# Ch9. Additional Results in Conformal Prediction

Hyerim Lee

Uncertainty Quantification Lab
Seoul National University

December 1, 2025

# Table of contents

# Contents

# Conformal prediction with randomization

- We have presented full conformal prediction as a deterministic algorithm—the output $C(X_{n+1})$ is a deterministic function of the training data and $X_{n+1}$
- In practice, model fitting may use **randomized algorithms** (e.g., SGD)
- To guarantee marginal coverage, it is necessary to modify the definition of the **symmetric score function**
- Overcoverage can occur due to ties or when $(1-\alpha)(n+1)$ is not an integer.
- Adding randomization to break ties leads to **exact** $1-\alpha$ **coverage**

# Allowing for randomization in the score function

## Definition (Symmetric randomized score function)

A randomized score function $s$ is symmetric if for any fixed dataset $D$, and any fixed test points $(x'_1, y'_1), \ldots, (x'_M, y'_M)$, and any permutation $\sigma$,

$$\Big(s\big((x'_m, y'_m); D, \xi\big)\Big)_{m\in[M]} \overset{d}{=} \Big(s\big((x'_m, y'_m); D_\sigma, \xi\big)\Big)_{m\in[M]} \tag{1}$$

with respect to drawing the random seed as $\xi \sim \mathrm{Unif}[0, 1]$.

- The score function may involve **randomness**, e.g., models trained with **SGD**.
- Then the score is no longer deterministic:

$$s((x, y); \mathcal{D}, \xi), \quad \xi \in [0, 1].$$

- The requirement of a **symmetric score** must be modified.

# Allowing for randomization in the score function

- This is weaker than $s\big((x'_m, y'_m); D, \xi\big) = s\big((x'_m, y'_m); D_\sigma, \xi\big)$
- The scores $S_1, \ldots, S_{n+1}$ are exchangeable even in this randomized setting

$$\begin{aligned}
(S_{\sigma(1)}, \ldots, S_{\sigma(n+1)}) &= \big(s((X_{\sigma(i)}, Y_{\sigma(i)}); \mathcal{D}, \xi)\big)_{i\in[n+1]} \\
&\overset{d}{=} \big(s((X_{\sigma(i)}, Y_{\sigma(i)}); \mathcal{D}_\sigma, \xi)\big)_{i\in[n+1]} \\
&\overset{d}{=} \big(s((X_i, Y_i); \mathcal{D}_\sigma, \xi)\big)_{i\in[n+1]} \\
&\overset{d}{=} \big(s((X_i, Y_i); \mathcal{D}, \xi)\big)_{i\in[n+1]} \\
&= (S_1, \ldots, S_{n+1}).
\end{aligned}$$

- Therefore, with a **symmetric randomized score function**,

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

# Randomized calibration for exact coverage

- Standard p-value and prediction set:

$$p^y = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbf{1}\{S_i^y \geq S_{n+1}^y\}, \quad C(X_{n+1}) = \{y \in \mathcal{Y} : p^y > \alpha\}.$$

- **Overcoverage** can occur when $(1-\alpha)(n+1) \notin \mathbb{Z}$ or when ties occur (see 3.11).
- If scores are distinct, $p^{Y_{n+1}}$ is uniform on the grid $\left\{\frac{1}{n+1}, \ldots, 1\right\}$ (see Theorem 8.2)

$$P(p^{Y_{n+1}} > \alpha) = \frac{k}{n+1}, \quad k = \lceil (1-\alpha)(n+1) \rceil$$

- Idea : the distribution of $p^{Y_{n+1}}$ **smoothing!**

$$\{\frac{1}{n+1}, \ldots, \frac{n}{n+1}, 1\} \Rightarrow \text{Unif}[0,1]$$

# Smoothed p-value and exact coverage

**Define smoothed p-value (randomized):**

$$p^y(\xi) = \frac{\sum_{i=1}^{n+1} \mathbf{1}\{S_i^y > S_{n+1}^y\} + \xi \cdot \sum_{i=1}^{n+1} \mathbf{1}\{S_i^y = S_{n+1}^y\}}{n+1},$$

$$\xi \sim \mathrm{Unif}[0,1].$$

**Randomized conformal prediction set:**

$$\mathcal{C}(X_{n+1}) = \{y : p^y(\xi) > \alpha\}.$$

**Theorem (9.2)**

*Suppose that $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable, $s$ is a symmetric score function. Let $\xi \sim \mathrm{Unif}[0,1]$ be drawn independently of the data. Then the randomized conformal prediction set*

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

# Proof of Theorem 9.2

## Lemma (9.3)

*Let $S_1, \ldots, S_{n+1}$ be exchangeable, and let $\xi \sim \text{Unif}[0,1]$ be drawn independently from $S_1, \ldots, S_{n+1}$. Define*

$$p(\xi) = \frac{\sum_{i=1}^{n+1} \mathbf{1}\{S_i > S_{n+1}\} + \xi \cdot \sum_{i=1}^{n+1} \mathbf{1}\{S_i = S_{n+1}\}}{n+1}.$$

*Then $P(p(\xi) \leq \tau) = \tau$ for all $\tau \in [0,1]$.*

(Proof) Define

$$p_j(\xi) = \frac{\sum_{i=1}^{n+1} \mathbf{1}\{S_i > S_j\} + \xi \cdot \sum_{i=1}^{n+1} \mathbf{1}\{S_i = S_j\}}{n+1}$$

for each $j \in [n+1]$. By score exchangeability, we have $p(\xi) \stackrel{d}{=} p_j(\xi)$ for each $j$.

