : \|x - x_0\| \leq \delta\}$ is the closed ball centered at x_0 of radius δ.

Proof. Fix $\epsilon > 0$, and let $\epsilon_n = 1 - (1 - \epsilon^2/2)^{1/n}$. Let x_0 be a non-atom on P_X and choose $\delta_n > 0$ such that $P_X(B(x_0, \delta_n)) < \epsilon_n$. Fix $K > 0$ and let $K_0 = \frac{K}{2(1-\alpha).}$ Given P, define another probability measure

$$Q(A) = P(A \cap S^c) + U(A \cap S),\ S = \{(x, y) : x \in B(x_0, \delta_n), y \in \mathbb{R}\}, \tag{3.12}$$

and U has total mass $P(S)$ and is uniform on $\{(x, y) : x \in B(x_0, \delta), |y| < K_0\}$. Then we have

$$d_{\mathrm{TV}}(P, Q) = \sup_A\{P(A) - Q(A)\} = \sup_A\{P(A \cap S) - U(A \cap S)\} \leq P(S) \leq \epsilon_n. \tag{3.13}$$

By Lemma 3.1, we have $d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n}) \leq \epsilon$. Moreover, for all $x \in B(x_0, \delta_n)$, note that

$$1 - \alpha \leq \int_{\widehat{C}_n(x)} \mathrm{d}Q_{Y|X}(y|x) = \int_{\widehat{C}_n(x)} \mathrm{d}U_{Y|X}(y|x) \leq \frac{\mu\{\widehat{C}_n(x)\}}{2K_0}, \tag{3.14}$$

then $\mu\left\{\widehat{C}_n(x)\right\} \geq 2(1 - \alpha)K_0 = K$. Therefore

$$Q^{\otimes n}\left\{\operatorname*{ess\,sup}_{x \in B(x_0,\delta)} \mu\left\{\widehat{C}_n(x)\right\} \geq K\right\} = 1, \tag{3.15}$$

and

$$P^{\otimes n}\left\{\operatorname*{ess\,sup}_{x \in B(x_0,\delta)} \mu\left\{\widehat{C}_n(x)\right\} \geq K\right\} \geq Q^{\otimes n}\left\{\operatorname*{ess\,sup}_{x \in B(x_0,\delta)} \mu\left\{\widehat{C}_n(x)\right\} \geq K\right\} - d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n}) \geq 1 - \epsilon. \tag{3.16}$$

Then (3.11) follows as a result of $\epsilon \to 0$ and $K \to \infty$. □

We can interpret the result as follows. In an arbitrarily small neighborhood $B(x_0, \delta)$ of a non-atom point $x_0 \in \mathcal{X}$, any prediction band, claiming to cover the response at almost every point in $B(x_0, \delta)$, for every joint distribution P, is infinite in size.

4 Conformal Classification

4.1 Likelihood scores

Consider a classification problems where the response Y is drawn from a label set $\mathcal{Y} = \{1, \cdots, K\}$. Similar to the idea of split conformal prediction discussed in section 2, we first train a probabilistic classifier $\widehat{f}_{n_1} = \{\widehat{f}_{n_1}(\cdot; k), k = 1, \cdots, K\}$ over the proper training set $\{(X_i, Y_i), i \in D_1\}$. To be specific, $\widehat{f}_{n_1}(x; k)$ predicts $\mathbb{P}(Y = k \,|\, X = x)$ for each $k = 1, \cdots, K$. Then, we can calculate likelihood scores on the calibration set:

$$\{R_j = \widehat{f}_{n_1}(X_j; Y_j),\ j \in D_2\}. \tag{4.1}$$

Note that this is an example of positively-oriented score, which indicates the probability assigned to the correct class. Then we can construct the conformal set as follows:

$$\widehat{C}_n(x) = \left\{k \in [K] : \widehat{f}_{n_1}(x; k) \geq \text{Quantile}\left(\frac{\lfloor \alpha(n_2 + 1) \rfloor}{n_2}; \frac{1}{n_2} \sum_{j \in D_2} \delta_{R_j}\right)\right\}. \tag{4.2}$$

4.2 Adaptive prediction sets

To make the conformal prediction sets more adaptive, Romano et al. (2020) propose a conformity score based on cumulative likelihood. For each $j \in D_2$, let π_j be the permutation of $1, \cdots, K$ that sorts the predicted probabilities in decreasing order:

$$\widehat{f}_{n_1}(X_j; \pi_j(1)) \geq \widehat{f}_{n_1}(X_j; \pi_j(2)) \geq \cdots \geq \widehat{f}_{n_1}(X_j; \pi_j(K)). \tag{4.3}$$

Then the cumulative likelihood is

$$R_i = \sum_{i=1}^{k_j} \widehat{f}_{n_1}(X_j; \pi_j(i)), \quad \text{where } \pi_j(k_j) = Y_j,\ j \in D_2, \tag{4.4}$$

which is a negatively-oriented score. Then the confidence set is defined as

$$\widehat{q}_{n_2} = \text{Quantile}\left(\frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2}; \frac{1}{n_2} \sum_{j \in D_2} \delta_{R_j} \right), \tag{4.5}$$

$$\widehat{C}_n(x) = \{\pi_x(1), \cdots, \pi_x(k_x)\}, \quad \text{where } k_x = \min\left\{ k : \sum_{j=1}^{k} \widehat{f}_{n_1}(x; \pi_x(j)) \leq \widehat{q}_{n_2} \right\}. \tag{4.6}$$

5 Likelihood-weighted conformal prediction

Motivation: covariate shift. In many instances, the covariates in test set is not identically distributed as in training set. Consider the following setting of covariate shift:

$$\begin{cases} (X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P = P_{Y|X} P_X, \quad i = 1, \cdots, n, \\ (X_{n+1}, Y_{n+1}) \sim \widetilde{P} = P_{Y|X} \widetilde{P}_X, \text{ independent of } \{(X_i, Y_i)\}_{i=1}^n. \end{cases} \tag{5.1}$$

In general, the distribution shift impacts the exchangeability among residuals and the effectiveness of usual conformal prediction.

Review: rank-based quantiles. In Lemma 1.3, we have shown that for exchangeable variables $R_1, \cdots, R_{n+1}$,

$$\mathbb{P}\left(R_{n+1} \leq \text{Quantile}\left(\frac{k}{n}; \frac{1}{n} \sum_{i=1}^{n} \delta_{R_i} \right) \right) \geq \frac{k}{n+1}. \tag{5.2}$$

Since R_{n+1} is never strictly greater than itself, that R_{n+1} is greater than the k smallest of $R_1, \cdots, R_n$ is equivalent to that R_{n+1} is greater than the k smallest of $R_1, \cdots, R_{n+1}$. Hence (5.2) can be rewritten as

$$\mathbb{P}\left(R_{n+1} \leq \text{Quantile}\left(\frac{k}{n+1}; \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{R_i} \right) \right) \geq \frac{k}{n+1}. \tag{5.3}$$

Let $k_\alpha = \lceil (1-\alpha)(n+1) \rceil$. Recall the definition of quantile function $Q_F(t) = \inf\{z : F(z) \geq t\},\ 0 \leq t \leq 1$, we know that the quantile of empirical distribution $\frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{R_i}$ only changes in increments of $1/(n+1)$. Hence, its $1-\alpha$ quantile is equivalent to its $k_\alpha/(n+1)$ quantile, and

$$\mathbb{P}\left(R_{n+1} \leq \text{Quantile}\left(1-\alpha; \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{R_i} \right) \right) \geq \frac{k_\alpha}{n+1} \geq 1-\alpha. \tag{5.4}$$

Finally, let's consider a discrete distribution F supported on m points $s_1, \cdots, s_n \in \mathbb{R}$. Fix $t \in [0,1]$, and denote $q_t = Q_F(t) = \inf\{z : F(z) \geq t\}$. If we reassign the points $s_i > q_t$ to arbitrary values strictly greater than q_t, yielding a new distribution F', then it still holds $q_t = Q_{F'}(t)$. Using this fact, we can rewrite (5.4) as

$$\mathbb{P}\left(R_{n+1} \leq \text{Quantile}\left(1-\alpha; \frac{1}{n+1}\sum_{i=1}^{n}\delta_{R_i} + \frac{\delta_\infty}{n+1}\right)\right) \geq 1-\alpha. \tag{5.5}$$

5.1 Weighted exchangeability

We first introduce a generalization of exchangeability for random variables.

