A Model-Based Perspective on Conformal Prediction

Heejoon Byun

Uncertainty Quantification Lab
Seoul National University

September 29, 2025

Table of contents

Contents

Conformal vs. Model-Based Methods

- **Q.** Would using conformal prediction lead to worse performance than using a model-based method?
- **A.** Not much, if we choose our score function wisely!
- Under stronger modeling assumptions, we may achieve stronger guarantees for conformal prediction by designing good conformal scores.

Contents

Recipe for Convergence Guarantees

Let the data points be drawn i.i.d. from P.

1. Define the aim
2. Define the **oracle set**

$$\mathcal{C}^*(x) = \{y : s^*(x, y) \leq q^*\},$$

   where s^* is an **oracle score function** and q^* is the $(1-\alpha)$-quantile of the distribution of $s^*(X, Y)$ under $(X, Y) \sim P$.
3. Choose a pretrained conformal score function s_n that mimics the oracle score function
4. State a model assumption that enables an asymptotic optimality guarantee

Q1

- **Q.** 책에서 model이라는 단어가 의미하는 것이 score function 인가요, 아니면 x값을 입력하면 y 예측값을 반환하는 모델($\widehat{f}$)을 말하는 것인가요? 아니면 다른 것을 의미하는 것인가요?
- **A.** 교재에서 말하는 model은 예측 모델 $\widehat{f}$가 맞습니다. 다만, 이 교재에서는 어차피 예측 모델이 score function을 정의하는 데에 사용되는 도구에 불과하고, 따라서 score function의 성질은 예측 모델의 성질에 강하게 의존하기 때문에 score function에 대한 가정도 model assumption이라고 부르는 것 같습니다.

Asymptotic Optimality (Informal)

Theorem (Informal)

Let $\mathcal{C}_n(x)$ be the usual split conformal set,

$$\mathcal{C}_n(x) = \{y : s_n(x, y) \leq \hat{q}_n\},$$

Then, under appropriate regularity conditions, if $s_n \to s^*$ then

$$\hat{q}_n \to q^*, \text{ and } \mathcal{C}_n \to \mathcal{C}^*.$$

i.e., if our base model is consistent for the true model, then conformal prediction will converge to the oracle prediction set.

- Note that the index n is the size of the calibration set $\mathcal{D}_n$.
- Typically, we also need the size of the pretraining set $\mathcal{D}_{\mathrm{pre},n}$ to grow with n in order to ensure convergence of s_n.

Contents

Minimal set size & Marginal coverage

Aim.

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{X \sim P_X}[|\mathcal{C}(X)|] \\ \text{subject to} \quad & \mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}(X)) \geq 1 - \alpha. \end{aligned} \tag{1}$$

Oracle.

$$\mathcal{C}^*(x) = \{y : s^*(x, y) \leq q^*\},$$

where

$$s^*(x, y) = -\pi^*(y \mid x),$$

and

$$q^* = \inf\left\{q \in \mathbb{R} : \mathbb{P}_{(X,Y)\sim P}(s^*(X, Y) \leq q) \geq 1 - \alpha\right\}.$$

Minimal set size & Marginal coverage

Choosing the Score.

$$\mathcal{C}_n(x) = \{y \in \mathcal{Y} : s_n(x, y) \le \hat{q}_n\} = \{y \in \mathcal{Y} : \hat{\pi}_n(y \mid x) \ge -\hat{q}_n\},$$

Proposition

Assume that $\pi^*(Y \mid X)$ *is continuously distributed under* $(X, Y) \sim P$*. Then the following claim holds almost surely: if*

$$\mathbb{E}_{X \sim P_X}\left[\mathrm{d}_{\mathrm{TV}}\left(\pi^*(\cdot \mid X), \hat{\pi}_n(\cdot \mid X)\right)\right] \to 0,$$

then

$\limsup_{n\to\infty} \mathbb{E}_{X\sim P_X}[|\mathcal{C}_n(X)|] \le \mathbb{E}_{X\sim P_X}[|\mathcal{C}^*(X)|]$ and

$\liminf_{n\to\infty} \mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}_n(X)) \ge 1 - \alpha,$

i.e., $\mathcal{C}_n$ *is asymptotically optimal for the aim* (1).

Minimal set size & Conditional coverage

Aim.