# Proof of Lemma 9.3

$$P\big(p(\xi) \le \alpha\big) = \frac{1}{n+1} \sum_{j \in [n+1]} P\big(p_j(\xi) \le \alpha\big)$$

$$= \mathbb{E}\left[ \frac{1}{n+1} \sum_{j \in [n+1]} P\big(p_j(\xi) \le \alpha \mid S_1, \ldots, S_{n+1}\big) \right].$$

Next, define

$$q = \mathit{Quantile}(S_1, \ldots, S_{n+1}; 1-\tau) = S_{(\lceil (1-\tau)(n+1) \rceil)},$$

$$N_+ = \sum_{j \in [n+1]} \mathbf{1}\{S_j > q\}, \quad N_= = \sum_{j \in [n+1]} \mathbf{1}\{S_j = q\}.$$

Then,

$$N_+ \le \tau(n+1), \quad N_= + N_+ \ge 1 + \tau(n+1).$$

# Proof of Lemma 9.3

For any $j$ with $S_j < q$,

$$p_j(\xi) \geq \frac{\sum_{i=1}^{n+1} \mathbf{1}\{S_i > S_j\}}{n+1} \geq \frac{N_= + N_+}{n+1} > \tau,$$

For any $j$ with $S_j > q$,

$$p_j(\xi) \leq \frac{\sum_{i=1}^{n+1} \mathbf{1}\{S_i \geq S_j\}}{n+1} \leq \frac{N_+}{n+1} \leq \tau.$$

If $S_j = q$,

$$p_j(\xi) = \frac{\sum_{i=1}^{n+1} \mathbf{1}\{S_i > q\} + \xi \cdot \sum_{i=1}^{n+1} \mathbf{1}\{S_i = q\}}{n+1} = \frac{N_+ + \xi N_=}{n+1},$$

$$p_j(\xi) \leq \tau \iff \xi \leq \frac{\tau(n+1) - N_+}{N_=}.$$

# Proof of Lemma 9.3

Finally,

$$P(p_j(\xi) \leq \tau \mid S_1, \ldots, S_{n+1}) = \mathbf{1}\{S_j > q\} + \mathbf{1}\{S_j = q\} \cdot \frac{\tau(n+1) - N_+}{N_=}.$$

Returning to our work above, we have

$$\begin{aligned} P(p(\xi) \leq \tau) &= E\left[\frac{1}{n+1}\sum_{j=1}^{n+1} P(p_j(\xi) \leq \tau \mid S_1, \ldots, S_{n+1}))\right] \\ &= \frac{1}{n+1} E\left[N_+ + N_= \cdot \frac{\tau(n+1) - N_+}{N_=}\right] = \tau, \end{aligned}$$

# Contents

# Computational shortcuts for full conformal prediction

- Full conformal prediction requires recomputing the score $s((x, y); \mathcal{D}^{y}_{n+1})$ for every possible test value $y$.
- If $\mathcal{Y}$ is continuous (e.g. $\mathbb{R}$), this becomes computationally infeasible.
- Thus, we derive shortcuts.

**Special case: linear regression.**

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

where

$$\hat{f}(x; \mathcal{D}) = x^{\top}\hat{\beta}, \qquad \hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{m} (y_i - x_i^{\top}\beta)^2.$$

# Linear regression: Explicit full conformal set

## Proposition 9.4 (Linear regression case)

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$. Let $X_{[n+1]} \in \mathbb{R}^{(n+1)\times d}$ denote the matrix whose $i$th row is $X_i$. Assume $X_{[n+1]}$ has full column rank and the score function is the least-squares residual.

