CONFORMAL PREDICTION FOR LONG-TAILED CLASSIFICATION

4/30/2026

Tiffany Ding, Jean-Bapiste Fermanian, Joseph Salmon

Motivation

- Long-tailed Data (skewed class distributions):
  - Plant identification
  - Disease Diagnosis
  - Biodiversity Monitoring
- Methods need to give:
  - Good class conditional coverage
  - Reasonable sized prediction sets
- Existing Methods on long-tailed data:
  - Poor class conditional coverage
  - Good class conditional coverage with a large size
- This work:
  - New conformal score function called prevalence adjusted softmax (optimizes for macro coverage)
  - New procedure that interpolates the thresholds between marginal and class-conditional coverage

Objective

- **Input:**
  - $X \in \mathcal{X}$ features
  - Unknown label $Y \in \mathcal{Y}$
- **Goal:** Construct a set-generating procedure with good class-conditional coverage: $\text{CondCov}(\mathcal{C}, y) = \mathbb{P}(Y \in \mathcal{C}(X) \mid Y = y)$
- **Landscape:**
  - Standard CP: Small sets but only guarantees marginal coverage and often has poor class-conditionalcoverage for some classes
  - Classwise CP and rank calibrated CP: Large Sets
  - Clustered CP: Defaults to Standard CP in rare classes

Table 1: Summary of the conformal methods considered. MargCov refers to the marginal coverage guarantee of the method.

	Score function	Threshold function	MargCov
STANDARD	any	$\hat{q}_{\text{STAND.}}$ (4)	$1-\alpha$
CLASSWISE	any	$\hat{q}_{\text{C.WISE}}$ (6)	$1-\alpha$
PAS*	s_{PAS} (11)	$\hat{q}_{\text{STAND.}}$ (4)	$1-\alpha$
WPAS*	s_{WPAS} (14)	$\hat{q}_{\text{STAND.}}$ (4)	$1-\alpha$
INTERP-Q*	any	$\hat{q}^{\text{IQ}}$ (15)	$1-2\alpha$

*Our methods

Recap: From the Class Conditional Conformal Paper

Standard CP

✅ Low variance

😕 No class-conditional coverage guarantee

Our method: Clustered CP

🔑 **Key idea:**

Combine data from classes that are "similar"

Classwise CP

😕 High variance

✅ Class-conditional coverage guarantee

$$C_{\text{CLUSTERED}}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{q}(\hat{h}(y))\}$$

where

- $\hat{h} : \mathcal{Y} \to \{1, \ldots, M\}$ is a clustering function
- $\hat{q}(m)$ is the conformal quantile computed using the calibration data in cluster m

Approaches

- **Target a relaxed version of class conditioned coverage:** Macro Coverage like macro accuracy (average over all classes):

$$\text{MacroCov}(\mathcal{C}) = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \mathbb{P}(Y \in \mathcal{C}(X) \mid Y = y) = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \text{CondCov}(\mathcal{C}, y).$$

We care equally about the coverage of all classes, but it is acceptable if a few classes have poor coverage, so long as the average class-conditional coverage is high

- **Target class-conditional coverage, then back off (until the set size is reasonable):** Interpolate between CLASSWISE CP and STANDARD CP by interpolating their quantile threshold

We want all classes to have good coverage

- User-defined parameter that can be tuned based on preference

Preliminaries: Conformal Prediction

Algorithm 1 Conformal prediction

Require: Score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, miscoverage level $\alpha \in [0, 1]$, calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, test point X_{n+1}, threshold function $\hat{\mathbf{q}} : \mathbb{R}^n \times [0, 1] \to \mathbb{R}^{|\mathcal{Y}|}$
Compute scores: $S_i \leftarrow s(X_i, Y_i)$ for $i = 1, \ldots, n$
Compute thresholds: $\mathbf{q} \leftarrow \hat{\mathbf{q}}((S_i)_{i=1}^n, \alpha)$
return Prediction set $\mathcal{C}(X_{n+1}) = \{y : s(X_{n+1}, y) \leq q_y\}$, where q_y is the y-th entry of $\mathbf{q}$

Conformal prediction as thresholded sets. Let $\mathbf{q} = (q_1, q_2, \ldots, q_{|\mathcal{Y}|})$ be a $|\mathcal{Y}|$-dimensional vector of score thresholds. Define the $\mathbf{q}$-*thresholded set* as

$$\mathcal{C}(X; \mathbf{q}) = \{y \in \mathcal{Y} : s(X, y) \leq q_y\}. \tag{3}$$