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{X \sim P_X}[|\mathcal{C}(X)|] \\ \text{subject to} \quad & \mathbb{P}_{Y \sim P_{Y|X}}(Y \in \mathcal{C}(X) \mid X) \geq 1 - \alpha \text{ almost surely.} \end{aligned} \tag{2}$$

Oracle.

$$\mathcal{C}^*(x) = \{y : s^*(x, y) < 1 - \alpha\}.$$

where

$$s^*(x, y) = \sum_{y' \in \mathcal{Y}} \pi^*(y' \mid x) \cdot \mathbb{1}\left\{\pi^*(y' \mid x) > \pi^*(y \mid x)\right\},$$

which is the (conditional) probability captured by the set $\{y' : \pi^*(y' \mid x) > \pi^*(y \mid x)\}$, i.e., all labels that are *strictly more likely* than the label y (given features x).

Minimal set size & Conditional coverage

Choosing the Score.

$$s_n(x, y) = \sum_{y' \in \mathcal{Y}} \hat{\pi}_n(y' \mid x) \cdot \mathbb{1}\left\{\hat{\pi}_n(y' \mid x) > \hat{\pi}_n(y \mid x)\right\}.$$

Proposition

Under some mild assumptions, if

$$\mathbb{E}_{X \sim P_X}\left[\mathrm{d}_{\mathrm{TV}}\left(\pi^*(\cdot \mid X), \hat{\pi}_n(\cdot \mid X)\right)\right] \to 0$$

then

$$\limsup_{n \to \infty} \mathbb{E}_{X \sim P_X}[|\mathcal{C}_n(X)|] \leq \mathbb{E}_{X \sim P_X}[|\mathcal{C}^*(X)|] \text{ and}$$

$$\lim_{n \to \infty} \mathbb{P}_{X \sim P_X}\left(\mathbb{P}_{Y \sim P_{Y|X}}(Y \in \mathcal{C}_n(X) \mid X) \geq 1 - \alpha - \epsilon\right) = 1, \ \forall \epsilon > 0,$$

almost surely. i.e., $\mathcal{C}_n$ is asymptotically optimal for the aim (2).

Contents

Minimal set size & Marginal coverage

Aim.

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{X \sim P_X}[\mathrm{Leb}(\mathcal{C}(X))] \\ \text{subject to} \quad & \mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}(X)) \geq 1 - \alpha \end{aligned} \tag{3}$$

Oracle. $\mathcal{C}^*(x) = \{y : s^*(x, y) \leq q^*\}$, where $s^*(x, y) = -f^*(y \mid x)$ and $q^* = -\sup\{t \in \mathbb{R} : \mathbb{P}_{(X,Y)\sim P}(f^*(Y \mid X) \geq t) \geq 1 - \alpha\}$.

Minimal set size & Marginal coverage

Choosing the Score.

$$\mathcal{C}_n(x) = \{y \in \mathbb{R} : s_n(x, y) \leq \hat{q}_n\} = \{y \in \mathbb{R} : \hat{f}_n(y \mid x) \geq -\hat{q}_n\},$$

Proposition

Assume that the conditional density $f^*(Y \mid X)$ has a continuous distribution under $(X, Y) \sim P$. Furthermore, assume that $\sup_{(x,y)} f^*(y \mid x) < \infty$. Then the following claim holds almost surely: if

$$\mathbb{E}_{X \sim P_X}\left[\mathrm{d}_{\mathrm{TV}}\left(f^*(\cdot \mid X), \hat{f}_n(\cdot \mid X)\right)\right] \to 0$$

then

$\limsup_{n\to\infty} \mathbb{E}_{X\sim P_X}[\mathrm{Leb}(\mathcal{C}_n(X))] \leq \mathbb{E}_{X\sim P_X}[\mathrm{Leb}(\mathcal{C}^*(X))]$ and
$\liminf_{n\to\infty} \mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}_n(X)) \geq 1 - \alpha$,
i.e., $\mathcal{C}_n$ is asymptotically optimal for the aim (3).

Minimal set size & Equal-tailed conditional coverage

Aim.