Definition 5.1 (Weighted exchangeability). A group of random variables $R_1, \cdots, R_{n+1}$ are said to be weighted exchangeable with respect to weight functions $w_1, \cdots, w_{n+1}$, if their joint density (more generally, Radon-Nikodym derivative with respect to an arbitrary base measure) admits the following representation:

$$f(r_1, \cdots, r_{n+1}) = \prod_{i=1}^{n+1} w_i(r_i) \cdot g(r_1, \cdots, r_{n+1}), \tag{5.6}$$

where g is a permutation invariant function, i.e. $g(r_1, \cdots, r_n) = g(r_{\sigma(1)}, \cdots, r_{\sigma(n+1)})$ for all permutation σ.

Then we have a weighted version of (5.5), which is stated as follows.

Lemma 5.2 (Quantile lemma). Let $\{Z_i, i = 1, \cdots, n+1\}$ be n exchangeable random variables with respect to weight functions $w_1, \cdots, w_{n+1}$. Let the scores be

$$R_i = V(Z_i; Z_1, \cdots, Z_n), \quad i = 1, \cdots, n+1, \tag{5.7}$$

where V is any score function symmetric in its last $n+1$ arguments. Define

$$p_i^w(z_1, \cdots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i}\prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}{\sum_\sigma \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}, \quad i = 1, \cdots, n+1, \tag{5.8}$$

where the sum is over permutations σ of numbers $1, \cdots, n+1$. Then for all $\alpha \in (0,1)$,

$$\mathbb{P}\left\{R_{n+1} \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w(Z_1, \cdots, Z_{n+1})\delta_{R_i} + p_{n+1}^w(Z_1, \cdots, Z_{n+1})\delta_\infty\right)\right\} \geq 1-\alpha \tag{5.9}$$

Proof. We fix $z_1, \cdots, z_{n+1}$ and denote by $E(z_1, \cdots, z_{n+1})$ the event that $\{Z_1, \cdots, Z_{n+1}\} = \{z_1, \cdots, z_{n+1}\}$. Let $r_i = V(z_i; z_1, \cdots, z_{n+1})$, $i = 1, \cdots, n+1$, and denote $\mathcal{S}(i) = \{j \in [n+1] : r_i = V(z_j; z_1, \cdots, z_{n+1})\}$ (which is introduced to deal with possible ties in $r_1, \cdots, r_{n+1}$). Using the joint density of $Z_1, \cdots, Z_{n+1}$ given in (5.4), for each i, it holds

$$\begin{aligned}
\mathbb{P}(R_{n+1} = r_i | E(z_1, \cdots, z_{n+1})) &= \mathbb{P}(Z_{n+1} \in \{z_j : j \in \mathcal{S}(i)\} | E(z_1, \cdots, z_{n+1})) \\
&= \frac{\sum_{\sigma:\sigma(n+1)\in\mathcal{S}(i)}\prod_{j=1}^{n} w_j(z_{\sigma(j)})g(z_{\sigma(1)}, \cdots, z_{\sigma(n+1)})}{\sum_\sigma \prod_{j=1}^{n} w_j(z_{\sigma(j)})g(z_{\sigma(1)}, \cdots, z_{\sigma(n+1)})} \\
&= \frac{\sum_{\sigma:\sigma(n+1)\in\mathcal{S}(i)}\prod_{j=1}^{n} w_j(z_{\sigma(j)})g(z_1, \cdots, z_{n+1})}{\sum_\sigma \prod_{j=1}^{n} w_j(z_{\sigma(j)})g(z_1, \cdots, z_{n+1})} \\
&= \sum_{k\in\mathcal{S}(i)} p_k^w(z_1, \cdots, z_{n+1}).
\end{aligned} \tag{5.10}$$

The the second equality follows from permutation invariance of g. Then we have

$$R_{n+1} \,|\, E(z_1, \cdots, z_{n+1}) \sim \sum_{i=1}^{n+1} p_i^w(z_1, \cdots, z_{n+1})\delta_{r_i}. \tag{5.11}$$

By Corollary 1.2, we have

$$\mathbb{P}\left\{ R_{n+1} \le \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n+1} p_i^w(z_1, \cdots, z_{n+1})\delta_{r_i}\right) \,\middle|\, E(z_1, \cdots, z_n)\right\} \ge 1-\alpha. \tag{5.12}$$

We can then replace each r_i with R_i in (5.12), and marginalize on E. Akin to the discussion above (5.5), we can change the point mass at R_{n+1} to one at ∞ and derive a form of (5.9), which concludes the proof. □

Following Lemma 5.2, we can design a weighted version of conformal prediction.

Theorem 5.3 (Weighted conformal prediction). Assume that $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \cdots, n+1$ are weighted exchangeable with respect to weight functions $w_1, \cdots, w_{n+1}$. Let V be an arbitrary score function that is symmetric in its last $n+1$ arguments. Define scores

$$\begin{cases} R_i^{(x,y)} = V\big((X_i, Y_i); Z_1, \cdots, Z_n, (x,y)\big), \quad i = 1, \cdots, n, \\ R_{n+1}^{(x,y)} = V\big((x,y); Z_1, \cdots, Z_n, (x,y)\big). \end{cases} \tag{5.13}$$

and a conformal set

$$\widehat{C}_n^w(x) = \left\{ y : R_{n+1}^{(x,y)} \le \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w\big(Z_{1:n}, (x,y)\big)\delta_{R_i^{(x,y)}} + p_{n+1}^w\big(Z_{1:n}, (x,y)\big)\delta_\infty\right)\right\}, \tag{5.14}$$

where $\{p_i^w, i = 1, \cdots, n+1\}$ are defined in (5.8). Then $\widehat{C}_n^w$ satisfies

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n^w(X_{n+1})\right) \ge 1-\alpha. \tag{5.15}$$

Proof. Abbreviate $R_i = R_i^{(X_{n+1}, Y_{n+1})}, i = 1, \cdots, n+1$. By construction of $\widehat{C}_n^w$ in (5.14), our conclusion (5.15) follows right from Lemma 5.2. □

Split version. The split conformal version of the above result can be viewed as a special case where the score function relies on a point predictor that has been fit on an external dataset. For example, if we take it to be $V(x,y) = |y - \widehat{\mu}_0(x)|$, where $\widehat{\mu}_0$ has been pre-trained on a data set $\mathbf{Z}_0$, then (5.14) simplifies to

$$\widehat{C}_n^w(x) = \widehat{\mu}_0(x) \pm \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w\big(Z_{1:n}, (x,y)\big)\delta_{|Y_i - \widehat{\mu}_0(X_i)|} + p_{n+1}^w\big(Z_{1:n}, (x,y)\big)\delta_\infty\right), \tag{5.16}$$

which has coverage at least $1-\alpha$, conditional on $\mathbf{Z}_0$.

