Ch3. Conformal Prediction Under Exchangeability

Hyerim Lee

Uncertainty Quantification Lab
Seoul National University

September 1, 2025

Table of contents

Contents

Exchangeable Sequence of Data Points

- We begin with an **exchangeable** sequence of data points:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}).$$

- $X_i \in \mathcal{X}$: feature, $Y_i \in \mathcal{Y}$: response.
- In prediction, Y_{n+1} is unobserved.
- Training data: $(X_1, Y_1), \ldots, (X_n, Y_n)$, Test feature: X_{n+1}.
- **Goal**: Construct prediction sets $C(X_{n+1}) \subseteq \mathcal{Y}$ with marginal coverage

$$\Pr\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha.$$

Score function

- $C(X_{n+1})$ Constructed using a **score function**:

$$s((x, y); D) \in \mathbb{R},$$

mapping data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and dataset $D \in (\mathcal{X} \times \mathcal{Y})^k$ to a real value.

- ex) Residual score:

$$s((x, y); D) = |y - \hat{f}(x; D)|$$

Symmetric Score Function

Definition 3.1: Symmetric score function

A score function s is **symmetric** if for any data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, any dataset $D \in (\mathcal{X} \times \mathcal{Y})^k$, and any permutation σ on $[k]$, we have

$$s((x, y); D) = s((x, y); D_\sigma).$$

- $\mathcal{D}_\sigma = \left((X_{\sigma(1)}, Y_{\sigma(1)}), \ldots, (X_{\sigma(n)}, Y_{\sigma(n)})\right)$
- For the residual score in (3.1), symmetry implies that the fitted model $\hat{f}(\cdot; D)$ is trained using a learning algorithm that is itself symmetric.

Contents

Full Conformal Prediction: Setup

- Training dataset:

$$\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$$

- Combined dataset with test:

$$\mathcal{D}_{n+1} = ((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}))$$

- For any hypothesized $y \in \mathcal{Y}$, define augmented dataset:

$$\mathcal{D}_{n+1}^y = ((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)).$$

- Scores (using score function s):

$$S_i^y = s((X_i, Y_i); \mathcal{D}_{n+1}^y), \quad i = 1, \ldots, n$$

$$S_{n+1}^y = s((X_{n+1}, y); \mathcal{D}_{n+1}^y).$$

Prediction Set Construction

- Define prediction set:

$$C(X_{n+1}) = \{y \in \mathcal{Y} : S_{n+1}^y \leq \hat{q}^y\},$$

where

$$\hat{q}^y = \text{Quantile}(S_1^y, \ldots, S_n^y; (1-\alpha)(1+1/n)).$$

- $\hat{q}^y$ is called the **conformal quantile**.
- Intuition: score가 매우 크다는 것은 data와 다르다는 것을 의미, 따라서y if S_{n+1}^y is too large compared to training scores를 제외하자

Theorem: Marginal Coverage Guarantee

Theorem 3.2

Suppose $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable and s is a symmetric score function. Then the prediction set $C(X_{n+1})$ satisfies

$$\Pr\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha.$$

- The guarantee is **marginal**.
- It holds on average over the distribution of the entire dataset
- See Chapter 4 for details.
- 증명을 4가지 방식으로 할 수 있음