Then the full conformal prediction set $C(X_{n+1})$ is

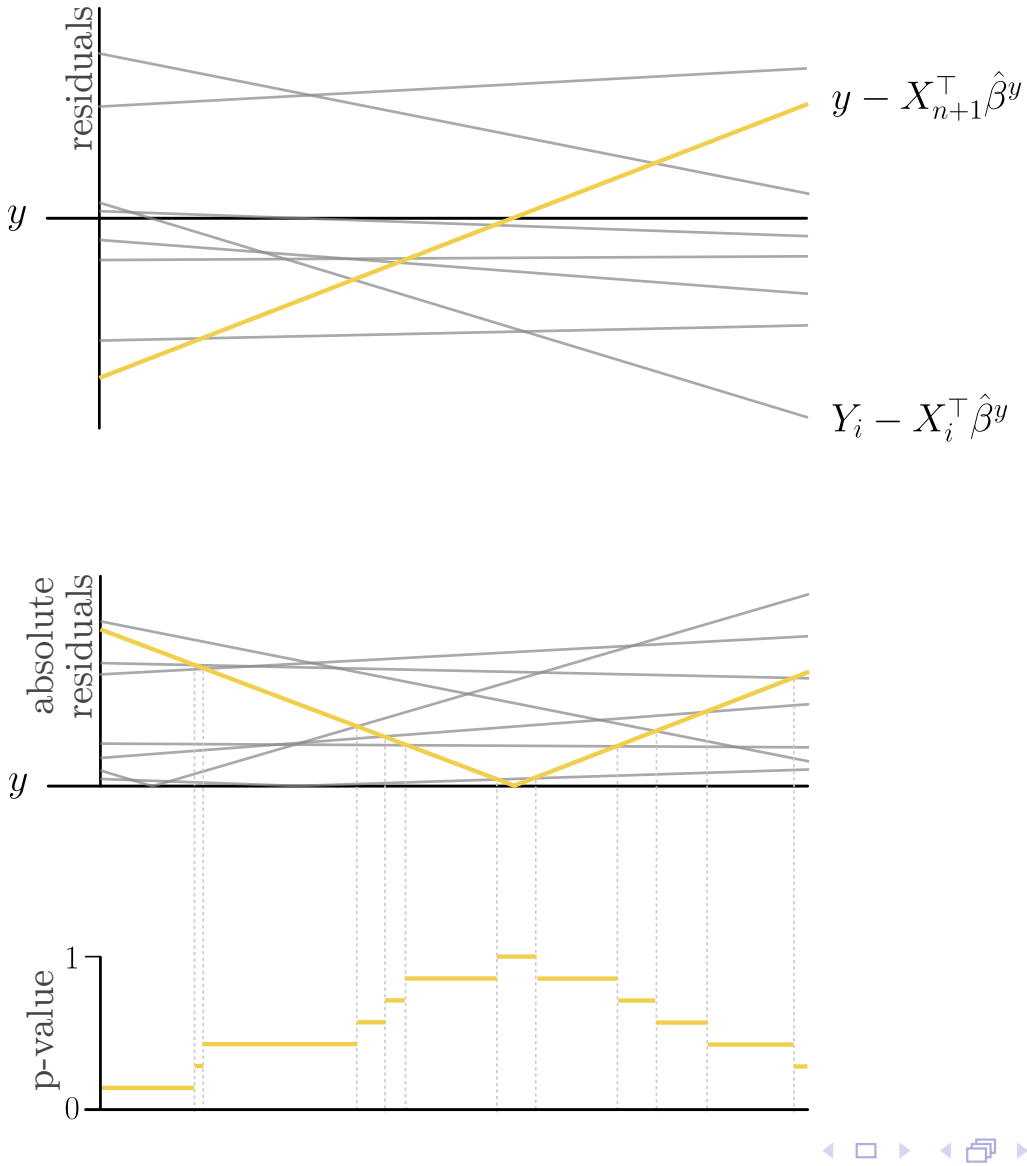
$$\left\{ y : \frac{1 + \sum_{i=1}^{n} \mathbf{1}(|y(-b_i) - (a_i - Y_i)| \geq |y(1 - b_{n+1}) - a_{n+1}|)}{n+1} > \alpha \right\}$$

Where

$$a_i = \sum_{j=1}^{n} H_{ij} Y_j, \qquad b_i = H_{i,n+1},$$

$$H = X_{[n+1]} \left( X_{[n+1]}^\top X_{[n+1]} \right)^{-1} X_{[n+1]}^\top.$$

# Visualization of the conformal p-value



# Proof of proposition 9.4

OLS solution:

$$\hat{\beta}^y = (X_{[n+1]}^\top X_{[n+1]})^{-1} X_{[n+1]}^\top \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ y \end{pmatrix}.$$

For each $i \in [n+1]$,

$$X_i^\top \hat{\beta}^y = X_i^\top \left( X_{[n+1]}^\top X_{[n+1]} \right)^{-1} X_{[n+1]}^\top \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ y \end{pmatrix}$$

$$= \sum_{j=1}^{n} H_{ij}\, Y_j \; + \; H_{i,n+1}\, y = a_i + b_i\, y \Rightarrow \text{linear!}$$

# Proof of proposition 9.4

We can then calculate the scores

$$S_i^y = s((X_i, Y_i); {}^y_{n+1}) = |Y_i - X_i^\top \hat{\beta}^y| = |y \cdot (-b_i) - (a_i - Y_i)|,$$

for $i = 1, \ldots, n$, and

$$S_{n+1}^y = s((X_{n+1}, y); {}^y_{n+1}) = |y - X_{n+1}^\top \hat{\beta}^y| = |y \cdot (1 - b_{n+1}) - a_{n+1}|.$$

The conformal p-value is

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

$$= \frac{1 + \sum_{i \in [n]} \mathbf{1}\{|y(-b_i) - (a_i - Y_i)| \geq |y(1 - b_{n+1}) - a_{n+1}|\}}{n+1}.$$

This completes the proof.

# QnA

Q. data의 오차가 정규분포 등 특정 분포를 따른다고 가정할 수 있다면, 선형회귀를 했을 때 전통적인 방식으로도 예측구간을 구할 수 있는데, 이러한 경우에 conformal prediction으로 구한 구간과 전통적 방식으로 구한 구간 중 어떤 것을 사용하는 것이 좋을까요?

