Volume Optimality in Conformal Prediction with Structured Prediction Sets

Chao Gao[1], Liren Shan[2], Vaidehi Srinivas[3], Aravindan Vijayaraghavan[3]

[1]University of Chicago, [2]Toyota Technological Institute at Chicago, [3]Northwestern University

Problem Setting

Goal: designing conformal methods with formal guarantees on the **size** of the output confidence sets

Problem statement: For some $\alpha \in (0,1)$, dataset of n labeled calibration samples $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$, and test example X_{n+1} that corresponds to an unknown true label Y_{n+1}, construct a prediction set $\widehat{C}(X_{n+1})$, satisfying the coverage requirement:

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

$\mathbb{P}$ refers to the joint distribution over all $n+1$ pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ including the test sample. In addition to achieving coverage, the conformal set $\widehat{C} = \widehat{C}(X_{n+1})$ should be efficient, i.e., small.

Overview of Results

Most conformal methods provide formal guarantees on coverage, and validate the size of sets empirically. We investigate the problem of formally achieving **volume optimality**:

Question: Given calibration samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn i.i.d. from distribution P, can we find the smallest (in volume/Lebesgue measure) data-dependent set $\widehat{C} \subset \mathcal{Y}$ that satisfies coverage for $(X_{n+1}, Y_{n+1}) \sim P$?

- **Impossibility result:** Any distribution-free method that satisfies coverage can only find a trivial solution whose volume is sub-optimal.
- **Structured Prediction Sets and Restricted Volume Optimality:** For $\mathscr{C}$ of bounded VC-dimension, it is possible to compete with the smallest $C \in \mathscr{C}$ that achieves coverage, via standard uniform convergence
- **Conformalized Dynamic Programming:** A new conformity score based on dynamic programming achieves volume optimality with respect to unions of k intervals, as long as a reasonable estimator of the conditional CDF is available. (Extension of the framework of [Izbicki, Shimizu, Stern, JMLR '22], [Chernozhukov, Wüthrich, Zhu, PNAS '21])

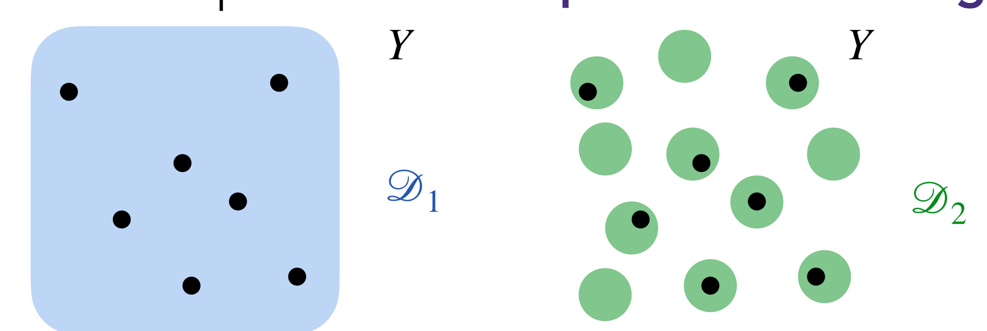
Informal Theorem: For $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \sim P$ drawn i.i.d., and given $\alpha \in (0,1), k \geq 1$, conformalized dynamic programming can find a set $\widehat{C}$ such that

$$\mathrm{vol}(\widehat{C}) \leq \mathrm{Opt}_k(P, 1 - \alpha + \varepsilon),$$

for some appropriately chosen small $\varepsilon > 0$. A conformalizing procedure ensures that we achieve finite sample coverage for exchangeable data.

Impossibility Result

Can be seen as consequence of a **nonparametric testing** lower bound.



From samples Y, we cannot distinguish between $\mathscr{D}_1$ (uniform on **blue** region), and $\mathscr{D}_2$ (uniform on **green** region). Thus any distribution-free method that achieves coverage on $\mathscr{D}_1$, must provide a very large conformal set on $\mathscr{D}_2$. (Construction holds even in 1 dimension.)

