Personalized Federated Conformal Prediction with Localization

Yinjie Min[1], Chuchen Zhang[1], Liuhua Peng[†2], Changliang Zou[†1]

1 School of Statistics and Data Science, Nankai University
2 School of Mathematics & Statistics, The University of Melbourne

NeurIPS 2025

Outline

Introduction

- Problems arise as heterogeneous data distributions across agents in distributed systems (e.g., medical diagnostics, autonomous driving) require personalized modeling.
- Global models are often ineffective; PFL balances shared knowledge transfer and **agent-specific** adaptation.
- In risk-sensitive domains (e.g., autonomous vehicles, clinical models), reliable uncertainty quantification is critical for safety.

Background: Conformal Prediction

Given calibration data $\{(X_i, Y_i)\}_{i=1}^{n+1} \sim P$ i.i.d., construct prediction set $\widehat{C}_\alpha(X_{n+1})$ such that $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - \alpha$.

Limitation: Marginal coverage may hide severe miscoverage for certain subpopulations.

Test-conditional coverage

Test-conditional coverage quantifies **instance-specific** uncertainty

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \mid X_{n+1} = x)$$

Our objective is to imporve the test-conditional property (towards $1 - \alpha$) of conformal prediction for individual agents under the multi-agent heterogeneity setting of PFL.

Problem Setup

- $K+1$ agents, each with dataset $\mathcal{D}_k = \{(X_{k,i}, Y_{k,i})\}_{i=1}^{n_k} \sim P_k$
- Target agent: $P_{K+1} = P$, with $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{cal}}$
- Goal: Build $\widehat{C}_\alpha(X)$ for target agent using data from all agents without sharing raw data, such that

$$\mathbb{P}_{(X,Y)\sim P}(Y \in \widehat{C}_\alpha(X)) \geq 1 - \alpha\,,$$

and the test-conditional coverage of $\widehat{C}(X)$ given $X = x$ is closed to $1 - \alpha$.

Notations

- Let $s(x, y)$ and $\{s_k(x, y)\}_{k=1}^{K}$ be pretrained score functions.
- Let $S_{k,i} = s_k(X_{k,i}, Y_{k,i})$ for $i \in [n_k]$, and let $S_i = s(X_i, Y_i)$ for $i \in [n]$ be the scores. Define $S_i^y = S_i$ for $i \in [n]$, and $S_{n+1}^y = s(X_{n+1}, y)$.
- Let $f(x, s)$ and $f_k(x, s)$ denote the joint density functions of (X_1, S_1) and $(X_{k,1}, S_{k,1})$. Define weight $\pi_k = n_k / \sum_{\ell=1}^{K} n_\ell$ and the mixture density as $f_{\text{mix}}(x, s) = \sum_{k=1}^{K} \pi_k f_k(x, s)$
- Let $f(s \mid x), f_{\text{mix}}(s \mid x)$ denote the conditional PDF of $f(x, s), f_{\text{mix}}(s, x)$, and $F(s \mid x), F_{\text{mix}}(s \mid x)$ the corresponding conditional CDF.
- Define the density ratios as $r(x, s) = f(x, s)/f_{\text{mix}}(x, s)$

Generalized Localized Conformal Prediction (GLCP)

- Estimate conditional distribution $\widehat{F}(s \mid x)$ using engression or other estimators
- Define transformed scores $V_i^y = \widehat{F}(S_i^y \mid X_i)$
- Prediction set is defined as:

$$\widehat{C}_{\alpha}^{\mathsf{GLCP}}(X_{n+1}) = \{y : V_{n+1}^y \leq Q(1-\alpha; \{V_i^y\}_{i=1}^{n+1})\}$$

- GLCP is designed to *generalize* and *unify* existing approaches. Choosing $\widehat{F}(s \mid x)$ as the NW estimator, GLCP reduces to the Localized Conformal Prediction (LCP). Specifying $s(x, y) = y$, GLCP becomes equivalent to Distributional Conformal Prediction (DCP).

Personalized Federated Conformal Prediction (PFCP)

The key idea of PFCP is to derive an enhanced estimator of the conditional distribution by incorporating data from source agents.