A.

$$\hat{y}_{n+1} \pm t_{n-2,1-\alpha/2} \times \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

만약 오차가 정규분포를 따르고, 등분산성 및 선형성 등의 가정이 만족한다면 당연히 더 간단한 전통적 방식의 예측구간을 사용하는 것이 적절. cp로 추정한 구간은 더 넓을 수 있기 때문이다 (marginal coverage $\geq 1 - \alpha$). 하지만 딥러닝 모델과 같이 비선형 모델을 사용하는 경우, 정규오차를 가정해도 t 분포를 활용해서 예측구간을 만들기 어려움. 이런 경우는 cp 사용

# Lasso regression and residual score

**Special case: Lasso regression**

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{2} \sum_{i=1}^{m} (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_1 \right\}.$$

**Key fact:**

$$y \mapsto \hat{\beta}^y \quad \text{and} \quad y \mapsto \hat{f}(X_i; \mathcal{D}_{n+1}^y)$$

are **piecewise linear**.

**Why?**

For each $y$, define support and signs

$$\hat{I}^y = \{j : \hat{\beta}_j^y \neq 0\}, \qquad \hat{\gamma}^y = \text{sign}(\hat{\beta}_{\hat{I}^y}^y).$$

# Support/sign partition and linear representation

For fixed support and signs,

$$(\hat{\beta}^y)_{\hat{I}^y} = \left(X^\top_{[n+1],\hat{I}^y} X_{[n+1],\hat{I}^y}\right)^{-1} \left(X^\top_{[n+1],\hat{I}^y} (Y_1, \ldots, Y_n, y)^\top - \lambda \hat{\gamma}^y\right).$$

Thus the fitted value is linear in $y$:

$$\hat{f}(X_i; \mathcal{D}^y_{n+1}) = a_i(\hat{I}^y, \hat{\gamma}^y) + b_i(\hat{I}^y, \hat{\gamma}^y)\, y.$$

Since $(\hat{I}^y, \hat{\gamma}^y)$ is **piecewise constant**, the prediction is **piecewise linear** over intervals

$$\mathbb{R} = I_1 \cup \cdots \cup I_R.$$

# Computing the full conformal prediction set

On each interval $I_r$,

$$\hat{f}(X_i; \mathcal{D}^y_{n+1}) = a_i(I_r, \gamma_r) + b_i(I_r, \gamma_r)\, y.$$

Thus the conformal set within $\mathcal{C}(X_{n+1}) \cap I_r$ is

$$\left\{ y \in I_r : \frac{1 + \sum_{i=1}^{n} \mathbf{1}\left\{ \left| y(-b_i) - (a_i - Y_i) \right| \geq \left| y(1 - b_{n+1}) - a_{n+1} \right| \right\}}{n+1} > \alpha \right\}.$$

**Algorithm:**

1. Pick $y_1$, compute $\hat{\beta}^{y_1}$, obtain interval $I_1((\hat{I}^y, \hat{\gamma}^y) = (I_1, \gamma_1)$ for all $y \in I_1$).
2. Compute $\mathcal{C}(X_{n+1}) \cap I_1$.
3. Pick $y_2 \notin I_1$, compute new interval $I_2$, and repeat.
4. After $R$ intervals,

$$\mathcal{C}(X_{n+1}) = \bigcup_{r=1}^{R} \left( \mathcal{C}(X_{n+1}) \cap I_r \right).$$

# QnA

**Q.** LASSO regression에서 support 와 sign이 고정되어 있으면 beta is linear w.r.t y 가 유지된다는 내용을 부연 설명해주시면 좋겠습니다. (with Karush-Kuhn-Tucker conditions). 추가로, 그렇다면 ridge regression은 절대 linearity 가 유지되지 않으니 모든 y에 대한 full CP를 진행해야 하는 건가요? 다른 대안은 없나요?

**A.**

$$\tilde{Y} := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ y \end{pmatrix} \qquad \hat{\beta}^y = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \| \tilde{Y} - X_{[n+1]} \beta \|_2^2 + \lambda \|\beta\|_1 \right\} \quad \leftarrow \text{Convex 문제}$$

$$\frac{1}{2} \| \tilde{Y} - X_{[n+1]} \beta \|_2^2 + \lambda \|\beta\|_1 \xrightarrow{\text{미분}} X_{[n+1]}' \left( X_{[n+1]} \beta^y - \tilde{Y} \right) + \lambda z, \quad z = \partial \|\beta\|_1$$

By KKT 조건의 Stationarity 에 의해 최적해 $\hat{\beta}^y$ 의 필요충분 조건은

$$\Leftrightarrow X_{[n+1]}' \left( X_{[n+1]} \hat{\beta}^y - \tilde{Y} \right) + \lambda z = 0, \quad z = \partial \|\hat{\beta}\|_1 \Rightarrow z_j = \begin{cases} \hat{\beta}_j^y \neq 0 \Rightarrow \operatorname{sign}(\hat{\beta}_j^y) = \sigma_j \\ \hat{\beta}_j^y = 0 \Rightarrow [-1, 1] \end{cases}$$