Algorithm 3.3: Full Conformal Prediction

Procedure

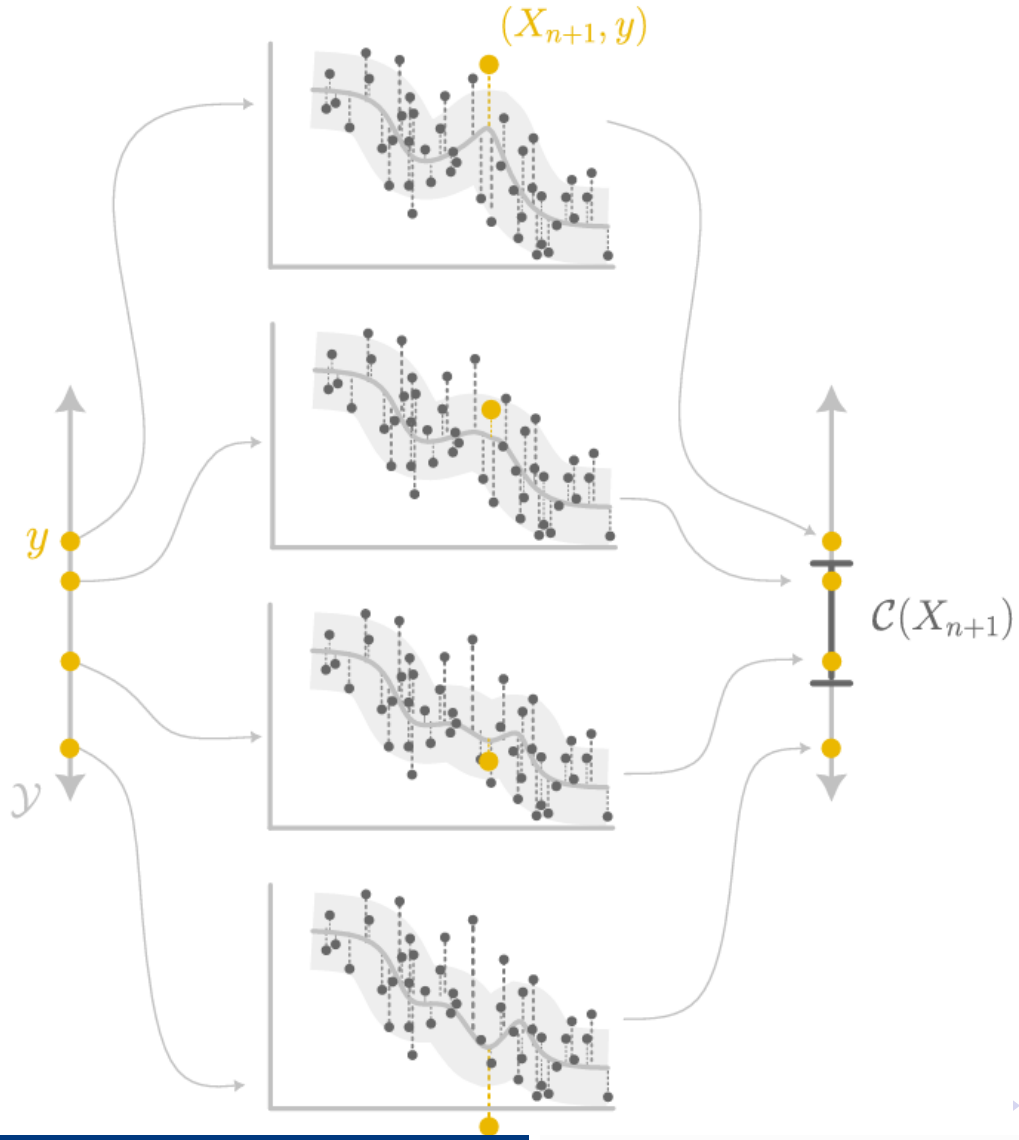
1. Input: training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, test point X_{n+1}, target coverage level $1 - \alpha$, score function s.
2. For each possible response $y \in \mathcal{Y}$:
   1. Compute $S_i^y = s((X_i, Y_i); \mathcal{D}_{n+1}^y),\ i = 1, \ldots, n.$
   2. Compute $S_{n+1}^y = s((X_{n+1}, y); \mathcal{D}_{n+1}^y).$
   3. Compute conformal quantile

$$\hat{q}^y = \text{Quantile}(S_1^y, \ldots, S_n^y;\ (1-\alpha)(1+1/n)).$$

3. Return prediction set:

$$C(X_{n+1}) = \{y \in \mathcal{Y} : S_{n+1}^y \leq \hat{q}^y\}.$$

Algorithm 3.3: Full Conformal Prediction



Contents

Idea behind the proof

- Naive procedure: a model trained on data D_n

$$s((X_1, Y_1); D_n), \ldots, s((X_n, Y_n); D_n), s((X_{n+1}, Y_{n+1}); D_n).$$

- Problem due to overfit:

$$s((X_{n+1}, Y_{n+1}); D_n) > s((X_i, Y_i); D_n) \text{ for } i = 1, ..., n$$

$\rightarrow$ low coverage

- Solution: include (X_{n+1}, Y_{n+1}) into training dataset.

$$S_i = s((X_i, Y_i); D_{n+1}), \quad i = 1, \ldots, n+1.$$

- Since Y_{n+1} is unknown, we add (X_i, y) instead of Y_{n+1}.