Structured Prediction Sets

We define **restricted volume optimality**. For a given k, let $\mathscr{C}_k$ be the set family of all unions of k intervals. Let

$$\mathrm{Opt}_k(P, 1-\alpha) = \inf_{C \in \mathscr{C}_k} \{\mathrm{vol}(C) : P(C) \geq 1 - \alpha\}.$$

(1) Since $\mathscr{C}_k$ is a set family of bounded VC-dimension, the coverage of sets in $\mathscr{C}_k$ over i.i.d. samples exhibits **uniform convergence**, thus it is statistically tractable to find a set $\widehat{C}$ such that, for a small $\varepsilon > 0$,
$$\mathrm{vol}(\widehat{C}) \leq \mathrm{Opt}_k(P, 1 - \alpha + \varepsilon).$$

(2) Restricted optimality coincides with true optimality for P that can be approximated by a distribution with at most k modes (e.g., P that admits a good KDE)

(3) Efficiently computable using **dynamic programming**

(4) Reasonable to restrict k from the perspective of interpretability ($k = 1$ is already interesting)

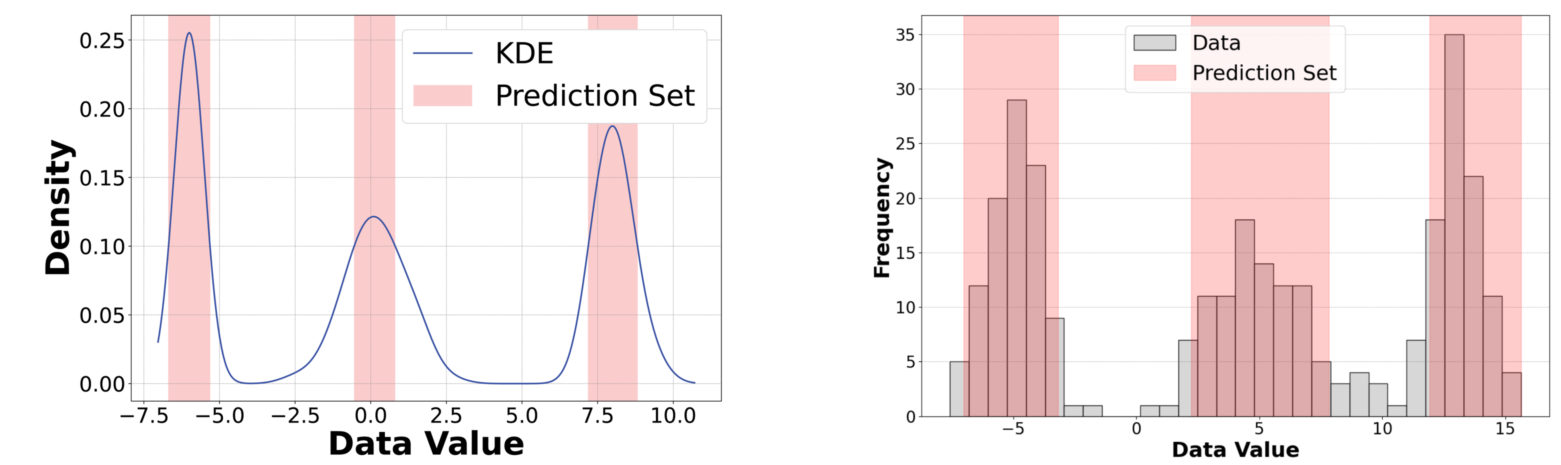
Conformalized Dynamic Programming (DP)

- Dynamic programming alone achieves a coverage guarantee over i.i.d. samples given number of samples scaling with k (VC-dimension of $\mathscr{C}_k$)
- For finite sample coverage over exchangeable samples, we design a **conformity score** based on dynamic programming, which plugs into a split conformal framework
- Dynamic programming procedure can be applied to the **estimated c.d.f.** of $Y|X$, achieve **approximate conditional coverage** and **conditional restricted volume optimality** when the estimated c.d.f has low error

Experiments

Comparison to density estimation (KDE)

Avoiding density estimation allows our method to achieve lower volume. KDE requires fine-tuning the bandwidth ρ, whereas Conformalized DP is robust to the setting of the parameter k.



Volume: 4.48 Volume: 3.14

Figure 1: Mixture of Gaussian data for target coverage 0.8
(Left) Conformalized KDE [Lei, Robbins, Wasserman JASA '13] with $\rho = 0.5$
(Right) Conformalized Dynamic Programming with $k = 3$