Support $\hat{I}$ 와 sign $\hat{\sigma}$ 로 고정 시,

For $j \in \hat{I}$, $X_{[n+1]}' \left( X_{[n+1]} \hat{\beta}^y - \tilde{Y} \right) + \lambda z = 0$

$$\Rightarrow X_{[n+1],\hat{I}}' \left( X_{[n+1],\hat{I}} (\hat{\beta}^y)_{\hat{I}} - \tilde{Y} \right) + \lambda \hat{\sigma} = 0$$

$$\Rightarrow (\hat{\beta}^y)_{\hat{I}} = \left( X_{[n+1],\hat{I}}' X_{[n+1],\hat{I}} \right)^{-1} \left( X_{[n+1],\hat{I}}' \tilde{Y} - \lambda \hat{\sigma} \right)$$

$$\Rightarrow (\hat{\beta}^y)_{\hat{I}} = \left( X_{[n+1],\hat{I}}' X_{[n+1],\hat{I}} \right)^{-1} \left( X_{[n+1],\hat{I}}' \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ y \end{pmatrix} - \lambda \hat{\sigma} \right)$$

# QnA

$$\Rightarrow \quad (\hat{\beta}^y)_{\hat{I}} = \left(X_{[n+1],\hat{I}}' X_{[n+1],\hat{I}}\right)^{-1} \left(X_{[n+1],\hat{I}}' \left(\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ y \end{pmatrix}\right) - \lambda \hat{s}\right)$$

$$\hat{f}(x_i; D_{n+1}^y) = x_i^T \hat{\beta}^y = (x_i)_{\hat{I}}' (\hat{\beta}^y)_{\hat{I}}$$

$$(x_i)_{\hat{I}}^T (\hat{\beta}^y)_{\hat{I}} = (x_i)_{\hat{I}}' \left(X_{[n+1],\hat{I}}' X_{[n+1],\hat{I}}\right)^{-1} \left(X_{[n+1],\hat{I}}' \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \end{pmatrix} - \lambda \hat{s}\right) \quad a_i(\hat{I}, \hat{s})$$

$$+ (x_i)_{\hat{I}}' \left(X_{[n+1],\hat{I}}' X_{[n+1],\hat{I}}\right)^{-1} \left(X_{[n+1],\hat{I}}' \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}\right) y \quad b_i(\hat{I}, \hat{s})$$

$\Rightarrow$ $\hat{I}, \hat{s}$ 고정시 $\hat{y}$은 $y$에 대해 고정된 선형식을 가짐!

Piecewise 인 이유. = $y$가 움직이면서 (1) $j \in \hat{I}$ 인 계수 $\hat{\beta}_j^y \to 0$ 이 되거나

(2) $j \notin \hat{I}$ 인 계수 $\hat{\beta}_j^y$ 가 더 이상 0이 아니게 될 때 $\hat{I}, \hat{s}$ 이 변함 $\Rightarrow$ $a_i, b_i$ 도 변함. $\Rightarrow$ piecewise 선형

(1) $(\hat{\beta}_{\hat{I}}^y)_j = 0$

(2) From KKT, for $j \notin \hat{I}$ $\left| x_j' (x_j \hat{\beta}_j^y - y) \right| = \lambda$

# QnA

<Ridge>

$$\hat{\beta}^y = (X_{[n+1]}' X_{[n+1]} + \lambda I)^{-1} X_{[n+1]}' \tilde{Y} \quad \leftarrow \text{구간 안나눠짐}$$

$$= (X_{[n+1]}' X_{[n+1]} + \lambda I)^{-1} X_{[n+1]}' \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \end{pmatrix} + (X_{[n+1]}' X_{[n+1]} + \lambda I)^{-1} X_{[n+1]}' \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} y$$

$$X_i' \hat{\beta}^y = \underbrace{X_i' (X_{[n+1]}' X_{[n+1]} + \lambda I)^{-1} X_{[n+1]}' \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \end{pmatrix}}_{a_i} + \underbrace{X_i' (X_{[n+1]}' X_{[n+1]} + \lambda I)^{-1} X_{[n+1]}' \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}}_{b_i} y$$

$a_i$, $b_i$ 에 대한 $X$

$\therefore$ 전 구간 선형 유지 $\Rightarrow$ 더 쉬움

# A general approach: discretization

**Naive discretization:** Choose grid points

$$y^{(1)} < \ldots < y^{(M)}, \qquad y^{(m+1)} - y^{(m)} = \Delta.$$

At each grid point:

$$s_m(x, y) = s((x, y); \mathcal{D}_{n+1}^{y^{(m)}}),$$

and include $y^{(m)}$ if

$$s_m(X_{n+1}, y^{(m)}) \leq \text{Quantile}((s_m(X_i, Y_i))_{i \leq n}; (1-\alpha)(1+1/n)).$$

Approximate:

$$\mathcal{C}(X_{n+1}) \approx \bigcup_{m: y^{(m)} \in \mathcal{C}} (y^{(m-1)}, y^{(m+1)}).$$

# Problem: naïve discretization breaks symmetry

- We only need to compute the score function at the $M$ grid points $\rightarrow$ **computationally efficient**.
- Only the test point $y^{(m)}$ is discretized and training responses $Y_1, \ldots, Y_n$ remain unchanged $\Rightarrow$ **Asymmetric**!
- Therefore, we lose the distribution-free guarantee.
- Therefore, we discretize not only the test point but also all training points.