Proof of Theorem 3.2 I

Step 1: Reformulation of the Prediction Set

Lemma 3.4

For $v_1, \ldots, v_{n+1} \in \mathbb{R}$ and $t \in [0, 1]$,
$v_{n+1} \leq \text{Quantile}(v_1, \ldots, v_{n+1}; t)$
$\iff v_{n+1} \leq \text{Quantile}(v_1, \ldots, v_n; t(1 + 1/n))$.

Applying this to the conformal setting:

$y \in C(X_{n+1}) \iff S^y_{n+1} \leq \hat{q}^y$

$\iff S^y_{n+1} \leq \text{Quantile}(S^y_1, \ldots, S^y_n, S^y_{n+1}; 1 - \alpha)$.

$Y_{n+1} \in C(X_{n+1}) \iff S_{n+1} \leq \text{Quantile}(S_1, \ldots, S_{n+1}; 1 - \alpha)$,

where $S_i = s((X_i, Y_i); D_{n+1})$ for $i = 1, \ldots, n + 1$

Proof of Theorem 3.2 II

QnA

Q. 27페이지의 Theorem 3.2 증명에서, Step1 마지막 문장을 어떤 맥락에서 이해?

insight: the coverage guarantee will depend only on these scores $S_1, \ldots, S_{n+1}$, which are values obtained when the model is fitted with $y = Y_{n+1}$, i.e., on the dataset $\mathcal{D}_{n+1} = ((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}))$. In other words, while running full conformal prediction requires us to train the model using values $y \neq Y_{n+1}$ (since we must train using each possible value $y \in \mathcal{Y}$), for the theoretical coverage guarantee these resulting scores are irrelevant.

A. step 1에서 coverage event를 재정의.

$Y_{n+1} \in C(X_{n+1}) \iff S_{n+1} \leq \text{Quantile}(S_1, \ldots, S_{n+1}; 1-\alpha)$.

즉, 커버리지 보장은 $(S_1, \ldots, S_{n+1})$에만 의존. $(S_1, \ldots, S_{n+1})$는 실제 test 값인 Y_{n+1}이 주어졌을때 계산된 값. 따라서 실제 알고리즘을 돌릴 때는 모든 후보 $y \in \mathcal{Y}$를 넣어보고 스코어를 계산해야하지만, 이론적인 증명에는 실제 test 값인 Y_{n+1}일때의 스코어만 이용하면 충분

Proof of Theorem 3.2 III

Step 2: Exchangeability of Scores

We show that the scores $S_1, \ldots, S_{n+1}$ are exchangeable

$$(S_1, \ldots, S_{n+1}) \overset{d}{=} (S_{\sigma(1)}, \ldots, S_{\sigma(n+1)})$$

$$\Big(s((X_i, Y_i); D_{n+1})\Big)_{i \in [n+1]} \overset{d}{=} \Big(s((X_{\sigma(i)}, Y_{\sigma(i)}); (D_{n+1})_\sigma)\Big)_{i \in [n+1]}.$$

By symmetry of s, for any permutation σ,

$$S_i = s((X_i, Y_i); D_{n+1}) = s((X_i, Y_i); (D_{n+1})_\sigma)$$

$$S_{\sigma(i)} = s((X_{\sigma(i)}, Y_{\sigma(i)}); D_{n+1}) = s((X_{\sigma(i)}, Y_{\sigma(i)}); (D_{n+1})_\sigma).$$

Due to exchangeability of the data (F : fixed func.),

$$\mathcal{D}_{n+1} \overset{d}{=} (\mathcal{D}_{n+1})_\sigma \Rightarrow F(\mathcal{D}_{n+1}) \overset{d}{=} F((\mathcal{D}_{n+1})_\sigma),$$

Proof of Theorem 3.2 IV

$$F(\mathcal{D}_{n+1}) = \Big(s((X_i, Y_i); D_{n+1})\Big)_{i\in[n+1]}$$

$$F((\mathcal{D}_{n+1})_\sigma) = \Big(s((X_{\sigma(i)}, Y_{\sigma(i)}); D_{n+1})\Big)_{i\in[n+1]}$$

$$= \Big(s((X_{\sigma(i)}, Y_{\sigma(i)}); (D_{n+1})_\sigma)\Big)_{i\in[n+1]}$$

Therefore,

$$\Big(s((X_i, Y_i); D_{n+1})\Big)_{i\in[n+1]} \overset{d}{=} \Big(s((X_{\sigma(i)}, Y_{\sigma(i)}); (D_{n+1})_\sigma)\Big)_{i\in[n+1]}.$$