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{X\sim P_X}[\mathrm{Leb}(\mathcal{C}(X))] \\ \text{subject to} \quad & \mathbb{P}_{Y\sim P_{Y|X}}(Y > \sup \mathcal{C}(X) \mid X) \le \alpha/2, \\ & \mathbb{P}_{Y\sim P_{Y|X}}(Y < \inf \mathcal{C}(X) \mid X) \le \alpha/2, \\ & \mathcal{C}(x) \text{ is an interval for all } x. \end{aligned} \tag{4}$$

Oracle.

$$\mathcal{C}^*(x) = [\tau^*(x;\alpha/2), \tau^*(x;1-\alpha/2)],$$

where we recall that $\tau^*(x;\beta)$ is the β-quantile of the conditional distribution of $Y \mid X = x$. This corresponds to $\mathcal{C}^*(x) = \{y \in \mathbb{R} : s^*(x,y) \le q^*\}$ where $q^* = 0$ and

$$s^*(x,y) = \max\{\tau^*(x;\alpha/2) - y, y - \tau^*(x;1-\alpha/2)\}.$$

Minimal set size & Equal-tailed conditional coverage

Choosing the Score.

$$s_n(x,y) = \max\{\hat{\tau}_n(x;\alpha/2) - y, y - \hat{\tau}_n(x;1-\alpha/2)\}$$
$$\mathcal{C}_n(x) = [\hat{\tau}_n(x;\alpha/2) - \hat{q}_n, \hat{\tau}_n(x;1-\alpha/2) + \hat{q}_n]$$

Proposition

Under some mild assumptions, the following holds almost surely: if $\mathbb{E}_{X\sim P_X}[|\hat{\tau}(X;\alpha/2) - \tau^*(X;\alpha/2)|] \to 0$ *and* $E_{X\sim P_X}[|\hat{\tau}(X;1-\alpha/2) - \tau^*(X;1-\alpha/2)|] \to 0$, *then*

$$\limsup_{n\to\infty} \mathbb{E}_{X\sim P_X}[\mathrm{Leb}(\mathcal{C}_n(X))] \le \mathbb{E}_{X\sim P_X}[\mathrm{Leb}(\mathcal{C}^*(X))],$$

$$\lim_{n\to\infty} \mathbb{P}_{X\sim P_X}\left(\mathbb{P}_{Y\sim P_{Y|X}}(Y > \sup \mathcal{C}_n(X) \mid X) \ge \alpha/2 + \epsilon\right) = 0,$$
$$\lim_{n\to\infty} \mathbb{P}_{X\sim P_X}\left(\mathbb{P}_{Y\sim P_{Y|X}}(Y < \inf \mathcal{C}_n(X) \mid X) \ge \alpha/2 + \epsilon\right) = 0$$

for all $\epsilon > 0$. *That is,* $\mathcal{C}_n$ *is asymptotically optimal for the aim* (4).

Contents

Settings

$$\underbrace{(X'_{n,1}, Y'_{n,1}), \ldots, (X'_{n,m_n}, Y'_{n,m_n})}_{\text{pretraining set } \mathcal{D}_{\mathrm{pre},n}}, \underbrace{(X_{n,1}, Y_{n,1}), \ldots, (X_{n,n}, Y_{n,n})}_{\text{calibration set } \mathcal{D}_n} \overset{\text{i.i.d.}}{\sim} P.$$

$$s_n : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \text{ is a function of } \mathcal{D}_{\mathrm{pre},n} = ((X'_{n,i}, Y'_{n,i}))_{i \in [m_n]}.$$

$$\hat{q}_n = \mathrm{Quantile}\left(s_n(X_{n,1}, Y_{n,1}), \ldots, s_n(X_{n,n}, Y_{n,n}); 1 - \alpha_n\right)$$

(where $1 - \alpha_n = (1 - \alpha)(1 + 1/n)$, as in our usual definition of the split conformal method).

$$\mathcal{C}_n(x) = \{y \in \mathcal{Y} : s_n(x, y) \le \hat{q}_n\}.$$

Weakly Converging Scores Imply Converging Quantiles

Theorem

Let $s^* : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any fixed score function, and define

$$q^* = \inf\{t : F_{P,s^*}(t) \geq 1-\alpha\} \quad \text{and} \quad q_+^* = \sup\{t : F_{P,s^*}(t) \leq 1-\alpha\}.$$

Then the following statement holds almost surely:

$$\text{If } s_n \xrightarrow{\text{CDF}} s^*, \ \text{ then } \ q^* \leq \liminf_{n\to\infty} \hat{q}_n \leq \limsup_{n\to\infty} \hat{q}_n \leq q_+^*.$$

If we also assume that $q^* = q_+^*$, then the following statement holds almost surely:

$$\text{If } s_n \xrightarrow{\text{CDF}} s^* \text{ then } \hat{q}_n \to q^*.$$

Proof Sketch

Proof of Theorem.