# Symmetry-preserving discretization

## Proposition 9.5

Let $y_1, \ldots, y_M \in$ be a prespecified collection of values and let $k :\to [M]$ be rounding function. Let $s$ be a symmetric score function, and for each $m \in [M]$, define a function

$$s_m(x, y) = s\Big((x; y); ((X_1, y_{k(Y_1)}), \ldots, (X_n, y_{k(Y_n)}), (X_{n+1}, y_m))\Big).$$

Define $C(X_{n+1})$

$$\bigcup_{m \in [M]} \left\{ y : k(y) = m,\ s_m(X_{n+1}, y) \leq Q\big((s_m(X_i, Y_i))_{i \in [n]}; (1-\alpha)(1+\frac{1}{n})\big) \right\}.$$

Then if $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable then

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

# Proof Proposition 9.5

Define a new score function $\tilde{s}$ as follows.
For any dataset $D = ((x'_1, y'_1), \ldots, (x'_\ell, y'_\ell))$ and data point $(x', y')$, let

$$\tilde{s}\big((x', y'); \mathcal{D}\big) = s\big((x', y'); (x'_1, y_{k(y'_1)}), \ldots, (x'_\ell, y_{k(y'_\ell)})\big).$$

Now suppose that we run full conformal prediction with $\tilde{s}$ as our score function. Then we have

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

by exchangeability ($s$ is symmetric $\Rightarrow$ $\tilde{s}$ is symmetric. But if we use naive method, $s$ is symmetric $\not\Rightarrow$ $\tilde{s}$ is symmetric).

# Proof Proposition 9.5

Then we have

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \tilde{S}^y_{n+1} \leq \mathit{Quantile}\big((\tilde{S}^y_i)_{i\in[n]}; (1-\alpha)(1+1/n)\big) \right\},$$

By partitioning, we can equivalently write

$$C(X_{n+1}) = \bigcup_{m\in[M]} \left\{ y : k(y) = m,\ \tilde{S}^y_{n+1} \leq \big((\tilde{S}^y_i)_{i\in[n]}; (1-\alpha)(1+1/n)\big) \right\}$$

Now fix any $m \in [M]$. For any $y \in \mathcal{Y}$ with $k(y) = m$,

$$\begin{aligned}\tilde{s}(\cdot; {}^{y}_{n+1}) &= s(\cdot; (X_1, y_{k(Y_1)}), \ldots, (X_n, y_{k(Y_n)}), (X_{n+1}, y_{k(y)})) \\ &= s(\cdot; (X_1, y_{k(Y_1)}), \ldots, (X_n, y_{k(Y_n)}), (X_{n+1}, y_m)) = s_m(\cdot).\end{aligned}$$

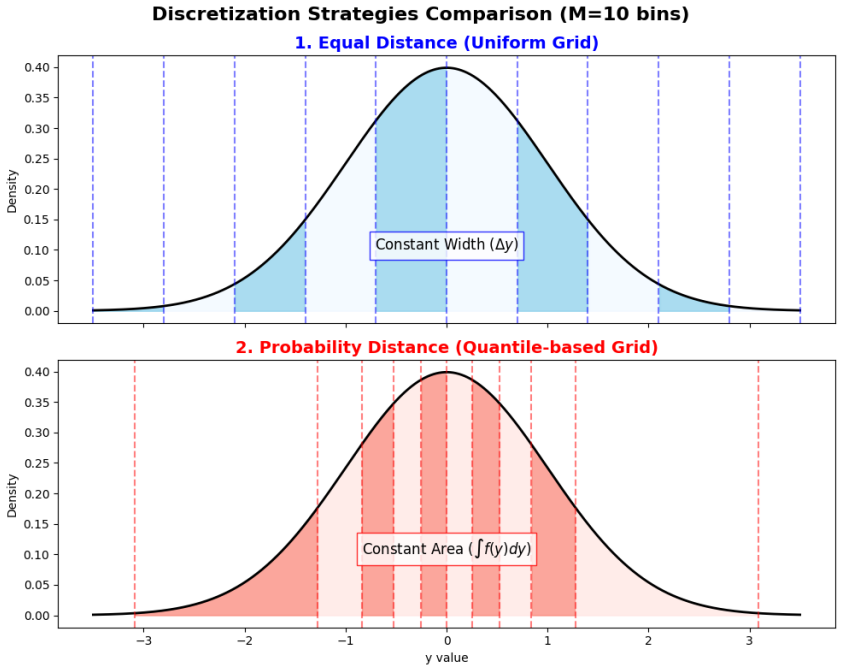
Therefore, for any $y \in$ with $k(y) = m$,

$$\tilde{S}^y_i = \begin{cases} s_m(X_i, Y_i), & i \in [n], \\ s_m(X_{n+1}, y), & i = n+1. \end{cases}$$

# QnA

**Q.** Discretization에서 y를 잡을 때 Equal Distance로 잡고 정리를 전개하고 있다는 생각이 듭니다. 만약 Y에 대한 분포가정을 추가한다면 분위수 분석과 비슷하게 해당 분포에 비례한 가중치를 준 distance를 적용해 볼 수 있을까요? (e.g. d(x, y) = integral x to y f(x)dx)

**A.** 가능하다! 정리 9.5는 데이터와 테스트 포인트 모두 동일한 라운딩 규칙을 쓰기만 하면 exchangeability가 보존. 따라서 grid point를 어떻게 잡는지는 상관없음. quantile 기반으로 grid를 잡으면 확률밀도가 높은 곳에 집중. 따라서 rounding 시 생기는 정보손실 최소화하고 예측집합이 더 정확해짐.