CDF form. The set (5.14) can be rewritten as

$$\widehat{C}_n^w(x) = \left\{ y : \sum_{i=1}^{n} p_i^w\big(Z_{1:n}, (x,y)\big)\mathbb{1}_{\left\{R_i^{(x,y)} \le R_{n+1}^{(x,y)}\right\}} \le \lceil 1-\alpha \rceil_w \right\},$$

where $\lceil 1-\alpha \rceil_w = \min\left\{\tau \in \text{Range}(\widehat{F}_n^w) : \tau \ge 1-\alpha\right\}$, and $\widehat{F}_n^w$ is the c.d.f. of the weighted empirical distribution $\sum_{i=1}^n p_i^w\big(Z_1, \cdots, Z_n, (x,y)\big)\delta_{R_i^{(x,y)}} + p_{n+1}^w\big(Z_1, \cdots, Z_n, (x,y)\big)\delta_\infty$.

Auxiliary randomization. Parallel to our discussion in secntion 2.2, we can construct a randomized conformal set which as exact $1-\alpha$ coverage:

$$\widehat{C}_n^{w,*}(x) = \left\{ y : \sum_{i=1}^{n} p_i^w\big(Z_{1:n},(x,y)\big)\mathbb{1}_{\left\{R_i^{(x,y)} < R_{n+1}^{(x,y)}\right\}} + U\sum_{i=1}^{n+1} p_i^w\big(Z_{1:n},(x,y)\big)\mathbb{1}_{\left\{R_i^{(x,y)} = R_{n+1}^{(x,y)}\right\}} \le 1-\alpha \right\}, \quad (5.17)$$

where U is an independent $\mathrm{Unif}(0,1)$ variable.

We can also randomize the quantile form in (5.14) using an external variable $B^w \sim \mathrm{Bernoulli}\left(\frac{1-\alpha-\lfloor 1-\alpha\rfloor_w}{\lceil 1-\alpha\rceil_w - \lfloor 1-\alpha\rfloor_w}\right)$, where $\lfloor 1-\alpha\rfloor_w = \max\left\{\tau \in \mathrm{Range}(\widehat{F}_n^w) : \tau \le 1-\alpha\right\}$. The randomized conformal set is

$$\begin{aligned}\widehat{C}_n^{w,*} = \Bigg\{ y : R_{n+1}^{(x,y)} \le & B^w \mathrm{Quantile}\Big(1-\alpha; \sum_{i=1}^{n} p_i^w\big(Z_{1:n},(x,y)\big)\delta_{R_i^{(x,y)}} + p_{n+1}^w\big(Z_{1:n},(x,y)\big)\delta_\infty\Big) \\ & + (1-B^w)\Big(\lfloor 1-\alpha\rfloor_w; \sum_{i=1}^{n} p_i^w\big(Z_{1:n},(x,y)\big)\delta_{R_i^{(x,y)}} + p_{n+1}^w\big(Z_{1:n},(x,y)\big)\delta_\infty\Big)\Bigg\}. \end{aligned} \quad (5.18)$$

Both the two randomized conformal set have exact $1-\alpha$ coverage.

5.1.1 Conformal prediction for covariate shift

We now demonstrate how to apply the above results to derive a covariate-shifted version of conformal prediction.

Proposition 5.4. Suppose that $\{Z_i = (X_i, Y_i), i = 1, \cdots, n+1\}$ is distributed according to model (5.1)., and $\widetilde{P}_X$ is absolutely continuous with respect to P_X with Radon-Nikodym derivative $w = \mathrm{d}\widetilde{P}_X/\mathrm{d}P_X$. Suppose V is a score function that is symmetric in its last $n+1$ arguments. Define

$$\pi_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)}, \quad i = 1, \cdots, n \quad \text{and} \quad \pi_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}. \quad (5.19)$$

Fix a nominal error level $\alpha \in (0,1)$, and define a weighted conformal set at a point $x \in \mathcal{X}$ by

$$\widehat{C}_n^w(x) = \left\{ y : R_{n+1}^{(x,y)} \le \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \pi_i^w(x)\delta_{R_i^{(x,y)}} + \pi_{n+1}^w(x)\delta_\infty\right)\right\}, \quad (5.20)$$

where $\{R_i^{(x,y)}\}$ is defined in (5.13). Then $\widehat{C}_n^w$ satisfies

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n^w(X_{n+1})\right) \ge 1-\alpha. \quad (5.21)$$

Proof. Since $\{Z_i = (X_i, Y_i), i = 1, \cdots, n\}$ are i.i.d., their joint distribution are symmetric. Then we can set $w_1, \cdots, w_n$ to be 1. Moreover, set w_{n+1} to be an importance ratio:

$$w = \frac{\mathrm{d}\widetilde{P}}{\mathrm{d}P} = \frac{\mathrm{d}\widetilde{P}_X}{\mathrm{d}P_X}, \quad (5.22)$$

which is the Radon-Nikodym derivative. Hence $\{Z_i = (X_i, Y_i), i = 1, \cdots, n\}$ is exchangeable with respect to $w_1 = 1, \cdots, w_n = 1$ and $w_{n+1} = w$. According to (5.8), for $\{z_i = (x_i, y_i), i = 1, \cdots, n+1\}$,

$$p_i^w(z_1, \cdots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} w_{n+1}(x_{\sigma(n+1)})}{\sum_\sigma w_{n+1}(x_{\sigma(n+1)})} = \frac{n! w(x_i)}{n!\sum_{j=1}^{n+1} w(x_j)} = \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)}, \quad (5.23)$$

namely, $\pi_i^w(x) = p_i^w(Z_1, \cdots, Z_n, (x, y)),\ i = 1, \cdots, n+1$. Then the coverage property of $\widehat{C}_n^w$ given in (5.20) and (5.21) follows from Theorem 5.3. □

5.1.2 Likelihood ratio estimation

In weighted conformal prediction, we need to estimate the likelihood ratio $w = \mathrm{d}\widetilde{P}_X/\mathrm{d}P_X$. Suppose we have access to unlabeled data $X_{n+1}, \cdots, X_{n+m} \in \mathcal{X}$ at prediction time. Then we can use any classifier like logistic regression or random forests to estimate probabilities of class membership.

- Add class labels to the training data $\{(X_i, C_i)\}_{i=1}^{m+n}$, where we assign $C_i = 0$ for $i = 1, \cdots, n$ and $C_i = 1$ for $i = n+1, \cdots, n+m$.
- Train a classifier $\widehat{p} : \mathbb{R} \to [0, 1]$ on $\{(X_i, C_i)\}_{i=1}^{m+n}$ such that $\widehat{p}(x)$ estimates the probability $\mathbb{P}(C = 1|X = x)$.

Note that the odds ratio

$$\frac{\mathbb{P}(C = 1|X = x)}{\mathbb{P}(C = 0|X = x)} = \frac{\mathbb{P}(C = 1)}{\mathbb{P}(C = 0)} \frac{\mathrm{d}\widetilde{P}_X}{\mathrm{d}P_X}. \tag{5.24}$$

By (5.19), it suffices to know the likelihood ratio up to a proportionally constant. Therefore, we can estimate our weight function by

$$\widehat{w}(x) = \frac{\widehat{p}(x)}{1 - \widehat{p}(x)}, \tag{5.25}$$

and construct a weighted conformal set according to (5.19)-(5.20).

