Distribution-free Uncertainty Quantification: Impossibility and Possibility

Part II

Shuo Li

Based on:

Distribution-free binary classification: prediction sets, confidence intervals and calibration

Distribution-free inference for regression: discrete, continuous, and in between.

Achieving asymptotic calibration via binning

- As shown in *part I*, it is impossible for an injective post-hoc calibration algorithm to be distribution-free asymptotically calibrated.
- However, many parametric calibration schemes are injective. These schemes include Platt scaling, temperature scaling, and beta calibration.
- How do we obtain finite partitions? Binning!

Important concepts in binning

- Suppose we have a fixed partition of $\mathcal{X}$ into B regions $\{\mathcal{X}_b\}_{b\in[B]}$, and let $\pi_b = \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]$ be the expected label probability in region $\mathcal{X}_b$.
- Denote the partition-identity function as $\mathcal{B} : \mathcal{X} \to [B]$ where $\mathcal{B}(x) = b$ if and only if $x \in \mathcal{X}_b$.

Important concepts in binning

- Given a calibration set $\{(X_i, Y_i)\}_{i\in[n]}$, let

$$N_b := \left|\{i \in [n] : \mathcal{B}\left(X_i\right) = b\}\right|$$

be the number of points from the calibration set that belong to region $\mathcal{X}_b$.

- Define

$$\widehat{\pi}_b := \frac{1}{N_b} \sum_{i:\mathcal{B}(X_i)=b} Y_i \quad \text{and} \quad \widehat{V}_b := \frac{1}{N_b} \sum_{i:\mathcal{B}(X_i)=b} \left(Y_i - \widehat{\pi}_b\right)^2$$

as the empirical average and variance of the Y values in a partition.

Asymptotic calibration can be achieved via binning

Theorem

For any $\alpha \in (0,1)$*, with probability at least* $1-\alpha$*,*

$$|\pi_b - \widehat{\pi}_b| \leqslant \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b},$$

simultaneously for all $b \in [B]$*.*

- As $N_b \to \infty$, $|\pi_b - \hat{\pi}_b| \to 0$ in bin b.

Proof sketch

Theorem

(Partial statement of Audibert et al.) Let $X_1, \dots, X_n$ be $i.i.d.$ random variables bounded in $[0, s]$, for some $s > 0$. Let $\pi = \mathbb{E}[X_1]$ be their common expected value. Consider the empirical mean $\bar{X}_n$ and variance V_n defined respectively by

$$\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \text{and} \quad V_n = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}{n}.$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left|\bar{X}_n - \pi\right| \leqslant \sqrt{\frac{2V_n \log(3/\delta)}{n}} + \frac{3s \log(3/\delta)}{n}.$$

Proof sketch

- Let $E_{\mathcal{B}(x)}$ be a partition of the calibration set: $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. The Y_i-s in each bin represent independent Bernoulli random variables that share the same mean $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$.
- By setting $\delta = \alpha/B$ and $s = 1$ in the above theorem, we obtain that:

$$P\left(|\pi_b - \widehat{\pi}_b| > \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \mid E_{\mathcal{B}(x)}\right) \leqslant \alpha/B.$$

- Applying a union bound over all regions, we get that:

$$P\left(\forall b \in [B] : |\pi_b - \widehat{\pi}_b| \leqslant \sqrt{\tfrac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \tfrac{3\ln(3B/\alpha)}{N_b} \mid E_{\mathcal{B}(x)}\right) \geq 1 - \alpha.$$

- Because this is true for any $E_{\mathcal{B}(x)}$, we can marginalize to obtain the assertion of the theorem in an unconditional form.

Proof sketch

- Let $E_{\mathcal{B}(x)}$ be a partition of the calibration set: $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. The Y_i-s in each bin represent independent Bernoulli random variables that share the same mean $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$.
- By setting $\delta = \alpha/B$ and $s = 1$ in the above theorem, we obtain that:

$$P\left(|\pi_b - \widehat{\pi}_b| > \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \mid E_{\mathcal{B}(x)}\right) \leqslant \alpha/B.$$

- Applying a union bound over all regions, we get that:

$$P\left(\forall b \in [B] : |\pi_b - \widehat{\pi}_b| \leqslant \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \mid E_{\mathcal{B}(x)}\right) \geq 1 - \alpha.$$

- Because this is true for any $E_{\mathcal{B}(x)}$, we can marginalize to obtain the assertion of the theorem in an unconditional form.