- Estimate $\widehat{F}(s \mid x)$ locally (target agent)
- Estimate $\widehat{F}_{\text{mix}}(s \mid x)$ federatedly (source agents)
- Estimate density ratio $\widehat{r}(x, s)$ via federated classification
- Aggregate $\widehat{F}_{\text{agg}}(s \mid x)$ by

$$\{1 + \widehat{r}(x)\}^{-1} \left\{ \widehat{F}(s \mid x) + \widehat{r}(x) \cdot \widehat{F}_{\text{mix}}(s \mid x) \right\}$$

- Prediction set $\widehat{C}_{\alpha}^{\text{PFCP}}(X_{n+1})$ is defined as:

$$\left\{ y : \widehat{F}_{\text{agg}}(S_{n+1}^{y} \mid X_{n+1}) \leq Q(1-\alpha; \{\widehat{F}_{\text{agg}}(S_{i}^{y} \mid X_{i})\}_{i=1}^{n+1}) \right\}$$

Theoretical Guarantees

Theorem 1 (Marginal Validity)

Under i.i.d. data and independence of training/calibration:

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha^{PFCP}(X_{n+1})) \geq 1 - \alpha$$

Lemma 2

Under certain conditions, let F_V *be the continuous CDF of* $\widehat{F}(s(X_1, Y_1) \mid X_1)$. *Let* $q = \lceil (1-\alpha)(n+1) \rceil / n$. *For* $\delta \in (0,1)$,

$$\left| \mathbb{P}\left(Y_{n+1} \notin \widehat{C}_\alpha^{\mathrm{GLCP}}(X_{n+1}) \mid X_{n+1} = x \right) - \alpha \right| \leq \delta_1(x; \widehat{F}) + \max\left\{ |1 - \alpha - F_V^{-1}(q - \epsilon) + \delta|, |1 - \alpha - F_V^{-1}(q + \epsilon + n^{-1}) - \delta| \right\}$$

where $\epsilon = \{(2n)^{-1} \ln(2/\delta)\}^{1/2}$, $\delta_1(x; \widehat{F}) = d_{\mathrm{TV}}(\widehat{F}(\cdot \mid x), F(\cdot \mid x))$.

Theoretical Guarantees

Theorem 3

Define $\widetilde{\delta}(x;\widehat{r}) = \int_0^\infty f_{\mathrm{mix}}(s \mid x)|\widehat{r}(x,s) - r(x,s)|ds$, *and* $\delta_2(x;\widehat{F}_{\mathrm{mix}})$ *as the* L_2*-distance between the density of* $\widehat{F}_{\mathrm{mix}}(\cdot \mid x)$ *and* $F_{\mathrm{mix}}(\cdot \mid x)$, *and* $L_2(x;\widehat{r}) = \{\int_0^\infty \widehat{r}^2(s \mid x)\,ds\}^{-1/2}$. *Under certain conditions,* $L_2(x;\widehat{r}) < \infty$ *and* $2\delta_1(x;\widehat{F}_{\mathrm{agg}})\{1+\widehat{r}(x)\}$ *is bounded by*

$$2\delta_1(x;\widehat{F}) + \widetilde{\delta}(x;\widehat{r}) + |\widehat{r}(x) - r(x)| + L_2(x;\widehat{r})\delta_2(x;\widehat{F}_{\mathrm{mix}})\,.$$

If we have sufficient data from source agents, $\delta_2(x;\widehat{F}_{\mathrm{mix}}) \to 0$. Density ratio estimation is a binary classification task, which is easier than conditional distribution estimation. Taken together, these observations suggest $\widehat{F}_{\mathrm{agg}}(s \mid x)$ is better than $\widehat{F}(s \mid x)$, and thus PFCP is expected to outperform GLCP.