For each $n \geq 1$, define the empirical CDF of the calibration scores,

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{s_n(X_{n,i}, Y_{n,i}) \leq t\right\}, \ t \in \mathbb{R}.$$

We can observe that, by definition, at each $n \geq 1$ the value $\hat{q}_n$ is defined as the $(1-\alpha_n)$-quantile of this empirical CDF $\hat{F}_n$.

Step 1: a deterministic result for the empirical CDFs. First we prove that, for any $q \in \mathbb{R}$,

$$\text{if } \limsup_{n\to\infty} \hat{F}_n(q) < 1-\alpha \text{ then } \liminf_{n\to\infty} \hat{q}_n \geq q,$$

and

$$\text{if } \liminf_{n\to\infty} \hat{F}_n(q) > 1-\alpha \text{ then } \limsup_{n\to\infty} \hat{q}_n \leq q.$$

Proof Sketch

Proof of Theorem.

Step 2: refining the deterministic result. Now we prove that

$$\text{if } s_n \xrightarrow{\text{CDF}} s^* \text{ and } \|\hat{F}_n - F_{P,s_n}\|_\infty \to 0 \text{ then } \liminf_{n\to\infty} \hat{q}_n \geq q^*,$$

and,

$$\text{if } s_n \xrightarrow{\text{CDF}} s^* \text{ and } \|\hat{F}_n - F_{P,s_n}\|_\infty \to 0 \text{ then } \limsup_{n\to\infty} \hat{q}_n \leq q^*_+.$$

Step 3: almost sure convergence. Finally, we show that $\|\hat{F}_n - F_{P,s_n}\|_\infty \overset{\text{a.s.}}{\to} 0$ by applying the Dvoretzky–Kiefer–Wolfowitz inequality, which tells us that for each n and for any $\epsilon > 0$,

$$\mathbb{P}\left(\|\hat{F}_n - F_{P,s_n}\|_\infty \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

$\square$

$\mathbb{P}$-Converging Scores Imply Converging Sets

Theorem

Under the setting and notation of the previous theorem, assume also that

$$\mathbb{P}_{(X,Y)\sim P}(s^*(X,Y) = q_+^*) = 0. \tag{5}$$

Define the oracle prediction set

$$\mathcal{C}^*(x) = \{y \in \mathcal{Y} : s^*(x,y) \leq q^*\}.$$

Then the following statement holds almost surely:

$$\text{If } s_n \xrightarrow{P} s^* \ \text{ then } \ \mathbb{P}_{(X,Y)\sim P}\left(Y \in \mathcal{C}_n(X) \triangle \mathcal{C}^*(X)\right) \to 0.$$

Proof Sketch

Proof of Theorem 7.

It suffices to prove the following *deterministic* statement:
If $s_n \xrightarrow{P} s^*$ and $q^* \leq \liminf_{n\to\infty} \hat{q}_n \leq \limsup_{n\to\infty} \hat{q}_n \leq q^*_+$,
then $\mathbb{P}_{(X,Y)\sim P}\left(Y \in \mathcal{C}_n(X) \triangle \mathcal{C}^*(X)\right)\right) \to 0$.

First, fix any (x, y), and any $\epsilon > 0$. For sufficiently large n, it holds that $q^* - \epsilon < \hat{q}_n < q^*_+ + \epsilon$. Therefore we have

$$y \in \mathcal{C}_n(x) \triangle \mathcal{C}^*(x)$$

$\implies q^* - 2\epsilon < s^*(x, y) < q^*_+ + 2\epsilon$ or $|s_n(x, y) - s^*(x, y)| > \epsilon$. Thus,

$$\begin{aligned}
&\mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}_n(X) \triangle \mathcal{C}^*(X)) \\
&\qquad \leq \mathbb{P}_{(X,Y)\sim P}(q^* - 2\epsilon < s^*(X, Y) < q^*_+ + 2\epsilon) \\
&\qquad\qquad + \mathbb{P}_{(X,Y)\sim P}(|s_n(X, Y) - s^*(X, Y)| > \epsilon).
\end{aligned}$$

Proof for First Case Study (Classification)

Proof of Proposition 2.