# QnA

Q. 본문에는 grid point들의 개수인 M을 선택하는 것의 중요성에 대한 언급이 있는데, 제 생각에 conformal prediction의 성능을 최대한으로 끌어내는 "좋은" grid를 만들기 위해서는 grid의 하한 및 상한인 $y^{(1)}, y^{(M)}$의 선택에도 신경을 써야 할 것 같은데 혹시 이들의 선택이 conformal prediction의 성능에 어떤 영향을 미치는지, 더 나아가 이들을 잘 선택하는 방법론이 있는지 간단하게 설명해주실 수 있을까요? 찾아보니 Bibliographic Notes에서 소개된 논문 중 하나인 (https://arxiv.org/abs/1611.09933) 에서 이와 비슷한 문제를 다루는 것 같습니다.

A. 계산비용과 예측집합의 폭에 영향을 미친다. 예를 들어 하한과 상한을 매우 넓게 잡은 경우,

- 양극단의 grid point들은 거의 의미가 없음 ⇒ grid point 낭비
- 동일한 M개의 grid point 사용시, 구간 폭이 넓어짐 ⇒ rounding 오차 증가 ⇒ 예측구간 폭 넓어짐
- rounding 오차를 줄이기 위해 grid point 많이 필요 ⇒ 계산비용 증가

반대의 경우, 관측될 수 있는 $Y_{n+1}$이 grid point 밖을 벗어날 가능성이 높아져 undercoverage 문제가 생길 수 있음. 따라서 가장 기본적인 상한과 하한 설정 방법은 $y_{\max} = \max_{1 \le i \le n} |Y_i|$. $[-y_{\max}, y_{\max}]$로 설정하는 것. 논문에서는 Trimmed Conformal Prediction(TCP) 방법을 제시. 간단히 설명하면 먼저 가볍고 빠른 모델을 사용해 cp를 통해 구한 예측집합을 구함. 그리고 구한 예측구간에 대해서 다시 원래 사용하려던 모델을 사용해 cp를 진행. 이 방법은 full cp의 이론적 보장은 거의 유지하면서 계산비용을 줄임.

# QnA

1. Trimming step:

$$\mathcal{Y}_{\text{trim}} = \Big\{ y \in \mathbb{R} : |((\widehat{r}_{\text{fast}})_y)_{n+1}| \text{ is in the bottom } (1-\alpha_{\text{trim}}) \text{ quantile of } |((\widehat{r}_{\text{fast}})_y)_1|, \ldots, |((\widehat{r}_{\text{fast}})_y)_{n+1}| \Big\}.$$

2. Prediction step:

$$\mathcal{Y}_{\text{predict}} = \Big\{ y \in \mathcal{Y}_{\text{trim}} : |((\widehat{r}_{\text{slow}})_y)_{n+1}| \text{ is in the bottom } (1-\alpha_{\text{predict}}) \text{ quantile of } |((\widehat{r}_{\text{slow}})_y)_1|, \ldots, |((\widehat{r}_{\text{slow}})_y)_{n+1}| \Big\}.$$

$\mathcal{Y}_{trim}$의 상한과 하한을 $y^{(1)}, y^{(M)}$로 설정할 수 있으며 $\alpha_{trim}$은 보통 1/(n+1)정도로 아주 작게 설정!

**Theorem 1.** *If the data points* $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ *are exchangeable, then either version of the TCP method gives coverage level*

$$\mathbb{P}\{Y_{n+1} \in \mathcal{Y}_{\text{predict}}\} \geq 1 - \alpha_{\text{trim}} - \alpha_{\text{predict}}.$$

# Contents

# The universality of conformal prediction

- Conformal prediction (CP) guarantees **distribution-free marginal coverage** under **exchangeability**.
- Question: Can any other method achieve distribution-free marginal coverage, and perhaps produce better intervals?
- The universality theorem says: **No—any symmetric distribution-free method is equivalent to CP**, for some choice of score function.

# Theorem: Universality of Full Conformal Prediction

## Theorem (9.6)

*Let $C$ be any predictive inference procedure, which maps any training dataset and test point to a prediction interval,*

$$(D, x) \mapsto C(x; D) \subseteq \mathcal{Y}.$$

*Assume $C$ is symmetric in the training data, i.e., $C(x; D) = C(x; D_\sigma)$ for any $x \in \mathcal{X}$, any dataset $D$, and any permutation $\sigma$. Assume also that $C$ satisfies distribution-free predictive coverage:*

$$\text{Under exchangeablity} \Rightarrow P(Y_{n+1} \in C(X_{n+1}; D_n)) \geq 1 - \alpha.$$

*Then there exists a symmetric conformal score function $s$ such that $C$ is equal to the full conformal prediction interval constructed with this score function.*

# QnA

Q. 9.3절에서의 내용을 "exchangeability보다 더 완화해서 CP를 할 수 있다"로 이해했는데, exchangeability를 만족하지 않으면서 Thm9.6 조건들을 만족해서 CP를 적용할 수 있는 예시가 있을까요?