Proof sketch

- Let $E_{\mathcal{B}(x)}$ be a partition of the calibration set: $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. The Y_i-s in each bin represent independent Bernoulli random variables that share the same mean $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$.
- By setting $\delta = \alpha/B$ and $s = 1$ in the above theorem, we obtain that:

$$P\left(|\pi_b - \widehat{\pi}_b| > \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \mid E_{\mathcal{B}(x)}\right) \leqslant \alpha/B.$$

- Applying a union bound over all regions, we get that:

$$P\left(\forall b \in [B] : |\pi_b - \widehat{\pi}_b| \leqslant \sqrt{\tfrac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \tfrac{3\ln(3B/\alpha)}{N_b} \mid E_{\mathcal{B}(x)}\right) \geq 1 - \alpha.$$

- Because this is true for any $E_{\mathcal{B}(x)}$, we can marginalize to obtain the assertion of the theorem in an unconditional form.

Proof sketch

- Let $E_{\mathcal{B}(x)}$ be a partition of the calibration set: $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. The Y_i-s in each bin represent independent Bernoulli random variables that share the same mean $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$.
- By setting $\delta = \alpha/B$ and $s = 1$ in the above theorem, we obtain that:

$$P\left(|\pi_b - \widehat{\pi}_b| > \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \mid E_{\mathcal{B}(x)}\right) \leqslant \alpha/B.$$

- Applying a union bound over all regions, we get that:

$$P\left(\forall b \in [B] : |\pi_b - \widehat{\pi}_b| \leqslant \sqrt{\tfrac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \tfrac{3\ln(3B/\alpha)}{N_b} \mid E_{\mathcal{B}(x)}\right) \geq 1 - \alpha.$$

- Because this is true for any $E_{\mathcal{B}(x)}$, we can marginalize to obtain the assertion of the theorem in an unconditional form.

The length is controlled by the partition

Let $b^* = \arg\min_{b \in [B]} N_b$ denote the index of the region with the minimum number of calibration examples.

Corollary

For $\alpha \in (0, 1)$, the function $h_n(\cdot) := \widehat{\pi}_{\mathcal{B}(\cdot)}$ is distribution-free (ε, α)-calibrated with

$$\varepsilon = \sqrt{\frac{2\widehat{V}_{b^\star} \ln(3B/\alpha)}{N_{b^\star}}} + \frac{3 \ln(3B/\alpha)}{N_{b^\star}}$$

Thus, $\{h_n\}_{n \in \mathbb{N}}$ is distribution-free asymptotically calibrated for any α.

Proof sketch

- For each bin, we construct an $(1-\alpha)$-confidence interval via

$$C_n(b) = \left[\widehat{\pi}_b - \left(\sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b}\right), \widehat{\pi}_b + \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b}\right], b \in [B].$$

- By Part I, we can convert this confidence interval to (ϵ, α)-calibrated with

$$\varepsilon = \sup_{b\in[B]} |C(b)|/2 = \sqrt{\frac{2\widehat{V}_{b^\star} \ln(3B/\alpha)}{N_{b^\star}}} + \frac{3\ln(3B/\alpha)}{N_{b^\star}}.$$

Proof sketch

- For each bin, we construct an $(1-\alpha)$-confidence interval via

$$C_n(b) = \left[\widehat{\pi}_b - \left(\sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b}\right), \widehat{\pi}_b + \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b}\right], b \in [B].$$

- By Part I, we can convert this confidence interval to (ϵ, α)-calibrated with

$$\varepsilon = \sup_{b\in[B]} |C(b)|/2 = \sqrt{\frac{2\widehat{V}_{b^\star} \ln(3B/\alpha)}{N_{b^\star}}} + \frac{3\ln(3B/\alpha)}{N_{b^\star}}.$$

Proof sketch

- Let $\min_{b \in [B]} P(\mathcal{B}(X) = b) = \tau > 0$. By Hoeffding's inequality, we have, with probability $1 - \alpha/B$,

$$N_b \geq n\tau - \sqrt{\frac{n \ln(B/\alpha)}{2}}.$$

- Taking a union bound, we have with probability $1 - \alpha$, simultaneously for every $b \in [B]$,

$$N_b \geq n\tau - \sqrt{\frac{n \ln(B/\alpha)}{2}} = \Omega(n).$$

and in particular $N_{b^\star} = \Omega(n)$ where $b^\star = \arg\min_{b \in [B]} N_b$.

Proof sketch

- Let $\min_{b\in[B]} P(\mathcal{B}(X) = b) = \tau > 0$. By Hoeffding's inequality, we have, with probability $1 - \alpha/B$,

$$N_b \geq n\tau - \sqrt{\frac{n\ln(B/\alpha)}{2}}.$$

- Taking a union bound, we have with probability $1 - \alpha$, simultaneously for every $b \in [B]$,

$$N_b \geq n\tau - \sqrt{\frac{n\ln(B/\alpha)}{2}} = \Omega(n).$$

and in particular $N_{b^\star} = \Omega(n)$ where $b^\star = \arg\min_{b\in[B]} N_b$.

