Conformal Prediction Meets Long-tail Classification

Shuqi Liu, Jianguo Huang, and Luke Ong

College of Computing and Data Science, Nanyang Technological University, Singapore

Introduction

- **Conformal Prediction (CP)** is a method for **uncertainty quantification** that converts a pretrained model's point prediction into a **prediction set**, with the set size reflecting the model's confidence.

$$X \xrightarrow{\text{Model } f} \underbrace{f(X)}_{\text{Point Prediction}} \xrightarrow{\text{CP}} \underbrace{\mathcal{C}(X)}_{\text{Prediction Set}}$$

- **Marginal Coverage Guarantee**: Assume the cal set $\{(X_i, Y_i)\}_{i=1}^n$ and test point (X_{n+1}, Y_{n+1}) are exchangeability, the following property holds for a user-defined level $1-\alpha$:

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

where prediction set is defined as $\mathcal{C}(X_{n+1}) = \{y : s(X_{n+1}, y) \leq \hat{\tau}_\alpha\}$.

  - **Non-conformity score** $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ quantifies the uncertainty of the pretrained model.
  - **Threshold** $\hat{\tau}_\alpha = \text{Quantile}\left(1-\alpha, \{s(X_i, Y_i)\}_{i=1}^n\right)$

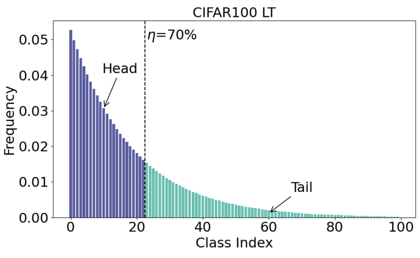
Motivation

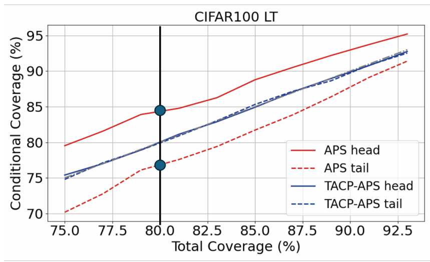
- Marginal coverage guarantees validity on average, while a stronger notion is **Class-conditional Coverage** [1]

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid Y_{n+1} = y) \geq 1-\alpha,$$

but most existing methods are **inefficient** and produce overly large prediction sets to achieve it.

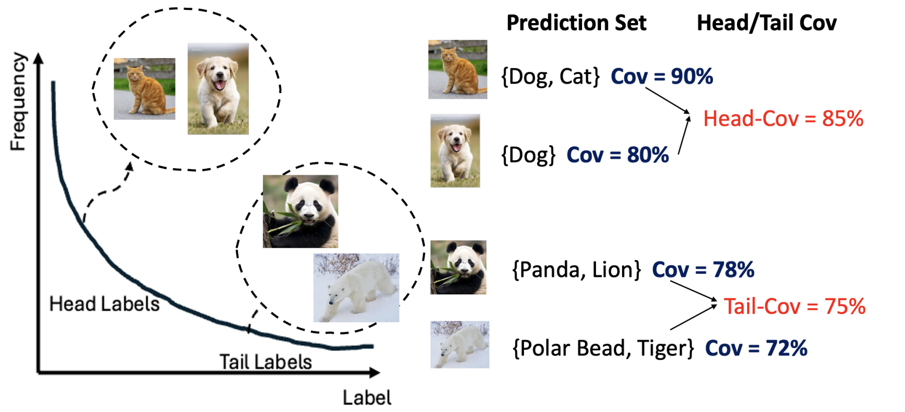
- Under **long-tailed label distributions**, existing CP methods exhibit **imbalance performance** across classes.





  - Head-/Tail-Cov $= \mathbb{P}(Y \in \mathcal{C}(X) \mid Y \in \text{Head/Tail})$
  - ★ Observation: Head-Cov$\geq 1-\alpha \geq$Tail-Cov, tending to over cover head classes at the expense of under covering tail classes.

- **Illustration:** Animal Classification; Total coverage at 80%.



!! Such under-coverage is concerning, as it undermines the reliability of the prediction sets for minority classes.