5.1.3 Generalization: Conformal prediction for structured-X settings

We now consider a more general case of covariate shift, in which we assume a joint distribution Λ of our training and test samples. This can be useful in certain structured-X settings, for example, where the sequence $X_1, \cdots, X_{n+1}$ has some kind of Markov structure.

Model. We assume $\{Z_i = (X_i, Y_i), i = 1, \cdots, n+1\}$ are distributed to

$$\begin{cases} (X_1, \cdots, X_{n+1}) \sim \Lambda, \\ Y_i \,|\, X_i \sim P_{Y|X}, \text{ independently, for } i = 1, \cdots, n+1. \end{cases} \tag{5.26}$$

Furthermore, let λ be the joint density (or more generally, Radon-Nikodym derivative with respect to a base measure) of $X_1, \cdots, X_n$.

Theorem 5.5. Let V be a score function that is symmetric in its last $n+1$ arguments. Define conformity scores $\{R_i^{(x,y)}\}$ as in (5.13), and

$$p_i^\lambda(x_1, \cdots, x_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \lambda(x_{\sigma(1)}, \cdots, x_{\sigma(n+1)})}{\sum_\sigma \lambda(x_{\sigma(1)}, \cdots, x_{\sigma(n+1)})}. \tag{5.27}$$

Define the conformal set at a point $x \in \mathcal{X}$ with nominal error level $\alpha \in (0, 1)$ by

$$\widehat{C}_n^\lambda(x) = \left\{ y : R_{n+1}^{(x,y)} \le \text{Quantile}\left(1 - \alpha; \sum_{i=1}^n p_i^\lambda(X_1, \cdots, X_n, x)\delta_{R_i^{(x,y)}} + p_{n+1}^\lambda(X_1, \cdots, X_n, x)\delta_\infty\right)\right\}. \tag{5.28}$$

Then $\widehat{C}_n^{\lambda}$ satisfies

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n^{\lambda}(X_{n+1})\right) \geq 1-\alpha. \tag{5.29}$$

Proof. We fix $\{z_i = (x_i, y_i)\}_{i=1}^n$ and denote by $E(z_1, \cdots, z_{n+1})$ the event that $\{Z_1, \cdots, Z_{n+1}\} = \{z_1, \cdots, z_{n+1}\}$. Let $r_i = V(z_i; z_1, \cdots, z_{n+1})$, $i = 1, \cdots, n+1$, and denote $\mathcal{S}(i) = \{j \in [n+1] : r_i = V(z_j; z_1, \cdots, z_{n+1})\}$. Then we have for all $i = 1, \cdots, n+1$ that

$$\begin{aligned}
\mathbb{P}(R_{n+1} = r_i | E(z_1, \cdots, z_{n+1})) &= \mathbb{P}(Z_{n+1} \in \{z_j : j \in \mathcal{S}(i)\} | E(z_1, \cdots, z_{n+1})) \\
&= \frac{\sum_{\sigma:\sigma(n+1)\in\mathcal{S}(i)} \lambda(x_{\sigma(1)}, \cdots, x_{\sigma(n+1)}) \prod_{i=1}^{n+1} p_{Y|X}(y_i|x_i)}{\sum_{\sigma} \lambda(x_{\sigma(1)}, \cdots, x_{\sigma(n+1)}) \prod_{i=1}^{n+1} p_{Y|X}(y_i|x_i)} \\
&= \sum_{k\in\mathcal{S}(i)} p_k^{\lambda}(z_1, \cdots, z_{n+1}).
\end{aligned} \tag{5.30}$$

Then we have

$$R_{n+1} \,|\, E(z_1, \cdots, z_{n+1}) \sim \sum_{i=1}^{n+1} p_i^{\lambda}(z_1, \cdots, z_{n+1}) \delta_{r_i}. \tag{5.31}$$

The remaining part is akin to the proof of Lemma 5.2. □

Theorem 5.5 constructs a conformal set (5.28) with general coverage (5.29). However, the computational expense can be extremely high because the cauculation (5.27) is complicated, even intractable when n is large.

6 Non-exchangeable Conformal Prediction

In this section we discuss non-exchangeable conformal prediction, as developed in Barber et al. (2022). This approach does not require the assumptions of exchangeability of the data. It is also referred to as custom-weighted conformal prediction, since the weights are fixed manually instead of being a function of the data.

6.1 Robust inference through weighted quantiles

As an extension, let $\{Z_i = (X_i, Y_i), i = 1, \cdots, n+1\}$ be data points (with the last one $Z_{n+1} = (X_{n+1}, Y_{n+1})$ serving as the test point) that are no longer changeable, and V a score function. For non-exchangeable conformal prediction, we choose a set of fixed weights $w_1, \cdots, w_n \in [0, 1]$ such that a higher weight is associated with a data point that undergoes less distribution shift from Z_{n+1} (for example, in sense of temporal or spatial proximity). To simplify notation, in what follows, given $w_i \in [0, 1], i = 1, \cdots, n$, we define normalized weights

$$\widetilde{w}_i = \frac{w_i}{\sum_{i=1}^n w_i + 1}, \; i = 1, \cdots, n, \quad \text{and } \widetilde{w}_{n+1} = \frac{1}{\sum_{i=1}^n w_i + 1}. \tag{6.1}$$

So far, we still assume that the score function V is symmetric in its last $n+1$ arguments. With the normalized weight given, we then define the full conformal set:

$$\widehat{C}_n^w(x) = \left\{ y : R_{n+1}^{(x,y)} \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^n \widetilde{w}_i \delta_{R_i^{(x,y)}} + \widetilde{w}_{n+1}\delta_{\infty}\right)\right\}, \tag{6.2}$$

where the conformity scores $\{R_i^{(x,y)}\}$ are defined in (5.13).

Split version. As before, the split conformal version of the above result can be viewed as a special case where the score function relies on a point predictor that has been fit on an external dataset:

$$\widehat{C}_n^w(x) = \widehat{\mu}_0(x) \pm \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \widetilde{w}_i \delta_{|Y_i - \widehat{\mu}_0(X_i)|} + \widetilde{w}_{n+1}\delta_\infty\right), \tag{6.3}$$

where $\widehat{\mu}_0$ is a pre-trained model.

Remark. In fact, we can recover the classical conformal prediction methods (unweighted version) by setting weights $w_1 = \cdots = w_n = 1$. Furthermore, as in the discussion of likelihood-weighted conformal prediction, we can derive the CDF form and auxiliary randomization (but here it is not clear we will achieve an exact coverage).

The theoretical results of these conformal sets will follow as a corollary of more general results that also accommodate nonsymmetric algorithms, which is going to be discussed in section 6.2. For brevity, we do not restate them in this section.

6.2 Enhanced predictions with nonsymmetric algorithms

Now, we will allow the score function V to be an arbitrary function of the data points, removing the requirement of being symmetric in the last $n+1$ arguments. Namely, the prediction algorithm does not treat all input data points in a symmetric way. For instance, the order of input data matters. We first introduce some notations:

- Denote by $\mathbf{Z} = (Z_1, \cdots, Z_{n+1})$ the data vector (an ordered sequence),
- $\mathbf{Z}^i = (Z_1, \cdots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \cdots, Z_n, Z_i)$ the sequence with components i and $n+1$ swapped, and
- $R(\mathbf{Z})$ the conformity score vector corresponding to data $\mathbf{Z}$, with components $R(\mathbf{Z})_j = V(Z_j; \mathbf{Z})$.