Step 3: Completing the Proof

From exchangeability and Fact 2.15(ii),

$$\Pr\Big(S_{n+1} \le \text{Quantile}(S_1, \ldots, S_{n+1}; \tau)\Big) \ge \tau, \quad \forall \tau \in [0, 1].$$

Proof of Lemma 3.4

If $t > \frac{n}{n+1}$, the result holds trivially since

$$\text{Quantile}(v_1, \ldots, v_{n+1}; t) = \max_i v_i \geq v_{n+1}.$$

Otherwise, let $v_{(1)} \leq \cdots \leq v_{(n)}$ and $v_{(1)} \leq \cdots \leq v_{(n+1)}$ be order statistics. Let $k = \lceil t(n+1) \rceil \in [n]$.
By 2.10,

$$\text{Quantile}(v_1, \ldots, v_{n+1}; t) = v_{(n+1;k)},$$

$$\text{Quantile}(v_1, \ldots, v_n; t(1+1/n)) = v_{(n;k)}.$$

Thus, we need to show

$$v_{n+1} \leq v_{(n+1;k)} \iff v_{n+1} \leq v_{(n;k)}.$$

Proof of Lemma 3.4 (continued)

By definition of order statistics:

$$v_{(n;k)} \geq v_{(n+1;k)}.$$

Therefore,

$$v_{n+1} \leq v_{(n+1;k)} \implies v_{n+1} \leq v_{(n;k)}.$$

Conversely, if $v_{n+1} > v_{(n+1;k)}$, then $v_{(n+1;k)} = v_{(n;k)}$, so we must have

$$v_{n+1} > v_{(n;k)}.$$

Hence,

$$v_{n+1} \leq v_{(n;k)} \implies v_{n+1} \leq v_{(n+1;k)}.$$

This completes the proof.

Contents

Split Conformal as a Special Case

- Split conformal prediction avoids retraining for each $y \in \mathcal{Y}$.
- Uses data splitting:
  - Pretraining set $\mathcal{D}_{pre}$: used for model training.
  - Calibration set $\mathcal{D}_n$: used to calibrate threshold for scores.
- $\mathcal{D}_{pre}$ and $\mathcal{D}_n$ are disjoint.
- $\mathcal{D}_n$ is called the **calibration set**.
- Split Conformal Prediction Set

$$C(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}\}$$

$$\hat{q} = \text{Quantile}(S_1, \ldots, S_n; (1-\alpha)(1+1/n)),$$

with $S_i = s(X_i, Y_i)$.

Algorithm 3.6: Split Conformal Prediction

Procedure

1. Input: pretraining set $\mathcal{D}_{pre}$, calibration data $(X_1, Y_1), \ldots, (X_n, Y_n)$, test point X_{n+1}, level $1 - \alpha$.
2. Train model on $\mathcal{D}_{pre}$ to obtain score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.
3. Compute scores on calibration set: $S_i = s(X_i, Y_i)$ for $i \in [n]$.
4. Compute quantile: $\hat{q} = \text{Quantile}(S_1, \ldots, S_n; (1-\alpha)(1+1/n))$.
5. Return prediction set:

$$C(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}\}.$$

The marginal coverage guarantee can apply to the split conformal prediction procedure as well.

Statistical vs Computational Efficiency

Full Conformal	Split Conformal
All data for both training + calibration	Disjoint sets for training + calibration
Retrains model for each y	One-time training only
Requires symmetric score s	Works for any pretrained s

Table: Comparison of full and split conformal methods

- Split conformal is more computationally efficient:
  - Train only once on $\mathcal{D}_{pre}$.
  - Full conformal requires retraining for each y.
- But full conformal is more statistically efficient.

QnA$_1$

Q. 데이터셋이 하나만 주어졌을 때 Full Conformal Prediction vs Split Conformal Prediction?

A. 데이터에 따라 다름. 보통 데이터가 많으면 split conformal을 사용하는게 컴퓨팅 효율적. 그러나 데이터가 작으면 모델 학습에 최대한 많이 활용해야하고, 모델 학습도 오래 걸리지 않으므로 full conformal이 나음. 일단 split으로 해보고 성능이 별로면 full을 해보는걸 추천. 아니면 full + split의 장점을 합쳐 만든 하이브리드 방식, cross-conformal prediction, 을 사용해보는 것도...