A.

> Throughout this book, we have seen that conformal prediction provides a strategy for ensuring distribution-free marginal predictive coverage under only an assumption of exchangeability. But are there alternative methods that might achieve the same goal—and, perhaps, offer more informative prediction intervals?
>
> The following result proves a *universality* property of full conformal prediction. It demonstrates that any method achieving distribution-free marginal coverage must actually be equivalent to running full conformal prediction (with some choice of score function), as long as we assume that the method is symmetric in the training data.

9.3절은 Exchangeability 가정을 완화하는 정리가 아니라 Exchangeability 가정 하에서 Symmetric한 방법론을 쓴다면, 그것은 결국 CP로 귀결된다는 CP의 유일성/보편성을 강조하는 장으로 이해할 수 있다. 즉, 정리 9.6도 일단 교환가능성은 가정하고 있음.

# Proof proposition 9.6

Define a score function:

$$s((x,y);\mathcal{D}') = \mathbf{1}\{y \notin \mathcal{C}(x;\mathcal{D}'_{\setminus(x,y)})\}.$$

- we define $D_{\setminus(x,y)}$ as the dataset obtained by removing one copy of $(x,y)$.
- $s$ is symmetric because $\mathcal{C}$ is symmetric.

By definition of $s$, we see that

$$S^y_{n+1} = \mathbf{1}\{y \notin C(X_{n+1};(D^y_{n+1})_{\setminus(X_{n+1},y)})\} = \mathbf{1}\{y \notin C(X_{n+1};D_n)\},$$

$$\begin{aligned} y \in C(X_{n+1};D_n) &\implies S^y_{n+1} = 0 \\ &\implies S^y_{n+1} \leq \mathit{Quantile}\left(S^y_1,\ldots,S^y_n;\,(1-\alpha)\left(1+\frac{1}{n}\right)\right) \\ &\implies y \in \mathcal{C}_{\mathrm{CP}}(X_{n+1}). \end{aligned}$$

# Proof proposition 9.6

**Lemma (9.7)**

*Let $D = ((x_1, y_1), \ldots, (x_m, y_m))$ be any dataset. Then, if $C$ satisfies the assumptions of Theorem 9.6, it holds that*

$$\sum_{i=1}^{m} \mathbf{1}\{y_i \in C(x_i; D_{\setminus(x_i,y_i)})\} \geq (1-\alpha)m.$$

*In particular, defining $s_i = \mathbf{1}\{y_i \notin C(x_i; D_{\setminus(x_i,y_i)})\} \in \{0,1\}$ for each $i \in [m]$, it holds that $Quantile(s_1, \ldots, s_m; 1-\alpha) = 0$.*

Therefore

$$y \notin \mathcal{C}(X_{n+1}; \mathcal{D}_n) \Longrightarrow S^y_{n+1} = 1 \Longrightarrow S^y_{n+1} > Q(S^y_1, \ldots, S^y_{n+1}; 1-\alpha),$$

$$\mathcal{C}(X_{n+1}; \mathcal{D}_n) \supseteq \mathcal{C}_{\mathrm{CP}}(X_{n+1}),$$

which completes the proof.

# Proof of Lemma 9.7

Let $\sigma$ be a uniformly random permutation and define

$$(\tilde{X}_i, \tilde{Y}_i) = (x_{\sigma(i)}, y_{\sigma(i)}).$$

Then $\{(\tilde{X}_i, \tilde{Y}_i)\}$ is **exchangeable**.

Distribution-free validity of $C$ then implies that

$$1 - \alpha \leq \mathbb{P}\left(\tilde{Y}_m \in \mathcal{C}(\tilde{X}_m; (\tilde{X}_i, \tilde{Y}_i)_{i<m})\right).$$

Rewrite in terms of the original data,

$$\tilde{Y}_m \in \mathcal{C}(\tilde{X}_m; \cdot) \iff y_{\sigma(m)} \in \mathcal{C}(x_{\sigma(m)}; \mathcal{D}_{\setminus(x_{\sigma(m)}, y_{\sigma(m)})}).$$

Thus $\sigma(m)$ is distributed uniformly over $\{1, \ldots, m\}$

$$P\left(y_{\sigma(m)} \in \mathcal{C}(x_{\sigma(m)}; \mathcal{D}_{\setminus(x_{\sigma(m)}, y_{\sigma(m)})})\right) = \mathbb{E}\left[\mathbf{1}\left\{y_{\sigma(m)} \in \mathcal{C}(x_{\sigma(m)}; \mathcal{D}_{\setminus(x_{\sigma(m)}, y_{\sigma(m)})})\right\}\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left\{y_i \in \mathcal{C}(x_i; \mathcal{D}_{\setminus(x_i, y_i)})\right\}.$$