With the added flexibility of a nonsymmetric prediction algorithm, we will need some key modification to the methods to maintain predictive coverage. Our modification requires that, before applying the model fitting algorithm, we first randomly swap the tags of two of the data points in the ordering.

We first draw a random index K from a multinomial distribution which puts mass $\widetilde{w}_i$ at value i:

$$K \sim \sum_{i=1}^{n+1} \widetilde{w}_i \delta_i. \tag{6.4}$$

Then we apply our prediction algorithm to data $\mathbf{Z}^K$ in place of $\mathbf{Z}$. Let $\mathbf{Z}^{(x,y)} = ((X_1, Y_1), \cdots, (X_n, Y_n), (x, y))$, our conformity scores are defined as follows:

$$\begin{cases} R_i^{(x,y),K} = V\left((X_i, Y_i); \left(\mathbf{Z}^{(x,y)}\right)^K\right), & i = 1, \cdots, n, \\ R_{n+1}^{(x,y),K} = V\left((x, y); \left(\mathbf{Z}^{(x,y)}\right)^K\right). \end{cases} \tag{6.5}$$

After drawing a random index K as in (6.4) and obtaining the conformity scores in (6.5), the prediction set is given by

$$\widehat{C}_n^w(x) = \left\{ y : R_{n+1}^{(x,y),K} \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \widetilde{w}_i \delta_{R_i^{(x,y),K}} + \widetilde{w}_{n+1}\delta_\infty\right)\right\}, \tag{6.6}$$

Now let's investigate the coverage property of the conformal set given by (6.6). The following theorem gives a lower bound on coverage, which can be seen as a generalization of its counterpart of exchangeable data points and symmetric score functions.

Theorem 6.1 (Lower bounds on coverage). Let V be an arbitrary score function. Define the random index and the conformity scores as in (6.4)-(6.5). Then the non-exchangeable full conformal set $\widehat{C}_n^w$ given by (6.6) satisfies

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n^w(X_{n+1})\right) \geq 1 - \alpha - \sum_{i=1}^{n} \widetilde{w}_i \cdot d_{\mathrm{TV}}\big(R(\mathbf{Z}), R(\mathbf{Z}^i)\big). \tag{6.7}$$

This result also holds for split conformal sets (6.2) and (6.3) with random index K dropped.

Proof. For brevity, we denote $R_i^K = R_i^{(X_{n+1},Y_{n+1}),K},\ i = 1, \cdots, n+1$. The definition of the non-exchangeable conformal set (6.6) implies

$$Y_{n+1} \notin \widehat{C}_n^w(X_n) \quad \Leftrightarrow \quad R_{n+1}^K > \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \widetilde{w}_i \delta_{R_i^K} + \widetilde{w}_{n+1}\delta_\infty\right). \tag{6.8}$$

Note that

$$R(\mathbf{Z}^K) = \left(R_1^K, \cdots, R_{K-1}^K, R_{n+1}^K, R_{K-1}^K, \cdots, R_n^K, R_K^K\right), \tag{6.9}$$

we have

$$\sum_{i=1}^{n} \widetilde{w}_i \delta_{R_i^K} + \widetilde{w}_{n+1}\delta_\infty = \sum_{i=1, i\neq K}^{n} \widetilde{w}_i \delta_{R_i^K} + \widetilde{w}_K\left(\delta_{R_K^K} + \delta_\infty\right) + \left(\widetilde{w}_{n+1} - \widetilde{w}_K\right)\delta_\infty, \tag{6.10}$$

$$\sum_{i=1}^{n+1} \widetilde{w}_i \delta_{(R(\mathbf{Z}^K))_i} = \sum_{i=1, i\neq K}^{n} \widetilde{w}_i \delta_{R_i^K} + \widetilde{w}_K\left(\delta_{R_K^K} + \delta_{R_{n+1}^K}\right) + \left(\widetilde{w}_{n+1} - \widetilde{w}_K\right)\delta_{R_K^K}. \tag{6.11}$$

Since $w_1, \cdots, w_n \in [0,1]$, we have $\widetilde{w}_{n+1} = \max\{\widetilde{w}_1, \cdots, \widetilde{w}_{n+1}\}$. Hence the distribution (6.10) is greater than (6.11), and

$$\text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} \widetilde{w}_i \delta_{R_i^K} + \widetilde{w}_{n+1}\delta_\infty\right) \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n+1} \widetilde{w}_i \delta_{(R(\mathbf{Z}^K))_i}\right). \tag{6.12}$$

Combining (6.8), (6.9) and (6.10) yields

$$Y_{n+1} \notin \widehat{C}_n^w(X_n) \ \Rightarrow\ \left(R(\mathbf{Z}^K)\right)_K > \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n+1} \widetilde{w}_i \delta_{(R(\mathbf{Z}^K))_i}\right) \ \Leftrightarrow\ K \in \mathcal{S}\left(R(\mathbf{Z}^K)\right), \tag{6.13}$$

where we define for any $\mathbf{r} = (r_1, \cdots, r_{n+1}) \in \mathbb{R}^{n+1}$ the strange point set:

$$\mathcal{S}(\mathbf{r}) = \left\{ i \in [n+1] : r_i > \text{Quantile}\left(1-\alpha; \sum_{j=1}^{n} \widetilde{w}_j \delta_{r_j}\right)\right\}. \tag{6.14}$$

Suppose $R \sim \sum_{j=1}^{n} \widetilde{w}_j \delta_{r_j}$, which is a multinomial distribution. By Corollary 1.2, we have

$$\sum_{i \in \mathcal{S}(\mathbf{r})} \widetilde{w}_i = \mathbb{P}\left(R > \text{Quantile}\left(1-\alpha; \sum_{j=1}^{n} \widetilde{w}_j \delta_{r_j}\right)\right) \leq \alpha, \tag{6.15}$$

which holds for all $\mathbf{r} \in \mathbb{R}^{n+1}$.

Recall that $K \sim \sum_{i=1}^{n+1} \widetilde{w}_i \delta_i$ is independent of $\mathbf{Z} := \mathbf{Z}^{(X_{n+1}, Y_{n+1})}$, we can bound the probability of the last event in (6.13) as follows:

$$\begin{aligned}
\mathbb{P}\left\{K \in \mathcal{S}\left(R(\mathbf{Z}^K)\right)\right\} &= \sum_{i=1}^{n+1} \widetilde{w}_i \cdot \mathbb{P}\left\{i \in \mathcal{S}\left(R(\mathbf{Z}^i)\right)\right\} \\
&\leq \sum_{i=1}^{n+1} \widetilde{w}_i \cdot \mathbb{P}\left\{i \in \mathcal{S}\left(R(\mathbf{Z})\right) + d_{\mathrm{TV}}\left(R(\mathbf{Z}), R(\mathbf{Z}^i)\right)\right\} \\
&= \mathbb{E}\left[\sum_{i \in \mathcal{S}(R(\mathbf{Z}))} \widetilde{w}_i\right] + \sum_{i=1}^{n+1} \widetilde{w}_i \cdot d_{\mathrm{TV}}\left(R(\mathbf{Z}), R(\mathbf{Z}^i)\right) \\
&\leq \alpha + \sum_{i=1}^{n+1} \widetilde{w}_i \cdot d_{\mathrm{TV}}\left(R(\mathbf{Z}), R(\mathbf{Z}^i)\right), \qquad (6.16)
\end{aligned}$$

where the last inequality follows from (6.15). By combining (6.13) and (6.16), we obtain the bound in (6.7). □

Theorem 6.2 (Upper bounds on coverage). Let V be an arbitrary score function. Define the random index and the conformity scores as in (6.4)-(6.5). Suppose the scores $R_i^{(X_{n+1}, Y_{n+1}), K}, i = 1, \cdots, n+1$ are almost surely distinct. Then the non-exchangeable full conformal set $\widehat{C}_n^w$ given by (6.6) satisfies

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n^w(X_{n+1})\right) < 1 - \alpha + \widetilde{w}_{n+1} + \sum_{i=1}^{n} \widetilde{w}_i \cdot d_{\mathrm{TV}}\big(R(\mathbf{Z}), R(\mathbf{Z}^i)\big). \qquad (6.17)$$

This result also holds for split conformal sets (6.2) and (6.3) with random index K dropped.