QnA$_2$

Q. Split Conformal 의 score function은 full conformal과 달리 symmetric이 아니어도 되는 이유

A. full conformal pred.은 D_n + hypothesized test point를 학습에 사용. 따라서 스코어 함수가 D_n에 의존. 즉,
$S_i^y = s\big((X_i, Y_i); D_{n+1}^y\big)$.
split conformal pred.은 D_{pre}를 학습에 사용하고 스코어 함수는 D_{pre}에만 의존 $S_i = s\big((X_i, Y_i); D_{\text{pre}}\big) = s\big(X_i, Y_i\big)$. 즉, 스코어 함수가 D_n의 순서에 영향 받지 않음. 따라서 이미 고정된 함수이므로 대칭성이 필요하지 않음

Contents

The conformal p-value

Definition 3.8 (The conformal p-value)

Given training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, a test feature X_{n+1}, and a score function s, the conformal p-value is

$$p^y = \frac{1 + \sum_{i=1}^{n} \mathbf{1}\{S_i^y \geq S_{n+1}^y\}}{n+1},$$

where

$$S_i^y = s((X_i, Y_i); D_{n+1}^y), \quad i \in [n], \quad S_{n+1}^y = s((X_{n+1}, y); D_{n+1}^y).$$

- p^y tests whether hypothesized point (X_{n+1}, y) is consistent with training data.
- If S_{n+1}^y is much larger than calibration scores, then p^y is small.

The conformal p-value

Proposition 3.9

The full conformal prediction set $C(X_{n+1})$ defined in (3.2) satisfies

$$C(X_{n+1}) = \{y \in \mathcal{Y} : p^y > \alpha\},$$

where p^y is the conformal p-value (Definition 3.8).

- The full conformal set can equivalently be constructed via the conformal p-value.
- p^y can be reinterpreted as the p-value for a permutation test
- The validity of permutation tests imply the marginal coverage.

Proof of Proposition 3.9

- By definition,

$$y \notin C(X_{n+1}) \iff S^y_{n+1} > \text{Quantile}(S^y_1, \ldots, S^y_n; (1-\alpha)(1+1/n)).$$

- By definition of the quantile of a finite list,

$$\text{Quantile}(S^y_1, \ldots, S^y_n; \tau) < t \iff \sum_{i=1}^{n} \mathbf{1}\{S^y_i < t\} \geq n\tau.$$

- Choosing $\tau = (1-\alpha)(1+1/n)$, $t = S^y_{n+1}$, we obtain

$$y \notin C(X_{n+1}) \iff \sum_{i=1}^{n} \mathbf{1}\{S^y_i < S^y_{n+1}\} \geq (1-\alpha)(n+1).$$

- This is equivalent to

$$p^y = \frac{1 + \sum_{i=1}^{n} \mathbf{1}\{S^y_i \geq S^y_{n+1}\}}{n+1} \leq \alpha.$$

Proof of Theorem 3.2 via permutation tests I

Permutation p-value

For data $Z_i = (X_i, Y_i)$ and test function $T : \mathcal{Z}^{n+1} \to \mathbb{R}$,

$$p_{\text{perm}} = \frac{1}{(n+1)!} \sum_{\sigma \in S_{n+1}} \mathbf{1}\left\{ T(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) \geq T(Z_1, \ldots, Z_{n+1}) \right\}.$$

- Thm 2.4. If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then

$$\Pr(p_{\text{perm}} \leq \alpha) \leq \alpha.$$

- By 2.4, we need to show that,

$$Y_{n+1} \in C(X_{n+1}) \iff p_{\text{perm}} > \alpha,$$

Proof of Theorem 3.2 via permutation tests II

- By 3.9, we need to show that $p_{\text{perm}} = p^{Y_{n+1}}$.

$$T(z_1, \ldots, z_{n+1}) = s(z_{n+1}; (z_1, \ldots, z_{n+1})).$$

By symmetry,

$$\begin{aligned} T(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) &= s(Z_{\sigma(n+1)}; (D_{n+1})_\sigma) \\ &= s(Z_{\sigma(n+1)}; D_{n+1}) = S_{\sigma(n+1)}. \end{aligned}$$

$$\begin{aligned} p_{\text{perm}} &= \frac{1}{(n+1)!} \sum_{\sigma \in S_{n+1}} \mathbf{1}\{S_{\sigma(n+1)} \geq S_{n+1}\} \\ &= \frac{1 + \sum_{i=1}^{n} \mathbf{1}\{S_i \geq S_{n+1}\}}{n+1} = p^{Y_{n+1}}. \end{aligned}$$