Proof sketch

- Given that
$$\varepsilon = \sqrt{\frac{2\widehat{V}_{b^\star} \ln(3B/\alpha)}{N_{b^\star}}} + \frac{3\ln(3B/\alpha)}{N_{b^\star}},$$
  and $N_{b^\star} = \Omega(n)$, we have
$$\varepsilon_n = O\left(\sqrt{n^{-1}}\right) = o(1).$$
  This concludes the proof.

Discussion

- Any finite partition of $\mathcal{X}$ leads to asymptotic calibration.
- However, the finite sample guarantee of the Corollary can be unsatisfactory if the sample-space partition is chosen poorly, since it might lead to small $N_{b^\star}$.
- This paper presents a data-dependent partitioning scheme that provably guarantees that $N_{b^\star}$ scales as $\Omega(n/B)$ with high probability.

Uniform-mass binning

This work proposes to construct the partition $\{\mathcal{X}_b\}_{b\in[B]}$ through binning, which uses a sample splitting strategy to learn the partition of $\mathcal{X}$.

- The labeled data is split at random into a training set $\mathcal{D}_{\text{tr}}$ and a calibration set $\mathcal{D}_{\text{cal}}$.
- Then $\mathcal{D}_{\text{tr}}$ is used to train a scoring function $g : \mathcal{X} \to [0, 1]$.
- The scoring function g usually does not satisfy a calibration guarantee out-of-the-box but can be calibrated using binning.

Uniform-mass binning

Then, *uniform-mass binning* is used to guarantee that each region $\mathcal{X}_b$ contains approximately equal numbers of calibration points. This is done by estimating the empirical quantiles of $g(X)$.

- First, the calibration set $\mathcal{D}_{\text{cal}}$ is randomly split into two parts, $\mathcal{D}^1_{\text{cal}}$ and $\mathcal{D}^2_{\text{cal}}$.
- For $j \in [B-1]$, the (j/B)-th quantile $\widehat{q}_j$ of $g(X)$ is estimated from $\left\{g\left(X_i\right), i \in \mathcal{D}^1_{\text{cal}}\right\}$.
- Then, the bins are defined as:

$$I_1 = [0, \widehat{q}_1), I_i = [\hat{q}_{i-1}, \hat{q}_i], i = 2, \ldots, B-1 \text{ and } I_B = (\hat{q}_{B-1}, 1].$$

Uniform-mass binning

Then, *uniform-mass binning* is used to guarantee that each region $\mathcal{X}_b$ contains approximately equal numbers of calibration points. This is done by estimating the empirical quantiles of $g(X)$.

- First, the calibration set $\mathcal{D}_{\mathsf{cal}}$ is randomly split into two parts, $\mathcal{D}^1_{\mathsf{cal}}$ and $\mathcal{D}^2_{\mathsf{cal}}$.
- For $j \in [B-1]$, the (j/B)-th quantile $\widehat{q}_j$ of $g(X)$ is estimated from $\left\{g\left(X_i\right), i \in \mathcal{D}^1_{\mathsf{cal}}\right\}$.
- Then, the bins are defined as:

$$I_1 = [0, \widehat{q}_1), I_i = [\hat{q}_{i-1}, \hat{q}_i], i = 2, \ldots, B-1 \text{ and } I_B = (\hat{q}_{B-1}, 1].$$

Uniform-mass binning

Then, *uniform-mass binning* is used to guarantee that each region $\mathcal{X}_b$ contains approximately equal numbers of calibration points. This is done by estimating the empirical quantiles of $g(X)$.

- First, the calibration set $\mathcal{D}_{\mathsf{cal}}$ is randomly split into two parts, $\mathcal{D}_{\mathsf{cal}}^1$ and $\mathcal{D}_{\mathsf{cal}}^2$.
- For $j \in [B-1]$, the (j/B)-th quantile $\widehat{q}_j$ of $g(X)$ is estimated from $\left\{g\left(X_i\right), i \in \mathcal{D}_{\mathsf{cal}}^1\right\}$.
- Then, the bins are defined as:

$$I_1=\left[0, \widehat{q}_1\right), I_i=\left[\hat{q}_{i-1}, \hat{q}_i\right], i=2, \ldots, B-1 \text{ and } I_B=\left(\hat{q}_{B-1}, 1\right].$$

Uniform-mass binning

Theorem

Fix $g : \mathcal{X} \to [0, 1]$ and $\alpha \in (0, 1)$. There exists a universal constant c such that if $\left|\mathcal{D}^1_{cal}\right| \geqslant cB \ln(2B/\alpha)$, then with probability at least $1 - \alpha$,

$$N_{b^\star} \geqslant \left|\mathcal{D}^2_{cal}\right| /2B - \sqrt{|\mathcal{D}^2_{cal}| \ln(2B/\alpha)/2}.$$

Thus even if $\left|\mathcal{D}^1_{\text{cal}}\right|$ does not grow with n, as long as $\left|\mathcal{D}^2_{\text{cal}}\right| = \Omega(n)$, uniform-mass binning is distribution-free $(\widetilde{O}(\sqrt{B \ln(1/\alpha)/n}), \alpha)$-calibrated, and hence distribution-free asymptotically calibrated for any α.