Step 1: verifying condition (5). Since $\pi^*(Y \mid X)$ is assumed to have a continuous distribution under $(X, Y) \sim P$, the score $s^*(X, Y) = -\pi^*(Y \mid X)$ is therefore also continuously distributed, which immediately implies (5).

Step 2: verifying that $s_n \xrightarrow{P} s^*$. We calculate

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim P}[|s_n(X, Y) - s^*(X, Y)|] \\
&\leq \sup_{(x,y)} \pi^*(y \mid x) \cdot \mathbb{E}_{X\sim P_X}\left[2\mathrm{d}_{\mathrm{TV}}(\hat{\pi}_n(\cdot \mid X), \pi^*(\cdot \mid X))\right],
\end{aligned}$$

Therefore, if we assume $\mathbb{E}_{X\sim P_X}\left[\mathrm{d}_{\mathrm{TV}}\left(\pi^*(\cdot \mid X), \hat{\pi}_n(\cdot \mid X)\right)\right] \to 0$ as in the proposition, this implies $\mathbb{E}_{(X,Y)\sim P}[|s_n(X, Y) - s^*(X, Y)|] \to 0$, which in turn implies $s_n \xrightarrow{P} s^*$.

Proof for First Case Study (Classification)

Proof of Proposition 2.

Step 3: establishing asymptotic optimality. To verify asymptotic optimality of the set size, it suffices to show that $\mathbb{E}_{X\sim P_X}[|\mathcal{C}_n(X)\backslash\mathcal{C}^*(X)|] \to 0$. Defining $c_n = \inf_{(x,y):y\in\mathcal{C}_n(x)} \hat{\pi}_n(y \mid x)$, we can derive that

$$\begin{aligned}\mathbb{E}_{X\sim P_X}[|\mathcal{C}_n(X)\backslash\mathcal{C}^*|] \leq c_n^{-1}\big(&\mathbb{E}_{X\sim P_X}\left[\mathrm{d}_{\mathrm{TV}}(\hat{\pi}_n(\cdot \mid X), \pi^*(\cdot \mid X))\right] \\ &+ \mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}_n(X)\backslash\mathcal{C}^*(X))\big).\end{aligned}$$

Next, for the asymptotic coverage guarantee, since $\mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}^*(X)) \geq 1-\alpha$ by definition of the oracle $\mathcal{C}^*$,

$$\begin{aligned}\mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}_n(X)) &\geq 1-\alpha-\mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}^*(X)\backslash\mathcal{C}_n(X)) \\ &\geq 1-\alpha-\mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}^*(X)\triangle\mathcal{C}_n(X)).\end{aligned}$$

Contents

Double Robustness

- **Q.** Do model-based assumptions allow for conformal methods to perform well *even if exchangeability does not hold*?
- **A.** Yes. For some cases, conformal prediction offers coverage guarantees as long as *either* exchangeability holds, or if instead we can rely on model-based assumptions.

- For example, let $(X_1, Y_1), (X_2, Y_2), \dots$ be a time series of **identically distributed** data points and assume a **positive and bounded** conditional density $f^*(y \mid x)$. Let $\tau^*(x; \beta)$ denote the β-quantile of this conditional distribution.
- Suppose that we construct quantile estimates $\hat{\tau}_n(x; \alpha/2)$ and $\hat{\tau}_n(x; 1 - \alpha/2)$ using past data points $(X_1, Y_1), \dots, (X_{\lfloor n/2 \rfloor}, Y_{\lfloor n/2 \rfloor})$, and use the more recent data points $(X_{\lfloor n/2 \rfloor + 1}, Y_{\lfloor n/2 \rfloor + 1}), \dots, (X_n, Y_n)$ to define $\hat{q}_n$ and return the corresponding **CQR prediction set**,

$$\mathcal{C}_n(X_{n+1}) = [\hat{\tau}_n(X_{n+1}; \alpha/2) - \hat{q}_n, \hat{\tau}_n(X_{n+1}; 1 - \alpha/2) + \hat{q}_n] .$$