Experiments: Setup

- **Datasets:** Synthetic (S1/S2), BIO, BIKE, CRIME, STAR, CONCRETE, DERMA
- **Baselines:** GLCP, FedCP, FedCP-QQ, CPlab, CPhet
- **Metrics:** Marginal coverage, test-conditional miscoverage, prediction set size
- **Coverage target:** $1 - \alpha = 90\%$

Results: Marginal & Conditional Coverage (Synthetic)

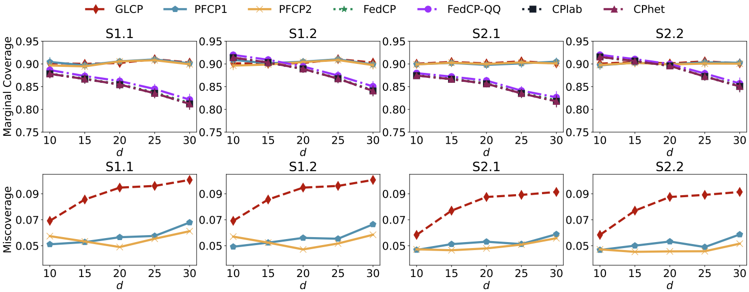

Figure 1: Marginal coverage under scenarios S1/S2. PFCP and GLCP achieve valid coverage; others fail. PFCP consistently outperforms GLCP in test-conditional miscoverage.

Real Data: Coverage and Set Size

Table 1: Marginal Coverage, miscoverage rate, and size of prediction sets.

	Dataset	GLCP	PFCP	FedCP	FedCP-QQ	CPlab	CPhet
Marginal	BIO	0.902	0.903	**0.993**×	**0.970**×	**0.978**×	**0.981**×
	BIKE	0.900	0.900	**0.885**×	0.890	**0.885**×	**0.883**×
	CRIME	0.900	0.896	**0.866**×	**0.852**×	**0.862**×	**0.864**×
	STAR	0.898	0.899	0.897	0.892	0.897	0.898
	CONCRETE	0.903	0.905	**0.947**×	**0.963**×	**0.949**×	**0.946**×
	DERMA	0.895	0.899	**0.824**×	**0.809**×	**0.880**×	**0.868**×
Miscoverage	BIO	0.0315	**0.0199**	0.0941	0.0753	0.0822	0.0844
	BIKE	0.0234	**0.0193**	0.0678	0.0647	0.0681	0.0690
	CRIME	0.0387	**0.0268**	0.0426	0.0495	0.0439	0.0429
	STAR	0.0392	**0.0244**	0.0502	0.0507	0.0498	0.0493
	CONCRETE	0.0366	**0.0238**	0.0582	0.0675	0.0600	0.0580
	DERMA	0.0300	**0.0230**	0.0884	0.0976	0.0616	0.0659
Size	BIO	0.5144	**0.5032**	0.9991	0.7087	0.8054	0.8271
	BIKE	4.0968	**3.9645**	4.0044	4.0910	3.9959	3.9671
	CRIME	5.1830	**4.4961**	3.9176	3.7724	3.8802	3.8855
	STAR	48.4987	43.2931	42.7556	**42.1185**	42.8081	43.0090
	CONCRETE	34.4859	**28.3797**	34.7115	38.9746	35.1545	34.6905
	DERMA	2.4926	**2.4430**	1.2490	1.1943	1.5649	1.4810

Figure 2: Real data results: both GLCP and PFCP achieve reliable marginal coverage close to the required $1 - \alpha = 90\%$ level, while others deviate significantly from the target in most scenarios. Across all scenarios, PFCP consistently outperforms GLCP.

Real Data: Number of Source Agents

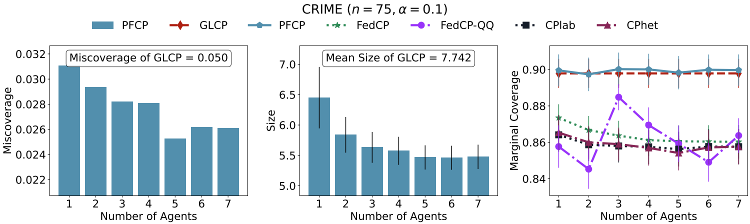

Figure 3: Target performance improves steadily as the number of source agents grows.

Thank You!