



Generative Conformal Prediction with Optimized Coverage Allocation

Minxing Zheng and Shixiang Zhu

Carnegie Mellon University

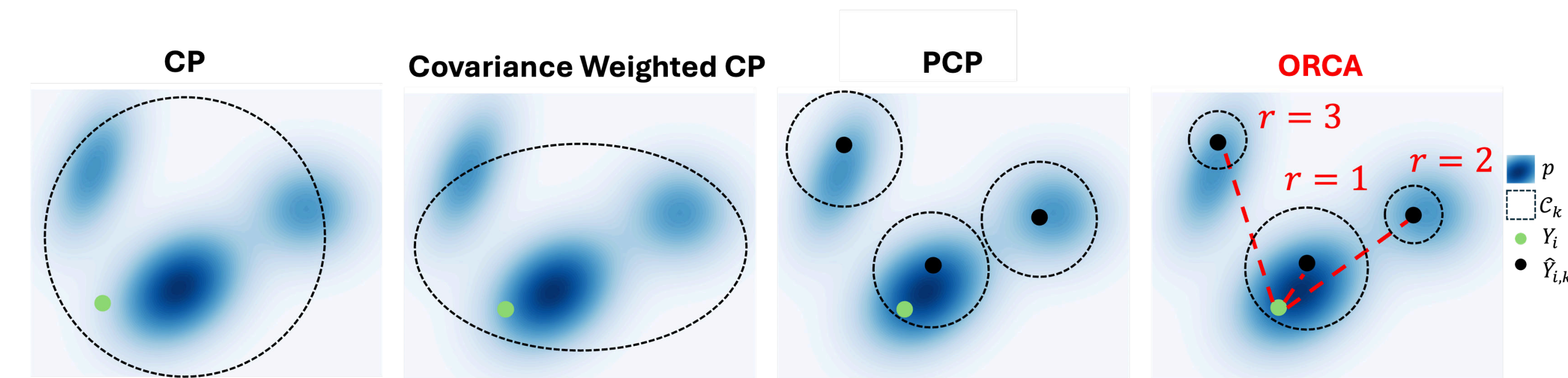
Problem Setup

Conformal Prediction: Given $(X_{\text{test}}, Y_{\text{test}}) \sim P$, the goal is to construct an uncertainty set satisfying $\hat{\mathcal{C}}(X_{\text{test}})$