Distribution-free calibration in the online setting

- We have considered the batch setting with a fixed calibration set of size n.
- In the online setting, previous methods are no longer valid.
- The calibration set size n is not know before hand; data points are observed sequentially.

Distribution-free calibration in the online setting

- For some value of n, let the calibration data be given as $\mathcal{D}_{\text{cal}}^{(n)}$.
- Let $\left\{\left(X_i^b, Y_i^b\right)\right\}_{i \in \left[N_b^{(n)}\right]}$ be examples from the calibration set that fall into the partition $\mathcal{X}_b$.
- Let the empirical label average and cumulative (unnormalized) empirical variance be denoted as

$$\widehat{V}_b^{+} = 1 \vee \sum_{i=1}^{N_b^{(n)}} \left(Y_i^b - \bar{Y}_{i-1}^b\right)^2, \text{ where } \bar{Y}_i^b := \frac{1}{i} \sum_{j=1}^{i} Y_j^b \text{ for } i \in \left[N_b^{(n)}\right]$$

Distribution-free calibration in the online setting

The following theorem constructs confidence intervals for $\{\pi_b\}_{b\in[B]}$ that are valid uniformly for any value of n.

Theorem

For any $\alpha \in (0,1)$, with probability at least $1-\alpha$,

$$|\pi_b - \widehat{\pi}_b| \leqslant \frac{7\sqrt{\widehat{V}_b^{+} \ln\left(1 + \ln \widehat{V}_b^{+}\right)} + 5.3 \ln\left(\frac{6.3B}{\alpha}\right)}{N_b^{(n)}},$$

simultaneously for all $b \in [B]$ and all $n \in \mathbb{N}$.

Summary

- This work proposes to utilize binning to obtain non-injective calibration algorithm.
- Using binning, distribution-free asymptotic calibration can be achieved.
- This work recommends some form of binning as the last step of calibrated prediction.

Distribution-free inference for regression: discrete, continuous, and in between

Yonghoon Lee and Rina Foygel Barber

Problem recap

Given an i.i.d. training data

$$\left(X_1, Y_1\right), \cdots, \left(X_n, Y_n\right) \overset{\text{iid}}{\sim} P = P_X \times P_{Y|X}$$

and a new test point X_{n+1}, we want to do

- estimation : estimate $\pi_P\left(X_{n+1}\right) = \mathbb{E}\left[Y_{n+1} \mid X_{n+1}\right]$.
- inference : quantify the uncertainty of the estimator - e.g. confidence interval.

Confidence interval definition

This paper aims at constructing a confidence interval $\hat{C}_n(x)$ for $\pi_P(x)$ that satisfies the following property:

Definition

An algorithm $\widehat{C}_n$ provides a distribution-free $(1-\alpha)$-confidence interval for the conditional mean if it holds that

$$\mathbb{P}_{(X_i,Y_i)\overset{\text{iid}}{\sim}P}\left\{\pi_P\left(X_{n+1}\right)\in\widehat{C}_n\left(X_{n+1}\right)\right\}\geq 1-\alpha$$

for all distributions P on $(X,Y)\in\mathbb{R}^d\times[0,1]$.

Findings from the last class

- Any distribution-free confidence interval for the mean has a non-vanishing length if P_X is nonatomic.
- By partitioning the feature space into finitely many sets (e.g., via binning), the confidence interval can have a vanishing length.

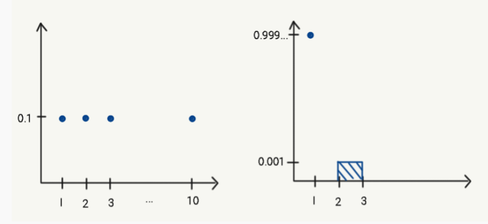
Important definitions

- For any distribution P_X on $X \in \mathbb{R}^d$, we first define the effective support size as

$$M_\gamma(P_X) = (\min \# \text{ of points needed to capture } \geq 1-\gamma \text{ probability})$$

- For any distribution P on $(X, Y) \in \mathbb{R}^d \times [0,1]$, we define $\sigma^2_{P,\beta} =$ the β-quantile of $\mathrm{Var}_P(Y \mid X)$, under the distribution $X \sim P_X$.

Effective support size example



- Left: $X \sim \mathrm{Unif}(\{1, 2, \cdots, 10\})$, $M = 10, M_{0.05}(P_X) = 10$.
- Right: $X = 0.999 \cdot \delta_1 + 0.001 \cdot$ Unif $[2, 3], M = \infty, M_{0.05}(P_X) = 1$

(Source: Yonghoon Lee's slides)

Main finding of this paper

- (Hard) $M_\gamma(P_X) \gg n^2$: distributional-free inference is as hard as the nonatomic case.
- (Easy) $M_\gamma(P_X) \ll n$: inference is trivial.
- (In-between) $n \ll M_\gamma(P_X) \ll n^2$: meaningful distributional-free inference is possible.