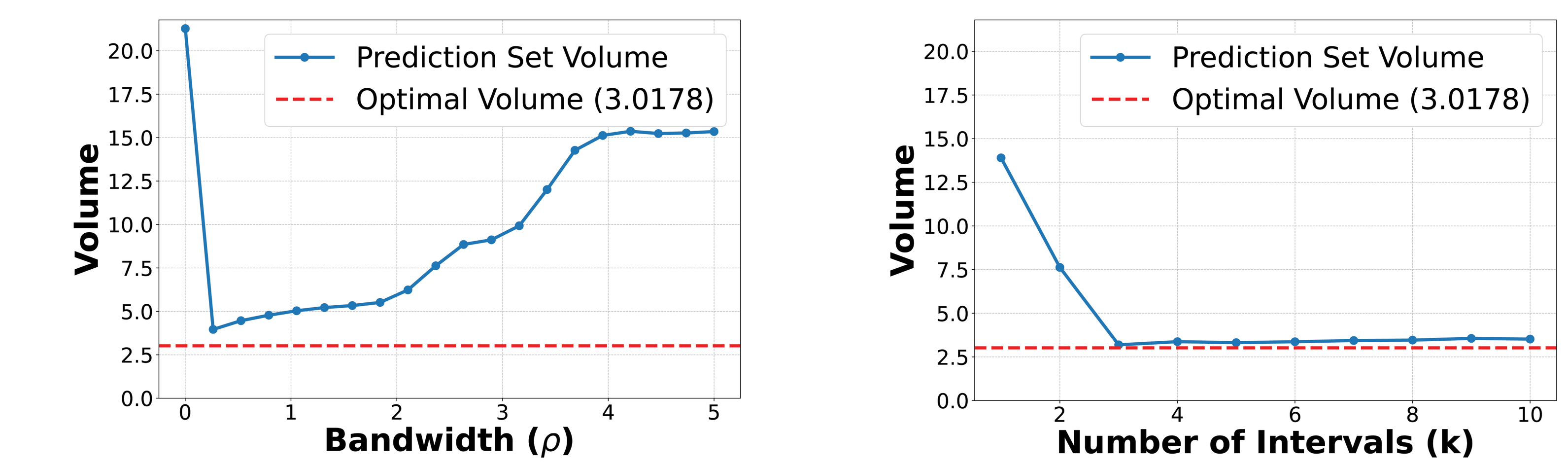
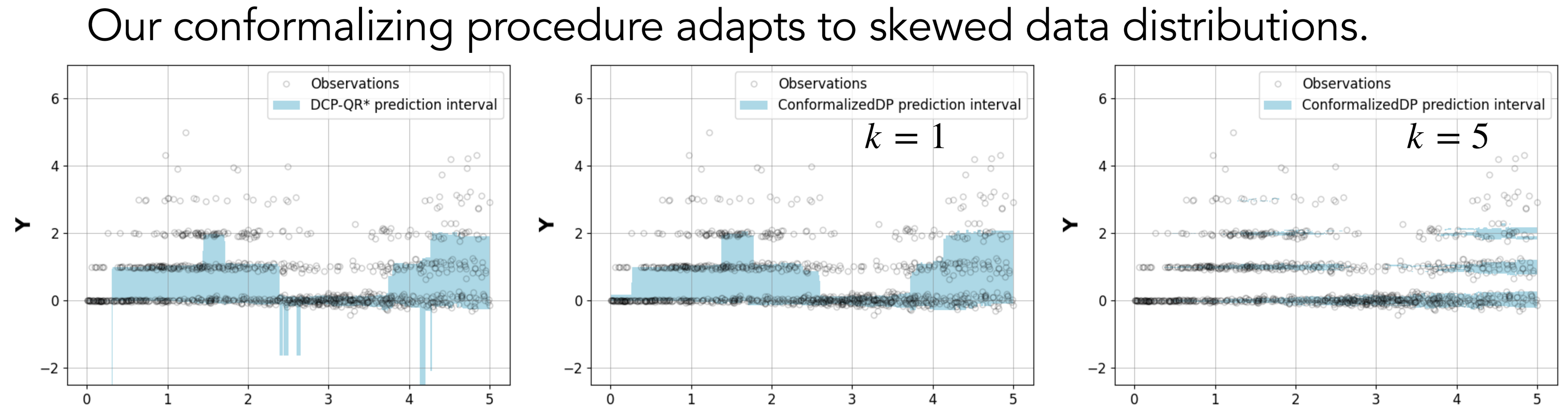


Figure 2: Performance of Conformalized KDE *(left)* and Conformalized DP *(right)* on mixture of Gaussian data for target coverage 0.8, with different parameter settings

Comparison to Conformalized Quantile Regression Methods

Our conformalizing procedure adapts to skewed data distributions.



Avg. vol.: 1.29 Avg. vol.: 1.14 Avg. vol: 0.45

Figure 3: *(Left)* DCP-QR* [Chernozhukov, Wüthrich, Zhu PNAS '21], *(Center)* Conformalized DP for $k = 1$, *(Right)* Conformalized DP for $k = 5$, on synthetic data from [Romano, Patterson, Candès NeurIPS '19]