Proof. For brevity, we denote $R_i^K = R_i^{(X_{n+1}, Y_{n+1}), K}$, $i = 1, \cdots, n+1$. The definition of the non-exchangeable conformal set (6.6) implies

$$\begin{aligned}
Y_{n+1} \in \widehat{C}_n^w(X_n) \quad &\Leftrightarrow \quad R_{n+1}^K \leq \text{Quantile}\left(1 - \alpha; \sum_{i=1}^{n+1} \widetilde{w}_i \delta_{R_i^K}\right) \qquad \text{(Replacing } \delta_\infty \text{ by } \delta_{R_{n+1}^K}\text{)} \\
&\Leftrightarrow \quad \left(R(\mathbf{Z}^K)\right)_K \leq \text{Quantile}\left(1 - \alpha; \sum_{i=1}^{n+1} \widetilde{w}_{\sigma_K(i)} \delta_{(R(\mathbf{Z}^K))_i}\right), \qquad (6.18)
\end{aligned}$$

where we denote by σ_K the permutation on $[n+1]$ after swapping $n+1$ and K. Since K is independent of $\mathbf{Z}$,

$$\begin{aligned}
&\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n^w(X_n)\right) \\
&= \sum_{k=1}^{n+1} \widetilde{w}_k \cdot \mathbb{P}\left(\left(R(\mathbf{Z}^k)\right)_k \leq \text{Quantile}\left(1 - \alpha; \sum_{i=1}^{n+1} \widetilde{w}_{\sigma_k(i)} \delta_{(R(\mathbf{Z}^k))_i}\right)\right) \\
&\leq \sum_{k=1}^{n+1} \widetilde{w}_k \cdot \left\{\mathbb{P}\left((R(\mathbf{Z}))_k \leq \text{Quantile}\left(1 - \alpha; \sum_{i=1}^{n+1} \widetilde{w}_{\sigma_k(i)} \delta_{(R(\mathbf{Z}))_i}\right)\right) + d_{\mathrm{TV}}\left(R(\mathbf{Z}), R(\mathbf{Z}^k)\right)\right\} \\
&\leq \underbrace{\mathbb{E}\left[\sum_{k=1}^{n+1} \widetilde{w}_k \cdot \mathbb{1}\left\{(R(\mathbf{Z}))_k \leq \text{Quantile}\left(1 - \alpha; \sum_{i=1}^{n+1} \widetilde{w}_{\sigma_k(i)} \delta_{(R(\mathbf{Z}))_i}\right)\right\}\right]}_{\text{(a)}} + \sum_{k=1}^{n+1} \widetilde{w}_k \cdot d_{\mathrm{TV}}\left(R(\mathbf{Z}), R(\mathbf{Z}^k)\right). \qquad (6.19)
\end{aligned}$$

The it remains to bound term (a). For any $\mathbf{r} = (r_1, \cdots, r_{n+1}) \in \mathbb{R}^{n+1}$, define the normal point set:

$$\mathcal{N}(\mathbf{r}) = \left\{k \in [n+1] : r_k \leq \text{Quantile}\left(1 - \alpha; \sum_{i=1}^{n+1} \widetilde{w}_{\sigma_k(i)} \delta_{r_i}\right)\right\}. \qquad (6.20)$$

Recall the form in (a), it suffices to show that for any $\mathbf{r} \in \mathbb{R}^{n+1}$ such that $r_1, \cdots, r_{n+1}$ are distinct,

$$\sum_{k \in \mathcal{N}(\mathbf{r})} \widetilde{w}_k < 1 - \alpha + \widetilde{w}_{n+1}. \tag{6.21}$$

Let $k^* = \text{argmax}_{k \in \mathcal{N}(\mathbf{r})} r_k$, which indices the greatest r_k over $k \in \mathcal{N}(\mathbf{r})$. Define $\mathcal{K}^* := \{k \in [n+1] : r_k \leq r_{k^*}\}$, and $\mathcal{K}^{**} := \{k \in [n+1] : r_k < r_{k^*}\}$. Since $\mathcal{N}(\mathbf{r}) \subseteq \mathcal{K}^*$, we have

$$\begin{aligned}\sum_{k \in \mathcal{N}(\mathbf{r})} \widetilde{w}_k \leq \sum_{k \in \mathcal{K}^*} \widetilde{w}_k &= \widetilde{w}_{k^*} + \sum_{k \in \mathcal{K}^{**}} \widetilde{w}_k \\ &= \widetilde{w}_{k^*} + \underbrace{\sum_{k \in \mathcal{K}^{**}} (\widetilde{w}_k - \widetilde{w}_{\sigma_{k^*}(k)})}_{\text{(b)}} + \underbrace{\sum_{k \in \mathcal{K}^{**}} \widetilde{w}_{\sigma_{k^*}(k)}}_{\text{(c)}}.\end{aligned} \tag{6.22}$$

To bound term (b), note that

$$\begin{aligned}\text{(b)} &= \sum_{k=1}^{n+1} (\widetilde{w}_k - \widetilde{w}_{\sigma_k(k^*)}) \mathbb{1}_{\{r_k < r_{k^*}\}} \\ &= \sum_{k=1, k \neq k^*}^{n} \cancel{(\widetilde{w}_k - \widetilde{w}_k) \mathbb{1}_{\{r_k < r_{k^*}\}}} + (\widetilde{w}_{n+1} - \widetilde{w}_{k^*}) \mathbb{1}_{\{r_{n+1} < r_{k^*}\}} + \cancel{(\widetilde{w}_{k^*} - \widetilde{w}_{n+1}) \mathbb{1}_{\{r_{k^*} < r_{k^*}\}}} \\ &\leq \widetilde{w}_{n+1} - \widetilde{w}_{k^*}.\end{aligned} \tag{6.23}$$

For the term (c), we have $k^* \in \mathcal{N}(\mathbf{r})$, hence

$$r_{k^*} \leq \text{Quantile}\left(1 - \alpha; \sum_{k=1}^{n+1} \widetilde{w}_{\sigma_{k^*}(k)} \delta_{r_k}\right) \quad \Rightarrow \quad \text{(c)} = \sum_{k=1}^{n+1} \widetilde{w}_{\sigma_k(k^*)} \mathbb{1}_{\{r_k < r_{k^*}\}} < 1 - \alpha. \tag{6.24}$$

Combining (6.22), (6.23) and (6.24) yields (6.21), which concludes the proof. □

By Theorem 6.1 and Theorem 6.2, the coverage of non-exchangeable conformal sets falls in an interval, akin to what we derived in exchangeable cases:

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_n^w(X_n)\right) \in \left[1 - \alpha - \sum_{i=1}^{n} \widetilde{w}_i \cdot d_{\text{TV}}\big(R(\mathbf{Z}), R(\mathbf{Z}^i)\big),\ 1 - \alpha + \widetilde{w}_{n+1} + \sum_{i=1}^{n} \widetilde{w}_i \cdot d_{\text{TV}}\big(R(\mathbf{Z}), R(\mathbf{Z}^i)\big)\right). \tag{6.25}$$