(Source: YongHoon Lee's slides)

A lower bound on the length of $\hat{C}_n$

Theorem

Fix any $\alpha > 0$, and let $\widehat{C}_n$ be a distribution-free $(1-\alpha)$-confidence interval. Then for any distribution P on $\mathbb{R}^d \times \mathbb{R}$, for any $\beta > 0$ and $\gamma > \alpha + \beta$

$$\mathbb{E}\left[|\widehat{C}_n\left(X_{n+1}\right)|\right] \geq \frac{1}{3}\sigma^2_{P,\beta}(\gamma - \alpha - \beta)^{1.5} \cdot \min\left\{\frac{\left(M_\gamma\left(P_X\right)\right)^{1/4}}{n^{1/2}}, 1\right\}$$

where the expected value is taken over data points $(X_i, Y_i) \overset{\text{iid}}{\sim} P$, for $i = 1, \ldots, n+1$.

Proof sketch

To show (length of $\widehat{C}_n(X_{n+1})) > \epsilon$,

- Construct a perturbation $\tilde{P}$ of P such that

$$\pi_P(X) - \pi_{\tilde{P}}(X) \asymp \epsilon$$

  and $d_{TV}\left(P^n, \tilde{P}^n\right)$: small
- $\widehat{C}_n(X_{n+1})$ should cover the mean under both P^n and $\tilde{P}^n$.
- Then $\widehat{C}_n(X_{n+1})$ should cover both $\pi_P(X_{n+1})$ and $\pi_{\tilde{P}}(X_{n+1})$ with substantial probability.

By $\pi_P(X) - \pi_{\tilde{P}}(X) \asymp \epsilon$, length of $\widehat{C}_n(X_{n+1})) > \epsilon$.

Proof sketch (more detail)

- Define $\mathcal{X}_1 = \left\{x \in \mathbb{R}^d : \mathbb{P}_{P_X}\{X = x\} > \frac{1}{M_\gamma(P_X)}\right\}$; partition $\mathbb{R}^d$ to $\mathcal{X}_2 \cup \mathcal{X}_3 \cup \dots$; write $p_m = \mathbb{P}_{P_X}\{X \in \mathcal{X}_m\}$.
- For $\epsilon \in (0, 0.5]$, $a = (a_1, a_2, \dots)$ of signs $a_1, a_2, \dots \in \{\pm 1\}$, define P_a as follows:

$$Y \mid X = x \sim \begin{cases} P_{Y|X=x}, & \text{if } x \in \mathcal{X}_1 \\ (0.5 + a_m\epsilon) \cdot P^1_{Y|X=x} + (0.5 - a_m\epsilon) \cdot P^0_{Y|X=x}, & \text{if } x \in \mathcal{X}_m \text{ for } m \geq 2 \end{cases}$$

- Define P_{mix} as the distribution over $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P_A$.

Proof sketch (more detail)

- We have

$$\pi_{P_a}(x) = \pi_P(x) + a_m \epsilon \Delta(x)$$

$$\mathrm{d_{TV}}\left(P_{\mathrm{mix}}, P^n\right) \leq 2n \sqrt{\sum_{m \geq 1} \epsilon_m^4 p_m^2} = 2n \sqrt{\sum_{m \geq 2} \epsilon^4 p_m^2} \leq \frac{2\epsilon^2 n}{\sqrt{M_\gamma\left(P_X\right)}}$$

- For a confidence interval of P_{mid}, we have

$$\mathbb{P}_{P_{\mathsf{mix}} \times P_X}\left\{\left\{P\left(X_{n+1}\right) \pm \epsilon\Delta\left(X_{n+1}\right)\right\} \cap \widehat{C}_n\left(X_{n+1}\right) \neq \emptyset\right\} \geq \gamma - \alpha$$

$$\mathbb{P}_{P^n \times P_X}\left\{\left\{P\left(X_{n+1}\right) \pm \epsilon\Delta\left(X_{n+1}\right)\right\} \cap \widehat{C}_n\left(X_{n+1}\right) \neq \emptyset\right\} \geq \gamma - \alpha - \frac{2\epsilon^2 n}{\sqrt{M_\gamma\left(P_X\right)}}$$

Proof sketch (more detail)