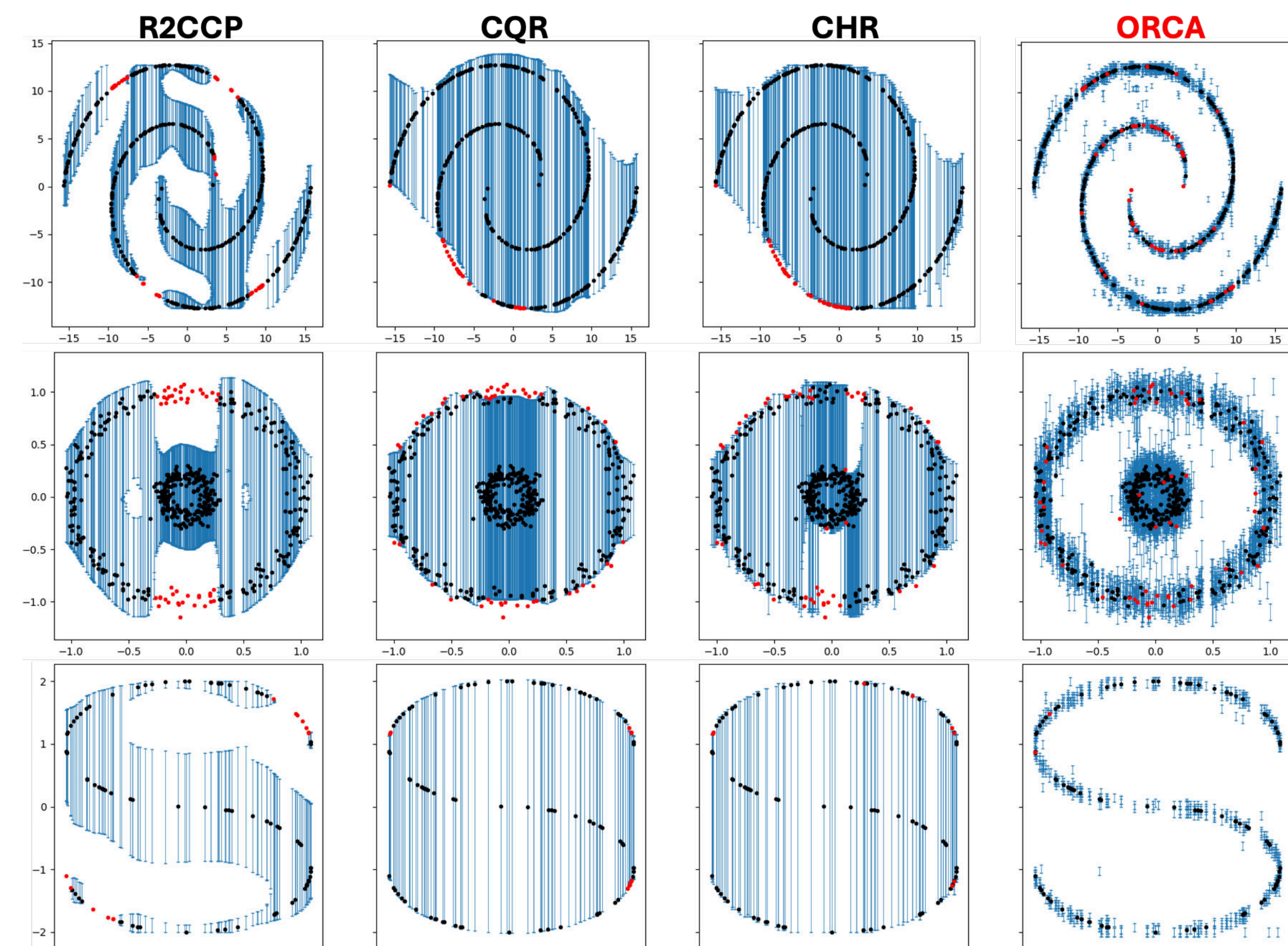
$$\mathbb{P}(Y_{\text{test}} \in \hat{\mathcal{C}}(X_{\text{test}})) \geq 1 - \alpha.$$

It provides distribution-free uncertainty quantification with finite-sample coverage guarantees. Efficiency is assessed by the expected set size $|\hat{\mathcal{C}}(X_{\text{test}})|$, motivating compact yet valid uncertainty sets.

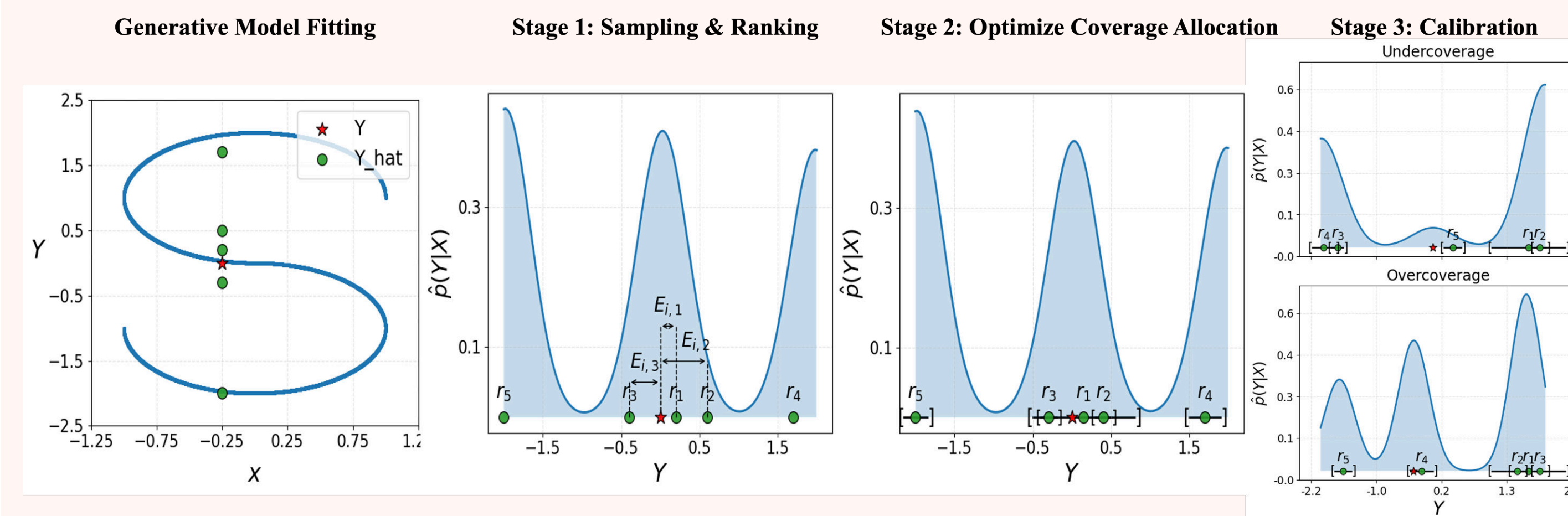
Generative Modeling: Unlike point estimation methods that predict means or quantiles of $Y|X$, generative models approximate the full conditional density $p(Y|X)$ through a learned sampler $\hat{p}(Y|X)$, it enables generative samples $\hat{Y}_{i,k} \sim \hat{p}(\cdot|X_i)$, $k = 1, \dots, K$, to capture complex distributional patterns in $Y|X$.