- CP finds principled ways to set q as a function of the calibration data

STANDARD *conformal prediction* constructs sets as $\mathcal{C}_{\text{STAND.}}(X) := \mathcal{C}(X; \hat{\mathbf{q}}_{\text{STAND.}})$ for $\hat{\mathbf{q}}_{\text{STAND.}} = (\hat{q}, \ldots, \hat{q})$ where

$$\hat{q} = \text{Quantile}_{1-\alpha}\left(\frac{1}{n+1}\sum_{i=1}^{n}\delta_{s(X_i,Y_i)} + \frac{1}{n+1}\delta_{\infty}\right), \tag{4}$$

CLASSWISE *conformal prediction* constructs sets as $\mathcal{C}_{\text{CLASSWISE}}(X) := \mathcal{C}(X; \hat{\mathbf{q}}_{\text{CLASSWISE}})$ where the y-th entry of $\hat{\mathbf{q}}_{\text{CLASSWISE}}$ is

$$\hat{q}_y^{\text{CW}} = \text{Quantile}_{1-\alpha}\left(\frac{1}{n_y+1}\sum_{i\in\mathcal{I}_y}\delta_{s(X_i,Y_i)} + \frac{1}{n_y+1}\delta_{\infty}\right). \tag{6}$$

CLASSWISE conformal prediction sets achieve a *class-conditional coverage* guarantee (Vovk et al., 2005):

$$\mathbb{P}(Y \in \mathcal{C}(X; \hat{\mathbf{q}}_{\text{CLASSWISE}}) \mid Y = y) \geq 1 - \alpha, \qquad \text{for all } y \in \mathcal{Y}. \tag{7}$$

Approach 1: Targeting Macro-Coverage

- Minimize expected set size subject to macro coverage constraint

$$\min_{\mathcal{C}:\mathcal{X}\mapsto 2^{\mathcal{Y}}} \mathbb{E}[|\mathcal{C}(X)|] \quad \text{subject to } \text{MacroCov}(\mathcal{C}) \geq \beta, \tag{8}$$

- Or its equivalent dual:

$$\max_{\mathcal{C}:\mathcal{X}\mapsto 2^{\mathcal{Y}}} \text{MacroCov}(\mathcal{C}) \quad \text{subject to } \mathbb{E}[|\mathcal{C}(X)|] \leq \kappa \tag{9}$$

Proposition 1 (Informal). *The solutions of* (8) *and* (9) *are of the form*

$$\mathcal{C}^*(x) = \{y \in \mathcal{Y} : p(y|x)/p(y) \geq t\},$$

- This proposition says that thresholding on this quantity optimally balances macro-coverage and expected set size
- While access to exact probabilities is not present the numerator can be obtained from the classifier and the denominator is the label distribution

Approach 1: Cont

- By creating prediction sets $\widehat{\mathcal{C}}(x) = \{y \in \mathcal{Y} : \hat{p}(y|x)/\hat{p}(y) \geq t\}$, approximate an oracle Pareto optimal set
- Choose t to achieve marginal coverage guarantee: Rewrite

$$\widehat{\mathcal{C}}(x) = \{y \in \mathcal{Y} : s_{\mathsf{PAS}}(x, y) \leq -t\} \quad \text{where} \quad s_{\mathsf{PAS}}(x, y) = -\hat{p}(y|x)/\hat{p}(y)$$

- Method:
  - Run STANDARD CP with PAS score function
  - Achieves the desired marginal coverage guarantee while (approximately) optimally trading off set size and macro-coverage
- Weighted Macro Coverage:

$$\mathrm{MacroCov}_\omega(\mathcal{C}) = \sum_{y \in \mathcal{Y}} \omega(y) \mathbb{P}(Y \in \mathcal{C}(X) \mid Y = y).$$

Proposition 2 (Informal). *The solutions of (8) and (9) when* MacroCov *is replaced with* $\mathrm{MacroCov}_\omega$ *are of the form*

$$\mathcal{C}^*(x) = \{y \in \mathcal{Y} : \omega(y) \cdot p(y|x)/p(y) \geq t\}, \tag{13}$$

for some threshold t *that depends on* ω *and* β *or* κ*, respectively.*

$$s_{\mathsf{WPAS}}(x, y) := -\omega(y)\frac{\hat{p}(y|x)}{\hat{p}(y)},$$

Approach 2: Interpolating between classwise and Standard CP

- Linearly interpolating their quantile thresholds

$$\mathcal{C}_{\text{INTERP-Q}}(X) := \mathcal{C}(X; \hat{\mathbf{q}}_{\text{IQ}})$$

- yth entry of the quantile is the weighted average of standard and classwise CP:

$$\hat{q}_y^{\text{IQ}} = \tau \hat{q}_y^{\text{CW}} + (1-\tau)\hat{q} \qquad \text{for all } y \in \mathcal{Y}.$$