- Finally, we can write the length of the confidence interval as

$$\mathrm{Leb}\left(\widehat{C}_n\left(X_{n+1}\right)\right) = \int_{t\in\mathbb{R}} \mathbb{I}\left\{t \in \widehat{C}_n\left(X_{n+1}\right)\right\} \mathrm{d}t$$
$$\geq 2\sigma^2_{P,\beta} \int_{\epsilon=0}^{\epsilon_0} \mathbb{I}\left\{\sigma^2_P\left(X_{n+1}\right) \geq \sigma^2_{P,\beta} \text{ and } \left\{_P\left(X_{n+1}\right) \pm \epsilon\Delta\left(X_{n+1}\right)\right\} \cap \widehat{C}_n\left(X_{n+1}\right) \neq \emptyset\right\} \mathrm{d}\epsilon,$$

- The first term on the right hand size is propotional to $\pi_P(X) - \pi_{P_{\text{mix}}}(X)$; the second term is the probability that the confidence interval covers both $\pi_P(X)$ and $\pi_{P_{\text{mix}}}(X)$.

Special Cases: Uniform discrete features

If P_X is a uniform distribution over M points, then for any $\gamma > 0$ the effective support size is $M_\gamma\left(P_X\right) = \lceil (1-\gamma)M \rceil$. Therefore, Theorem 1 implies that for any P with nonatomic marginal P_X,

$$\mathbb{E}\left[\left|\widehat{C}_n\left(X_{n+1}\right)\right|\right] \geq \frac{1}{3}\sigma_{P,\beta}^2(\gamma-\alpha-\beta)^{1.5}(1-\gamma)^{0.25} \cdot \min\left\{\frac{M^{1/4}}{n^{1/2}}, 1\right\}$$

for any $\beta \in (0, \gamma - \alpha)$.

Special Cases: Uniform discrete features

$$\boxed{\mathbb{E}\left[|\widehat{C}_n\left(X_{n+1}\right)|\right] \geq \frac{1}{3}\sigma^2_{P,\beta}(\gamma-\alpha-\beta)^{1.5}(1-\gamma)^{0.25}\cdot\min\left\{\frac{M^{1/4}}{n^{1/2}},1\right\}}$$

- $M \gg n^2$ implies a *constant* lower bound on:

$$\mathbb{E}\left[|\widehat{C}_n\left(X_{n+1}\right)|\right] \geq \frac{1}{3}\sigma^2_{P,\beta}(\gamma-\alpha-\beta)^{1.5}(1-\gamma)^{0.25}.$$

- $M \ll n^2$ allows for the possibility of a *vanishing* length for $\mathbb{E}\left[|\widehat{C}_n\left(X_{n+1}\right)|\right]$.

Special cases: binary response

If the response Y is known to be binary (i.e., $Y \in \{0,1\}$), we might relax the requirement of distribution-free coverage to only include distributions of this type, i.e., we require

$$\mathbb{P}_{(X_i,Y_i) \overset{\text{iid}}{\sim} P} \left\{ \pi_P \left(X_{n+1} \right) \in \widehat{C}_n \left(X_{n+1} \right) \right\} \geq 1 - \alpha$$

for all distributions P on $\mathbb{R}^d \times \{0,1\}$.

Special cases: binary response

- This condition is strictly weaker than the original definition.
- If we could construct a confidence interval $\widehat{C}_n$ in the binary response case, then we could also construct a confidence interval the general case.

Special cases: binary response

- Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, $\ldots, (X_n, Y_n)$, for each $i = 1, \ldots, n$, draw a binary response $\tilde{Y}_i \sim \mathrm{Bernoulli}\,(Y_i)$.
- Then we clearly have n i.i.d. draws from a distribution on $(X, \tilde{Y}) \in \mathbb{R}^d \times \{0, 1\}$, where

$$\mathbb{E}[\tilde{Y} \mid X] = \mathbb{E}[Y \mid X] = \pi_P(X).$$

- $\widehat{C}_n$ on the new data $\left(X_1, \tilde{Y}_1\right), \ldots, \left(X_n, \tilde{Y}_n\right)$ satisfies the coverage property in the general case.

Special cases: binary response

- Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, $\ldots, (X_n, Y_n)$, for each $i = 1, \ldots, n$, draw a binary response $\tilde{Y}_i \sim \mathrm{Bernoulli}(Y_i)$.
- Then we clearly have n i.i.d. draws from a distribution on $(X, \tilde{Y}) \in \mathbb{R}^d \times \{0, 1\}$, where

$$\mathbb{E}[\tilde{Y} \mid X] = \mathbb{E}[Y \mid X] = \pi_P(X).$$

- $\widehat{C}_n$ on the new data $\left(X_1, \tilde{Y}_1\right), \ldots, \left(X_n, \tilde{Y}_n\right)$ satisfies the coverage property in the general case.