Strongly Mixing Stationary Time Series

Proposition

Under the above settings, if the time series is strongly mixing, i.e.,

$$\lim_{m\to\infty}\left\{\sup_{k\geq 1}\sup_{\substack{A\in\mathcal{A}_{\leq k}\\ A'\in\mathcal{A}_{\geq k+m}}}\left|\mathbb{P}(A\cap A')-\mathbb{P}(A)\mathbb{P}(A')\right|\right\}=0,$$

and if the quantile estimates $\hat{\tau}_n$ satisfy

$$\lim_{n\to\infty}\mathbb{E}\left[\int_{\mathcal{X}}\left|\hat{\tau}_n(x;\beta)-\tau^*(x;\beta)\right|\,\mathrm{d}P_X(x)\right]=0$$

for each $\beta\in\{\alpha/2, 1-\alpha/2\}$, then

$$\lim_{n\to\infty}\mathbb{E}\left[\left|\mathbb{P}\left(Y_{n+1}\in\mathcal{C}_n(X_{n+1})\mid X_{n+1}\right)-(1-\alpha)\right|\right]=0.$$

Q2

- **Q.** Proposition 5.9 의 Strongly mixing condition이 의미하는 바가 무엇인가요?
- **A.** 두 σ-field $\mathcal{A}, \mathcal{B}$ 사이의 strong mixing coefficient $\sup_{\substack{A \in \mathcal{A} \\ B \in \mathcal{B}}} \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right|$는 $\mathcal{A}$와 $\mathcal{B}$가 얼마나 종속되어 있는지를 나타내는 값으로 해석할 수 있습니다. 따라서 Strongly mixing condition은 직관적으로 time series에서 임의의 두 시점을 선택했을 때, 그 두 시점 사이의 시간차가 길수록 두 시점에서의 time series 값이 독립에 가까워진다는 것을 의미합니다.
  이 조건을 만족하는 확률과정으로는 iid data, block-independent time series 이외에 (적절한 조건 하에서의) stationary & invertible ARMA process, stationary GARCH process, irreducible & aperiodic positive recurrent Markov chain 등이 있습니다.

Proof Sketch

Proof of Proposition 8.

Let $Z = (X, Y) \sim P$ denote an independent data point, and let $Z_i = (X_i, Y_i)$ as usual. We will use the following notation:

$$\gamma_m = \sup_{k \geq 1} \mathrm{d}_{\mathrm{TV}}\big((Z_1, \ldots, Z_k, Z_{k+m}), (Z_1, \ldots, Z_k, Z)\big).$$

By the strongly mixing assumption, we must have $\lim_{m \to \infty} \gamma_m = 0$. Note that since $\hat{\tau}_n$ is trained on $(Z_1, \ldots, Z_{\lfloor n/2 \rfloor})$, we therefore have

$$\mathrm{d}_{\mathrm{TV}}\big((\hat{\tau}_n, Z_{\lfloor n/2 \rfloor + m}), (\hat{\tau}_n, Z)\big) \leq \gamma_m$$

for all n and all m. In other words, for large m (i.e., if $\gamma_m \approx 0$), the trained model $\hat{\tau}_n$ is nearly independent of the future data point. □

Proof Sketch

Proof of Proposition 8.

Step 1: show that convergence of $\hat{q}_n$ is sufficient. First we assume $\lim_{n\to\infty} \mathbb{P}(|\hat{q}_n| > \epsilon) = 0$ holds for any $\epsilon > 0$. Then

$$\mathbb{E}\left[\left|\mathbb{P}\left(Y_{n+1} \in \mathcal{C}_n(X_{n+1}) \mid X_{n+1}\right) - (1-\alpha)\right|\right] \leq$$
$$\mathbb{P}\left(|s_n(Z_{n+1}) - s^*(Z_{n+1})| > \epsilon\right) + \mathbb{P}\left(|s^*(Z_{n+1})| \leq 2\epsilon\right) + \mathbb{P}\left(|\hat{q}_n| > \epsilon\right).$$