6.3 Remarks

Choosing the weights. Theoretical findings presented earlier validate the intuition that assigning higher weights, denoted as w_i, to data points (X_i, Y_i) believed to be drawn from a distribution similar to (X_{n+1}, Y_{n+1}) is beneficial, while lower weights should be allocated to less reliable points. However, optimal weight selection involves a tradeoff. If many weights w_i are set to be quite low, it shrinks the effective sample size of the method. For instance, in split conformal prediction, this reduction in effective sample size affects the estimation of the empirical quantile of the residual distribution, often resulting in broader prediction intervals. Striking the right balance is crucial, as excessively low weights may lead to overly wide prediction intervals. At the extreme end, setting all weights to zero ($w_1 = \cdots = w_n = 0$) eliminates the coverage gap but yields an uninformative prediction interval, denoted as $\widehat{C}_n^w(X_{n+1}) \equiv \mathbb{R}$. The optimal choice of weights, and how to quantify optimality, pose intriguing and important questions for future exploration.

Coverage gap bounds. We define the coverage gap as the loss in coverage compared to what is achieved under exchangeability. At a desired error level of $1-\alpha$, we have:

$$\text{Coverage gap} := 1-\alpha-\mathbb{P}\left(Y_{n+1}\in \widehat{C}_n^w(X_{n+1})\right), \tag{6.26}$$

By Theorem 6.1, we can bound the coverage gap as

$$\text{Coverage gap} \le \sum_{i=1}^n \widetilde{w}_i \cdot d_{\mathrm{TV}}\left(R(\mathbf{Z}), R(\mathbf{Z}^i)\right) \le \sum_{i=1}^n \widetilde{w}_i \cdot d_{\mathrm{TV}}\left(\mathbf{Z}, \mathbf{Z}^i\right). \tag{6.27}$$

where the second inequality follows from the data processing inequality.

Exchangeable setting. When $Z_i=(X_i,Y_i), i=1,\cdots,n+1$ are exchangeable, we are back to the classical setting for conformal prediction. Since we have $R(\mathbf{Z}) \overset{\mathrm{d}}{=} R(\mathbf{Z}^i)$ by exchangeability, the slack in (6.7) vanishes and the coverage collapses to an exact $1-\alpha$ lower bound.

Independent setting. When $Z_i=(X_i,Y_i), i=1,\cdots,n+1$ are independent (but not necessarily identically distributed), the bound for coverage gap in (6.27) can be slacked to

$$\text{Coverage gap} \le \sum_{i=1}^n \widetilde{w}_i \cdot d_{\mathrm{TV}}\left(\mathbf{Z}, \mathbf{Z}^i\right) \le 2\sum_{i=1}^n \widetilde{w}_i \cdot d_{\mathrm{TV}}(Z_i, Z_{n+1}). \tag{6.28}$$

To prove this result, we need to introduce the theory of coupling.

Definition 6.3 (Coupling). Let P and Q be two probability measures on the same measurable space $(\mathbb{R},\mathcal{B})$. A coupling of P and Q is a probability measure μ on the product space $(\mathbb{R}^2,\mathcal{B}^2)$ such that the marginals of μ coincide with P and Q: $\mu(A\times\mathbb{R})=P(A),\ \mu(\mathbb{R}\times A)=Q(A)$. In other words, if random variables $X\sim P$ and $Y\sim Q$, then (X',Y') is a coupling of X and Y if $X'\overset{\mathrm{d}}{=}X$ and $Y'\overset{\mathrm{d}}{=}Y$.

Lemma 6.4 (Maximal coupling). For any coupling μ of P and Q, we have

$$\mu(\{(x,y):x=y\}) \le 1-d_{\mathrm{TV}}(P,Q). \tag{6.29}$$

Moreover, there exists a maximal coupling μ^* such that (6.29) becomes an equality after replacing μ by μ^*.

Proof. Using Hahn decomposition theorem, we know that there exists a partition (E,E^c) of $\mathbb{R}$ such that for any Borel set $A\subseteq E$, $P(A)\le Q(A)$, and for any Borel set $B\subseteq E^c$, $P(B)\ge Q(B)$. Then any coupling μ of P and Q must satisfy

$$\begin{aligned}
\mu(\{(x,y):x=y\}) &= \mu(\{(x,y):x=y, x\in E\}) + \mu(\{(x,y):x=y, x\in E^c\}) \\
&\le \mu(\{(x,y):x\in E\}) + \mu(\{(x,y):y\in E^c\}) \\
&= P(E)+1-Q(E) = 1-d_{\mathrm{TV}}(P,Q).
\end{aligned} \tag{6.30}$$

For brevity, we denote $\gamma = 1-d_{\mathrm{TV}}(P,Q)$. Then we define probability measures F,G,H on $(\mathbb{R},\mathcal{B})$ as follows:

for all $A \in \mathcal{B}$,

$$\begin{aligned} F(A) &= \frac{P(A \cap E) + Q(A \cap E^c)}{\gamma}, \\ G(A) &= \frac{P(A \cap E^c) - Q(A \cap E^c)}{1-\gamma}, \\ H(A) &= \frac{Q(A \cap E) - P(A \cap E)}{1-\gamma}. \end{aligned} \tag{6.31}$$

For any $S \in \mathcal{B}^2$, we define the marginals $S^x = \{(x', y') \in S : x' = x\}$, and $S_y = \{(x', y') \in S : y' = y\}$. Then we define the maximal coupling μ^* as

$$\mu^*(S) = \gamma F(\{(x, y) \in S : x = y\}) + (1-\gamma)(G \times H)(S), \tag{6.32}$$

where $(G \times H)(S) = \int H(S^x) \mathrm{d}G(x) = \int G(S_y) \mathrm{d}H(y)$ by Fubini's theorem. Then for all $A \in \mathcal{B}$,

$$\begin{aligned} \mu^*(A \times \mathbb{R}) &= \gamma F(A) + (1-\gamma) G(A) = P(A), \\ \mu^*(\mathbb{R} \times A) &= \gamma F(A) + (1-\gamma) H(A) = Q(A), \end{aligned} \tag{6.33}$$

and

$$\mu^*(\{(x, y) : x = y\}) = \gamma F(\mathbb{R}) = 1 - d_{\mathrm{TV}}(P, Q). \tag{6.34}$$

Hence μ^* is a valid maximal coupling of P and Q, and we complete the proof. □

Lemma 6.5. Let $Z_i = (X_i, Y_i)$, $i = 1, \cdots, n+1$ be independent random variables. Then for any $i \in [n]$,

$$d_{\mathrm{TV}}\left(\mathbf{Z}, \mathbf{Z}^i\right) = 2 d_{\mathrm{TV}}\left(Z_i, Z_{n+1}\right) - d_{\mathrm{TV}}^2\left(Z_i, Z_{n+1}\right). \tag{6.35}$$

Proof. By Lemma 6.4, there exists a distribution μ^* on a pair of random variable (U_i, U_{n+1}) such that, marginally, $U_i \overset{\mathrm{d}}{=} Z_i$ and $U_{n+1} \overset{\mathrm{d}}{=} Z_{n+1}$, and such that $\mathbb{P}(U_i = U_{n+1}) = 1 - d_{\mathrm{TV}}(Z_i, Z_{n+1})$. Let (V_i, V_{n+1}) be an independent copy of (U_i, U_{n+1}). Denote

$$\mathbf{U} = (Z_1, \cdots, Z_{i-1}, U_i, Z_{i+1}, \cdots, Z_n, V_{n+1}), \mathbf{V} = (Z_1, \cdots, Z_{i-1}, V_i, Z_{i+1}, \cdots, Z_n, U_{n+1}). \tag{6.36}$$