Proposition 3. *If $\hat{q}$ and $\hat{q}_y^{\text{CW}}$ (for $y \in \mathcal{Y}$) are the* STANDARD *and* CLASSWISE *conformal quantiles for $\alpha \in [0,1]$, then $\mathcal{C}_{\text{INTERP-Q}}$ achieves a marginal coverage of at least $1-2\alpha$.*

Experiments

- Datasets:
  - Pl@ntNet-300k
  - iNaturalist-2018

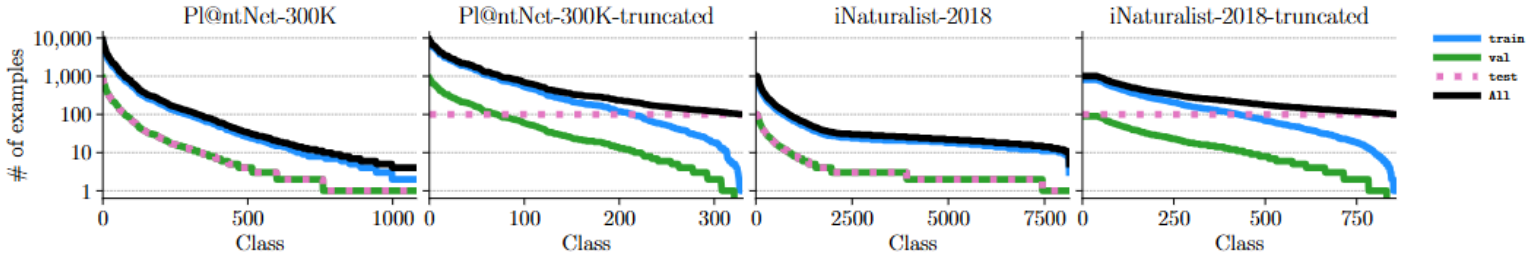


Figure 2: Class distributions (sorted by prevalence), plotted using a logarithmic scale, of the classical `train`, `val`, and `test` sets in the datasets we experiment on. We further randomly split 30% of `val` to use for model validation and use the remaining 70% as the calibration set $\mathcal{D}_{\text{cal}}$. We use the truncated versions to obtain good estimates of class-conditional metrics.

- Truncated versions retain classes with ≥101 test examples (100 each). Used for reliable per-class evaluation. 69% of Pl@ntNet and 90% of iNaturalist classes have fewer than 10 test examples!
- Metrics:
  - FracBelow50%: fraction of classes with coverage < 50% (smaller is better)
  - UnderCovGap: mean max(1−α − $\hat{c}_\gamma$, 0) across classes (smaller is better)
  - MacroCov — average class-conditional coverage (larger is better)
  - MarginalCov — overall coverage (should be ≥ 0.9)
  - Avg set size — mean |C(X)| (smaller is better)

Result

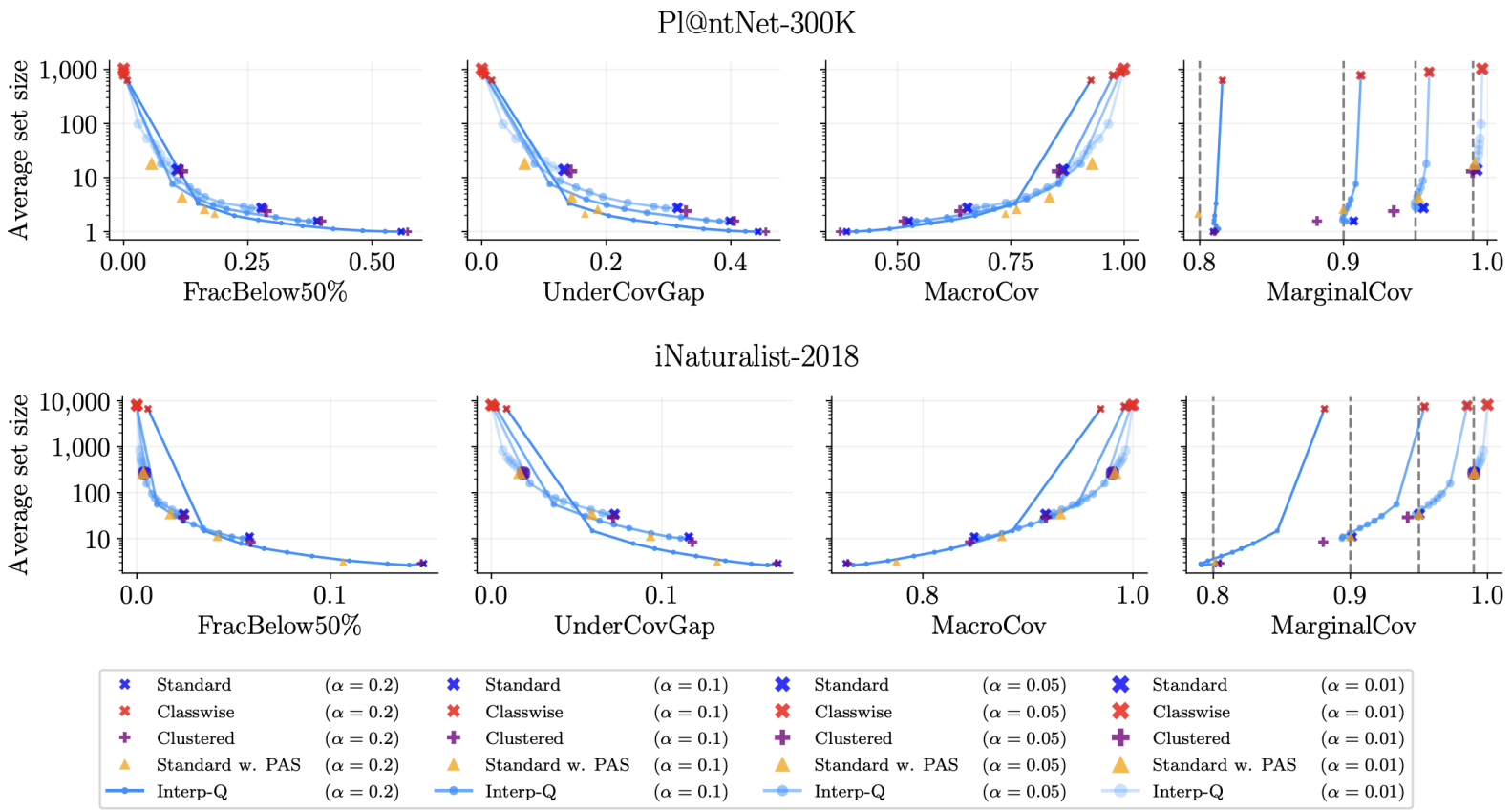
Table 3: Set size and coverage metrics for Pl@ntNet-300K using the s_{softmax} score and $\alpha = 0.1$. The arrows next to the coverage metric names indicate whether it is better for the metric to be smaller (↓) or larger (↑).