Proof of Theorem 3.2 via Empirical Distribution I

Let $h : \mathcal{Z} \to \mathbb{R}$ be any function. By Proposition 2.2, we have

$$Z_{n+1} \mid \hat{P}_{n+1} \sim \hat{P}_{n+1}.$$

This implies

$$h(Z_{n+1}) \mid \hat{P}_{n+1} \sim \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{h(Z_i)}.$$

By definition quantile,

$$\Pr\left(h(Z_{n+1}) \leq \text{Quantile}((h(Z_i))_{i \in [n+1]}; 1-\alpha) \,\middle|\, \hat{P}_{n+1} \right) \geq 1-\alpha.$$

Since this calculation is carried out conditionally on $\hat{P}_{n+1}$, we can take $h(z) = s(z; \hat{P}_{n+1})$

Proof of Theorem 3.2 via Empirical Distribution II

By symmetry, $S_i = s(Z_i; D_{n+1}) = s(Z_i; \hat{P}_{n+1})$ for each $i \in [n+1]$,

$$\Pr\left(s(Z_{n+1}; \hat{P}_{n+1}) \leq \text{Quantile}\left((s(Z_i; \hat{P}_{n+1}))_{i\in[n+1]}; 1-\alpha\right) \,\middle|\, \hat{P}_{n+1}\right)$$

$$= \Pr\left(Y_{n+1} \in C(X_{n+1}) \,\middle|\, \hat{P}_{n+1}\right) \geq 1-\alpha.$$

Marginalizing over $\hat{P}_{n+1}$, we have proved the claim.

Plug-in Estimate of the Error Rate

In split conformal, prediction set:

$$C(X_{n+1}; \lambda) = \{y : s(X_{n+1}, y) \leq \lambda\}.$$

Empirical miscoverage (using Calibration set):

$$\hat{R}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \notin C(X_i; \lambda)\} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{s(X_i, Y_i) > \lambda\}.$$

Population miscoverage:

$$R(\lambda) = \mathbb{E}_P[\mathbf{1}\{Y \notin C(X; \lambda)\}] = 1 - \Pr(Y \in C(X; \lambda)) = \Pr(s(X, Y) > \lambda).$$

Goal: choose smallest λ such that $R(\lambda) \leq \alpha$.

Adjusted Plug-in Threshold

Because $\hat{R}(\lambda)$ is noisy, require slightly stronger condition:

$$\hat{\lambda} = \inf\{\lambda : \hat{R}(\lambda) \leq \alpha'\}, \quad \alpha' = \alpha - \frac{1-\alpha}{n}.$$

Proposition 3.10

The split conformal prediction set $C(X_{n+1})$ defined in 3.8 satisfies

$$C(X_{n+1}) = \{y : s(X_{n+1}, y) \leq \hat{\lambda}\}$$

- Proof :

$$\hat{R}(\lambda) \leq \alpha' \iff \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{S_i \leq \lambda\} \geq (1-\alpha)(1+1/n).$$

Plugging this into the definition of $\hat{\lambda}$,

$$\hat{\lambda} = \hat{q}$$

Proof of Theorem 3.2 via plug-in estimate (Split Conformal) I

We will begin by defining an oracle threshold:

$$\tilde{\lambda} = \inf \left\{ \lambda \in \mathbb{R} : \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{S_i > \lambda\} \leq \alpha \right\}.$$

The quantity $\tilde{\lambda}$ depends on the full dataset $\mathcal{D}_{n+1}$, including both calibration and test data
Exchangeability of the data implies,

$$\Pr(S_{n+1} > \tilde{\lambda}) = \Pr(S_i > \tilde{\lambda}), \quad \forall i \in [n+1].$$

Proof of Theorem 3.2 via plug-in estimate (Split Conformal) II

Below, we show $\tilde{\lambda} \leq \hat{\lambda}$. Then,

$$\begin{aligned}
\Pr(Y_{n+1} \notin C(X_{n+1})) &= \Pr(S_{n+1} > \hat{\lambda}) \\
&\leq \Pr(S_{n+1} > \tilde{\lambda}) \\
&= \mathbb{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1} \mathbf{1}\{S_i > \tilde{\lambda}\}\right] \\
&\leq \alpha \text{ by definition of } \tilde{\lambda}
\end{aligned}$$