Next, the definition of s^* implies

$$\mathbb{P}(|s^*(Z_{n+1})| \leq 2\epsilon) \leq 2 \cdot 4\epsilon \cdot \sup_{x,y} f^*(y \mid x),$$

Furthermore, by definition of s_n and s^*, we have

$$\begin{aligned} &\mathbb{P}\left(|s_n(Z_{n+1}) - s^*(Z_{n+1})| > \epsilon\right) \\ &\leq \sum_{\beta \in \{\alpha/2, 1-\alpha/2\}} \left(\mathbb{P}\left(|\hat{\tau}_n(X;\beta) - \tau^*(X;\beta)| > \epsilon\right) + \gamma_{\lceil n/2 \rceil + 1}\right). \end{aligned}$$

Proof Sketch

Proof of Proposition 8.

Step 2: prove convergence of $\hat{q}_n$. We will prove that, for any $\epsilon > 0$, $\mathbb{P}(\hat{q}_n \leq \epsilon) \to 1$. First, for each $\beta \in \{\alpha/2, 1-\alpha/2\}$, we have

$$\mathbb{P}\left(\frac{1}{\lceil n/2 \rceil} \sum_{i=\lfloor n/2 \rfloor+1}^{n} \mathbb{1}\left\{|\hat{\tau}_n(X_i;\beta) - \tau^*(X_i;\beta)| > \epsilon/2\right\} > \delta\right) \to 0 \tag{6}$$

Next, by the LLN for strongly mixing time series, we have

$$\mathbb{P}\left(\frac{1}{\lceil n/2 \rceil} \sum_{i=1}^{\lceil n/2 \rceil} \mathbb{1}\left\{Y_i \in \mathcal{C}^{*,\epsilon/2}(X_i)\right\} < 1 - \alpha' - \delta\right) \to 0 \tag{7}$$

for any $\delta > 0$, where $1 - \alpha' = \mathbb{P}_{(X,Y)\sim P}(Y \in \mathcal{C}^{*,\epsilon/2}(X))$, and

$$\mathcal{C}^{*,\epsilon/2}(x) = \left[\tau^*(x;\alpha/2) - \epsilon/2, \tau^*(x;1-\alpha/2) + \epsilon/2\right].$$

Proof Sketch

Proof of Proposition 8.

Next, by definition of the CQR score,

$$\mathbb{1}\left\{s_n(X_i, Y_i) \leq \epsilon\right\} \geq \mathbb{1}\left\{Y_i \in \mathcal{C}^{*,\epsilon/2}(X_i)\right\} - \sum_{\beta \in \{\alpha/2, 1-\alpha/2\}} \mathbb{1}\left\{|\hat{\tau}_n(x;\beta) - \tau^*(x;\beta)| > \epsilon/2\right\}$$

By (6) and (7), for any $\delta > 0$, we therefore have

$$\mathbb{P}\left(\frac{1}{\lceil n/2 \rceil} \sum_{i=\lfloor n/2 \rfloor+1}^{n} \mathbb{1}\left\{s_n(X_i, Y_i) \leq \epsilon\right\} \geq 1 - \alpha' - 3\delta\right) \to 1.$$

Choosing δ sufficiently small so that $1 - \alpha' - 3\delta > (1-\alpha)\left(1 + \frac{1}{\lceil n/2 \rceil}\right)$ implies $\mathbb{P}(\hat{q}_n \leq \epsilon) \to 1$. □

Contents

Q3

- **Q.** 이 챕터에서는 데이터셋을 Model Train용하고 Conformal Prediction을 테스트하는 셋으로 쪼개는데, 어느 비율로 쪼개는 것이 가장 좋은 (혹은 robust)한 성능을 가지게 되는 지 궁금합니다.
- **A.** "좋은 성능"을 어떻게 정의하냐에 따라 달라질 것 같습니다. 예를 들어 Training conditional coverage가 $1-\alpha$ 근처에 모여있기를 원한다면, 데이터가 연속적인 분포에서 iid로 추출되었을 때 Training conditional coverage가 Beta 분포를 따르는 것을 이용하여 (Ch4 참고) 해당 Beta 분포의 $\delta/2, 1-\delta/2$ 번째 분위수가 각각 $1-\alpha \pm \epsilon$에 오게끔 calibration set size를 먼저 정한 후, 나머지를 training set으로 사용하여 원하는 바를 이룰 수 있습니다.