Synthetic Data Results



Method Overview: ORCA



Stage 1: Sampling & Ranking. For each $(X_i, Y_i) \in D$, where $D = D_{\text{explore}} \cup D_{\text{calib}}$, draw K samples $\{\hat{Y}_{i,k}\}_{k=1}^K \sim \hat{p}(\cdot | X_i)$. Rank the samples by their m -NN distances and denote the r -th ranked sample as $\hat{Y}_{i,j(r)}$. Define the ranked distance vector $E_i = (E_{i,1}, \dots, E_{i,K})$, where $E_{i,r} = \|Y_i - \hat{Y}_{i,j(r)}\|$.

Stage 2: Optimal Coverage Allocation. Use D_{explore} to find efficient rank-dependent quantile threshold vector $T(\alpha) = (Q_1(\beta_1), \dots, Q_K(\beta_K))$:

$$\min_{\beta} \sum_{r=1}^K (Q_r(\beta_r))^d \quad \text{s.t.} \quad \frac{1}{n_1} \sum_{i=1}^{n_1} \max_r \mathbb{1}\{E_{i,r} \leq Q_r(\beta_r)\} \geq 1 - \alpha,$$

where $Q_r(\beta_r)$ denotes the $1 - \beta_r$ -th sample quantile for $\{E_{i,r}\}_{i=1}^{n_1}$

Stage 3: Conformal Calibration. Use D_{calib} to calibrate the threshold vector $T(\alpha)$ obtained in Stage 2. Define $S_i := \min_{r \in [K]} \frac{E_{i,r}}{T_r(\alpha)}$, $i \in D_{\text{calib}}$ and let

$$R(\alpha) = t_{\alpha} T(\alpha), \quad t_{\alpha} = \min \left\{ \tau : \frac{1}{n_2+1} \sum_i \mathbb{1}\{S_i \leq \tau\} \geq 1 - \alpha \right\}.$$

Uncertainty set: $\hat{\mathcal{C}}_{\text{ORCA}}(X_{\text{test}}) = \bigcup_{r=1}^K \{y : \|y - \hat{Y}_{\text{test}, j(r)}\| \leq R_r(\alpha)\}$.

Theory & Guarantees

The optimization in Stage 2 admits an equivalent **mixed-integer linear program**:

$$\min_{z, y} \sum_{r, \ell} c_{r, \ell} z_{r, \ell} \quad \text{s.t.} \quad y_i \leq \sum_{r, \ell} a_{i, r, \ell} z_{r, \ell}, \quad \sum_i y_i \geq \tau, \\ z_{r, \ell} \geq z_{r, \ell+1}, \quad z_{r, \ell}, y_i \in \{0, 1\},$$

where $E_{(\ell), r}$ is order statistics, $z_{r, \ell} = 1$ if $E_{(\ell), r}$ is covered by the chosen threshold, $y_i = 1$ if sample i is covered, $c_{r, \ell} = E_{(\ell), r}^d - E_{(\ell-1), r}^d$, $a_{i, r, \ell} = \mathbb{1}\{E_{i, r} \leq E_{(\ell), r}\}$, and $\tau = \lceil n_1(1 - \alpha) \rceil$.