Proof of Theorem 3.2 via plug-in estimate (Split Conformal) III

To conclude, we show $\tilde{\lambda} \leq \hat{\lambda}$. For all $\lambda \in \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{S_i > \lambda\} \leq \alpha' \iff \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{S_i > \lambda\} + \frac{1}{n+1} \leq \alpha$$

$$\implies \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbf{1}\{S_i > \lambda\} \leq \alpha.$$

This means the infimum in the definition of $\tilde{\lambda}$ is taken over a smaller set than that in $\hat{\lambda}$, hence $\tilde{\lambda} \leq \hat{\lambda}$.

Contents

Theorem 3.11: Upper Bound on Coverage

Theorem 3.11

Under the conditions of Theorem 3.2,

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \leq 1 - \alpha + \frac{1}{n+1} + \epsilon_{\text{tie}},$$

where

$$\epsilon_{\text{tie}} = \mathbb{P}(\exists j \in [n],\ S_{n+1} = S_j)\,.$$

- ϵ_{tie}: probability of score ties between test point and training data.
- Ensures coverage is not much larger than $1 - \alpha$ (conformal not overly conservative).

QnA

Q. Thm 3.11에 대한 "the bound says that this set is not conservative on the scale of coverage" 라는 말을 어떻게 이해?

of the scores is unlikely to produce ties. Without making further assumptions on the model, however, this does not directly translate to a bound on the *size* of the prediction set $\mathcal{C}(X_{n+1})$. Instead, the bound says that this set is not conservative on the scale of coverage. For example, for a residual score s of the form $s((x, y); \mathcal{D}) = |y - \hat{f}(x; \mathcal{D})|$, a model $\hat{f}(\cdot; \mathcal{D})$ that is a very poor fit to the data distribution will necessarily lead to wide prediction intervals. However, this result is telling us that the prediction intervals are no wider than is needed to compensate for the errors in $\hat{f}(\cdot; \mathcal{D})$.

A. 3.11은 커버리지 차원에서 보수성이 심하지 않다고 주장. 즉, prediction interval은 모델이 안좋으면 넓을 수 있음. 다만, prediction interval이 넓은 것이 모델 에러 때문이지 cp 커버리지가 보수적이어서 넓은 것이 아님. 다시말해, prediction interval이 너무 넓어도 커버리지 자체는 약 $1-\alpha$ 정도이다. 만약 prediction interval이 너무 넓다면 그것은 cp 문제가 아니라, 본인의 모델이 문제임.

Proof of Theorem 3.11.

From Theorem 3.2,

$$Y_{n+1} \in C(X_{n+1}) \iff S_{n+1} \leq \text{Quantile}(S_1, \ldots, S_{n+1}; 1-\alpha).$$

Then,

$$Y_{n+1} \in C(X_{n+1}) \iff S_{n+1} \leq S_{(k)}, \quad k = \lceil (1-\alpha)(n+1) \rceil.$$

$$S_{n+1} \leq S_{(k)} \iff \text{either } S_{n+1} < S_{(k+1)} \text{ or } S_{n+1} = S_{(k)} = S_{(k+1)}.$$

$$\Pr(S_{n+1} \in C(X_{n+1})) \leq \Pr(S_{n+1} < S_{(k+1)}) + \epsilon_{\text{tie}}.$$

Finally, using Fact 2.15,

Assume $Z \in \mathbb{R}^n$ is exchangeable, and fix any $i \in [n]$. Then we have that

(i) For any $k \in [n]$, $\mathbb{P}(Z_i \leq Z_{(k)}) \geq k/n$ and $\mathbb{P}(Z_i < Z_{(k)}) \leq (k-1)/n$.

$$\Pr(S_{n+1} < S_{(k+1)}) \leq \frac{(k+1)-1}{n+1} = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1}.$$

Summary

- Introduced the **full and split conformal prediction**
  - Full conformal: statistically efficient
  - Split conformal: computationally efficient
- Proved the **marginal coverage guarantee**:

$$\Pr\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha.$$

  - Permutation test interpretation
  - Conditioning on empirical distribution
  - Plug-in estimate of error rate
- It provides **distribution-free** guarantees

QnA

Q.
크기 n_1인 dataset 1과 이것의 부분집합인 크기 n_2인 dataset 2가 있을때 $(n_1 > n_2)$,
1. Split conformal prediction을 각각 dataset 1, 2를 calibration set으로 하여 적용했을 때, 각각의 경우에 만들어지는 신뢰집합 사이에 포함관계가 존재한다고 말할 수 있을까요?
2. Split conformal prediction 대신 두 데이터셋 모두 full conformal prediction을 적용했을 때에도 비슷한 결론을 내릴 수 있을까요?