TACP Method

- **Key idea:** utilize the LT information and selectively penalize label rankings for head classes
- **Tail-Aware Conformal Prediction (TACP)**

$$s_{\text{TACP}}(x, y) = s(x, y) + \lambda \mathbb{I}(y \in \mathcal{G}_h)(o_x(y) - k_r)^+$$

where $\lambda \in \mathbb{R}_+$ and $k_r \in \mathbb{N}$ are hyperparameters, and $o_x(y)$ denotes the rank of class y at input x, ordered by the predicted class probabilities of model f, and the label prediction set:

$$\mathcal{C}_{\text{TACP}}(x_{n+1}) := \{y \in \mathcal{Y} : s_{\text{TACP}}(x_{n+1}, y) \leq \hat{\tau}_\alpha\}$$

- TACP is **score-agnostic**. E.g., TACP-LAC with $\lambda = 1$ and $k_r = 2$.

$$s_{\text{TACP-LAC}}(x, y) = \underbrace{1 - \hat{\pi}_y(x)}_{\text{LAC Score [2]}} + \underbrace{\mathbb{I}(y \in \mathcal{G}_h)(o_x(y) - 2)^+}_{\text{Selective Rank Regularization}}$$

Theoretical Results

Theorem 1.(Marginal Coverage Guarantee) If we assume additionally a uniform random variable u to ensure the scores $\{s_{\text{TACP}}(X_i, Y_i)\}_{i=1}^{n+1}$ are almost surely distinct, then

$$1-\alpha \leq \mathbb{P}(Y_{n+1} \in \mathcal{C}_{\text{TACP}}(X_{n+1})) \leq (1-\alpha) + \frac{1}{n+1}$$

Theorem 2.(Improved Coverage Gap) TACP narrows the Head-Tail coverage gap compared with STANDARD method when using the same non-conformity score.

$$\begin{aligned} P(Y_{n+1} \in & \mathcal{C}_{\text{TACP}}(X_{n+1}) \mid Y_{n+1} \in \mathcal{G}_h, E_{xy}) \\ & - P(Y_{n+1} \in \mathcal{C}_{\text{TACP}}(X_{n+1}) \mid Y_{n+1} \in \mathcal{G}_t, E_{xy}) \\ \leq P(Y_{n+1} \in & \mathcal{C}_{\text{STANDARD}}(X_{n+1}) \mid Y_{n+1} \in \mathcal{G}_h, E_{xy}) \\ & - P(Y_{n+1} \in \mathcal{C}_{\text{STANDARD}}(X_{n+1}) \mid Y_{n+1} \in \mathcal{G}_t, E_{xy}). \end{aligned}$$

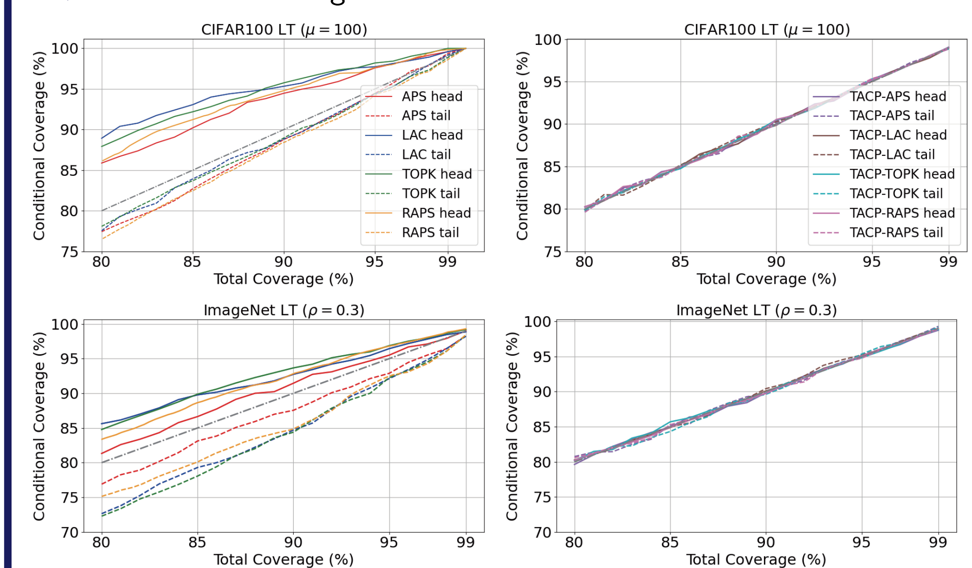
sTACP Method

- TACP effectively reduces the head–tail coverage gap, however, the indicator term in TACP enforces a head–tail partition, hindering general class-conditional coverage balance.
- We further proposed **Soft TACP (sTACP)**, which generalizes TACP by **continuously reweighting penalties** according to the estimated class prior $\hat{p}(y) \in [0, 1]$:

$$s_{\text{sTACP}}(x, y) = s(x, y) + \lambda \hat{p}(y)(o_x(y) - k_r)^+$$

Experimental Results

- Head-Tail Coverage Results



- Head-Tail Coverage Gap Results

Score	Method	$\rho = 0.6$		$\rho = 0.3$		$\mu = 100$		$\mu = 50$	
		CovGap-HT	AvgSize	CovGap-HT	AvgSize	CovGap-HT	AvgSize	CovGap-HT	AvgSize
APS	STA	2.18±1.21	36.43±2.21	2.79±0.85	35.29±2.29	4.81±0.89	9.93±0.45	3.87±0.81	8.58±0.32
	PW	2.01±1.29	38.10±2.65	1.43±1.20	38.35±2.63	1.38±1.19	10.28±0.43	1.22±0.84	8.79±0.25
	TACP	**1.11±0.71**	33.98±2.40	**0.76±0.60**	34.07±2.22	**0.78±0.64**	10.25±0.46	**0.76±0.59**	8.55±0.26
LAC	STA	7.45±1.33	15.36±0.96	7.50±0.73	14.81±1.00	6.35±1.01	7.11±0.29	4.88±0.91	6.16±0.26
	PW	1.96±1.28	17.22±1.30	1.25±1.25	17.72±1.20	1.65±1.18	7.44±0.35	1.56±1.13	6.36±0.25
	TACP	**1.18±0.94**	17.92±1.40	**0.81±0.67**	19.16±1.91	**0.77±0.59**	7.86±0.36	**0.66±0.52**	6.47±0.25
TOPK	STA	8.83±1.69	23.15±1.45	9.45±0.89	20.77±1.22	6.47±0.84	10.90±0.50	5.51±0.82	8.58±0.36
	PW	1.87±1.53	24.06±2.46	1.31±0.98	24.25±1.98	1.33±1.16	11.01±0.47	1.27±0.97	8.71±0.40
	TACP	**1.02±0.84**	25.63±2.44	**0.58±0.48**	26.26±2.24	**0.82±0.61**	11.88±0.57	**0.77±0.57**	9.38±0.39
RAPS	STA	7.76±1.25	16.49±1.30	7.34±0.71	15.98±1.04	6.34±0.80	7.67±0.52	4.58±0.75	6.62±0.18
	PW	1.91±1.43	20.32±1.56	1.27±1.00	20.57±1.48	1.49±1.16	8.93±0.50	1.24±0.88	6.48±0.23
	TACP	**1.10±0.87**	20.67±1.58	**0.61±0.41**	21.76±2.03	**0.82±0.70**	9.62±0.68	**0.80±0.58**	6.76±0.24

- Class-conditional Coverage Gap Results

Score	Method	$\rho = 0.6$			$\rho = 0.3$		
		CovGap	Coverage	AvgSize	CovGap	Coverage	AvgSize
APS	STA	19.00±0.00	89.78±0.01	35.60±2.27	13.37±0.35	90.07±0.61	38.39±1.54
	CLUS	17.84±0.00	89.66±0.01	37.54±1.43	13.41±0.00	90.62±0.00	42.80±1.28
	RC3P	18.99±0.61	89.31±0.65	30.20±1.88	14.21±0.31	89.45±0.45	43.50±2.22
	sTACP	**15.86±0.81**	89.53±1.12	36.52±3.66	**12.76±0.31**	90.04±0.61	35.48±2.08
LAC	STA	18.87±0.01	90.20±0.01	15.30±0.81	14.35±0.36	89.90±0.63	16.94±0.67
	CLUS	17.80±0.00	89.10±0.00	16.19±0.05	14.05±0.00	90.82±0.00	19.59±0.10
	RC3P	19.17±0.59	89.19±0.70	30.45±2.09	14.24±0.32	89.55±0.45	43.10±2.00
	sTACP	**15.98**±0.78	89.54±1.07	37.57±3.40	**12.66±0.28**	90.09±0.59	41.72±1.72
TOPK	STA	18.87±0.01	90.12±0.01	23.28±1.23	14.38±0.41	89.96±0.61	23.55±1.02
	CLUS	17.59±0.00	89.57±0.00	24.24±0.15	13.89±0.00	91.61±0.00	27.96±0.20
	RC3P	18.99±0.61	89.39±0.67	31.01±2.15	14.21±0.31	89.67±0.47	45.62±2.32
	sTACP	**15.83±0.77**	89.55±1.08	38.99±3.54	**12.80±0.28**	90.05±0.58	37.38±1.71
RAPS	STA	18.45±0.93	89.98±1.07	16.49±1.30	14.25±0.33	89.98±0.60	18.32±1.04
	CLUS	17.69±0.00	90.13±0.00	18.53±0.09	13.92±0.00	90.99±0.00	23.54±1.25
	RC3P	18.99±0.61	89.35±0.66	30.92±2.14	14.19±0.31	89.70±0.46	45.56±2.31
	sTACP	**16.04±0.82**	89.54±1.10	33.68±3.45	**12.90±0.31**	90.04±0.61	32.85±1.65

References

[1] V Vovk: Conditional validity of inductive conformal predictors. Mach. Learn. 2013.

[2] M Sadinle et al.: Least ambiguous set-valued classifiers with bounded error levels. J. Am. Stat. Assoc., 2019.

Paper