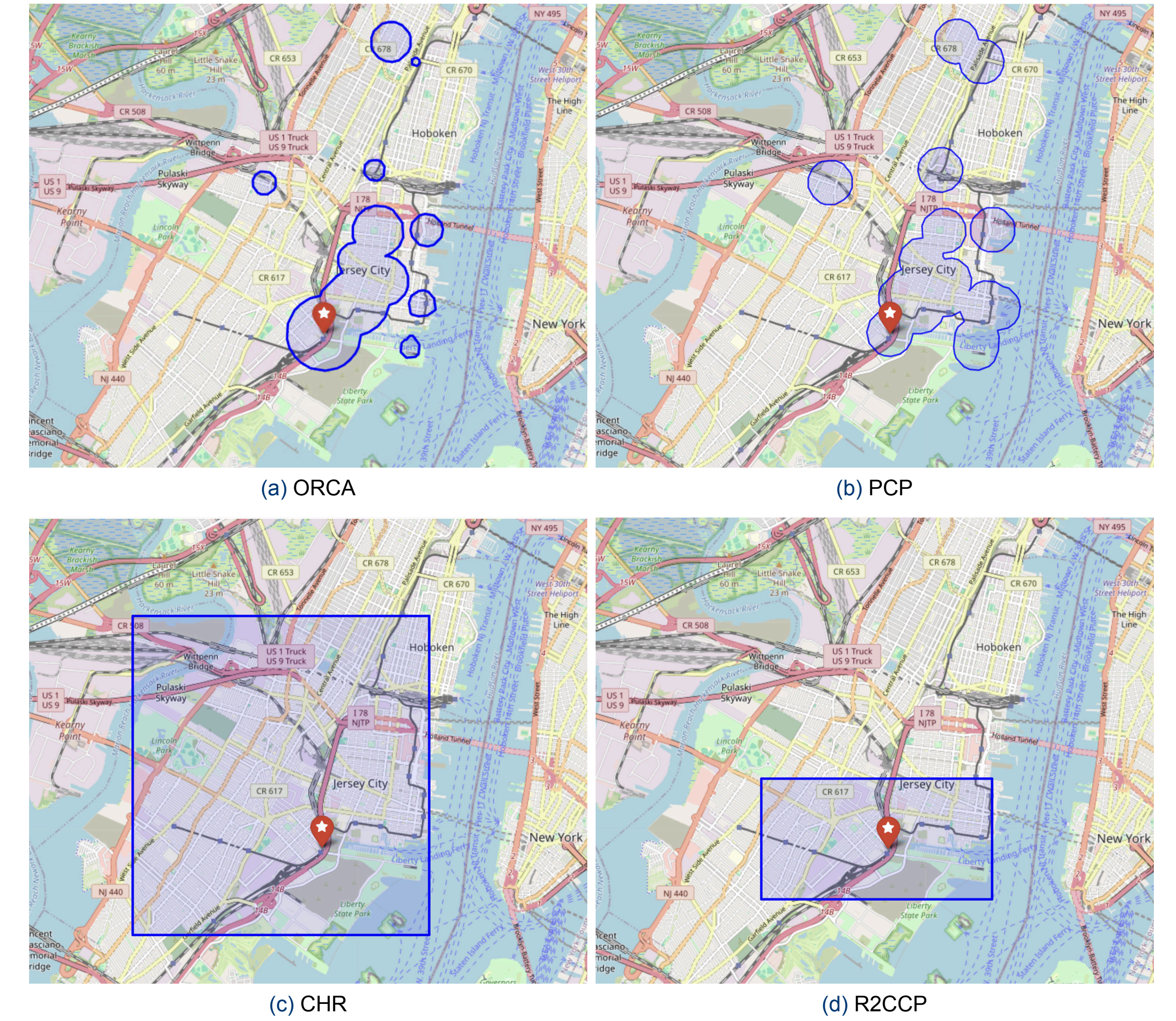
The proposed method achieves a **finite-sample coverage guarantee**,

$$\mathbb{P}(Y_{\text{test}} \in \hat{\mathcal{C}}_{\text{ORCA}}(X_{\text{test}})) \geq 1 - \alpha.$$

Moreover, as $K, n_1, n_2 \rightarrow \infty$, the uncertainty set converges to the oracle highest-density region:

$$\mu(\hat{\mathcal{C}}_{\text{ORCA}}(x) \triangle L_{\tau^*}(x)) \rightarrow 0, \quad L_{\tau^*}(x) = \{y : p(y|x) \geq \tau^*(x)\}.$$

Bike Sharing Destination Prediction



MNIST