Q4

- **Q.** 5.5에서 Model-based assumption으로 double-robustnesss type guarantee가 보증된다 하던데 이게 오히려 assumption만 강화되는 안좋은 케이스가 되는거 아닌가요...? model assumption이 들어가는 것으로 더 좋아지는지 사실 잘 이해가 되진 않습니다. 이에 대한 장점이 있을까요?
- **A.** 대략적으로 설명하자면 Double-Robustness type guarantee란 어떤 방법론에 대해 두 개의 가정 중 하나만 성립해도 좋은 이론적 결과들이 성립하게 되는 성질을 말합니다. 5.5단원에서 다루는 CQR prediction set의 경우에 대입해서 보면, model-based assumption 없이 exchangibility만 있으면 finite sample marginal coverage가 보장되는 한편, 반대로 exchangibility가 어느 정도 깨져도 (stationary & strongly mixing) 적절한 model-based assumption 하에서 asymptotic test-conditional coverage가 보장되는 것을 확인할 수 있습니다.

Q5

- **Q.** p.66에서 (...)It is straightforward to prove that the solution to the above optimization problem then has the form $\mathcal{C}^*(x) = \{y : \pi^*(y \mid x) \geq t^*\}$, for some appropriate value t^*.(...) 라고 이야기하는데, 여기의 두번째 ∼ 세번째 문장과 관련해서 좀 더 자세하게 설명해 주실 수 있나요?
- **A.** 해당 단락에서 언급하는 제약 조건이 있는 최적화 문제에 대한 라그랑지안을 구해보면
$\mathcal{L}(\mathcal{C}, \lambda) = \sum_y \mathbb{1}\{y \in \mathcal{C}(x)\}(1 - \lambda\pi^*(y \mid x)) \quad (\lambda \geq 0)$이고, 이 값을 최소화 시키려면 $\mathcal{C}(x) \supset \{\pi^*(y \mid x) > 1/\lambda\}$여야 한다는 사실을 이용하여 $\mathcal{C}^*(x)$의 형태를 직관적으로 유추할 수 있습니다.
엄밀한 증명을 위해서는 Neyman-Pearson lemma나 베이즈 통계학에서 HPD region이 가장 작은 credible region인 것을 증명할 때와 마찬가지로 $\mathcal{L}(\mathcal{C}, \lambda) - \mathcal{L}(\mathcal{C}^*, \lambda) \geq 0$임을 보이면 됩니다.

Q6

- **Q.** p.68 에서 $\mathcal{C}^*(x) = \{y : \pi^*(y \mid x) \geq t^*(x)\}$이 $\mathcal{C}^*(x) = \{y : s^*(x, y) < 1 - \alpha\}$으로 동등하게 표현된다는데, 왜 그런지 잘 모르겠습니다.
- **A.** x를 고정하였을 때 $\pi^*(Y \mid X = x)$의 cdf를 $F_{\pi^*_x}$라 하면, $t^*(x) = \sup\{t : F_{\pi^*_x}(t-) \leq \alpha\}$를 만족해야 하고, $s^*(x, y) = 1 - F_{\pi^*_x}(\pi^*(y \mid x))$로 나타낼 수 있으므로

$$\begin{aligned}
&\pi^*(y \mid x) \geq t^*(x) \\
&\iff \forall \epsilon > 0 : \exists \delta > 0 \text{ s.t. } F_{\pi^*_x}((\pi^*(y \mid x) + \epsilon)-) \geq \alpha + \delta \\
&\iff \forall \epsilon > 0 : \exists \delta > 0 \text{ s.t. } F_{\pi^*_x}(\pi^*(y \mid x) + \epsilon) \geq \alpha + \delta \\
&\iff F_{\pi^*_x}(\pi^*(y \mid x)) > \alpha \text{ or } \pi^*(y \mid x) = t^*(x) \\
&\iff s^*(x, y) < 1 - \alpha \text{ or } \pi^*(y \mid x) = t^*(x)
\end{aligned}$$

입니다. 그러나 Proposition 5.3의 조건 하에서 $\pi^*(y \mid x) = t^*(x)$는 measure zero event이므로 이 부분을 제거하여도 주어진 문제의 해가 됩니다.

Thank you!