Then $\mathbf{U} \overset{\mathrm{d}}{=} \mathbf{V} \overset{\mathrm{d}}{=} \mathbf{Z}$. Again applying Lemma 6.4, we have

$$\begin{aligned} d_{\mathrm{TV}}(\mathbf{Z}, \mathbf{Z}^i) = d_{\mathrm{TV}}\left(\mathbf{U}, \mathbf{V}^i\right) &\leq 1 - \mathbb{P}\left(\mathbf{U} = \mathbf{V}^i\right) \\ &= 1 - \mathbb{P}\left(U_i = U_{n+1}, V_i = V_{n+1}\right) \\ &= 1 - \mathbb{P}\left(U_i = U_{n+1}\right) \mathbb{P}\left(V_i = V_{n+1}\right) \\ &= 1 - \left(1 - d_{\mathrm{TV}}\left(Z_i, Z_{n+1}\right)\right)^2 = 2 d_{\mathrm{TV}}\left(Z_i, Z_{n+1}\right) - d_{\mathrm{TV}}^2\left(Z_i, Z_{n+1}\right), \end{aligned} \tag{6.37}$$

which concludes the proof. □

By applying Lemma 6.5 to (6.27), we immediately obtain (6.28).

7 Adaptive conformal inference

Adaptive conformal inference (ACI), proposed by Gibbs and Candès (2021), is a common-used algorithm in conformal-like sequential prediction.

Setting. Let $\{(X_t, Y_t), t \in \mathbb{N}_0\} \subseteq \{\Omega \to \mathcal{X} \times \mathcal{Y}\}$ be a stochastic process indexed by time. For each step t we have an algorithm that produces a prediction set $C_t^\beta \subseteq \mathcal{Y}$ for Y_t based on tha past data $\{(X_s, Y_s), s < t\}$, at any nominal error level $\beta \in \mathbb{R}$. We assume that our prediction sets saturate at any level below 0 or above 1, namely, for any $t \in \mathbb{N}$,

$$C_t^\beta = \emptyset, \ \beta \leq 0, \quad \text{and} \quad C_t^\beta = \mathcal{Y}, \ \beta \geq 1. \tag{7.1}$$

ACI update. During the prediction process, ACI adjusts the working level $1 - \alpha_t$ of the prediction sets over time $t \in \mathbb{N}$ in order to maintain a realized coverage as close to $1 - \alpha$ as possible, where $\alpha \in (0, 1)$ is some prespecified error tolerance. Let $\eta > 0$ be a step size and $e_t = \mathbb{1}_{\{Y_t \notin C_t^{1-\alpha_t}\}}$ the error indicator. Set $\alpha_0 = \alpha$, the update formula is

$$\alpha_t = \alpha_{t-1} - \eta(e_{t-1} - \alpha), \ t = 1, 2, \cdots. \tag{7.2}$$

This formula has an intuitive interpretation. If we cover, then we shrink the prediction sets by increasing the working error level by $\eta\alpha$. If we miscover, then we inflate the prediction sets by decreasing the working error level by $\eta(1-\alpha)$. In fact, such a self-correcting property makes the working error levels $\{\alpha_t\}$ produced by (7.2) uniformly bounded.

Lemma 7.1 (Boundedness of ACI iterates). The iterates from ACI (7.2) is uniformly bounded by $[-\eta, 1+\eta]$.

Proof. Argue by contradiction. Suppose $\exists t \geq 1$ such that $\alpha_t < -\eta$. If $e_{t-1} = 1$, then $Y_{n-1} \notin C_{t-1}^\beta$, and $\alpha_{t-1} = \alpha_t + \eta(1-\alpha) < -\eta\alpha < 0$. Recall the saturation property (7.1), we have $C_{t-1}^\beta = \mathcal{Y}$, a contradiction! Hence $e_{t-1} = 0$, and $\alpha_{t-1} = \alpha_t - \eta\alpha < \alpha_t$. Following this, we have $0 > -\eta > \alpha_t > \alpha_{t-1} > \alpha_{t-2} > \cdots > \alpha_0 = \alpha > 0$, again a contradiction! Therefore $\inf_{t\in\mathbb{N}} \alpha_t \geq -\eta$, and similarly we can prove $\sup_{t\in\mathbb{N}} \alpha_t \leq 1 + \eta$, which concludes the proof. □

Theorem 7.2 (Asymptotic coverage of ACI). For any $t_0 \geq 0$ and $T \geq 1$, the errors from the ACI iterates (7.2) satisfy

$$\left| \frac{1}{T} \sum_{t=t_0+1}^{t_0+T} e_t - \alpha \right| \leq \frac{1 + 2\eta}{T\eta}. \tag{7.3}$$

Proof. By (7.2), we have

$$\frac{1}{T} \sum_{t=t_0+1}^{t_0+T} (e_t - \alpha) = \frac{\alpha_{t_0+T+1} - \alpha_{t_0+1}}{T\eta}. \tag{7.4}$$

By Lemma 7.1, $\alpha_{t_0+1}, \alpha_{t_0+T+1} \in [-\eta, 1 + \eta]$, which immediately yields the result in (7.3). □

We can rewrite Theorem 7.2 to a limit form:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=t_0}^{t_0+T} e_t = \alpha. \tag{7.5}$$

Then it can be seen that over all time, the prediction bands adjusted by ACI has an exact coverage of $1 - \alpha$. This is a distribution-free result, since we do not pose any assumption on our sequence $\{(X_t, Y_t)\}$.

ACI as online gradient descent. The update formula (7.2) is an instance of online gradient descent applied to a proper convex function. To see this, we define

$$\beta_t = \sup\left\{\beta : Y_t \in C_t^{1-\beta}\right\}, \tag{7.6}$$

and

$$f_t(a) = \phi_{1-\alpha}(1 - \beta_t - (1-a)) = \phi_{1-\alpha}(a - \beta_t), \tag{7.7}$$

where $\phi_{1-\alpha}$ is the tilted ℓ_1-loss at quantile level $1-\alpha$: $\phi_{1-\alpha}(x) = \begin{cases} (1-\alpha)|x|, \ x \geq 0, \\ \alpha|x|, \ x < 0. \end{cases}$

We can straightforwardly calculate the subgradient of f_t:

$$\partial f_t(a) = \begin{cases} \{1-\alpha\}, & a > \beta_t, \\ [-\alpha, 1-\alpha], & a = \beta_t, \\ \{-\alpha\}, & a < \beta_t. \end{cases} \tag{7.8}$$

Furthermore, (7.6) implies $a > \beta_t \Leftrightarrow Y_t \notin C_t^{1-a} \Leftrightarrow e_t = 1$. Then by (7.8), we have $e_t - \alpha \in \partial f_t(\alpha_t)$. Therefore, (7.2) is the online gradient descent step for minimizing the convex function $\sum_{t=1}^{T} f_t(a)$ with arbitrarily large horizon T.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 2020.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. arXiv: 2202.13415, 2022.

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

Ryan J. Tibshirani. *Conformal Prediction*. Notes for Advanced Topics in Statistical Learning, Spring 2023.

Ryan J. Tibshirani. *Conformal Prediction Under Distribution Shift*. Notes for Advanced Topics in Statistical Learning, Spring 2023.