Method	**FracBelow50%** ↓	**UnderCovGap** ↓	**MacroCov** ↑	**MarginalCov** (desired ≥ 0.9)	**Avg. set size** ↓
STANDARD	0.389	0.398	0.525	0.907	1.57
CLASSWISE	0.000	0.006	0.976	0.912	780.00
CLUSTERED	0.398	0.406	0.513	0.882	1.57
STANDARD W. PAS	0.167	0.193	0.755	0.902	2.57
INTERP-Q ($\tau = 0.9$)	0.248	0.265	0.671	0.901	2.24
INTERP-Q ($\tau = 0.99$)	0.151	0.168	0.785	0.905	3.95
INTERP-Q ($\tau = 0.999$)	0.098	0.109	0.856	0.908	7.58

Table 4: Set size and coverage metrics for iNaturalist-2018 using the s_{softmax} score and $\alpha = 0.1$.

Method	**FracBelow50%** ↓	**UnderCovGap** ↓	**MacroCov** ↑	**MarginalCov** (desired ≥ 0.9)	**Avg. set size** ↓
STANDARD	0.058	0.116	0.849	0.902	10.9
CLASSWISE	0.000	0.002	0.992	0.954	7430.0
CLUSTERED	0.059	0.118	0.845	0.880	8.4
STANDARD W. PAS	0.042	0.093	0.875	0.900	11.3
INTERP-Q ($\tau = 0.9$)	0.034	0.081	0.891	0.907	16.8
INTERP-Q ($\tau = 0.99$)	0.020	0.055	0.924	0.923	31.1
INTERP-Q ($\tau = 0.999$)	0.010	0.037	0.947	0.934	55.8

Result: Set size vs Notions of coverage



- When targeting set size and class-conditional or macro-coverage, it is more effective to optimize for this trade-off directly than trading off set size and marginal coverage
- CLASSWISE should generally be avoided, as comparable class-conditional and macro-coverage can be achieved with significantly smaller sets using our proposed methods
- INTERP-Q produces reasonable set sizes even for large values of τ

Case Study: Coverage on Endangered Plant Species

Species: *Metasequoia glyptostroboides Hu & W.C.Cheng*
# of examples: 410

Method	Coverage	Size
Standard	0.00	1.3
Classwise	1.00	781.3
Std w. PAS	1.00	7.3

Species: *Vanilla planifolia Jacks. ex Andrews*
# of examples: 35

Method	Coverage	Size
Standard	0.60	3.2
Classwise	1.00	782.2
Std w. PAS	0.80	10.0

Species: *Abeliophyllum distichum Nakai*
# of examples: 4

Method	Coverage	Size
Standard	0.00	4.0
Classwise	1.00	784.0
Std w. PAS	0.00	6.0