Special cases: binary response

- Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, $\ldots, (X_n, Y_n)$, for each $i = 1, \ldots, n$, draw a binary response $\tilde{Y}_i \sim \mathrm{Bernoulli}\,(Y_i)$.
- Then we clearly have n i.i.d. draws from a distribution on $(X, \tilde{Y}) \in \mathbb{R}^d \times \{0, 1\}$, where

$$\mathbb{E}[\tilde{Y} \mid X] = \mathbb{E}[Y \mid X] = \pi_P(X).$$

- $\widehat{C}_n$ on the new data $\left(X_1, \tilde{Y}_1\right), \ldots, \left(X_n, \tilde{Y}_n\right)$ satisfies the coverage property in the general case.

Special cases: nonatomic features

We now consider the setting where the marginal distribution of X is nonatomic. In this case, for any $\gamma > 0$ the effective support size is $M_\gamma(P_X) = \infty$.

Special cases: nonatomic features

Therefore, Theorem 1 implies that for any P with nonatomic marginal P_X, for any $\beta \in (0, 1-\alpha)$,

$$\mathbb{E}\left[\left|\widehat{C}_n\left(X_{n+1}\right)\right|\right] \geq \frac{1}{3}\sigma^2_{P,\beta}(1-\alpha-\beta)^{1.5}$$

- This lower bound does not depend on n.
- The width of any distributionfree confidence interval is non-vanishing even for arbitrarily large sample size n.

Comparing to a similar Theorem in Part I

Theorem

For any nonatomic P, len$_{n,\alpha}(\hat{C}_n, P) \geq L_\alpha(P) > 0$ where

$$L_\alpha(P) = \inf_{a:\mathbb{R}^d \to [0,1]} \{\mathbb{E}_P[\ell(\pi_P(X), a(X))] : \mathbb{E}_P[a(X)] \leq \alpha\}$$

with $\ell : [0,1] \to [0,1]$ fixed.

- An interesting question is to see which bound is tighter.

Special Cases: unbounded

Would it be possible for us to instead consider the general case, where P is an unknown distribution on $\mathbb{R}^d \times \mathbb{R}$? The following result shows that this more general question is not meaningful:

Proposition

Suppose an algorithm $\widehat{C}_n$ satisfies

$$\mathbb{P}_{(X_i,Y_i) \overset{\text{iid}}{\sim} P} \left\{ \pi_P \left(X_{n+1} \right) \in \widehat{C}_n \left(X_{n+1} \right) \right\} \geq 1 - \alpha$$

for all distributions P on $\mathbb{R}^d \times \mathbb{R}$. Then for all distributions P, for all $y \in \mathbb{R}$ it holds that

$$\mathbb{P}_{(X_i,Y_i) \overset{\text{iid}}{\sim} P} \left\{ y \in \widehat{C}_n \left(X_{n+1} \right) \right\} \geq 1 - \alpha$$

Special Cases: unbounded

$$\boxed{\mathbb{P}_{(X_i,Y_i)\overset{\text{iid}}{\sim}P}\left\{y \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1-\alpha}$$

This means that if we require $\widehat{C}_n$ to have distribution-free coverage over distributions with unbounded response, then inevitably, every point in the real line is contained in the resulting confidence interval a substantial portion of the time.

Propose an algorithm achieving the lower bound

To achieve the lower bound, this work proposes an algorithm that, for certain "nice" distributions P, can achieve a confidence interval length that matches the rate of the lower bound.

Propose an algorithm achieving the lower bound

This algorithm requires two inputs:

- A hypothesized ordered support set $\left\{x^{(1)}, x^{(2)}, \ldots\right\} \subset \mathbb{R}^d$ for the marginal P_X.
- A hypothesized mean function $\hat{\pi} : \mathbb{R}^d \to [0, 1]$.

The hypothesized support set $\left\{x^{(1)}, x^{(2)}, \ldots\right\}$ should aim to list the highest-probability values of X early in the list, while the hypothesized mean function $\hat{\pi}(\cdot)$ should aim to be as close to the true conditional mean π_P as possible.

Propose an algorithm achieving the lower bound

Step 1 : **estimate the effective support size.**

We compute a probabilistic upper bound of the effective support size:

$$\widehat{M}_\gamma = \min\left\{m : \sum_{i=1}^{n} \mathbb{I}\left\{X_i \in \left\{x^{(1)}, \ldots, x^{(m)}\right\}\right\} \geq (1-\gamma)n + \sqrt{\frac{n\log(2/\delta)}{2}}\right\}.$$

The estimated effective support size satisfies

$$\mathbb{P}\left\{\widehat{M}_\gamma \geq M_\gamma\left(P_X\right)\right\} \geq 1 - \delta/2$$