QnA

A. split의 경우, 그렇지 않다. (full conformal은 모르겠음)
그럼 n 감소 시 왜 wide interval? quantile 추정이 불안정 → 가끔 큰 임계값을 내놓기도. 그렇지만 n이 커지면 안정적인 임계값을 출력해서 일반적으로 interval이 작다.

예측집합 $C(x_{test})_1 = \{y : s(x_{test}, y) \le \hat{q}_{n1}\}$

$C(x_{test})_2 = \{y : s(x_{test}, y) \le \hat{q}_{n2}\}$

같은 test point 이니까 $s(x_{test}, y)$ 는 같음.

즉, 만약 $s(x,y) = |y - \hat{f}(x)|$ 이면

$|y - \hat{f}(x_{test})| \le \hat{q} \Leftrightarrow [\hat{f}(x_{test}) - \hat{q}, \hat{f}(x_{test}) + \hat{q}]$

따라서 만약 C_1이 C_2에 포함된다면,

$\hat{q}_1 = \text{Quantile}(s_1, \ldots, s_{n1}; (1-\alpha)(1+1/n_1))$

$\hat{q}_2 = \text{Quantile}(s_1, \ldots, s_{n2}; (1-\alpha)(1+1/n_2))$

$\hat{q}_{n1} \le \hat{q}_{n2}$ 이어야 하는데 이게 성립 X

QnA

```
[ ]  q_hat_dict1 = {
         name: compute_sr_qhat(val_preds, y_val, alpha=0.05)
         for name, val_preds in pred_dic_val.items()}
     q_hat_dict1
```

$n_1 = 3750$

```
{'MC-Dropout': array([3.01743017, 6.48573485, 1.61652161, 2.31367331]),
 'CVAE-Gaussian': array([2.50509479, 3.72556946, 0.91650464, 1.30045072]),
 'CVAE-SB': array([4.10133205, 6.71655918, 1.53649719, 3.16095638]),
 'CVAE-Softmax': array([3.93935405, 6.19837   , 3.85647105, 3.29218884]),
 'CVAE-Dirichlet': array([4.1959973 , 5.59601413, 1.96362816, 2.6124006 ]),
 'CVAE-InverseCDF': array([2.13607142, 3.3094246 , 1.12364295, 1.44209985])}
```

4번째 y

```
q_hat_dict_sub = {
    name: compute_sr_qhat(val_preds, y_val_sub, alpha=0.05)
    for name, val_preds in pred_dic_val_sub.items()}
q_hat_dict_sub
```

$n_2 = 2500$

```
{'MC-Dropout': array([2.9358327 , 6.79217748, 2.14701696, 2.18402785]),
 'CVAE-Gaussian': array([2.12237101, 3.73252273, 0.91713179, 1.37850648]),
 'CVAE-SB': array([3.43665358, 9.80354772, 1.44449948, 3.33086118]),
 'CVAE-Softmax': array([3.46428754, 7.11889529, 4.27690662, 3.52711324]),
 'CVAE-Dirichlet': array([4.22007581, 5.67843786, 1.4959645 , 3.27477257]),
 'CVAE-InverseCDF': array([1.9150402 , 3.35670377, 1.09884359, 1.48287605])}
```

Figure: studentized residual score을 이용해서 구한 임계값

QnA

Q. 실제 분석 사례 중에 symmetric가정 혹은 exchangable가정이 만족하지 않는 경우가 많이 있는지 궁금!

A. symmetric 가정은 잘 모르겠는데, 교환가능성 가정은 많이 깨질 것으로 예상. 예를 들어, 시계열 데이터나 공간 데이터, 혹은 train data와 test data 분포가 너무 다른 경우 등이 실제 상황에서 많이 발생. 따라서 이러한 가정이 깨지는 상황을 보완하기 위해 다양한 변형 conformal prediction 기법들이 활발히 연구 중...