Propose an algorithm achieving the lower bound

Step 2 : **estimate error at each repeated X value.**

Define

$$Z = \sum_{\substack{m=1,2,\ldots \\ \text{s.t. } n_m \geq 2}} (n_m - 1) \cdot \left(\left(\bar{y}_m - \left(\hat{\pi}(x^{(m))} \right) \right)^2 - n_m^{-1} s_m^2 \right)$$

where

$$\begin{aligned} n_m &= \textstyle\sum_{i=1}^n \mathbb{I}\left\{ X_i = x^{(m)} \right\}, \bar{y}_m = \frac{1}{n_m} \sum_{i=1}^n Y_i \cdot \mathbb{I}\left\{ X_i = x^{(m)} \right\} \\ s_m^2 &= \textstyle\frac{1}{n_m - 1} \sum_{i=1}^n \left(Y_i - \bar{y}_m \right)^2 \cdot \mathbb{I}\left\{ X_i = x^{(m)} \right\}. \end{aligned}$$

Note: $\mathbb{E}\left[\left(\bar{y}_m - \left(\hat{\pi}(x^{(m)}) \right) \right)^2 - n_m^{-1} s_m^2 \right] = \left(\left(\hat{\pi}(x^{(m)}) \right) - \pi_P \left(x^{(m)} \right) \right)^2$

Propose an algorithm achieving the lower bound

Step 3 : **construct the confidence interval**

Let

$$\widehat{\Delta} = \sqrt{\frac{2\widehat{M}_\gamma + n}{n(n-1)}} \cdot \sqrt{4Z_+ + 4\sqrt{N_{\geq 2} \cdot 2/\delta} + 16/\delta},$$

where

$$N_{\geq 2} = \sum_{m=1}^{\infty} \mathbb{I}\left\{n_m \geq 2\right\} \text{ and } Z_+ = \max\{Z, 0\},$$

and define

$$\widehat{C}_n(x) = \left[\max\left\{0, \hat{\pi}(x) - \frac{\widehat{\Delta}}{\alpha - \delta - \gamma}\right\}, \min\left\{1, \hat{\pi}(x) + \frac{\widehat{\Delta}}{\alpha - \delta - \gamma}\right\}\right].$$

Theorem 2

Theorem

The confidence interval constructed by the above algorithm is a distribution-free $(1-\alpha)$-confidence interval.

Theorem 3

Theorem

Suppose the distribution P on $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ has marginal P_X that is supported on $\left\{x^{(1)}, \ldots, x^{(M)}\right\}$ and satisfies $\mathbb{P}_{P_X}\left\{X = x^{(m)}\right\} \leq \eta / M$ for all m, and suppose that P has conditional mean $\pi_P : \mathbb{R}^d \to \mathbb{R}$ that satisfies $\mathbb{E}_{P_X}\left[\left(\pi_P(X) - \hat{\pi}(X)\right)^2\right] \leq \mathrm{err}_{\hat{\pi}}^2$. Then the confidence interval constructed in step 3 satisfies

$$\mathbb{E}\left[\left|\widehat{C}_n\left(X_{n+1}\right)\right|\right] \leq c\left(\mathrm{err}_{\hat{\pi}} + \frac{M^{1/4}}{n^{1/2}}\right)$$

where c depends only on the parameters $\alpha, \delta, \gamma, \eta$.

Theorem 3

- There exists a setting where a distribution-free confidence interval satisfies

$$\mathbb{E}\left[\mathrm{Leb}\left(\widehat{C}_n\left(X_{n+1}\right)\right)\right] \leq c\left(\mathrm{err}_{\hat{\pi}} + \frac{M^{1/4}}{n^{1/2}}\right)$$

- i.e., the lower bound can be achieved.
- If $M \ll n^2$, it is possible to have a confidence interval of vanishing length.

Theorem 3 Example

- If $\hat{\pi}$ is constructed via logistic regression, and the distribution P follows this model, then we have $\mathrm{err}_{\hat{\pi}} = \mathcal{O}(\sqrt{d/n})$.
- In a k-sparse regression setting where we use logistic lasso we might instead obtain $\mathrm{err}_{\hat{\pi}} = \mathcal{O}(\sqrt{k\log(d)/n})$.
- If $x \mapsto \pi_P(x)$ is β-Hölder smooth, then we have $\mathrm{err}_{\hat{\pi}} = \mathcal{O}\left(n^{-\beta/(\beta+d)}\right)$.

Conclusion

- This work derives a lower bound for the confidence interval:

$$\mathbb{E}\left[\left|\widehat{C}_n\left(X_{n+1}\right)\right|\right] \geq \frac{1}{3}\sigma_{P,\beta}^2(\gamma-\alpha-\beta)^{1.5} \cdot \min\left\{\frac{\left(M_\gamma\left(P_X\right)\right)^{1/4}}{n^{1/2}}, 1\right\}.$$

  - $M_\gamma(P_X) \gg n^2$: distribution-free confidence interval does not have vanishing length.
  - $n \ll M_\gamma(P_X) \ll n^2$: distribution-free confidence interval can have vanishing length.
- This work proposes an algorithm to construct an $(1-\alpha)$ confidence interval that can achieve the length lower bound.

(Source: YongHoon Lee's slides)

Thank you!